Matemaattinen tilastotiede

Erkki Liski
Matematiikan, Tilastotieteen ja Filosofian Laitos
Tampereen Yliopisto

1. 9. 2004

Sisältö

Luku 1

Johdanto

1.1 Todennäköisyys ja tilastotiede

Tämä kurssi käsittelee sekä todennäköisyyslaskentaa että tilastotiedettä. Uhkapelurien ongelmat inspiroivat todennäköisyyslaskennan uranuurtajien ajattelua, mutta nykyisin todennäköisyyslaskennan sovellusalue on erittäin monipuolinen ja jatkuvasti laajeneva. Tilastotieteessä laaditaan satunnaisilmiöille todennäköisyysmalleja ja tutkitaan sitten havaintojen perusteella, miten hyvin mallit kuvaavat todellisuutta.

1.2 Havaitut frekvenssit ja empiiriset jakaumat

Jatkossa käytämme termiä *koe* tai *satunnaiskoe,* kun puhumme menettelystä tai prosessista, joka tuottaa (generoi) havaintoja. Esimerkkejä satunnaiskokeista ovat lantin heitto tai kännykkään tulevien viestien lukumäärä seuraavan tunnin aikana. Heitetään lanttia esimerkiksi 100 kertaa ja saadaan 56 klaavaa (L). Tapahtuman 'klaava' frekvenssi 100:n heiton sarjassa on tässä tapauksessa 56 ja suhteellinen frekvenssi $56/100 = 0.56$. Merkitään tapahtuman A lukumäärää eli frekvenssiä n:n kokeen sarjassa $N_n(A)$. Useimmissa sovelluksissa näyttää käyvän niin, että suhteellinen frekvenssi

$$\frac{N_n(A)}{n} \quad \text{lähenee lukua} \quad P(A), \tag{1.2.1}$$

kun toistojen lukumäärä n kasvaa. On helppo todeta, että $0 \leq P(A) \leq 1$. Tätä lukua $P(A)$ kutsumme tapahtuman A todennäköisyydeksi.

Vaikka emme olekaan vielä määritelleet todennäköisyyttä, voimme todeta, että suhteellinen frekvenssi on ominaisuuksiltaan todennäköisyyden kaltainen ja antaa siksi hyvän intuitiivisen käsityksen todennäköisyydestä. Suhteellisen frekvenssin avulla voidaan myös arvioida todennäköisyyksiä numeerisesti. Näin tehdään esimerkiksi simulointikokeissa. Huomattakoon, että

suhteellinen frekvenssi ei ole todennäköisyyden määritelmä vaan todennäköisyyden eräs tulkinta. Todennäköisyys määritellään aksiomaattisesti. Kun todennäköisyys on määritelty, seuraa tulos (1.2.1) näistä aksioomeista. Itse asiassa (1.2.1) voidaan perustella *vahvan suurten lukujen lain* avulla. Se on tilastotieteen kannalta yksi todennäköisyyslaskennan tärkeimpiä lauseita.

Olkoon $x_1, x_2, \ldots, x_n$ jokin lukujono. Tavallisesti nämä luvut $x_1, x_2, \ldots, x_n$ ovat jonkin suureen, kuten esimerkiksi pituuden tai painon, mittalukuja. Jos esimerkiksi n tilastoyksikköä on mitattu, niin silloin x_i on i. tilastoyksikön mittaluku ja luvut x_1, x_2, …, x_n muodostavat havaintoaineiston. Lukujen x_1, x_2, …, x_n (havaintoaineiston) *empiirinen kertymäfunktio* (ekf) reaalilukuakselilla $(-\infty, \infty)$ on

$$F_n(a) = \frac{1}{n}|\{\, i : 1 \le i \le n,\, x_i \le a \,\}|,$$

missä $-\infty < a < \infty$ ja $|\,.\,|$ on joukon alkioiden lukumäärä.

Lukujen $x_1, x_2, \ldots, x_n$ *empiirinen jakaumafunktio* tai lyhyesti *empiirinen jakauma* (ej) on

$$P_n(a, b) = F_n(b) - F_n(a).$$

$P_n(a, b)$ on siis puoliavoimelle välille $(a, b]$ kuuluvien lukujen suhteellinen osuus lukujoukossa $\{x_1, x_2, \ldots, x_n\}$:

$$P_n(a, b) = \frac{1}{n}|\{\, i : 1 \le i \le n,\, a < x_i \le b \,\}|.$$

Esimerkki 1.1 Olkoon hatussa n arpalippua ja i. lippuun on kirjoitettu luku x_i. Valitaan hatusta satunnaisesti yksi arpa. Silloin todennäköisyys, että arvan numero sattuu välille $(a, b]$ on $P_n(a, b)$. Tässä tilanteessa empiiriselle jakaumalle voidaan siis antaa todennäköisyystulkinta. □

Empiirisen jakauman kuvaajana käytetään tavallisesti histogrammia. Histogrammin piirtäminen aloitetaan valitsemalla ensin *jakopisteet* $b_1 < b_2 < \cdots < b_m$ siten, että kaikki luvut x_i sisältyvät avoimelle välille (b_1, b_m) ja mikään jakopiste ei ole mittaluku. Jakopisteet määrittelevät $m - 1$ osaväliä (b_j, b_{j+1}), $1 \le j \le m - 1$. Histogrammi piirretään asettamalla vierekkäin $m - 1$ pylvästä (suorakaidetta) siten, että j. pylvään kannan (luokan) leveys on $b_{j+1} - b_j$ ja pylvään korkeus on

$$\frac{P_n(b_j, b_{j+1})}{(b_{j+1} - b_j)} = \frac{|\{\, i : 1 \le i \le n,\, b_j < x_i < b_{j+1} \,\}|}{n(b_{j+1} - b_j)}.$$

Korkeus on siis j. osaväliin kuuluvien *havaintojen suhteellinen osuus pituusyksikköä kohti.* Pylvään korkeutta kutsutaan *havaintotiheydeksi* tai lyhyesti *tiheydeksi.* Vastaavasti j. *pylvään pinta-ala* on $P_n(b_j, b_{j+1})$ ja kaikkien pylväiden yhteenlaskettu pinta-ala on 1.

Käytännön sovelluksissa mittaustarkkuus on aina äärellinen, sanokaamme Δx. Jokainen mittaluku on silloin muotoa *kokonaisluku* $\cdot\, \Delta x$. Kahden

mittaluvun pienin mahdollinen erotus on Δx. Jakopisteet valitaan siten, että ne ovat muotoa

$$kokonaisluku \cdot \Delta x + \frac{\Delta x}{2}.$$

Silloin jakopiste ei voi olla mittaluku. Jakopisteet muodostavat aineistoon *luokituksen* ja puhumme silloin *luokitellusta aineistosta.* Jakopisteet b_j, b_{j+1} ovat silloin j. luokan ns. *todelliset luokkarajat* ja pisteet $b_j + \frac{\Delta x}{2}$, $b_{j+1} - \frac{\Delta x}{2}$ ovat ns. pyöristetyt luokkarajat.

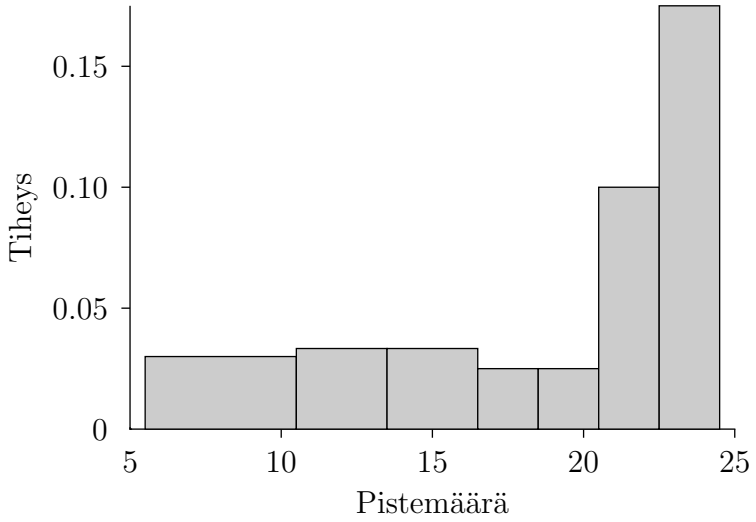
Esimerkki 1.2 Kurssin 1. välikokeen pistemäärät $x_i, 1 \leq i \leq 20$ olivat

18, 12, 14, 11, 24, 14, 24, 22, 24, 10, 8, 19, 21, 22, 24, 24, 24, 6, 24, 21.

Kokeeseen osallistui siis 20 opiskelijaa. Valitaan todellisiksi luokkarajoiksi

5.5, 10.5, 13.5, 16.5, 18.5, 20.5, 22.5, 24.5.

Nyt siis $b_1 = 5.5$ ja $b_8 = 24.5$. Luokkarajat määrittelevät 7 luokkaa.



Kuvio 1.1. Koepistemäärän histogrammi ($n = 20$).

Esimerkiksi $P_{20}(20.5, 22.5) = \frac{4}{20} = 0.2$ ja havaintotiheys luokassa $(20.5, 22.5)$ on

$$\frac{P_{20}(20.5, 22.5)}{22.5 - 20.5} = \frac{0.2}{2} = 0.1.$$

□

1.3 Todennäköisyysmallit

1.3.1 Satunnaiskoe

Todennäköisyyslaskenta on *satunnaisilmiöiden* matemaattista teoriaa. Kun tarkastelemme satunnaisilmiöitä, puhumme *satunnaiskokeista,* vaikka kyse

on tavallisesti vain ajatelluista satunnaiskokeista. Se on siis matemaattinen abstraktio. Satunnaiskokeessa on oletuksena, että kokeen alkutila ei määritä tulosta deterministisesti, vaan väliintuleva tekijä, sattuma, vaikuttaa kokeen tulokseen. Satunnaiskokeen mahdolliset tulosvaihtoehdot tiedetään, mutta yksittäisen kokeen tulosta ei voida varmuudella ennustaa. Ainoa tapa saada tietoa satunnaisilmiöistä on tehdä satunnaiskokeita (eli havainnoida satunnaisilmiöitä).

Oletetaan nyt, että koe (ilmiö) on sellainen, että sen tulos ei ole varmuudella ennustettavissa, mutta kaikki mahdolliset tulosvaihtoehdot ovat tiedossa. Jos tällainen koe voidaan toistaa samoissa olosuhteissa, sitä kutsutaan satunnaiskokeeksi. Satunnaiskokeen kaikkien mahdollisten tulosten joukkoa kutsutaan *otosavaruudeksi* ja merkitään Ω:lla. Satunnaiskokeen yksittäistä mahdollista tulosta kutsutaan *alkeistapaukseksi* (satunnaiskokeeseen liittyvän otosavaruuden Ω yksi piste). Jos otosavaruus on äärellinen, merkitään

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\},$$

missä alkeistapaukset ovat ω_1, ω_2, $\ldots$, ω_n ja Ω:n alkeistapausten lukumäärä $|\Omega| = n$. Otosavaruus voi olla myös ääretön.

Tapahtuma on otosavaruuden Ω osajoukko. Otosavaruuden osajoukkoja merkitään isoilla kirjaimilla $A, B, C, \ldots$ Sanomme, että tapahtuma A sattuu, jos kokeen tulos ω kuuluu joukkoon A eli $\omega \in A$. Ω on ns. *varma tapahtuma,* koska jokin mahdollisista vaihtoehdoista sattuu varmasti.

Esimerkki 1.3 Heitetään lanttia. Tulosvaihtoehdot ovat klaava (L) ja kruunu (R), joten otosavaruus $\Omega = \{\mathrm{L}, \mathrm{R}\}$ ja $|\Omega| = 2$.

Heitetään lanttia, kunnes saadaan ensimmäinen klaava. Silloin otosavaruus

$$\Omega = \{\mathrm{L}, \mathrm{RL}, \mathrm{RRL}, \mathrm{RRRL}, \ldots\}$$

ja $|\Omega| = \infty$. Jos tapahtuma A on 'enintään kaksi kruunua ennen 1. klaavaa', niin $A = \{\mathrm{L}, \mathrm{RL}, \mathrm{RRL}\}$.

Olkoon $\omega > 0$ laitteen kestoikä (tunteina). Tällöin

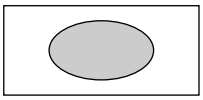
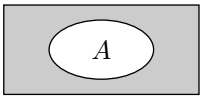
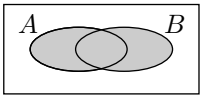
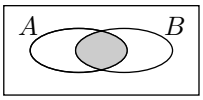
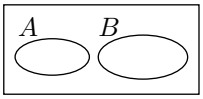
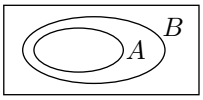
$$\Omega = \{\, \omega \in \mathbb{R} \mid \omega > 0 \,\}.$$

Esimerkiksi tapahtuma 'kestoikä ainakin 100 tuntia' on $[100, \infty)$ ja 'kestoikä yli 150, mutta korkeintaan 200 tuntia' on $(150, 200]$. □

1.3.2 Joukko-operaatiot

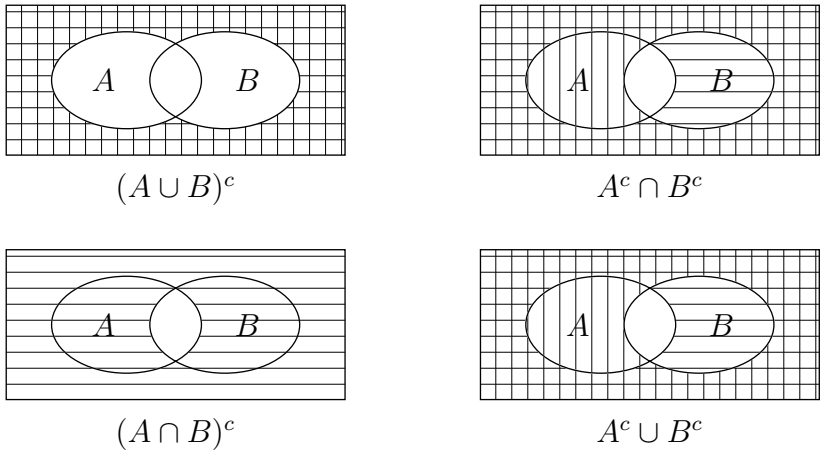
Oletetaan, että satunnaiskokeen $\mathcal{E}$ otosavaruus Ω on annettu. Kaikki tarkastelun kohteena olevat tapahtumat esitetään Ω:n osajoukkoina. Olkoon A tapahtuma. Jos A sattuu, se tarkoittaa, että kokeen $\mathcal{E}$ tulos ω kuuluu joukkoon A eli $\omega \in A$. Tulkitse Vennin diagrammi siten, että valitset suorakaiteesta (Ω:sta) satunnaisesti pisteen. Jokainen suorakaiteen piste on alkeistapaus. Jokainen suorakaiteen osa-alue on tapahtuma.

Taulukko 1.1. Joukko-opillisen ja todennäköisyyslaskennan terminologian vastaavuus.

Tapahtumat	Joukot	Joukkojen merkintä	Vennin diagrammi
otosavaruus	perusjoukko	Ω	
tapahtuma	Ω:n osajoukko	A, B, C jne.	
mahdoton tapahtuma	tyhjä joukko	$\emptyset$	
ei A, A ei satu	A:n komplementti	A^c	
joko A tai B tai molemmat	A:n ja B:n yhdiste	$A \cup B$	
sekä A että B	A:n ja B:n leikkaus	AB, $A \cap B$	
A ja B toisensa poissulkevat	A ja B pistevieraat	$A \cap B = \emptyset$	
jos A niin B	A on B:n osajoukko	$A \subset B$	

Taulukossa 1.1 on esitetty joukko-opilliset operaatiot *komplementti, yhdiste* ja *leikkaus.* Nämä operaatiot toteuttavat ns. *De Morganin lait:*

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$



Kuvio 1.2. De Morganin lait.

Kaksinkertaisen komplementin sääntö

$$(A^c)^c = A$$

on myös usein käyttökelpoinen.

Joukkojen A ja B *erotukseen* $A \setminus B$ kuuluvat ne A:n pisteet, jotka eivät kuulu joukkoon B:

$$A \setminus B = A \cap B^c = \{\, \omega \mid \omega \in A \text{ ja } \omega \notin B \,\}.$$

Jos $B \subset A$, käytämme merkinnän $A \setminus B$ sijasta myös merkintää $A - B$. Tätä merkintää käyttäen

$$A \setminus B = A - (A \cap B)$$

ja

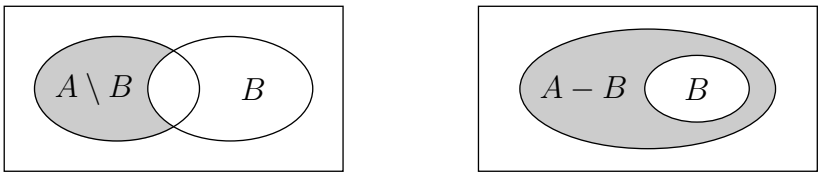
$$A^c = \Omega - A.$$

Sanomme, että tapahtumat A_1, A_2, …, A_m muodostavat tapahtuman A *osituksen* (tai jaon), jos $A = A_1 \cup A_2 \cup \cdots \cup A_m$ ja tapahtumat A_1, A_2, …, A_m ovat toisensa poissulkevat ($A_i \cap A_j = \emptyset$, kun $i \neq j$). Esimerkiksi A, A^c muodostaa otosavaruuden Ω osituksen ja $A \setminus B$, $A \cap B$ muodostaa A:n osituksen. Jos joukot A ja B ovat pistevieraat ($A \cap B = \emptyset$), niin voimme merkinnän $A \cup B$ sijasta käyttää merkintää $A + B$. Silloin esimerkiksi

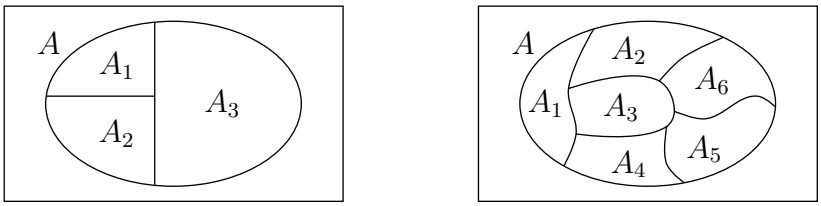
$$\Omega = A + A^c.$$

Jos A_1, A_2, A_3 on A:n jako, niin

$$A = A_1 + A_2 + A_3.$$



Kuvio 1.3. Joukkojen erotus.



Kuvio 1.4. Joukon A osituksia.

1.3.3 Todennäköisyys

Oletetaan, että satunnaiskoe ja siihen liittyvä otosavaruus on annettu. Tarkastellaan nyt todennäköisyyden määritelemistä. Oletamme aluksi, että otosavaruus on äärellinen. Silloin todennäköisyys voidaan määritellä alkeistapahtumien avulla.

Määritelmä 1.1 Olkoon $\mathcal{E}$ satunnaiskoe ja Ω sen äärellinen otosavaruus. Todennäköisyys on otosavaruudessa Ω määritelty reaaliarvoinen kuvaus

$$P\colon \Omega \to [0,1],$$

jolla on seuraavat ominaisuudet:

1. $P(\omega) \geq 0$ kaikilla $\omega \in \Omega$, ja
2. $\sum\limits_{\omega\in\Omega} P(\omega) = 1.$

Sanomme, että $P(\omega)$ on *alkeistapahtuman ω todennäköisyys.* Tapahtuman A eli Ω:n osajoukon todennäköisyys määritellään lukuna

$$P(A) = \sum_{\omega\in A} P(\omega).$$

Näin funktio P voidaan laajentaa joukkofunktioksi, joka liittää jokaiseen tapahtumaan $A \subset \Omega$ luvun $0 \leq P(A) \leq 1$. Koska todennäköisyys on joukkofunktio, pitäisi alkeistapahtuman todennäköisyyttä oikeastaan merkitä $P(\{\omega\})$, mutta käytämme kuitenkin yleensä lyhyempää merkintää $P(\omega)$. Ominaisuuksiensa nojalla todennäköisyyttä kutsutaan yleisessä teoriassa todennäköisyysmitaksi. Jos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, niin

$$\sum_{\omega_i\in\Omega} P(\omega_i) = \sum_{i=1}^{n} P(\omega_i) = 1.$$

Esimerkiksi tapahtuman $A = \{\omega_1, \omega_3, \omega_5\}$ todennäköisyys $P(A) = P(\omega_1) + P(\omega_3) + P(\omega_5)$. Lisäksi määrittelemme *mahdottoman tapahtuman,* jota merkitään tyhjällä joukolla $\emptyset$, todennäköisyyden $P(\emptyset)$ nollaksi. Satunnaiskokeen todennäköisyysmalli määritellään antamalla kokeen otosavaruus Ω ja siihen liittyvä funktio P, joka toteuttaa Määritelmän 1.1 ehdot. *Todennäköisyysmalli* on siis pari (Ω, P).

Määritelmän mukaan $P(\emptyset) = 0$. Mahdoton tapahtuma $\emptyset$ on varman tapahtuman Ω komplementti eli $\Omega^c = \emptyset$. Tapahtuman A komplementti on joukko, johon kuuluvat kaikki ne alkeistapaukset, jotka eivät kuulu joukkoon A. Koska jokainen alkeistapaus ω kuuluu joukkoon A tai sen komplementtiin, mutta ei molempiin samanaikaisesti, niin

$$\sum_{\omega \in A} P(\omega) + \sum_{\omega \in A^c} P(\omega) = \sum_{\omega \in \Omega} P(\omega) = 1.$$

Tästä seuraa, että $P(A) + P(A^c) = 1$, joten

$$P(A^c) = 1 - P(A).$$

Määritelmän 1.1 oletukset toteuttava funktio määrittelee *todennäköisyysjakauman* Ω:ssa. Jos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, niin voimme esittää todennäköisyysjakauman muodossa

$$\begin{array}{cccc} \omega_1 & \omega_2 & \ldots & \omega_n \\ p_1 & p_2 & \ldots & p_n, \end{array}$$

missä $p_i = P(\omega_i)$ ja $\sum_{i=1}^n p_i = 1$. Mikä tahansa Määritelmän 1.1 ehdot toteuttava reaalilukujoukko $\{\, p_i \mid p_i = P(\omega_i),\ 1 \le i \le n \,\}$ määrittelee todennäköisyysjakauman Ω:ssa.

Esimerkki 1.4 Heitetään harhatonta noppaa. Silloin silmälukujen muodostama otosavaruus on $\Omega = \{1, 2, 3, 4, 5, 6\}$. Jos jokainen silmäluku on yhtä mahdollinen, niin määritellään todennäköisyys P siten, että

$$P(i) = \frac{1}{6}, \qquad i = 1, \ldots, 6.$$

Tapahtuman 'silmäluku pariton' todennäköisyys on

$$P(\{1, 3, 5\}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

□

1.3.4 Äärettömät otosavaruudet

Edellä on käsitelty vain äärellisiä otosavaruuksia. Esimerkissä 1.3 esitettiin myös äärettömiä otosavaruuksia, jotka ovat sovelluksissa tavallisia. Jos Ω on numeroituvasti ääretön, niin

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}.$$

Silloin jakaumafunktio voidaan määritellä samalla tavalla kuin äärellisen otosavaruuden tapauksessa. Määritelmä 1.1 siis soveltuu myös numeroituvasti äärettömiin otosavaruuksiin. Silloin Määritelmän 1.1 2. ehdossa äärellinen summa korvataan äärettömällä summalla

$$\sum_{i=1}^{\infty} p_i = p_1 + p_2 + p_3 + \cdots = 1,$$

missä $P(\omega_i) = p_i$. Jos Ω ei ole numeroituva (eli on ylinumeroituva), niin Määritelmä 1.1 ei sovellu tapahtumien todennäköisyyden määrittelemiseen, vaan tarvitaan uusia käsitteitä. Niihin palataan myöhemmin.

1.3.5 Todennäköisyyden tulkinnat

Todennäköisyyslaskenta ei ole riippuvainen todennäköisyyksien eli lukujen p tulkinnoista eikä siitä, miten näitä lukuja mitataan tai arvioidaan. Todennäköisyyslaskenta on aksiomaattinen matemaattinen teoria. Esimerkiksi diskreetti todennäköisyyslaskenta perustuu Määritelmän 1.1 esittämiin todennäköisyyden ominaisuuksiin. Sovelluksissa tulkitsemme todennäköisyydet usein suureiksi, joita voidaan estimoida suhteellisilla frekvensseillä.

Tapahtuman A *mahdollisuus* (*odds*) määritellään suhteena

$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}. \tag{1.3.1}$$

Tapahtuman A mahdollisuus kertoo, kuinka monta kertaa todennäköisempää on, että A sattuu, verrattuna siihen, että A ei satu. Jos tapahtuman A mahdollisuus $\text{odds}(A)$ on annettu, niin A:n todennäköisyys on

$$P(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}.$$

Esimerkki 1.5 Jos 1000 henkilön populaatiossa on 600 naista ja 400 miestä, niin naisten suhteellinen osuus on

$$\frac{600}{600 + 400} = 0.6.$$

Jos tästä populaatista valitaan satunnaisesti yksi henkilö, niin naisen valitsemisen todennäköisyys on 0.6. Naisen mahdollisuus (odds) tulla valituksi on 6 vastaan 4. Mahdollisuus, että nainen ei tule valituksi on 4 vastaan 6. Jos $A = \{\text{nainen}\}$ ja $B = \{\text{mies}\}$, niin naisen mahdollisuus tulla valituksi on

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)} = \frac{0.6}{0.4} = \frac{3}{2}.$$

□

Uhkapelurit ovat kiinnostuneita hieman erityyppisestä mahdollisuudesta, nimittäin *voiton mahdollisuudesta* (*payoff odds*). Pelikasinot ja vedonlyönnin välittäjät tarjoavat näitä mahdollisuuksia. Jos tapahtuman mahdollisuus on 10 vastaan 1 ja lyöt euron vetoa tapahtuman A puolesta, niin tapahtuman A sattuessa voitat 10 euroa. Jos A ei satu, häviät sen yhden euron. Kasinossa maksat pelimaksuna yhden euron. Jos A sattuu, saat takaisin 11 euroa, joka on voittosi plus euron palautus. Jos A ei satu, kasino pitää maksamasi euron. *Panoksesi* on 1 euro, *kasinon panos* 10 euroa ja *kokonaispanos* 11 euroa.

Voiton mahdollisuuden ja tapahtuman mahdollisuuden välillä on yhteys, joka on ymmärretty uhkapelin yhteydessä paljon ennen varsinaisen todennäköisyyslaskennan syntyä. Puhutaan esimerkiksi ns. *reilun pelin säännöstä,* joka toteutuu silloin, kun tapahtumaa A koskevassa vedolyönnissä voiton mahdollisuus on sama kuin A:n mahdollisuus eli

$$\frac{\text{panos}}{\text{kasinon panos}} = \text{odds}(A).$$

Reilun pelin säännön mukaan panoksen suhteellisen osuuden kokonasipanoksesta tulee olla $P(A)$.

Eivät ainoastaan tapahtumien mahdollisuudet vaan myös mahdollisuuksien suhteet ovat keskeisiä pelitilanteiden analysoinnissa. Ne ovat tärkeitä käsiteitä myös esimerkiksi frekvenssiaineistojen analyysissa ja logistisessa regressiossa. Olkoon A:n mahdollisuus $\text{odds}(A)$ ja B:n mahdollisuus $\text{odds}(B)$. Silloin *mahdollisuuksien suhde* (*odds ratio*) $\theta(A, B)$ on

$$\theta(A, B) = \frac{\text{odds}(A)}{\text{odds}(B)} = \frac{P(A)/[1 - P(A)]}{P(B)/[1 - P(B)]}. \tag{1.3.2}$$

Vedonlyöntiterminologian mukaan θ on *vedonlyöntisuhde.* Todennäköisyyksien arviointi vedonlyönnissä perustuu pitkälti henkilökohtaisiin uskomuksiin ja kokemuksiin. Myös esimerkiksi liiketoiminnan päätöksenteossa henkilökohtaiset todennäköisyyden tulkinnat voivat olla käyttökelpoisia.

1.4 Ehdollinen todennäköisyys

Ehdollistaminen on varsin tehokas ja hyödyllinen tekniikka todennäköisyyslaskennassa ja tilastotieteessä. Käsittelemme tässä luvussa ensimmäisen kerran lyhyesti ehdollista todennäköisyyttä, joka tulee olemaan tärkeä käsite läpi koko kurssin.

Esimerkki 1.6 Heitetään harhatonta noppaa kuten Esimerkissä 1.4. Meille kerrotaan, että on saatu pariton silmäluku, mutta emme tiedä, mikä niistä. Mikä on silmäluvun 5 todennäköisyys? Olkoon B 'silmäluku pariton' ja A 'silmäluku 5'. Tiedämme siis, että silmäluku on 1, 3 tai 5. Nämä alkeistapaukset ovat yhtä todennäköisiä, joten silmäluvun 5 todennäköisyys on $1/3$. Sanomme, että tapahtuman A *ehdollinen todennäköisyys* ehdolla B on $1/3$. Tätä ehdollista todennäköisyyttä merkitään $P(A \mid B)$. Huomaamme, että ainakin tässä esimerkissä $P(A \mid B) \neq P(A) = 1/6$. □

Kun tarkastellaan tapahtuman A ehdollista todennäköisyyttä $P(A \mid B)$, rajoitutaan tarkastelemaan tapahtuman B alkeistapauksia. Sitten katsotaan, kuinka usein B:ssä sattuu myös A. Tämä on tapahtuma 'sekä A että B sattuvat', jota merkitään $A \cap B$. Edellisessä esimerkissä laskimme itse asiassa ehdollisen todennäköisyyden $P(A \mid B)$ kaavalla

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.4.1}$$

Todennäköisyys $P(A \mid B)$ on määritelty, kun $P(B) > 0$.

Esimerkki 1.7 Eloonjäämistaulukoissa esitetään eri ikäisenä elossa olevien odotettu lukumäärä 100000 elävänä syntynyttä kohti. Esimerkiksi seuraavassa taulukossa on annettu 20-, 45- ja 65-vuotiaana elossa olevien naisten lukumäärät eräässä väestössä 100000 elävänä syntynyttä tyttölasta kohti.

Ikä	20	45	65
Elossa	98040	95662	84483

Tässä voidaan ajatella, että alkuperäinen otosavaruus Ω on 100000 tyttölasta. Mikä on todennäköisyys, että 20-vuotias elää 45-vuotiaaksi (tarkoittaa itse asiassa, että elää ainakin 45-vuotiaaksi)? Olkoon A = 'elää 45-vuotiaaksi' ja B = 'elää 20-vuotiaaksi'. Koska 20-vuotiaaksi on elänyt 98040 naista ja näistä 45-vuotiaaksi 95662, niin kysytty todennäköisyys on $95662/98040 = 0.97574$. Laskettaessa ehdollista todennäköisyyttä valitaan perusjoukoksi B ja katsotaan kuinka moni näistä selviää 45-vuotiaaksi.

Nyt tapahtuma $A \cap B$ on 'elää 45-vuotiaaksi', koska 45-vuotiaksi eläneet ovat eläneet myös 20-vuotiaksi. Koska 20-vuotiaaksi elää 98040, niin $P(B) = 98040/100000 = 0.98040$. Vastaavasti $P(A \cap B) = 95662/100000 = 0.95662$. Ehdollinen todennäköisyys

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.95662}{0.98040} = 0.97574.$$

□

1.4.1 Ehdollisen todennäköisyyden frekvenssitulkinta

Olkoot A ja B jotkut satunnaiskokeen $\mathcal{E}$ otosavaruuteen Ω liittyvät tapahtumat ja $N_n(A \cap B)$ on tapahtuman $A \cap B$ frekvenssi ja $N_n(B)$ tapahtuman B frekvenssi, kun satunnaiskoe $\mathcal{E}$ toistetaan n kertaa. Voimme ajatella, että

$$P(A \mid B) \approx \frac{N_n(A \cap B)}{N_n(B)} = \frac{N_n(A \cap B)/n}{N_n(B)/n} \approx \frac{P(A \cap B)}{P(B)}, \tag{1.4.2}$$

kun toistojen lukumäärä n on suuri.

1.4.2 Kertolaskusääntö

Koska ehdollisen todennäköisyyden kaavassa (1.4.1) $P(B) > 0$, saadaan siitä kertolaskusääntö

$$P(A \cap B) = P(B)\, P(A \mid B) \tag{1.4.3}$$

tapahtuman $A \cap B$ todennäköisyyden laskemiseksi.

1.4.3 Riippumattomuus

Sanomme, että tapahtumat A ja B ovat *riippumattomat,* jos

$$P(A \cap B) = P(A)\, P(B). \tag{1.4.4}$$

Huomaa, että ehdollinen todennäköisyys (1.4.1) ei ole määritelty, jos $P(B) = 0$, mutta riippumattomuuden määritelmä (1.4.4) on silloinkin voimassa. Jos $P(B) \neq 0$ ja (1.4.4) pitää paikkansa, niin

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A).$$

Jos A ja B ovat riippumattomat, niin tieto B:n sattumisesta ei vaikuta A:n todennäköisyyteen. Jos $P(A) > 0$, niin myös $P(B \mid A) = P(A \cap B)/P(A) = P(B)$, kun A ja B ovat riippumattomat.

1.5 Odotetut frekvenssit

Kokeen $\mathcal{E}$ todennäköisyysmalli (Ω, P) on teoreettinen konstruktio. Mallin hyvyys käytännön sovelluksissa on tutkittava empiirisesti. Tämä tehdään vertailemalla kokeen (empiirisen ilmiön) havaittuja tuloksia mallin perusteella odotettavissa oleviin tuloksiin. Oletetaan, että koe toistetaan n kertaa. Jos tapahtuman A todennäköisyys on mallin mukaan p, niin silloin A:n *odotettu frekvenssi* eli *teoreettinen frekvenssi* on np. Jos A sattui suoritetussa toistokokeessa n_A kertaa, niin tätä *havaittua frekvenssiä* verrataan odotettuun frekvenssiin. Jos n_A poikkeaa "liian paljon" odotetusta frekvenssistä np, niin malli (teoria) joutuu kyseenalaiseksi. Havainnot eivät silloin tue teoriaa. Siihen, mikä on "liian suuri" poikkeama, pyrimme vastaamaan todennäköisyyslaskennan ja tilastotieteen avulla.

Johdanto: Yhteenveto

- Empiirinen kertymäfunktio. Lukujen $x_1, x_2, \ldots, x_n$ *empiirinen kertymäfunktio* on
$$F_n(a) = \frac{1}{n} |\{\, i : 1 \leq i \leq n,\, x_i \leq a \,\}|,$$
missä $-\infty < a < \infty$ ja $|\,.\,|$ on joukon alkioiden lukumäärä.

- *Empiirinen jakaumafunktio* tai lyhyesti *empiirinen jakauma* on

$$P_n(a, b) = F_n(b) - F_n(a).$$

- Otosavaruus Ω on satunnaiskokeen (tai satunnaisilmiön) mahdollisten tulosten (alkeistapausten ω) joukko. Satunnaiskokeessa voi sattua yksi ja vain yksi alkeistapaus.

- Tapahtuma on otosavaruuden Ω osajoukko.

A ja B tapahtumia	$A \subset \Omega$ ja $B \subset \Omega$
Ω	varma tapahtuma
$\emptyset$	mahdoton tapahtuma
$A \subset B$	jos A sattuu, niin B sattuu
A^c	A ei satu
$A \cup B$	A tai B sattuu (tai molemmat)
$A \cap B$, AB	sekä A että B sattuvat
$A \setminus B = A \cap B^c$	A sattuu, mutta ei B
$A \cap B = \emptyset$	A ja B pistevieraat (toisensa poissulkevat)
A:n ositus	$A = A_1 \cup A_2 \cup \cdots \cup A_m$ ja $A_i \cap A_j = \emptyset$, $i \neq j$

- De Morganin lait

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c.$$

- *Todennäköisyys* P on otosavaruudessa Ω (numeroituva) määritelty funktio $P\colon \Omega \to [0, 1]$, jolla on seuraavat ominaisuudet:

  1. $P(\omega) \geq 0$ kaikilla $\omega \in \Omega$, ja
  2. $\sum_{\omega \in \Omega} P(\omega) = 1$.

- Tapahtuman A todennäköisyys $P(A) = \sum_{\omega \in A} P(\omega)$.

- Tapahtuman A mahdollisuus

$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}.$$

- Vedonlyöntisuhde

$$\theta(A, B) = \frac{\text{odds}(A)}{\text{odds}(B)}.$$

- A:n todennäköisyys ehdolla B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0.$$

- Kertolaskusääntö $P(A \cap B) = P(B)\,P(A \mid B)$.
- Riippumattomuus: A ja B ovat riippumattomat, jos $P(A \cap B) = P(A)\,P(B)$.
- Todennäköisyysmalli: Kokeen $\mathcal{E}$ todennäköisyysmalli on otosavaruuden Ω ja todennäköisyyden P muodostama kaksikko (Ω, P).

Harjoituksia

1. Liitteessä 1 (ja tiedostossa `mtt/datat/hsarjat200.dat`) on kolme 200:n heiton sarjaa, joista yksi on tuotettu heittämällä harhatonta lanttia (200 riippumatonta toistokoetta, jossa kruunun tn = 1/2). Muut sarjat poikkeavat selvästi (?) "oikeasta" rahanheittokokeen tuloksesta. Koeta päätellä tai arvata, mikä on se aito rahanheiton tulos (Vrt. Mustonen: SURVO MM, Opetusohjelmat/Todennäköisyyksien laskentaa). Laske jokaisesta sarjasta kruunujen lkm. Onko Liitteen 1 tulosten perusteella uskottavaa, että sarjat on saatu harhattomalla rahalla (kruunu = 1 ja klaava = 0).

2. Aineistossa `kaivos_onn.dat` on aikajärjestyksessä pahojen (yli 10 kuollutta) peräkkäisten kaivosonnettomuuksien väliajat (päivinä) ajanjaksolta 6. 12. 1875 – 29. 5. 1951. Piirrä väliaikojen frekvenssihistogramma koko aineistosta ja erilliset histogrammat 56:sta ensimmäisestä ja 53:sta viimeisestä havainnosta. Kommentoi eroja ja yhtäläisyyksiä.

3. Oletetaan, että histogrammassa kahden vierekkäisen suorakaiteen kannan leveydet ovat k_1 ja k_2 sekä korkeudet h_1 ja h_2. Yhdistetään suorakaiteet yhdeksi suorakaiteeksi. Esitä uuden suorakaiteen korkeuden h lauseke ja osoita, että h on korkeuksien h_1 ja h_2 välissä.

4. Heitä harhatonta noppaa (R-ohjelma) 60, 120, 240, 480, 960 ja 2000 kertaa ja laske eri silmälukujen suhteelliset frekvenssit eri heittosarjoissa. Piirrä myös suhteellisten frekvenssien histogrammat. Miten heittojen lkm:n n kasvattaminen vaikuttaa suhteellisiin frekvensseihin?

5. Henkilöille X, Y, Z ja W on kullekin osoitettu kirje. Jokaiselle kirjeelle on varattu osoitteella varustettu kirjekuori. Kirjeet pannaan satunnaisesti kirjekuoriin.

(a) Mikä on tämän kokeen 24 alkeistapahtuman otosavaruus.

(b) Luettele seuraaviin tapahtumiin liittyvät alkaistapahtumat.

A: "X:n kirje menee oikeaan kuoreen";

B: "Mikään kirje ei mene oikeaan kuoreen";

C: "Täsmälleen kaksi kirjettä menee oikeaan kuoreen";

D: "Täsmälleen kolme kirjettä menee oikeaan kuoreen";

(c) Laske edellisessä kohdassa mainittujen tapahtumien todennäköisyydet, jos oletetaan, että kaikki alkeistapaukset ovat yhtä todennäköisiä. Määritä tapahtumien A, C ja D mahdollisuudet tapahtumaa B vastaan.

6. Kaksi joukkuetta pelaa paras seitsemästä sarjaa. Se joukkue voittaa, joka on ensiksi voittanut neljä peliä. Mikä on kokeen otosavaruus? Jos joukkueet ovat tasavahvoja (ja pelien tulokset toisistaan riippumattomia), niin mitkä ovat eri alkeistapahtumien todennäköisyydet? Mikä on todennäköisyys, että voittoon tarvitaan 7 peliä?

7. Tarkastellaan sellaista noppaa, että $p_1 = p_2 = p_3 = p_4 = p$ ja $p_5 = p_6 = q$. Kirjoitetaan tn p muodossa $p = \frac{1}{6} + \theta$.

(a) Lausu q θ:n avulla.

(b) Heitetään noppaa n kertaa ja saadaan silmälukujen 1, 2, 3, 4, 5, 6 lukumääriksi n_1, n_2, n_3, n_4, n_5, n_6. Miten estimoisit θ:n arvon?

(c) Heitettiin noppaa 30, 120, 600 ja 1200. Silmälukujen frekvenssit olivat.

	Silmäluvut					
n	1	2	3	4	5	6
30	6	10	6	5	0	3
120	29	17	35	25	9	5
600	126	119	141	124	50	40
1200	255	278	231	254	90	92

Laske θ:n, p:n ja q:n estimaatit.

8. (a) Mikä on tn-malli, kun heitetään samanaikaisesti kolmea harhatonta lanttia.

(b) Määritä tn saada x kruunua.

(c) Heitettiin kolmea lanttia 80 kertaa ja saatiin seuraavat kruunujen lukumäärät.

```
1 1 1 1 2 1 1 2 2 1 1 2 2 3 2 1 1 2 1 2 0 1 1 0 2 1 0
1 1 3 0 3 0 1 2 1 2 1 2 2 1 3 1 2 2 0 1 1 1 3 2 0 3 2
0 2 0 1 0 1 1 3 2 2 1 1 2 1 2 1 1 1 2 3 3 2 0 2 1 3
```

Määritä kruunujen lukumäärän odotetut ja havaitut frekvenssit. Ovatko havainnot sopusoinnussa mallin kanssa (Heitot tiedostossa `H1.8_heitot.dat`)?

9. (a) Heitetään samanaikaisesti kahta noppaa ja olkoon tulos silmälukujen summa. Olkoot kaikki 36 alkeistapausta ovat yhtä todennäköisiä. Osoita, että tuloksen tn-jakauma on:

Tulos	2	3	4	5	6	7	8	9	10	11	12
$36 \times$ tn	1	2	3	4	5	6	5	4	3	2	1

(b) Heitä kahta noppaa 100 kertaa. Vertaa tuloksen havaittuja frekvenssejä odotettuihin frekvensseihin.

10. Vuoden 2003 jääkiekon pudotuspelijoukkueet olivat HPK (1/3), Jokerit (1/2), Kärpät (1/3), Espoon BLUES (1/6), Tappara (1/3), JYP (1/7), HIFK (1/6) ja TPS (1/9). Eräällä työpaikalla järjestettiin ennen pudotuspelien alkua vuoden mestaria koskeva vedonlyönti käyttäen suluissa ilmoitettuja voiton mahdollisuuksia. Jos veikkasit esimerkiksi Tapparaa mestariksi, niin voitit panoksesi kolminkertaisena.

(a) Laske annettujen voiton mahdollisuuksien (payoff odds) avulla joukkueiden voiton todennäköisyydet kaavalla (1.3.2). Laske todennäköisyyksien summa S.

(b) Skaalaa edellisessä kohdassa lasketut "todennäköisyydet" jakamalla ne summalla S. Miksi skaalaus on tarpeellinen?

(c) Oleta, että skaalatut todennäköisyydet ovat "oikeita". Laske odotettu voittosi, jos veikkasit Tapparaa [voitto $\times P(A) +$ panoksesi $\times (1 - P(A))$]. Toteuttaako veikkaus reilun pelin säännön?

11. Eräässä kyselyssä tutkittiin suhtautumista lailliseen aborttiin ja saatiin oheisessa taulukossa esitetyt tulokset.

	Asenne		
Sukupuoli	Myönteinen	Kielteinen	Yhteensä
Nainen	309	191	500
Mies	319	281	600
Yhteensä	628	472	1100

Käytä todennäköisyyksien estimaatteina suhteellisia frekvenssejä.

(a) Laske todennäköisyys, että (i) nainen (ii) mies suhtautuu aborttiin positiivisesti (tarkasteltavassa otosavaruudessa).

(b) Laske mahdollisuudet (odds), että (i) nainen (ii) mies suhtautuu aborttiin positiivisesti.

(c) Laske mahdollisuuksien suhde (odds ratio, vedonlyöntisuhde).

12. Esimerkissä 1.2 (luennot) on annettu erään kurssin 1. välikokeen pistemäärät.

(a) Laske empiirisen kertymäfunktion (ekf) arvo pisteessä 15.3.

(b) Lausu empiirisen jakauman arvo $P_{20}(18.5, 20.5)$ ekf:n avulla.

(c) Laske histogrammissa luokkaa $[18.5, 20.5]$ kuvaavan pylvään korkeus.

Liite 1

1 1 0 1 0 1 0 0 1 1 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 1 1 0 1 0 0 1
1 1 0 1 1 1 0 0 0 1 1 1 1 1 1 0 1 0 0 1 0 0 0 1 1 0 1 1 0 1 1 0
1 1 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1 0 1 0 0 0 1 0 1 0 0
1 0 1 1 0 1 0 1 0 1 0 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0 1 0 0 1 0 1
0 0 1 1 0 1 1 0 0 0 1 0 0 0 1 0 1 1 0 1 1 1 0 0 1 0 0 1 0 1 0 1
0 0 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 1 0 1 0
1 0 1 1 0 1 0 0

1 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 1 1 0 0 0 1 1 1 1 0 1 1 0 0 1
0 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1
1 1 0 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 1 0 1 1 0 1 1 1 1 1 1 0 0 0 1
0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 1 1 1 0 1 0
1 1 1 0 1 1 0 1 0 0 1 1 1 1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 1 0
0 0 1 1 0 0 1 1 1 1 1 1 0 1 0 0 0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1
1 1 1 0 1 0 1 0

0 0 1 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 1 0 1 1 0 1
1 1 0 1 0 1 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 0 0 0 0 0 1 0 1 0 1 0
0 0 0 1 0 1 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0
1 0 1 0 0 1 0 0 1 0 0 0 0 1 1 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1 0 1
1 0 0 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 1 1 1 1 0 1 0 0 1 0 1 0 0
1 0 1 0 1 1 0 1 0 1 0 1 0 1 0 1 0 1 1 1 0 0 1 0 0 0 1 0 0 1 0 0
0 1 1 1 0 1 1 0

Luku 2

Todennäköisyyslaskenta ja kombinatoriikka

Tässä luvussa käsitellään lähinnä vain äärellisiä ja numeroituvasti äärettömiä otosavaruuksia Ω. Lopuksi esitetään todennäköisyyden aksioomat, jotka soveltuvat myös silloin, kun Ω ei ole numeroituva.

2.1 Todennäköisyyden ominaisuuksia

Seuraavassa lauseessa on esitetty todennäköisyyden keskeiset ominaisuudet. Erityisesti numeroituvien otosavaruuksien tapauksessa lauseen tulokset on helppo todistaa.

Lause 2.1 *Oletetaan, että Ω on numeroituva otosavaruus ja P on Ω:ssa määritelty todennäköisyys. Todennäköisyydellä P on seuraavat ominaisuudet:*

1. $P(A) \geq 0$ *kaikilla* $A \subset \Omega$.
2. $P(\Omega) = 1$.
3. *Jos* $A \subset B \subset \Omega$, *niin* $P(A) \leq P(B)$.
4. *Jos A ja B ovat erilliset* ($A \cap B = \emptyset$), *niin* $P(A \cup B) = P(A) + P(B)$.
5. $P(A^c) = 1 - P(A)$ *kaikilla* $A \subset \Omega$.

Todistus. Jokaisen tapahtuman $A \subset \Omega$ todennäköisyys on Määritelmän 1.1 mukaan

$$P(A) = \sum_{\omega \in A} P(\omega).$$

Koska $P(\omega) \geq 0$ kaikilla $\omega \in \Omega$, niin $P(A) \geq 0$. Näin on 1. kohta todistettu.

Toinen kohta pitää paikkansa, koska Määritelmän 1.1

$$P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1.$$

Ominaisuuksien 3–5 todistaminen jätetään harjoitustehtäväksi. □

Todennäköisyyden additiivisuus (Ominaisuus 4) voidaan suoraviivaisesti yleistää useammalle kuin kahdelle erilliselle joukolle.

Lause 2.2 *Olkoot A_1, A_2, ..., A_n parittain pistevieraat (erilliset) Ω:n osajoukot eli tapahtumat (ts. $A_i \cap A_j = \emptyset$, kun $i \neq j$). Silloin*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n).$$

Itse asiassa additiivisuus yleistyy myös ärettömän monelle parittain erilliselle tapahtumalle A_1, A_2, A_3, ... Silloin

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots .$$

Jos A_1, A_2, ..., A_n ovat parittain erilliset (ts. $A_i \cap A_j = \emptyset$, kun $i \neq j$) ja $\Omega = A_1 \cup A_2 \cup \cdots \cup A_n$, niin joukkokokoelma $A_1, A_2, \ldots, A_n$ on *otosavaruuden Ω ositus.*

Lause 2.3 *Olkoon kokoelma $A_1, A_2, \ldots, A_n$ otosavaruuden Ω ositus ja $E \subset \Omega$ on jokin tapahtuma. Silloin*

$$P(E) = \sum_{i=1}^{n} P(E \cap A_i).$$

Seuraus 2.1 *Mille tahansa kahdelle tapahtumalle A ja B pitää paikkansa, että*

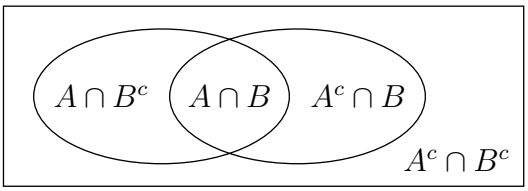
$$P(A) = P(A \cap B) + P(A \cap B^c).$$

Lauseen 2.1 kohta 4 voidaan yleistää myös joukoille, jotka eivät ole erillisiä. Tällöin saadaan seuraava *yhteenlaskulause.*

Lause 2.4 *Jos $A \subset \Omega$ ja $B \subset \Omega$, niin*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.1.1}$$

Todistus. Kuten Kuvio 2.1 osoittaa, joukot $A \cap B^c$, $A \cap B$, $A^c \cap B$ muodos-



Kuvio 2.1. Tapahtuman $A \cup B$ ositus.

tavat tapahtuman $A \cup B$ osituksen. Siksi

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B). \tag{2.1.2}$$

Lauseen 2.2 mukaan vastaavasti

$$P(A) = P(A \cap B^c) + P(A \cap B)$$
$$P(B) = P(A^c \cap B) + P(A \cap B),$$

joten

$$P(A) + P(B) = P(A \cap B^c) + P(A^c \cap B) + 2\,P(A \cap B). \tag{2.1.3}$$

Kun identiteetistä (2.1.3) vähennetään puolittain $P(A \cap B)$, saadaan lauseke

$$P(A) + P(B) - P(A \cap B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B),$$

jonka oikea puoli on (2.1.2):n mukaan $P(A \cup B)$. Näin yhteenlaskulause on todistettu. □

Tämä todennäköisyyksien yhteenlaskulause voidaan edelleen yleistää mielivaltaisen monelle tapahtumalle. Esitämme aluksi yleistyksen, kun tapahtumia on kolme. Yleinen tapaus saadaan samalla periaatteella, mutta se esitetään vasta myöhemmin.

Lause 2.5 *Jos A_1, A_2 ja A_3 ovat Ω:n osajoukkoja (tapahtumia), niin*

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = {} & P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) \\ & - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3). \end{aligned} \tag{2.1.4}$$

Bonferronin epäyhtälö. Koska $P(A \cup B) \leq 1$, seuraa Lauseesta 2.4 epäyhtälö

$$P(A \cap B) \geq P(A) + P(B) - 1. \tag{2.1.5}$$

Epäyhtälöä 2.1.5 sanotaan *Bonferronin epäyhtälöksi.*

Esimerkki 2.1 Bonferronin epäyhtälö saattaa olla käyttökelpoinen silloin, kun ei pystytä laskemaan todennäköisyyttä $P(A \cap B)$ tarkasti, mutta tunnetaan todennäköisyydet $P(A)$ ja $P(B)$. Olkoon esimerkiksi $P(A) = P(B) = 0.95$. Silloin

$$P(A \cap B) \geq 0.95 + 0.95 - 1 = 0.90.$$

Jos $P(A) + P(B) < 1$, niin alaraja (2.1.5):ssa on negatiivinen ja epäyhtälö pitää triviaalisti paikkansa. □

2.2 Symmetriaan perustuva todennäköisyys

Jos äärellisen otosavaruuden $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ jokainen alkeistapaus on yhtä mahdollinen, niin jakaumafunktio on

$$p_i = P(\omega_i) = \frac{1}{n}, \quad 1 \leq i \leq n$$

missä n on alkeistapausten lukumäärä. Silloin jokaisen tapahtuman todennäköisyys on yksinkertaisesti

$$P(A) = \sum_{\omega_i \in A} p_i = \sum_{\omega_i \in A} \frac{1}{n} = \frac{|A|}{n}, \tag{2.2.1}$$

missä $|A|$ on A:n alkioiden lukumäärä. Tapahtuman A todennäköisyys saadaan siis jakamalla A:n alkeistapausten lukumäärä kaikkien alkeistapahtumien lukumäärällä n. Tätä '*suotuisat per kaikki*' -sääntöä kutsutaan myös klassiseksi todennäköisyyden määritelmäksi.

Esimerkki 2.2 Heitetään harhatonta noppaa. Silloin eri silmälukuja voidaan pitää yhtä mahdollisina ja jakaumafunktio on perusteltua määritellä $p_i = \frac{1}{6}$, $i = 1, \ldots, 6$ otosavaruudessa $\Omega = \{1, 2, 3, 4, 5, 6\}$. Jos heitetään kahta noppaa, voidaan symmetrisiksi alkeistapauksiksi valita järjestetyt parit

$$(1,1), (1,2), (1,3), \ldots, (6,6).$$

Siinä tulokset on annettu muodossa (1. nopan silmäluku, 2. nopan silmäluku). Tämän satunnaiskokeen otosavaruus on siis

$$\Omega = \{\, (i,j) \mid i, j \in \{1,2,3,4,5,6\} \,\}.$$

Koska $|\Omega| = 36$, niin $P\big(\{(i,j)\}\big) = \frac{1}{36}$ kaikilla $i, j \in \{1,2,3,4,5,6\}$.

Väitetään, että ranskalainen aatelismies ja uhkapeluri Chevalier de Méré havaitsi kokeellisesti seuraavan tuloksen:

(i) Heitettäessä noppaa 4 kertaa kannattaa lyödä vetoa siitä, että saadaan ainakin yksi kuutonen.

(ii) Heitettäessä kahta noppaa 24 kertaa *ei kannata* lyödä vetoa siitä, että saadaan ainakin yksi kuutospari.

De Méré huomasi jäävänsä pitkässä pelisarjassa häviölle lyödessään vetoa kuutosparin puolesta. Hän ei kuitenkaan pystynyt teoreettisesti selittämään havaintoaan (de Mérén ongelma) ja niinpä hän kääntyi ranskalaisen filosofin ja matemaatikon Pascalin puoleen (n. 1650). De Mérén ongelman uskotaan antaneen alkusysäyksen kuuluisaan Pascalin ja Fermatin väliseen kirjeenvaihtoon, joka johti todennäköisyyslaskennan syntyyn. □

Klassisen määritelmän mukaan tapahtuman A todennäköisyys saadaan jakamalla joukon A alkioiden lukumäärä kaikkien alkeistapausten lukumäärällä. Vaikka tehtävä on periaatteessa helppo, se voi käytännössä osoittautua yllättävän hankalaksi. Lukumäärien laskemisen helpottamiseksi esitämme seuraavassa joitain kombinatoriikan periaatteita ja tuloksia.

2.3 Aksiomaattinen lähestymistapa

Kutsumme tästä lähtien joukkoa *numeroituvaksi,* jos se on *äärellinen* tai *numeroituvasti ääretön.* Oletamme tässä, että otosavaruus Ω on numeroituva, jotta voisimme esittää todennäköisyyden aksioomat mahdollisimman yksinkertaisessa muodossa. Todennäköisyys luonnehditaan Ω:n osajoukkojen joukossa määriteltynä funktiona.

Määritelmä 2.1 Todennäköisyys(mitta) P on otosavaruudessa Ω määritelty Ω:n osajoukkojen kuvaus eli funktio, joka toteuttaa seuraavat kolme aksioomaa:

(i) Funktion P arvo on kaikilla osajoukoilla $A \subset \Omega$ epänegatiivinen:

$$P(A) \geq 0.$$

(ii) Minkä tahansa kahden toisensa poissulkevan tapahtuman A ja B unionin $A + B$ kuva on A:n ja B:n kuvien summa:

$$P(A + B) = P(A) + P(B), \qquad \text{kun } A \cap B = \emptyset.$$

(iii) Koko otosavaruuden Ω kuva on 1 eli

$$P(\Omega) = 1.$$

2.4 Kombinatoriikkaa

2.4.1 Summa- ja tuloperiaate

Olkoot kokeiden $\mathcal{E}_1$ ja $\mathcal{E}_2$ otosavaruudet Ω_1 ja Ω_2. Silloin kokeiden tulosvaihtoehtojen lukumäärät ovat $|\Omega_1|$ ja $|\Omega_2|$. Merkitään $|\Omega_1| = n_1$ ja $|\Omega_2| = n_2$.

Summaperiaate. Tehdään joko koe $\mathcal{E}_1$ tai $\mathcal{E}_2$. Silloin mahdollisten tulosten lukumäärä on $n_1 + n_2$.

Tuloperiaate. Tehdään ensin koe $\mathcal{E}_1$ ja sitten $\mathcal{E}_2$. Silloin yhdistetyn kokeen $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$ tulosvaihtoehtojen lukumäärä on $n_1 n_2$.

2.4.2 Valinta järjestyksessä

Tarkastellaan ensin *valintaa palauttaen.* Olkoon perusjoukon Ω alkioiden lukumäärä $|\Omega| = n$ ja voimme siis ajatella, että alkiot on numeroitu 1:stä n:ään. Valitaan Ω:sta peräkkäin r alkiota ja jokainen valittu alkio palautetaan takaisin Ω:aan ennen seuraavaa valintaa. Valinnan tuloksena saatua järjestettyä jonoa kutsutaan *järjestetyksi r-otokseksi* $(a_1, a_2, \ldots, a_r)$, jossa jokainen alkio $1 \leq a_j \leq n$. Järjestetyssä r-otoksessa sama alkio voi siis toistua monta

kertaa. Tehdään esimerkiksi järjestetty 3-otos joukosta $A = \{a, b\}$. Silloin kaikki mahdolliset järjestetyt 3-otokset ovat *aaa*, *aab*, *aba*, *abb*, *baa*, *bab*, *bba*, *bbb*. Järjestettyjen 3-otosten lukumäärä on tuloperiaatteen mukaan $2^3 = 8$. Samalla tavalla tuloperiaatteesta seuraa, että Ω:sta valittujen järjestettyjen r-otosten lukumäärä on n^r.

Tarkastellaan nyt *valintaa palauttamatta.* Jos Ω:sta valitaan järjestyksessä r alkiota ($r \leq n$) palauttamatta, saadaan järjestetty r-otos, jossa sama alkio voi esiintyä vain kerran. Tällaista järjestettyä r-otosta kutsutaan Ω:n *r-permutaatioksi.* Esimerkiksi joukon $B = \{a, b, c, d\}$ 2-permutaatiot ovat

$$ab,\ ba,\ ac,\ ca,\ ad,\ da,\ bc,\ cb,\ bd,\ db,\ cd,\ dc,$$

joiden lukumäärä on tuloperiaatteen nojalla $4 \cdot 3 = 12$. Yleisesti r-permutaatioiden lukumäärä Ω:sta on

$$n^{(r)} = n(n-1)(n-2)\cdots(n-r+1), \qquad 0 < r \leq n.$$

Merkintä $n^{(r)}$ luetaan "n:n r-kertoma". Kun $r = n$, saadaan *joukon* Ω *n-permutaatio,* jota kutsutaan yksinkertaisesti joukon *permutaatioksi.* Permutaatio on siis joukon alkioiden järjestetty jono. Joukon Ω permutaatioiden lukumäärä on siis

$$n^{(n)} = n(n-1)(n-2)\cdots 2 \cdot 1$$

ja sitä merkitään $n!$ ja luetaan "n-kertoma".

Esimerkki 2.3 (Syntymäpäiväongelma) Kutsuilla on r henkilöä. Henkilöiden syntymäpäivät muodostavat r:n päivämäärän jonon, jossa sama päivämäärä voi toistua. Vuoden päivien lukumäärä $n = 365$, jos karkausvuotta ei oteta huomioon. Oletetaan, että kaikki mahdolliset 365^r syntymäpäiväjonoa ovat yhtä todennäköiset. Mikä on todennäköisyys, että ainakin kahdella henkilöllä on sama syntymäpäivä? Ensinnäkin 365:n päivän r-permutaatioiden lukumäärä on $365^{(r)}$, mikä on siis kaikkien r:n pituisten eri syntymäpäivistä muodostettujen jonojen lukumäärä. Todennäköisyys, että kaikilla on eri syntymäpäivä, on kaavan (2.2.1) mukaan

$$P(\text{'Eri syntymäpäivät'}) = \frac{365^{(r)}}{365^r}.$$

Silloin todennäköisyys, että ainakin kahdella sama syntymäpäivä on

$$1 - \frac{365^{(r)}}{365^r}.$$

□

2.4.3 Osajoukon valinta

Kun Ω:sta valitaan r alkiota ($r \leq n$) palauttamatta, saadaan Ω:n osajoukko. Nyt ei siis kiinnitetä huomiota alkioiden järjestykseen, vaan ainoastaan

siihen, mitkä alkiot osajoukkoon kuuluvat. Joukon r:n alkion osajoukkoa kutsutaan joukon r-kombinaatioksi. Esimerkiksi joukon $B = \{a, b, c, d\}$ 2-kombinaatiot ovat

$$\{a, b\},\ \{a, c\},\ \{a, d\},\ \{b, c\},\ \{b, d\},\ \{c, d\}.$$

Jokaista 2-kombinaatiota kohti on olemassa kaksi 2-permutaatiota. Esimerkiksi 2-kombinaatioon $\{a, b\}$ liittyvät 2-permutaatiot ovat ab, ba. Koska 2-permutaatiota on $4 \cdot 3 = 12$ kappaletta, niin 2-kombinaatiota on $\frac{4 \cdot 3}{2} = 6$ kappaletta. Ja yleisesti: Koska r-permutaatioden lukumäärä jokaista r-kombinaatiota kohti on $r!$ ja r-permutaatioden lukumäärä on $n^{(r)}$, niin r-kombinaatioiden lukumäärä on

$$\frac{n^{(r)}}{r!} = \frac{n!}{r!\,(n-r)!},$$

jota merkitään $\binom{n}{r}$ ja se luetaan "n r:n yli".

Joukossa $A = \{a, a, a, a, b, b, c, c, c, d\}$ on 4 a-kirjainta, 2 b:tä, 3 c:tä ja yksi d. Kuinka monta erilaista 10-kirjaimista sanaa näistä kirjaimista voidaan muodostaa? Sanassa on kirjaimille 10 eri paikkaa ja jokainen kirjain voidaan sijoittaa johonkin 10:stä mahdollisesta paikasta. Ensiksikin a-kirjaimien paikka voidaan valita $\binom{10}{4}$ tavalla, jäljelle jääneisiin 6:een paikkaan voidaan b sijoittaa $\binom{6}{2}$ tavalla, sen jäkeen c $\binom{4}{3}$ tavalla ja lopuksi d:lle jää $\binom{1}{1} = 1$ paikka. Kertolaskuperiaatteen mukaan kaikkien mahdollisten sanojen lukumäärä on

$$\binom{10}{4}\binom{6}{2}\binom{4}{3}\binom{1}{1} = \frac{10!}{4!\,2!\,3!\,1!} = 12600.$$

Olkoon joukossa n alkiota, joista n_1 kuuluu 1. ryhmään, n_2 2. ryhmään ja lopulta n_k alkiota k. ryhmään, joten $n = n_1 + n_2 + \cdots + n_k$. Joukosta valitaan peräkkäin palauttamatta alkioita kunnes kaikki on valittu. Kuinka monta tunnistettavasti erilaista alkiojonoa voidaan saada? Nyt ajatellaan, että kunkin ryhmän alkiot ovat keskenään samanlaisia. Emme voi siis tunnistaa erilaisia ryhmän sisäisiä järjestyksiä. Vastaus saadaan samalla tavalla kuin edellisessä esimerkissä, joten valintojen lukumäärä on

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k} \quad = \quad \frac{n!}{n_1!\,n_2!\cdots n_k!}.$$

Tätä lauseketta sanotaan *multinomikertoimeksi* ja sitä merkitään

$$\frac{n!}{n_1!\,n_2!\cdots n_k!} = \binom{n}{n_1\ n_2\ \ldots\ n_k}. \tag{2.4.1}$$

Kun $k = 2$, saadaan erikoistapauksena binomikerroin

$$\binom{n}{n_1\ n_2} = \frac{n!}{n_1!\,n_2!} = \frac{n!}{n_1!\,(n-n_1)!} = \binom{n}{n_1}.$$

2.4.4 Otanta palauttaen, kun järjestystä ei oteta huomioon

Valitaan r palloa uurnasta, jossa on k erilaista (esimerkiksi eriväristä) palloa. Jokaisessa valinnassa rekisteröidään pallon väri ja pallo palautetaan uurnaan ennen seuraavaa valintaa. Olkoon uurnassa esimerkiksi 3 erilaista palloa: $\odot$, $\ominus$, $\oplus$. Valitaan uurnasta 3 palloa palauttaen ($r = k = 3$).

Taulukko 2.1. Erilaiset valinnat palauttaen, kun $r = k = 3$.

Tulos	$\odot$	$\ominus$	$\oplus$	
$\odot\odot\odot$	***			***||
$\odot\odot\ominus$	**	*		**|*|
$\odot\odot\oplus$	**		*	**||*
$\odot\ominus\ominus$	*	**		*|**|
$\odot\ominus\oplus$	*	*	*	*|*|*
$\odot\oplus\oplus$	*		**	*||**
$\ominus\ominus\ominus$		***		|***|
$\ominus\ominus\oplus$		**	*	|**|*
$\ominus\oplus\oplus$		*	**	|*|**
$\oplus\oplus\oplus$			***	||***

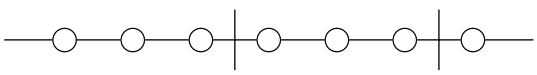
Jokaisen valinnan jälkeen pistetään merkki "$*$" kyseisen pallon kohdalle. Kaikkien valintojen jälkeen meillä on 3 (r) merkkiä. Huomaa, että r voi olla suurempi kuin k, vaikka esimerkissä $r = k = 3$. Taulukon 2.1 viimeisellä sarakkeella pallojen valinta on esitetty merkkien "$*$" ja "$|$" jonona ilmeisellä tavalla. Jonossa on yhteensä 5 ($= r + k - 1$) merkkiä. Kuinka monella tavalla 2 "$|$"-merkkiä voi jakaa 3 "$*$"-merkkiä ryhmiin? Vastaus on $\frac{5!}{3!\,2!} = \binom{5}{2} = 10$. Vastaavasti yleisessä tapauksessa erilaisten tulosvaihtoehtojen lukumäärä on

$$\binom{k+r-1}{r} = \binom{k+r-1}{k-1}.$$

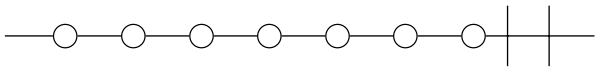
Esimerkki 2.4 Tarkastellaan yhtälöä

$$x_1 + x_2 + \cdots x_k = r, \tag{2.4.2}$$

missä k ja r ovat annettuja positiivisia kokonaislukuja ja muuttujat x_1, x_2, $\ldots$, x_k voivat saada arvoikseen epänegatiivisia kokonaislukuja. Montako erilaista ratkaisua yhtälöllä (2.4.2) on? Olkoon esimerkiksi $r = 7$ ja $k = 3$. Tarkastellaan kysymystä 'helmitaululla',



jossa on esitetty ratkaisu $x_1 = 3$, $x_2 = 3$, $x_3 = 1$. Ratkaisu $x_1 = 7$, $x_2 = 0$, $x_3 = 0$ on helmitaululla muotoa



Helmitaululla on 7 helmeä ($r = 7$) ja 2 jakoviivaa ($k - 1 = 2$) eli yhteensä 9 ($k + r - 1 = 9$) objektia. Jakoviivat voidaan sijoittaa helmitaululla $\binom{9}{2} = 36$ tavalla. Analogisesti voimme päätellä, että yhtälön (2.4.2) epänegatiivisten kokonaislukuratkaisujen lukumäärä on $\binom{k+r-1}{k-1}$. □

2.4.5 Kombinatoriikan merkintöjä ja identiteettejä

Olkoon r epänegatiivinen kokonaisluku ja n reaaliluku. Nyt n:n r-kertoma määritellään kaikille reaaliluvuille samalla tavalla kuin epänegatiivisille kokonaisluvuille:

$$\begin{aligned} &n^{(r)} = n(n-1)(n-2)\cdots(n-r+1), \qquad r > 0; \\ &n^{(0)} = 1. \end{aligned} \tag{2.4.3}$$

Jos n on epänegatiivinen kokonaisluku, niin $n^{(r)}$ on n:n alkion joukon $\{1, 2, \ldots, n\}$ kaikkien r:n ($r \leq n$) kokoisten *järjestettyjen osajoukkojen lukumäärä.* Eritysesti n-kertoma on

$$n! = n^{(n)} = n(n-1)(n-2)\cdots 2 \cdot 1$$

ja 0-kertoma määritellään $0! = 0^{(0)} = 1$.

Olkoon r epänegatiivinen kokonaisluku ja n reaaliluku. Määritellään

$$\binom{n}{r} = \begin{cases} \dfrac{n^{(r)}}{r!} = \dfrac{n!}{(n-r)!\,r!}, & r \geq 0; \\ 0, & r < 0. \end{cases} \tag{2.4.4}$$

Huomaa, että yllä esitetyt $n^{(r)}$ ja $\binom{n}{r}$ on määritelty kaikilla reaaliluvuilla $n \in \mathbb{R}$. Lausekkeille esitettiin kombinatorinen tulkinta, kun n on positiivinen kokonaisluku.

Esimerkki 2.5 Kertoman ja binomikertoimen laskuesimerkkejä:

$$\begin{aligned} 3^{(5)} &= 3 \cdot 2 \cdot 1 \cdot 0 \cdot (-1) = 0 \\ (0.5)^{(4)} &= 0.5 \cdot (-0.5)(-1.5)(-2.5) = -0.9375 \\ \binom{3}{-1} &= 0 \qquad \text{määritelmän perusteella} \\ \binom{n}{0} &= \frac{n^{(0)}}{0!} = 1 \qquad \text{kaikilla } n \in \mathbb{R} \\ \binom{0.5}{4} &= \frac{0.5^{(4)}}{4!} = \frac{0.5 \cdot (-0.5)(-1.5)(-2.5)}{6} = -\frac{5}{128} \\ \binom{-2}{3} &= \frac{-2^{(3)}}{3!} = \frac{(-2)(-3)(-4)}{6} = -4. \end{aligned}$$

□

Määritelmien perusteella on suoraviivaista todeta, että

$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r},$$

$$r\binom{n}{r} = n\binom{n-1}{r-1}. \tag{2.4.5}$$

Jos s on epänegatiivinen kokonaisluku, niin silloin

$$r^{(s)}\binom{n}{r} = n^{(s)}\binom{n-s}{r-s}. \tag{2.4.6}$$

Stirlingin kaava

$$n! \approx \sqrt{2\pi n} \cdot n^n \mathrm{e}^{-n}. \tag{2.4.7}$$

antaa kertomalle hyvän likiarvon.

2.4.6 Binomilause, hypergeometrinen identiteetti ja multinomilause

Lause 2.6 (Binomilause) *Olkoon n mikä tahansa positiivinen kokonaisluku. Silloin*

$$(1+t)^n = \sum_{r=0}^{n} \binom{n}{r} t^r \tag{2.4.8}$$

kaikilla reaaliluvuilla $t \in \mathbb{R}$.

Kertoimia $\binom{n}{r}$ kutsutaan binomikertoimiksi. Binomisarja on binomilauseen yleistys.

Lause 2.7 (Binomisarja) *Olkoon $\alpha \in \mathbb{R}$ nollasta poikkeava reaaliluku. Silloin sarja*

$$(1+t)^\alpha = \sum_{r=0}^{\infty} \binom{\alpha}{r} t^r \tag{2.4.9}$$

suppenee kaikilla $|t| < 1$ ja hajaantuu, kun $|t| > 1$. Jos $\alpha > -1$, niin sarja suppenee myös pisteessä $t = +1$, ja jos $\alpha > 0$, niin sarja suppenee myös pisteessä $t = -1$.

Lause 2.8 (Hypergeometrinen identiteetti) *Olkoot a ja b reaalilukuja ja n positiivinen kokonaisluku. Silloin*

$$\sum_{r=0}^{\infty} \binom{a}{r}\binom{b}{n-r} = \binom{a+b}{n}. \tag{2.4.10}$$

Lause 2.9 (Multinomilause) *Olkoon annettu positiivinen kokonaisluku n ja reaaliluvut $t_1, t_2, \ldots, t_k$. Silloin*

$$(t_1 + t_2 + \cdots + t_k)^n = \sum \sum \cdots \sum \binom{n}{n_1\ n_2\ \ldots\ n_k} t_1^{n_1} t_2^{n_2} \cdots t_k^{n_k}, \tag{2.4.11}$$

missä summa käy yli kaikkien sellaisten epänegatiivisten kokonaislukujen n_1, $n_2, \ldots, n_k$, että $n_1 + n_2 + \cdots + n_k = n$.

2.4.7 Gammafunktio

Gammafunktio Γ määritellään integraalina

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x}\, \mathrm{d}x \tag{2.4.12}$$

kaikilla positiivisilla reaaliluvuilla α. Epäoleellinen integraali (2.4.12) suppenee, kun $\alpha > 0$. Tällä funktiolla on tärkeitä sovelluksia ei vain tilastotieteessä ja todennäköisyyslaskennassa, vaan yleisemminkin sovelletussa matematiikassa.

Kun $\alpha > 1$, voidaan (2.4.12) lausua osittaisintegraalina

$$\int_0^\infty x^{\alpha-1} \mathrm{e}^{-x}\, \mathrm{d}x = \int_0^\infty (-x^{\alpha-1} \mathrm{e}^{-x}) + (n-1) \int_0^\infty x^{\alpha-2} \mathrm{e}^{-x}\, \mathrm{d}x,$$

josta seuraan relaatio

$$\Gamma(\alpha + 1) = \alpha\, \Gamma(\alpha). \tag{2.4.13}$$

Jos $\alpha = n$ on positiivinen kokonaisluku, niin

$$\Gamma(n+1) = n\, \Gamma(n) = n(n-1) \cdots 2 \cdot 1 \cdot \Gamma(1) = n!\, \Gamma(1).$$

Koska

$$\Gamma(1) = \int_0^\infty \mathrm{e}^{-x}\, \mathrm{d}x = \int_0^\infty (-\mathrm{e}^{-x}) = 1,$$

niin positiivisilla kokonaisluvuilla

$$\Gamma(n+1) = n!$$

Tämän ominaisuuden perusteella gammafunktiota kutsutaan joskus *yleistetyksi kertomaksi.* Voidaan myös osoittaa, että $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, joten

$$\Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right) \cdots \frac{1}{2}\sqrt{\pi},$$

kun n on positiivinen kokonaisluku.

2.5 Satunnaismuuttuja

Satunnaiskokeiden tulokset esitetään tavallisesti numeeristen muuttujien avulla. Nämä muuttujat luonnehtivat tarkasteltavan satunnaiskokeen tuloksia. Tilastollisen tarkastelun kannalta onkin oleellista osata määritellä 'oikeat' muuttujat.

Määritelmä 2.2 Olkoon Ω jonkin satunnaiskokeen otosavaruus. *Satunnaismuuttuja* (SM) X on kuvaus (funktio) Ω:lta reaalilukujen joukkoon $\mathbb{R}$.

Satunnaismuuttujia merkitään isoilla kirjaimilla X, Y, Z, ... Voimme kirjoittaa

$$X\colon \Omega \to \mathbb{R},$$

missä $X(\omega)$ on reaaliluku. Satunnaismuuttuja X liittää siis jokaiseen alkeistapaukseen $\omega \in \Omega$ yhden ja vain yhden reaaliluvun $X(\omega) \in \mathbb{R}$. Satunnaismuuttujien $X, Y, Z, \ldots$ arvoja merkitään pienillä kirjaimilla $x, y, z, \ldots$ Merkitään siis $X(\omega) = x$. Jos X:n arvojen muodostama joukko $S \subset \mathbb{R}$ (arvojoukko) on numeroituva (äärellinen tai ääretön), niin X on *diskreetti.* Jos otosavaruus Ω on numeroituva, niin Ω:lla määritellyt satunnaismuuttujat ovat välttämättä diskreettejä. Seuraavassa tarkastellaan diskreettejä satunnaismuuttujia.

Esimerkki 2.6 Heitetään harhatonta lanttia 3 kertaa. Satunnaismuuttuja X on 'kruunujen lukumäärä'. Merkitään R = 'kruunu' ja L = 'klaava'. Silloin

ω:	RRR	RRL	RLR	RLL	LRR	LRL	LLR	LLL
$X(\omega)$:	3	2	2	1	2	1	1	0

Silloin esimerkiksi $X(\text{RRL}) = X(\text{RLR}) = 2$. □

Esimerkki 2.7 Mielipidekyselyssä tiedustellaan 100:lta satunnaisesti valitulta suomalaiselta, millainen kanta heillä on Suomen NATO-jäsenyyteen. Mahdolliset kannanotot ovat: kannattaa (K), ei kantaa (E) ja vastustaa (V). Mahdollisten vastausten lukumäärä eli otosavaruuden koko on silloin 3^{100}. Jos olemme kuitenkin kiinnostuneita, esimerkiksi 'kannattajien lukumäärästä' X, niin silloin X:n mahdollinen arvojoukko on $\{0, 1, \ldots, 100\}$, jonka alkioiden lukumäärä on 101. Alkeistapaus ω on 100:n pituinen tyyppiä "KEVV-VE...EV" oleva jono. Satunnaismuuttujan X arvo $X(\omega)$ on alkeistapauksesta ω laskettu kannattajien lukumäärä, esimerkiksi 36. □

Olemme jo edellä implisiittisesti soveltaneet satunnaismuuttujan käsitettä. Nopan ja lantin heittoon sekä korttipakkaan liittyvillä satunnaiskokeilla on perinteisesti havainnollistettu todennäköisyyslaskennan käsitteitä. Taulukossa 2.2 on esitetty muutamia tuttuja satunnaismuuttujia.
Määritelmässä 2.2 olemme todenneet, että satunnaismuuttujan arvot ovat reaalilukuja. Näin ei aina välttämättä ole. Esimerkiksi Taulukon 2.2 satunnaismuuttujan W arvo on valitun kortin 'maa'. Nämä arvot voidaan kuitenkin

Taulukko 2.2. Joitakin satunnaismuuttujia ja niiden arvoalueet.

Satunnais-muuttuja	Kuvaus	Arvojoukko S
X	Nopan silmäluku	$\{1, 2, 3, 4, 5, 6\}$
Y	Kruunujen lukumäärä 3:ssa lantin heitossa	$\{0, 1, 2, 3\}$
Z	Heittojen lukumäärä kunnes saadaan 1. kruunu	$\{1, 2, 3 \ldots\}$
W	Korttipakasta satunnaisesti valitun kortin maa	$\{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$

aina tarvittaessa koodata numeerisesti. Joissain yhteyksissä tarkastelemme esimerkiksi satunnaispareja, satunnaisjonoja tai satunnaisjärjestyksiä. Näihin satunnaismuuttujan yleistyksiin palataan tuonnempana.

Huomautus 2.1 Jos X ja Y ovat satunnaismuuttujia, niin

$$aX, \qquad X + Y, \qquad X - Y, \qquad XY \qquad \text{ja} \qquad \frac{X}{Y} \quad (Y \neq 0)$$

ovat satunnaismuuttujia, missä a on reaaliluku. Nämä tulokset seuraavat siitä, että satunnaismuuttuja on funktio.

Matematiikan analyysin kursseilla opitun perusteella tiedämme, että *funktion funktio* on edelleen funktio:

$$x \to \sin(\log x) \qquad \text{tai} \qquad x \to f[h(x)] = (f \circ h)(x).$$

Yhdistetty satunnaismuuttuja on siis edelleen satunnaismuuttuja. Jos W (Taulukko 2.2) on esimerkiksi satunnaisesti valitun kortin maa ja V maiden joukossa $S = \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$ määritelty väri, niin satunnaismuuttujan *kortin väri* $V(W) = V[W(\omega)]$ arvoalue on $S_V = \{\text{musta, punainen}\}$. Korttipakan kortit (52 kpl) muodostavat alkeistapahtumien joukon ω. Olkoon Y kruunujen lukumäärä 3:ssa lantin heitossa (Taulukko 2.2). Silloin esimerkiksi

$$g(Y) = Y - \frac{3}{2} \qquad \text{tai} \qquad h(Y) = \left(Y - \frac{3}{2}\right)^2$$

ovat satunnaismuuttujia.

Kahden tai useamman satunnaismuuttujan funktio on edelleen satunnaismuuttuja. Jos siis X ja Y ovat satunnaismuuttujia, niin

$$\omega \to h[X(\omega), Y(\omega)]$$

määrittelee satunnaismuuttujan, kun kahden muuttujan funktio h on määritelty arvojoukossa $\{\, (X(\omega), Y(\omega)) \mid \omega \in \Omega \,\} \subset \mathbb{R}^2$. Tätä satunnaismuuttujaa merkitään lyhyesti $h(X, Y)$.

Määritelmä 2.3 (Indikaattorifunktio) Olkoon A tapahtuma otosavaruudessa Ω. Tapahtuman A indikaattorifunktio I_A saa arvon 0 tai 1 seuraavasti:

$$I_A(\omega) = \begin{cases} 1, & \text{jos } \omega \in A; \\ 0, & \text{jos } \omega \notin A. \end{cases}$$

Jos tapahtuma A sattuu, niin $I_A = 1$, muutoin $I_A = 0$. Indikaattorifunktio on satunnaismuuttuja ja

$$P(I_A = 1) = P(A) \qquad \text{ja} \qquad P(I_A = 0) = P(A^c) = 1 - P(A).$$

Voimme käyttää indikaattorifunktiota vaikkapa lukumäärien laskemiseen. Heitetään lanttia n kertaa ja olkoon X_k tapahtuman 'kruunu k. heitossa' $(1 \leq k \leq n)$ indikaattorifunktio. Silloin satunnaismuuttuja

$$X = X_1 + X_2 + \cdots + X_n \tag{2.5.1}$$

on kruunujen lukumäärä n:ssä heitossa, koska summa on ykkösten (kruunujen) lukumäärä n:ssä heitossa.

2.5.1 Satunnaismuuttujan jakauma

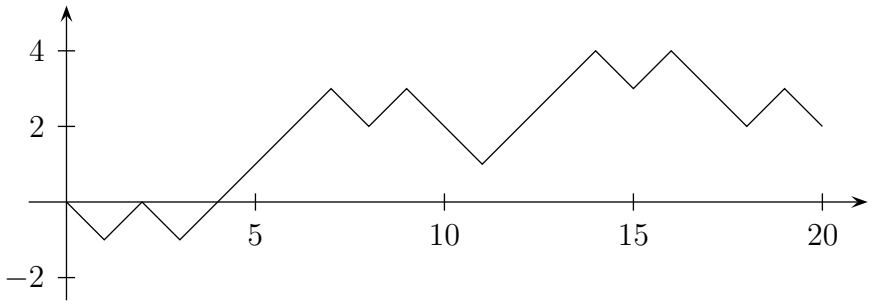
Olkoon Y kruunujen lukumäärä 3:ssa lantin heitossa (Esimerkki 2.6). Silloin satunnaismuuttujaa koskevat väittämät, kuten 'täsmälleen yksi kruunu' $\equiv$ "$Y = 1$" tai 'korkeintaan 2 kruunua' $\equiv$ "$Y \leq 2$" määrittelevät tapahtuman. Tapahtumat voidaan silloin kirjoittaa muodossa "$Y \in A$". Jos $S = \{0, 1, 2, 3\}$, $A = \{1\}$ ja $B = \{0, 1, 2\}$, niin "$Y = 1$" $\equiv$ "$Y \in A$" ja "$Y \leq 2$" $\equiv$ "$Y \in B$". Jatkossa tulemme pääsääntöisesti tarkastelemaan *satunnaismuuttujien määrittämiä tapahtumia.*

Tapahtuman todennäköisyyttä merkitään $P(Y \in B)$ tai lyhyesti $P(B)$. Merkintä $P(Y \in B)$ osoittaa, että tapahtuma on määritelty satunnaismuuttujan Y avulla. Koska B voi olla mikä tahansa Y:n arvoalueen S_Y osajoukko, todennäköisyydet $P(Y \in B)$ määrittelevät *satunnaismuuttujan jakauman.* Jos $y \in S_Y$, silloin Y:n arvon $Y = y$ todennäköisyys on $P(Y = y)$. Vastaavasti tapahtuman "$Y \in B$" todennäköisyys on

$$P(Y \in B) = \sum_{y \in B} P(Y = y).$$

Esimerkki 2.8 (Satunnaiskävely, Random Walk) Pekka ja Paavo pelaavat "kruunua ja klaavaa". Tässä pelissä heitetään peräkkäin lanttia n kertaa – tässä esimerkissä $n = 20$. Aina kun tulee kruunu (R), Pekka voittaa euron Paavolta. Kun tulee klaava (L), Pekka häviää euron Paavolle. Kuviossa 2.2 esitetyn pelin tulos $(n = 20)$ on

L R L R R R R L R L L R R R L R L L R L.



Kuvio 2.2. ”Kruunu ja klaava” -pelin tuloksen kehitys, kun pelin pituus on 20 heittoa.

Pekka voittaa 2 euroa.

Mikä on todennäköisyys, että Pekka voittaa s euroa, kun $n = 20$ ($-20 \leq s \leq 20$)? On helppo nähdä, että mahdollinen voitto on parillinen. Voitto S voidaan määritellä satunnaismuuttujien X_i ($i = 1, 2, \ldots, 20$) summana:

$$S_{20} = X_1 + X_2 + \cdots + X_{20},$$

missä

$$X_i = \begin{cases} 1, & \text{kun kruunu } i. \text{ heitossa;} \\ -1, & \text{kun klaava } i. \text{ heitossa.} \end{cases}$$

Minkä voiton $S_{20} = s$ todennäköisyys on suurin (pienin)?

On mielenkiintoista tarkastella myös sitä, kuinka usein Pekka on voitolla pelin aikana. Jos pelaajat ovat tasoissa (voitto 0), määrittelemme, että Pekka on johdossa, jos hän oli edellisellä heitolla johdossa. Jos Pekka oli tappiolla edellisellä heitolla ja pääsi tasoihin, sovimme, että hän on edelleen tappiolla. Jokainen peli tuottaa vastaavan kuvaajan kuin Kuviossa 2.2. Kuvaajassa on yhdistetty pisteet $(0, 0)$, $(1, S_1)$, $(2, S_2)$, $\ldots$, $(20, S_{20})$.

Tällaista prosessia kutsutaan satunnaiskävelyksi (random walk). Eräs tapa havainnollistaa satunnaiskävelyä on ajatella, että satunnaiskävelijä RW (Random Walker) lähtee origosta (itään) ja astuu sekunnissa askeleen oikealle (etelään) tai vasemmalle (pohjoiseen). Esimerkiksi Kuviossa 2.2 kuvaaja kulkee pisteen $(5, 1)$ kautta. RW on 5 sekunnin kävelyn jälkeen yhden askeleen pohjoiseen $x-$akselista. On helppo todeta, että kaikkien mahdollisten pelin kulkujen lukumäärä on 2^{20}. Koska raha on harhaton ja heitot ovat toisistaan riippumattomat, kaikki 2^{20} pelin kulkua ovat yhtä todennäköiset. □

2.5.2 Kertymäfunktio

Edellä olemme käsitelleet otosavaruuden osajoukkojen eli tapahtumien todennäköisyyksiä. Nämä joukot määritellään tavallisesti satunnaismuuttujien avulla. Esimerkiksi

$$\{a \leq X \leq b\} = \{\, \omega \mid a \leq X(\omega) \leq b \,\}, \tag{2.5.2}$$

missä X on satunnaismuuttuja, a ja b ovat annettuja vakioita. Koska otosavaruus on numeroituva, kaikilla muotoa (2.5.2) olevilla joukoilla on todennäköisyys ja sitä merkitään

$$P(a \leq X \leq b) = P(\{\, \omega \mid a \leq X(\omega) \leq b \,\}).$$

Jos $A \subset \mathbb{R} = (-\infty, \infty)$ on jokin reaalilukujoukko, niin merkitään

$$P(X \in A) = P(\{\, \omega \mid X(\omega) \in A \,\}). \tag{2.5.3}$$

Joukko A voi olla esimerkiksi suljettu väli $[a, b]$, avoin väli (a, b) puoliavoin väli $(a, b]$ tai $[a, b)$, ääretön väli $(-\infty, b]$ tai $[a, \infty)$, usean välin yhdiste tai kokonaislukujen $\{k, k+1, \ldots, k+n\}$ joukko. Yhden pisteen x joukko $\{x\}$ on tärkeä erikoistapaus. Olkoon X esimerkiksi henkilön ikä, kun Ω on suomalaisten joukko. Silloin $\{X = 20\} = \{\, \omega \mid X(\omega) = 20 \,\}$ on 20-vuotiaiden suomalaisten joukko. Todennäköisyys on

$$P(X = 20) = P(X \in \{20\}) = P(\{\, \omega \mid X(\omega) = 20 \,\}).$$

Olkoon X:n arvojoukko $S_X = \{x_1, x_2, \ldots, x_n, \ldots,\}$, missä X:n arvot on lueteltu jossain järjestyksessä. Määritellään

$$p_n = P(X = x_n),$$

kun $x_n \in S_X$. Jos $x \notin S_X$, niin $P(X = x) = 0$. Toisaalta voi olla, että $P(X = x_i) = 0$ jollakin X:n arvolla $x_i \in S_X$. Jos tunnemme kaikki todennäköisyydet p_n, on ilmeistä, että voimme laskea kaikki satunnaismuuttujaan X liittyvät todennäköisyydet. Silloin todennäköisyydet (2.5.2) ja (2.5.3) ovat

$$P(a \leq X \leq b) = \sum_{a \leq x_n \leq b} p_n \qquad \text{ja} \qquad P(X \in A) = \sum_{x_n \in A} p_n.$$

Kun A on ääretön väli $(-\infty, x]$, niin jokaista reaalilukua x kohti voidaan määritellä funktio

$$F_X(x) = P(X \leq x) = \sum_{x_n \leq x} p_n.$$

Tätä funktiota F_X kutsutaan X:n *kertymäfunktioksi* (kf). Jokaiseen satunnaismuuttujaan X liittyy siis kertymäfunktio, jota merkitään $F_X(x)$.

Määritelmä 2.4 (Kertymäfunktio) Satunnaismuuttujan X kertymäfunktio $F_X(x)$ on sellainen kuvaus $F_X \colon \mathbb{R} \to [0, 1]$, että

$$F_X(x) = P(X \leq x), \qquad \text{kaikilla } x \in \mathbb{R}.$$

Merkintä $P(X \leq x)$ on lyhennys merkinnästä $P(\{X \leq x\})$, missä $P(\{X \leq x\}) = P(\{\, \omega \mid X(\omega) \leq x \,\})$. Merkitään X:n arvoaluetta $S = X(\Omega) = \{\, x \mid X(\omega),\ \omega \in \Omega \,\}$.

Esimerkki 2.9 Esimerkissä 2.6 heitettiin harhatonta lanttia 3 kertaa. Satunnaismuuttuja X on 'kruunujen lukumäärä' ja X:n arvojoukko $S = \{0, 1, 2, 3\}$. Nyt

$$\begin{aligned}
\{\,\omega \mid X = 0\,\} &= \{\mathrm{LLL}\},\\
\{\,\omega \mid X = 1\,\} &= \{\mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}\},\\
\{\,\omega \mid X = 2\,\} &= \{\mathrm{RRL}, \mathrm{LRR}, \mathrm{RLR}\},\\
\{\,\omega \mid X = 3\,\} &= \{\mathrm{RRR}\},
\end{aligned}$$

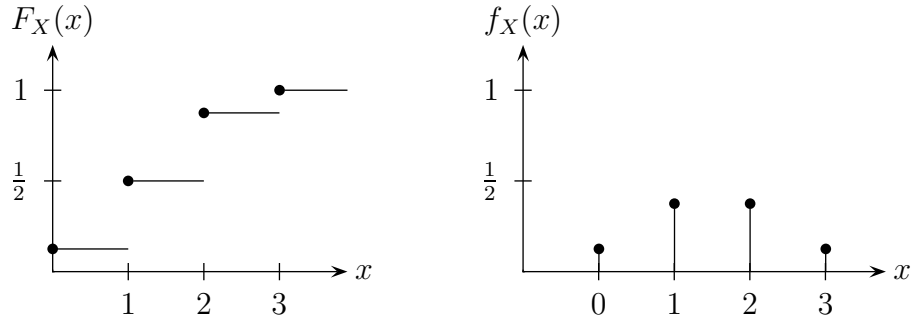
missä kaikki alkeistapaukset ovat yhtä todennäköisiä. Silloin

$$\begin{aligned}
F_X(0) &= P(X \le 0) = P(\{\mathrm{LLL}\}) = 1/8,\\
F_X(1) &= P(X \le 1) = P(\{\mathrm{LLL}, \mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}\}) = 4/8,\\
F_X(2) &= P(X \le 2)\\
&= P(\{\mathrm{LLL}, \mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}, \mathrm{RRL}, \mathrm{LRR}, \mathrm{RLR}\}) = 7/8,\\
F_X(3) &= P(X \le 3)\\
&= P(\{\mathrm{LLL}, \mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}, \mathrm{RRL}, \mathrm{LRR}, \mathrm{RLR}, \mathrm{RRR}\}) = 1.
\end{aligned}$$

Satunnaismuuttujan X kertymäfunktio on siis

$$F_X(x) = \begin{cases} 0, & \text{kun } x < 0;\\ \frac{1}{8}, & \text{kun } 0 \le x < 1;\\ \frac{4}{8}, & \text{kun } 1 \le x < 2;\\ \frac{7}{8}, & \text{kun } 2 \le x < 3;\\ 1, & \text{kun } x \ge 3. \end{cases}$$

□



Kuvio 2.3. Satunnaismuuttujan X kertymäfunktion $F_X(x)$ ja todennäköisyysfunktion $f_X(x)$ kuvaajat.

Jos $x_1 \le x_2$, niin $\{X \le x_1\} \subset \{X \le x_2\}$ ja todennäköisyyden monotonisuusominaisuuden perusteella [Lause (2.1), kohta 3)] $P(X \le x_1) \le P(X \le x_2)$, joten kertymäfunktio $F(x)$ on *kasvava* (ei vähenevä). Seuraavassa lauseessa esitetään kertymäfunktion ominaisuudet.

Lause 2.10 *Funktio $F(x)$ on satunnaismuuttujan X kertymäfunktio, jos ja vain jos seuraavat kolme ehtoa pitävät paikkansa:*

1. $\lim_{x\to-\infty} F(x) = 0$ *ja* $\lim_{x\to\infty} F(x) = 1$.
2. $F(x)$ *on x:n kasvava funktio.*
3. $F(x)$ *on oikealta jatkuva, ts. kaikilla $x_0 \in \mathbb{R}$ on* $\lim_{x\to x_0+} F(x) = F(x_0)$ *($x \to x_0+$ tarkoittaa, että x_0:aa lähestytään oikealta).*

Esimerkissä 2.9 esitetty kertymäfunktio on porrasfunktio. Pitää yleisestikin paikkansa, että diskreetin satunnaismuuttujan kertymäfunktio on porrasfunktio. Voimme sanoa, että satunnaismuuttuja on diskreetti satunnaismuuttuja, jos sen kertymäfunktio on porrasfunktio.

2.5.3 Diskreetin satunnaismuuttujan todennäköisyysfunktio

Olkoon X otosavaruudessa Ω määritelty diskreetti satunnaismuuttuja. Satunnaismuuttujan X todennäköisyysfunktio (tnf) $f_X(x)$ määritellään siten, että

$$f_X(x) = P(X = x). \tag{2.5.4}$$

Jos merkitään X:n arvoaluetta $S_X = X(\Omega) = \{\, x \mid X(\omega) = x,\ \omega \in \Omega \,\}$, niin todennäköisyysfunktio on kuvaus

$$f_X(x)\colon S_X \to [0, 1].$$

Huomattakoon, että $f_X(x)$ on määritelty kaikilla reaaliluvuilla, mutta $f_X(x) = 0$ aina, kun $x \notin S$. Diskreetin satunnaismuuttujan arvojoukko on numeroituva, joten arvoja on korkeintaan yhtä paljon kuin kokonaislukuja. Niinpä diskreettien satunnaismuuttuien arvojoukko on tavallisimmin jokin kokonaislukujen ja erityisesti positiivisten kokonaislukujen osajoukko.

Olkoon X diskreetti satunnaismuuttuja, jonka arvoalue on S_X. Silloin X:n todennäköisyysfunktio toteuttaa seuraavat ehdot:

$$f_X(x) > 0, \qquad \text{kun} \quad x \in S_X; \tag{2.5.5}$$

$$f_X(x) = 0, \qquad \text{kun} \quad x \notin S_X; \tag{2.5.6}$$

$$\sum_{x\in S_X} f_X(x) = 1. \tag{2.5.7}$$

Esimerkki 2.10 Jatketaan esimerkiä 2.9, jossa heitettiin harhatonta lanttia 3 kertaa. Satunnaismuuttuja X on 'kruunujen lukumäärä'. Määritetään X:n

todennäköisyysfunktio. Nyt

$$\begin{aligned}
X^{-1}(0) &= \{\mathrm{LLL}\},\\
X^{-1}(1) &= \{\mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}\},\\
X^{-1}(2) &= \{\mathrm{RRL}, \mathrm{LRR}, \mathrm{RLR}\},\\
X^{-1}(3) &= \{\mathrm{RRR}\},
\end{aligned}$$

missä merkintä $X^{-1}(x) = \{\, \omega \mid X(\omega) = x \,\}$ on kaikkien sellaisten alkeistapausten ω joukko, jotka kuvautuvat pisteeseen x. Koska alkeistapaukset ovat yhtä todennäköisiä, satunnaismuuttujan arvojen todennäköisyydet ovat

$$\begin{aligned}
P(X = 0) &= P(X^{-1}(0)) = P(\{\mathrm{LLL}\}) = 1/8,\\
P(X = 1) &= P(X^{-1}(1)) = P(\{\mathrm{RLL}, \mathrm{LRL}, \mathrm{LLR}\} = 3/8,\\
P(X = 2) &= P(X^{-1}(2)) = P(\{\mathrm{RRL}, \mathrm{LRR}, \mathrm{RLR}\} = 3/8,\\
P(X = 3) &= P(X^{-1}(3)) = P(\{\mathrm{RRR}\} = 1/8.
\end{aligned}$$

Satunnaismuuttujan X todennäköisyysfunktio on siis

$$f_X(x) = \begin{cases} \frac{1}{8}, & \text{kun } x = 0;\\ \frac{3}{8}, & \text{kun } x = 1;\\ \frac{3}{8}, & \text{kun } x = 2;\\ \frac{1}{8}, & \text{kun } x = 3;\\ 0, & \text{muutoin.} \end{cases}$$

□

Jos X on diskreetti satunnaismuuttuja, niin X:n arvojoukko $S_X = \{x_1, x_2, \ldots\}$. Silloin X:n kertymäfunktio voidaan lausua todennäköisyysfunktion avulla seuraavasti:

$$F_X(x) = \sum_{x_i \le x} f_X(x_i),$$

joka on porrasfunktio. Porrasfuntion $F(x)$ 'hyppäykset' ovat pisteissä x_1, x_2, ... ja hyppäysten suuruudet ovat $f_X(x_1)$, $f_X(x_2)$, ... Jos esimerkiksi

$$f(x) = \frac{1}{2^x}, \qquad x = 1, 2, \ldots,$$

niin esimerkiksi

$$F(3.5) = P(x \le 3.5) = f(1) + f(2) + f(3) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}.$$

2.5.4 Diskreetti tasajakauma

Olkoon X satunnaismuuttuja, jonka arvojoukko on $S = \{x_1, x_2, \ldots, x_N\}$. Jos

$$f(x_k) = \frac{1}{N} \qquad \text{kaikilla } k = 1, 2, \ldots, N,$$

niin X noudattaa diskreettiä tasajakaumaa ja merkitään $X \sim \text{Tasd}(x_1, x_2, \ldots, x_N)$. Hyvin usein X:n arvojoukko on $S = \{1, 2, \ldots, N\}$, jolloin merkitään $X \sim \text{Tasd}(1, 2, \ldots, N)$. Esimerkiksi nopanheitossa silmäluvun X arvojoukko on $S = \{1, 2, 3, 4, 5, 6\}$ ja todennäköisyysfunktio

$$f(x) = \frac{1}{6}, \qquad x = 1, 2, 3, 4, 5, 6.$$

2.6 Otanta palauttamatta

Tarkastellaan nyt koetta, jossa valitaan n alkiota N:n alkion joukosta ($n \leq N$), jota kutsutaan *populaatioksi.* Valintaprosessia kutsutaan *otannaksi.* Halutaan esimerkiksi tietää ennen presidentin vaaleja, mikä on ehdokkaiden kannatus. Kannatuksesta voi saada tietoa tiedustelemalla äänestäjiltä, ketä he aikovat äänestää. On käytännössä mahdotonta haastatella kaikkia potentiaalisia äänestäjiä. Siksi tehdään otos, eli valitaan vain osa mahdollisista äänestäjistä ja haastatellaan heidät. Populaation muodostavat siis äänioikeutut kansalaiset. Nimitämme menetelmää, jolla otos valitaan, *otantamenetelmäksi.* Tässä esityksessä tarkastellaan vain *yksinkertaista satunnaisotantaa* (YSO). YSO:ssa kaikki mahdolliset n:n kokoiset otokset ovat yhtä todennäköisiä. YSO:lla valittua otosta kutsutaan *yksinkertaiseksi satunnaisotokseksi.*

Yksinkertaisessa satunnaisotannassa *palauttamatta* otos valitaan siten, että kukin alkio voi tulla otokseen korkeintaan kerran. Valitaan n:n alkion otos N:stä. Ajatellaan alkiot valituiksi järjestyksessä. Silloin 1. alkio voidaan valita N:llä tavalla ja 2. alkio $(N-1)$:llä tavalla, koska toisen täytyy olla eri alkio kuin ensimmäinen, jne. Lopulta n. alkio voidaan valita $[N-(n-1)]$:llä tavalla. Kaikkien mahdollisten *järjestettyjen otosten* lukumäärä on

$$N(N-1)(N-2)\cdots(N-n+1) = N^{(n)}.$$

Otos on *valittu satunnaisesti,* jos jokainen $N^{(n)}$:stä järjestetystä otoksesta on yhtä todennäköinen. Silloin jokaisen järjestetyn otoksen todennäköisyys on $1/N^{(n)}$.

Koska kaikkien mahdollisten otosten eli osajoukkojen lukumäärä on $\binom{N}{n}$ ja otokset oletetaan yhtä todennäköisiksi, niin jokaisen otoksen todennäköisyys on

$$\frac{1}{\binom{N}{n}}, \qquad \text{missä otosten lukumäärä on } \binom{N}{n} = \frac{N^{(n)}}{n!}.$$

Otoksia on siis $\binom{N}{n}$ kappaletta ja YSO:ssa ne ovat yhtä todennäköisiä.

Esimerkki 2.11 Monissa korttipeleissä jaetaan n:n kortin käsi (otos) pakasta (populaatio), jossa on N korttia. Pakka on *hyvin sekoitettu,* jos pakan korttien kaikki $N!$ järjestystä ovat yhtä todennäköisiä. Oletetaan, että n korttia on jaettu hyvin sekoitetusta pakasta. Sellaisia pakan järjestyksiä, joissa

nämä n korttia ovat tietyssä järjestyksessä (esimerkiksi pakan päällä), on $(N-n)!$ kappaletta. Todennäköisyys saada n korttia tietyssä järjestyksessä on

$$\frac{(N-n)!}{N!} = \frac{1}{N^{(n)}}.$$

Jokaisen järjestetyn otoksen todennäköisyys on siis $1/N^{(n)}$ ja jokaisen otoksen eli käden todennäköisyys on $1/\binom{N}{n}$.

Mikä on esimerkiksi todennäköisyys, että tavallisesta korttipakasta ($N = 52$) saadaan patasuora (♠1, ♠2, ♠3, ♠4, ♠5)? Erilaisten viiden käsien lukumäärä on $\binom{52}{5} = 2598960$, joten patasuoran todennäköisyys on $1/2598960$. Jos lisäksi korttien pitää jaossa tulla annetussa järjestyksessä $(1, 2, 3, 4, 5)$, niin järjestettyjen otosten lukumäärä $52^{(5)} = 311875200$ ja järjestetyn otoksen (♠1, ♠2, ♠3, ♠4, ♠5) todennäköisyys on $1/311875200$. □

2.6.1 Hypergeometrinen jakauma

Oletetaan, että populaatiossa on $a + b$ alkiota – esimerkiksi väestössä on a miestä ja b naista tai tuotepopulaatiossa on a viallista ja b virheetöntä tuotetta. Valitaan populaatiosta n:n kokoinen satunnaisotos palauttamatta. Mikä on todennäköisyys, että otokseen tulee x kappaletta tyyppiä 1 olevia alkiota ja $n - x$ kappaletta tyyppiä 2? Tavanomainen todennäköisyyslaskennassa käytetty satunnaiskoe on pallojen valinta uurnasta. Uurnassa on a valkoista ja b mustaa palloa. Valitaan uurnasta satunnaisesti palauttamatta n palloa. Mikä on todennäköisyys, että otokseen tulee x valkoista palloa?

Taulukko 2.3. Valinta palauttamatta äärellisestä populaatiosta

	Tyyppi 1	Tyyppi 2	Yhteensä
Populaatio	a	b	$a+b$
Otos	x	$n-x$	n

Populaatiosta voidaan valita kaikkiaan $\binom{a+b}{n}$ yhtä todennäköistä n:n kokoista otosta palauttamatta. Koska a:sta tyyppiä 1 olevasta alkiosta voidaan valita x kappaletta $\binom{a}{x}$ tavalla ja $n - x$ alkiota b:sta $\binom{b}{n-x}$ tavalla, saadaan kaikkiaan $\binom{a}{x}\binom{b}{n-x}$ sellaista otosta, joissa on x kappaletta tyyppiä 1 ja $n - x$ kappaletta tyyppiä 2 olevaa alkiota. Olkoon nyt satunnaismuuttuja X tyyppiä 1 olevien alkioiden lukumäärä otoksessa. Silloin satunnaismuuttujan X todennäköisyysfunktio $f(x)$ on

$$f(x) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}, \qquad x = 0, 1, 2, \ldots \tag{2.6.1}$$

Edellä esitetystä otanta-asetelmasta tietysti seuraa, että ehtojen $x \leq a$ ja $n - x \leq b$ täytyy olla voimassa. Jos ehdot eivät ole voimassa, niin $f(x) =$

0. Jakaumaa (2.6.1) kutsutaan *hypergeometriseksi jakaumaksi.* Se on tärkeä myös esimerkiksi joidenkin ns. tarkkojen testien konstruoinnissa.

2.6.2 Tarkistusotanta teollisuudessa

Mikään teollisuusprosessi ei ole täydellinen, siksi myös virheellisiä tuotteita on odotettavissa. Yrityksillä on käytössä erilaisia laadunvarmistusjärjestelmiä, jotta voitaisiin pitää yllä riittävän hyvä laatu. Virheelliset tuotteet olisi havaittava ja poistettava, jotta ne eivät joutuisi asiakkaalle saakka. Tietysti voitaisiin tarkistaa jokainen tuote riittävän tarkasti. Täydellinen tarkistus ei ole käytännössä yleensa realistinen – se ei ole taloudellisesti kannattavaa tai se on jopa mahdotonta, jos tarkistus esimerkiksi tuhoaa tuotteen. On siis yleensä käytettävä tarkistusotantaa.

Oletetaan, että tuotteet ovat joko virheellisiä tai hyväksyttäviä ja ne tulevat laadun tarkistukseen N:n tuotteen erissä. Valitaan jokaisesta erästä satunnaisesti n tuotetta tarkistukseen. Oletetaan, että löydetään x viallista. Jos x on suuri, todennäköisesti erässä on paljon viallisia ja erä pitäisi hylätä tai panna jatkotarkistukseen. Voimme käyttää päätössääntöä:

$$\text{Hyväksy erä, jos } x \leq h, \text{ muutoin hylkää erä (tai testaa lisää).}$$

Nyt olisi valittava hyväksymisraja h mahdollisimman viisaasti. On tietysti mahdollista, että otoksessa $x > h$, vaikka viallisten lukumäärä v tuote-erässä ei olisikaan ”liian” suuri. Toisaalta voi ehto $x \leq c$ toteutua, vaikka tuote-erässä olisi ”liikaa” viallisia. Edellä mainitut päätäntävirheet ovat siis seuraavat:

1. lajin virhe – erä, jossa on vähän viallisia, hylätään;

2. lajin virhe – erä, jossa on paljon viallisia, hyväksytään.

Jos hyväksymisrajaa h kasvatetaan, pienenee 2. lajin virhe, mutta 1. lajin virhe kasvaa. Molempia virheitä voidaan pienentää samanaikaisesti, kasvattamalla otoskokoa n, mutta se taas nostaa tarkistuskustannuksia. Jotta h:n ja n:n arvot voitaisiin määrittää optimaalisesti, olisi tunnettava tarkistuskustannukset sekä 1. ja 2. lajin virheiden aiheuttamat kustannukset. Olisi myös tiedettävä virheelisten lukumäärän v jakauma yli tuote-erien.

2.7 Otanta palauttaen

Valitaan n:n alkion otos populaatiosta, jossa on N alkiota. Ajattelemme, että populaation alkiot on numeroitu juoksevasti $\{1, 2, \ldots, N\}$. Otannassa palauttaen populaation alkio voidaan valita otokseen useammin kuin kerran. On esimerkiksi mahdollista, että otokseen tulee sama alkio toistuvasti n kertaa. Voimme ajatella valinnan prosessina, jossa alkiot valitaan peräkkäin. Jokaisen valinnan jälkeen alkio palautetaan populaatioon, mutta sitä ennen

saatu alkio merkitään muistiin. Silloin 1. alkio voidaan valita N:llä tavalla, 2. alkio myös N:llä tavalla ja lopulta n. alkio N:llä tavalla, koska edellisissä valinnoissa valitut voivat tulla uudestaan otokseen. Kaikkien mahdollisten palauttaen valittujen *järjestettyjen otosten* lukumäärä on siis N^n. Sanomme, että otos on *valittu satunnaisesti palauttaen,* jos kaikki mahdolliset N^n järjestettyä jonoa ovat yhtä todennäköiset. Näin valittu otos on *yksinkertainen satunnaisotos* (*YSO*) *palauttaen.*

Oletetaan esimerkiksi, että valitaan kolme numeroa palauttaen numeroista 0, 1, 2, ..., 9. Silloin voidaan saada $10^3 = 1000$ yhtä mahdollista järjestettyä jonoa 000, 001, 002, ..., 999. Osajoukko $\{1, 2, 3\}$ voidaan valita $3! = 6$ tavalla, joten otoksen $\{1, 2, 3\}$ todennäköisyys on 0.006. Otos $\{1, 1, 3\}$ voidaan saada 3:lla tavalla, koska järjestetyt jonot $(1, 1, 3)$, $(1, 3, 1)$ ja $(3, 1, 1)$ sisältävät samat alkiot. Otoksen $\{1, 1, 3\}$ todennäköisyys on 0.003. Otos $\{1, 1, 1\}$ saadaan vain yhdellä tavalla, joten sen todennäköisyys on 0.001. Otannassa palauttaen (järjestämättömät) otokset *eivät ole yhtä todennäköisiä* kuten otannassa palauttamatta.

Olkoon A_i tapahtuma, että valitaan i. alkio, $i = 1, 2, \ldots, N$. Koska valinnan (kokeen) tulos on varmasti yksi ja vain yksi tapahtumista A_1, A_2, ..., A_N, niin $\Omega = A_1 \cup A_2 \cup \cdots \cup A_N$ on kokeeseen (valinta palauttaen) liittyvän otosavaruuden ositus. Valinta toistetaan n kertaa. Oletetaan, että populaation i. alkio toistuu otoksessa n_i $(0 \leq n_i \leq n)$ kertaa $(i = 1, 2, \ldots, N)$. Silloin $\sum_{i=1}^{n} n_i = n$ ja erilaisten järjestettyjen otosten lukumäärä on tuloksen (2.4.1) mukaan

$$\binom{n}{n_1\ n_2\ \ldots\ n_N} = \frac{n!}{n_1!\, n_2! \cdots n_N!}.$$

Olkoon X_i alkion i toistojen lukumäärä otoksessa. Nyt siis jokaisen X_i:n arvoalue on $\{0, 1, 2, \ldots, n\}$ ja $X_1 + X_2 + \cdots + X_N = n$. Merkitään todennäköisyyttä $P(X_1 = n_1, X_2 = n_2, \ldots, X_N = n_N)$ yksinkertaisesti $P(n_1, n_2, \ldots, n_N)$, joka voidaan siis laskea kaavalla

$$P(n_1, n_2, \ldots, n_N) = \binom{n}{n_1\ n_2\ \ldots\ n_N} \frac{1}{N^n}.$$

Esimerkki 2.12 Valitaan populaatiosta $\{A_1, A_2, A_3\}$ $(N = 3)$ 5 kertaa $(n = 5)$ alkio palauttaen. Silloin $A_1A_1A_3A_1A_3$ on eräs mahdollinen tulosjono (otos palauttaen), missä $X_1 = 3$, $X_2 = 0, X_3 = 2$ ja $X_1 + X_2 + X_3 = 5$. Jonon $A_1A_1A_3A_1A_3$ todennäköisyys, samoin kuin jokaisen viiden pituisen järjestetyn otoksen, todennäköisyys on $1/3^5$. Koska erilaisia tulosjonoja, joissa $X_1 = 3$, $X_2 = 0$ ja $X_3 = 2$, on

$$\binom{5}{3\ 0\ 2} = \frac{5!}{3!\, 0!\, 2!} = 10,$$

niin

$$P(X_1 = 3, X_2 = 0, X_3 = 2) = \binom{5}{3\ 0\ 2} \frac{1}{3^5} = \frac{10}{243} = 0.04115.$$

Mikä on todennäköisyys, että n:n kokoiseen järjestettyyn otokseen tulee populaation n ensimmäistä alkiota ($n \leq N$) missä tahansa järjestyksessä? Kyseinen tapahtuma sattuu täsmälleen silloin, kun $X_1 = X_2 = \cdots = X_n = 1$ ja $X_{n+1} = \cdots = X_N = 0$. Tämän tapahtuman todennäköisyys on siis

$$P(1, 1, \ldots, 1, 0, \ldots, 0) = \frac{n!}{(1!)^n (0!)^{N-n}} \frac{1}{N^n} = \frac{n!}{N^n}.$$

□

Otoksen kaikki alkiot erilaisia

Sellaisia järjestettyjä otoksia, joissa mikään alkio ei toistu, on

$$N^{(n)} = N(N-1)\cdots(N-n+1)$$

kappaletta. Jos otos valitaan palauttaen, niin todennäköisyys, että otoksessa mikään alkio ei toistu, on

$$P(\text{'Sama ei toistu'}) = \frac{N^{(n)}}{N^n} = \frac{N!}{(N-n)!\, N^n}. \tag{2.7.1}$$

On selvää, että todennäköisyys (2.7.1) on 0, jos $n > N$. Huomaa, että $N^{(n)} = 0$, jos $n > N$. Syntymäpäiväongelmassa (Esimerkki 2.3) $N = 365$ ja $n = r$.

Soveltamalla Stirlingin kaavaa (2.4.7) kertoimiin $N!$ ja $(N-1)!$ saadaan likiarvo

$$\frac{N!}{(N-n)!\, N^n} \approx \left(\frac{N}{N-n}\right)^{N-n+1/2} \mathrm{e}^{-n}. \tag{2.7.2}$$

Kun $N \to \infty$ ja n on kiinnitetty, niin lauseke (2.7.2) lähestyy ykköstä. Jos siis hyvin suuresta populaatiosta valitaan n alkiota ($n \ll N$) palauttaen, niin on hyvin epätodennäköistä, että sama alkio valitaan usemmin kuin kerran. Otanta palauttaen ja palauttamatta ovat käytännöllisesti katsoen jokseenkin identtiset, kun populaation koko N on paljon suurempi kuin otoskoko n.

2.8 Binomijakauma

Oletetaan, että populaatiossa on kahdenlaisia alkioita: a kappaletta tyyppiä A ja b kappaletta tyyppiä B. Valitaan populaatiosta n alkiota palauttaen. Mikä on todennäköisyys, että otokseen tulee x alkiota tyyppiä A ja $n-x$ alkiota tyyppiä B? Voimme käyttää vastavaa uurnamallia kuin hypergeometrisen jakauman yhteydessä. Uurnassa on a valkoista ja b mustaa palloa. Valitaan uurnasta satunnaisesti palauttaen n palloa. Mikä on todennäköisyys, että otokseen tulee x valkoista palloa? Koska otanta tehdään palauttaen, uurnan sisältö ei muutu.

Taulukko 2.4. Otanta palauttaen

	Tyyppi A	Tyyppi B	Yhteensä
Populaatio	a	b	$a+b$
Otos	x	$n-x$	n

Kaikkien mahdollisten n:n kokoisten yhtä todennäköisten järjestettyjen jonojen lukumäärä on $(a+b)^n$. Sellaisia järjestettyjä otoksia, joissa on ensin x kappaletta tyyppiä A olevia alkioita ja sitten $n-x$ tyyppiä B, on $a^x b^{n-x}$. Tyyppiä A olevan x:n alkion paikka n:n pituisessa jonossa voidaan valita $\binom{n}{x}$ tavalla. Otoksia (järjestämättömiä), joissa on x kappaletta tyyppiä A ja $n-x$ kappaletta tyyppiä B olevia alkioita, on $\binom{n}{x}a^x b^{n-x}$ kappaletta. Olkoon satunnaismuuttuja X tyyppiä A olevien alkioiden lukumäärä otoksessa. Silloin X:n todennäköisyysfunktio on

$$f(x) = \binom{n}{x}\frac{a^x b^{n-x}}{(a+b)^n}, \qquad x = 0, 1, 2, \ldots, n.$$

Merkitään tyyppiä A olevien alkioiden suhteellista osuutta $p = \frac{a}{a+b}$ ja $1-p = \frac{b}{a+b}$ on tyyppiä B olevien suhteellinen osuus. Nyt X:n todennäköisyysfunktio voidaan kirjoittaa sen tavallisimmassa esitysmuodossa

$$f(x) = \binom{n}{x}p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{2.8.1}$$

Funktio (2.8.1) on *binomijakauman* todennäköisyysfunktio. Kun X noudattaa binomijakaumaa, merkitsemme $X \sim \mathrm{Bin}(n, p)$.

2.8.1 Binomijakauma hypergeometrisen jakauman likiarvona

Kun populaation koko on paljon suurempi kuin otoskoko, on tuloksen kannalta jokseenkin samantekevää, tehdäänkö otanta palauttaen vai palauttamatta. Kun $a+b$ on paljon suurempi kuin n (merkitään $a+b \gg n$), niin binomijakauma (2.8.1) on hypergeometrisen jakauman (2.6.1) hyvä likiarvo. Otanta palauttamatta voidaan luonnehtia hypergeometrisen jakauman avulla ja otanta palauttaen binomijakauman avulla.

Lause 2.11 *Jos $a+b \gg n$, niin*

$$\frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} \approx \binom{n}{x}p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, \tag{2.8.2}$$

missä $p = a/(a+b)$.

Koska binomitodennäköisyydet on helpompi laskea kuin hypergeometriset todennäköisyydet, voidaan relaatiota (2.8.2) käyttää laskennassa hyväksi, kun $a + b \gg n$. Tosin nykyisillä ohjelmilla on helppo laskea tarkat todennököisyydet suoraan hypergeometrisesta jakaumasta, vaikka $a + b$ on suuri.

2.9 Todennäköisyyden yleiset aksioomat

Numeroituvien otosavaruuksien tapauksessa on kaikkiin tapahtumiin helppo liittää todennäköisyydet alkeistapahtumien todennäköisyyksien avulla. Todennäköisyyden keskeiset ominaisuudet (Lause 2.1) seuraavat sitten helposti Määritelmästä 1.1. Jos tapahtumat A_1, A_2, ..., A_n ovat parittain erilliset, niin Lauseen 2.2 mukaan

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n).$$

Additiivisuus voidaan todistaa (numeroituvan otosavaruuden Ω tapauksessa) myös äärettömän monelle parittain erilliselle tapahtumalle A_1, A_2, A_3, ... (vrt. Pykälä 1.3.4). Silloin

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots .$$

Ylinumeroituvasti äärettömille otosavaruuksilla ei todennäköisyyttä voida määritellä alkeistapausten todennäköisyyksien avulla. Silloin todennäköisyys määritellään aksiomaattisesti sopivasti määritellyille otosavaruuden Ω osajoukoille.

Määritelmä 2.5 Olkoon Ω otosavaruus ja $P(A)$ on Ω:n sopivasti valituilla osajoukoilla $A \subset \Omega$ määritelty reaaliarvoinen funktio. Funktio $P(A)$ on todennäköisyysmitta, jos se toteuttaa seuraavat aksioomat:

1. $0 \leq P(A) \leq 1$.
2. $P(\emptyset) = 0 \quad \text{ja} \quad P(\Omega) = 1$.
3. Jos $A_1, A_2, A_3, \ldots$ on parittain pistevieraitten Ω:n osajoukkojen jono, niin
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Näistä aksioomista voidaan johtaa samat lauseet kuin numeroituvien otosavaruuksien tapauksessa. Yleisessä teoriassa Ω:n *kaikki osajoukot eivät ole tapahtumia.* Määrittelimme edellä todennäköisyyden Ω:n 'sopivasti valituille osajoukoille'. Niiden osajoukkojen, jotka ovat tapahtumia, täytyy muodostaa ns. *σ-algebra.*

Määritelmä 2.6 Kokoelma $\mathcal{F}$ on otosavaruuden Ω osajoukkojen muodostama σ-algebra, jos

1. $\Omega \in \mathcal{F}$.
2. Jos $A \in \mathcal{F}$, niin $A^c \in \mathcal{F}$.
3. Jos $A_1, A_2, \ldots \in \mathcal{F}$, niin $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Kolmikko $(\Omega, \mathcal{F}, P)$ on *todennäköisyysavaruus,* missä Ω on ei-tyhjä otosavaruus, $\mathcal{F}$ on σ-algebra ja $P\colon \mathcal{F} \to [0,1]$ on todennäköisyys(mitta). Tämän todennäköisyyden aksiomatisoinnin esitti venäläinen matemaatikko A. N. Kolmogorov (1903–87) vuonna 1929.

Todennäköisyyslaskenta ja kombinatoriikka: Yhteenveto

Todennäköisyyden ominaisuuksia

- Epänegatiivisuus $P(A) \geq 0, \quad A \subset \Omega.$
- Monotonisuus $P(A) \leq P(B), \quad \text{kun } A \subset B \subset \Omega.$
- Additiivisuus $P(A) = \sum_{i=1}^{n} P(A_i), \quad \text{jos } A_1, A_2, \ldots A_n \text{ on } A\text{:n jako.}$
- Komplementti $P(A^c) = 1 - P(A).$
- Yhteenlaskulause $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

Symmetriaan perustuva todennäköisyys

$$p(\omega_i) = \frac{1}{n}, \qquad \text{kaikilla } \omega_i \in \Omega = \{\omega_1, \omega_2, \ldots, \omega_n\};$$

$$P(A) = \sum_{\omega_i \in A} p(\omega_i) = \frac{|A|}{n} = \frac{\text{'suotuisat'}}{\text{'kaikki'}}.$$

Kombinatoriikkaa

- Järjestettyjen n-otosten lukumäärä, kun perusjoukon koko on N:

  1) valinta paluttaen: N^n,
  2) valinta palauttamatta: $N^{(n)} = N(N-1)(N-2)\cdots(N-n+1),\ 0 \leq n \leq N,$
     $N^{(N)} = N!.$

- Otokset (palauttamatta) eli n-kombinaatiot

$$\binom{N}{n} = \frac{N^{(n)}}{n!} = \frac{N!}{n!\,(N-n)!}.$$

- Multinomikerroin

$$\binom{n}{n_1\ n_2\ \dots\ n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!}.$$

- Binomilause

$$(1+t)^n = \sum_{r=0}^{n} \binom{n}{r} t^r,$$

kaikilla $t \in \mathbb{R}$ ja positiivisilla kokonaisluvuilla n.

Satunnaismuuttuja

- Satunnaismuuttuja X on kuvaus $\quad X \colon \Omega \to \mathbb{R}$.
- X:n arvojoukko S: $\quad X(\omega) \in S \subset \mathbb{R}$.
- Jos S on numeroituva, niin X on diskreetti satunnaismuuttuja.
- Jos X ja Y ovat satunnaismuuttujia, niin

$$aX, \qquad X+Y, \qquad X-Y, \qquad XY \qquad \text{ja} \qquad \frac{X}{Y}, \quad Y \neq 0$$

ovat satunnaismuuttujia, missä a on reaalivakio.

- Diskreetin satunnaismuuttujan X jakauma

$$P(X \in A) = \sum_{x \in A} P(X = x) \qquad \text{kaikilla } A \subset S.$$

- X:n kertymäfunktio F: $\quad F(x) = P(X \leq x)$.
- Jos X:n todennäköisyysfunktio f on diskreetti, niin $f(x) = P(X = x)$.
- Hypergeometrinen jakauman todennäköisyysfunktio

$$f(x) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}, \qquad x \leq a \quad \text{ja} \quad n - x \leq b.$$

- Binomijakauman todennäköisyysfunktio

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \dots, n.$$

Kolmogorovin aksioomat

Olkoon $\mathcal{F}$ jokin Ω:n osajoukkojen muodostama σ-algebra. Kuvaus $P\colon \mathcal{F} \to \mathbb{R}$ määrittelee todennäköisyysmitan, jos

1. $0 \leq P(A) \leq 1$ kaikilla $A \in \mathcal{F}$.
2. $P(\emptyset) = 0$ ja $P(\Omega) = 1$.
3. Jos tapahtumat $A_i \in \mathcal{F}$ $(i = 1, 2, \ldots)$ ovat parittain erilliset, niin

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Harjoituksia

1. Laske seuraavat lausekkeet:

(a) $6^{(3)}$, $0^{(5)}$, $5^{(0)}$, $7!$, $\binom{7}{-3}$, $\binom{-7}{3}$.

(b) $\binom{10}{7\ 3}$, $\binom{14}{2\ 3\ 5\ 4}$, $\binom{-1.5}{4}$.

2. Olkoon n positiivinen kokonaisluku ja $0 \leq p \leq 1$. Osoita, että

(a) $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$;
(Vihje: Merkitse $(1-p) + p = (1-p)(1 + \frac{p}{1-p})$ ja käytä binomilausetta.)

(b) $\sum_{x=0}^{n} x\binom{n}{x} p^x (1-p)^{n-x} = np$.
(Vihje: Käytä hyväksi tulosta $x\binom{n}{x} = n\binom{n-1}{x-1}$ ja binomilausetta.)

3. (a) Valitaan satunnaisesti 2 lukua luvuista 1, 2, ..., 39 palauttamatta. Millä todennäköisyydellä saadaan peräkkäiset luvut?

(b) Valitaan 7 lukua luvuista 1, 2, ..., 39 palauttamatta (Lotto). Millä todennäköisyydellä saadaan peräkkäiset luvut?

(c) Valitaan 2 lukua luvuista 1, 2, ..., n palauttamatta. Millä todennäköisyydellä saadaan peräkkäiset luvut?

4. Kahdestatoista verinäytteestä 4 oli positiivisia ja 8 negatiivisia. Sekaannuksen takia näytteet unohtuivat merkitsemättä, joten ne oli analysoitava uudestaan yksitellen (satunnaisessa järjestyksessä).

(a) Millä todennäköisyydellä tarvitaan vain 4 analyysia (4 ensimmäistä positiivisia)?

(b) Millä todennäköisyydellä tarvitaan täsmälleen 5 analyysia?

(c) Millä todennäköisyydellä positiiviset tulokset saadaan peräkkäin?

5. Erääseen lääketieteelliseen hoitokokeeseen osallistui 15 miestä ja 20 naista. Kymmenen satunnaiseti valittua potilasta sai tutkittavaa uutta hoitoa (hoitoryhmä) ja loput kuuluivat vertailuryhmään. Mikä on todennäköisyys, että hoitoryhmään tulee

(a) ainakin yksi kumpaakin sukupuolta?

(b) ainakin kolme kumpaakin sukupuolta?

6. (a) Valitaan 30 kännykän tuote-erästä 4 satunnaisesti palauttamatta tarkastukseen. Jos tuote-erässä on 3 viallista, niin millä todennäköisyydellä otoksessa on

i. täsmälleen 2 viallista?

ii. ainakin 2 viallista?

(b) Olkoon 30:n kännykän tuote-erässä d viallista. Tuote-erästä tarkastetaan n:n kännykän otos. Erä lähetetään myyntiin, jos otoksessa ei ole yhtään viallista, muutoin erä palautetaan. Halutaan, että 5 viallista sisältävät tuote-erät palautetaan todennäköisyydellä $p \geq 0.95$. Kuinka suuri otoskoko silloin tarvitaan?

7. (a) Sijoitetaan 22 palloa satunnaisesti 120 laatikkoon. Mikä on todennäköisyys, että yhdessäkään laatikossa ei ole enempää kuin yksi pallo?

(b) Eräässä 120 päivän jaksossa kaapattiin 22 liikennekonetta. Mikä on todennäköisyys, että samana päivänä kaapataan ainakin 2 konetta, jos eri kaappaukset ajoittuvat täysin satunnaisesti ja toisistaan riippumatta.

8. Valitaan satunnaisesti palauttaen 3 palloa laatikosta, jossa on 3 punaista, 4 keltaista ja 5 sinistä palloa. Laske todennäköisyys, että

(a) pallot ovat samanvärisiä;

(b) pallot ovat erivärisiä.

Laske vastaavat todennäköisyydet, kun otanta on palauttamatta.

9. Eräässä 10000 vaalikelpoisen asukkaan kaupungissa tehtiin juuri ennen vaalia mielipidekysely valitsemalla 100 henkilön otos vaalikelpoisten populaatiosta. Ehdokkaat olivat A ja B. Vaalin tuloksen perusteella tiedetään, että A:n kannatus oli 45 % ja B:n kannatus 55 %. Mikä on todennäköisyys, että kyselyssä

(a) 51 henkilöä kannattaa A:ta?

(b) yli puolet kannattaa A:ta?

(c) Kuinka suuri otos on tehtävä, jotta otoksessa olisi B:n kannattajia enemmän kuin A:n kannattajia vähintään todennäköisyydellä 0.9?

10. Oletetaan, että neliönmuotoinen maa-alue on jaettu kolmeen pinta-alaltaan yhtä suureen kaistaleeseen A, B ja C. Lisäksi oletetaan, että kaistaleiden yksikköhinnat ovat toisiinsa suhteessa $1 : 2 : 3$. Jokaisen (mitallisen) osa-alueen M suhteellinen hinta verrattuna koko maa-alueen hintaan saadaan kaavalla

$$V(M) = \frac{P(M \cap A) + 2\,P(M \cap B) + 3\,P(M \cap C)}{2},$$

missä $P(M) = \frac{|M|}{|\Omega|}$, $|M|$ on M:n pinta-ala ja $|\Omega|$ on koko maa-alueen pinta-ala. Osoita, että $V(M)$ on todennäköisyys(mitta) (Määritelmä 2.1).

11. Olkoon otosavaruus Ω äärellinen. Osoita, että Määritelmän 2.1 aksioomeista seuraavat Määritelmän 1.1 mukaiset todennäköisyysfunktion ominaisuudet.

12. Monivalintatehtävässä on 6 väittämää, joista jokaiseen on vastattava tosi (T) tai epätosi (E). Vastaus on oikein tai väärin ja oikeasta vastauksesta saa 1 pistettä ja väärästä -1 pistettä. Oletetaan, että Mr RW (Random Walker) vastaa väittämiin täysin satunnaisesti (Heittää esimerkiksi lanttia).

(a) Millä todennäköisyydellä RW saa negatiivisen pistemäärän?

(b) Mikä on RW:n pistemäärän todennäköisyysjakauma?

(c) Jos kolmantena vaihtoehtona on mahdollisuus vastata ”en tiedä” (N), niin millä todennäköisyydellä RW saa negatiivisen pistemäärän?

13. Mitä voit sanoa tapahtumasta A, joka on riippumaton itsensä kanssa? Miten luonnehdit tapahtumia A ja B, jotka ovat toisensa poissulkevat ja riippumattomat?

14. Todista Yhteenlaskulause 2.4 osoittamalla ensin, että

$$(A \cup B) - A = B - (A \cap B).$$

15. Oletetaan, että 550 omenan laatikossa on 2 % pilaantuneita.

(a) Millä todennäköisyydellä 25 omenan satunnaisotoksessa (otanta palauttamatta) on 2 pilaantunutta?

(b) Millä todennäköisyydellä 25 omenan satunnaisotoksessa on korkeintaan 2 pilaantunutta?

(c) Halutaan, että ainakin 2 % pilaantuneita sisältävät laatikot hylätään todennäköisyydellä $p \geq 0.95$. Kuinka suureksi otoskoko on valittava?

16. Ovessa on kaksi (erilaista) lukkoa ja avaimet ovat niiden kuuden joukossa, joita kannat aina mukanasi. Olet kiireessä pudottanut yhden näistä kuudesta jonnekin.

(a) Mikä on todennäköisyys, että vielä saat oven auki avaimillasi?

(b) Millä todennäköisyydellä saat oven auki heti kahdella ensiksi kokeilemallasi avaimella (Oletetaan, että avaimet näyttävät täysin samanlaisilta.)

17. (a) Kuinka monta kokonaislukuarvoista ratkaisua yhtälöllä $x_1 + x_2 = 5$ on, kun ratkaisujen tulee olla epänegatiivisia?

(b) Investoit 20 tuhatta euroa 4:ään mahdolliseen kohteeseen. Jokaisen sijoituksen tulee olla 100 euron monikerta. Montako investointistrategiaa on, jos koko summa on sijoitettava?

(c) Montako strategiaa on silloin, jos koko summaa ei tarvitse investoida?

18. Tietokoneessa on n prosessoria ja r työtä jaetaan prosessoreille satunnaisesti. Eri prosessoreille tulevien töiden lukumäärät ovat r_1, r_2, $\ldots$, r_n, $r_i \geq 0$, $i = 1, 2, \ldots, n$ ja

$$r_1 + r_2 + \cdots + r_n = r. \tag{2.9.1}$$

(a) Mikä on erilaisten varausjakaumien (Yhtälön (2.9.1) ratkaisujen lukumäärä)?

(b) Mikä on todennäköisyys, että tietyllä prosessorilla on k, $0 \leq k \leq r$ työtä?

(c) Oletetaan, että annetut n lukua r_1, r_2, $\ldots$, r_n toteuttavat yhtälön (2.9.1) ja kaikki mahdolliset n^r töiden sijoittelua prosessoreille ovat yhtä mahdollisia. Mikä on todennäköisyys saada varausluvut r_1, r_2, $\ldots$, r_n?

Luku 3

Satunnaismuuttujat, ehdollistaminen ja riippumattomuus

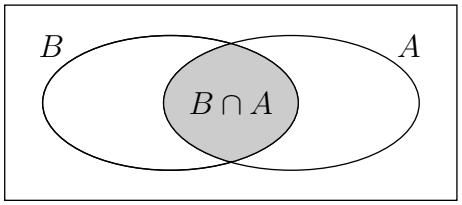
Tässä luvussa käsitellään satunnaismuuttujien ominaisuuksia ja täydennetään todennäköisyyslaskennan tietoja. Erityisesti satunnaismuuttujien odotusarvo on keskeinen käsite. Satunnaismuuttujien tarkastelussa rajoitutaan diskreettiin tapaukseen, mutta vastaavat tulokset pitävät paikkansa myös jatkuville satunnaismuuttujille. Tulosten todistaminen ja soveltaminen on huomattavasti helpompaa diskreettien satunnaismuuttujien yhteydessä.

3.1 Ehdollinen todennäköisyys

Määritelmä 3.1 (Ehdollinen todennäköisyys) Olkoot A ja B otosavaruuden Ω tapahtumia. Jos $P(A) > 0$, niin tapahtuman B ehdollinen todennäköisyys ehdolla A on

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}. \tag{3.1.1}$$

Lauseke $P(B \mid A)$ luetaan "B:n todennäköisyys ehdolla A".



Voidaan ajatella, että $P(A)$ on alueen A pinta-ala ja $P(B \cap A)$ alueen $B \cap A$ pinta-ala. Ehdollinen todennäköisyys $P(B \mid A)$ on siis alueen $B \cap A$ pinta-alan suhteellinen osuus A:n pinta-alasta.

Esimerkki 3.1 Mikä on todennäköisyys, että saat pokerissa kuninkaallisen värisuoran K (samaa maata olevat kortit 10, 11, 12, 13 ja 14 = ässä)? Jos oletetaan, että kaikki 5 kortin kädet ovat yhtä todennäköisiä, niin

$$P(K) = \frac{4}{\binom{52}{5}} = \frac{1}{649740}.$$

Oletetaan, että jakaja jakaa 4 ensimmäistä korttia pöytään kuvapuoli alaspäin ja 5. kortin kuvapuoli ylöspäin. Viimenen korttisi on herttaässä (H_{14}). Millä todennäköisyydellä tämä käsi on kuninkaallinen värisuora? Ehdollisen todennäköisyyden (3.1.1) mukaan

$$P(K \mid H_{14}) = \frac{P(K \cap H_{14})}{P(H_{14})} = \frac{1/\binom{52}{5}}{\binom{51}{4}/\binom{52}{5}} = \frac{1}{\binom{51}{4}}.$$

Voimme nyt helposti todeta, että

$$P(K \mid H_{14}) = \frac{13}{5}\, P(K).$$

Kuninkaallisen värisuoran mahdollisuus siis yli kaksinkertaistuu, kun saat tietää, että viimeinen kortti on herttaässä. □

3.1.1 Todennäköisyyksien tulosääntö

Ehdollisen todennäköisyyden määritelmästä saadaan tulosääntö tapahtuman 'A ja B sattuvat' todennäköisyyden laskemiseksi. Jos tiedetään todennäköisyydet $P(A)$ ja $P(B \mid A)$, saadaan tulokaava

$$P(A \cap B) = P(A)\, P(B \mid A), \tag{3.1.2}$$

ja vastaavasti $P(A^c \cap B) = P(A^c)\, P(B \mid A^c)$. Lauseen 2.3 perusteella

$$P(B) = P(A \cap B) + P(A^c \cap B),$$

joten saamme *kokonaistodennäköisyyden* kaavan

$$P(B) = P(A)\, P(B \mid A) + P(A^c)\, P(B \mid A^c). \tag{3.1.3}$$

Ehdollisen todennäköisyyden määritelmän mukaan

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{kun} \quad P(B) > 0.$$

Kun tämän lausekkeen oikealle puolelle sijoitetaan $P(A \cap B)$:n paikalle (3.1.2) ja $P(B)$:n paikalle vastaavasti (3.1.3), saadaan *Bayesin kaava*

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(A)\, P(B \mid A) + P(A^c)\, P(B \mid A^c)}.$$

Jos siis tunnetaan todennäköisyydet $P(A)$, $P(B \mid A)$ ja $P(B \mid A^c)$, voidaan todennäköisyys $P(A \mid B)$ laskea Bayesin kaavan avulla.

Tulokaava (3.1.2) yleistyy myös useammalle kuin kahdelle tapahtumalle. Esimerkiksi

$$P(A \cap B \cap C) = P(A)\, P(B \mid A)\, P(C \mid A \cap B).$$

Tulokaavan, kokonaistodennäköisyyden ja Bayesin kaavan yleistykset käsitellään luvun loppupuolella.

Esimerkki 3.2 Suuri teollisuuskonserni valmistaa kännyköitä kolmessa eri maassa, jotka ovat nimeltään Fahru, Russo ja Swedla. Ostat kännykän, mutta et tiedä, missä se on valmistettu. Olkoon V tapahtuma, että tuote on viallinen. F on tapahtuma, että tuote on valmistettu Fahrussa. Vastaavasti R ja S viittaavat valmistusmaihin Russo ja Swedla. Lasketaan todennäköisyydet (a) $P(F \mid S^c)$, (b) $P(V \mid S^c)$, (c) $P(V)$, (d) $P(F \mid V)$. Oletetaan, että kaikki valmistetut 6000000 kännykkää ovat yhtä todennäköisiä.

Taulukko 3.1. Kokonaistuotanto ja viallisten %-osuus eri maissa.

	Maa		
	Fahru	Russo	Swedla
Kokonaistuotanto	1000000	2000000	3000000
Viallisten %-osuus	20 %	10 %	5 %

Ratkaisu.

(a)
$$\begin{aligned} P(F \mid S^c) &= \frac{P(F \cap S^c)}{P(S^c)} \\ &= \frac{P(F)}{P(S^c)} \qquad (\text{koska } F \subseteq S^c) \\ &= \frac{1000000/6000000}{3000000/6000000} = \frac{1}{3}. \end{aligned}$$

(b)
$$\begin{aligned} P(V \mid S^c) &= \frac{V \cap S^c}{P(S^c)} \\ &= \frac{P[V \cap (F \cup R)]}{P(S^c)} \qquad (\text{koska } S^c = F \cup R) \\ &= \frac{P(V \cap F) + P(V \cap R)}{P(S^c)} \qquad (\text{koska } F \cap R = \emptyset) \\ &= \frac{P(V \mid F)\, P(F) + P(V \mid R)\, P(R)}{P(S^c)} \\ &= \frac{\frac{1}{5} \cdot \frac{1}{6} + \frac{1}{10} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{15}. \end{aligned}$$

Kohdat (c) ja (d) jätetään harjoitustehtäviksi. □

Esimerkki 3.3 (Väärä positiivinen) Oletetaan, että eräs verinäytteiden laboratoriotesti antaa kaksi ja vain kaksi tulosta: positiivisen ja negatiivisen. Tiedetään, että 95 % tautia A sairastavista saa testissä positiivisen tuloksen. Myös 2 % niistä, joilla ei ole tautia A, saa positiivisen tuloksen (väärän positiivisen!). Oletetaan, että 1 % populaatiosta sairastaa tautia A. Jos satunnaisesti valitun henkilön testitulos on positiivinen, mikä on todennäköisyys, että hän sairastaa tautia A?

Olkoon nyt $T = \{\text{sairastaa tautia}\}$ ja $+$ tarkoittaa positiivista testitulosta. Tiedämme, että

$$P(+ \mid T) = 0.95, \quad P(+ \mid T^c) = 0.02, \quad P(T) = 0.01 \quad \text{ja} \quad P(T^c) = 0.99.$$

Soveltamalla Bayesin kaavaa (3.7.4) saadaan

$$\begin{aligned} P(T \mid +) &= \frac{P(T)\,P(+ \mid T)}{P(T)\,P(+ \mid T) + P(T^c)\,P(+ \mid T^c)} \\ &= \frac{0.01 \cdot 0.95}{0.01 \cdot 0.95 + 0.99 \cdot 0.02} = \frac{95}{293} \approx 0.32. \end{aligned}$$

Todennäköisyys vaikuttaa ensi näkemältä kovin pieneltä. Alhainen todennäköisyys selittyy sillä, että positiiviset tulevat joukosta, joka on pieni verrattuna siihen joukkoon, josta väärät positiiviset tulevat. □

3.1.2 Riippumattomuus

Milloin käy niin, että ehdollinen todennäköisyys $P(B \mid A)$ on sama kuin ehdollistamaton todennäköisyys $P(B)$? Silloin on voimassa identiteetti

$$P(B) = P(B \mid A) = \frac{P(B \cap A)}{P(A)}.$$

Tämä kysymys johtaa riippumattomuuden määritelmään.

Määritelmä 3.2 Tapahtumat A ja B ovat *riippumattomat*, jos

$$P(A \cap B) = P(A)\,P(B) \tag{3.1.4}$$

Jos tapahtumat A ja B ovat riippumattomat, niin silloin identiteetit

$$P(A \mid B) = P(A) \qquad \text{ja} \qquad P(B \mid A) = P(B)$$

pitävät paikkansa. Tapahtumien A ja B riippumattomuudesta seuraa, että myös niiden komplementit ovat riippumattomat.

Lause 3.1 *Jos tapahtumat A ja B ovat riippumattomat, niin myös*

1. *A ja B^c,*
2. *A^c ja B,*
3. *A^c ja B^c*

ovat riippumattomat.

Todistus. Todistetaan 1. kohta. On siis näytettävä, että A:n ja B:n riippumattomuudesta seuraa identiteetti $P(A \cap B^c) = P(A)\,P(B^c)$. Seurauslauseen 2.1 mukaan

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)\,P(B) && [A \text{ ja } B \text{ riippumattomat}] \\ &= P(A)[1 - P(B)] \\ &= P(A)\,P(B^c) && [\text{Lause 2.1(5)}], \end{aligned}$$

joten A ja B^c ovat riippumattomat. Muut kohdat todistetaan vastaavalla tavalla. □

Esimerkki 3.4 Gynekologisen irtosolunäytteen eli Papa-kokeen avulla voidaan todeta kohdun kaulaosan syöpää edeltävät kudosmuutokset. Oletetaan, että 30–65-vuotiaista naisista $100p$ %:lla on epänormaaleja (muuntuneita) soluja (kohdunsuussa ja kohdunkaulassa). Papa-kokeen suorittamiseen liittyvät seuraavat virheet:

1. Tapahtuma B: Kohdunkaulassa on epänormaaleja soluja, mutta ne *eivät osu* otokseen. Olkoon $P(B) = b$.
2. Tapahtuma C: Otoksessa on poikkeavia soluja, mutta niitä *ei havaita*. Olkoon $P(C) = c$.
3. Tapahtuma D: Pelkästään normaaleja soluja sisältävä otos *luokitellaan väärin* poikkeavaksi. Olkoon $P(D) = d$.

Oletetaan, että kaikki mainitut otanta- ja määritysvirheet ovat toisistaan riippumattomat. Jos satunnaisesti valitulle 30–65-vuotiaalle naiselle tehdään Papa-koe, niin

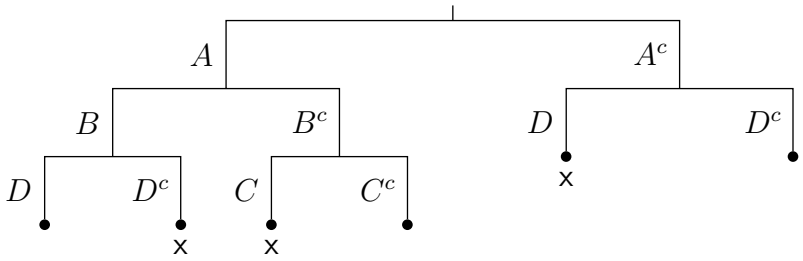
(a) millä todennäköisyydellä koe antaa väärän tuloksen?

(b) Jos testitulos osoittaa poikkeavia soluja löytyneen, millä todennäköisyydellä henkilöllä *ei ole* poikkeavia soluja?

Ratkaisu. (a) Tarkastellaan tapahtumia

V: Testi antaa virheellisen tuloksen,

A: Poikkeavia soluja on kohdunkaulassa



Kuvio 3.1. Kaaviokuva eri tulosvaihtoehdoista. Rastilla (x) merkityissä tilanteissa saadaan virheellinen testitulos.

ja tapahtumaa B (Poikkeavia soluja on, mutta ne eivät osu otokseen). Oletuksen mukaan $P(A) = p$, joten (Seurauslause 2.1)

$$\begin{aligned} P(V) &= P(A)\,P(V \mid A) + P(A^c)\,P(V \mid A^c) \\ &= p\,P(V \mid A) + (1-p)\,P(V \mid A^c). \end{aligned}$$

Virhetodennäköisyyden 3 mukaan $P(V \mid A^c) = d$. Toisaalta

$$P(V \mid A) = P(V \cap B \mid A) + P(V \cap B^c \mid A).$$

Virhetodennäköisyyksien 1 ja 3 mukaan

$$P(V \cap B \mid A) = (1-d)b$$

ja vastaavasti virheiden 1 ja 2 seurauksena

$$P(V \cap B^c \mid A) = c(1-b),$$

joten

$$P(V) = p[(1-d)b + c(1-b)] + (1-p)d.$$

(b) Jätetään harjoitustehtäväksi. □

Useamman kuin kahden tapahtuman riippumattomuuden määrittely vaatii hieman harkintaa. Milloin tapahtumat A, B ja C ovat riippumattomat? Ehdosta $P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$ ei nimittäin seuraa, että tapahtumat ovat parittain riippumattomat.

Määritelmä 3.3 Tapahtumat A, B ja C ovat keskenään riippumattomat, jos

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B), \\ P(A \cap C) &= P(A)\,P(C), \\ P(B \cap C) &= P(B)\,P(C) \\ \text{ja} \qquad P(A \cap B \cap C) &= P(A)\,P(B)\,P(C). \end{aligned}$$

Esimerkki 3.5 Keskinäinen riippumattomuus ei seuraa parittaisesta riippumattomuudesta. Olkoon Ω otosavaruus, jonka alkeistapahtumia ovat tavallisen korttipakan kortit. Valitaan pakasta satunnaisesti yksi kortti. Olkoon $A = \{\spadesuit, \heartsuit\}$ tapahtuma, että saadaan pata tai hertta. Vastaavasti määritellään $B = \{\spadesuit, \clubsuit\}$ ja $C = \{\spadesuit, \diamondsuit\}$. Tapahtumien todennäköisyydet ovat $P(A) = P(B) = P(C) = \frac{26}{52} = \frac{1}{2}$. Mutta $A \cap B = A \cap C = B \cap C = \{\spadesuit\}$, joten

$$P(A \cap B) = P(A \cap C) = P(B \cap C) = P(\{\spadesuit\}) = \frac{13}{52} = \frac{1}{4}.$$

Nyt A, B ja C ovat parittain riippumattomat, sillä $P(A \cap B) = P(A)\,P(B)$, $P(A \cap C) = P(A)\,P(C)$ ja $P(B \cap C) = P(B)\,P(C)$. Koska $A \cap B \cap C = \{\spadesuit\}$ ja

$$P(A \cap B \cap C) = P(\{\spadesuit\}) = \frac{1}{4} \neq P(A)\,P(B)\,P(C) = \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

niin A, B ja C eivät ole keskenään riippumattomat. □

Esimerkki 3.6 Valitaan korttipakasta satunnaisesti yksi kortti. Määritellään tapahtumat $A = \{$ässä tai punainen kuningas tai punainen kuningatar$\}$, $M = \{$musta$\}$ ja $R = \{$risti$\}$. Silloin $P(A) = \frac{8}{52}$, $P(M) = \frac{1}{2}$ ja $P(R) = \frac{1}{4}$. Tapahtuma $A \cap M \cap R = \{$ristiässä$\}$ ja

$$P(A \cap M \cap R) = P(A)\,P(M)\,P(R) = \frac{8}{52} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{52}.$$

Toisaalta

$$\begin{aligned} P(M \cap R) = P(R) = \frac{1}{4} &\neq P(M)\,P(R) = \frac{1}{8}, \\ P(A \cap M) = \frac{2}{52} &\neq P(A)\,P(M) = \frac{8}{52} \cdot \frac{1}{2} = \frac{4}{52}, \\ P(A \cap R) = \frac{1}{52} &\neq P(A)\,P(R) = \frac{8}{52} \cdot \frac{1}{4} = \frac{2}{52}, \end{aligned}$$

joten tapahtumat A, M ja R eivät ole parittain riippumattomia. Identiteetistä $P(A \cap M \cap R) = P(A)\,P(M)\,P(R)$ ei siis seuraa tapahtumien parittainen riippumattomuus. □

Tapahtumien keskinäinen riippumattomuus vaatii toteutuakseen varsin voimakkaita ehtoja.

Määritelmä 3.4 Tapahtumat $A_1, \ldots, A_n$ ovat keskenään riippumattomat, jos jokainen tapahtumien osakokoelma $A_{i_1}, \ldots, A_{i_k}$ $(1 \leq k \leq n)$ toteuttaa ehdon

$$P\left(\bigcup_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Ehdollinen riippumattomuus. Tapahtumat A ja B ovat riippumattomat ehdolla C, jos $P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$.

3.1.3 Joukko-oppi ja todennäköisyys

Todennäköisyyslaskennan kannalta hyödylliset joukko-opin merkinnät esitettiin 1. luvussa. Tapahtumat A ja sen komplementti A^c eivät voi sattua samanaikaisesti, sillä $A \cap A^c = \emptyset$ ja $P(A \cap A^c) = P(\emptyset) = 0$. Toisaalta A, A^c on otosavaruuden Ω ositus, joten $A \cup A^c = \Omega$ ja $P(A \cup A^c) = P(\Omega) = 1$. Tapahtuma "$A$ tai A^c" sattuu varmasti. Lauseen 2.1 (kohta 4) perusteella tiedämme tuloksen $P(A \cup A^c) = P(A) + P(A^c)$, josta seuraa erittäin käyttökelpoinen sääntö (Lause 2.1, kohta 5)

$$P(A) = 1 - P(A^c).$$

De Morganin sääntö

$$(A \cap B)^c = A^c \cup B^c \tag{3.1.5}$$

on tärkeä apuväline todennäköisyyslaskennassa. Se pitää paikkansa myös mielivaltaisen monille tapahtumille. Tapahtuma-avaruuden kielellä luemme identiteetin (3.1.5) seuraavasti

Vasen puoli: Ei ole totta, että sekä A että B sattuvat.

Oikea puoli: Ainakin toinen tapahtumista A, B ei satu.

Soveltamalla kaksinkertaisen komplementin sääntöä $(A^c)^c = A$ saadaan De Morganin säännöstä (3.1.5) toinen vastaava sääntö

$$(A \cup B)^c = A^c \cap B^c.$$

3.2 Ehdolliset jakaumat

Olkoon X jossakin (numeroituvassa) otosavaruudessa Ω määritelty satunnaismuuttuja ja $P(\cdot)$ samassa otosavaruudessa määritelty todennäköisyys. Oletetaan, että tapahtuma $A \subset \Omega$, $P(A) > 0$, on sattunut. Määrittelemme nyt ehdollisen jakauman ehdollisen todennäköisyyden määritelmää mukaillen.

Jokaista X:n arvoa $x \in \mathbb{R}$ kohti voimme määritellä joukon

$$B_x = \{\, \omega \mid X(\omega) = x \,\}.$$

Ehdollisen todennäköisyyden määritelmän mukaan

$$P(X(\omega) = x \mid A) = P(B_x \mid A) = \frac{P(B_x \cap A)}{P(A)} \geq 0. \tag{3.2.1}$$

Koska $\bigcup_x B_x = \Omega$ ja $B_x \cap B_y = \emptyset$ kaikilla $x \neq y$, niin

$$\sum_x P(B_x \mid A) = \sum_x \frac{P(B_x \cap A)}{P(A)} = \frac{P(\Omega \cap A)}{P(A)} = 1. \tag{3.2.2}$$

Määritellään nyt funktio

$$f(x \mid A) = P(B_x \mid A) = P(X = x \mid A), \tag{3.2.3}$$

joka on (3.2.1):n ja (3.2.2):n perusteella todennäköisyysfunktio. Funktio (3.2.3) on X:n *ehdollinen todennäköisyysfunktio ehdolla* A.

Esimerkki 3.7 Oletetaan, että X noudattaa diskreettiä tasajakaumaa Tasd$(1, N)$. Silloin X:n arvojoukko on $S_X = \{1, 2, \ldots, N\}$ ja $P(X = i) = 1/N$ kaikilla $i \in S_X$. Määritellään tapahtuma $A = \{\, \omega \mid a \leq X \leq b \,\}$, missä a, b ja N, $1 \leq a < b \leq N$, ovat kokonaislukuja. Silloin

$$P(A) = \sum_{i=a}^{b} \frac{1}{N} = \frac{b - a + 1}{N}$$

ja

$$P(\{X = k\} \cap A) = \begin{cases} 1/N; & a \leq k \leq b \\ 0; & \text{muutoin.} \end{cases}$$

Siksi X:n ehdollinen todennäköisyysfunktio ehdolla A on

$$f(x \mid A) = \begin{cases} \dfrac{1}{b - a + 1}; & a \leq x \leq b \\ 0; & \text{muutoin.} \end{cases}$$

□

3.3 Satunnaismuuttujien ominaisuuksia

3.3.1 Diskreetin satunnaismuuttujan odotusarvo

Numeroituvassa otosavaruudessa Ω määritellyn satunnaismuuttujan X *odotusarvo* on

$$E(X) = \sum_{\omega \in \Omega} X(\omega)\, P(\{\omega\}), \tag{3.3.1}$$

jos

$$\sum_{\omega \in \Omega} |X(\omega)\, P(\{\omega\})| < \infty. \tag{3.3.2}$$

Jos ehto (3.3.2) toteutuu, sarja (3.3.1) suppenee itseisesti. Tässä tapauksessa sanomme, että satunnaismuuttujalla X on odotusarvo. Muutoin satunnaismuuttujalla ei ole odotusarvoa. Jos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ on äärellinen, niin

$$E(X) = \sum_{i=1}^{n} X(\omega_i)\, P(\{\omega_i\})$$

on aina olemassa.

Tarkastellaan nyt odotusarvon laskemista yleisemmin numeroituvassa otosavaruudessa. Olkoon A_1, A_2, $\ldots$ sellainen otosavaruuden jako

$$\Omega = \bigcup_i A_i,$$

että X saa saman arvon x_i koko joukossa A_i. Voimme kirjoittaa

$$X(\omega) = x_i, \quad \text{kun} \quad \omega \in A_i.$$

Merkitään nyt $P(A_i) = P(X = x_i) = p_i$, joten

$$E(X) = \sum_i P(A_i) x_i = \sum_i p_i x_i. \tag{3.3.3}$$

Tämä kaava saadaan ryhmittelemällä alkeistapaukset kaavassa (3.3.1) osajoukkoihin A_i ja summaamalla sitten yli indeksin i.

Kaavasta (3.3.1) saadaan myös minkä tahansa satunnaismuuttujan X funktion $h(X)$ odotusarvo. Koska $h(X)$ on satunnaismuuttuja, niin

$$E[h(X)] = \sum_{\omega \in \Omega} h[X(\omega)]\, P(\{\omega\}) = \sum_i p_i h(x_i).$$

Näin siis X:n jakauma määrittää $h(X)$:n odotusarvon. Jos erityisesti $h(X) = X^r$, saamme X:n *r. momentin*

$$E(X^r) = \sum_i p_i x_i^r. \tag{3.3.4}$$

Määrittelemme seuraavassa diskreetin satunnaismuuttujan *odotusarvon* todennäköisyysfunktion avulla. Jatkossa kutsumme satunnaismuuttujan odotusarvoa myös satunnaismuuttujan *keskiarvoksi.*

Määritelmä 3.5 (Odotusarvo) Olkoon X diskreetti satunnaismuuttuja, jonka arvojoukko on S ja todennäköisyysfunktio $f_X(x)$. Silloin X:n odotusarvo μ_X on

$$\mu_X = E(X) = \sum_{x \in S} x f_X(x) = \sum_{x \in S} x\, P(X = x), \tag{3.3.5}$$

jos summa suppenee itseisesti.

Odotusarvo μ_X on siis X:n arvojen todennäköisyyksillä painotettu keskiarvo. Jätämme usein merkinnästä satunnaismuuttujaan viittaavan alaindeksin X pois ja merkitsemme lyhyesti $f_X(x) = f(x)$ ja $\mu = E(X)$. Jos summan $\sum_{x \in S} x f_X(x)$ yhteenlaskettavien määrä on äärellinen, niin odotusarvo on aina olemassa. Mikäli yhteenlaskettavien määrä on ääretön, tulee summan supeta itseisesti.

Lause 3.2 *Oletetaan, että otosavaruudessa Ω määritellyllä diskreeteillä satunnaismuuttujilla X ja Y on odotusarvo ja $a \in \mathbb{R}$ on vakio. Silloin*

1. *$E(aX) = a\,E(X)$ ja $E(X+Y) = E(X) + E(Y)$, joten odotusarvo on lineaarinen operaattori.*

Olkoot $h(x)$, $h_1(x)$ ja $h_2(x)$ sellaisia funktioita, että satunnaismuuttujilla $h(X)$, $h_1(X)$ ja $h_2(X)$ on odotusarvo. Silloin seuraavat tulokset pitävät paikkansa:

2. *$E[h(X)] = \sum_x h(x) f_X(x) = \sum_x h(x)\,P(X = x)$*

3. *Jos $h_1(x) \geq h_2(x)$ kaikilla x, niin $E[h_1(X)] \geq E[h_2(X)]$.*

Todistus. 1. Todistetaan ensin $E(aX) = a\,E(X)$. Määritelmän mukaan

$$\begin{aligned} E(aX) &= \sum_x ax\,P(aX = ax) = a \sum_x x\,P(aX = ax) \\ &= a \sum_x x\,P(X = x) = a\,E(X). \end{aligned}$$

Identiteetti $P(aX = ax) = P(X = x)$ pitää paikkansa kaikilla $a \neq 0$, koska $\{\,\omega \mid aX(\omega) = ax\,\} = \{\,\omega \mid X(\omega) = x\,\}$. Jos $a = 0$, niin $aX = 0$ ja $E(aX) = 0 = 0 \cdot E(X)$. Odotusarvo $E(aX)$ on olemassa, koska $E(X)$ on olemassa (oletus). Huomaa, että X:n arvojoukko S_X on numeroituva ja merkintä $\sum_x$ tarkoittaa summaa yli arvojen S_X eli $\sum_x \equiv \sum_{x \in S_X}$.

Todistetaan $E(X+Y) = E(X) + E(Y)$:

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y)\,P(X = x,\, Y = y) \\ &= \sum_x \sum_y [x\,P(X = x,\, Y = y) + y\,P(X = x,\, Y = y)] \\ &= \sum_x \sum_y x\,P(X = x,\, Y = y) + \sum_x \sum_y y\,P(X = x,\, Y = y) \\ &= \sum_x \sum_y x\,P(X = x)\,P(Y = y \mid X = x) \\ &\qquad + \sum_x \sum_y y\,P(Y = y)\,P(X = x \mid Y = y) \end{aligned}$$

$$= \sum_x x\, P(X = x) \Big[\sum_y P(Y = y \mid X = x) \Big] + \sum_y y\, P(Y = y) \Big[\sum_x P(X = x \mid Y = y) \Big]$$
$$= \sum_x x\, P(X = x) + \sum_y y\, P(Y = y) = E(X) + E(Y).$$

Viimeistä edellinen yhtäsuuruus seuraa siitä, että $P(Y = y \mid X = x)$ on Y:n ehdollinen todennäköisyysfunktio ehdolla $X = x$ ja $P(X = x \mid Y = y)$ on X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$. Odotusarvon $E(X+Y)$ olemassaolo seuraa siitä, että $E(X)$ ja $E(Y)$ ovat olemassa ja $|x+y| \leq |x|+|y|$.

2. Seuraa suoraan odotusarvon määritelmästä.

3. Jos $h_1(x) \geq h_2(x)$ kaikilla $x \in \mathbb{R}$, niin

$$E[h_1(X)] - E[h_2(X)] = E[h_1(X) - h_2(X)]$$

1. kohdan mukaan. Nyt

$$E[h_1(X) - h_2(X)] = \sum_x [h_1(x) - h_2(x)]\, P(X = x) \geq 0,$$

koska $h_1(x) - h_2(x) \geq 0$ ja $P(X = x) \geq 0$ kaikilla $x \in \mathbb{R}$. Näin väite on todistettu. □

Olkoon I_A tapahtuman A indikaattorifunktio. Silloin

$$E(I_A) = P(A) \cdot 1 + [1 - P(A)] \cdot 0 = P(A).$$

Huomaa, että $1 - I_A = I_{A^c}$ on A:n komplementin indikaattorifunktio ja $I_\Omega = I_A + I_{A^c} = 1$ kaikilla $\omega \in \Omega$. Määritellään vastaavasti tapahtuman 'kruunu k. heitossa' indikaattorifunktio X_k:

$$X_k(\omega) = \begin{cases} 1, & \text{kun } \omega = \text{kruunu;} \\ 0, & \text{kun } \omega = \text{klaava.} \end{cases}$$

Oletetaan, että kruunun sattumisen todennäköisyys $P(X_k = 1) = p$, $k = 1, 2, \ldots, n$. Nyt satunnaismuuttuja

$$X = X_1 + X_2 + \cdots + X_n$$

on kruunujen lukumäärä, kun heitetään lanttia n kertaa. Silloin odotusarvon lineaarisuuden nojalla

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = p + p + \cdots + p = np.$$

Kruunujen lukumäärän odotusarvo n:ssä heitossa on heittojen lukumäärä kertaa kruunun todennäköisyys. Jos lantti on harhaton, niin $E(X) = \frac{n}{2}$.

Esimerkki 3.8 Olkoon satunnaismuuttujan X arvoalue $S_X = \{-1, 0, 1\}$ ja arvojen todennäköisyydet

$$P(X = -1) = 0.2, \qquad P(X = 0) = 0.5 \qquad \text{ja} \qquad P(X = 1) = 0.3.$$

Lasketaan odotusarvo $E(X^2)$. Merkitään $Y = X^2$. Satunnaismuuttuja Y on siis X:n funktio. Y:n arvoalue on $S_Y = \{0, 1\}$, koska

$$Y(\omega) = \begin{cases} 1, & \text{kun } X(\omega) = 1 \text{ tai } X(\omega) = -1; \\ 0, & \text{kun } X(\omega) = 0. \end{cases}$$

Y:n arvojen 1 ja 0 todennäköisyydet ovat

$$\begin{aligned} P(Y = 1) &= P(X = -1) + P(X = 1) = 0.5, \\ P(Y = 0) &= P(X = 0) = 0.5. \end{aligned}$$

Siksi

$$E(X^2) = E(Y) = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5.$$

Olemme siis ensin määrittäneet X^2:n jakauman ja laskeneet siitä odotusarvon $E(X^2)$.

Voimme kuitenkin laskea $E(X^2)$:n määrittämättä ensin X^2:n jakaumaa. Soveltamalla Lausetta 3.9 (kohta 2) saadaan

$$\begin{aligned} E(X^2) &= (-1)^2 \cdot 0.2 + 0^2 \cdot 0.5 + 1^2 \cdot 0.3 \\ &= 1 \cdot (0.2 + 0.3) + 0 \cdot 0.5 = 0.5. \end{aligned}$$

Määritellään nyt satunnaismuuttuja

$$h(X) = [X - E(X)]^2 = (X - 0.5)^2 = X^2 - X + 0.25.$$

Satunnaismuuttuja $h(X)$ saa arvot $h(-1) = 2.25$, $h(0) = 0.25$ ja $h(1) = 0.25$. Odotusarvo on

$$\begin{aligned} E\big([X - E(X)]^2\big) &= 0.2 \cdot 2.25 + 0.5 \cdot 0.25 + 0.3 \cdot 0.25 \\ &= 0.2 \cdot 2.25 + 0.8 \cdot 0.25 = 0.65. \end{aligned}$$

Odotusarvo $E\big([X - E(X)]^2\big)$ on satunnaismuuttujan X varianssi. □

Esimerkki 3.9 Indikaattorifunktio (Määritelmä 2.3) on käyttökelpoinen myös todennäköisyyksien tarkastelussa. Jos A ja B ovat tapahtumia, niin silloin

$$I_{A^c} = 1 - I_A \qquad \text{ja} \qquad I_{A \cap B} = I_A I_B.$$

Koska $E(I_A) = P(A)$ ja $E(I_{A^c}) = P(A^c)$, niin odotusarvon lineaarisuuden nojalla (Lause 3.9, 1. kohta)

$$E(I_{A^c}) = 1 - E(I_A),$$

josta saamme tutun tuloksen $P(A^c) = 1 - P(A)$. De Morganin sääntöjen avulla saadaan myös identiteetti

$$I_{A \cup B} = I_A + I_B - I_A I_B.$$

□

Esimerkki 3.10 Satunnaismuuttuja X noudattaa diskreettiä tasajakaumaa Tasd$(1, N)$, kun $P(X = i) = \frac{1}{N}$, $i = 1, 2, \ldots, N$ (ks. alaluku 2.5.4). Silloin

$$\begin{aligned} E(X) &= \sum_{x=1}^{N} x \frac{1}{N} = \frac{1}{N} \sum_{x=1}^{N} x \\ &= \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}. \end{aligned}$$

Vastaavasti

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{N} x^2 \frac{1}{N} = \frac{1}{N} \sum_{x=1}^{N} x^2 \\ &= \frac{1}{N} \cdot \frac{N(N+1)(2N+1)}{6} = \frac{(N+1)(2N+1)}{6}. \end{aligned}$$

□

Esimerkki 3.11 Hypergeometrinen jakauma esiteltiin tarkasteltaessa otantaa palauttamatta (alaluku 2.6.1). Esimerkiksi tarkistusotannassa tuotteet luokitellaan viallisiksi tai hyväksyttäviksi. Olkoon tuote-erässä N tuotetta, joista viallisia a ja hyväksyttäviä $N - a$ kappaletta. Tehdään n:n alkion satunnaisotos palauttamatta. Viallisten lukumäärä X otoksessa noudattaa hypergeometrista jakaumaa parametrein n, N ja p, missä $p = \frac{a}{N}$ on viallisten suhteellinen osuus tuote-erässä. Merkitään $X \sim \text{HGeo}(n, N, p)$. Hypergeometrisen jakauman todennäköisyysfunktio on

$$P(X = x; N, n, p) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, \ldots, n, \tag{3.3.6}$$

missä $a = pN$. Huomaa, että $x \leq \min(a, n)$ ja $x \geq \max(0, a + n - N)$, joten X:n todellinen arvoalue saattaa olla suppeampi kuin (3.3.6):ssä annettu.

Tarkistamme ensin, että kyseessä on todennäköisyysjakauma. Selvästikin $P(X = x) \geq 0$, kun $x = 0, 1, \ldots, n$. Mutta identiteetin

$$\sum_{x=0}^{n} P(X = x) = \frac{1}{\binom{N}{n}} \sum_{x=0}^{n} \binom{a}{x}\binom{N-a}{n-x} = 1$$

oikeellisuuden tarkistaminen ei ole täysin vaivaton tehtävä. Voimme kuitenkin tässä nojautua hypergeometriseen identiteettiin (2.4.10), jonka mukaan

$$\sum_{x=0}^{n} \binom{a}{x}\binom{N-a}{n-x} = \binom{N}{n}.$$

Lasketaan nyt hypergeometrisen jakauman odotusarvo

$$E(X) = \sum_{x=0}^{n} x \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} = \sum_{x=1}^{n} x \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}.$$

Identiteetin (2.4.5) nojalla saadaan

$$x\binom{a}{x} = a\binom{a-1}{x-1}$$

ja

$$\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1},$$

joten

$$E(X) = \sum_{x=1}^{n} \frac{a\binom{a-1}{x-1}\binom{N-a}{n-x}}{\frac{N}{n}\binom{N-1}{n-1}} = \frac{na}{N}\sum_{x=1}^{n} \frac{\binom{a-1}{x-1}\binom{N-a}{n-x}}{\binom{N-1}{n-1}}.$$

Kun merkitään $y = n-1$, voidaan kirjoittaa

$$\begin{aligned}\sum_{x=1}^{n} \frac{\binom{a-1}{x-1}\binom{N-a}{n-x}}{\binom{N-1}{n-1}} &= \sum_{y=0}^{n-1} \frac{\binom{a-1}{y}\binom{N-a}{n-1-y}}{\binom{N-1}{n-1}} \\ &= \sum_{y=0}^{n-1} P(Y = y; N-1, n-1, p_1) = 1,\end{aligned}$$

missä $p_1 = \frac{a-1}{N-1}$. Satunnaismuuttuja Y noudattaa siis jakaumaa HGeo$(n-1, N-1, p_1)$. Siksi hypergeometrisen jakauman HGeo(n, N, p) odotusarvo on

$$E(X) = n\frac{a}{N} = np.$$

Summa laskettiin muuntamalla alkuperäinen jakauma hypergeometriseksi jakaumaksi, jonka parametrit ovat $n-1$, $N-1$ ja $p_1 = \frac{a-1}{N-1}$. Vastaavilla laskelmilla voidaan osoittaa, että

$$\mathrm{Var}(X) = \frac{na}{N} \cdot \frac{(N-a)(N-n)}{N(N-1)} = np(1-p)\frac{N-n}{N-1}.$$

□

3.3.2 Ehdollisen jakauman odotusarvo

Koska $f(x \mid A)$ on todennäköisyysfunktio (ks. identiteetti (3.2.3)), niin sen avulla voidaan määritellä odotusarvo. Jos $\sum_x |x| f(x \mid A) < \infty$, niin *$X$:n ehdollinen odotusarvo ehdolla A* on

$$E(X \mid A) = \sum_x x f(x \mid A). \tag{3.3.7}$$

Esimerkki 3.12 Oletetaan, että $X \sim \mathrm{Tasd}(1, N)$ ja $A = \{\, \omega \mid a \le X(\omega) \le b \,\}$, $1 \le a < b \le N$, kuten Esimerkissä 3.7. Nyt X:n ehdollinen odotusarvo ehdolla A on

$$E(X \mid A) = \sum_x x f(x \mid A) = \sum_{x=a}^{b} x \frac{1}{b-a+1} = \frac{a+b}{2}.$$

□

Ehdollisen odotusarvon ja odotusarvon välillä on olemassa seuraavassa lauseessa esitetty erittäin tärkeä yhteys.

Lause 3.3 *Olkoon satunnaismuuttujan X odotusarvo $E(X)$ ja olkoon A sellainen tapahtuma, että $P(A)\,P(A^c) > 0$. Silloin*

$$E(X) = P(A)\,E(X \mid A) + P(A^c)\,E(X \mid A^c).$$

Todistus. Seurauslauseen 2.1 mukaan

$$P(X = x) = P(\{X = x\} \cap A) + P(\{X = x\} \cap A^c)$$

ja ehdollisen todennäköisyyden määritelmän nojalla

$$P(\{X = x\} \cap A) = P(A)\,P(X = x \mid A)$$

ja

$$P(\{X = x\} \cap A^c) = P(A^c)\,P(X = x \mid A^c).$$

Tästä seuraa, että

$$f(x) = P(X = x) = P(A)f(x \mid A) + P(A^c)f(x \mid A^c).$$

Siksi

$$\begin{aligned} E(X) &= \sum_x xf(x) = P(A)\sum_x xf(x \mid A) + P(A^c)\sum_x xf(x \mid A^c) \\ &= P(A)\,E(x \mid A) + P(A^c)\,E(x \mid A^c), \end{aligned}$$

niinkuin väitettiin. □

Jos joukkokokoelma $\{A_i; i \geq 1\}$ muodostaa otosavaruuden Ω osituksen (ks. alaluku 1.3.2), niin voidaan todistaa seuraava yleinen tulos:

$$E(X) = \sum_i P(A_i)\,E(X \mid A_i).$$

Alaluvussa 1.3.2 tarkasteltiin vain äärellisiä osituksia. On syytä huomata, että joukkokokoelma $\{A_i; i \geq 1\}$ voi olla numeroituvasti ääretön. Koska $\{A_i; i \geq 1\}$ on Ω:n ositus, niin

(i) $\bigcup_{i=1}^{\infty} A_i = \Omega$,

(ii) $A_i \cap A_j = \emptyset$, kun $i \neq j$, ja

(iii) $P(A_i) > 0$, $i \geq 1$.

3.3.3 Satunnaismuuttujan varianssi

Varianssin laskemiseksi tarvitaan funktion $h(X) = X^2$ odotusarvo (Vertaa Lauseen 3.9 kohta 2). Odotusarvoa $E(X^2)$ sanotaan satunnaismuuttujan X 2. *momentiksi.* Vastaavasti odotusarvo $E(X)$ on X:n 1. *momentti.* Ennen varianssin määrittelyä esitetään muutamia jatkossa tärkeitä aputuloksia.

Apulause 3.1 *Oletetaan, että satunnaismuuttujilla X ja Y on 2. momentti ja $c \in \mathbb{R}$ on vakio. Silloin odotusarvot*

$$E[(cX)^2], \quad E[(X+Y)^2], \quad E(X), \quad E(Y) \quad ja \quad E(XY) \tag{3.3.8}$$

ovat olemassa.

Todistus.

1. Koska $E[(cX)^2] = c^2\, E(X^2)$ ja $E(X^2)$ on oletuksen mukaan olemassa, niin $E[(cX)^2]$ on olemassa.
2. Koska $0 \le (X+Y)^2 = 2(X^2+Y^2)-(X-Y)^2 \le 2(X^2+Y^2)$ ja oletuksen mukaan $E(X^2+Y^2) = E(X^2) + E(X^2)$ on olemassa, niin Lauseen 3.9 (kohta 3) mukaan $E[(X+Y)^2]$ on olemassa.
3. Koska $0 \le (|X| - |Y|)^2 = |X|^2 + |Y|^2 - 2|X||Y|$, niin niin Lauseen 3.9 (kohta 3) mukaan

   $$E(|XY|) \le \frac{1}{2}\, E(X^2+Y^2),$$

   joten $E(XY)$ on olemassa.

□

Lause 3.4 (Cauchyn ja Schwarzin epäyhtälö) *Jos satunnaismuuttujilla X ja Y on 2. momentti, niin*

$$[E(XY)]^2 \le E(X^2)\, E(Y^2). \tag{3.3.9}$$

Yhtäsuuruus on voimassa jos ja vain jos $P(aX + bY = 0) = 1$, joillain $a, b \in \mathbb{R}$, joista ainakin toinen poikkeaa nollasta.

Todistus. (1) Oletetaan, että $E(X^2) \neq 0$. Koska oletuksen mukaan $E(X^2)$ ja $E(Y^2)$ ovat olemassa, niin Apulauseen 3.1 mukaan myös $E(XY)$ on olemassa. Merkitään nyt $c = E(XY)/E(X^2)$. Silloin

$$0 \le E[(Y-cX)^2] = E(Y^2) - \frac{[E(XY)]^2}{E(X^2)},$$

mistä väite seuraa. Yhtäsuuruus on voimassa silloin ja vain silloin kun

$$P(Y - cX = 0) = 1.$$

(2) Jos $E(X^2) = 0$, niin $P(X = 0) = 1$. Silloin $P(XY = 0) = 0$ ja $E(XY) = 0$, joten epäyhtälö (3.3.9) pitää triviaalisti paikkansa. □

Yhtäsuuruus (3.3.9):ssä vallitsee silloin, kun $aX = -bY$ (todennäköisyydellä 1). Silloin $Y = -\frac{a}{b}X$, jos $b \neq 0$. Epäyhtälössä (3.3.9) pätee siis yhtäsuuruus, kun X ja Y ovat lineaarisesti riippuvia. Epäyhtälö (3.3.9) voidaan lausua myös muodossa

$$|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}.$$

Määritelmä 3.6 (Varianssi) Jos satunnaismuuttujalla X on 2. momentti $E(X^2)$, niin sillä on odotusarvo μ_X ja X:n varianssi on

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{3.3.10}$$

Merkintöjen μ_X ja σ_X^2 sijasta käytämme tavallisesti lyhyempiä versioita μ ja σ^2, jos sekaannuksen vaaraa ei ole. Odotusarvon lineaarisuutta soveltaen voidaan todeta, että

$$\begin{aligned} E[(X-\mu)^2] &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu\, E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2, \end{aligned}$$

joten

$$\sigma^2 = \mathrm{Var}(X) = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2. \tag{3.3.11}$$

satunnaismuuttujan X hajonta $\sigma_X = \sqrt{\mathrm{Var}(X)}$. Odotusarvon määritelmästä ja identiteetistä (3.3.11) saamme erittäin käyttökelpoisen tuloksen:

$$\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X), \qquad E(X^2) = \mu^2 + \mathrm{Var}(X). \tag{3.3.12}$$

Esimerkki 3.13 Lasketaan diskreettiä tasajakaumaa Tasd$(1, N)$ noudattavan satunnaismuuttujan varianssi. Esimerkin 3.10 mukaan

$$E(X) = \frac{N+1}{2} \qquad \text{ja} \qquad E(X^2) = \frac{(N+1)(2N+1)}{6}.$$

Soveltamalla kaavaa (3.3.11) saadaan

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12}. \end{aligned}$$

□

3.3.4 Kovarianssi ja korrelaatio

Oletetaan, että satunnaismuuttujilla X ja Y on 2. momentti. Silloin odotusarvot $E(XY)$ ja $E[(X-\mu_X)(Y-\mu_Y)]$ ovat olemassa Apulauseen 3.1 nojalla.

Määritelmä 3.7 (Kovarianssi) Satunnaismuuttujien X ja Y *kovarianssi* σ_{XY} määritellään odotusarvona

$$\begin{aligned}\sigma_{XY} &= \operatorname{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY) - \mu_X\mu_Y.\end{aligned} \tag{3.3.13}$$

Kovarianssin avulla voidaan sitten määritellä korrelaatiokerroin.

Määritelmä 3.8 (Korrelaatiokerroin) Satunnaismuuttujien X ja Y *korrelaatiokerroin*

$$\rho_{XY} = \operatorname{Cor}(X,Y) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \tag{3.3.14}$$

Sanomme, että X ja Y ovat positiivisesti (negatiivisesti) korrelotuneita, jos $\rho_{XY} > 0$ (< 0). X ja Y eivät korreloi (korreloimattomia), jos $\rho_{XY} = 0$.

Apulause 3.2 (Summan varianssi) *Oletetaan, että satunnaismuuttujilla X ja Y on varianssi. Silloin*

1. $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$.
2. *Jos satunnaismuuttujalla X_1, X_2, …, X_n on varianssi, niin*

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n\sum_{j=1}^n \operatorname{Cov}(X_i,X_j) \\ &= \sum_{i=1}^n \operatorname{Var}(X_i) + \sum_{i=1}^n\sum_{j\neq i}^n \operatorname{Cov}(X_i,X_j).\end{aligned} \tag{3.3.15}$$

Todistus. Todistetaan 1. kohta. Määritelmän mukaan

$$\operatorname{Var}(X+Y) = E[X+Y-(\mu_X+\mu_Y)]^2$$

ja

$$\begin{aligned}[X+Y-(\mu_X+\mu_Y)]^2 &= [(X-\mu_X)+(Y-\mu_Y)]^2 \\ &= (X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y),\end{aligned}$$

missä $\mu_X = E(X)$ ja $\mu_Y = E(Y)$. Odotusarvon lineaarisuuden nojalla

$$\begin{aligned}E[X+Y-(\mu_X+\mu_Y)]^2 &= E(X-\mu_X)^2 + E(Y-\mu_Y)^2 \\ &\quad + 2\,E[(X-\mu_X)(Y-\mu_Y)] \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y).\end{aligned}$$

Kaava (3.3.15) voidaan todistaa induktiolla. □

3.3.5 Satunnaismuuttujan funktion jakauma

Lauseen 3.9 kohdassa 2 esitetään satunnaismuuttujan X funktion odotusarvo X:n jakauman avulla. Jos Y on X:n funktio, voidaan Y:n todennäköisyysjakauma johtaa X:n jakaumasta. Olkoon $Y = h(X)$ satunnaismuuttujan X funktio ja S_Y satunnaismuuttujan Y arvoalue. Jos $A \subset S_Y$, niin

$$P(Y \in A) = P(h(X) \in A).$$

Esimerkki 3.14 Olkoon X diskreetti satunnaismuuttuja, jonka arvoalue on $S = \{-1, 0, 1, 2\}$ ja todennäköisyysfunktio määritellään seuraavasti:

x:	-1	0	1	2
$f_X(x)$:	0.2	0.3	0.4	0.1

Jos $Y = X^2$, niin Y:n todennäköisyysfunktio on

y:	0	1	4
$f_Y(y)$:	0.3	0.6	0.1

Nyt siis esimerkiksi $P(Y = 1) = P(X = -1) + P(X = 1) = 0.2 + 0.4 = 0.6$. Y:n todennäköisyysfunktion määrittäminen X:n todennäköisyysfunktion avulla on suoraviivainen, vaikkakin joskus työläs prosessi.

Tarkastellaan vielä satunnaismuuttujaa $V = g(X) = (X - \mu_X)^2 = (X - 0.4)^2$, missä $\mu_X = 0.4$. V:n todennäköisyysfunktio on

v:	1.96	0.16	0.36	2.56
$f_Y(v)$:	0.2	0.3	0.4	0.1

ja $E(V) = E[(X - 0.4)^2] = \mathrm{Var}(X)$. □

Olkoot S_X ja S_Y satunnaismuuttujien X ja Y otosavaruudet (arvoalueet). Silloin funktio $h(x)$ määrittelee kuvauksen

$$h\colon S_X \to S_Y.$$

Määritellään *joukon A alkukuva* kuvauksessa h seuraavasti:

$$h^{-1}(A) = \{\, x \in S_X \mid h(x) \in A \,\}. \tag{3.3.16}$$

Joukko A voi olla myös yhden pisteen muodostama joukko eli $A = \{y\}$. Silloin

$$h^{-1}(\{y\}) = \{\, x \in S_X \mid h(x) = y \,\}.$$

Tässä tapauksessa merkitsemme $h^{-1}(y)$ merkinnän $h^{-1}(\{y\})$ sijasta. Huomaa, että $h^{-1}(y)$ on edelleen monen pisteen joukko, jos on useita sellaisia X:n arvoja x, että $h(x) = y$. Jos on vain yksi sellainen x, että $h(x) = y$, niin $h^{-1}(y)$ on yhden pisteen muodostama joukko $\{x\}$ ja kirjoitamme silloin $h^{-1}(y) = x$.

3.3.6 Identtisesti jakautuneet satunnaismuuttujat

Määritelmä 3.9 satunnaismuuttujat X ja Y ovat *identtisesti jakautuneet* eli noudattavat samaa jakaumaa, jos jokaiselle tapahtumalle $A \subset \Omega$ pätee $P(X \in A) = P(Y \in A)$.

Kun X ja Y noudattavat samaa jakaumaa, merkitään $X \sim Y$. Jos $X \sim Y$, niin siitä ei seuraa, että X ja Y ovat sama satunnaismuuttuja. Satunnaismuuttujat X ja Y ovat *identtiset* ($X \equiv Y$) eli samat, jos ne on määritelty samassa otosavaruudessa Ω ja $X(\omega) = Y(\omega)$ kaikilla $\omega \in \Omega$.

Esimerkki 3.15 Esimerkissä 2.6 heitettiin harhatonta lanttia 3 kertaa ja määriteltiin satunnaismuuttuja X = 'kruunujen lukumäärä'. Määritellään myös satunnaismuuttuja Y = 'klaavojen lukumäärä'. Merkitään R = 'kruunu' ja L = 'klaava'. Satunnaismuuttujilla X ja Y on sama jakauma, mutta $X \neq Y$, sillä esimerkiksi $X(\text{RRL}) = 2 \neq Y(\text{RRL}) = 1$. Satunnaismuuttujien X ja Y määritelmistä seuraa, että $X + Y \equiv 3$. $X + Y$ on vakio todennäköisyydellä 1: $P(X + Y = 3) = 1$. □

Satunnaismuuttujan jakauma voidaan luonnehtia kertymäfunktion avulla.

Lause 3.5 *Seuraavat kaksi väitettä ovat yhtäpitävät:*

1. *Satunnaismuuttujat X ja Y noudattavat samaa jakaumaa.*
2. *$F_X(x) = F_Y(x)$ kaikilla $x \in \mathbb{R}$, missä F_X on X:n ja F_Y on Y:n kertymäfunktio.*

Kun X ja Y ovat diskreettejä, niin $X \sim Y$, jos $f_X(x) = f_Y(x)$ kaikilla $x \in \mathbb{R}$.

Esimerkki 3.16 Heitetään harhatonta lanttia 4 kertaa. Olkoon kruunun todennäköisyys p. X ja Y on määritelty samoin kuin Esimerkissä 3.15. Mikä on tapahtuman $\{X = Y\}$ todennäköisyys? Tapahtuma $\{X = Y\}$ on

$$\{\, \omega \mid X(\omega) = Y(\omega) \,\} = \{\text{RRLL}, \text{LRRL}, \text{LLRR}, \text{LRLR}, \text{RLLR}, \text{RLRL}\}.$$

Jokaisen yksittäisen alkeistapahtuman (jonon) todennäköisyys on $p^2(1-p)^2$ ja jonoja on $\binom{4}{2} = 6$ kappaletta, joten

$$P(X = Y) = \binom{4}{2} p^2 (1-p)^2.$$

Milloin $X \sim Y$? Koska

$$f_X(x) = \binom{4}{x} p^x (1-p)^{4-x}, \qquad x = 0, 1, 2, 3, 4$$

ja

$$f_Y(y) = \binom{4}{y} p^y (1-p)^{4-y}, \qquad y = 0, 1, 2, 3, 4,$$

niin $f_X(x) = f_Y(x)$ kaikilla $x = 0, 1, 2, 3, 4$ jos ja vain jos $p = \frac{1}{2}$. Siis $X \sim Y$, kun $p = \frac{1}{2}$. □

3.3.7 Satunnaismuuttujien riippumattomuus

Määrittelimme tapahtumien riippumattomuuden alaluvussa 3.1.2. Tarkastelemme nyt satunnaismuuttujien riippumattomuutta.

Määritelmä 3.10 (Satunnaismuuttujien riippumattomuus) Satunnaismuuttujat X ja Y ovat riippumattomat jos

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B) \tag{3.3.17}$$

kaikilla joukoilla $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$.

Merkintä $P(X \in A,\, Y \in B)$ on lyhennys merkinnästä $P(\{X \in A\} \cap \{Y \in B\})$. Satunnaismuuttujat X ja Y ovat siis riippumattomat, jos tapahtumat $\{X \in A\}$ ja $\{X \in B\}$ ovat riippumattomat kaikilla $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$. Riippumattomuuden määritelmästä seuraa esimerkiksi, että kaikilla $x, y \in \mathbb{R}$

$$P(X = x, Y = y) = P(X = x)\, P(Y = y) = f_X(x) f_Y(y), \tag{3.3.18}$$

missä $f_X(x)$ on X:n ja $f_Y(y)$ on Y:n todennäköisyysfunktio.

Lause 3.6 *Jos X ja Y ovat riippumattomat, niin $U = g(X)$ ja $V = h(Y)$ ovat riippumattomat, missä $g(x)$ on pelkästään x:n (ts. X:n arvojen) funktio ja $h(y)$ pelkästään y:n funktio.*

Todistus. Määritellään $A_u = \{\, x \mid g(x) = u \,\}$ ja $A_v = \{\, y \mid h(y) = v \,\}$. Silloin kaikilla u ja v

$$\begin{aligned} P(U = u, V = v) &= P[g(X) = u, h(Y) = v] \\ &= P(X \in A_u, Y \in A_v) \\ &= P(X \in A_u)\, P(Y \in A_v) \qquad (X \text{ ja } Y \text{ riippumattomat}) \\ &= P(U = u)\, P(V = v), \end{aligned}$$

joten U ja V ovat riippumattomat. □

Määritelmä 3.10 pitää täsmälleen paikkansa vain diskreeteille satunnaismuuttujille. Koska yleisessä tapauksessa kaikki Ω:n osajoukot eivät ole tapahtumia, niin silloin on rajoituttava sopivasti määriteltyyn Ω:n osajoukkokokoelmaan. Yhtälö (3.3.17) pitää myös paikkansa, jos toinen oikean puolen tekijöistä on nolla. Huomaa, että $P(X \in A) = 0$ tarkoittaa, että $\{\, \omega \mid X(\omega) \in A \,\} = \emptyset$. Silloin

$$\{X \in A,\, Y \in B\} = \{\, \omega \mid X(\omega) \in A \,\} \cap \{\, \omega \mid Y(\omega) \in B \,\} = \emptyset,$$

joten $P(X \in A,\, Y \in B) = 0$.

Identiteettiä (3.3.18) voidaan myös pitää diskreettien satunnaismuuttujien X ja Y riippumattomuuden määritelmänä, sillä siitä seuraa identiteetti (3.3.17). Jos valitaan kaksi mielivaltaista numeroituvaa joukkoa $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$ sekä oletetaan (3.3.18), saadaan

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x_i \in A} \sum_{y_j \in B} P(X = x_i, Y = y_j) \\ &= \sum_{x_i \in A} \sum_{y_j \in B} P(X = x_i)\, P(Y = y_j) \qquad [(3.3.18)] \\ &= \sum_{x_i \in A} P(X = x_i) \sum_{y_j \in B} P(Y = y_j) \\ &= P(X \in A)\, P(Y \in B). \end{aligned}$$

Näin olemme todenneet, että ehdot (3.3.17) ja (3.3.18) ovat yhtäpitävät.

Tämän luvun alussa määritelty tapahtumien riippumattomuus on itse asiassa satunnaismuuttujien riippumattomuuden erikoistapaus. Olkoon I_A tapahtuman A ja I_B tapahtuman B indikaattorifunktio. Huomaa, että I_A ja I_B ovat satunnaismuuttujia. Koska indikaattorifunktio saa vain arvot 1 tai 0, niin esimerkiksi

$$\{I_A = 1\} = A \qquad \text{ja} \qquad \{I_A = 0\} = A^c.$$

Jos I_A ja I_B ovat riippumattomat, niin

$$P(I_A = x, I_B = y) = P(I_A = x)\, P(I_B = y) \tag{3.3.19}$$

kaikilla $x, y \in \mathbb{R}$. Nyt siis $\{I_A = x\}$ on joko A, A^c tai $\emptyset$ ja $\{I_B = y\}$ on joko B, B^c tai $\emptyset$. Tästä seuraa mm. tapahtumien A ja B riippumattomuuden määritelmä

$$P(A, B) = P(A \cap B) = P(A)\, P(B).$$

Lisäksi saadaan identiteetit

$$\begin{aligned} P(A \cap B^c) &= P(A)\, P(B^c), \\ P(A^c \cap B) &= P(A^c)\, P(B), \\ P(A^c \cap B^c) &= P(A^c)\, P(B^c). \end{aligned}$$

Lauseen 3.1 nojalla jokainen näistä identiteeteistä kelpaa A:n ja B:n riippumattomuuden määritelmäksi.

3.3.8 Useiden satunnaismuuttujien riippumattomuus

Satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomat, jos

$$\begin{aligned} P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) \\ = P(X_1 \in A_1)\, P(X_2 \in A_2) \cdots P(X_n \in A_n) \end{aligned} \tag{3.3.20}$$

kaikilla (sopivasti valituilla) joukoilla $A_i \subset \mathbb{R}$, $1 \leq i \leq n$. Jos $X_1, \ldots, X_n$ ovat diskreettejä, niin (3.3.20) pitää paikkansa kaikille joukoille $A_i \subset \mathbb{R}$, $1 \leq i \leq n$. Yleisessä tapauksessa on A_i:t ($1 \leq i \leq n$) valittava niin, että joukot $\{X_i \in A_i\} = \{\, \omega \mid X_i(\omega) \in A_i \,\}$ ovat tapahtumia. Huomaa, että riippumattomien satunnaismuuttujien $X_1, \ldots, X_n$ jokainen osajono $X_{i_1}, \ldots, X_{i_k}$ on riippumaton [$1 \leq k \leq n$ ja $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$]. Jos esimerkiksi X_1, X_2 ja X_3 ovat riippumattomat, niin myös X_1 ja X_2 ovat riippumattomat. Tämä nähdään, kun valitaan $A_3 = \mathbb{R}$. Silloin $\{X_3 \in \mathbb{R}\} = \Omega$ ja

$$\begin{aligned}\{X_1 \in A_1, X_2 \in A_2, X_3 \in \mathbb{R}\} &= \{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \Omega \\ &= \{X_1 \in A_1, X_2 \in A_2\},\end{aligned}$$

joten identiteetin (3.3.20) mukaan

$$\begin{aligned}P(X_1 \in A_1, X_2 \in A_2) &= P(X_1 \in A_1)\, P(X_2 \in A_2)\, P(\Omega) \\ &= P(X_1 \in A_1)\, P(X_2 \in A_2).\end{aligned}$$

3.4 Suurten lukujen laki

Riippumattomat, samoin jakautuneet satunnaismuuttujat (rsj). Riippumattomien satunnaismuuttujien jono $X_1, X_2, \ldots$ (äärellinen tai ääretön) on samoin jakautunut, jos jokaisella jonon satunnaismuuttujalla on sama jakauma. Sanomme lyhyesti, että jono $X_1, X_2, \ldots$ on *rsj.* Silloin jonon satunnaismuuttujilla on sama kertymäfunktio F, joten

$$P(X_k \leq x) = F(x) \qquad \text{kaikilla } x \in \mathbb{R}.$$

Jos siis yhden satunnaismuuttujan X_k odotusarvo on μ ja varianssi σ^2, silloin niiden kaikkien kaikkien odotusarvo on μ ja varianssi σ^2.

Lause 3.7 (Markovin epäyhtälö) *Olkoon $X \geq 0$ epänegatiivinen satunnaismuuttuja. Silloin*

$$P(X \geq a) \leq \frac{E(X)}{a}, \qquad \textit{kun } a > 0.$$

Todistus. Olkoon I_A joukon $A = \{\, \omega \mid X(\omega) \geq a \,\}$ indikaattorifunktio [ks. (2.3)]. Koska sekä indikaattorifunktio että X ovat epänegatiiviset ja $I_A + I_{A^c} = 1$, niin

$$X = I_A X + I_{A^c} X \geq I_A X \geq a I_A.$$

Viimeinen epäyhtälö seuraa siitä, että $X(\omega) \geq a$ ja $I_A(\omega) = 1$, kun $\omega \in A$. Jos taas $\omega \notin A$, niin $I_A(\omega) = 0$, joten $I_A(\omega) X(\omega) = I_A(\omega) a = 0$. Keskiarvon monotoonisuuden (Lause 3.9, 3. kohta) ja lineaarisuuden (1. kohta) nojalla saadaan

$$E(X) \geq E(aI_A) = a\, E(I_A) = a\, P(X \in A) = a\, P(X \geq a),$$

koska tapahtumat $\{X \in A\}$ ja $\{X \geq a\}$ ovat määritelmän mukaan ekvivalentteja. □

Markovin epäyhtälön avulla on helppo todistaa erittäin käyttökelpoinen *Tšebyševin epäyhtälö.*

Lause 3.8 (Tšebyševin epäyhtälö) *Olkoon X satunnaismuuttuja, jonka keskiarvo on μ ja varianssi σ^2. Silloin*

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \qquad \text{kaikilla } \varepsilon > 0. \tag{3.4.1}$$

Todistus. Määritellään satunnaismuuttuja $Y = h(X) = (X-\mu)^2$ ja valitaan $a = \varepsilon^2 > 0$. Koska $Y \geq 0$ ja $E(Y) = \sigma^2$, seuraa Tšebyševin epäyhtälö (3.4.1) suoraan Markovin epäyhtälöstä. □

Lause 3.9 *Oletetaan, että otosavaruudessa Ω määritellyllä diskreeteillä satunnaismuuttujilla X ja Y on odotusarvo ja $a \in \mathbb{R}$ on vakio. Silloin*

1. *$E(aX) = a\,E(X)$ ja $E(X+Y) = E(X) + E(Y)$, joten odotusarvo on lineaarinen operaattori.*

Olkoot $h(x)$, $h_1(x)$ ja $h_2(x)$ sellaisia funktioita, että satunnaismuuttujilla $h(X)$, $h_1(X)$ ja $h_2(X)$ on odotusarvo. Silloin seuraavat tulokset pitävät paikkansa:

2. *$E[h(X)] = \sum_x h(x) f_X(x) = \sum_x h(x)\,P(X=x)$*

3. *Jos $h_1(x) \geq h_2(x)$ kaikilla x, niin $E[h_1(X)] \geq E[h_2(X)]$.*

Lause 3.10 (Tulon odotusarvo, riippumattomat SM:t) *Olkoot satunnaismuuttujat X ja Y riippumattomat.*

1. *Jos $E(X)$ ja $E(Y)$ ovat olemassa, niin $E(XY) = E(X)\,E(Y)$.*

Olkoot satunnaismuuttujat $X_1, X_2, \ldots, X_n$ riippumattomat.

2. *Jos satunnaismuuttujilla $X_1, X_2, \ldots, X_n$ on odotusarvo, niin*

$$E(X_1 X_2 \cdots X_n) = E(X_1)\,E(X_2) \cdots E(X_n).$$

Todistus. 1. Odotusarvon määritelmän mukaan

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy\,P(X=x,\,Y=y) \\ &= \sum_x \sum_y xy\,P(X=x)\,P(Y=y) \qquad [X \text{ ja } Y \text{ riippumattomat}] \\ &= \Big[\sum_x x\,P(X=x)\Big]\Big[\sum_y y\,P(Y=y)\Big] \\ &= E(X)\,E(Y). \end{aligned}$$

Koska $\sum_x x\,P(X=x)$ ja $\sum_y y\,P(Y=y)$ suppenevat itseisesti odotusarvojen olemassaolon nojalla, pitää 3. yhtäsuuruus paikkansa ja myös odotusarvon $E(XY)$ olemassaolo seuraa odotusarvojen $E(X)$ ja $E(Y)$ olemassaolosta.

Kohta 2. voidaan todistaa soveltamalla toistuvasti 1. kohdan tulosta. □

Apulause 3.3 (Summan varianssi, riippumattomat SM:t) *Oletetaan, että X_1, X_2, …, X_n ovat riippumattomat ja niillä on varianssi. Silloin*

$$\mathrm{Cov}(X_i, X_j) = 0, \qquad i \neq j,$$

ja

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Todistus. Jos $i \neq j$, niin

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)\,E(X_j) \\ &= E(X_i)\,E(X_j) - E(X_i)\,E(X_j) = 0,\end{aligned}$$

koska X_i:n ja X_j:n riippumattomuuden nojalla $E(X_i X_j) = E(X_i)\,E(X_j) = 0$. Summan varianssin $\mathrm{Var}(\sum_{i=1}^n X_i)$ lauseke seuraa nyt suoraan Apulauseesta 3.2. □

Apulause 3.4 (Otoskeskiarvon odotusarvo ja varianssi) *Olkoot X_1, X_2, …, X_n RSJ satunnaismuuttujat, joiden keskiarvo on μ ja varianssi σ^2. Määritellään satunnaismuuttujat*

$$S_n = X_1 + X_2 + \cdots + X_n, \qquad \overline{X}_n = \frac{S_n}{n}.$$

Silloin

$$E(S_n) = n\mu, \qquad \mathrm{Var}(S_n) = n\sigma^2, \qquad E(\overline{X}_n) = \mu, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Voimme nyt todistaa Tšebyševin epäyhtälön avulla ns. *heikon suurten lukujen lain* (HSLL).

Lause 3.11 (Heikko suurten lukujen laki (HSLL)) *Olkoon X_1, X_2, …, X_n ääretön RSJ satunnaismuuttujien jono, jossa jokaisen satunnaismuuttujan keskiarvo on μ ja varianssi σ^2. Olkoon $S_n = X_1 + X_2 + \cdots + X_n$ ja*

$$\overline{X}_n = \frac{S_n}{n}.$$

Silloin jokaisella $\varepsilon > 0$,

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \to 0, \qquad \textit{kun } n \to \infty.$$

Todistus. Apulauseen 3.4 ja Tšebyševin epäyhtälön mukaan

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Kun $n \to \infty$, niin $\sigma^2/(n\varepsilon^2) \to 0$, joten

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \to 0.$$

Näin on lause todistettu. □

Heikko suurten lukujen laki sanoo, että otoskeskiarvo lähenee todennäköisyyden mielessä todellista keskiarvoa, kun otoskoko kasvaa.

3.5 Generoivat funktiot ja momentit

3.5.1 Momentit

Eräs tapa luonnehtia satunnaismuuttujan jakaumaa, on laskea jakauman momentit. Ne määritellään odotusarvon avulla.

Määritelmä 3.11 Olkoon r positiivinen kokonaisluku. Jos odotusarvo

$$\alpha_r = E(X^r)$$

on olemassa, se on satunnaismuuttujan X (tai X:n jakauman) r. momentti. Vastaavasti X:n r. keskusmomentti on

$$\mu_r = E[(X-\mu)^r],$$

missä $\mu = E(X) = \alpha_1$.

Momenttia α_r kutsutaan joskus myös *origomomentiksi.* Jakauman keskiarvo on siis 1. origomomentti ja varianssi 2. keskusmomentti. Satunnaismuuttujan X *tekijämomentit* g_r, $r = 1, 2, \ldots$ määritellään seuraavasti:

$$g_r = E[X^{(r)}] = E[X(X-1)\cdots(X-r+1)].$$

Ensimmäiset kaksi tekijämomenttia ovat

$$g_1 = E(X) = \alpha_1 = \mu,$$
$$g_2 = E[X(X-1)] = E(X^2 - X) = E(X^2) - E(X) = \alpha_2 - \mu.$$

Koska $\sigma^2 = \alpha_2 - \mu^2$, niin

$$\sigma^2 = g_2 + \mu - \mu^2.$$

3.5.2 Momenttifunktio

Esittelemme nyt uuden todennäköisyysjakaumaan liittyvän funktion, *momentteja generoivan funtion,* jota kutsutaan lyhyesti *momenttifunktioksi* (mf). Momenttifunktio tarjoaa erään yleisen menetelmän momenttien laskemiseksi, vaikka se ei aina ole siihen tarkoitukseen helpoin tai tehokkain menetelmä. Momenttien laskemista tärkeämpää on se, että jakaumat voidaan luonnehtia kätevästi momenttifunktion avulla (mikäili se on olemassa).

Määritelmä 3.12 Olkoon X diskreetti satunnaismuuttuja, jonka todennäköisyysfunktio on $f(x)$ ja arvoavaruus S. Silloin reaalimuuttujan t funktio

$$M(t) = E(\mathrm{e}^{tX})$$

on satunnaismuuttujan X (tai X:n jakauman) momenttifunktio (mf), jos odotusarvo

$$E(\mathrm{e}^{tX}) = \sum_{x \in S} \mathrm{e}^{tx} f(x)$$

on olemassa jollain avoimella välillä $-a < t < a$, missä $a > 0$.

Määritelmän perusteella on selvää, että

$$M(0) = E(\mathrm{e}^{0\cdot X}) = \sum_{x\in S} f(x) = 1.$$

Olkoon $S = \{x_1, x_2, \ldots\}$. Silloin

$$M_X(t) = \mathrm{e}^{tx_1} f(x_1) + \mathrm{e}^{tx_2} f(x_2) + \cdots,$$

missä e^{tx_k}:n kertoimet

$$f(x_k) = P(X = x_k), \qquad k = 1, 2, \ldots$$

ovat todennäköisyyksiä. Olkoon $f(x)$ satunnaismuuttujan X todennäköisyysfunktio, $g(y)$ satunnaismuuttujan Y todennäköisyysfunktio ja $S = \{a_1, a_2, \ldots\}$ X:n ja Y:n yhteinen arvoavaruus. Jos

$$M_X(t) = M_Y(t), \qquad \text{kaikilla } t,\ -h < t < h,$$

niin matemaattisen analyysin teorian nojalla

$$f(a_k) = g(a_k), \qquad k = 1, 2, \ldots$$

Jos siis kahdella satunnaismuuttujalla on sama momenttifunktio, niin niillä täytyy olla sama jakauma. Olkoon $F_X(u)$ X:n ja $F_Y(u)$ Y:n kertymäfunktio. Esitetään nyt momenttifunktion yksikäsitteisyyttä koskeva tulos lauseen muodossa.

Lause 3.12 *Olkoot satunnaismuuttujien X ja Y momenttifunktiot $M_X(t)$ ja $M_Y(t)$. Jos $M_X(t) = M_Y(t)$ kaikilla t jossain nollan ympäristössä, niin $F_X(u) = F_Y(u)$ kaikilla u:n arvoilla eli X:llä ja Y:llä on sama jakauma.*

Esimerkki 3.17 Jos $X \sim \mathrm{Ber}(p)$, niin

$$M(t) = E(\mathrm{e}^{tX}) = \mathrm{e}^{t\cdot 1} p + \mathrm{e}^{t\cdot 0} q = \mathrm{e}^t p + q,$$

missä $q = 1 - p$. □

Lause 3.13 *Olkoot X ja Y riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat $M_X(t)$ ja $M_Y(t)$. Silloin satunnaismuuttujan $Z = X + Y$ momenttifunktio on*

$$M_Z(t) = M_X(t) M_Y(t). \tag{3.5.1}$$

Todistus. Koska e^{tX} on pelkästään x:n (X:n arvojen) funktio ja e^{tY} pelkästään y:n funktio, niin Lauseen 3.6 mukaan e^{tX} ja e^{tY} ovat riippumattomat. Väite

$$E(\mathrm{e}^{tZ}) = E[\mathrm{e}^{t(X+Y)}] = E[\mathrm{e}^{tX}\mathrm{e}^{tY}] = E(\mathrm{e}^{tX})\, E(\mathrm{e}^{tY})$$

seuraa sitten suoraan Lauseesta 3.10. □

Usean satunnaismuuttujan tapauksessa on voimassa vastaava tulos.

Seuraus 3.1 *Olkoot X_1, X_2, ..., X_n riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat $M_{X_i}(t)$, $i = 1, 2, \ldots, n$. Silloin summan*

$$S_n = X_1 + X_2 + \cdots + X_n$$

momenttifunktio on

$$M_{S_n}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t).$$

Jos momenttifunktio $M(t)$ on olemassa välillä $(-h, h)$, niin momenttifunktiolla on kaikkien kertalukujen derivaatat pisteessä $t = 0$. Kun identiteetti

$$M(t) = \sum_{x\in S} \mathrm{e}^{tx} f(x) \tag{3.5.2}$$

derivoidaan puolittain, voidaan oikea puoli derivoida termeittäin ja yhtäsuuruus säilyy. Derivoimalla lauseke (3.5.2) puolittain muuttujan t suhteen saadaan

$$M(t)' = \sum_{x\in S} x\mathrm{e}^{tx} f(x),$$
$$M(t)'' = \sum_{x\in S} x^2\mathrm{e}^{tx} f(x)$$

ja jokaisella positiivisella kokonaisluvulla r

$$M(t)^{(r)} = \sum_{x\in S} x^r\mathrm{e}^{tx} f(x).$$

Sijoittamalla $t = 0$ saadaan

$$M(0)' = \sum_{x\in S} xf(x) = E(X),$$
$$M(0)'' = \sum_{x\in S} x^2 f(x) = E(X^2)$$

ja yleisesti

$$M(0)^{(r)} = \sum_{x\in S} x^r f(x) = E(X^r).$$

Erityisesti

$$\mu = M(0)' \qquad \text{ja} \qquad \sigma^2 = M(0)'' - [M(0)']^2.$$

Lause 3.14 *Olkoon $M_X(t)$ satunnaismuuttujan X momenttifunktio ja $Y = aX + b$, missä a ja b ovat annettuja reaaliarvoisia vakioita. Silloin $M_Y(t) = \mathrm{e}^{bt}M_X(at)$.*

Lause 3.15 (Momenttifunktioiden suppeneminen) *Olkoon $X_1, X_2, X_3, \ldots$ satunnaismuuttujien jono, jossa jokaisella X_n:llä on momenttifunktio $M_{X_n}(t)$, $n = 1, 2, 3, \ldots$ Oletetaan lisäksi, että*

$$M_{X_n}(t) \to M_X(t)$$

kaikilla t:n arvoilla jossain nollan ympäristössä $(-h, h)$, kun $n \to \infty$. Jos $M_X(t)$ on momenttifunktio, niin silloin on olemassa yksikäsitteinen kertymäfunktio $F_X(x)$, jonka momenttifunktio on $M_X(t)$ ja

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

kaikissa pisteissä x, joissa $F_X(x)$ on jatkuva.

Satunnaismuuttujien momenttifunktioiden suppenemisesta seuraa siis satunnaismuuttujien kertymäfunktioiden suppeneminen.

3.5.3 Todennäköisyydet generoiva funktio (tgf)

Diskreetin satunnaismuuttujan X todennäköisyydet generoiva funktio (tgf) $G(t)$ määritellään seuraavasti:

$$G(t) = E(t^X) = \sum_{i=1}^{\infty} f(x_i) t^{x_i}.$$

Nähdään helposti, että $G(1) = \sum_{i=1}^{\infty} f(x_i) = 1$. Sarja suppenee ainakin silloin, kun $|t| < 1$. Kun sarja derivoidaan termeittäin, saadaan

$$G'(t) = \sum_{i=1}^{\infty} x_i f(x_i) t^{x_i - 1}.$$

Jos $G(t)$ on olemassa jollain välillä $(-h-1, h+1)$, $h > 0$, niin

$$G'(1) = E(X)$$

ja yleisesti

$$G^{(r)}(1) = E(X^{(r)}) = E[X(X-1)\cdots(X-r+1)]$$

kaikilla positiivisilla kokonaisluvuilla r. Todennäköisyydet generoiva funktio liittyy läheisesti momenttifunktioon, sillä

$$G(\mathrm{e}^t) = E(\mathrm{e}^{tX}) = M(t).$$

3.6 Kokeiden yhdistäminen ja tulomallit

Tarkastellaan nyt satunnaiskokeita $\mathcal{E}_1$ ja $\mathcal{E}_2$, joiden otosavaruudet ovat vastaavasti Ω_1 ja Ω_2. Olkoot satunnaiskokeisiin liittyvät todennäköisyysjakaumat $\{p_i\}$ ja $\{q_i\}$ $i = 1, 2, \ldots$ Tarkastelemme seuraavassa vain numeroituvia otosavaruuksia. Yhdistetään kokeet siten, että tehdään kokeet $\mathcal{E}_1$ ja $\mathcal{E}_2$. Merkitään yhdistettyä koetta $\mathcal{E}_1 \times \mathcal{E}_1$. Yhdistetyn kokeen tulos esitetään järjestettynä parina (ω_i, ω_j), missä $\omega_i \in \Omega_1$ on kokeen $\mathcal{E}_1$ tulos ja $\omega_j \in \Omega_2$ on kokeen $\mathcal{E}_2$ tulos. Yhdistetyn kokeen otosavaruus on siis otosavaruuksien Ω_1 ja Ω_2 *karteesinen tulo* $\Omega_1 \times \Omega_2 = \{\, (\omega_i, \omega_j) \mid \omega_i \in \Omega_1 \text{ ja } \omega_j \in \Omega_2 \,\}$. Vastaavalla tavalla voidaan yhdistää useampiakin kokeita.

Määrittelemme nyt yhdistettyyn kokeeseen $\mathcal{E}_1 \times \mathcal{E}_2$ liittyvän todennäköisyysjakauman $\Omega_1 \times \Omega_2$:ssa. *Kokeet ovat riippumattomat* jos ja vain jos

$$P(\omega_i, \omega_j) = p_i q_j \tag{3.6.1}$$

kaikilla $\omega_i \in \Omega_1$ ja $\omega_j \in \Omega_2$, missä $p_i = p(\omega_i)$ on ω_i:n todennäköisyys Ω_1:ssä ja $q_j = p(\omega_j)$ on ω_j:n todennäköisyys Ω_2:ssä. Selvästikin $P(\omega_i, \omega_j) \geq 0$ kaikilla $(\omega_i, \omega_j) \in \Omega_1 \times \Omega_2$. Koska $\sum\limits_{\omega_i \in \Omega_1} p_i = \sum\limits_{\omega_j \in \Omega_2} q_j = 1$, niin

$$\sum_{(\omega_i, \omega_j) \in \Omega_1 \times \Omega_2} P(\omega_i, \omega_j) = \sum_{\omega_i \in \Omega_1} \sum_{\omega_j \in \Omega_2} p_i q_j = \Big(\sum_{\omega_i \in \Omega_1} p_i \Big) \Big(\sum_{\omega_j \in \Omega_2} q_j \Big) = 1.$$

Identiteetti (3.6.1) siis määrittelee todennäköisyysjakauman $\Omega_1 \times \Omega_2$:ssa. Sitä kutsutaan yhdistetyn kokeen $\mathcal{E}_1 \times \mathcal{E}_1$ *tulomalliksi.*

Riippumattomat toistot

Tulomallin tärkeä erikoistapaus saadaan toistamalla n kertaa koe $\mathcal{E}$, jonka otosavaruus on Ω. Tällaista koetta sanotaan *toistokokeeksi* ja sitä merkitään $\mathcal{E}^n$. Yhdistetyn kokeen otosavaruus on $\Omega \times \Omega \times \cdots \times \Omega$, jonka alkeistapaukset ovat muotoa $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)$, missä ω_i on i. toiston tulos. Olkoon $p(\omega)$ satunnaiskokeeseen $\mathcal{E}$ liittyvässä otosavaruudessa Ω määritelty jakaumafunktio. Toistokokeeseen $\mathcal{E}^n$ liittyvä jakaumafunktio määritellään seuraavasti:

$$p(\boldsymbol{\omega}) = p(\omega_1) p(\omega_2) \cdots p(\omega_n).$$

Bernoullin koe

Bernoullin koe (nimetty James Bernoullin mukaan) on koe, jossa on täsmälleen kaksi tulosvaihtoehtoa. Usein toista tulosvaihtoehtoa kutsutaan onnistumiseksi (O) ja toista epäonnistumiseksi (E), joten Bernoullin kokeen otosavaruus $\Omega = \{\mathrm{O}, \mathrm{E}\}$. Satunnaismuuttuja X noudattaa *Bernoullin jakaumaa,* kun

$$X = \begin{cases} 1, & \text{todennäköisyydellä } P(\mathrm{O}) = p; \\ 0, & \text{todennäköisyydellä } 1 - p, \end{cases} \tag{3.6.2}$$

missä $0 \leq p \leq 1$. Myös satunnaismuuttujan arvoa $X = 1$ kutsutaan onnistumiseksi ja p:tä onnistumistodennäköisyydeksi. Vastaavasti arvoa $X = 0$ kutsutaan epäonnistumiseksi. Huomaa, että X on 'onnistumisen' indikaattorifunktio. Bernoullin kokeen riippumattomat toistot muodostavat *Bernoullin toistokokeen.*

Esimerkki 3.18 Esimerkissä 2.6 heitetään harhatonta lanttia 3 kertaa. Yhdessä lantin heitossa otosavaruus $\Omega = \{\mathrm{R}, \mathrm{L}\}$. Voidaan sopia esimerkiksi, että kruunu (R) on onnistuminen ja klaava (L) on epäonnistuminen. Vastaavan Bernoullin jakaumaa noudattavaan satunnaismuuttujaan liittyvä otosavaruus $S = \{1, 0\}$. Lantin heitto on Bernoullin koe. Tehdään kolme riippumatonta Bernoullin koetta. Tähän yhdistettyyn kokeeseen liittyvä otosavaruus on $S \times S \times S = \{\, (s_1, s_2, s_3) \mid s_i \in S \,\} = \{111, 110, 101, 100, 011, 010, 001, 000\}$. □

Kun toistetaan Bernoullin koe n kertaa (riippumattomat toistot), ovat kokeen mahdolliset tulokset n:n pituisia 1:n ja 0:n muodostamia jonoja. Tyypillinen jono on muotoa $111011000\ldots110$, jonka todennäköisyys on

$$ppp(1-p)p(1-p)(1-p)ppp\cdots pp(1-p) = p^k(1-p)^{n-k},$$

missä k on onnistumisten lukumäärä ja $n-k$ epäonnistumisten lukumäärä. Erilaisten mahdollisten jonojen lukumäärä on 2^n.

Binomijakauma voidaan määritellä Bernoullin toistokokeen avulla. Olkoon $X_1, X_2, \ldots, X_n$ samaa Bernoullin jakaumaa noudattavien riippumattomien satunnaismuuttujien jono, missä $P(X_i = 1) = p$ ja $P(X_i = 0) = 1 - p = q$, $i = 1, 2, \ldots, n$. Silloin $E(X_i) = p$ ja $\mathrm{Var}(X_i) = pq$. Onnistumisten lukumäärä n:ssä riippumattomassa Bernoullin kokeessa on

$$X = X_1 + X_2 + \cdots + X_n.$$

Mikä on todennäköisyys, että onnistumisia on x $(0 \leq x \leq n)$ kappaletta? Jos jonossa on täsmälleen x ykköstä, niin jonon todennäköisyys on $p^x(1-p)^{n-x}$. Tällaisia jonoja on yhteensä $\binom{n}{x}$ kappaletta. Onnistumisten lukumäärä n:ssä Bernoullin kokeessa noudattaa *binomijakaumaa*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

missä siis $f(x) = P(X = x)$.

Onnistumisten lukumäärän otoskeskiarvo on

$$\overline{X}_n = \frac{S_n}{n}.$$

Se on onnistumisten suhteellinen frekvenssi n:ssä riippumattomassa Bernoullin kokeessa, esimerkiksi kruunujen suhteellinen frekvenssi lantin heitossa. Apulauseen 3.4 mukaan $E(\overline{X}_n) = p$ ja $\mathrm{Var}(\overline{X}_n) = pq/n$. HSSL:n mukaan

$$P(|\overline{X}_n - p| > \varepsilon) \to 0$$

kaikilla $\varepsilon > 0$, kun n kasvaa. Kruunujen suhteellinen frekvenssi lähenee p:tä todennäköisyyden mielessä, kun heittojen määrä kasvaa. Bernoulli todisti tämän tuloksen 1713. Tulosta kutsutaan hänen mukaansa *Bernoullin suurten lukujen laiksi.* Ensimmäisessä luvussa tarkasteltiin suhteellisen frekvenssin raja-arvoa todennäköisyyden tulkintana ja eräänlaisena perusteluna todennäköisyydelle. Nyt näemme, että tämä suhteellisen frekvenssin raja-arvotulos on yksi todennäköisyyslaskennan perustuloksista.

3.6.1 Yleinen tulokaava

Yleensä todennäköisyysongelmat koskevat useita tapahtumia tai satunnaismuuttujia, joiden keskinäisiä riippuvuuksia tarkastellaan. Tietyssä mielessä kaikki todennäköisyydet ovat ehdollisia, mutta tavallisesti selvänä pidetyt ehdot jätetään mainitsematta. Rahanheitossa mainitsemme vain vaihtoehdot 'kruunu' ja 'klaava', vaikka lantti voi jäädä myös reunalleen. Presidenttiehdokkaasta tulee presidentti vain sillä ehdolla, että säilyy hengissä vaalikampanjan ajan. Valitsemistodennäköisyyttä laskettaessa ei hengissäpysymisen todennäköisyyttä tavallisesti oteta huomioon.

Seuraavassa esitetään yleinen tulokaava. Huomaa, että jatkossa leikkausta $A_1 \cap A_2$ merkitään kaavojen yksinkertaistamiseksi lyhyesti A_1A_2.

Väittämä 3.1 (Tulokaava) *Olkoot A_1, A_2, …, A_n mitä tahansa tapahtumia. Silloin*

$$P(A_1A_2\cdots A_n) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1A_2)\cdots \cdot P(A_n \mid A_1A_2\cdots A_{n-1}), \tag{3.6.3}$$

jos $P(A_1A_2\cdots A_{n-1}) > 0$.

Todistus. Jos $P(A_1A_2\cdots A_{n-1}) > 0$, niin kaavassa (3.6.3) esitetyt ehdolliset todennäköisyydet ovat hyvin määritellyt, koska

$$P(A_1) \geq P(A_1A_2) \geq \cdots \geq P(A_1A_2\cdots A_{n-1}) > 0.$$

Kun yhtälön (3.6.3) oikea puoli kirjoitetaan auki ehdollisen todennäköisyyden kaavaa (3.1.1) soveltaen, saadaan

$$\frac{P(A_1)}{P(\Omega)} \cdot \frac{P(A_1A_2)}{P(A_1)} \cdot \frac{P(A_1A_2A_3)}{P(A_1A_2)} \cdots \frac{P(A_1A_2\cdots A_n)}{P(A_1A_2\cdots A_{n-1})},$$

joka supistuu todennäköisyydeksi $P(A_1A_2\cdots A_n)$. □

Kutsumme kaavaa (3.6.3) tapahtumien yhdisteen *yleiseksi tulokaavaksi.* Jos A_1, A_2, …, A_n ovat keskenään riippumattomat, niin saadaan

$$P(A_1A_2\cdots A_n) = P(A_1)\,P(A_2)\cdots P(A_n).$$

Oletetaan, että satunnaismuuttujien $X_1, X_2, \ldots, X_n, \ldots$ arvoalueet S_i ovat numeroituvia. Määritellään tapahtumat

$$A_i = \{X_i = x_i\}, \qquad i = 1, 2, \ldots,$$

missä $x_i \in S_i$. Silloin voimme kirjoittaa kertolaskukaavan (3.6.3) avulla

$$\begin{aligned}(3.6.4)\quad & P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P(X_1 = x_1)\, P(X_2 = x_2 \mid X_1 = x_1)\, P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2) \cdots \\ &\qquad \cdot P(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}).\end{aligned}$$

$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ on satunnaismuuttujien $X_1, X_2, \ldots, X_n$ *yhteistodennäköisyys,* joka on lausuttu peräkkäisten ehdollisten todennäköisyyksien avulla.

Esimerkki 3.19 (Syntymäpäiväongelma uudelleen) Olemme jo aikaisemmin implisiittisesti soveltaneet yleistä tulokaavaa (3.6.3). Tarkastellaan uudelleen Esimerkin 2.3 syntymäpäiväongelmaa. Kutsuilla on r henkilöä. Millä todennäköisyydellä ainakin kahdella henkilöllä on sama syntymäpäivä? Käytössämme on osanottajalista, johon syntymäpäivät on merkitty (karkausvuotta ei oteta huomioon). Käydään listaa läpi alusta lähtien järjestyksessä niin pitkälle, kunnes löydetään syntymäpäivä, joka jo oli listalla aikaisemmin. Silloin etsintä lopetetaan siihen ja todetaan, että ainakin kahdella vieraalla on sama syntymäpäivä. Jos lista päästään läpi löytämättä toistoa, kellään ei ole samaa syntymäpäivää.

Olkoon B_j tapahtuma, että tarkistus lopetetaan j. vieraaseen, koska hänen kohdallaan huomataan 1. toistuva syntymäpäivä. Olkoon A_j tapahtuma, että j:llä ensimmäisellä on eri syntymäpäivä. Silloin

$$A_r^c = B_2 \cup B_3 \cup \cdots \cup B_r$$

on tapahtuma, että ainakin kahdella on sama syntymäpäivä. Koska tapahtumat $B_2, B_3, \ldots, B_r$ ovat toisensa poissulkevat, niin

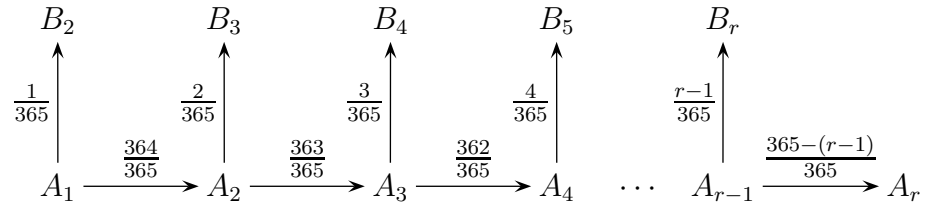
$$P(A_r^c) = P(B_2) + P(B_3) + \cdots + P(B_r),$$

missä

$$P(A_r^c) = 1 - P(A_r).$$

Lasketaan kysytty todennäköisyys $P(A_r^c)$ todennäköisyyden $P(A_r)$ avulla.

Kuvataan tarkistusprosessi toistokokeena:



Jotta tarkistusprosessi menee koko listan läpi, sattuu tapahtuma A_r, eli kaikilla vierailla on eri syntymäpäivä. Sitä ennen ovat sattuneet A_2, A_3, ..., A_{r-1}. Esimerkiksi A_2 on tapahtuma, että tarkistusprosessi ei pysähdy 2. vieraaseen, vaan hänellä on eri syntymäpäivä kuin 1. vieraalla. Todennäköisyys

$$P(A_2) = \frac{364}{365} = 1 - \frac{1}{365} = 1 - P(B_2).$$

koska valittavana on 364 päivää, jotka poikkeavat 1. vieraan syntymäpäiväpäivästä. Jos j:n ensimmäisen syntymäpäivän joukossa ei ole samoja, niin ei myöskään i:n ensimmäisen, jos $i < j$, jolloin $A_i \subset A_j$. Tästä seuraa, että $A_2A_3 \cdots A_j = A_j$ ja

$$P(A_{j+1} \mid A_2A_3 \cdots A_j) = P(A_{j+1} \mid A_j) = \frac{365 - j}{365} = 1 - \frac{j}{365}.$$

Soveltamalla tapahtumien yhdisteen tulokaavaa saadaan

$$\begin{aligned} P(A_r) &= P(A_2A_3A_4 \cdots A_r) \\ &= P(A_2)\, P(A_3 \mid A_2)\, P(A_4 \mid A_2A_3) \cdots P(A_r \mid A_2 \cdots A_{r-1}) \\ &= P(A_2)\, P(A_3 \mid A_2)\, P(A_4 \mid A_3) \cdots P(A_r \mid A_{r-1}) \\ &= \frac{364}{365} \cdot \frac{362}{365} \cdot \frac{362}{365} \cdots \frac{365 - r + 1}{365} = \frac{365^{(r)}}{365^r}. \end{aligned}$$

□

3.7 Bayesin lause

Pastori Thomas Bayesin (1763) mukaan nimetty lause seuraa suoraan ehdollisen todennäköisyyden määritelmästä. Bayesilainen lähestymistapa tilastotieteeseen perustuu tähän lauseeseen. Olkoot H_1, H_2, ..., H_k sellaiset tapahtumat, että

$$H_iH_j = \emptyset \quad (i \neq j) \qquad \text{ja} \qquad \sum_{i=1}^{k} H_i = \Omega.$$

Nyt siis tapahtumajoukko H_1, H_2, ..., H_k muodostaa otosavaruuden Ω osituksen. Tämä tarkoittaa sitä, että yksi ja vain yksi tapahtumista H_1, H_2, ..., H_k sattuu, kun tehdään satunnaiskoe $\mathcal{E}$, jonka otosavaruus on Ω. Oletamme lisäksi, että $P(H_i) > 0$ kaikilla $i = 1, 2, \ldots, k$.

Lause 3.16 *Olkoon*

$$\Omega = \sum_i H_i$$

jokin otosavaruuden ositus. Silloin minkä tahansa tapahtuman $T \subset \Omega$ todennäköisyys voidaan lausua muodossa

$$P(T) = \sum_i P(H_i)\, P(T \mid H_i). \tag{3.7.1}$$

Todistus. Joukko-opin sääntöjen nojalla saadaan

$$T = \Omega T = \Big(\sum_i H_i\Big)T = \sum_i H_i T,$$

josta todennäköisyyden P additiivisuuden (Määritelmä 2.5) perusteella seuraa kaava

$$P(T) = P\Big(\sum_i H_i T\Big) = \sum_i P(H_i T).$$

Kun kaavaan sijoitetaan

$$P(H_i T) = P(H_i)\, P(T \mid H_i),$$

saadaan (3.7.1). □

Jos kaavassa (3.7.1) jokin $P(H_i) = 0$, vastaava summan termi on 0, vaikka $P(T \mid H_i)$ ei olekaan määritelty. Kaavaa (3.7.1) kutsutaan *kokonaistodennäköisyyden kaavaksi.*

Olkoot X ja Y kokonaislukuarvoiset satunnaismuuttujat ja k jokin kokonaisluku. Soveltamalla kaavaa (3.7.1) tapahtumiin

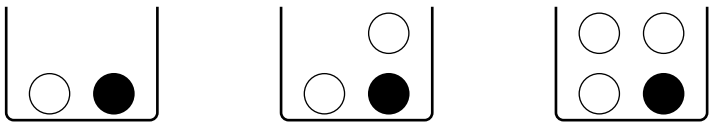
$$H_i = \{X = i\}, \qquad T = \{Y = k\}$$

saadaan

$$P(Y = k) = \sum_i P(X = i)\, P(Y = k \mid X = i), \tag{3.7.2}$$

missä summa käy yli kaikkien kokonaislukujen. Jos $P(X = i) = 0$, niin vastaava yhteenlaskettava summassa on 0. Kaava on helppo yleistää mille tahansa satunnaismuuttujalle X, jonka arvojoukko S_X on numeroituva. Y voi olla jokin yleisempi satunnaismuuttuja, ei välttämättä kokonaislukuarvoinen, ja tapahtuma $T = \{Y = k\}$ voidaan korvata vaikkapa tapahtumalla $T = \{Y > a\}$, $a \in \mathbb{R}$.

Esimerkki 3.20 (Mikä laatikko?) Meillä on 3 samanlaista laatikkoa. Laatikossa i on i valkoista palloa ja yksi musta, $i = 1, 2, 3$. Tilanne on siis oheisen kuvion kaltainen



Laatikko 1 Laatikko 2 Laatikko 3

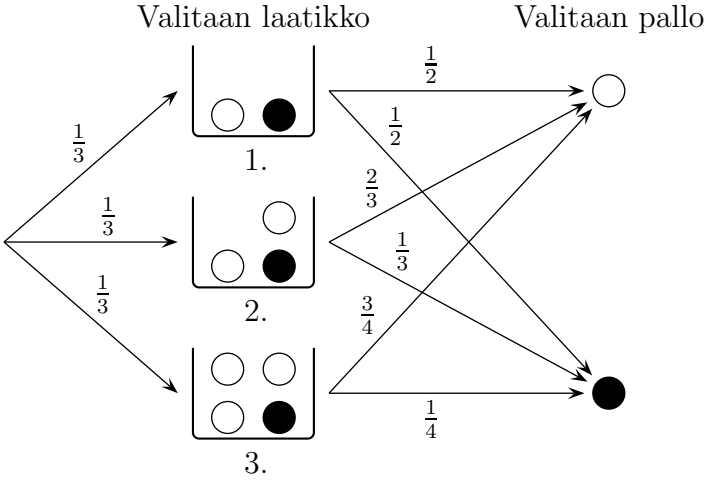
Laatikko valitaan harhattoman nopan heitolla. Jos silmäluku on k, valitaan laatikko, jonka numero $i = \lceil \frac{k}{2} \rceil$ on $\frac{k}{2}$ pyöristettynä lähimpään (suurempaan) kokonaislukuun. Jos esimerkiksi $k = 3$, niin $\lceil \frac{k}{2} \rceil = 2$ ja valitaan Laatikko

2. Kun valitun pallon väri on tiedossa, arvataan, mistä laatikosta pallo on valittu.

Mikä on arvauksesi, jos valittu pallo on valkoinen? Tuntuisi järkevältä arvata Laatikko 3, koska siellä on suhteellisesti eniten valkoisia. Olkoon $H_i = \{\text{Pallo Laatikosta } i\}$, $T = \{\text{Pallo valkoinen}\}$. Arvion varmentamiseksi lasketaan todennäköisyydet

$$P(H_i \mid T) = \frac{P(H_i T)}{P(T)}, \qquad i = 1, 2, 3. \tag{3.7.3}$$

Seuraavassa kuviossa on esitetty havainnollisesti tilanteeseen liittyvät todennäköisyydet.



Kaavassa (3.7.3) osoittaja on

$$P(H_i T) = P(H_i)\, P(T \mid H_i) = \frac{1}{3} \cdot \frac{i}{i+1}, \qquad i = 1, 2, 3.$$

Koska $\sum_{i=1}^{3} H_i T = T$ ja T_1, T_2 ja T_3 muodostavat T:n osituksen, niin yhteenlaskulauseen perusteella

$$P(T) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{3}{4} = \frac{23}{36}.$$

Kaavasta (3.7.3) saadaan

$$P(H_i \mid T) = \frac{\frac{1}{3} \cdot \frac{i}{i+1}}{\frac{23}{36}} = \frac{12}{23} \cdot \frac{i}{i+1}, \qquad i = 1, 2, 3.$$

Jos veikkaat Laatikkoa 3, todennäköisyys osua oikeaan on $\frac{9}{23}$. Laatikolla 1 vastaava todennäköisyys on $\frac{6}{23}$ ja Laatikolla 2 se on $\frac{8}{23}$. Intuitiivisesti oikealta tuntunut Laatikon 3 valinta on siis paras arvaus. □

Väittämä 3.2 (Bayesin lause) *Olkoon H_1, H_2, …, H_k otosavaruuden Ω ositus ja T sellainen tapahtuma, että $P(T) > 0$. Silloin*

$$P(H_i \mid T) = \frac{P(H_i)\, P(T \mid H_i)}{\sum_{j=1}^{k} P(H_j)\, P(T \mid H_j)}. \tag{3.7.4}$$

Todistus. Todennäköisyyksien tulokaavan nojalla sadaan

$$P(H_i T) = P(H_i)\,P(T \mid H_i) = P(T)\,P(H_i \mid T),$$

mistä seuraa

$$P(H_i \mid T) = \frac{P(H_i)\,P(T \mid H_i)}{P(T)}.$$

Väittämän 3.16 mukaan $P(T) = \sum_j P(H_j)\,P(T \mid H_j)$, joten kaava (3.7.4) on todistettu. □

Kaavaa (3.7.4) kutsutaan Bayesin säännöksi. Tapahtumat $H_1, H_2, \ldots, H_k$ voidaan usein ajatella *hypoteeseiksi*, joista täsmälleen yksi on tosi. T on taas jokin tunnettu tieto satunnaiskokeen tuloksesta: tiedämme, että tapahtuma T on sattunut. Todennäköisyydet $P(H_i)$, $i = 1, 2, \ldots, k$ ovat hypoteeseja koskevia ns. *prioritodennäköisyyksiä,* jotka voivat kuvastaa uskoa tai luottamusta kyseisiin hypoteeseihin. Ehdollista todennäköisyyttä $P(H_i \mid T)$ kutsutaan hypoteesin H_i *posterioritodennäköisyydeksi* tai posterioriluottamukseksi hypoteesiin H_i. Tapahtuman T todennäköisyys $P(T \mid H_i)$ ehdolla, että hypoteesi H_i on tosi, on tapahtuman T *uskottavuus* (likelihood) ehdolla H_i.

3.7.1 Peräkkäisotanta

Populaatiossa on N henkilöä, joista Np $(0 \le p \le 1)$ henkilöä kannattaa puoluetta B ja loput $N - Np$ eivät kannata B:tä (ts. kannattavat jotain muuta puoluetta, eivät kannata mitään puoluetta, eivät ota kantaa yms.). Haluamme estimoida kannattajien suhteellisen osuuden p, joka on tuntematon parametri. Haastattelija kysyy n:n satunnaisesti valitun henkilön mielipiteen (otanta palauttamatta). Määritellään

$$X_i = \begin{cases} 1, & \text{jos } i.\text{ haastateltava kannattaa } B\text{:tä;} \\ 0 & \text{muutoin,} \end{cases}$$

missä $1 \le i \le n$ ja $1 \le n \le N$. Tarkastellaan siis satunnaismuuttujien jonoa $\{X_1, X_2, \ldots, X_n\}$ tai lyhyesti $\{\, X_i \mid 1 \le i \le n \,\}$. Tällaista jonoa kutsutaan *stokastiseksi prosessiksi,* mikä on satunnaismuuttujien perheestä käytetty nimitys.

Merkitään nyt $A_i = \{X_i = 1\}$ ja $A_i^c = \{X_i = 0\}$. Silloin kokonaistodennäköisyyden kaavan mukaan

$$P(A_2) = P(A_1)\,P(A_2 \mid A_1) + P(A_1^c)\,P(A_2 \mid A_1^c). \tag{3.7.5}$$

Helposti nähdään, että

$$P(A_1) = \frac{Np}{N} = p \qquad \text{ja} \qquad P(A_1^c) = \frac{N - Np}{N} = 1 - p.$$

Toisaalta

$$P(A_2 \mid A_1) = \frac{Np-1}{N-1} \qquad \text{ja} \qquad P(A_2 \mid A_1^c) = \frac{Np}{N-1},$$

koska 1. haastatellun jälkeen jäljellä on $N-1$ haastateltavaa, joiden joukossa on $Np-1$ B:n kannattajaa, jos 1. haastateltava oli B:n kannattaja. Jos 1. haastateltava ei ollut B:n kannattaja, niin jäljellä on vielä Np B:n kannattajaa. Kun nämä todennäköisyydet sijoitetaan kaavaan (3.7.5), saadaan

$$P(A_2) = p\frac{Np-1}{N-1} + (1-p)\frac{Np}{N-1} = p.$$

Näin olemme osoittaneet, että $P(A_1) = P(A_2)$. Mutta tämä tulos pitää paikkansa yleisesti:

$$P(A_i) = p, \qquad i = 1, 2, \ldots, n; \quad 1 \leq n \leq N. \tag{3.7.6}$$

Näytämme nyt, että tämä yleinen tulos pitää paikkansa. Voimme ajatella, että B:n kannattajat on numeroitu 1, 2, ..., Np ja muut $Np+1$, $Np+2$, ..., N. Kysymyksessä on otanta palauttamatta, kun järjestys otetaan huomioon. Tarkastellaan tapahtumaa A_{i+1}, että $(i+1)$. haastateltava on B:n kannattaja. Kaikkien $(i+1)$:n kokoisten järjestettyjen jonojen (otosten) lukumäärä on $N^{(i+1)}$. Sellaisia jonoja, joissa $(i+1)$. alkio on 1 (B:n kannattaja) on $Np(N-1)^{(i)}$ kappaletta, koska B:n kannattaja voidaan valita Np tavalla ja loput i otosalkiota $(N-1)^{(i)}$ tavalla. Tuloperiaatteen mukaan suotuisia otoksia on siis $Np(N-1)^{(i)}$ kappaletta. Tästä seuraa, että

$$\begin{aligned} P(A_{i+1}) = \frac{Np(N-1)^{(i)}}{N^{(i+1)}} &= \frac{pN(N-1)\cdots(N-1-i+1)}{N^{(i+1)}} \\ &= \frac{pN^{(i+1)}}{N^{(i+1)}} = p. \end{aligned}$$

Olemme näin todistaneet tuloksen (3.7.6)

Määritellään nyt satunnaismuuttuja

$$X = X_1 + X_2 + \cdots + X_n,$$

joka on B:n kannattajien lukumäärä otoksessa. Tiedämme aikaisempien tarkastelujen perusteella, että X noudattaa hypergeometrista jakaumaa HGeo(n, N, p). Johdimme Esimerkissä 3.11 hypergeometrisen jakauman odotusarvon. Nyt tämä odotusarvo on helppo laskea satunnaismuuttujan X avulla, koska

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= p + p + \cdots + p = np, \end{aligned}$$

koska

$$E(X_i) = 1 \cdot p + 0 \cdot (1-p) = p, \qquad i = 1, 2, \ldots, n.$$

Jos satunnaismuuttuja X/n valitaan p:n estimaattoriksi, voimme todeta, että

$$E\left(\frac{X}{n}\right) = \frac{1}{n}\,E(X) = \frac{1}{n} \cdot np = p.$$

Sanomme, että X/n on *harhaton estimaattori*.

3.8 Usean tapahtuman unionin todennäköisyys

Lauseessa 2.5 esitettiin kolmen tapahtuman A_1, A_2 ja A_3 unionin todennäköisyyden $P(A_1 \cup A_2 \cup A_3)$ lauseke. Yleistetään nyt tämä tulos n:n tapahtuman A_1, A_2, ..., A_n unionin tapaukseen.

Lause 3.17 *Samassa otosavaruudessa määriteltyjen tapahtumien A_1, A_2, ..., A_n unionin $\bigcup_{i=1}^n A_i$ todennäköisyys on*

$$P\Big(\bigcup_{i=1}^n A_i\Big) = \sum_{i=1}^n P(A_i) - \sum_{j>i}^n P(A_iA_j) + \sum_{k>j>i}^n P(A_iA_jA_k) - + \cdots + (-1)^{n-1} P(A_1A_2\cdots A_n). \tag{3.8.1}$$

Todistus. Olkoon $\alpha_i = I_{A_i}$ tapahtuman A_i indikaattorifunktio, eli

$$\alpha_i(\omega) = \begin{cases} 1, & \text{kun } \omega \in A_i \\ 0, & \text{kun } \omega \in A_i^c. \end{cases}$$

Silloin tapahtuman $A_1^cA_2^c\cdots A_n^c$ indikaattorifunktio on $\prod_{i=1}^n(1-\alpha_i)$. Koska $\bigcup_{i=1}^n A_i = (A_1^cA_2^c\cdots A_n^c)^c$, niin sen indikaattorifunktio on

$$\begin{aligned} I_{\bigcup A_i} &= 1 - \prod_{i=1}^n (1-\alpha_i) \\ &= \sum_{i=1}^n \alpha_i - \sum_{j>i}^n \alpha_i\alpha_j + \sum_{k>j>i}^n \alpha_i\alpha_k\alpha_k \\ &\qquad - + \cdots + (-1)^{n-1}\alpha_1\alpha_2\cdots\alpha_n. \end{aligned} \tag{3.8.2}$$

Kun nyt yhtälössä (3.8.2) otetaan odotusarvot puolittain ja käytetään hyväksi odotusarvon lineaarisuutta, saadaan tulos (3.8.1). Huomaa, että indikaattorifunktion I_A odotusarvo $E(I_A) = P(A)$ on vastaavan tapahtuman todennäköisyys. Silloin $E(I_{\bigcup A_i}) = P\big(\bigcup_{i=1}^n A_i\big)$, $E(\alpha_i) = P(A_i)$, $E(\alpha_i\alpha_j) = P(A_iA_j)$, ..., $E(\alpha_1\alpha_2\cdots\alpha_n) = P(A_1A_2\cdots A_n)$. □

Esimerkki 3.21 (Yhteensopivuusongelma) Meillä on kaksi n:n kortin korttipakkaa, joiden kortit on numeroitu juoksevasti 1:stä n:ään. Asetetaan 1. pakan kortit pöydälle riviin numerojärjestyksessä 1, 2, ..., n. Sekoitetaan 2. pakka ja asetetaan kortit riviin pöydälle saadussa satunnaisjärjestyksessä. Mikä on todennäköisyys, että i. kortin numero on i? Silloin molemmissa riveissä i. kortti on i eli on saatu i-pari. Mikä on todennäköisyys, että saadaan ainakin yksi pari?

Ratkaisu. Olkoon A_i tapahtuma, että saadaan i-pari. Pakan 2 kortit voidaan asettaa $n!$ erilaiseen järjestykseen. Jos numero i kiinnitetään i. paikalle, niin loput kortit voidaan asettaa $(n-1)!$ erilaiseen järjestykseen, joten

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}. \tag{3.8.3}$$

Jos kiinnitetään i-pari ja j-pari ($i \neq j$), niin loput $(n-2)$ korttia voidaan permutoida $(n-2)!$ tavalla. Silloin

$$P(A_iA_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}. \tag{3.8.4}$$

Vastaavalla tavalla voidaan laskea todennäköisyys, että saadaan i-pari, j-pari ja k-pari ($i \neq j \neq k$):

$$P(A_iA_jA_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)} \tag{3.8.5}$$

ja yleisesti

$$P(A_{i_1}A_{i_2}\dots A_{i_m}) = \frac{(n-m)!}{n!} = \frac{1}{n(n-1)\cdots(n-m+1)}, \qquad 1 \leq m \leq n.$$

Todennäköisyys, että saadaan ainakin yksi pari on siis Lauseen 3.17 mukaan

$$\begin{aligned} P\Big(\bigcup_{i=1}^{n} A_i\Big) &= \binom{n}{1}\frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} \\ &\qquad - + \cdots + (-1)^{n-1}\frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - + \cdots + (-1)^{n-1}\frac{1}{n!}. \end{aligned}$$

Huomaa, että

$$1 - \frac{1}{2!} + \frac{1}{3!} - + \cdots + (-1)^{n-1}\frac{1}{n!} + \cdots = \sum_{i=1}^{\infty}\frac{(-1)^{i-1}}{i!} = 1 - \mathrm{e}^{-1} = 0.632\dots$$

Kun siis n on suuri, niin

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) \approx 1 - \mathrm{e}^{-1} = 0.632\dots$$

Suurilla n:n arvoilla todennäköisyys saada ainakin yksi pari on hyvin lähellä lukua 0.632. □

Satunnaismuuttujat, ehdollistaminen ja riippumattomuus: Yhteenveto

Todennäköisyys

- Ehdollinen todennäköisyys

$$P(B \mid A) = \frac{P(AB)}{P(A)}, \qquad P(A) \neq 0.$$

- Tulosääntö $P(AB) = P(A)\,P(B \mid A)$.

- Yleinen tulokaava

$$P(A_1A_2A_3 \cdots A_{n-1}A_n) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1A_2) \cdots \\ \cdot P(A_n \mid A_1A_2 \cdots A_{n-1}).$$

- Riippumattomuus. A ja B ovat riippumattomat, jos $P(AB) = P(A)\,P(B)$.

- $P(A_1$ tai A_2 tai $A_3)$

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) \\ - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3).$$

Satunnaismuuttujat

- Odotusarvo

$$E(X) = \sum_{\omega \in \Omega} X(\omega)\,P(\{\omega\}),$$
$$E(X) = \sum_{x_i \in S} x_i\,P(X = x_i),$$

  missä S on X:n arvojoukko.

  $E(X)$ on todennäköisyyksillä painotettu X:n arvojen keskiarvo.

- Odotusarvon lineaarisuus

$$E(X + Y) = E(X) + E(Y) \quad \text{ja} \quad E(cX) = c\,E(X), \text{ missä } c \text{ on vakio.}$$

- Varianssi

$$\mathrm{Var}(X) = E(X - \mu)^2 = E(X^2) - \mu^2, \qquad \mu = E(X).$$

- Lineaarinen muunnos $cX + b$

$$E(cX + b) = c\,E(X) + b, \quad \mathrm{Var}(cX + b) = c^2\,\mathrm{Var}(X), \quad b \text{ ja } c \text{ vakioita.}$$

- Cauchyn ja Schwarzin epäyhtälö

$$[E(XY)]^2 \leq E(X^2)\,E(Y^2).$$

- Kovarianssi

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y,$$

missä $\mu_X = E(X)$ ja $\mu_Y = E(Y)$.

- Summat

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y),$$
$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum\sum_{i\neq j} \mathrm{Cov}(X_i,X_j).$$

- Identtiset jakaumat. Diskreeteillä satunnaismuuttujilla X ja Y on sama jakauma, jos niillä on sama arvoalue S ja kaikilla $v \in S$

$$P(X=v) = P(Y=v).$$

- Samat satunnaismuuttujat. X ja Y ovat identtiset, jos $X(\omega) = Y(\omega)$ kaikilla $\omega \in \Omega$. Jos $P(X=Y)=1$, niin $X=Y$ (X ja Y diskreettejä).

- Riippumattomuus. X ja Y ovat riippumattomat, jos

$$P(X\in A, Y\in B) = P(X\in A)\,P(Y\in B)$$

kaikilla $A \subset S_X$ ja $B \subset S_Y$.

Jos X ja Y ovat riippumattomat, niin

1) $g(X)$ ja $h(Y)$ ovat riippumattomat,
2) $E(XY) = E(X)\,E(Y)$,
3) $\mathrm{Cov}(X,Y) = 0$,
4) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

- Markovin epäyhtälö

$$P(X \geq a) \leq \frac{E(X)}{a}, \qquad \text{missä } X \geq 0 \text{ ja } a > 0.$$

- Tšebyševin epäyhtälö

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2},$$

missä $\varepsilon > 0$, $\mu = E(X)$, $\sigma^2 = \mathrm{Var}(X)$.

- Otoskeskiarvo

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

$$E(\overline{X}_n) = \mu \qquad \text{ja} \qquad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n},$$

jos $E(X_i) = \mu$ ja $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$.

- Suurten lukujen laki (heikko): $P(|\overline{X}_n - \mu| \geq \varepsilon) \to 0$, kun $n \to \infty$ ja X_1, X_2, …, X_n ovat riippumattomat ja noudattavat samaa jakaumaa.

Generoivat funktiot ja momentit

- Satunnaismuuttujan momentit

X:n r. momentti	$\alpha_r = E(X^r)$,
r. keskusmomentti	$\mu_r = E(X - \mu)^r$,
r. tekijämomentti	$g_r = E[X^{(r)}] = E[X(X-1)\cdots(X-r+1)]$.

- Momenttifunktio

$$M_X(t) = E(\mathrm{e}^{tX}); \qquad t \in (-a, a),\ a > 0.$$

- Summan $Z = X + Y$ momenttifunktio $M_Z(t) = M_X(t)M_Y(t)$, jos X ja Y ovat riippumattomat.

- r. momentti. Momenttifunktion r. derivaatta pisteessä $t = 0$ on r. momentti: $M(0)^{(r)} = E(X^r)$.

- Todennäköisyydet generoiva funktio

$$G(t) = E(t^X), \qquad X \text{ on diskreetti.}$$

- r. tekijämomentti. G:n r. derivaatta pisteessä $t = 1$ on X:n r. tekijämomentti: $G^{(r)}(1) = E[X^{(r)}]$.

- $G(t)$ vs. $M(t)$: $G(\mathrm{e}^t) = E(\mathrm{e}^{tX}) = M(t)$.

Bayesin lause

- Kokonaistodennäköisyys

$$P(T) = \sum_{i=1}^{k} P(H_i)\, P(T \mid H_i),$$

missä $T \subset \Omega$ ja H_1, H_2, …, H_k on Ω:n ositus.

- Bayesin kaava
$$P(H_i \mid T) = \frac{P(H_i)\,P(T \mid H_i)}{\sum_{j=1}^{k} P(H_j)\,P(T \mid H_j)}.$$
- Prioritodennäköisyydet $P(H_i)$.
- Posterioritodennäköisyydet $P(H_i \mid T)$, $i = 1, 2, \ldots, n$.
- Uskottavuus. $P(T \mid H_i)$ on tapahtuman T uskottavuus ehdolla, että H_i on tosi.

Harjoituksia

1. Oletetaan, että $P(X = 0) = 1 - P(X = 1)$ ja $E(X) = 3\,\mathrm{Var}(X)$. Laske $P(X = 0)$.

2. Olkoon satunnaismuuttujan X todennäköisyysfunktio
$$f(x) = \frac{(|x| + 1)^2}{9}, \qquad x = -1, 0, 1.$$
Laske $E(X)$, $E(X^2)$ ja $E(3X^2 - 2X + 4)$.

3. Olkoon $h(x) = (x - b)^2$, missä b ei ole X:n funktio. Millä b:n arvolla odotusarvo $E[(X - b)^2]$ saavuttaa miniminsä, kun oletetaan, että odotusarvo on olemassa. (Vihje: Tarkastele funktiota $g(b) = E[(X - b)^2] = E(X^2) - 2b\,E(X) + b^2$.)

4. Olkoon $\Omega = \{\omega_1, \omega_2, \omega_3\}$ ja $P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$. Määritellään satunnaismuuttujat X, Y ja Z seuraavasti:
$$\begin{array}{lll} X(\omega_1) = 1, & X(\omega_2) = 2, & X(\omega_3) = 3, \\ Y(\omega_1) = 2, & Y(\omega_2) = 3, & Y(\omega_3) = 1, \\ Z(\omega_1) = 3, & Z(\omega_2) = 1, & Z(\omega_3) = 2. \end{array}$$
(a) Osoita, että satunnaismuuttujilla X, Y ja Z on sama todennäköisyysjakauma.
(b) Määritä satunnaismuuttujien $X + Y$, $Y + Z$, $X + Z$ ja
(c) satunnaismuuttujien $\sqrt{(X^2 + Y^2)Z}$ ja $Z/|X - Y|$ todennäköisyysjakauma.

5. Populaatiossa on M miestä ja N naista. Miehissä on m ja naisissa n tupakoitsijaa. Populaatiosta valitaan satunnaisesti yksi. A on tapahtuma, että valittu on mies ja B tapahtuma, että on valittu tupakoitsija. Mitkä ehdot lukumäärien M, N, m ja n on toteutettava, jotta A ja B ovat toisistaan riippumattomat?

6. Tarkastellaan Esimerkin 2.8 tilannetta, jossa Pekka ja Paavo pelaavat "kruunua ja klaavaa" (satunnaiskävely, $n = 20$).

(a) Mikä on todennäköisyys, että Pekka on 5 heiton jälkeen voitolla yhden euron, 10 heiton jälkeen 2 euroa, 20 heiton jälkeen 2 euroa?

(b) Mikä on Pekan voiton odotusarvo 20 heiton sarjassa?

(c) Jos Pekka on 5. heiton jälkeen voitolla euron, mikä on Pekan voiton odotusarvo 20. heiton jälkeen?

7. Tarkastellaan Tehtävän 6 peliä simuloimalla ($n = 20$).

(a) Mikä on todennäköisin voittosumma? Epätodennäköisin voittosumma? Hahmottele voittosumman todennäköisyysjakauma.

(b) Kuinka usein Pekka on voitolla pelin aikana? Hahmottele tämän satunnaismuuttujan todennäköisyysjakauma.

8. Oletetaan, että $X \sim \text{Tasd}(1, N)$ noudattaa diskreettiä tasajakaumaa.

(a) Jos $E(x) = 6$, niin mitä on $\text{Var}(X)$?

(b) Olkoon $X \sim \text{Tasd}(3, 8)$. Laske $E(X)$ ja $\text{Var}(X)$.

9. Suuressa tehtaassa sattuu 5:n päivän jakson aikana 3 onnettomuutta. Oletetaan, että kaikki mahdolliset 5^3 erilaista 3:n onnettomuuden sijoittumista 5:n päivän jaksolle ovat yhtä todennäköisiä.

Olkoon $Y = \{\text{Onnettomuuspäivien lukumäärä jakson aikana}\}$ ja X niiden päivien lukumäärä, jolloin onnettomuuksia ei satu.

(a) Määritä satunnaismuuttujan $X = 5 - Y$ todennäköisyysfuntio.

(b) Laske $E(X)$ ja $\text{Var}(X)$.

10. Lennolla Havannasta Helsinkiin laukkuni eivät olleet perillä Helsingissä samaan aikaan kuin minä. Laukkuja on reitillä siirretty koneesta toiseen 3 kertaa ja todennäköisyydet, että siirtoa ei ole tehty ajoissa tai oikein, ovat siirtojärjestyksessä 0.4, 0.2 ja 0.1. Mikä on todennäköisyys, että moka sattui jo ensimmäisessä siirrossa?

11. Olkoot X ja Y riippumattomat kokonaislukuarvoiset satunnaismuuttujat, joilla on sama todennäköisyysfunktio $f_X(n) = f_Y(n) = p_n$, $n \geq 1$. Laske todennäköisyydet $P(X = Y)$ ja $P(X \leq Y)$.

12. Tarkastellaan kaksilapsisia perheitä. Oletetaan pojat ja tytöt yhtä todennäköisiksi ja 2. lapsen sukupuoli on riippumaton 1. lapsen sukupuolesta.

Tarkastellaan neljää tapahtumaa:

$$A = 1.\text{ lapsi on poika},$$
$$B = \text{lapset ovat eri sukupuolta},$$
$$C = 1.\text{ lapsi on tyttö},$$
$$D = 2.\text{ lapsi on poika}.$$

(a) Mitkä tapahtumaparit $\{A, B\}$, $\{A, C\}$, $\{B, C\}$ ovat keskenään riippumattomat?

(b) Ovatko tapahtumat A, B ja D keskenään eli täydellisesti riippumattomat?

13. A, B ja C ampuvat maaliin 20 laukausta. Yhden laukauksen osumistodennäköisyys on A:lla 0.4, B:llä 0.3, C:llä 0.1 ja laukaukset ovat toisistaan riippumattomat. Olkoot X_A, X_B ja X_C vastaavasti A:n, B:n ja C:n osumien lukumäärät ja X osumien kokonaismäärä.

(a) Määrittele X riippumattomien satunnaismuuttujien summana ja laske sen avulla X:n odotusarvo ja varianssi.

(b) Määritä Tšebyševin epäyhtälön avulla väli, jolle osumien kokonaismäärä osuu vähintään todennäköisyydellä $\frac{8}{9}$.

14. Liukuhihnalta tulevat pullot ovat vikaantuneita, toisistaan riippumatta, todennäköisyydellä 0.2. Hihnalta tulevat pullot tarkistetaan, vikaantuneet poistetaan ja loput pakataan 12 pullon laatikoihin.

(a) Millä todennäköisyydellä on tutkittava täsmälleen 17 pulloa, kunnes laatikko saadaan täyteen?

(b) Ainakin 17 pulloa, kunnes laatikko saadaan täyteen?

15. Lääkärillä oli oheisessa taulukossa esitetty uuden hoidon vaikutusta koskeva potilasaineisto:

	Asuu kaupungissa		Asuu maaseudulla	
	Saanut hoidon	Ei hoitoa	Saanut hoidon	Ei hoitoa
Elossa	1000	50	95	5000
Kuollut	9000	950	5	5000

Tarkastellaan tapahtumia A = 'potilas elossa', B = 'saanut hoidon' ja C = 'asuu kaupungissa'. Estimoi tarvittavat todennäköisyydet taulukon frekvenssien avulla ja laske

(a) $P(A \mid B)$ ja $P(A \mid B^c)$ sekä

(b) $P(A \mid BC)$, $P(A \mid B^cC)$, $P(A \mid BC^c)$ ja $P(A \mid B^cC^c)$.

(c) Oliko hoidosta apua?

16. Olkoot X ja Y sellaiset satunnaismuuttujat, että $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{Var}(Y) = \sigma_Y^2$ ja $\rho = \mathrm{Cor}(X, Y)$. Käytetään satunnaismuuttujan Y arvioimiseen regressioennustetta $\hat{Y} = \alpha + \beta X$, missä α ja β ovat vakioita. Ennusteen keskineliövirhe määritellään

$$\mathrm{MSE}(\hat{Y}) = E\big([Y - (\alpha + \beta X)]^2\big).$$

(a) Osoita laskemalla, että

$$\mathrm{MSE}(\hat{Y}) = [\mu_Y - (\alpha + \beta\mu_X)]^2 + \mathrm{Var}(Y - \beta X).$$

(b) Valitse edellisessä $\mathrm{MSE}(\hat{Y})$:n lausekkeessa $\alpha = \mu_Y - \beta\mu_X$ ja näytä, että silloin

$$\mathrm{MSE}(\hat{Y}) = (\beta\sigma_X - \rho\sigma_Y)^2 + \sigma_Y^2(1 - \rho^2).$$

(c) Päättele nyt, että $\mathrm{MSE}(\hat{Y})$ saavuttaa miniminsä $\sigma_Y^2(1 - \rho^2)$, kun $\alpha = \mu_Y - \beta\mu_X$ ja $\beta = \rho\sigma_Y/\sigma_X$.

17. Olkoon X sellainen diskreetti satunnaimuuttuja, että sen todennäköisyysfunktio on $P(X = x_i) = p_i$, $i \geq 1$ ja 2. momentti $E(X^2) = \sum_i p_i x_i^2$ on olemassa. Olkoon $A = \{\, i \mid |x_i| \geq \varepsilon \,\}$, missä $\varepsilon > 0$.

(a) Osoita, että

$$P(|X| \geq \varepsilon) = \sum_{i \in A} p_i \text{ ja } E(X^2) \geq \sum_{i \in A} p_i x_i^2,$$

(b) $\sum_{i \in A} p_i x_i^2 \geq \sum_{i \in A} p_i \varepsilon^2$

(c) ja lopuksi $P(|X| \geq \varepsilon) \leq E(X^2)/\varepsilon^2$.

Luku 4

Diskreetit jakaumat

Diskreetti satunnaismuuttuja määriteltiin alaluvussa 2.5. Olemme jo edellisissä luvuissa käsitelleet hypergeometrista jakaumaa (alaluku 2.6.1), binomijakaumaa (alaluvut 2.8 ja 3.6) ja sen erikoistapauksena Bernoullin jakaumaa sekä diskreettiä tasajakaumaa (alaluku 2.5.4), jotka kaikki ovat esimerkkejä *diskreeteistä jakaumista.*

4.1 Diskreetti satunnaismuuttuja

Määritelmä 4.1 Otosavaruudessa Ω määritelty satunnaismuuttuja X on diskreetti, jos sen arvojoukko $S \subset \mathbb{R}$ on numeroituva ja $P(X \in S) = 1$. Joukon S pisteillä on positiivinen todennäköisyys ja ne ovat X:n kertymäfunktion F *hyppypisteitä* ja näiden pisteiden todennäköisyydet ovat F:n hyppyjä.

Määritellään nyt yksinkertainen *hyppyfunktio* $\varepsilon(x)$ seuraavasti:

$$\varepsilon(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Olkoon X:n arvoalue $S = \{1, 2, 3, \ldots\}$ ja $P(x = i) = p_i$, $i \geq 1$. Silloin X:n kertymäfunktio $F(X)$ voidaan kirjoittaa muodossa

$$F(x) = \sum_{i=1}^{\infty} p_i \, \varepsilon(x - i). \tag{4.1.1}$$

Vaikka usein tarkastelemme vain kokonaislukuarvoisia satunnaismuuttujia, se ei ole teoreettiselta kannalta oleellinen rajoitus. Olkoon $S^* = \{x_1, x_2, x_3, \ldots\}$ diskreetin satunnaismuuttujan arvojoukko. Silloin joukkojen S ja S^* välillä on bijektiivinen vastaavuus $g(x_i) = i$ ja $P(X = x_i) = P\big(g(x_i) = i\big)$, joten voimme aina tarvittaessa siirtyä tarkastelemaan vastaavaa kokonaislukuarvoista satunnaismuuttujaa.

Esimerkki 4.1 Yksinkertaisin satunnaismuuttuja X on sellainen, jonka arvoalue $S = \{c\}$ on yksi piste, jolloin $P(X = c) = 1$. Silloin X:n kertymäfunktio on

$$F(x) = \varepsilon(x - c) = \begin{cases} 1, & x \geq c; \\ 0, & x < c. \end{cases}$$

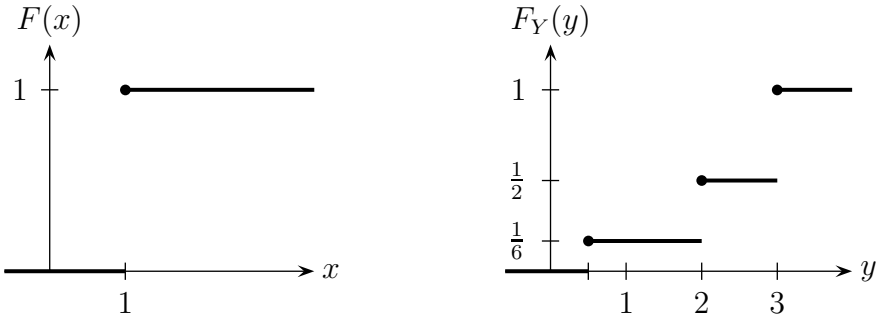
Olkoon Y:n todennäköisyysfunktio

$$P\left(Y = \tfrac{1}{2}\right) = \tfrac{1}{6}, \qquad P(Y = 2) = \tfrac{1}{3} \qquad \text{ja} \qquad P(Y = 3) = \tfrac{1}{2}.$$

Silloin Y:n kertymäfunktio on

$$F_Y(y) = \tfrac{1}{6}\,\varepsilon\left(y - \tfrac{1}{2}\right) + \tfrac{1}{3}\,\varepsilon(y - 2) + \tfrac{1}{2}\,\varepsilon(y - 3).$$

□



Kuvio 4.1. Funktioiden $F(x) = \varepsilon(x - 1)$ ja $F_Y(y)$ kuvaajat.

Esimerkki 4.2 Hatussa on N arpalippua, jotka on numeroitu juoksevasti ykkösestä lähtien. Valitaan hatusta arpa satunnaisesti palauttaen n kertaa ja merkitään valittujen arpojen numerot muistiin. Olkoon X suurin valittujen arpojen numeroista. Silloin $P(X \leq r) = (r/N)^n$ ja

$$\begin{aligned} P(X = r) &= P(X \leq r) - P(X \leq r - 1) \\ &= \left(\frac{r}{N}\right)^n - \left(\frac{r-1}{N}\right)^n. \end{aligned}$$

Määritelmän mukaan X:n odotusarvo on

$$\begin{aligned} E(X) &= N^{-n} \sum_{r=1}^{N} [r^n - (r-1)^n] r \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n r\right] \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n \big((r-1) + 1\big)\right] \end{aligned}$$

$$= N^{-n} \sum_{r=1}^{N} \big[r^{n+1} - (r-1)^{n+1} - (r-1)^n \big]$$
$$= N^{-n} \Big[N^{n+1} - \sum_{r=1}^{N} (r-1)^n \Big].$$

□

4.2 Bernoullin kokeet ja binomijakauma

Alaluvussa 2.8 binomijakauma esiteltiin tarkastelemalla otantaa palauttaen ja alaluvussa 3.6 binomijakauma liitettiin Bernoullin kokeisiin. Bernoullin koe on satunnaiskoe, jolla on täsmälleen kaksi toisensa poissulkevaa tulosvaihtoehtoa (onnistuminen ja epäonnistuminen — lyhyesti O ja E). Esimerkiksi mielipidetiedustelussa henkilö kannattaa tai ei kannata ehdokasta, laatukontrollissa tuote on virheetön tai viallinen, hoidon tuloksena potilas paranee tai ei parane.

Satunnaismuuttuja X noudattaa *Bernoullin jakaumaa,* kun

$$X = \begin{cases} 1 & \text{todennäköisyydellä } p, \\ 0 & \text{todennäköisyydellä } 1-p, \end{cases} \tag{4.2.1}$$

missä $0 \le p \le 1$. Nyt siis X on 'onnistumisen' indikaattorifunktio. Onnistumistodennäköisyys on $P(X=1) = p$ ja vastaavasti epäonnistumisen todennäköisyys on $P(X=0) = 1-p$, jota merkitään usein $q = 1-p$. Bernoullin jakaumaa noudattavan satunnaismuuttujan X odotusarvo ja varianssi ovat

$$E(X) = p \qquad \text{ja} \qquad \operatorname{Var}(X) = pq,$$

sillä

$$E(X) = p \cdot 1 + q \cdot 0 = p, \qquad E(X^2) = p \cdot 1^2 + q \cdot 0^2 = p$$

ja

$$\operatorname{Var}(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p) = pq.$$

Merkitsemme $X \sim \operatorname{Ber}(p)$, kun X noudattaa Bernoullin jakaumaa, jonka odotusarvo on p.

Jos $X \sim \operatorname{Ber}(p)$, niin X:n kertymäfunktio on

$$F(x) = (1-p)\,\varepsilon(x) + p\,\varepsilon(x-1).$$

Yleisesti X:n r. momentti

$$E(X^r) = (1-p) \cdot 0^r + p \cdot 1^r = p$$

on tässä tapauksessa hyvin helppo laskea. Bernoullin jakauman $\operatorname{Ber}(p)$ momenttifunktio on

$$M(t) = E(\mathrm{e}^{tX}) = P(X=0)\mathrm{e}^{t\cdot 0} + P(X=1)\mathrm{e}^{t\cdot 1}$$
$$= (1-p) + p\mathrm{e}^t = 1 + p(\mathrm{e}^t - 1),$$

joka on määritelty kaikilla $t \in \mathbb{R}$.

Esimerkki 4.3 (Sabharwal 1969). Olkoon n:n Bernoullin kokeen jonossa $X_1, X_2, \ldots, X_n$ onnistumistodennäköisyys $P(\mathrm{O}) = p$ ja vastaavasti $P(\mathrm{E}) = 1 - p$ (E = epäonnistuminen). Olkoon Y_n tapahtuman OE (osajono) esiintymisten lukumäärä koejonossa. Mikä on tällaisten osajonojen lukumäärän odotusarvo $E(Y_n)$? Määritellään ensin uusi satunnaismuuttuja

$$Z_i = h(X_i, X_{i+1}) = \begin{cases} 1, & \text{jos } X_i = \mathrm{O} \text{ ja } X_{i+1} = \mathrm{E}; \\ 0 & \text{muulloin,} \end{cases}$$

kun $i = 1, 2, \ldots, n-1$. Silloin

$$Y_n = \sum_{i=1}^{n-1} Z_i$$

ja

$$\begin{aligned} E\, Y_n &= \sum_{i=1}^{n-1} E(Z_i) \\ &= \sum_{i=1}^{n-1} p(1-p) = (n-1)p(1-p). \end{aligned}$$

Jos esimerkiksi $p = \frac{1}{2}$ ja $n = 101$, niin

$$E(Y_n) = \frac{n-1}{4} = 25.$$

□

Tehdään n riippumatonta Bernoullin koetta, joissa jokaisessa onnistumistodennäköisyys on p. Olkoon i. Bernoullin kokeen tulos satunnaismuuttuja X_i, joka saa arvon 1 tai 0. Silloin koesarjan tulos on riippumattomien samaa Bernoullin jakaumaa noudattavien satunnaismuuttujien jono $X_1, X_2, \ldots, X_n$, missä $P(X_i = 1) = p$ ja $P(X_i = 0) = q$, $i = 1, 2, \ldots, n$. Kun koe on tehty, tulos voisi olla esimerkiksi 111011000 . . . 110. Tällaisen tuloksen todennäköisyys (ennen koetta) olisi

$$ppp(1-p)p(1-p)(1-p)ppp \cdots pp(1-p) = p^k(1-p)^{n-k},$$

missä k on onnistumisten lukumäärä ja $n - k$ epäonnistumisten lukumäärä. Olkoon X onnistumisten lukumäärä n:ssä riippumattomassa Bernoullin kokeessa. Alaluvussa 3.6 totesimme, että X noudattaa binomijakaumaa parametrein n ja p. Silloin merkitään $X \sim \mathrm{Bin}(n, p)$. Binomijakauman todennäköisyysfunktio on

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{4.2.2}$$

Esitetään nyt edellä mainittu binomijakauman luonnehdinta Bernoullin kokeiden avulla lauseen muodossa. Jatkossa oletetaan, että Bernoullin kokeet ovat toisistaan riippumattomat, vaikkei oletusta erikseen mainittaisikaan.

Lause 4.1 *Tehdään n riipumatonta Bernoullin koetta, joissa jokaisessa onnistumistodennäköisyys on p. Olkoon X onnistumisten lukumäärä. Silloin*

$$X \sim \mathrm{Bin}(n, p).$$

Todistus. Koska X on onnistumisten lukumäärä n:ssä riipumatomassa Bernoullin kokeessa, niin $X = X_1 + X_2 + \cdots + X_n$, missä $X_i \sim \mathrm{Ber}(p) = \mathrm{Bin}(1, p)$, $i = 1, 2, \ldots, n$ ovat riippumattomat ja noudattavat samaa Bernoullin jakaumaa. Merkitään nyt $X = S_n$ ja

$$S_n = X_1 + X_2 + \cdots + X_n = S_{n-1} + X_n.$$

Todistamme väitteen induktiolla.

Kun $n = 1$, niin oletuksen mukaan $X = X_1 \sim \mathrm{Ber}(p) = \mathrm{Bin}(1, p)$, joten väite pitää paikkansa tapauksessa $n = 1$. Teemme nyt induktio-oletuksen $S_{n-1} \sim \mathrm{Bin}(n-1, p)$ ja näytämme, että $S_n \sim \mathrm{Bin}(n, p)$.

Tapahtuma $\{S_{n-1} + X_n = k\}$ voidaan lausua yhdisteenä

$$\{S_{n-1} + X_n = k\} = \{S_{n-1} = k,\, X_n = 0\} \cup \{S_{n-1} = k-1,\, X_n = 1\},$$

missä $\{S_{n-1} = k,\, X_n = 0\}$ ja $\{S_{n-1} = k-1,\, X_n = 1\}$ ovat erillisiä tapahtumia. Silloin yhteenlaskusäännön nojalla

$$P(S_{n-1} + X_n = k) = P(S_{n-1} = k,\, X_n = 0) + P(S_{n-1} = k-1,\, X_n = 1).$$

Satunnaismuuttujat S_{n-1} ja X_n ovat oletuksen mukaan riippumattomat, joten

$$\begin{aligned}
&P(S_{n-1} + X_n = k) \\
&\quad = P(S_{n-1} = k)\, P(X_n = 0) + P(S_{n-1} = k-1)\, P(X_n = 1) \\
&\quad = \binom{n-1}{k} p^k (1-p)^{n-1-k}(1-p) + \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} p \\
&\quad = \binom{n-1}{k} p^k (1-p)^{n-k} + \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&\quad = \left[\binom{n-1}{k} + \binom{n-1}{k-1}\right] p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k},
\end{aligned}$$

missä viimeinen yhtäsuuruus seuraa siitä, että $\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$ [Pascalin kolmio]. Näin on lause todistettu. □

Esimerkki 4.4 Erään kasvin siementen itämistodennäköisyydeksi on ilmoitettu 0.8. Siemenen itäminen on tässä ”onnistuminen” ja itämistodennäköisyys on onnistumistodennäköisyys. Jos kylvetään 10 siementä ja siementen itämistapahtumat ovat toisistaan riippumattomat, niin kylvöä voidaan pitää

kymmenenä riippumattomana Bernoullin kokeena, joissa onnistumistodennäköisyys on 0.8. Silloin itävien siementen lukumäärä $X \sim \text{Bin}(10, 0.8)$, eli

$$f(x) = \binom{10}{x} 0.8^x \cdot 0.2^{10-x}, \qquad x = 0, 1, \ldots, 10.$$

Mikä on todennäköisyys, että vähemmän kuin 9 jyvää itää? Todennäköisyys

$$\begin{aligned} P(X < 9) = P(X \leq 8) &= 1 - \sum_{k=9}^{10} P(X = k) \\ &= 1 - 10 \cdot 0.8^9 \cdot 0.2 - 0.8^{10} = 0.6242. \end{aligned}$$

□

Laskemme usein muotoa $P(X \leq x)$ olevia todennäköisyyksiä, kuten edellisessä esimerkissä. Todennäköisyydet $P(X \leq x)$ määrittelevät jakauman kertymäfunktion

$$F(x) = P(X \leq x).$$

Kertymäfunktio määriteltiin alaluvussa 2.5.2. Binomijakauman kertymäfunktion arvot pisteissä $x = 0, 1, \ldots, n$ ovat

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}.$$

Lause 4.2 *Jos $X \sim \text{Bin}(n, p)$, niin*

1. *X:n todennäköisyysfunktio $f(x)$ on*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

kaikilla $n \in \mathbb{N}$ ja kaikilla $p \in [0, 1]$;

2. *X:n kertymäfunktio $F(y)$ on*

$$F(y) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x}\, \varepsilon(y - x)$$

kaikilla $y \in \mathbb{R}$, missä $\varepsilon(y)$ on hyppyfunktio;

3. *X:n odotusarvo, varianssi ja momenttifunktio ovat*

$$\begin{aligned} \mu = E(X) = np, \qquad \text{Var}(X) = np(1-p), \\ M(t) = E(\mathrm{e}^{tX}) = (1 - p + p\mathrm{e}^t)^n, \qquad -\infty < t < \infty. \end{aligned}$$

Todistus. 1. Binomijakauman todennäköisyysfunktio johdettiin Lauseen 4.1 todistuksessa.

2. Odotusarvo ja varianssi. Koska $X = X_1 + X_2 + \cdots + X_n$ on riippumattomien Bernoullin muuttujien $X_i \sim \text{Ber}(p)$ summa, niin

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= p + p + \cdots + p = np \end{aligned}$$

ja

$$\begin{aligned} \text{Var}(X) &= \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p). \end{aligned}$$

3. Momenttifunktio on

$$\begin{aligned} M(t) &= E\big(\mathrm{e}^{tX}\big) \\ &= E\big(\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}\big) = E\big(\mathrm{e}^{tX_1+tX_2+\cdots+tX_n}\big) \\ &= E\big(\mathrm{e}^{tX_1}\mathrm{e}^{tX_2}\cdots\mathrm{e}^{tX_n}\big) \\ &= E\big(\mathrm{e}^{tX_1}\big)\,E\big(\mathrm{e}^{tX_2}\big)\cdots E\big(\mathrm{e}^{tX_n}\big), \end{aligned}$$

missä viimeinen yhtäsuuruus seuraa lauseista 3.6 ja 3.10. Koska X_i ja X_j ($i \neq j$) ovat riippumattomat, niin e^{tX_i} ja e^{tX_j} ovat riippumattomat (Lause 3.6) ja riippumattomien satunnaismuuttujien e^{tX_1}, e^{tX_2}, …, e^{tX_n} tulon odotusarvo on yksittäisten tulon tekijöiden odotusarvojen tulo (Lause 3.10). Koska

$$M_{X_i}(t) = E(\mathrm{e}^{tX_i}) = 1 - p + p\mathrm{e}^t, \qquad i = 1, 2, \ldots, n,$$

niin

$$M(t) = (1 - p + p\mathrm{e}^t)^n \qquad \text{kaikilla } t \in \mathbb{R}.$$

Momenttifunktio itse asiassa määrittelee yksikäsitteisesti todennäköisyysfunktion (Lause 3.12). Näytämme kuitenkin vielä eksplisiittisesti, että binomitodennäköisyydet määrittelevät todennäköisyysfunktion. Koska Binomilauseen 2.6 perusteella

$$[p + (1-p)]^n = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$$

kaikilla $p \in [0,1]$, niin todennäköisyydet $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ määrittelevät todennäköisyysfunktion kaikilla $p \in [0,1]$ ja $n \geq 1$. Huomaa myös, että

$$M(0) = (1 - p + p\mathrm{e}^0)^n = [p + (1-p)]^n.$$

□

Seuraus 4.1 *Jos $X_1 \sim \text{Bin}(n_1, p)$ ja $X_2 \sim \text{Bin}(n_2, p)$ ovat riippumattomat, niin $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$.*

Todistus. Koska Lauseen 4.2 mukaan X_1:n momenttifunktio on $(1-p+pe^t)^{n_1}$ ja X_2:n momenttifunktio on $(1-p+pe^t)^{n_2}$, niin satunnaismuuttujan X_1+X_2 momenttifunktio on Lauseen 3.13 mukaan $(1-p+pe^t)^{n_1+n_2}$. Mutta Lauseen 4.2 perusteella $(1-p+pe^t)^{n_1+n_2}$ on binomijakuman $\text{Bin}(n_1+n_2,p)$ momenttifunktio. Tästä seuraa momenttifunktion yksikäsitteisyyden (Lause 3.12) nojalla, että $X_1+X_2 \sim \text{Bin}(n_1+n_2,p)$. □

Seurauslauseen 4.1 todistuksessa on käytetty esimerkin vuoksi yleistä momenttifunktiotekniikkaa. Tässä tapauksessa tulos saadaan kuitenkin helposti turvautumatta noin voimakkaisiin menetelmiin. Koska X_1 esittää onnistumisten lukumäärää n_1:ssä Bernoullin kokeessa ja X_2 onnistumisten lukumäärää n_2:ssa kokeessa, missä p on jokaisen kokeen onnistumistodennäköisyys, niin riippumattomien satunnaismuuttujien X_1 ja X_2 summa X_1+X_2 esittää onnistumisen lukumäärää (n_1+n_2):ssa kokeessa. Tämän perusteella saadaan tulos $X_1+X_2 \sim \text{Bin}(n_1+n_2,p)$. Analyyttisesti tulos voidaan tarkistaa laskemalla lauseke

$$\begin{aligned} P(X_1+X_2=k) &= \sum_{i=0}^{n_1} P(X_1=i,\, X_2=k-i) \\ &= \sum_{i=0}^{n_1} P(X_1=i)\, P(X_2=k-i) \\ &= \sum_{i=0}^{n_1} \binom{n_1}{i} p^i (1-p)^{n_1-i} \binom{n_2}{k-i} p^{k-i} (1-p)^{n_2-k+i}, \end{aligned}$$

missä $\binom{n_2}{k-i} = 0$ kaikilla $k-i > n_2$. Tästä seuraa

$$P(X_1+X_2=k) = p^k (1-p)^{n_1+n_2-k} \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{k-i}.$$

Soveltamalla hypergeometrista identiteettiä (ks. Lause 2.8)

$$\binom{n_1+n_2}{k} = \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{k-i}$$

saadaan kaivattu tulos.

4.3 Odotusaikojen jakaumat

Monissa sovelluksissa on kiinnostuksen kohteena odotusaika siihen hetkeen, että jokin tietty tapahtuma sattuu. Tässä alaluvussa käsitellään Bernoullin kokeisiin ja yksinkertaiseen satunnaisotantaan liittyviä odotusaikatehtäviä.

4.3.1 Odotusajat Bernoullin kokeissa

Tarkastellaan riippumattomien samaa Bernoullin jakaumaa noudattavien satunnaismuuttujien jonoa $X_1, X_2, \ldots, X_n$, missä $X_i \sim \text{Ber}(p)$. Määritellään satunnaismuuttujat S_n ja W_r seuraavasti:

$$S_n = X_1 + X_2 + \cdots + X_n,$$
$$W_r = r\text{:ään onnistumiseen tarvittavien yritysten määrä.}$$

Jos ajattelemme, että yhteen Bernoullin kokeeseen kuluu yhden yksikön pituinen aika, niin S_n vie n aikayksikköä. Nyt siis W_r on r:n onnistumisen saavuttamiseen tarvittava aika eli *odotusaika* ja sen mahdolliset arvot ovat r, $r+1$, $r+2$, ... Tiedämme, että $S_n \sim \text{Bin}(n,p)$, mutta mikä on W_r:n jakauma?

Esimerkki 4.5 Heitetään harhatonta lanttia, kunnes saadaan kruunu (R). Olkoon W_1 tarvittavien heittojen lukumäärä. Tapahtuma $\{W_1 = x\}$ sattuu vain silloin, kun $(x-1)$:llä ensimmäisellä heitolla on saatu pelkkiä klaavoja (L) ja x. heitolla saadaan kruunu:

$$\underbrace{\text{LLL}\ldots\text{L}}_{x-1 \text{ kertaa}}\ \text{R}.$$

Tästä seuraa, että

$$P(W_1 = x) = \frac{1}{2^x}, \qquad x = 1, 2, \ldots.$$

Satunnaismuuttujan W_1 odotusarvo on määritelmän mukaan

$$E(W_1) = \sum_{x=1}^{\infty} \frac{x}{2^x}. \tag{4.3.1}$$

Tiedämme, että

$$\sum_{x=0}^{\infty} p^x = 1 + p + p^2 + p^3 + \cdots = \frac{1}{1-p}, \qquad \text{kun } |p| < 1. \tag{4.3.2}$$

Kun derivoimme sarjan (4.3.2) termeittäin, saamme

$$0 + 1 + 2p + 3p^2 + \cdots = \sum_{x=0}^{\infty} (x+1)p^x = \frac{1}{(1-p)^2}, \qquad \text{kun } |p| < 1. \tag{4.3.3}$$

Koska sarjan (4.3.2) suppenemissäde on 1, suppenee derivointioperaation tuloksena saatu sarja (4.3.3) arvoilla $|p| < 1$. Sijoittamalla $p = \frac{1}{2}$ sarjaan (4.3.3) saadaan

$$\sum_{x=0}^{\infty} (x+1)\left(\frac{1}{2}\right)^x = 4,$$

joka voidaan esittää muodossa

$$\sum_{x=0}^{\infty} x\left(\frac{1}{2}\right)^x + \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} x\left(\frac{1}{2}\right)^x + 2 = 4,$$

missä summa $\sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x = 2$ saadaan kaavasta (4.3.2). Nyt siis odotusarvo (4.3.1) on 2.

Jos kruunun todennäköisyys on p, niin silloin

$$P(W_1 = x) = \underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ kertaa}}p = (1-p)^{x-1}p$$

ja

$$\begin{aligned} E(W_1) &= \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=0}^{\infty}(x+1)(1-p)^x \\ &= p \cdot \frac{1}{[1-(1-p)]^2} = \frac{1}{p}, \end{aligned}$$

missä sarjan summa saadaan (4.3.3):n avulla. Satunnaismuuttuja W_1 on siis kruunun tai yleisemmin 'onnistumisen' odotusaika. Jakaumaa

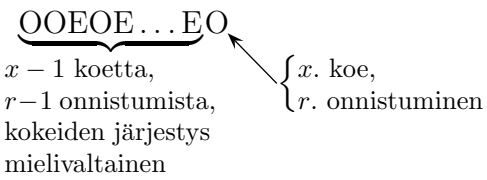
$$P(W_1 = x) = (1-p)^{x-1}p, \qquad x = 1, 2, \ldots \tag{4.3.4}$$

kutsutaan *geometriseksi jakaumaksi*. Todennäköisyydet (4.3.4) todellakin määrittelevät jakauman, koska

$$\sum_{x=1}^{\infty} P(W_1 = x) = \sum_{x=1}^{\infty}(1-p)^{x-1}p = p \cdot \sum_{x=0}^{\infty}(1-p)^x = p \cdot \frac{1}{p} = 1.$$

□

Tapahtuma $\{W_r = x\}$ sattuu, kun $(x-1)$:ssä ensimmäisessä kokeessa on saatu $r-1$ onnistumista ja x. kokeessa saadaan onnistuminen:



Nyt siis $\{W_r = x\} = \{S_{x-1} = r-1,\ X_x = 1\}$. Koska X_i:t $(i = 1, 2, \ldots, x)$ ovat riippumattomat, niin myös S_{x-1} ja X_x ovat riippumattomat. Silloin

$$\begin{aligned} P(W_r = x) &= P(S_{x-1} = r-1)\,P(X_x = 1) \\ &= \binom{x-1}{r-1}p^{r-1}(1-p)^{x-r}p = \binom{x-1}{r-1}p^r(1-p)^{x-r}, \end{aligned} \tag{4.3.5}$$

koska $S_{x-1} \sim \text{Bin}(x-1, p)$. Todennäköisyydet (4.3.5) määrittelevät ns. *negatiivisen binomijakauman.* Soveltamalla identiteettiä [ks. (2.4.5)]

$$\frac{r}{x}\binom{x}{r} = \binom{x-1}{r-1}$$

saadaan

$$P(W_r = x) = \frac{r}{x}\, P(S_x = r).$$

Toinen usein käyttökelpoinen identiteetti on

$$P(W_r > x) = P(S_x < r).$$

4.3.2 Geometrinen jakauma ja negatiivinen binomijakauma

Sanomme, että satunnaismuuttuja X noudattaa *negatiivista binomijakaumaa* parametrein r ja p, jos

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, r+2, \ldots. \tag{4.3.6}$$

Merkitsemme silloin

$$X \sim \text{NBin}(r, p).$$

Edellisessä pykälässä huomasimme, että odotusaika $W_r \sim \text{NBin}(r, p)$. Kun $r = 1$, sanomme negatiivista binomijakaumaa *geometriseksi jakaumaksi.* Geometrisen jakauman todennäköisyysfunktio on siis

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots. \tag{4.3.7}$$

Kun siis $X \sim \text{NBin}(1, p)$, niin X:n noudattaa geometrista jakaumaa parametrilla p. Merkitsemme silloin $X \sim \text{Geo}(p)$.

Lause 4.3 *Oletetaan, että* $X \sim \text{NBin}(r, p)$.

1. *Funktio (4.3.6) on negatiivisen binomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla* r *ja kaikilla* $0 < p < 1$ *ja*

2. $$E(X) = \frac{r}{p}, \qquad \text{Var}(X) = \frac{r(1-p)}{p^2},$$
$$M(t) = E(\mathrm{e}^{tX}) = \frac{(p\mathrm{e}^t)^r}{[1-(1-p)\mathrm{e}^t]^r}, \qquad t < -\log(1-p).$$

Todistus. Johdamme ensin negatiivisen binomijakauman momenttifunktion suoraan määritelmän nojalla. Koska $M(t) = E(\mathrm{e}^{tX})$, niin momenttifunktio on

$$\begin{aligned}
E(\mathrm{e}^{tX}) &= \sum_{x=r}^{\infty} \mathrm{e}^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= p^r \sum_{y=0}^{\infty} \mathrm{e}^{t(y+r)} \binom{r+y-1}{r-1} p^r (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} \binom{r+y-1}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} (-1)^y \binom{-r}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\mathrm{e}^t\right]^y \\
&= p^r \mathrm{e}^{tr} \left[1-(1-p)\mathrm{e}^t\right]^{-r} = \left[\frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}\right]^r.
\end{aligned}$$

Binomisarja $\sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\mathrm{e}^t\right]^y$ suppenee (Lause 2.7), kun $(1-p)\mathrm{e}^t < 1$, joka on yhtäpitävä epäyhtälön $t < -\log(1-p)$ kanssa.

Koska $M(0) = 1$ kaikilla positiivisilla kokonaisluvuilla r ($r \in \mathbb{N}$) ja kaikilla $0 < p < 1$, niin (4.3.6) on todennäköisyysfunktio kaikilla $r \in \mathbb{N}$ ja kaikilla $0 < p < 1$. Odotusarvo ja varianssi saadaan laskemalla ensin $M(t)$:n 1. ja 2. derivaatta ja niiden avulla

$$E(X) = M'(0) \qquad \text{ja} \qquad \mathrm{Var}(X) = M''(0) - [M'(0)]^2.$$

□

Seuraus 4.2 *Jos* $X \sim \mathrm{Geo}(p)$*, niin* $X \sim \mathrm{NBin}(1, p)$ *ja*

1. *funktio (4.3.7) on geometrisen jakauman todennäköisyysfunktio kaikilla* $0 < p < 1$ *ja*

2. $$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2},$$
$$M(t) = E(\mathrm{e}^{tX}) = \frac{p\mathrm{e}^t}{[1-(1-p)\mathrm{e}^t]}, \qquad t < -\log(1-p).$$

Olkoon Y epäonnistumisten lukumäärä Bernoullin toistokokeessa, ennen kuin saadaan r. onnistuminen. Koska r. onnistumiseen tarvittavien yritysten määrä $W_r \sim \mathrm{NBin}(r, p)$, niin

$$Y = W_r - r \qquad \text{ja} \qquad E(Y) = E(W_r) - r = \frac{r}{p} - r = \frac{r(1-p)}{p}.$$

Y:n varianssi on tietysti sama kuin W_r:n varianssi. Nyt siis $P(Y = y) = P(W_r = r + y)$ kaikilla $y = 0, 1, 2, \ldots$.

Nimitys "negatiivinen binomijakauma" on peräisin esitystavasta

$$1 = p^r \cdot p^{-r} = p^r \left[1 - (1-p)\right]^{-r} = p^r \sum_{y=0}^{\infty} \binom{-r}{y} \left[-(1-p)\right]^y,$$

mistä saadaan todennäköisyydet $P(W_r = y + r)$, $y = 0, 1, 2, \ldots$. Merkintä $\binom{-r}{y}$ on määritelmänsä mukaan

$$\binom{-r}{y} = \frac{(-r)^{(y)}}{y!} = (-1)^y \binom{r+y-1}{y},$$

missä $r > 0$ ja $y \geq 0$ ovat kokonaislukuja.

Esimerkki 4.6 Geometrisella jakaumalla ja negatiivisella binomijakaumalla on tärkeä merkitys esimerkiksi jonoteoriassa. Oletetaan, että joukko asiakkaita jonottaa pääsyä palvelutiskille. Olkoon todennäköisyys p, että jokaisella pienellä aikavälillä tulee 1 uusi asiakas (0 uutta asiakasta todennäköisyydellä $1 - p = q$). Silloin seuraavan asiakkaan odotusaika $W \sim \text{Geo}(p)$. Todennäköisyys $P(W > k)$, että seuraavan k:n aikayksikön aikana ei tule asiakasta, on

$$\begin{aligned} P(W > k) = \sum_{j=k+1}^{\infty} q^{j-1}p &= q^k(p + qp + q^2p + \cdots) \\ &= q^k = 1 - P(W \leq k). \end{aligned}$$

□

Geometrisen jakauman kertymäfunktio on siis

$$\begin{aligned} F(k) = P(W \leq k) &= \sum_{i=1}^{k} (1-p)^{i-1}p \\ &= 1 - P(W > k) = 1 - q^k, \end{aligned}$$

missä $q = 1 - p$ ja $k = 1, 2, \ldots$. Geometrisen jakauman kertymäfunktion arvot saadaan geometrisesta sarjasta, josta jakauman nimi tulee.

Usein oletetaan, että myös asiakkaan palvelemiseen käytetty aika (palveluaika) noudattaa geometrista jakaumaa. Palveluajan jakaumalla on tietysti yleensä eri parametrin p arvo kuin palvelun odotusajan jakaumalla. Geometrisella jakaumalla on "unohtamisominaisuus", joka havaitaan laskemalla seuraava ehdollinen todennäköisyys:

$$P(W > k + s \mid W > k) = \frac{P(W > k + s)}{P(W > k)} = \frac{q^{k+s}}{q^k} = q^s. \tag{4.3.8}$$

Nyt siis todennäköisyys, että asiakkaan palveleminen kestää vielä s aikayksikköä, ei riipu siitä, kuinka kauan häntä on jo palveltu. Onneksi kuitenkin käytännössä palveluaika ei aina täysin noudata geometrista jakaumaa.

Esimerkki 4.7 Banachin tulitikkuongelma. Piippua polttelevalla matemaatikolla oli tapana pitää yksi tulitikkulaatikko oikeassa ja yksi vasemmassa taskussa. Joka kerta tikkua tarvitessaan hän valitsi taskun täysin satunnaisesti, joten kummankin taskun valintatodennäköisyys on $\frac{1}{2}$. Tarkastellaan tapahtumaa, että matemaatikko huomaa laatikon olevan tyhjä. Oletetaan, että kummassakin laatikossa oli alunperin N tikkua. Mikä on todennäköisyys, että toisessa laatikossa on täsmälleen k tikkua ($k = 0, 1, \ldots, N$) silloin, kun matemaatikko havaitsee toisen laatikon olevan tyhjä?

Olkoon A tapahtuma, että matemaatikko huomaa oikeanpuoleisen laatikon olevan tyhjä ja samalla vasemman taskun laatikossa on k tikkua. Tapahtuma voi sattua täsmälleen silloin, kun oikeanpuoleisen taskun laatikosta valitaan tikku $(N+1)$. kerran ja yhteensä valintoja on tehty $N+1+N-k$ kappaletta. Teemme siis valintoja palauttamatta. Molemmissa laatikoissa on N tikkua, joten tapahtuma A on ekvivalentti tapahtuman $\{W_{N+1} = N+1+N-k\}$ kanssa. Saamme kaavalla (4.3.6) todennäköisyydeksi

$$P(W_{N+1} = N + 1 + N - k) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k+1}.$$

Koska myös todennäköisyys, että vasemmanpuoleinen laatikko huomataan tyhjäksi ja oikeanpuoleisessa on k tikkua, on $P(W_{N+1} = N+1+N-k)$, niin vastaus kysymykseen on

$$2\,P(W_{N+1} = N + 1 + N - k) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}.$$

□

4.3.3 Odotusajat peräkkäisotannassa

Oletetaan, että populaatiossa on kahdenlaisia alkioita. Valitaan populaatiosta peräkkäisotos. Käytetään nyt apuna uurnamallia. Olkoon uurnassa a *valkoista* palloa ja b *mustaa* palloa eli yhteensä $a + b = N$ palloa. Poimitaan satunnaisvalinnalla palloja uurnasta yksitellen. Määritellään satunnaismuuttujat

$$\begin{aligned} S_n &= \text{valkoisten pallojen (onnistumisten) lukumäärä} \\ &\quad\ n\text{:ssä ensimmäisessä nostossa;} \\ W_r &= r\text{:n valkoisen pallon saamiseksi tarvittavien nostojen määrä.} \end{aligned}$$

Jos ajatellaan, että nostoon menee yksi aikayksikkö, niin W_r on r:n valkoisen pallon saamiseksi tarvittava odotusaika.

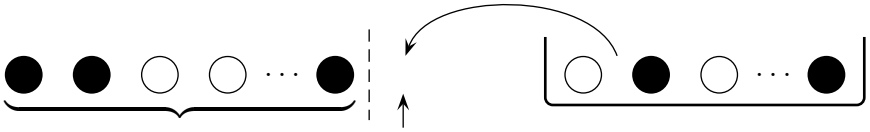
Jos otanta tehdään *palauttaen,* niin peräkkäiset nostot ovat riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on $p = a/N$. Tässä tapauksessa voidaan suoraan soveltaa edellä esitettyjä Bernoullin kokeita koskevia tuloksia.

Kun otanta tehdään *palauttamatta,* peräkkäiset nostot eivät ole riippumattomia, koska valkoisten pallojen suhteellinen osuus uurnassa riippuu siitä, mitä sieltä on jo valittu. Alaluvussa 2.6.1 osoitimme, että S_n noudattaa hypergeometrista jakaumaa, kun otanta tehdään palauttamatta (ks. myös alaluku 3.7.1). Silloin

$$P(S_n = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \tag{4.3.9}$$

kun $x = 0, 1, \ldots, n$. Mikä on todennäköisyys, että saamme x. nostossa r. valkoisen pallon?

Tapahtuma $\{W_r = x\}$ sattuu täsmälleen silloin, kun $x-1$ ensimmäisessä nostossa on saatu $r-1$ valkoista ja x. nostossa saadaan valkoinen:



Valittu $x-1$ palloa, joista valkoisia $r-1$; valintajärjestys mielivaltainen

x. valinta, r. valkoinen

Uurnassa jäljellä $N-x+1$ palloa, joista valkoisia $a-r+1$.

Voimme siis kirjoittaa $\{W_r = x\} = \{S_{x-1} = r-1,\, X_x = 1\}$, missä $S_{x-1} \sim \mathrm{HGeo}(x-1, N, a/x)$ [ks. Esimerkki 3.11 ja (3.3.6)] ja $X_x = 1$, kun valitaan valkoinen pallo x. nostossa. Tästä seuraa, että

$$\begin{aligned} P(W_r = x) &= P(S_{x-1} = r-1, X_x = 1) \\ &= P(S_{x-1} = r-1)\, P(X_x = 1 \mid S_x = r-1) \\ &= \frac{\binom{a}{r-1}\binom{N-a}{x-r}}{\binom{N}{x-1}} \cdot \frac{a-r+1}{N-x+1}, \end{aligned} \tag{4.3.10}$$

kun $x = r, r+1, \ldots, N$.

Todennäköisyys (4.3.10) voidaan kirjoittaa lausekkeena

$$P(W_r = x) = \binom{x-1}{r-1} \frac{\binom{N-x}{a-r}}{\binom{N}{a}}, \tag{4.3.11}$$

joka on *negatiivisen hypergeometrisen jakauman* todennäköisyysfunktio. Koska $\binom{x-1}{r-1} = \frac{r}{x}\binom{x}{r}$, niin

$$P(W_r = x) = \frac{r}{x} \cdot \frac{\binom{x}{r}\binom{N-x}{a-r}}{\binom{N}{a}} = \frac{r}{n}\, P(S_x = r),$$

missä $S_x \sim \mathrm{HGeo}(x, N, a/N)$. Vastaavanlainen tulos saatiin otannassa palauttaen. Samoin on jälleen helppo nähdä, että

$$P(W_r > x) = P(S_x < r).$$

Merkitään $W_r \sim \mathrm{NHGeo}(r, N, p)$, missä $p = a/N$.

4.3.4 Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma

Olemme esitelleet hypergeometrisen jakauman tarkastelemalla otantaa palauttamatta (alaluku 2.6.1). Jakauman avulla voidaan siis ratkaista otantaan liittyviä todennäköisyystehtäviä. Hypergeometrisen jakauman momenttifunktiolla $M(t)$ ei ole olemassa siistiä lauseketta, vaikka se tietysti voidaan lausua määritelmänsä mukaan äärellisenä summana, koska satunnaismuuttujan arvojoukko on äärellinen. Hypergeometrisen jakauman odotusarvon ja varianssin laskeminen ei myöskään ole aivan helppo tehtävä.

Olemme merkinneet populaation alkioiden lukumäärää $N = a + b$, joista a kappaletta on tyyppiä A ja b kappaletta tyyppiä B. Esimerkiksi tuotepopulaatiossa on a viallista. Tyyppiä A olevien alkioden suhteellinen osuus on $p = a/N$. Tyyppiä A olevan alkion valinta on ”onnistuminen” ja tyypin B valinta ”epäonnistuminen”. Valitaan populaatiosta n:n alkion otos palauttamatta. Olkoon X onnistuneiden valintojen lukumäärä otoksessa. On selvää, että $0 \leq X \leq n$. Koska populaatiossa on pN kappaletta tyyppiä A olevia alkioita ja $(1-p)N$ kappaletta tyyppiä B, niin $X \leq pN$ ja $n - X \leq (1-p)N$. Siksi X:n arvoalue S on ehdon

$$\max\{0, n - (1-p)N\} \leq x \leq \min\{n, pN\}$$

toteuttavien kokonaislukujen x joukko.

Kun X noudattaa hypergeometrista jakaumaa $\mathrm{HGeo}(n, N, p)$, niin X:n todennäköisyysfunktio on

$$f(x) = P(X = x) = \frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}, \qquad x \in S. \tag{4.3.12}$$

Huomattakoon, että todennäköisyys (4.3.12) on määritelty myös arvoilla $x \notin S$, mutta silloin $f(x) = 0$.

Lause 4.4 *Oletetaan, että* $X \sim \mathrm{HGeo}(n, N, p)$. *Silloin*

$$E(X) = np \qquad ja \qquad \mathrm{Var}(X) = \frac{N-n}{N-1} np(1-p).$$

Todistus. Hypergeometrisen jakauman odotusarvo laskettiin esimerkissä 3.11 ja alaluvussa 3.7.1. Varianssi voidaan laskea vastaavalla tavalla. □

Lause 4.5 *Oletetaan, että* $Y \sim \mathrm{NHGeo}(r, N, p)$. *Silloin*

$$E(Y) = r \cdot \frac{N+1}{Np+1} \qquad ja \qquad \mathrm{Var}(Y) = \frac{rN(N+1)(1-p)(Np+1-r)}{(Np+1)^2(Np+2)}.$$

Mainitsimme jo alaluvussa 2.8.1, että binomijakaumaa voidaan käyttää hypergeometrisen jakauman likiarvona, kun N on suuri. Erityisesti, kun N on ääretön tai hyvin suuri (verrattuna otoskokoon), on yhdentekevää, käytetäänkö otantaa palauttaen vai palauttamatta. Oletetaan nyt, että

$$X_N \sim \mathrm{HGeo}(n, N, p) \qquad \text{ja} \qquad X \sim \mathrm{Bin}(n, p).$$

Kun parametrit n ja p ovat annettuja vakioita ja N kasvaa rajatta, voimme osoittaa, että X_N:n jakauma lähestyy X:n jakaumaa. Silloin siis

$$X_N \xrightarrow{d} X, \quad \text{kun} \quad N \to \infty.$$

Koska $X \sim \mathrm{Bin}(n, p)$, niin

$$X_N \xrightarrow{d} \mathrm{Bin}(n, p),$$

eli X_N:n jakauma lähestyy binomijakaumaa, jonka parametrit ovat n ja p. Sanomme myös, että X_N:n jakauma suppenee kohti X:n jakaumaa N:n kasvaessa. Kutsumme X:n jakaumaa X_N:n *asymptoottiseksi jakaumaksi.*

Lauseen 3.5 mukaan satunnaismuuttujilla on sama jakauma, jos niillä on sama kertymäfunktio. Voimme nyt tarkastella satunnaismuuttujien jonoa

$$\{\, X_N;\ N = 1, 2, \dots \} = X_1, X_2, \dots$$

ja vastaavaa kertymäfunktioiden jonoa

$$\{\, F_N;\ N = 1, 2, \dots \} = F_1, F_2, \dots,$$

missä $F_N(x)$ on X_N:n kertymäfunktio.

Määritelmä 4.2 Jono $\{\, X_N;\ N = 1, 2, \dots \}$ suppenee jakaumaltaan kohti satunnaismuuttujaa X, jos

$$\lim_{N \to \infty} F_N(x) = F(x)$$

kaikissa pisteissä $x \in \mathbb{R}$, joissa X:n kertymäfunktio $F(x)$ on jatkuva.

Diskreettien satunnaismuuttujien tapauksessa voidaan helposti todistaa tulos, joka osoitaa, että suppenemista jakaumamielessä voidaan tarkastella yhtä hyvin myös todennäköisyysfunktioiden avulla.

Lause 4.6 *Olkoon $\{\, X_N;\ N = 1, 2, \dots \}$ sellainen epänegatiivisten kokonaislukuarvoisten satunnaismuuttujien jono, että X_N:n todennäköisyysfunktio on $f_N(k)$, $N = 1, 2, \dots$. Olkoon X epänegatiivinen kokonaislukuarvoinen satunnaismuuttuja, jonka todennäköisyysfunktio on $f(k)$. Silloin*

$$X_N \xrightarrow{d} X \Leftrightarrow \lim_{N \to \infty} f_N(k) = f(k)$$

kaikilla epänegatiivisilla kokonaisluvuilla k.

Todistus. Jätetään harjoitustehtäväksi. □

Lause 4.7 *Jos* $X_N \sim \mathrm{HGeo}(n, N, p)$*, niin*

$$X_N \xrightarrow{d} \mathrm{Bin}(n, p), \quad \textit{kun} \quad N \to \infty.$$

Todistus. Käytetään lausetta 4.6 ja osoitetaan, että $P(X_N = k) = f_N(k) \to f(k)$ kaikilla epänegatiivisilla kokonaisluvuilla k, kun $N \to \infty$. Yksityiskohdat jätetään lukijan pohdittavaksi. □

4.3.5 Tasajakauma

Diskreetti tasajakauma esiteltiin ensimmäisen kerran alaluvussa 2.5.4. Satunnaismuuttuja X, jonka arvoavaruus on $S = \{1, 2, \ldots, N\}$, noudattaa diskreettiä tasajakaumaa, jos

$$P(X = x) = \frac{1}{N}, \qquad k = 1, 2, \ldots, N.$$

Silloin merkitään $X \sim \mathrm{Tasd}(1, 2, \ldots, N)$, missä $N \geq 1$ on annettu positiivinen kokonaisluku. Jos $X \sim \mathrm{Tasd}(1, 2, \ldots, N)$, niin

$$E(X) = \frac{N+1}{2} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{(N+1)(N-1)}{12}.$$

4.4 Poissonin jakauma

Satunnaismuuttuja X, jonka todennäköisyysfunktio on

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots \tag{4.4.1}$$

noudattaa *Poissonin jakaumaa* parametrilla $\lambda > 0$, joka on Poissonin jakauman odotusarvo. Silloin merkitään

$$X \sim \mathrm{Poi}(\lambda).$$

Poissonin jakaumalla on runsaasti sovelluksia eri aloilla. Sitä voidaan käyttää myös binomijakauman $\mathrm{Bin}(n, p)$ likiarvona, kun n on suuri ja p pieni. Silloin siis pätee

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\mathrm{e}^{-np}(np)^x}{x!}.$$

Lause 4.8 *Olkoon* $X \sim \mathrm{Poi}(\lambda)$*. Silloin*

1. *funktio (4.4.1) on Poissonin jakauman todennäköisyysfunktio kaikilla* $\lambda > 0$ *ja*

2.

$$\mu = E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda,$$
$$M(t) = E(\mathrm{e}^{tX}) = \exp(\lambda \mathrm{e}^t - \lambda).$$

Todistus. Sovelletaan eksponenttifunktion sarjakehitelmää

$$\exp(\lambda) = \mathrm{e}^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}. \tag{4.4.2}$$

1. Ensinnäkin $f(x) \geq 0$ kaikilla $x = 0, 1, 2, \ldots$, ja eksponenttifunktion sarjakehitelmän (4.4.2) perusteella

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^\lambda = 1.$$

2. Johdetaan ensin momenttifunktion $M(t)$ lauseke:

$$\begin{aligned} M(t) = E(\mathrm{e}^{tX}) &= \sum_{x=0}^{\infty} \mathrm{e}^{tx} \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda} \\ &= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda \mathrm{e}^t)^x}{x!} \\ &= \mathrm{e}^{-\lambda} \cdot \exp(\lambda \mathrm{e}^t) = \exp(\lambda \mathrm{e}^t - \lambda). \end{aligned}$$

Odotusarvo ja varianssi saadaan sitten laskemalla $M(t)$:n 1. ja 2. derivaatta ja soveltamalla identiteettejä

$$E(X) = M'(0) \qquad \text{ja} \qquad \mathrm{Var}(X) = M''(0) - [M'(0)]^2.$$

□

Riippumattomien Poissonin jakaumaa noudattavien satunnaismuuttujien summa noudattaa myös Poissonin jakaumaa.

Lause 4.9 *Olkoot X_1, X_2, …, X_n riippumattomat ja $X_i \sim \mathrm{Poi}(\lambda_i)$, $i = 1, 2, \ldots, n$. Olkoon $Y = X_1 + X_2 + \cdots + X_n$. Silloin*

$$Y \sim \mathrm{Poi}(\lambda),$$

missä $\lambda = \sum_{i=1}^{n} \lambda_i$.

Todistus. Seurauslauseen 3.1 mukaan

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n} \exp(\lambda_i \mathrm{e}^t - \lambda_i) = \exp[(\mathrm{e}^t - 1)\lambda], \end{aligned}$$

missä $\lambda = \sum_{i=1}^{n} \lambda_i$. Lauseesta 3.12 seuraa sitten väite $Y \sim \mathrm{Poi}(\lambda)$. □

Jos riippumattomat $X_1, X_2, \ldots, X_n$ noudattavat samaa Poissonin jakaumaa Poi(λ), niin Lauseen 4.9 mukaan niiden summa $Y = X_1 + X_2 + \cdots + X_n$ noudattaa Poissonin jakaumaa Poi($n\lambda$). Poissonin jakauma on hyvä binomijakauman Bin(n, p) likiarvo silloin, kun n on suuri ja p pieni.

Kun $X \sim \text{Bin}(n, p)$, niin binomitodennäköisyys on

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{4.4.3}$$

Annetaan nyt p:n riippua n:stä ja merkitään lausekkeessa (4.4.3) $p = p_n$. Valitaan erityisesti

$$p_n = \frac{\lambda}{n}, \qquad n \geq 1.$$

Tarkastellaan nyt binomijakaumien jonoa

$$\text{Bin}(1, p_1),\ \text{Bin}(2, p_2),\ \text{Bin}(3, p_3), \ldots$$

ja vastaavaa satunnaismuuttujien $X_1, X_2, X_3, \ldots$ jonoa, missä $X_n \sim \text{Bin}(n, p_n)$, $n \geq 1$. Nyt siis

$$P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}, \qquad 0 \leq x \leq n. \tag{4.4.4}$$

Merkitään todennäköisyyttä (4.4.4) lyhyesti $b_x(n)$

Kiinnitetään nyt x ja annetaan n:n kasvaa rajatta. Osoittautuu, että $b_x(n)$ suppenee kaikilla x. Valitaan ensin $x = 0$. Silloin saamme

$$\lim_{n\to\infty} b_0(n) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \mathrm{e}^{-\lambda}. \tag{4.4.5}$$

Se on eräs keskeinen eksponenttifunktioon liittyvä kaava, joka pitäisi analyysin kurssien perusteella muistaa. Tulos (4.4.5) saadaan esimerkiksi Taylorin sarjan

$$\log(1-p) = -\sum_{n=1}^{\infty} \frac{p^n}{n}$$

avulla, kun sijoitetaan $p = \frac{\lambda}{n}$:

$$\begin{aligned} \log\left(1 - \frac{\lambda}{n}\right)^n &= n \log\left(1 - \frac{\lambda}{n}\right) = n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \frac{\lambda^3}{3n^3} - \cdots\right) \\ &= -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \cdots \\ &= -\lambda - \frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \cdots\right). \end{aligned} \tag{4.4.6}$$

Kun $n \to \infty$, niin $\frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \cdots\right) \to 0$ ja siksi $\log\left(1 - \frac{\lambda}{n}\right)^n \to -\lambda$.

Lasketaan seuraavaksi $b_x(n)$:n raja-arvo, kun $x > 0$. Tarkastellaan peräkkäisten binomitodennäköisyyksien suhdetta

$$\frac{b_{x+1}(n)}{b_x(n)} = \frac{n-x}{x+1}\left(\frac{\lambda}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-1} = \frac{\lambda}{x+1}\left(\frac{n-x}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-1},$$

missä $\frac{n-x}{n} \to 1$ ja $1 - \frac{\lambda}{n} \to 1$, kun $n \to \infty$. Tästä seuraa, että

$$\lim_{n\to\infty} \frac{b_{x+1}(n)}{b_x(n)} = \frac{\lambda}{x+1}. \tag{4.4.7}$$

Kun lähdetään tuloksesta (4.4.5) ja käytetään hyväksi raja-arvoa (4.4.7), saadaan

$$\begin{aligned}
\lim_{n\to\infty} b_1(n) &= \frac{\lambda}{1} \lim_{n\to\infty} b_0(n) = \lambda \mathrm{e}^{-\lambda}, \\
\lim_{n\to\infty} b_2(n) &= \frac{\lambda}{2} \lim_{n\to\infty} b_1(n) = \frac{\lambda^2}{1\cdot 2}\mathrm{e}^{-\lambda}, \\
&\vdots \\
\lim_{n\to\infty} b_x(n) &= \frac{\lambda}{x} \lim_{n\to\infty} b_{x-1}(n) = \frac{\lambda^x}{1\cdot 2\cdots x}\mathrm{e}^{-\lambda}.
\end{aligned}$$

Olemme siis näyttäneet, että

$$\lim_{n\to\infty} b_x(n) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \tag{4.4.8}$$

missä raja-arvo on $P(X = x)$, kun $X \sim \mathrm{Poi}(\lambda)$. Tulos (4.4.8) tunnetaan *Poissonin raja-arvolakina.*

Satunnaismuuttujat noudattavat samaa jakaumaa, kun niillä on sama kertymäfunktio (Lause 3.5). Jos diskreetit satunnaismuuttujat noudattavat samaa jakaumaa, niin niillä on sama todennäköisyysfunktio. Jos satunnaismuuttujan X_n jakauma lähenee X:n jakaumaa n:n kasvaessa rajatta, niin X_n:n todennäköisyysfunktio lähenee X:n todennäköisyysfunktiota, mikäli jakaumat ovat diskreettejä (Lause 4.6). Vaikka edellä olemmekin johtaneet Poissonin raja-arvolain (4.4.8), esitetään tulos vielä *Poissonin lauseena.*

Lause 4.10 (Poissonin lause) *Olkoon $X_n \sim \mathrm{Bin}(n, p)$. Silloin*

$$X_n \xrightarrow{d} \mathrm{Poi}(\lambda),$$

kun $n \to \infty$ siten, että $np = \lambda$.

Todistus. Koska $np = \lambda$, voimme merkitä $p = \lambda/n$. Todistus perustuu

Lauseeseen 4.6. Jos $X_n \sim \mathrm{Bin}(n,p)$, niin

(4.4.9)
$$\begin{aligned} f_{X_n}(x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \frac{n!}{(n-x)!\,n^x}\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Kiinteällä x:n arvolla

$$\lim_{n\to\infty}\left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right] = 1$$

ja

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1.$$

Näistä tuloksista yhdessä raja-arvon (4.4.5) kanssa seuraa

$$\lim_{n\to\infty} f_{X_n}(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}.$$

Satunnaismuuttujan X_n jakauma lähestyy siis Poissonin jakaumaa $\mathrm{Poi}(\lambda)$, kun $n \to \infty$. □

Poissonin jakaumaa sanotaan usein harvinaisten tapahtumien laiksi. Tämä luonnehdinta perustuu edellisessä lauseessa esitettyyn ominaisuuteen. Jos tehdään suuri määrä riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on hyvin pieni, niin silloin Lauseen 4.10 mukaan onnistumisten lukumäärä noudattaa likimain Poissonin jakaumaa. Esimerkiksi suuri määrä ihmisiä on päivittäin alttiina liikenneonnettomuuksille. Yksittäisen henkilön todennäköisyys (onnistumistodennäköisyys!) joutua onnettomuuteen on pieni, mutta onnettomuuksille alttiina olevien henkilöiden lukumäärä n on suuri. Silloin onnettomuuksien lukumäärä noudattaa likimain Poissonin jakaumaa.

Lause 4.11 *Olkoot X ja Y sellaiset riippumattomat satunnaismuuttujat, että $X \sim \mathrm{Poi}(\lambda_1)$ ja $Y \sim \mathrm{Poi}(\lambda_2)$. Silloin X:n ehdollinen jakauma ehdolla $X+Y$ on binomijakauma.*

Todistus. Olkoot m ja n sellaiset epänegatiiviset kokonaisluvut, että $m < n$.

Silloin

$$\begin{aligned}
P(X = m \mid X + Y = n) &= \frac{P(X = m, X + Y = n)}{P(X + Y = n)} \\
&= \frac{P(X = m, Y = n - m)}{P(X + Y = n)} \\
&= \frac{P(x = m)\, P(Y = n - m)}{P(X + Y = n)} \\
&= \frac{\mathrm{e}^{-\lambda_1}(\lambda_1^m/m!)\mathrm{e}^{-\lambda_2}[\lambda_2^{n-m}/(n-m)!]}{\mathrm{e}^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n/n!} \\
&= \binom{n}{m} \frac{\lambda_1^m \lambda_2^{n-m}}{(\lambda_1+\lambda_2)^n} \\
&= \binom{n}{m} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-m}
\end{aligned}$$

on binomitodennäköisyys kaikilla $m = 0, 1, \ldots, n$. Näin on lause todistettu. □

Lauseella 4.11 on tärkeä merkitys esimekiksi frekvenssiaineistojen analyysissa.

Esimerkki 4.8 Tiedetään, että auto-onnettomuuksien lukumäärä aikayksikössä (esimerkiksi kuukaudessa) noudattaa Poissonin jakaumaa. Tarkastellaan eräällä tieosuudella lokakuussa sattuvien onnettomuuksien lukumäärää. Aikaisempien tilastojen perusteella voidaan olettaa, että auto-onnettomuuksien lukumäärä Z kyseisellä tieosuudella (kuukaudessa) noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda)$. Onnettomuudet luokitellaan mahdollisten henkilövahinkojen mukaan vakaviin ja lieviin (jokainen onnettomuus kuuluu toiseen näistä luokista). Vakavien onnettomuuksien lukumäärä $X \sim \mathrm{Poi}(\lambda_1)$ ja lievien lukumäärä $Y \sim \mathrm{Poi}(\lambda_2)$. Lisäksi X ja Y ovat toisistaan riippumattomat. Koska $Z = X + Y$, niin $E(Z) = E(X) + E(Y)$ eli $\lambda = \lambda_1 + \lambda_2$.

Tutkijat valitsivat poliisin tiedostoista satunnaisesti valitun kuukauden (vuonna 2003) onnettomuudet. He havaitsivat onnettomuuksien lukumääräksi 120 ($n = 120$), mutta he eivät olleet vielä luokitelleet onnettomuuksia. Mitä jakaumaa noudattaa vakavien onnettomuuksien lukumäärä? Lauseen 4.11 perusteella

$$P(X = m \mid Z = 120) = \binom{120}{m} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{120-m},$$

$m = 0, 1, \ldots, 120$. Vakavien onnettomuuksien lukumäärä noudattaa siis binomijakaumaa $\mathrm{Bin}\big(120, \frac{\lambda_1}{\lambda_1+\lambda_2}\big)$. Aikaisempien onnettomuustilastojen perusteella voimme arvioida parametrit λ_1 ja λ_2, joiden avulla saamme estimaatin parametrille $\frac{\lambda_1}{\lambda_1+\lambda_2}$. Kun tutkijat olivat luokitelleet nuo 120 onnettomuutta, aineistossa havaittiin 15 vakavaa onnettomuutta. Koska $E(X \mid Z = 120) = \lambda_1$, niin havainnon 15 pitäisi osua "melko lähelle" arvoa λ_1. □

4.5 Poissonin prosessi

4.5.1 Laskuriprosessi

Stokastinen prosessi $\{ N(t),\ t \geq 0 \}$ on *laskuriprosessi,* jos $N(t)$ on ajankohtaan t mennessä sattuneiden "tapahtumien" lukumäärä.

Esimerkki 4.9 Seuraavassa luetellaan esimerkkejä laskuririprosesseista.

1. Jos $N(t)$ on annetulla tieosuudella hetkeen t mennessä sattuneiden onnettomuuksien lukumäärä, niin $\{ N(t),\ t \geq 0 \}$ on tapahtumaan "onnettomuus" liittyvä laskuriprosessi.

2. Olkoon $N(t)$ palvelutiskille tulleiden asiakkaiden lukumäärä hetkeen t mennessä. Tapahtuma on "asiakkaan tulo palvelutiskille" ja $\{ N(t),\ t \geq 0 \}$ on tapahtumaan liittyvä laskuriprosessi.

3. $N(t)$ on vuoden alusta hetkeen t mennessä syntyneiden lasten lukumäärä kaupungissa A.

4. $N(t)$ on jalkapallojoukkueen A tekemien maalien lukumäärä kauden alusta ajankohtaan t mennessä.

□

Laskuriprosessin tulee toteuttaa seuraavat ominaisuudet:

1. $N(t) \geq 0$.

2. $N(t) \in \mathbb{N}$, eli $N(t)$ on kokonaislukuarvoinen.

3. Jos $s < t$, niin $N(s) \leq N(t)$.

4. Kun $s < t$, niin $N(t) - N(s)$ on välillä $(s, t]$ sattuneiden tapahtumien lukumäärä.

Laskuriprosessi on *riippumattomien lisäysten* prosessi, jos erillisillä aikaväleillä sattuvien tapahtumien lukumäärät ovat riippumattomat. Esimerkiksi satunnaismuuttujat $N(2)$ ja $N(10) - N(2)$ ovat riippumattomat, jos $N(t)$ on riippumattomien lisäysten laskuriprosessi. Laskuriprosessin *lisäykset ovat stationaariset,* jos millä tahansä välillä sattuvien tapahtumien lukumäärän jakauma riippuu vain välin pituudesta. Jos $N(t)$ on stationaarinen laskuriprosessi, niin satunnaismuuttujilla $N(t_2) - N(t_1)$ ja $N(t_2 + s) - N(t_1 + s)$ on sama jakauma kaikilla väleillä $(t_1, t_2]$ ja $(t_1 + s, t_2 + s]$, missä $t_2 > t_1$ ja $s > 0$.

4.5.2 Poissonin prosessin määritely

Poissonin prosessi on yksi tärkeimpiä laskuriprosesseja. Se määritellään seuraavasti:

Määritelmä 4.3 Laskuriprosessi $\{ N(t),\ t \geq 0 \}$ on Poissonin prosessi, jonka intensiteetti on λ ($\lambda > 0$), jos

1. $N(0) = 0$.
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella h:n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on λh:

$$P[N(h+t) - N(t) = x] = \mathrm{e}^{-\lambda h} \frac{(\lambda h)^x}{x!}, \qquad x = 0, 1, \ldots$$

kaikilla $h, t \geq 0$.

Laskuriprosessin osoittaminen Poissonin prosessiksi Määritelmän 4.3 avulla saattaa olla hankalaa. Ei ole mitään yksinkertaista keinoa tarkistaa esimerkiksi ehdon 3 pätevyyttä. Siksi esitetään vielä toinen määritelmä, jonka avulla voi olla helpompaa tunnistaa prosessi. Voidaan osoitaa, että määritelmät 4.3 ja 4.4 ovat yhtäpitävät.

Määritelmä 4.4 Laskuriprosessi $\{ N(t),\ t \geq 0 \}$ on Poissonin prosessi, jonka intensiteetti on λ ($\lambda > 0$), jos

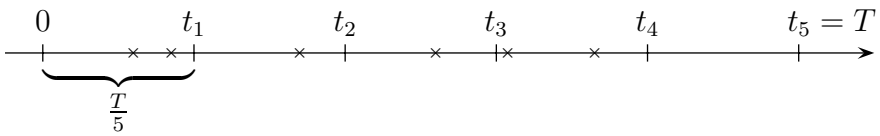
1. $N(0) = 0$.
2. Prosessin lisäykset ovat stationaariset ja riippumattomat.
3. $P\big(N(t+h) - N(t) = 1\big) = \lambda h + o(h)$.
4. $P\big(N(t+h) - N(t) \geq 2\big) = o(h)$.

Määritelmässä 4.4 käytetään merkintää $o(h)$. Sanomme, että funktio $f(\cdot) = o(h)$, jos

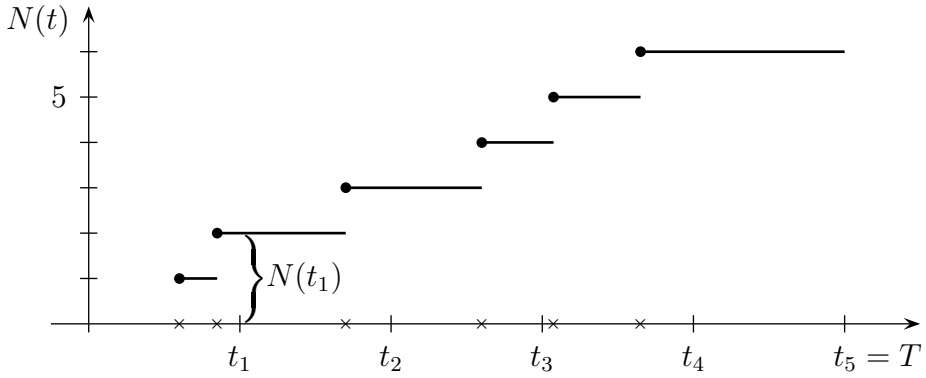
$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

Esimerkki 4.10 Tieliikenneonnettomuudet. Havainnoidaan esimerkiksi jollain tieosuudella sattuvien auto-onnettomuuksien lukumäärää. Onnettomuuksien määrä noudattaa tavallisesti varsin hyvin Poissonin prosessia. □

Tarkastellaan nyt hieman lähemmin Poissonin prosessin oletuksia. Oletetaan, että onnettomuuksien lukumäärä eräällä tieosuudella noudattaa aikavälillä $(0, T)$ Poissonin prosessia, jonka intensiteetti on λ. Aikaväli voi olla esimerkiksi ruuhka-aika tiettynä perjantai-iltapäivänä klo 15–19 ja tieosuus jokin ulosmenotie. Oheisessa kuviossa on havaitut onnettomuudet merkitty aika-akselille.

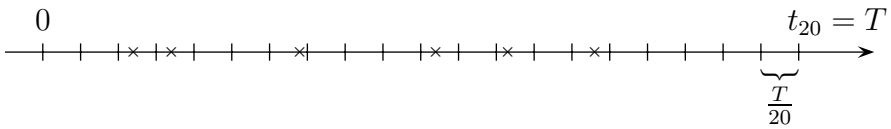


Tarkasteluväli $(0, T]$ on jaettu viiteen yhtä pitkään osaväliin, joiden pituudet ovat $T/5$. Nyt esimerkiksi 1. osavälillä sattuneiden onnettomuuksien lukumäärä on $N(t_1) - N(0) = N(t_1)$, joka on siis hetkeen t mennessä sattuneiden onnettomuuksien lukumäärä. Kuvioon 4.2 on piirretty prosessin $\{\, N(t),\ t \in (0, T] \,\}$ realisaatio, missä havaintoina ovat kyseiset onnettomuudet.



Kuvio 4.2. Poissonin prosessin $\{\, N(t),\ t \in (0, T] \,\}$ erään realisaation kuvaaja.

Määritelmän 4.4 oletuksen 2 mukaan lisäykset $N(t_1) - N(0)$, $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, $N(t_4) - N(t_3)$ ja $N(t_5) - N(t_4)$ ovat riippumattomat ja noudattavat samaa jakaumaa. Määritelmän 4.4 oletukset 3 ja 4 tarkoittavat, että tapahtumat (onnettomuudet) sattuvat yksittäin ja samalla intensiteetillä koko tarkastelujakson ajan. Koska tapahtumat ovat erillisiä pisteitä, niin aina voidaan valita niin hienojakoinen välin ositus, että kullakin osavälillä on korkeintaan 1 tapahtuma. Jos tarkastelemassamme esimerkkitapauksessa valitaan osavälin pituudeksi $T/20$, sattuu tässä osituksessa kullekin osavälille korkeintaan 1 tapahtuma. Riippuen tietysti kulloisestakin havaintojaksosta, kuinka hienojakoinen ositus tarvitaan.



Todennäköisyys, että T/n:n pituiselle osavälille sattuu havainto, on Määritelmän 4.4 oletuksen 3 mukaan

$$P\left[N\left(t + \frac{T}{n}\right) - N(t) = 1\right] = \lambda \cdot \frac{T}{n} + o\left(\frac{T}{n}\right).$$

Vastaavasti todennäköisyys, että osavälillä sattuu enemmän kuin yksi havainto, on häviävän pieni, sillä Määritelmän 4.4 oletuksen 4 mukaan

$$P\left[N\left(t+\frac{T}{n}\right)-N(t)\geq 2\right]=o\left(\frac{T}{n}\right).$$

Voimme siis olettaa, että kullakin osavälillä sattuu vain 0 tai 1 tapahtumaa, kun n on riittävän suuri.

Määritellään nyt satunnaismuuttujat

$$X_i=N\left(\frac{iT}{n}\right)-N\left(\frac{(i-1)T}{n}\right),\qquad i=1,2,\ldots,n.$$

Muuttujia X_i voidaan käsitellä toisistaan riippumattomina Bernoullin jakaumaa noudattavina satunnaismuuttujina:

$$X_i\sim \mathrm{Ber}\left(\frac{\lambda T}{n}\right),\qquad i=1,2,\ldots,n.$$

Koko välillä $(0,T]$ havaittujen tapahtumien lukumäärä on

$$S_n=X_1+X_2+\cdots+X_n,$$

joka noudattaa binomijakaumaa $\mathrm{Bin}\big(n,\frac{\lambda T}{n}\big)$. Koska $E(S_n)=n\cdot\frac{\lambda T}{n}=\lambda T$ kaikilla $n\in\mathbb{N}$, niin $E(S_n)=\lambda T$, kun $n\to\infty$. Voimme siis soveltaa Poissonin lausetta (Lause 4.10), jonka mukaan S_n noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda T)$, kun n kasvaa rajatta. Näin esimerkiksi todennäköisyys, että välillä $(0,T]$ sattuu x onnettomuutta, on

$$P\big(N(T)=x\big)=\frac{\mathrm{e}^{-\lambda T}(\lambda T)^x}{x!}.$$

Todennäköisyys riippuu vain välin pituudesta T ja intensiteetistä $\lambda>0$.

4.5.3 Satunnaistapahtumat tila-avaruudessa

Poissonin prosessilla mallinnetaan myös ilmiöitä, jotka tapahtuvat satunnaisesti tila-avaruudessa. Silloin Määritelmän 4.4 ehdot voidaan luonnehtia seuraavasti:

1. *Riippumattomuus.* Erillisillä alueilla sattuvien tapahtumien lukumäärät ovat riippumattomat.

2. *Yksittäisyys.* Todennäköisyys, että alueella sattuu enemmän kuin yksi tahtuma, on häviävän pieni.

3. *Homogeenisuus.* Tapahtumat sattuvat samalla intensiteetillä koko tarkasteltavalla alueella.

Tarkastellaan esimerkiksi Poissonin prosessia tasossa. Silloin todennäköisyys, että pinta-alaltaan A:n kokoisella alueella sattuu x tapahtumaa, on

$$f_A(x) = \frac{\mathrm{e}^{-\lambda A}(\lambda A)^x}{x!}, \qquad x = 0, 1, \ldots,$$

missä λ on tapahtumien lukumäärän odotusarvo yhtä pinta-alayksikköä kohti. Jos Poissonin prosessia noudattavat tapahtumat sattuvat kolmiulotteisessa avaruudessa, niin silloin V:n kokoiseen tilaan osuu x tapahtumaa todennäköisyydellä

$$f_V(x) = \frac{\mathrm{e}^{-\lambda V}(\lambda V)^x}{x!}, \qquad x = 0, 1, \ldots,$$

missä λ on tapahtumien lukumäärän odotusarvo yhtä tilavuus-yksikköä kohti.

Esimerkki 4.11 Leipomo valmistaa suuren erän pullataikinaa, josta tehdään rusinapullia. Leipuri haluaa, että ainakin 95 % pullista sisältää vähintään 2 rusinaa. Kuinka monta rusinaa pullaa kohti pitäisi sekoittaa taikinaan?

Olkoon pullan tilavuus $V = 1$. Kun rusinat sekoitetaan hyvin taikinaan, on kaikilla pullilla sama todennäköisyys sisältää rusinoita (homogeenisuus). Koska taikina on suuri, ovat eri pulliin sattuvien rusinoiden lukumäärät toisistaan riippumattomat. Todennäköisyys, että pieneen pullaan sattuu enemmän kuin yksi rusina, on hyvin pieni.

Tässä tilanteessa on kyse Poissonin prosessista 3-ulotteisessa tila-avaruudessa. Pullassa on x rusinaa todennäköisyydellä

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

ja ainakin 2 rusinaa todennäköisyydellä

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \mathrm{e}^{-\lambda} - \mathrm{e}^{-\lambda}\lambda. \end{aligned}$$

Leipuri vaatii, että

$$1 - \mathrm{e}^{-\lambda} - \mathrm{e}^{-\lambda}\lambda \geq 0.95.$$

Epäyhtälö toteutuu, kun $\lambda \geq 4.74$, joten rusinoita on sekoitettava taikinaan 5 rusinaa pullaa kohti. □

4.6 Kaksiulotteiset jakaumat

Tilastollisissa sovelluksissa tarkastellaan tavallisesti useita muuttujia samanaikaisesti. Esimerkiksi haastattelututkimuksessa valitaan opiskelijoista satunnaisotosotos. Jokaiselta otokseen osuneelta kysytään useita kysymyksiä ja lisäksi saadaan haastateltavien taustatiedot kuten ikä, sukupuoli, asuinpaikka jne. Otosavaruudessa on siis määritelty useita muuttujia (kysymykset ja

taustamuuttujat). Tällainen asetelma mahdollistaa muuttujien välisten riippuvuuksien tarkastelun. Seuraavassa esitellään usean muuttujan jakaumiin liittyvää käsitteistöä. Ensin käsitellään kahden muuttujan tapaus yksityiskohtaisesti. Sen jälkeen on suoraviivaista yleistää tarkastelu usean muuttujan tapaukseen.

Määritelmä 4.5 Olkoot X ja Y samassa otosavaruudessa määritellyt diskreetit satunnaismuuttujat ja olkoon kaksiulotteisen diskreetin satunnaismuuttujan (X, Y) arvoavaruus S. Tapahtuman "$\{X = x\}$ ja $\{Y = y\}$ sattuvat" todennäköisyyttä merkitään $P(X = x, Y = y) = f(x, y)$. Funktio $f(x, y)$ on (X, Y):n todennäköisyysfunktio (tnf), jolla on seuraavat ominaisuudet:

1. $0 \leq f(x, y) \leq 1$,
2. $\sum\limits_{(x,y)\in S} f(x, y) = 1$ ja
3. $P[(X, Y) \in A] = \sum\limits_{(x,y)\in A} f(x, y)$, missä $A \subset S$.

Funktiota $f(x, y)$ sanotaan myös X:n ja Y:n yhteisjakauman todennäköisyysfunktioksi. Moniulotteista satunnaismuuttujaa kutsutaan satunnaisvektoriksi (SV).

Esimerkki 4.12 Olkoon (X, Y) satunnaisvektori, jonka arvoavaruus on

$$S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$$

ja todennäköisyysfunktio

$$f(x, y) = c(x + 2y), \qquad (x, y) \in S.$$

Todennäköisyysfunktion ominaisuuksista seuraa, että

$$\sum_{(x,y)\in S} c(x + 2y) = c(2 + 4 + 1 + 3 + 2) = 12c = 1,$$

joten $c = \frac{1}{12}$. Silloin esimerkiksi

$$P(X > Y) = f(1, 0) + f(2, 0) = \frac{3}{12}$$

ja

$$P(X \geq Y) = f(1, 0) + f(2, 0) + f(1, 1) = \frac{6}{12}.$$

□

4.6.1 Reunajakauma ja ehdollinen jakauma

Jos (X, Y) on kaksiulotteinen satunnaisvektori, niin X ja Y ovat satunnaismuuttujia. Satunnaismuuttujan X *reunajakauman todennäköisyysfunktio,* jota merkitään $f_X(x)$, on on X:n todennäköisyysfunktio, kun Y:tä ei oteta huomioon. Satunnaismuuttujan X *ehdollisen jakauman todennäköisyysfunktio,* jota merkitään $f_1(x \mid y)$, on on X:n todennäköisyysfunktio, kun Y:n arvo $Y = y$ on kiinnitetty.

Määritelmä 4.6 Olkoon diskreetin satunnaisvektorin (X, Y) todennäköisyysfunktio $f(x, y)$ ja arvoavaruus S. Silloin satunnaismuuttujat X ja Y ovat diskreettejä ja niiden reunajakaumien todennäköisyysfunktiot ovat

$$f_X(x) = \sum_{y \in S_Y} f(x, y), \quad x \in S_X; \qquad f_Y(y) = \sum_{x \in S_X} f(x, y), \quad y \in S_Y,$$

missä S_X on X:n ja S_Y Y:n arvoavaruus. Satunnaismuuttujat X *ja* Y *ovat riippumattomat* jos ja vain jos

$$P(X = x,\, Y = y) = P(X = x)\, P(Y = y) \tag{4.6.1}$$

kaikilla $x \in S_X$ ja $y \in S_Y$. Jos X ja Y eivät ole riippumattomia, niin ne ovat riippuvia. Todennäköisyysfunktion avulla ehto (4.6.1) voidaan lausua muodossa:

$$f(x, y) = f_X(x) f_Y(y) \qquad \text{kaikilla } x \in S_X \text{ ja } y \in S_Y.$$

Merkitään $A = \{X = x\}$ ja $B = \{Y = y\}$, missä $(x, y) \in S$. Silloin $A \cap B = \{X = x,\, Y = y\}$. Koska

$$P(A \cap X) = P(X = x, Y = y) = f(x, y)$$

ja

$$P(B) = P(Y = y) = f_Y(y) > 0 \qquad (\text{koska } y \in S_Y),$$

niin

$$P(A \mid B) = \frac{P(A \cap X)}{P(B)} = \frac{f(x, y)}{f_Y(y)}.$$

Siksi voimme määritellä ehdollisen todennäköisyysfunktion seuraavasti:

Määritelmä 4.7 Jos diskreetin satunnaisvektorin (X, Y) todennäköisyysfunktio on $f(x, y)$ ja arvoavaruus S, niin *X:n ehdollinen todennäköisyysfunktio ehdolla* $Y = y$ on

$$f_1(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad (x, y) \in S$$

ja *Y:n ehdollinen todennäköisyysfunktio ehdolla* $X = x$ on

$$f_2(y \mid x) = \frac{f(x, y)}{f_X(x)}, \qquad (x, y) \in S.$$

Esimerkki 4.13 Esimerkissä 4.12 käsitellyn satunnaisvektorin (X, Y) todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \qquad \text{kun } (x, y) \in S,$$

missä $S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$. Huomaa, että $f(x, y) = 0$, kun $(x, y) \notin S$. X:n ja Y:n reunajakaumien todennäköisyysfunktiot ovat

$$f_X(x) = \sum_{y=0}^{3} f(x, y) \qquad \text{ja} \qquad f_Y(y) = \sum_{x=0}^{3} f(x, y).$$

X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$ on

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad x = 0, 1, 2.$$

Saadaan siis kolme X:n ehdollista todennäköisyysfunktiota:

$$\begin{aligned} f_1(x \mid 0) &= \frac{x}{3}, && x = 0, 1, 2; \\ f_1(x \mid 1) &= \frac{x + 2}{9}, && x = 0, 1; \\ f_1(x \mid 2) &= 1, && x = 0. \end{aligned}$$

Vastaavalla tavalla saadaan kolme Y:n ehdollista todennäköisyysfunktiota ehdolla $X = x$. Satunnaismuuttujat X ja Y eivät ole riippumattomat, koska esimerkiksi

$$f(0, 1) = \frac{1}{6} \neq f_X(0) f_Y(1) = \frac{1}{2} \cdot \frac{5}{12} = \frac{5}{24}.$$

X:n ja Y:n riippuvuutta (vs. riippumattomomuutta) on luontevaa tarkastella ehdollisen jakauman avulla. Y:n ehdollinen todennäköisyysfunktio ehdolla $X = 1$ on

$$f_2(y \mid 1) = \frac{f(1, y)}{f_X(1)} = \frac{1 + 2y}{12} \Big/ \frac{1}{3} = \frac{1 + 2y}{4}, \qquad y = 0, 1.$$

Koska

$$f_2(y \mid 1) \neq f_Y(y),$$

voimme jälleen päätellä, että X ja Y eivät ole riippumattomat. □

Ehdollisen jakauman odotusarvoa kutsutaan jakauman *ehdolliseksi odotusarvoksi.* X:n ehdollinen odotusarvo ehdolla $Y = y$ on

$$E(X \mid Y = y) = \sum_{x \in S_X} x f(x \mid y) \tag{4.6.2}$$

ja Y:n ehdollinen odotusarvo ehdolla $X = x$ on

$$E(Y \mid X = x) = \sum_{y \in S_Y} y f(y \mid x). \tag{4.6.3}$$

Näitä odotusarvoja merkitään myös

$$E(X \mid y) = \mu_{X|y} \qquad \text{ja} \qquad E(Y \mid x) = \mu_{Y|x}.$$

Vastaavasti määritellään *X:n ehdollinen varianssi* ehdolla $Y = y$ ja Y:n ehdollinen varianssi ehdolla $X = x$. Y:n ehdollinen varianssi ehdolla $X = x$ on

$$\begin{aligned} \operatorname{Var}(Y \mid x) &= E\big[(Y - \mu_{Y|x})^2 \mid x\big] \\ &= \sum_{y \in S_Y} (y - \mu_{Y|x})^2 f(y \mid x), \end{aligned} \tag{4.6.4}$$

jota merkitään myös $\operatorname{Var}(Y \mid x) = \sigma^2_{Y|x}$. Samalla periaatteella voidaan ehdollisen jakauman avulla määritellä mikä tahansa jakauman ehdollinen tunnusluku, kuten esimerkiksi ehdolliset momentit tai ehdollinen mediaani.

Kun $E(Y \mid x)$ lasketaan eri x:n arvoilla, riippuu tulos yleensä x:n arvosta. Jos halutaan korostaa $E(Y \mid x)$:n riippuvuutta x:stä, merkitään esimerkiksi $E(Y \mid x) = g(x)$. Silloin ehdollinen odotusarvo määrittelee funktion $g(x)$.

Esimerkki 4.14 Esimerkissä 4.13 määritettiin ehdolliset todennäköisyysfunktiot $f_1(x \mid 0), f_1(x \mid 1)$ ja $f_1(x \mid 1)$, kun (X, Y):n yhteisjakauman todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \qquad (x, y) \in S.$$

Lasketaan nyt ehdolliset odotusarvot $E(X \mid y)$, $y = 0, 1, 2$:

$$\begin{aligned} E(X \mid 0) &= 0 + 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3}, \\ E(X \mid 1) &= 0 \cdot \frac{2}{5} + 1 \cdot \frac{3}{5} + 2 \cdot 0 = \frac{3}{5}, \\ E(X \mid 2) &= 0 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 = 0. \end{aligned}$$

Ehdolliset varianssit $\operatorname{Var}(X \mid y)$, $y = 0, 1, 2$, ovat vastaavasti:

$$\begin{aligned} \operatorname{Var}(X \mid 0) &= \left(1 - \frac{5}{3}\right)^2 \cdot \frac{1}{3} + \left(2 - \frac{5}{3}\right)^2 \cdot \frac{2}{3} = \frac{2}{9}, \\ \operatorname{Var}(X \mid 1) &= \left(0 - \frac{3}{5}\right)^2 \cdot \frac{2}{2} + \left(1 - \frac{3}{5}\right)^2 \cdot \frac{3}{5} + \left(2 - \frac{3}{5}\right)^2 \cdot 0 = \frac{6}{25}, \\ \operatorname{Var}(X \mid 2) &= (0 - 0)^2 \cdot 1 + (1 - 0)^2 \cdot 0 + (2 - 0)^2 \cdot 0 = 0. \end{aligned}$$

□

4.6.2 Satunnaismuuttujien funktion jakauma

Usein tarvitaan satunnaismuuttujien X ja Y jonkin funktion $h(X,Y)$ jakaumaa. Funktio $h(X,Y)$ voi olla esimerkiksi muotoa $X+Y$, XY, X^2+Y^2 jne. Jos h on jokin reaaliarvoinen funktio $h(x,y)$, voimme määritellä uuden satunnaismuuttujan $Z = h(X,Y)$. Olkoon satunnaisvektorin (X,Y) arvoalue S. Merkitään

$$A_z = \{\, (x,y) \in S \mid h(x,y) = z \,\}.$$

Silloin todennäköisyys $P(Z=z)$ voidaan laskea seuraavasti:

$$P(Z=z) = \sum_{(x,y)\in A_z} f(x,y).$$

Lasketaan siis yhteen todennäköisyydet $f(x,y)$ kaikkissa pisteissa (x,y), jotka toteuttavat ehdon $h(x,y)=z$. Tällä tavalla voidaan johtaa Z:n todennäköisyysfunktio.

Esimerkki 4.15 Oletetaan, että satunnaisvektorin (X,Y) (Esimerkki 4.12) todennäköisyysfunktio on

$$f(x,y) = \frac{x+2y}{12}, \qquad (x,y) \in S,$$

missä $S = \{(0,1),(0,2),(1,0),(1,1),(2,0)\}$. Määritellään kokonaislukuarvoinen satunnaismuuttuja

$$Z = h(X,Y) = XY.$$

Silloin Z:n arvojen joukko on $S_z = \{0,1\}$, ja vastaavasti

$$\begin{aligned} A_1 &= \{\, (x,y) \mid xy = 1 \,\} = \{(1,1)\}, \\ A_0 &= \{\, (x,y) \mid xy = 0 \,\} = \{(0,1),(0,2),(1,0),(2,0)\}. \end{aligned}$$

Nyt siis

$$\begin{aligned} P(Z=0) &= P\big((X,Y) \in A_0\big) = \frac{3}{4}, \\ P(Z=1) &= P\big((X,Y) \in A_1\big) = \frac{1}{4}, \end{aligned}$$

joten $Z \sim \mathrm{Ber}\big(\frac{1}{4}\big)$. □

4.6.3 Ehdollinen odotusarvo

Ehdollinen odotusarvo esiteltiin jo alaluvussa 3.3.2 (identiteetti 3.3.7). Alaluvussa 4.6.1 ehdollinen odotusarvo luonnehdittiin ehdollisen jakauman odotusarvona (ks. (4.6.2) ja (4.6.4)). Määritelmän mukaan

$$E(X \mid Y=y) = \sum_x x f(x \mid y) = \sum_x x \frac{f(x,y)}{f_Y(y)}.$$

Huomaa, että $E(X \mid Y = y)$ on y:n funktio, eli $E(X \mid Y = y) = h(y)$. Merkitään $E(X \mid Y) = h(Y)$, missä siis $E(X \mid Y)$ on satunnaismuuttuja, joka saa arvoja $E(X \mid Y = y),\ y \in S_Y$. Voimme nyt laskea satunnaismuuttujan $E(X \mid Y)$ odotusarvon, joka on $E(X)$. Monissa sovelluksissa odotusarvon laskeminen on luontevinta ehdollistamisen kautta.

Lause 4.12 *Olkoot X ja Y mitkä tahansa kaksi satunnaismuuttujaa, joilla on odotusarvo. Silloin $E[E(X \mid Y)] = E(X)$.*

Todistus. Odotusarvon määritelmän mukaan

$$\begin{aligned} E[E(X \mid Y)] &= \sum_y E(X \mid Y = y) f_Y(y) \\ &= \sum_y \sum_x x \frac{f(x,y)}{f_Y(y)} f_Y(y) \\ &= \sum_y \sum_x x f(x,y) \\ &= \sum_x x \sum_y f(x,y) \\ &= \sum_x x f_X(x) = E(X), \end{aligned}$$

missä $\sum_y f(x,y) = f_X(x)$ on X:n reunajakauma. □

Ehdollinen varianssi

$$\begin{aligned} Var(X \mid Y = y) &= E[(X - E(X \mid Y = y))^2 \mid Y = y] \\ &= E(X^2 \mid Y = y) - [E(X \mid Y = y)]^2 \end{aligned}$$

määriteltiin alaluvussa 4.6.1 (ks. identiteetti (4.6.4)). Ehdollinen varianssi $Var(X \mid Y)$ on satunnaismuuttuja, joka saa arvoja $Var(X \mid Y = y)$. Koska

$$Var(X \mid Y) = E(X^2 \mid Y) - [E(X \mid Y)]^2,$$

niin

$$\begin{aligned} E[Var(X \mid Y)] &= E[E(X^2 \mid Y)] - E[E(X \mid Y)]^2 \\ &= E(X^2) - E[E(X \mid Y)]^2. \end{aligned} \tag{4.6.5}$$

Lauseen 4.12 mukaan $E[E(X \mid Y)] = E(X)$ ja $E[E(X^2 \mid Y)] = E(X^2)$, joten

$$Var[E(X \mid Y)] = E[E(X \mid Y)]^2 - [E(X)]^2. \tag{4.6.6}$$

Laskemalla yhtälöt (4.6.5) ja (4.6.6) puolittain yhteen, saadaan seuraavassa lauseessa esitettävä tulos.

Lause 4.13 *Mille tahansa satunnaismuutujille X ja Y pitää paikkansa identiteetti*

$$Var(X) = E[Var(X \mid Y)] + Var[E(X \mid Y)],$$

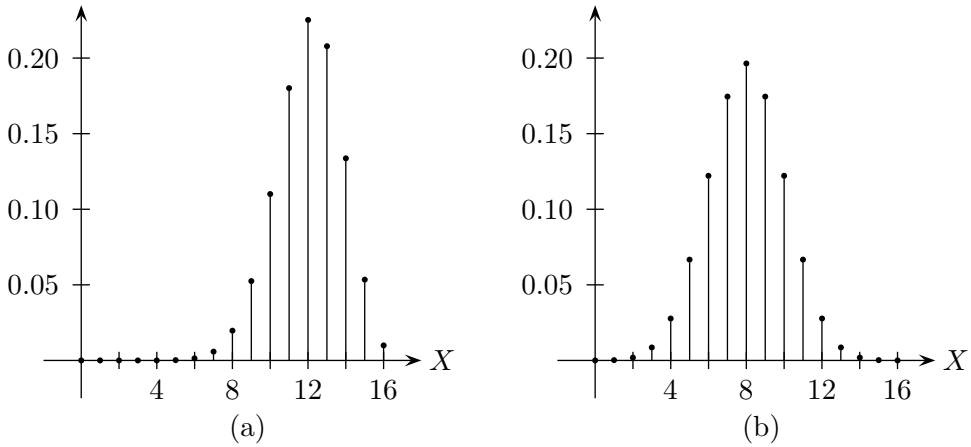
jos odotusarvot ovat olemassa.

4.6.4 Symmetrinen jakauma

Symmetriaan perustuvaa argumentointia voidaan usein hyödyntää todennäköisyyksien laskemisessa.

Symmetria pisteen suhteen. Jos $P(X = b+x) = P(X = b-x)$ kaikilla x, niin X:n jakauma on *symmetrinen pisteen b suhteen*. Satunnaismuuttuja X on symmetrinen b:n suhteen jos ja vain jos $X - b$ on symmetrinen origon suhteen. Silloin

$$P(X \leq b - x) = P(X \geq b + x).$$



Kuvio 4.3. Binomijakauman kuvaajat, kun (a) $X \sim \text{Bin}(16, 0.75)$ (b) $X \sim \text{Bin}(16, 0.50)$.

Esimerkiksi binomijakauma $\text{Bin}(16, 0.50)$ on symmetrinen pisteen 8 suhteen, mutta binomijakauma $\text{Bin}(16, 0.75)$ ei ole symmetrinen (Kuvio 4.3). Binomijakaumassa $\text{Bin}(16, 0.50)$ on

$$P(X = 8 + x) = P(X = 8 - x)$$

kaikilla x. Silloin jokaista $a \in \mathbb{R}$ kohti

$$P(X \leq 8 - a) = P(X \geq 8 + a).$$

Esimerkki 4.16 Hatussa on 3 korttia, jotka on numeroitu yhdestä kolmeen. Valitaan hatusta peräkkäin satunnaisesti palauttamatta 2 korttia. Olkoon X ensiksi valitun kortin numero ja Y toisen kortin numero. Selvästikin

$$P(X = i) = f_X(i) = \frac{1}{3}, \qquad i = 1, 2, 3.$$

On helppo havaita, että toisen valinnan tulos Y riippuu 1. valinnan tulokses-

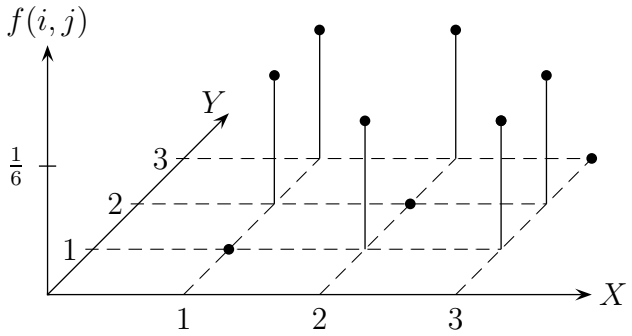
ta:

$$P(Y=1 \mid X=1)=0, \qquad P(Y=i \mid X=1)=\frac{1}{2}, \quad i=2,3;$$
$$P(Y=2 \mid X=2)=0, \qquad P(Y=i \mid X=2)=\frac{1}{2}, \quad i=1,3;$$
$$P(Y=3 \mid X=3)=0, \qquad P(Y=i \mid X=3)=\frac{1}{2}, \quad i=1,2.$$

Koska $P(X=i,Y=j)=P(X=i)\,P(Y=j \mid X=i)$, niin X:n ja Y:n yhteisjakauman todennäköisyysfunktio on

$$f(i,j)=f_X(i)f_2(j \mid i)=\begin{cases}\frac{1}{6}, & \text{kun } i\neq j;\\ 0, & \text{kun } i=j;\ 1\le i\le 3,\ 1\le j\le 3.\end{cases} \tag{4.6.7}$$

Satunnaisvektorin (X,Y) arvojoukko $S=\{(1,2),(1,3),(2,1),(2,3),(3,1),$



Kuvio 4.4. Satunnaisvektorin (X,Y) yhteisjakauman todennäköisyysfunktio $f(i,j)$, kun X on 1. valinta ja Y on 2. valinta palauttamatta joukosta $\{1,2,3\}$.

$(3,2)\}$, sillä $P(X=i,Y=j)>0$ kaikilla $(i,j)\in S$ ja $P(X=i,Y=j)=0$, jos $(i,j)\notin S$. □

Jakauma (4.6.7) on esimerkki symmetrisestä 2-ulotteisesta jakaumasta. Diskreetin satunnaisvektorin (X,Y) *jakauma on symmetrinen,* jos sen todennäköisyysfunktio $f(x,y)$ on *symetrinen funktio.* Se tarkoittaa sitä, että

$$f(x,y)=f(y,x) \qquad \text{kaikilla } (x,y)\in S,$$

missä S on (X,Y):n arvoalue.

4.6.5 Kaksiulotteinen Bernoullin jakauma

Bernoullin jakaumaa noudattava satunnaismuuttuja $X\sim \mathrm{Ber}(p)$ on eräs yksinkertaisimpia ajateltavissa olevia satunnaismuuttujia. Sen todennäköisyysfunktio on

$$f_X(x)=p^x(1-p)^{1-x}, \qquad x\in\{0,1\},$$

kun $0 \leq p \leq 1$. Bernoullin jakauma on binomijakauman erikoistapaus siten, että $X \sim \text{Bin}(1, p)$.

Kaksiulotteista Bernoullin jakaumaa noudattava satunnaismuuttuja (X, Y) voi saada arvot $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. Sen todennäköisyysfunktio on

$$f(x, y) = p_{xy}, \qquad x \in \{0, 1\} \quad \text{ja} \quad y \in \{0, 1\}, \tag{4.6.8}$$

missä $p_{00} + p_{01} + p_{10} + p_{11} = 1$. Todennäköisyysfunktio voidaan esittää myös muodossa

$$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy},$$

kun $x \in \{0, 1\}$ ja $y \in \{0, 1\}$; muualla $f(x, y) = 0$. Todennäköisyydet $P(X = x, Y = y) = p_{xy}$ on esitetty Taulukossa 4.1

Taulukko 4.1. Kaksiulotteisen Bernoullin jakauman todennäköisyysfunktio.

$f(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	p_{00}	p_{01}	$1-p_1$
$x=1$	p_{10}	p_{11}	p_1
$f_Y(y)$	$1-p_2$	p_2	1

”Reunatodennäköisyydet” määritellään $p_{00}+p_{01} = p_1$ ja $p_{00}+p_{10} = p_2$. On helppo havaita, että $X \sim \text{Ber}(p_1)$ ja $Y \sim \text{Ber}(p_2)$. Näiden reunajakaumien todennäköisyysfunktiot ovat siis

$$f_X(x) = p_1^x (1 - p_1)^{1-x}, \qquad x \in \{0, 1\}$$

ja

$$f_Y(y) = p_2^y (1 - p_2)^{1-y}, \qquad y \in \{0, 1\}.$$

Nyt esimerkiksi Y:n ehdollinen todennäköisyysfunktio ehdolla $X = 1$ on

$$f_2(y \mid 1) = \frac{p_{1y}}{p_1}, \qquad y \in \{0, 1\} \tag{4.6.9}$$

kun $p_1 > 0$. Satunnaismuuttujan Y ehdollinen jakauma ehdolla $X = 1$ on siis $\text{Ber}(p_{11}/p_1)$. Satunnaismuuttujat X ja Y ovat riippumattomat täsmälleen silloin, kun $P(X = x,\, Y = y) = P(X = x)\, P(Y = y)$ kaikilla $x \in \{0, 1\}$ ja $y \in \{0, 1\}$.

Koska $p_{00} + p_{01} + p_{10} + p_{11} = 1$, niin kaksiulotteinen Bernoullin jakauma voidaan luonnehtia kolmella parametrilla. Jakauman kolme ”luonnollista” parametria ovat

$$\begin{aligned} p_1 &= E(X) = P(X = 1), \\ p_2 &= E(Y) = P(Y = 1), \\ p_{11} &= E(XY) = P(X = 1, Y = 1). \end{aligned}$$

Kun (X, Y) noudattaa kaksiulotteista Bernoullin jakaumaa parametrein p_1, p_2 ja p_{11}, niin merkitään $(X, Y) \sim \text{Ber}(p_1, p_2, p_{11})$.

4.7 Satunnaismuuttujien funktion odotusarvo

Olkoot X ja Y diskreetit satunnaismuuttujat, joiden yhteinen todennäköisyysfunktio $f(x,y)$ on on määritelty arvoavaruudessa S. Olkoon $h(X,Y)$ satunnaismuuttujien X ja Y reaaliarvoinen funktio. Silloin

$$E[h(X,Y)] = \sum_{(x,y)\in S} h(x,y)f(x,y)$$

on satunnaismuuttujan $h(X,Y)$ odotusarvo, mikäli summa on olemassa.

Huomaa, että odotusarvon $E[h(X,Y)]$ olemassaolo tarkoittaa sitä, että summa

$$\sum_{(x,y)\in S} h(x,y)f(x,y)$$

suppenee itseisesti eli summa

$$\sum_{(x,y)\in S} |h(x,y)|f(x,y)$$

suppenee ja on äärellinen. Tästä seuraa, että $E[h(X,Y)]$ on olemassa. Funktio $V = h(X,Y)$ on satunnaismuuttuja, jonka todennäköisyysfunktio $g(v)$ on määritelty arvoavaruudessa $S_v = \{\, v \mid v = h(x,y),\ (x,y) \in S \,\}$. Silloin

$$E[h(X,Y)] = E(V) = \sum_{v\in S_v} vg(v).$$

4.7.1 Momentit

Monilla odotusarvoilla on omat nimensä, koska niillä on tärkeä rooli jakaumateoriassa. Olkoot X_1 ja X_2 diskreetit satunnaismuuttujat, joiden yhteisjakauman todennäköisyysfunktio $f(x_1,x_2)$ on on määritelty arvoavaruudessa S. Olkoon $h(X_1,X_2)$ satunnaismuuttujien X_1 ja X_2 reaaliarvoinen funktio. Määritellään esimerkiksi seuraavat odotusarvot:

1. Jos $h(X_1,X_2) = X_i$, niin

   $$E[h(X_1,X_2)] = E(X_i) = \mu_i$$

   on X_i:n *odotusarvo,* $i = 1,2$.

2. Jos $h(X_1,X_2) = (X_i - \mu_i)^2$, niin

   $$E[h(X_1,X_2)] = E[(X_i - \mu_i)^2] = \sigma_i^2$$

   on X_i:n *varianssi,* $i = 1,2$.

3. Jos $h(X_1, X_2) = (X_1 - \mu_1)(X_2 - \mu_2)$, niin

$$E[h(X_1, X_2)] = E[(X_1 - \mu_1)(X_2 - \mu_2)] = \sigma_{12}$$

on X_1:n ja X_2:n *kovarianssi.*

Odotusarvo μ_i ja varianssi σ_i^2 voidaan laskea joko yhteisjakauman todennäköisyysfunktion $f(x_1, x_2)$ tai reunajakauman todennäköisyysfunktion $f_i(x_i)$ avulla.

Vastaavalla tavalla voidaan määritellä kaikkien kertalukujen momentit: Olkoon r positiivinen kokonaisluku.

1. Jos $h(X_1, X_2) = X_i^r$, niin

$$E[h(X_1, X_2)] = E(X_i^r)$$

on X_i:n *r. momentti,* $i = 1, 2$.

2. Jos $h(X_1, X_2) = (X_i - \mu_i)^r$, niin

$$E[h(X_1, X_2)] = E[(X_i - \mu_i)^r]$$

on X_i:n *r. keskusmomentti,* $i = 1, 2$.

3. Jos $h(X_1, X_2) = X_1^r X_2^s$, niin

$$E[h(X_1, X_2)] = E(X_1^r X_2^s)$$

on X_1:n ja X_2:n kertalukua $r + s$ oleva *yhteismomentti.*

Esimerkiksi kovarianssin laskemisessa tarvitaan X_1:n ja X_2:n yhteismomentti $E(X_1 X_2)$.

4.7.2 Satunnaisvektorin momenttifunktio

Satunnaisvektorin (X, Y) yhteisjakauman momenttifunktio on

$$\begin{aligned} M(t, s) &= E[\exp(tX + sY)] \\ &= \sum_{x_i \in S_X} \sum_{y_j \in S_Y} \exp(tx_i + sy_j) f(x_i, y_j), \end{aligned}$$

mikäli odotusarvo on olemassa nollan ympäristössä. Silloin on siis olemassa sellainen positiiviluku $a > 0$, että odotusarvo $E[\exp(tX + sY)]$ on olemassa kaikilla $(t, s) \in \{ (t, s) \mid t^2 + s^2 < a \}$ jollain $a > 0$. Edellä on käytetty merkintää $\exp(tX + sY) = \mathrm{e}^{tX+sY}$.

4.8 Riippumattomat satunnaismuuttujat

Riippumattomuuden määritelmän mukaan tapahtumat $\{X = x\}$ ja $\{Y = y\}$ ovat riippumattomat jos ja vain jos $P(X = x,\, Y = y) = P(X = x)\, P(Y = y)$ eli

$$f(x, y) = f_X(x) f_Y(y), \tag{4.8.1}$$

missä $f(x, y)$ on X:n ja Y:n yhteisjakauman todennäköisyysfunktio, $f_X(x)$ on X:n ja $f_Y(y)$ Y:n todennäköisyysfunktio. Satunnaismuuttujat X *ja* Y *ovat riippumattomat* jos ja vain jos yhtäsuuruus (4.8.1) pitää paikkansa kaikilla $x \in S_X$ ja $y \in S_Y$, missä S_X on X:n ja S_Y on Y:n arvojoukko. Voidaan helposti osoittaa, että X ja Y ovat riippumattomat, jos ja vain jos

$$F(x, y) = F_X(x) F_Y(y) \tag{4.8.2}$$

kaikilla $x \in S_X$ ja $y \in S_Y$, missä $F_X(x)$ on X:n ja $F_Y(y)$ Y:n kertymäfunktio (reunakertymäfunktio).

Satunnaismuuttujien X ja Y ovat riippumattomuus voidaan luonnehtia myös ehdollisten jakaumien avulla. Jos Määritelmässä 4.7

$$f(y \mid x) = f_Y(y)$$

kaikilla $x \in S_X$ ja $y \in S_Y$, kun $f_X(x) \neq 0$, niin X ja Y ovat riippumattomat. Tämä tarkoittaa sitä, että tieto X:n arvosta ei vaikuta Y:n todennäköisyyteen. Vastaavasti pitää paikkansa, että X ja Y ovat riippumattomat jos ja vain jos X:n ehdollinen todennäköisyysfunktio ehdolla $Y = y$ ei riipu y:stä.

Olkoot $Y_1,\, Y_2,\, \ldots,\, Y_n$ jossain otosavaruudessa määritellyt diskreetit satunnaismuuttujat. Muuttujien $Y_1,\, Y_2,\, \ldots,\, Y_n$ *yhteisjakauman todennäköisyysfunktio* on

$$f(y_1, y_2, \ldots, y_n) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n).$$

Muuttujat $Y_1,\, Y_2,\, \ldots,\, Y_n$ ovat riippumattomat, jos

$$f(y_1, y_2, \ldots, y_n) = f_1(y_1) f_2(y_2) \cdots f_n(y_n) \tag{4.8.3}$$

kaikilla $y_i \in S_i$, $i = 1, 2, \ldots, n$, missä $f_i(y_i)$ on Y_i:n todennäköisyysfunktio ja S_i on Y_i:n arvoavaruus.

4.8.1 Riippumattomat kokeet

Olkoot $\mathcal{E}_1$ ja $\mathcal{E}_2$ riippumattomat satunnaiskokeet. Oletetaan, että satunnaismuuttujan X arvo määräytyy vain satunnaiskokeen $\mathcal{E}_1$ tuloksen ja Y:n arvo vain satunnaiskokeen $\mathcal{E}_2$ tuloksen perusteella. Silloin tapahtumat $\{X = x\}$ ja $\{Y = y\}$ määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat (katso alaluku 3.6 ja määritelmä (3.6.1)). Siksi tapahtumat $\{X = x\}$ ja $\{Y = y\}$

ovat riippumattomat. Koska riippumattomuus pätee kaikilla mahdollisilla x:n ja y:n arvoilla, niin X ja Y ovat riippumattomat. *Jos satunnaismuuttujien arvot määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat, niin satunnaismuuttujat ovat riippumattomat.*

Tehdään esimerkiksi Bernoullin toistokoe, jossa on $r + s$ toistoa ja onnistumistodennäköisyys p. Olkoon X onnistumisten lukumäärä r:ssä ensimmäisessä kokeessa ja Y onnistumisten lukumäärä s:ssä viimeisessä kokeessa. Koska X ja Y riippuvat eri kokeista, jotka ovat riippumattomat, niin X ja Y ovat riippumattomat. Silloin (4.8.1):n mukaan

$$f(x,y) = f_X(x)f_Y(y) = \binom{r}{x}\binom{s}{y}p^{x+y}(1-p)^{r+s-x-y},$$

missä $x = 0, 1, \ldots, r$ ja $y = 0, 1, \ldots, s$.

4.8.2 Samoin jakautuneet riippumattomat (SJR) satunnaismuuttujat

Riippumattomia satunnaismuuttujia Y_1, Y_2, $\ldots$, Y_n, joista jokainen noudattaa samaa jakaumaa, sanotaan samoin jakautuneiksi riippumattomiksi (sjr) satunnaismuuttujiksi. Silloin puhutaan usein lyhyesti sjr satunnaismuuttujista. Vastaava englanninkielinen termi on iid (independent, identically distributed).

Jos esimerkiksi $Y_i \sim \text{Poi}(\lambda)$, $i = 1, 2, \ldots, n$, niin

$$f_1(y) = f_2(y) = \cdots = f_n(y) = \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!}.$$

Silloin (4.8.3):n nojalla

$$\begin{aligned} f(y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{\lambda^{y_i}\mathrm{e}^{-\lambda}}{y_i!} \\ &= \frac{1}{y_1!\, y_2! \ldots y_n!}\lambda^y \mathrm{e}^{-n\lambda}, \end{aligned}$$

missä $y = \sum_{i=1}^{n} y_i$.

4.8.3 Riippumattomien satunnaismuuttujien funktio

Olkoot X ja Y riippumattomat satunnaismuuttujat. Silloin

$$f(x,y) = f_X(x)f_Y(y) \qquad \text{kaikilla } x, y \in S,$$

missä S on satunnaisvektorin (X, Y) arvoavaruus. Määritellään nyt satunnaismuuttujat $U = g(X)$ ja $V = h(Y)$, missä $g(\cdot)$ riippuu vain X:stä ja $h(\cdot)$ vain Y:stä. Silloin Lauseen 3.6 mukaan U ja V ovat riippumattomat.

Väite todistettiin tarkastelemalla U:n ja V:n mielivaltaisten arvojen u ja v todennäköisyyttä. Olkoon $A_u = \{\, x \in S_X \mid g(x) = u \,\}$ ja $B_v = \{\, y \in S_Y \mid g(y) = v \,\}$. Koska kaikilla U:n ja V:n arvoilla u ja v

$$\begin{aligned} (4.8.4) \qquad & P(U = u, V = v) \\ & = P(X \in A_u, Y \in B_v) \\ & = P(X \in A_u)\, P(Y \in B_v) \qquad (X \text{ ja } Y \text{ riippumattomat}) \\ & = P(U = u)\, P(V = v), \end{aligned}$$

niin U ja V ovat riippumattomat. Kun sijoitetaan identiteettiin (4.8.4)

$$P(U = u, V = v) = \sum_{x \in A_u} \sum_{y \in B_v} f(x, y),$$

$$P(U = u) = \sum_{x \in A_u} f_X(x) \qquad \text{ja} \qquad P(V = v) = \sum_{y \in B_v} f_Y(y),$$

niin saadaan

$$\sum_{x \in A_u} \sum_{y \in B_v} f(x, y) = \Big(\sum_{x \in A_u} f_X(x) \Big) \Big(\sum_{y \in B_v} f_Y(y) \Big).$$

4.9 Multinomijakauma ja moniulotteinen hypergeometrinen jakauma

Binomijakauma ja multinomijakauma ovat keskeisen tärkeitä tilastollisissa sovelluksissa, koska niitä tarvitaan esimerkiksi riippumattomien koetoistojen tulosten frekvenssijakaumien käsittelyssä. Alaluvussa 3.6 esitettiin, miten binomijakauma saadaan Bernoullin kokeiden avulla. Kun toistetaan kokeita, joissa on useampia kuin kaksi tulosvaihtoehtoa, tulosten frekvenssijakauma voidaan kuvata multinomijakauman avulla. Laajennetaan ensin binomijakauma *trinomijakaumaksi.*

Tarkastellaan koetta, jossa on kolme toisensa poissulkevaa tulosvaihtoehtoa. Esimerkiksi tuotantoprosessissa syntyvä tuote luokitellaan yhteen ja vain yhteen seuraavista kategorioista: ensiluokkainen (1), sekunda (2) tai viallinen (3). Olkoot ensiluokkaisen, sekundan ja viallisen todennäköisyydet vastaavasti p_1, p_2 ja $p_3 = 1 - p_1 - p_2$. Valmistetaan n tuotetta. Olkoon X_1 = ensiluokkaisten lukumäärä, X_2 = sekundatuotteiden lukumäärä ja $X_3 = n - X_1 - X_2$ = viallisten lukumäärä tuote-erässä. Jos x_1 ja x_2 ovat sellaiset epänegatiiviset kokonaisluvut, että $x_1 + x_2 \le n$, niin todennäköisyys saada x_1 ensiluokkaista, x_2 sekundaa ja $n - x_1 - x_2$ viallista jossain annetussa järjestyksessä on

$$p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x2}.$$

Sellaisia n:n tuotteen järjestyksiä, joissa on x_1 ensiluokkaista, x_2 sekundaa ja $n - x_1 - x_2$ viallista, on

$$\binom{n}{x_1} \binom{n - x_1}{x_2} = \frac{n!}{x_1!\, x_2!\, (n - x_1 - x_2)!}$$

kappaletta. Siksi *trinomijakauman todennäköisyysfunktio* on

$$f(x_1, x_2) = \frac{n!}{x_1!\, x_2!\, (n - x_1 - x_2)!}\, p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}, \tag{4.9.1}$$

missä $f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$. Kun (X_1, X_2) noudattaa trinomijakaumaa parametrein n, p_1 ja p_2, merkitään

$$(X_1, X_2) \sim \mathrm{Tri}(n, p_1, p_2).$$

On helppo todeta, että $X_1 \sim \mathrm{Bin}(n, p_1)$ (X_1:n reunajakauma) ja $X_2 \sim \mathrm{Bin}(n, p_2)$.

Multinomijakauma voidaan johtaa samalla periaatteella kuin trinomijakauma. Toistetaan n kertaa koe, jossa on k toisensa poissulkevaa tulosvaihtoehtoa. Merkitään tulosvaihtoehtoja $1, 2, \ldots, k$ ja olkoon p_i = tuloksen i todennäköisyys ja X_i on tuloksen i lukumäärä n:n kokeen sarjassa. Silloin k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ arvoalue on

$$S = \{\, (x_1, x_2, \ldots, x_k) \mid 0 \le x_i \le n,\ x_1 + x_2 + \cdots + x_k = n \,\}.$$

X_i:t ovat siis epänegatiivisia kokonaislukuarvoisia satunnaismuuttujia, joiden summa on n. Satunnaisvektori $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ noudattaa *k-ulotteista multinomijakaumaa* parametrein n ja $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, jota merkitään $\mathrm{Mult}(n, \boldsymbol{p})$. Multinomijakauman todennäköisyysfunktio on

$$f(x_1, x_2, \ldots, x_k) = \binom{n}{x_1\ x_2\ \ldots\ x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \tag{4.9.2}$$

missä $p_1 + p_2 + \cdots + p_k = 1$ ja $\binom{n}{x_1\ x_2\ \ldots\ x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!}$. Multinomilauseen 2.9 nojalla voidaan helposti osoittaa, että todennäköisyyksien (4.9.2) summa yli arvoalueen S on 1, joten kyseinen funktio on todellakin todennäköisyysfunktio. Multinomijakaumassa jokaisen X_i:n *reunajakauma on binomijakauma,* eli $X_i \sim \mathrm{Bin}(n, p_i)$, $i = 1, 2, \ldots, k$. Esitetään nyt multinomijakaumaa koskevat perustulokset lauseen muodossa.

Lause 4.14 *1. Funktio (4.9.2) on multinomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla n ja kaikilla sellaisilla p_1, $\ldots$, p_k, että $0 \le p_i \le 1$ ja $p_1 + \cdots + p_k = 1$.*

2. Jos $\boldsymbol{X} \sim \mathrm{Mult}(n, \boldsymbol{p})$, niin

$$X_i \sim \mathrm{Bin}(n, p_i) \qquad \textit{ja} \qquad (X_i, X_j) \sim \mathrm{Tri}(n, p_i, p_j),$$

3.

$$E(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i), \qquad \mathrm{Cov}(X_i, X_j) = -np_ip_j,$$
$$M(t) = E[\exp(t_1X_1 + \cdots + t_kX_k)] = \left(p_1\mathrm{e}^{t_1} + \cdots + p_k\mathrm{e}^{t_k}\right)^n.$$

Multinomijakauma liittyy otantaan palauttaen. Olkoon uurnassa erivärisiä palloja yhteensä N, värien lukumäärä on k ja väriä i olevia palloja on N_i kappaletta ($i = 1, 2, \ldots, k$) ja $N_1 + N_2 + \cdots + N_k = N$. Otannassa palauttaen todennäköisyys p_i saada väri i on N_i/N jokaisessa nostossa. Valitaan uurnasta n palloa palauttaen ja olkoon X_i väriä i olevien pallojen lukumäärä otoksessa. Silloin satunnaismuuttujien X_1, X_2, …, X_k yhteisjakauma on multinomijakauma.

Otannassa palauttamatta uurnan sisältö muuttuu ja siten myös valintatodennäköisyydet muuttuvat valintaprosessin aikana. Satunnaismuuttujien X_1, X_2, …, X_k yhteisjakauman johtamiseksi meidän on yleistettävä alaluvussa 2.6.1 esitetty hypergeometrinen jakauma.

Olkoon x_i väriä i ($1 \leq i \leq k$) olevien pallojen lukumäärä otannassa palauttamatta. Millä todennäköisyydellä saadaan otos, jossa eri väriä olevien lukumäärät ovat $(x_1, x_2, \ldots, x_k)$? Koska uurnassa on N_i kappaletta väriä i, niin $0 \leq x_i \leq N_i$. Otoskoko on n ja $n = x_1 + x_2 + \cdots + x_k$. Nyt väriä 1 olevat x_1 palloa voidaan valita $\binom{N_1}{x_1}$ tavalla, väriä 2 olevat $\binom{N_2}{x_2}$ tavalla ja lopulta väriä k olevat pallot $\binom{N_k}{x_k}$ tavalla. Suotuisten otosten lukumäärä on tuloperiaatteen nojalla

$$\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}.$$

Koska kaikkien mahdollisten n:n kokoisten otosten lukumäärä on $\binom{N}{n}$, niin todennäköisyys saada lukumäärät $(x_1, x_2, \ldots, x_k)$ erivärisiä palloja on

$$f(x_1, x_2, \ldots, x_k) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}, \tag{4.9.3}$$

missä $x_1 + x_2 + \cdots + x_k = n$. Tämä on *moniulotteisen hypergeometrisen jakauman* todennäköisyysfunktio.

Diskreetit jakaumat: Yhteenveto

Bernoulli
Ber(p)

$$f(x) = p^x(1-p)^x, \qquad x = 0, 1$$
$$E(X) = p, \qquad \mathrm{Var}(X) = p(1-p)$$
$$M(t) = 1 - p + p\mathrm{e}^t$$

Binomi
Bin(n, p)

$$f(x) = \binom{n}{x}p^x(1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$
$$E(X) = np, \qquad \mathrm{Var}(X) = np(1-p)$$
$$M(t) = (1 - p + p\mathrm{e}^t)^n$$

Negatiivinen binomi
NBin(r,p)

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \ldots$$
$$E(X) = \frac{r}{p}, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$
$$M(t) = \left[\frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}\right]^r, \qquad t < -\log(1-p)$$

Geometrinen
Geo(p)

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots$$
$$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$
$$M(t) = \frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}, \qquad t < -\log(1-p)$$

Hypergeometrinen
HGeo(n,N,p)

$$f(x) = \frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}, \qquad \begin{array}{l} X \le pN \quad \text{ja} \\ n - X \le N - Np \end{array}$$
$$E(X) = np, \qquad \mathrm{Var}(X) = \frac{N-n}{N-1} np(1-p),$$

Negatiivinen hypergeometrinen
NHGeo(r,N,p)

$$f(x) = \binom{x-1}{r-1} \frac{\binom{N-x}{Np-r}}{\binom{N}{Np}}, \qquad x = r, r+1, \ldots, N$$
$$E(X) = r \cdot \frac{N+1}{Np+1}$$
$$\mathrm{Var}(X) = \frac{r(1-p)N(N+1)(Np+1-r)}{(Np+1)^2(Np+2)}$$

Poisson
Poi(λ)

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots$$
$$E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda$$
$$M(t) = \exp(\lambda \mathrm{e}^t - \lambda), \qquad -\infty < t < \infty$$

Poissonin prosessi. Laskuriprosessi $\{\, N(t),\ t \ge 0 \,\}$, jonka intensiteetti λ.

1. $N(0) = 0$.
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella t:n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on λt:

$$P[N(t+s) - N(s) = x] = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^x}{x!}, \qquad x = 0, 1, \ldots$$

kaikilla $s, t \ge 0$.

Kaksiulotteiset jakaumat

Satunnaisvektorin (X, Y) yhteisjakauma.

Todennäköisyysfunktio $f(x, y)$ toteuttaa ehdot

1. $0 \le f(x, y) \le 1, \quad$ kaikilla $(x, y) \in S$ ja
2. $\sum\limits_{(x,y)\in S} f(x, y) = 1,$

missä S on satunnaisvektorin (X, Y) arvojoukko.

Reunajakaumien todennäköisyysfunktiot:

$$f_X(x) = \sum_{y\in S_Y} f(x, y), \quad x \in S_X; \qquad f_Y(y) = \sum_{x\in S_X} f(x, y), \quad y \in S_Y.$$

Ehdolliset todennäköisyysfunktiot:

$$f_1(x \mid y) = \frac{f(x, y)}{f_Y(y)} \qquad \text{ja} \qquad f_2(y \mid x) = \frac{f(x, y)}{f_X(x)}, \qquad (x, y) \in S.$$

Bernoulli
Ber2(p_1, p_2, p_{11})

$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy}, \qquad$ missä
$p_{00} + p_{01} + p_{10} + p_{11} = 1, \quad x \in \{0, 1\} \quad$ ja $\quad y \in \{0, 1\}$
$E(X) = p_1 = P(X = 1), \qquad E(Y) = p_2 = P(Y = 1)$
$E(XY) = p_{11} = P(X = 1,\, Y = 1)$

Multinomi
Mult$(n, \boldsymbol{p})$

$\boldsymbol{p} = (p_1, p_2, \ldots, p_k), \qquad$ missä
$0 \le p_i \le 1 \quad$ ja $\quad p_1 + p_2 + \cdots + p_k = 1$
$E(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i)$
$\mathrm{Cov}(X_i, X_j) = -np_ip_j$
$X_i \sim \mathrm{Bin}(n, p_i), \quad (X_i, X_j) \sim \mathrm{Mult}(n; p_i, p_j, 1 - p_i - p_j)$
$M(t) = (p_1 \mathrm{e}^{t_1} + \cdots + p_k \mathrm{e}^{t_k})^n$

Hypergeometrinen

$f(x_1, x_2, \ldots, x_k) = \dfrac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}, \qquad$ missä
$N_1 + N_2 + \cdots + N_k = N \quad$ ja $\quad x_1 + x_2 + \cdots + x_k = n$

Harjoituksia

1. Olkoon X satunnaismuuttuja, jonka arvojoukko on $S_X = \{x_1, x_2\}$ ja todennäköisyysfunktio $P(X = x_1) = p$, $P(X = x_2) = 1 - p$.

(a) Laske $E(X^r)$, $r = 1, 2$ ja

(b) $\mathrm{Var}(X)$.

(c) Määritä X:n momenttifunktio.

2. Olkoon $X \sim \text{Ber}(p)$ ja Y sellainen satunnaismuuttuja, jonka arvojoukko on $S_Y = \{y_1, y_2\}$ ja todennäköisyysfunktio $P(Y = y_1) = p$, $P(Y = y_2) = 1 - p$. Lausu Y satunnaismuuttujan X avulla.

3. Heitetään lanttia n kertaa (n riippumatonta Bernoullin koetta). Olkoon kruunun (R) todennäköisyys p ja X toistosten RR lukumäärä heittosarjassa.

(a) Mitä on $E(X)$? Mikä on $E(X)$:n arvo, kun kun $n = 200$?

(b) Laske $\text{Var}(X)$.

(c) Mitä on toistosten RRRR lukumäärän odotusarvo?

(Vihje: Katso Esimerkki 4.2.)

4. Jos X noudattaa binomijakaumaa, jonka odotusarvo on 6 ja varianssi 2.4, niin mitä on $P(X = 5)$?

5. Hatussa on N yhdestä lähtien juoksevasti numeroitua arpalippua. Valitaan hatusta n:n arvan satunnaisotos palauttamatta (ks. Esimerkki 4.2). Olkoon X suurin valittujen arpalippujen järjestysnumeroista.

(a) Piirrä X:n todennäköisyys- ja kertymäfunktion kuvaajat, kun $N = 100$ ja $n = 10$.

(b) Piirrä X:n odotusarvon kuvaaja n:n funktiona, kun $N = 100$.

6. Valitaan satunnaisesti ja toisistaan riippumatta 2000 pistettä yksikköneliöstä $\{ (x, y) \mid 0 \leq x \leq 1,\ 0 \leq y \leq 1 \}$. Olkoon Z yksikköympyrään $\{ (x, y) \mid x^2 + y^2 \leq 1 \}$ osuvien pisteiden lukumäärä.

(a) Mitä jakaumaa Z noudattaa?

(b) Laske Z:n odotusarvo ja hajonta.

(c) Satunnaismuuttujan $\frac{Z}{500}$ odotusarvo?

(d) Generoi 2000 satunnaislukuparia. Määritä Z:n arvo ja laske sen avulla π:n likiarvo.

7. Erääseen 90:n virheettömän kännykän tuote-erään oli sekaantunut 10 viallista. Valitaan tästä 100:n kännykän joukosta 30 kännykän otos palauttamatta. Olkoon X viallisten lukumäärä otoksessa.

(a) Määritä X:n todennäköisyysfunktio.

(b) Laske $P(X = 10)$.

(c) Valitaan kännyköitä testaukseen satunnaisotannalla yksitellen palauttamatta, kunnes kaikki vialliset on löydetty. Olkoon Y tarvittavien testien lukumäärä. Laske $P(Y \geq 20)$, eli todennäköisyys, että tarvitaan ainakin 20 testiä.

8. Heitetään harhatonta lanttia, kunnes havaitaan toistos RR (kaksi kruunua peräkkäin). Olkoon X tarvittavien heittojen lukumäärä. Olkoon f_n n. Fibonaccin luku, joka määritellään siten, että $f_1 = f_2 = 1$ ja $f_n = f_{n-1} + f_{n-2}$, $n = 3, 4, \ldots$.

(a) Osoita, että X:n todennäköisyysfunktio on

$$f(x) = \frac{f_{x-1}}{2^x}, \qquad x = 2, 3, 4, \ldots$$

(b) Osoita tuloksen

$$f_x = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^x - \left(\frac{1-\sqrt{5}}{2}\right)^x\right]$$

avulla, että $\sum_{x=2}^{\infty} = 1$.

(c) Osoita, että $E(X) = 6$.

(d) Osoita, että $E[X(X-1)] = 52$ ja $\mathrm{Var}(X) = 22$.

(e) Simuloi X:n arvoja ja tarkastele, vastaavatko simuloinnin tulokset teoreettisia tuloksia.

9. Eräässä vaalissa 4000:sta äänestäjästä 100 kannatti ehdokasta A. Jos valitaan 50 alkion otos äänestäjistä esitutkimukseen, niin millä todennäköisyydellä haastatelluista korkeintaan 5 kannattaa A:ta?

10. Yritykseen tulee lähetys, joka sisältää 1000 varaosaa. Tarkistussuunnitelman mukaan $n = 100$ satunnaisesti valittua (palauttamatta) varaosaa on tarkistettava. Tuote-erä hyväksytään, jos tarkistuksessa ei löydy kahta viallista enempää. Mikä on todennäköisyys, että tuote-erä hyväksytään? Laske todennäköisyys

(a) hypergeometrisen jakauman avulla.

(b) Laske sitten sama todennäköisyys käyttäen hypergeometrisen jakauman likiarvona binomijakaumaa

(c) ja Poissonin jakaumaa.

11. Oletetaan, että $X \sim \mathrm{Poi}(\lambda)$. Osoita, että $E(X) = \lambda$ ja $\mathrm{Var}(X) = \lambda$.

12. Leipomossa valmistetaan suuri taikina, josta tehdään rusinaleivoksia. Leipoyrittäjä haluaa, että 95 % leivoksista sisältää ainakin 2 rusinaa. Kuinka monta rusinaa leivosta kohti hänen pitää sekoittaa taikinaan?

13. Laboratoriohiiriin ruiskutetaan kahta eri liuosta. Ensimmäisessä liuoksessa on keskimäärin c kappaletta C-tyypin organismeja millitrassa ja toisessa liuoksessa keskimäärin d kappaletta D-tyypin organismeja millitrassa. Organismit ovat jakautuneet nesteeseen täysin satunnaisesti. Jokaiseen hiireen ruiskutetaan kumpaakin liuosta yksi millitra. Hiiri säilyy hengissä jos ja vain jos kummassakaan ruiskeessa ei ole yhtään organismia.

(a) Millä todennäköisyydellä hiiri jää eloon?

(b) Millä todennäköisyydellä kuolleista hiiristä löytyy molempia organismeja?

(Vihje: Käytä Poissonin jakaumaa.)

14. Tehtaalla sattuu keskimäärin 1.5 onnettomuutta kuukaudessa. Määritä seuraavien tapahtumien todennäköisyydet:

(a) Ei onnettomuuksia tammikuussa,

(b) yhteensä neljä onnettomuutta helmikuussa ja maaliskuussa,

(c) ainakin yksi onnettomuus vuoden jokaisena kuukautena.

(Vihje: Käytä Poissonin jakaumaa.)

15. Olkoot X ja Y toisistaan riippumattomat Poissonin jakaumaa noudattavat satunnaismuuttujat. Olkoon $E(X) = 1$ ja $E(Y) = 2$.

(a) Laske todennäköisyys $P(X + Y) = 5$.

(b) Millä kokonaislukuarvolla n todennäköisyys $P(X + Y) = n$ saavuttaa maksiminsa?

(c) Lausu todennäköisyys $P(X + Y) = 5$ satunnaismuuttujien X ja Y todennäköisyysfunktioiden avulla.

16. Kirjassa on 200 sivua. Painovirheiden lukumäärä jokaisella sivulla noudattaa Poissonin jakaumaa, jonka keskiarvo on 0.01. Painovirheiden lukumäärät eri sivuilla ovat toisistaan riippumattomat.

(a) Mikä on virheettömien sivujen lukumäärän odotusarvo ja hajonta?

(b) Kirjan oikolukija havaitsee minkä tahansa annetun virheen todennäköisyydellä 0.9. Mikä on oikolukijan havaitsemien virheellisten sivujen lukumäärän odotusarvo?

17. Määritellään X:n ja Y: yhteisjakauman todennäköisyysfunktio seuraavasti:

$$f(x, y) = \frac{x + y}{32}, \qquad \text{kun} \quad x = 1, 2; \quad y = 1, 2, 3, 4.$$

Määritä

(a) X:n reunajakauman todennäköisyysfunktio ja

(b) Y:n reunajakauman todennäköisyysfunktio.

(c) Laske $P(X > Y)$,

(d) $P(Y = 2X)$ ja

(e) $P(X + Y = 3)$.

18. X:n ja Y:n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.

(a) Laske odotusarvot μ_X ja μ_Y,

(b) varianssit σ_X^2 ja σ_Y^2 sekä

(c) korrelaatiokerroin ρ.

(d) Ovatko X ja Y riippumattomat?

19. X:n ja Y:n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.

(a) Määritä X:n ehdolliset todennäköisyysfunktiot $f_1(x \mid y)$ ehdolla $y = 1, 2, 3$ ja $y = 4$.

(b) Määritä Y:n ehdolliset todennäköisyysfunktiot $f_2(y \mid x)$ ehdolla $x = 1$ ja $x = 2$.

(c) Laske $P(1 \leq Y \leq 3 \mid X = 1)$, $P(Y \leq 2 \mid X = 2)$, ja $P(X = 2 \mid Y = 3)$.

(d) Laske $E(Y \mid X = 1)$ ja $\mathrm{Var}(Y \mid X = 1)$.

20. Testataan kymmenen suojakypärän iskukestävyys. Kypärät jaetaan kahteen viiden ryhmään. Ensimmäisen ryhmän kypärille annetaan isku, joka särkee kypärän todennäköisyydellä 0.1. Toisen ryhmän kypäriä isketään voimalla, joka särkee kypärän todennäköisyydellä 0.3. Millä todennäköisyydellä ensimmäisen ryhmän kypäriä rikkoontuu enemmän kuin toisen ryhmän kypäriä?

21. Oletetaan, että satunnaismuuttujat X_1, X_2, X_3 noudattavat multinomijakaumaa $\mathrm{Mult}(5; 0.1, 0.3, 0.6)$ [Toisin sanoen $(X_1, X_2) \sim \mathrm{Tri}(5; 0.1, 0.3)$].

(a) Määritä X_1:n reunajakauma ja X_2:n

(b) ehdollinen todennäköisyysfunktio $f_2(x_2 \mid x_1 = 1)$.

22. Nostetaan tavallisesta korttipakasta (52 korttia) satunnaisesti palauttamatta 13 korttia. Olkoon X_1 patojen lukumäärä, X_2 herttojen lukumäärä ja $13 - X_1 - X_2$ ruutujen ja ristien lukumäärä otoksessa.

(a) Määritä X_1:n ja X_2:n yhteisjakuman todennäköisyysfunktio. Mikä on satunnaisvektorin (X_1, X_2) arvoalue?

(b) X_1:n todennäköisyysfunktio ja arvoalue?

23. Oletetaan, että (X, Y) noudattaa trinomijakaumaa $\mathrm{Tri}\big(3, \frac{1}{6}, \frac{1}{2}\big)$.

(a) Laske odotusarvot μ_X ja μ_Y,

(b) varianssit σ_X^2 ja σ_Y^2 sekä

(c) kovarianssi $\mathrm{Cov}(X, Y)$ ja

(d) korrelaatiokerroin ρ.

24. Satunnaismuuttujat $X \geq 0$ ja $Y \geq 0$ ovat riippumattomat ja saavat vain kokonaislukuarvoja. Osoita, että

(a) $P(X+Y=n) = \sum_{k=0}^{n} P(X=0)\,P(Y=n-k)$.

(b) Heitetään 4:ää harhatonta noppaa ja lasketaan silmälukujen summa. Mikä on todennäköisyys, että summa on 8? (Vihje: Olkoon X kahden nopan silmälukujen summa ja Y kahden muun nopan silmälukujen summa.)

Luku 5

Jatkuvat jakaumat

Sellaiset suureet kuten esimerkiksi aika, lämpötila, pituus ja paino ajatellaan tavallisesti jatkuviksi muuttujiksi, ts. muuttujiksi, jotka voivat saada mitä tahansa reaaliarvoja annetulla välillä. Esimerkiksi henkilön ikä on *jatkuva satunnaismuuttuja,* joka voi saada positiivisia reaalilukuarvoja. Diskreetin satunnaismuuttujan arvoavaruus on äärellinen tai numeroituva, mutta jatkuvan satunnaismuuttujan arvoavaruus on ylinumeroituva.

5.1 Jatkuvat satunnaismuuttujat

Jokaiseen satunnaismuuttujaan liittyy kertymäfunktio. Satunnaismuuttujan X kertymäfunktio määriteltiin alaluvussa 2.5.2 (Määritelmä 2.4) funktiona

$$F(x) = P(X \leq x), \qquad x \in \mathbb{R}.$$

Diskreetin satunnaismuuttujan kertymäfunktio on porrasfunktio, joka voidaan lausua hyppyfunktioiden summana (4.1.1) [ks. alaluku 4.1].

Lauseen 2.10 mukaan funktio $F(x)$ on kertymäfunktio jos ja vain jos seuraavat kolme ehtoa toteutuvat:

1. $\lim_{x \to -\infty} F(x) = 0$ ja $\lim_{x \to \infty} F(x) = 1.$
2. $F(x)$ on kasvava (ei-vähenevä) funktio.
3. $F(x)$ on oikealta jatkuva eli $\lim_{x \to x_0+} F(x) = F(x_0)$ kaikilla $x_0 \in \mathbb{R}$.

Jos meillä on jokin satunnaismuuttuja X, niin ominaisuudet 1.–3. voidaan todeta todennäköisyysfunktion $P(X \leq x)$ ominaisuuksien avulla. Jos jokin funktio $F(x)$ toteuttaa ehdot 1.–3., ei ole aivan helppoa todistaa, että $F(x)$ on todella jonkin satunnaismuuttujan kertymäfunktio. Todistus löytyy vaativista todennäköisyyslaskennan oppikirjoista.

Esimerkki 5.1 Funktio

$$F(x) = \frac{1}{1 + e^{-x}} \tag{5.1.1}$$

on esimerkki jatkuvasta kertymäfunktiosta, joka siis toteuttaa Lauseen 2.10 ehdot 1.–3. Koska

$$\lim_{x\to-\infty} e^{-x} = \infty, \quad \text{niin} \quad \lim_{x\to-\infty} F(x) = 0$$

ja

$$\lim_{x\to\infty} F(x) = 1, \quad \text{koska} \quad \lim_{x\to\infty} e^{-x} = 0.$$

Funktio $F(x)$ on kasvava, koska sen 1. derivaatta

$$F'(x) = \frac{e^{-x}}{(1+e^{-x})^2} > 0.$$

On myös helppo todeta, että $F(x)$ ei ole ainoastaan oikealta jatkuva vaan *jatkuva*. □

Satunnaismuuttujan jatkuvuus voidaankin määritellä siihen liittyvän kertymäfunktion jatkuvuuden avulla.

Määritelmä 5.1 Satunnaismuuttuja X on jatkuva, jos sen kertymäfunktio $F_X(x)$ on x:n jatkuva funktio. Satunnaismuuttuja X on diskreetti, jos sen kertymäfunktio on x:n porrasfunktio.

Vastaavalla tavalla kuin diskreetin satunnaismuuttujan kertymäfunktio voidaan lausua summana, voidaan jatkuvan satunnaismuuttujan kertymäfunktio lausua integraalina:

$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t. \tag{5.1.2}$$

Jos $f_X(t)$ on jatkuva, niin integraalilaskennan peruslauseen mukaan

$$F'_X(x) = f_X(x), \tag{5.1.3}$$

missä $F'_X(x)$ on kertymäfunktion $F_X(x)$ derivaatta.

Määritelmä 5.2 Jatkuvan satunnaismuuttujan X tiheysfunktio $f_X(x)$ on funktio, joka toteuttaa yhtälön

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \quad \text{kaikilla} \quad x \in \mathbb{R}. \tag{5.1.4}$$

Esimerkki 5.2 Olkoon X tiettyyn palvelunumeroon tulevien puheluiden pituus. Oletetaan, että X:n tiheysfunktio on

$$f(x) = \frac{1}{20}\mathrm{e}^{-x/20}, \qquad 0 \le x < \infty.$$

Silloin X noudattaa ns. *eksponenttijakaumaa* keskiarvolla 20. Nyt

$$S = \{\, x \mid 0 \le x < \infty \,\} \quad \text{ja} \quad f(x) > 0 \quad \text{kun} \quad x \in S.$$

Kertymäfunktio on

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{20} \mathrm{e}^{-t/20}\,\mathrm{d}x = \int_{0}^{x} \frac{1}{20} \mathrm{e}^{-t/20}\,\mathrm{d}x \\ &= \int_{0}^{x} -\mathrm{e}^{-t/20} = 1 - \mathrm{e}^{-x/20}. \end{aligned}$$

Silloin

$$F'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\big(1 - \mathrm{e}^{-x/20}\big) = \frac{1}{20}\mathrm{e}^{-x/20} = f(x), \qquad x \ge 0$$

ja $f(x) = 0$, kun $x < 0$. □

Huomaa, että yksittäisen pisteen $a \in \mathbb{R}$ todennäköisyys $P(X = a)$ on aina nolla, jos X on jatkuva satunnaismuuttuja. Silloin erityisesti kaikilla reaaliluvuilla $b > a$

$$\begin{aligned} F(b) - F(a) &= P(a \le X \le b) = P(a < X \le b) \\ &= P(a \le X < b) = P(a < X < b). \end{aligned}$$

Esimerkki 5.3 Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on $f(x) = 2x$, kun $0 < x < 1$. Silloin X:n kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ x^2, & 0 \le x < 1; \\ 1, & 1 \le x. \end{cases}$$

Huomaa, että

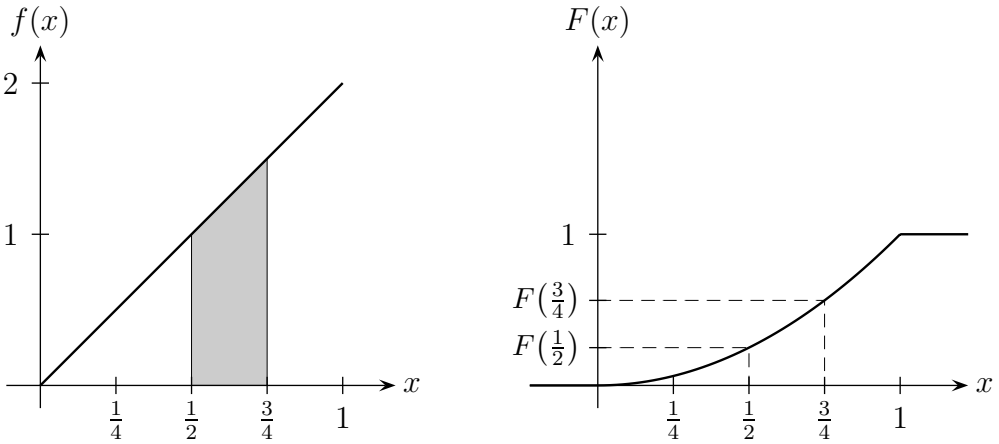
$$F(x) = \int_{0}^{x} 2t\,\mathrm{d}t = x^2, \qquad \text{kun } 0 \le x < 1.$$

Jos kertymäfunktio on annettu, niin tiheysfunktio saadaan derivoimalla kertymäfunktio:

$$F'(x) = \frac{\mathrm{d}}{\mathrm{d}x} x^2 = 2x, \qquad 0 \le x < 1.$$

Kertymäfunktion avulla voidaan laskea todennäköisyyksiä. Esimerkiksi todennäköisyys

$$P\left(\frac{1}{2} < X \le \frac{3}{4}\right) = F\left(\frac{3}{4}\right) - F\left(\frac{1}{2}\right) = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$



Kuvio 5.1. Jatkuvan satunnaismuuttujan X tiheysfunktio $f(x) = 2x$ ja kertymäfunktio $F(x) = x^2$.

ja

$$P\left(\frac{3}{4} < X \leq \frac{3}{2}\right) = F\left(\frac{3}{2}\right) - F\left(\frac{3}{4}\right) = 1 - \left(\frac{3}{4}\right)^2 = \frac{7}{16}.$$

Toisaalta tietysti $P\left(\frac{1}{2} \leq X \leq \frac{3}{4}\right)$ voidaan laskea suoran $y = 2x$ ja x-akselin väliin jäävänä pinta-alana:

$$P\left(\frac{1}{2} \leq X \leq \frac{3}{4}\right) = \int_{1/2}^{3/4} 2x \, \mathrm{d}x = \frac{5}{15},$$

joka tietysti voidaan esittää kertymäfunktion avulla. □

Jatkuvan satunnaismuuttujan momentit määritellään vastaavasti kuin diskreetin satunnaismuuttujan tapauksessa, mutta määritelmässä summa korvataan integraalilla. Jatkuvan satunnaismuuttujan $r.$ *momentti* on

$$\alpha_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) \, \mathrm{d}x,$$

missä $f(x)$ on X:n tiheysfunktio. Satunnaismuuttujan X $r.$ *keskusmomentti* on

$$\mu_r = E[(X - \mu)^r],$$

missä $\mu = E(X) = \alpha_1$ on X:n *odotusarvo*. Satunnaismuuttujan X odotusarvo on siis integraali

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x$$

ja X:n *varianssi* σ^2 on 2. keskusmomentti

$$\begin{aligned}\sigma^2 = \mu_2 &= E[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x.\end{aligned}$$

Merkitsemme myös $E\big[(X-\mu)^2\big] = \mathrm{Var}(X)$, jolloin X:n *hajonta* on

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Momenttifunktio on

$$M(t) = E(\mathrm{e}^{tx}) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\,\mathrm{d}x, \tag{5.1.5}$$

jos integraali 5.1.5 on olemassa jollakin avoimella välillä $(-a, a)$, missä $a > 0$. Tietysti esimerkiksi tulokset

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2, \\ \mu &= M'(0), \\ \alpha_2 &= E(X^2) = M''(0)\end{aligned}$$

pitävät edelleen paikkansa samalla tavalla kuin diskreettien satunnaismuuttujien tapauksessa.

Esimerkki 5.4 Lasketaan nyt Esimerkissä 5.3 määritellyn satunnaismuuttujan X odotusarvo ja varianssi:

$$\mu = E(X) = \int_0^1 x(2x)\,\mathrm{d}x = \frac{2}{3}\Big/_0^1 x^3 = \frac{2}{3}$$

ja

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 \\ &= \int_0^1 x^2(2x)\,\mathrm{d}x - \left(\frac{2}{3}\right)^2 = \frac{1}{2}\Big/_0^1 x^4 - \frac{4}{9} = \frac{1}{18}.\end{aligned}$$

Kolmas momentti on

$$\alpha_3 = E(X^3) = \int_0^1 x^3(2x)\,\mathrm{d}x = \frac{2}{5}\Big/_0^1 x^5 = \frac{2}{5}$$

ja 3. keskusmomentti on

$$\begin{aligned}
\mu_3 &= E\big[(X-\mu)^3\big] = \int_0^1 (x-\mu)^3(2x)\,\mathrm{d}x \\
&= \int_0^1 (x^3 - 3\mu x^2 + 3\mu^2 x - \mu^3)(2x)\,\mathrm{d}x \\
&= \int_0^1 x^3(2x)\,\mathrm{d}x - 3\mu\int_0^1 x^2(2x)\,\mathrm{d}x + 3\mu^2\int_0^1 x(2x)\,\mathrm{d}x - \mu^3\int_0^1 2x\,\mathrm{d}x \\
&= \alpha_3 - 3\mu\alpha_2 + 3\mu^3 - \mu^3 = \alpha_3 - 3\mu\alpha_2 + 2\mu^3 \\
&= \frac{2}{5} - 3\cdot\frac{2}{3}\cdot\frac{1}{2} + 2\cdot\left(\frac{2}{3}\right)^3 = -\frac{3}{5} + \frac{16}{27} = \frac{1}{15}.
\end{aligned}$$

□

Myös prosenttipisteet ovat tärkeitä jakauman tunnuslukuja. Jakauman $100p$-prosenttipiste π_p määritellään seuraavasti:

$$p = \int_{-\infty}^{\pi_p} f(x)\,\mathrm{d}x = F(\pi_p), \qquad 0 \le p \le 1.$$

Prosenttipistettä $\pi_{0.50}$ kutsutaan *mediaaniksi* ja pistettä $\pi_{0.25}$ ja $\pi_{0.75}$ *alakvartiiliksi* ja *yläkvartiiliksi.* Esimerkissä 5.3 käsitellyn jakauman 36 %:n piste on 0.6, koska

$$F(\pi_{0.36}) = \pi_{0.36}^2 = 0.6^2 = 0.36.$$

Esimerkki 5.5 Olkoon satunnaismuuttujan X kertymäfunktio määritelty seuraavasti

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{x^2}{2}, & 0 \le x \le 1; \\ 1 - \frac{(2-x)^2}{2}, & 1 \le x < 2; \\ 1, & 2 \le x. \end{cases}$$

Tarkistamme ensin, että F on todella kertymäfunktio. Toteamme helposti, että

1) $\lim_{x\to-\infty} F(x) = 0$ ja $\lim_{x\to\infty} F(x) = 1$,

2) $F(x)$ on x:n kasvava (ei-vähenevä) funktio ja

3) $F(x)$ on oikealta jatkuva, koska se on jatkuva.

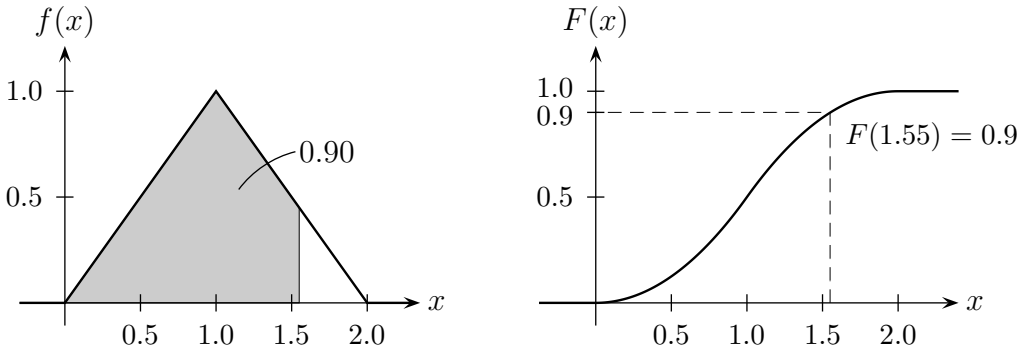
Tiheysfunktio saadaan derivoimalla $F(x)$. Nyt siis $F'(x) = x$ välillä $0 < x \leq 1$ ja $F'(x) = 2 - x$ välillä $1 \leq x \leq 2$. Näin siis tiheysfunktio on

$$f(x) = \begin{cases} x, & 0 < x \leq 1; \\ 2 - x, & 1 \leq x \leq 2; \\ 0 & \text{muualla.} \end{cases}$$

Tiheysfunktio voidaan kirjoittaa lyhyesti muodossa

$$f(x) = 1 - |x - 1|, \qquad 0 \leq x \leq 2.$$

Koska X:n tiheysfunktion kuvaaja on kolmion muotoinen, X:n jakaumaa kutsutaan kolmiojakaumaksi.



Kuvio 5.2. Kolmiojakauman tiheysfunktion ja kertymäfunktion kuvaajat.

Kolmiojakauman odotusarvo on

$$\begin{aligned} \mu = \int_0^2 x f(x)\,\mathrm{d}x &= \int_0^1 x \cdot x\,\mathrm{d}x + \int_1^2 x(2-x)\,\mathrm{d}x \\ &= \Big/_0^1 \frac{x^3}{3} + \Big/_1^2 \left(x^2 - \frac{x^3}{3}\right) = \frac{1}{3} + \left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right) \\ &= 1. \end{aligned}$$

Koska jakauma on symmetrinen odotusarvon 1 suhteen, on 1 myös jakauman mediaani $\pi_{0.50}$. Se voidaan todeta helposti myös määritelmän perusteella, sillä $F(1) = \frac{1^2}{2} = 0.5$. Jakauman 90 %:n piste $\pi_{0.90}$ saadaan ratkaisemalla yhtälö

$$1 - \frac{(2 - \pi_{0.90})^2}{2} = 0.90.$$

Ratkaisu on $\pi_{0.90} = 2 - \sqrt{0.2} = 1.55$. □

Itse asiassa relaatio (5.1.4) ei välttämättä ole voimassa kaikilla x:n arvoilla, sillä $F(x)$ voi olla jatkuva, mutta ei derivoituva. Jos $f(x)$ on jatkuva, niin silloin tietysti yhtälö (5.1.4) pitää paikkansa. Huomattakoon, että jatkuvan satunnaismuuttujan tiheysfunktio ei välttämättä ole jatkuva, mutta kertymäfunktio on.

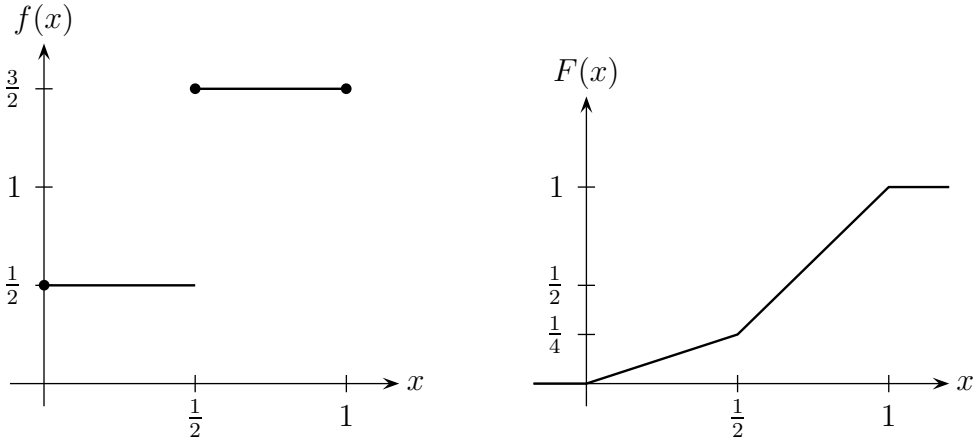
Esimerkki 5.6 Tarkastellaan nyt satunnaismuuttujaa X, jonka tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{2}, & 0 \le x < \frac{1}{2}; \\ \frac{3}{2}, & \frac{1}{2} \le x \le 1. \end{cases}$$

Vastaavasti X:n kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{2}x, & 0 \le x < \frac{1}{2}; \\ \frac{1}{4} + \frac{3}{2}\big(x - \frac{1}{2}\big), & \frac{1}{2} \le x \le 1; \\ 1, & 1 \le x. \end{cases}$$

Havaitsemme nyt, että X:n tiheysfunktio ei ole jatkuva. Nyt myöskään F ei



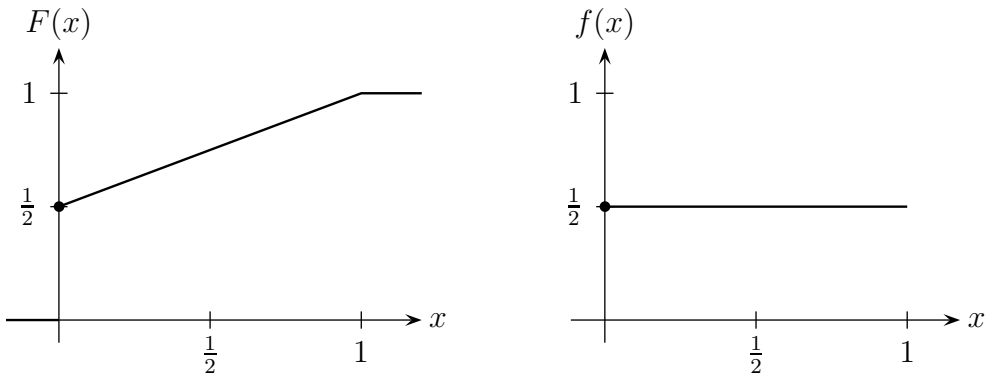
Kuvio 5.3. Satunnaismuuttujan X tiheysfunktion ja kertymäfunktion kuvaajat.

ole derivoituva pisteessä $\frac{1}{2}$. Pisteessä $x = \frac{1}{2}$ *ei ole voimassa,* että $F'(x) = f(x)$. Tässä on esimerkki jatkuvasta satunnaismuuttujasta, jonka tiheysfunktio ei ole jatkuva ja jonka kertymäfunktio ei ole koko määrittelyalueella S derivoituva. □

Jatkuvan satunnaismuuttujan tiheysfunktiolla voi olla äärellinen määrä epäjatkuvuuspisteitä, mutta kertymäfunktio on jatkuva. Esimerkin 5.6 satunnaismuuttujan tiheysfunktiolla on määrittelyalueellaan yksi epäjatkuvuuspiste ja kertymäfunktio on jatkuva. Relaatio (5.1.3) pitää paikkansa vain tiheysfunktion jatkuvuuspisteissä, mutta ei epäjatkuvuuspisteissä.

Esimerkki 5.7 Määritellään satunnaismuuttuja X siten, että sen kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{2}, & x = 0; \\ \frac{1}{2} + \frac{x}{2}, & 0 < x < 1; \\ 1, & 1 \leq x. \end{cases} \tag{5.1.6}$$



Kuvio 5.4. Satunnaismuuttujan X kertymäfunktion ja 'tiheysfunktion' kuvaajat.

Kertymäfunktio ei ole nyt jatkuva, koska funktio hyppää pisteessä $x = 0$. Kertymäfunktio ei ole myöskään porrasfunktio. Nyt myös yksittäisellä pisteellä $X = 0$ on positiivinen todennäköisyys $P(X = 0) = \frac{1}{2}$, joten $f(x)$ ei ole tiheysfunktio. Itse asiassa kertymäfunktio (5.1.6) voidaan kirjoittaa porrasfunktion (kertymäfunktio) ja jatkuvan kertymäfunktion summana. Alaluvussa 4.1 määriteltiin hyppyfunktio $\varepsilon(x)$ siten, että $\varepsilon(x) = 1$ epänegatiivisilla x:n arvoilla ja $\varepsilon(x) = 0$, kun $x < 0$. Funktio $\varepsilon(x)$ on porrasfunktio ja siis diskreetin satunnaismuuttujan kertymäfunktio. Puoliavoimella välillä $(0, 1]$ tasajakaumaa noudattavan satunnaismuuttujan kertymäfunktio on

$$F_c(x) = \begin{cases} 0, & x \leq 0; \\ x, & 0 < x < 1; \\ 1, & 1 \leq x. \end{cases}$$

Nyt kertymäfunktio (5.1.6) voidaan kirjoittaa muodossa

$$F(x) = \frac{1}{2}\,\varepsilon(x) + \frac{1}{2}F_c(x).$$

Esimerkiksi todennäköisyys

$$\begin{aligned} P\left(X \leq \frac{1}{2}\right) &= \frac{1}{2}\,\varepsilon\left(\frac{1}{2}\right) + \frac{1}{2}F_c\left(\frac{1}{2}\right) \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Satunnaismuuttuja X ei ole diskreetti eikä jatkuva. □

Yleisesti jatkuva satunnaismuuttuja voidaan määritellä identiteetin (5.1.4) avulla olettamatta tiheysfunktion $f(x)$ jatkuvuutta. Jos on olemassa sellainen epänegatiivinen funktio $f(x)$ [ts. $f(x) \geq 0$ kaikilla $x \in \mathbb{R}$], että (5.1.4) pitää paikkansa kaikilla $x \in \mathbb{R}$, niin kertymäfunktion $F(x)$ sanotaan olevan *absoluuttisesti jatkuva.* Absoluuttisesti jatkuva funktio on jatkuva. Kaikkien tässä luvussa käsiteltäviät jatkuvien satunnaismuuttujien kertymäfunktiot ovat absoluuttisesti jatkuvia.

5.2 Tasajakauma ja eksponenttijakauma

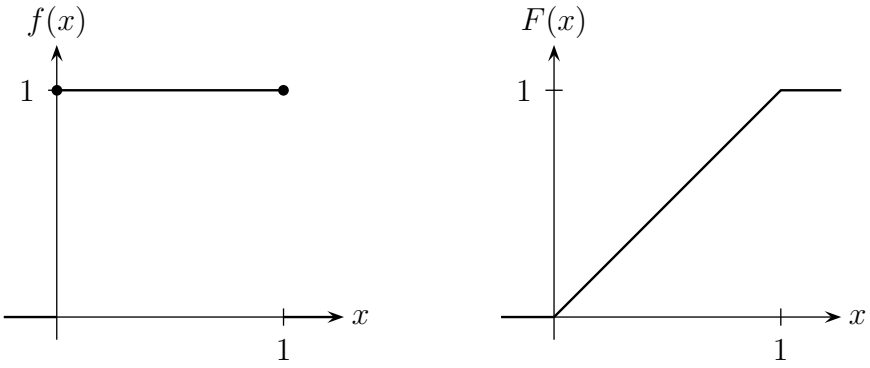
5.2.1 Tasajakauma

Jatkuva satunnaismuuttuja X noudattaa *tasajakaumaa* välillä $[0, 1]$, jos sen tiheysfunktio on 1 tällä välillä ja 0 muualla:

$$f(x) = \begin{cases} 1, & \text{kun } x \in [0,1], \\ 0 & \text{muualla.} \end{cases} \tag{5.2.1}$$

Silloin merkitään $X \sim \mathrm{Tas}(0, 1)$. On helppo todeta, että $f(x)$ on tiheysfunktio, koska $f(x) \geq 0$ ja

$$\int_0^1 f(x)\,\mathrm{d}x = \int_0^1 \mathrm{d}x = 1.$$



Kuvio 5.5. Tasajakauman $\mathrm{Tas}(0, 1)$ tiheysfunktio ja kertymäfunktio.

Tasajakauman keskiarvo ja varianssi ovat:

$$E(X) = \int_0^1 x\,\mathrm{d}x = \frac{1}{2}$$

ja

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \int_0^1 x^2\,\mathrm{d}x - \frac{1}{4} = \frac{1}{12}.$$

Satunnaismuuttujan X momenttifunktio on

$$M_X(t) = \int_0^1 \mathrm{e}^{tx}\,\mathrm{d}x = \Big/_0^1 \frac{1}{t}\mathrm{e}^{tx} = \frac{\mathrm{e}^t - 1}{t}.$$

Huomaa, että $M_X(0) = 1$.

Olkoon $[a, b]$ annettu suljettu väli, $a < b$. Silloin satunnaismuuttuja $U = (b-a)X + a$ noudattaa tasajakaumaa välillä $[a, b]$. Silloin merkitään $U \sim \mathrm{Tas}(a, b)$. Koska $E(U) = (b-a)\,E(X) + a$ ja $\mathrm{Var}(U) = (b-a)^2\,\mathrm{Var}(X)$, niin

$$E(U) = \frac{a+b}{2} \qquad \text{ja} \qquad \mathrm{Var}(U) = \frac{(b-a)^2}{12}.$$

Satunnaismuuttujan U tiheysfunktio on

$$f(u) = \begin{cases} \frac{1}{b-a}, & \text{kun } u \in [a, b]; \\ 0 & \text{muualla} \end{cases} \tag{5.2.2}$$

ja U:n momenttifunktio on

$$M_U(t) = \begin{cases} \dfrac{\mathrm{e}^{tb} - \mathrm{e}^{ab}}{t(b-a)}, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

5.2.2 Eksponenttijakauma

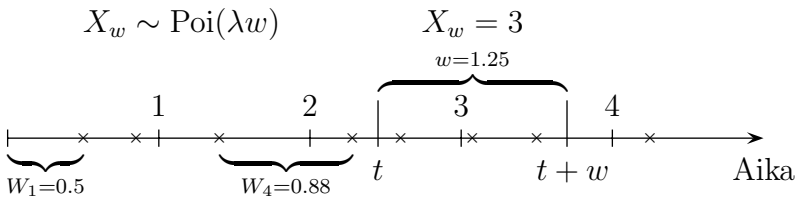
Poissonin prosessissa tarkastellaan, montako tapahtumaa (lisäystä) sattuu jollain aikavälillä. Merkitään w:n pituisella välillä sattuvien tapahtumien lukumäärää satunnaismuuttujalla X_w. Jos Poissonin prosessin intensiteetti on λ, niin Määritelmän 4.3 mukaan todennäköisyys, että w:n pituisella välillä sattuu x tapahtumaa, on

$$P(X_w = x) = \mathrm{e}^{-\lambda w}\frac{(\lambda w)^x}{x!}. \tag{5.2.3}$$

Poissonin prosessilla voidaan mallintaa esimerkiksi asiakkaiden saapumista palvelupisteeseen, puheluiden tuloa vaihteeseen, onnettomuuksien sattumista tarkasteltavalla tieosuudella tai autojen kulkua liikenteen tarkkailupisteen ohi. Tällöin ajatellaan, että yksittäiset tapahtumat sattuvat toisistaan riippumatta täysin satunnaisesti.

Tarkkaillaan nyt Poissonin prosessia, jonka intensiteetti on λ. Olkoon W odotusaika siihen hetkeen, kunnes seuraava tapahtuma sattuu. Odotusaika on jatkuva satunnaismuuttuja. Jos tarkkailemme prosessia hetkestä t hetkeen $t+w$ eli w:n pituisen aikavälin $[t, t+w]$, niin tapahtuma $\{\, W > w \,\}$ sattuu jos ja vain jos Poissonin prosessissa ei satu yhtään tapahtumaa välillä $[t, t+w]$. Siksi identiteetin (5.2.3) mukaan

$$P(W > w) = P(X_w = 0) = \mathrm{e}^{-\lambda w}.$$



Kuvio 5.6. Kaaviokuva esittää Poissonin saapumisprosessia, esimerkiksi autojen kulkemista liikenteen tarkkailupisteen ohi. Esimerkiksi W_1 on 1. auton odotusaika ja W_4 on 3. ja 4. auton välinen aika. Kiinnitetyllä w:n pituisella välillä on kulkenut ohi $X_w = 3$ autoa. Peräkkäiset odotusajat W_1, W_2, W_3, ... ovat toisistaan riippumattomat ja noudattavat samaa jakaumaa.

Odotusajan W kertymäfunktio on siis

$$\begin{aligned} F(w) &= P(W \le w) \\ &= 1 - P(W > w) = 1 - P(X_w = 0) \\ &= 1 - \mathrm{e}^{-\lambda w}. \end{aligned}$$

Koska odotusaika W on epänegatiivinen, niin $F(w) = 0$, kun $w < 0$.

Odotusajan W tiheysfunktio on

$$F'(w) = f(w) = \lambda \mathrm{e}^{-\lambda w}$$

derivointisäännön (5.1.3) nojalla. Usein merkitään $\lambda = \frac{1}{\theta}$, missä $\theta > 0$. Sanomme, että W noudattaa *eksponenttijakaumaa* parametrilla θ ja merkitsemme $W \sim \mathrm{Exp}(\theta)$. Parametri θ on jakauman keskiarvo. Eksponenttijakauman tiheysfunktio on silloin muotoa

$$f(w) = \frac{1}{\theta} \mathrm{e}^{-w/\theta}. \tag{5.2.4}$$

Eksponenttijakauman $\mathrm{Exp}(\theta)$ momenttifunktio on

$$\begin{aligned} M(t) &= \int_0^\infty \frac{1}{\theta} \mathrm{e}^{-w/\theta} \, \mathrm{d}w = \int_0^\infty -\frac{\mathrm{e}^{-(1-\theta t)w/\theta}}{1-\theta t} \\ &= \frac{1}{1-\theta t}, \qquad t < \frac{1}{\theta}. \end{aligned}$$

Eksponenttijakaumalla on vastaava "unohtamisominaisuus" kuin geometrisella jakaumalla. Jos $T \sim \mathrm{Exp}(\theta)$, niin

$$P(T > a + b \mid T > a) = P(T > b) \tag{5.2.5}$$

kaikilla epänegatiivisilla a ja b. Tulos voidaan todistaa laskemalla ehdollinen todennäköisyys

$$\begin{aligned} P(T > a + b \mid T > a) &= \frac{P(T > a,\ T > a + b)}{P(T > a)} = \frac{P(T > a + b)}{P(T > a)} \\ &= \frac{\mathrm{e}^{-(a+b)/\theta}}{\mathrm{e}^{-a/\theta}} = \mathrm{e}^{-b/\theta} = P(T > b). \end{aligned}$$

Huomattakoon, että edellä on käytetty tulosta

$$P(T > t) = 1 - P(T \leq t) = 1 - F(t) = \mathrm{e}^{-t/\theta}, \qquad t \geq 0.$$

Esimerkki 5.8 Oletetaan, että asiakkaiden saapuminen liikkeeseen noudattaa Poissonin prosessia intensiteetillä 20 asiakasta tunnissa. Mikä on todennäköisyys, että myyjä joutuu odottamaan seuraavaa asiakasta yli 5 minuuttia? Olkoon X odotusaika, kunnes seuraava asiakas saapuu. Silloin prosessissa (5.2.3) $\lambda = 1/3$ asiakasta minuutissa ja $X \sim \mathrm{Exp}(3)$, koska eksponenttijakauman keskiarvo $\theta = 1/\lambda$. Jakauman Exp(3) tiheysfunktio on

$$f(x) = \frac{1}{3}\mathrm{e}^{-x/3}, \qquad 0 \leq x < \infty$$

ja

$$P(X > 5) = \int_5^\infty \frac{1}{3}\mathrm{e}^{-x/3}\,\mathrm{d}x = \Big/_5^\infty -\mathrm{e}^{-x/3} = \mathrm{e}^{-5/3} \approx 0.1889.$$

Jatkuvan jakauman *mediaani* m on sellainen piste, että $F(m) = 1/2$. Nyt jakauman Exp(3) mediaanin m tulee toteuttaa ehto $F(m) = 1 - \mathrm{e}^{-m/3} = \frac{1}{2}$, joten

$$m = 3\log(2) \approx 2.0794.$$

□

5.2.3 Elinaikajakauma

Ominaisuuden (5.2.5) perusteella eksponenttijakauma on sopiva elinajan jakauma silloin, kun jäljellä oleva elinaika ei riipu tämänhetkisestä iästä. Olkoon T esimerkiksi jonkin elektronisen komponentin ikä tunteina. Silloin $P(T > b)$ on todennäköisyys, että uusi komponentti kestää ainakin b tuntia, kun taas $P(T > a + b \mid T > a)$ on todennäköisyys, että a tuntia käytössä ollut komponentti kestää vielä b tuntia. Jos elinaika noudattaa eksponenttijakaumaa, niin ominaisuuden (5.2.5) nojalla todennäköisyydet $P(T > b)$ ja $P(T > a + b \mid T > a)$ ovat samat kaikilla a ja b. Todennäköisyys, että komponentti rikkoontuu b:n seuraavan tunnin aikana, ei riipu lainkaan siitä, kuinka kauan komponentti on jo ollut käytössä.

Funktiota $G(t) = P(T > t)$ kutsutaan *eloonjäämisfunktioksi.* Eksponenttijakauma määrittelee eloonjäämisfunktion $G(t) = \mathrm{e}^{-t/\theta}$, jolla on unohtamisominaisuus

$$G(t + s) = G(t)G(s), \qquad t > 0, \quad s > 0. \tag{5.2.6}$$

Määritelmänsä nojalla $G(0) = 1$ ja $G(t) \to 0$, kun t kasvaa. Onko eksponenttifunktion lisäksi muita eloonjäämisfunktioita, joilla on unohtamisominaisuus (5.2.6)? Voidaan osoittaa, että ehdon (5.2.6) toteuttavat eloonjäämisfunktiot ovat aina muotoa $\mathrm{e}^{-\lambda t}$, $\lambda > 0$.

Jos elinaika T noudattaa eksponenttijakaumaa $\mathrm{Exp}(\theta)$, niin vakio $\lambda = \frac{1}{\theta}$ on hetkellinen *kuolleisuusaste* tai *vaaran aste*. Parametri λ säätelee todennäköisyyttä kuolla hetken $T = t$ jälkeisellä yksikön pituisella aikavälillä.

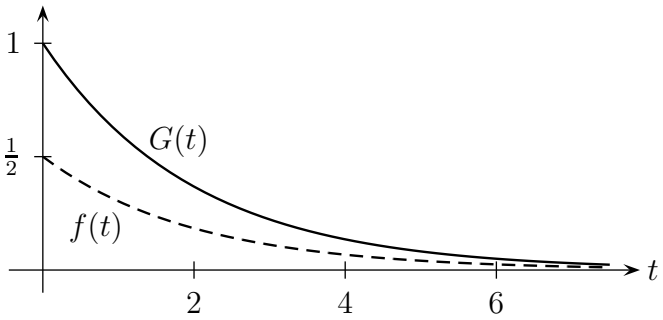
Olkoon Δ tarkasteltavan aikavälin pituus. Määritelläään todennäköisyys

$$\begin{aligned} P(T \le t + \Delta \mid T > t) &= 1 - P(T > t + \Delta \mid T > t) \\ &= 1 - P(T > \Delta) = 1 - \mathrm{e}^{-\lambda\Delta}, \end{aligned}$$

missä viimeistä edellinen yhtäsuuruus saadaan unohtamisominaisuuden (5.2.6) nojalla. Kun funktiota $\mathrm{e}^{-\lambda\Delta}$ arvioidaan Taylorin polynomin avulla, saadaan

$$\begin{aligned} 1 - \mathrm{e}^{-\lambda\Delta} &= 1 - (1 - \lambda\Delta + \frac{1}{2}\lambda^2\Delta^2 - \cdots) \\ &= \lambda\Delta - \frac{1}{2}\lambda^2\Delta^2 + \cdots \\ &\approx \lambda\Delta, \qquad \text{kun } \Delta \text{ on pieni.} \end{aligned}$$

Arviointivirhe pienenee merkityksettömäksi verrattuna Δ:aan, kun $\Delta \to 0$. Silloin siis $P(T \le t + \Delta \mid t > t) \approx \lambda\Delta$.



Kuvio 5.7. Eksponenttijakauman $\mathrm{Exp}(2)$ tiheysfunktio $f(t) = \frac{1}{2}\mathrm{e}^{-t/2}$ ja vastaava eloonjäämisfunktio $G(t) = \mathrm{e}^{-t/2}$.

Nyt nähdään, että

$$\lim_{\Delta \to 0} \frac{P(T \le t + \Delta \mid T > t)}{\Delta} = \lambda$$

on riippumaton ajasta t. Eksponentiaalisesti jakautuneen elinajan tapauksessa kuolleisuusaste λ on iästä riippumaton vakio. Yleisesti kuolleisuusaste $\lambda(t)$ on tietysti iän funktio.

5.3 Gammajakauma ja χ^2-jakauma

Gammajakaumajakauma on välillä $[0, \infty)$ määritelty jakauma tai jakaumaperhe, koska parametrien vaihdellessa saadaan hyvinkin erinäköisiä jakaumia,

vaikka ne ovat matemaattisesti samaa muotoa. Gammafunktio

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\,\mathrm{d}x \tag{5.3.1}$$

määriteltiin jo Pykälässä 2.4.7. Jos $\alpha > 0$, niin $\Gamma(\alpha)$ on äärellinen. Jos α on positiivinen kokonaisluku, niin $\Gamma(\alpha)$ voidaan lausua suljetussa muodossa, muutoin ei.

Gammafunktio toteuttaa rekursiivisen relaation

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha),$$

joka voidaan osoittaa osittaisintegroinnilla. Jos $\alpha = n$ on positiivinen kokonaisluku, niin

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\cdots 2\cdot 1\cdot\Gamma(1) = n!\Gamma(1).$$

Koska $\Gamma(1) = 1$, niin

$$\Gamma(n+1) = n!$$

kaikilla positiivisilla kokonaisluvuilla. Myös $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ on tärkeä erikoistapaus.

Funktio

$$f(t) = \frac{t^{\alpha-1}\mathrm{e}^{-t}}{\Gamma(\alpha)}, \qquad 0 < t < \infty \tag{5.3.2}$$

määrittelee tiheysfunktion, sillä gammafunktiossa integroitava on positiivinen välillä $(0,\infty)$. Sanokaamme, että (5.3.2) on satunnaismuuttujan T tiheysfunktio. Kaikkien gammajakaumien perhe saadaan määrittelemällä satunnaismuuttuja $X = \beta T$, missä β on positiivinen vakio. X:n tiheysfunktio voidaan johtaa soveltamalla Lauseen 5.5 muunnostekniikkaa. Merkitsemme $X \sim \text{Gamma}(\alpha,\beta)$ ja sanomme, että X noudattaa gammajakaumaa parametrein α ja β. Jakauman $\text{Gamma}(\alpha,\beta)$ tiheysfunktioksi saadaan

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\mathrm{e}^{-x/\beta}, \qquad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0. \tag{5.3.3}$$

Esitämme nyt gammajakauman perusominaisuudet seuraavassa lauseessa.

Lause 5.1 *Oletetaan, että* $X \sim \text{Gamma}(\alpha,\beta)$.

1. *Funktio (5.3.3) määrittelee tiheysfunktion kaikilla* $\alpha > 0$, $\beta > 0$.

2. $$E(X) = \alpha\beta, \qquad \text{Var}(X) = \alpha\beta^2$$

   ja

   $$M(t) = E(\mathrm{e}^{tX}) = \left(\frac{1}{1-\beta t}\right)^\alpha, \qquad t < \frac{1}{\beta}.$$

3.

$$E(X^c) = \frac{\Gamma(\alpha + c)\beta^c}{\Gamma(\alpha)}$$

kaikilla $c > -\alpha$.

4. Olkoon $U = bX$, $b > 0$. *Silloin* $U \sim \text{Gamma}(\alpha, b\beta)$.

Eksponettijakauma on gammajakauman erikoistapaus. Kun sijoitetaan tiheysfunktioon (5.3.3) $\alpha = 1$, saadaan

$$f(x;\beta) = \frac{1}{\beta}\mathrm{e}^{-x/\beta}, \qquad x > 0.$$

Havaitaan siis, että $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$.

χ^2-jakauma

Toinen tärkeä gammajakauman erikoistapaus on χ^2-jakauma. Jos valitaan $\alpha = \frac{r}{2}$, missä r on positiivinen kokonaisluku, ja $\beta = 2$, tulee tiheysfunktio (5.3.3) muotoon

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}x^{(r/2)-1}\mathrm{e}^{-x/2}, \qquad 0 < x < \infty, \tag{5.3.4}$$

mikä on χ^2-jakauman tiheysfunktio vapausastein r. Jos X noudattaa χ^2-jakaumaa vapausastein r, merkitään $X \sim \text{Khi2}(r)$. χ^2-jakauman keskiarvo, varianssi ja momenttifunktio saadaan nyt suoraan gammajakauman avulla. Jos $X \sim \text{Khi2}(r)$, niin

$$E(X) = r, \qquad \text{Var}(X) = 2r$$

ja

$$M(t) = (1 - 2t)^{-r/2}, \qquad t < \frac{1}{2}.$$

Odotusaika Poissonin prosessissa

Seuraavan tapahtuman odotusaika Poissonin prosessissa noudattaa eksponettijakaumaa. Olkoon W nyt odotusaika, kunnes sattuu α tapahtumaa, missä α on siis positiivinen kokonaisluku. Jos Poissonin prosessin intensiteetti on λ, niin todennäköisyys, että w:n pituisella aikavälillä sattuu x tapahtumaa, saadaan kaavalla (5.2.3):

$$P(X_w = x) = \mathrm{e}^{-\lambda w}\frac{(\lambda w)^x}{x!}.$$

Odotusajan W kertymäfunktio, kun $W \geq 0$, on

$$\begin{aligned} F(w) &= P(W \leq w) = 1 - P(W > w) \\ &= 1 - P(\text{vähemmän kuin } \alpha \text{ tapahtumaa välillä } [t, t+w]) \\ &= 1 - \sum_{x=0}^{\alpha-1} \mathrm{e}^{-\lambda w} \frac{(\lambda w)^x}{x!}, \end{aligned}$$

koska tapahtumien lukumäärä aikavälillä $[t, t+w]$ noudattaa Poissonin jakaumaa keskiarvolla λw [ks. (5.2.3)]. Laskemalla derivaatta $F'(w) = f(w)$ saadaan tiheysfunktio

$$f(w) = \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \mathrm{e}^{-\lambda w}.$$

Jos $w < 0$, niin $F(w) = 0$ ja $f(w) = 0$. Nyt huomaamme, että

$$W \sim \text{Gamma}\Big(\alpha, \frac{1}{\lambda}\Big).$$

5.4 Normaalijakauma

5.4.1 Standardimuotoinen normaalijakauma

Tarkastelemme nyt todennäköisyysteorian ja tilastotieteen tärkeintä jakaumaa, normaalijakaumaa. Olkoon Z jatkuva satunnaismuuttuja, jonka tiheysfunktio on

$$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \qquad -\infty < z < \infty. \tag{5.4.1}$$

Silloin Z noudattaa standardimuotoista normaalijakaumaa. Käytetään myös sanontaa "Z noudattaa standardoitua normaalijakaumaa".

Tarkistamme nyt, että (5.4.1) on todellakin tiheysfunktio. Koska $f(z) > 0$, pitää vain osoittaa, että

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-z^2/2} \, \mathrm{d}z = 1.$$

Osoitamme siis, että

$$\int_{-\infty}^{\infty} \mathrm{e}^{-z^2/2} \, \mathrm{d}z = \sqrt{2\pi}. \tag{5.4.2}$$

Emme pysty suoraan integroimaan funktiota $\mathrm{e}^{-z^2/2}$, koska sen integraalifunktio ei ole lausuttavissa suljetussa muodossa. Osoittautuu kuitenkin, että integraalin (5.4.2) neliö on helppo laskea.

Integraalin arvo ei muutu, jos integrointimuuttuja nimetään uudelleen, joten

$$I = \int_{-\infty}^{\infty} \mathrm{e}^{-z^2/2}\,\mathrm{d}z = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}\,\mathrm{d}y.$$

Riittää osoittaa, että $I^2 = 2\pi$. Nyt

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right) \left(\int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}\,\mathrm{d}y \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y = \left(\int_0^{\infty} r\mathrm{e}^{-r^2/2}\,\mathrm{d}r \right) \left(\int_0^{2\pi} \mathrm{d}\theta \right) \\ &= 2\pi \int_0^{\infty} \mathrm{e}^{-u}\,\mathrm{d}u = 2\pi. \end{aligned}$$

Näin siis tulos (5.4.2) pitää paikkansa. Edellä kolmas yhtäsuuruus saadaan siirtymällä napakoordinaatteihin:

$$x = r\cos\theta \qquad \text{ja} \qquad y = r\sin\theta.$$

Silloin $x^2 + y^2 = r^2$, $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}\theta\,\mathrm{d}r$ ja integrointirajat ovat $0 < r < \infty$, $0 < \theta < 2\pi$.

Integraalilla (5.4.2) on myös läheinen yhteys gammafunktioon. Koska integraalissa (5.4.2) integroitava on symmetrinen nollan suhteen, niin integraalit yli välien $(-\infty, 0)$ ja $(0, \infty)$ ovat yhtä suuret. Siksi

$$\int_0^{\infty} \mathrm{e}^{-z^2/2}\,\mathrm{d}z = \sqrt{\frac{\pi}{2}}. \tag{5.4.3}$$

Tekemällä sijoitus $x = \frac{1}{2}z^2$ integraaliin (5.4.3) saadaan integraali, joka on $\Gamma\left(\frac{1}{2}\right)$. Silloin

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} x^{-1/2}\mathrm{e}^{-x}\,\mathrm{d}x = \sqrt{\pi}. \tag{5.4.4}$$

Lause 5.2 *Oletetaan, että Z noudattaa standardoitua normaalijakaumaa. Silloin*

1. *Z:n momenttifunktio on*

$$M(t) = \mathrm{e}^{t^2/2}, \qquad -\infty < t < \infty.$$

2. *$E(Z) = 0$ ja $\mathrm{Var}(Z) = 1$.*

Todistus. *1.* Määritelmän mukaan

$$M(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\,\mathrm{d}z.$$

Tehdään sijoitus $x = z - t$. Silloin $\mathrm{d}z = \mathrm{d}x$ ja $\mathrm{e}^{tz}\mathrm{e}^{-z^2/2} = \mathrm{e}^{(t^2-x^2)/2}$, joten

$$M(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{(t^2-x^2)/2}\,\mathrm{d}x = \mathrm{e}^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \mathrm{e}^{t^2/2}.$$

Viimeinen yhtäsuuruus seuraa siitä, että integraali yli normaalijakauman tiheysfunktion $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ on 1.

2. Koska $M(t) = \mathrm{e}^{t^2/2}$, niin $M'(t) = t\mathrm{e}^{t^2/2}$ ja $M''(t) = \mathrm{e}^{t^2/2} + t^2\mathrm{e}^{t^2/2}$. Silloin $M'(0) = 0$, $M''(0) = 1$ ja $\mathrm{Var}(Z) = M''(0) - [M'(0)]^2 = 1$. □

Merkitään $Z \sim \mathrm{N}(0,1)$, missä siis $E(Z) = 0$ ja $\mathrm{Var}(Z) = 1$. Seuraavassa pykälässä määritellään normaalijakauma, jonka keskiarvo on μ ja varianssi σ^2.

5.4.2 Yleinen normaalijakauma

Satunnaismuuttuja X *noudattaa normaalijakaumaa* keskiarvolla μ ja varianssilla $\sigma^2 > 0$, jos se voidaan esittää muodossa

$$X = \mu + \sigma Z,$$

missä $Z \sim \mathrm{N}(0,1)$. Silloin merkitään $X \sim \mathrm{N}(\mu, \sigma^2)$. Jos $X \sim \mathrm{N}(\mu, \sigma^2)$, niin vastaavasti

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0,1).$$

Seuraavassa lauseessa esitetään jakaumaa koskevat perustulokset.

Lause 5.3 *Jos $X \sim \mathrm{N}(\mu, \sigma^2)$, niin*

1. $E(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2 \quad$ *ja*

2.

$$M_X(t) = E\big(\mathrm{e}^{tX}\big) = \mathrm{e}^{\mu t + \sigma^2 t^2/2}, \qquad -\infty < t < \infty.$$

3. X:n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty.$$

Todistus. *1.* Koska $X \sim \mathrm{N}(\mu, \sigma^2)$, niin $X = \mu + \sigma Z$, missä $Z \sim \mathrm{N}(0,1)$. Silloin

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma\, E(Z) = \mu$$

ja

$$\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2\, \mathrm{Var}(Z) = \sigma^2.$$

2. Määritelmän mukaan (ks. myös Lause 3.14)

$$\begin{aligned} M_X(t) &= E\left(\mathrm{e}^{tX}\right) = E\left[\mathrm{e}^{t(\mu+\sigma Z)}\right] = \mathrm{e}^{t\mu}\, E\left(\mathrm{e}^{t\sigma Z}\right) \\ &= \mathrm{e}^{t\mu} M_Z(t\sigma) = \mathrm{e}^{t\mu}\mathrm{e}^{t^2\sigma^2/2} = \mathrm{e}^{t\mu + t^2\sigma^2/2}. \end{aligned}$$

3. Tehdään muunnos $x = h(z) = \mu + \sigma z$. Silloin h:lla on käänteisfunktio g ja $z = g(x) = \frac{x-\mu}{\sigma}$ sekä $g'(x) = \frac{1}{\sigma}$. Alaluvussa 5.5 esitettävän muunnostekniikan avulla saadaan X:n tiheysfunktioksi

$$\begin{aligned} f_X(x) &= f_Z\left(\frac{x-\mu}{|\sigma|}\right)\frac{1}{|\sigma|} \\ &= \frac{1}{\sqrt{2\pi}|\sigma|}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}. \end{aligned} \tag{5.4.5}$$

Tavallisesti tiheysfunktio kirjoitetaan muodossa

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/2\sigma^2},$$

missä

$$\sigma = +\sqrt{\mathrm{Var}(X)} = +\sqrt{\sigma^2}$$

on X:n hajonta. Todistuksessa ei oletettu, että $\sigma > 0$. □

Esimerkki 5.9 Jos X:n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{32\pi}}\mathrm{e}^{-(x+7)^2/32}, \qquad -\infty < x < \infty,$$

niin $X \sim \mathrm{N}(-7, 16)$ ja

$$M_X(t) = \mathrm{e}^{-7t+8t^2}.$$

□

Esimerkki 5.10 Jos X:n momenttifunktio on

$$M_X(t) = \mathrm{e}^{5t+12t^2},$$

niin $X \sim \mathrm{N}(5, 24)$ ja X:n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{48\pi}}\mathrm{e}^{-(x-5)^2/48}, \qquad -\infty < x < \infty.$$

□

Jos $X \sim \mathrm{N}(\mu, \sigma^2)$, niin X:n tiheysfunktio saavuttaa maksimin pisteessä $x = \mu$ ja käänteispisteet ovat $x = \mu \pm \sigma$. Todennäköisyysmassa on jakautunut siten, että

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = 0.6826, \\ P(|X - \mu| \leq 2\sigma) &= P(|Z| \leq 2) = 0.9544, \\ P(|X - \mu| \leq 3\sigma) &= P(|Z| \leq 3) = 0.9974, \end{aligned}$$

missä $Z \sim \mathrm{N}(0, 1)$. Esimerkiksi

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = P(-1 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-1) = 0.8413447 - 0.1586553 = 0.6826895, \end{aligned}$$

missä

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-v^2/2} \, \mathrm{d}v$$

on standardimuotoisen normaalijakauman kertymäfunktio. Sen arvot on taulukoitu ja se saadaan laskettua useilla ohjelmistoilla. Edellä esitettyjen todennäköisyyksien kahden numeron likiarvoina käytetään tavallisesti lukuja 0.68, 0.95 ja 0.99, jotka eivät ole pyöristettyjä vaan katkaistuja arvoja. Myös yllä esitetyt neljän numeron likiarvot ovat katkaistuja arvoja.

Lause 5.4

1. *Olkoon $X \sim \mathrm{N}(\mu, \sigma^2)$ ja $U = aX + b$, missä $a \neq 0$ ja b ovat annettuja vakioita. Silloin*
$$U \sim \mathrm{N}(a\mu + b, a^2\sigma^2).$$

2. *Olkoot X_1, X_2, …, X_n riippumattomat, $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$ ja a_1, a_2, …, a_n, b ovat annetut vakiot, joista ainakin yksi a_i poikkeaa nollasta. Silloin $Y = \sum_{i=1}^{n} a_i X_i + b$ noudattaa normaalijakaumaa*
$$Y \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i + b, \ \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).$$

Esimerkki 5.11 Riippumattomat satunnaismuuttujat X_1, X_2, X_3 noudattavat normaalijakaumaa siten, että $X_i \sim \mathrm{N}(2^i, i^i)$, $i = 1, 2, 3$. Silloin $Y = X_1 + X_2 + X_3 \sim \mathrm{N}(14, 32)$, sillä

$$E(Y) = 2 + 2^2 + 2^3 = 14 \qquad \text{ja} \qquad \mathrm{Var}(Y) = 1 + 2^2 + 3^3 = 32$$

ja Lauseen 5.4 mukaan Y noudattaa normaalijakaumaa. Satunnaismuuttuja $Y = X_1 + 2X_2 + 3X_3 \sim \mathrm{N}(34, 260)$, koska

$$E(Y) = 2 + 2 \cdot 2^2 + 3 \cdot 2^3 = 34$$

ja

$$\mathrm{Var}(Y) = 1 + 2^2 \cdot 2^2 + 3^2 \cdot 3^3 = 260.$$

□

5.5 Muuttujien vaihto

Oletetaan, että X on jatkuva satunnaismuuttuja, jonka kertymäfunktio on $F(x)$. Lukuisissa sovelluksissa tarvitaan satunnaismuuttujan X jonkin funktion $Y = h(X)$ jakaumaa, kun X:n jakauma tunnetaan. Tehtävänämme on nyt siis määrittää satunnaismuuttujan $Y = h(X)$ jakauma, missä $h(x)$ on x:n reaaliarvoinen funktio.

5.5.1 Muunnos kertymäfunktio avulla

Voimme pyrkiä johtamaan Y:n kertymäfunktion

$$G(y) = P(Y \leq y)$$

suoraan X:n kertymäfunktion $F(x)$ avulla. Y:n tiheysfunktio $g(y)$ voidaan määrittää sitten identiteetin (5.1.3) avulla, kun $G(y)$ on derivoituva.

Esimerkki 5.12 Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on

$$f(x) = \frac{3x^2}{2}, \qquad -1 \leq x \leq 1.$$

Tarkastellaan satunnaismuuttujan $Y = X^2$ jakaumaa. Silloin Y:n arvoavaruus on $S_Y = [0, 1]$ ja Y:n kertymäfunktio on

$$\begin{aligned} G(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \geq X \geq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{3x^2}{2}\,\mathrm{d}x = \Big/_{-\sqrt{y}}^{\sqrt{y}} \frac{x^3}{2} = y^{3/2}, \qquad 0 \leq y \leq 1. \end{aligned}$$

Derivoimalla saadaan Y:n tiheysfunktioksi

$$g(y) = G'(y) = \frac{3y^{1/2}}{2}, \qquad 0 \leq y \leq 1.$$

□

Esimerkki 5.13 Olkoon X jatkuva satunnaismuuttuja, jonka kertymäfunktio on

$$F(x) = 1 - (1 + x)\mathrm{e}^{-x}, \qquad x > 0.$$

Johdetaan satunnaismuuttujan $Y = \mathrm{e}^{-X}$ jakauma. Merkitään Y:n kertymäfunktiota G:llä. Silloin

$$\begin{aligned} G(y) &= P(Y \leq y) = P\big(\mathrm{e}^{-X} \leq y\big) = P[-X \leq \log(y)] \\ &= P[X \geq -\log(y)] = 1 - P[X < -\log(y)] \\ &= 1 - F[-\log(y)], \end{aligned}$$

missä $F(x)$ on X:n kertymäfunktio. Sijoittamalla $x = -\log(y)$ X:n kertymäfunktioon saadaan

$$G(y) = [1 - \log(y)]\mathrm{e}^{\log(y)} = [1 - \log(y)]y.$$

Koska $S_X = (0, \infty)$, niin $S_Y = (0, 1)$. Y on jatkuva satunnaismuuttuja, koska $G(y)$ on jatkuva ja sillä on jatkuva derivaatta muualla paitsi pisteessä $y = 0$. Y:n tiheysfunktio on

$$g(y) = G'(y) = \begin{cases} -\log(y), & \text{kun } 0 < y < 1; \\ 0 & \text{muualla.} \end{cases}$$

Huomaa, että $-\log(y) > 0$, kun $0 < y < 1$. Nyt siis $g(y) \geq 0$ kaikilla $y \in S_Y = (0, 1)$. □

5.5.2 Muunnos tiheysfunktion avulla

Seuraavaksi esitetään yleinen menetelmä, jonka avulla voidaan johtaa satunnaismuuttujan X funktion $Y = h(X)$ tiheysfunktio suoraan X:n tiheysfunktion $f_X(x)$ avulla. Menetelmän edellyttää kuitenkin, että funktiolla $h(x)$ on tarkasteltavalla välillä *käänteisfunktio.* Esimerkiksi funktion $y = \mathrm{e}^x$ käänteisfunktio on $x = \log(y)$. Myös funktio $y = x^2$ on *kääntyvä,* kun $x > 0$, sillä silloin $x = \sqrt{y}$. Funktio $y = x^2$ *ei ole kääntyvä* koko reaaliakselilla, koska silloin $x = \pm\sqrt{y}$, joka ei ole funktio. Huomattakoon, että jatkuva funktio $h(x)$ on *kääntyvä,* jos ja vain jos se on joko aidosti kasvava tai aidosti vähenevä.

Lineaarinen munnos

Tarkastellaan ensin yksinkertaista lineaarista muunnosta $Y = aX + b$, missä a ja b ovat annettuja vakioita. Nyt siis $h(X) = aX + b$. Funktion $y = h(x)$ derivaatta on

$$\frac{\mathrm{d}y}{\mathrm{d}x} = h'(x) = a.$$

Funktiolla $h(x)$ on käänteisfunktio

$$g(y) = \frac{y - b}{a}, \qquad a \neq 0$$

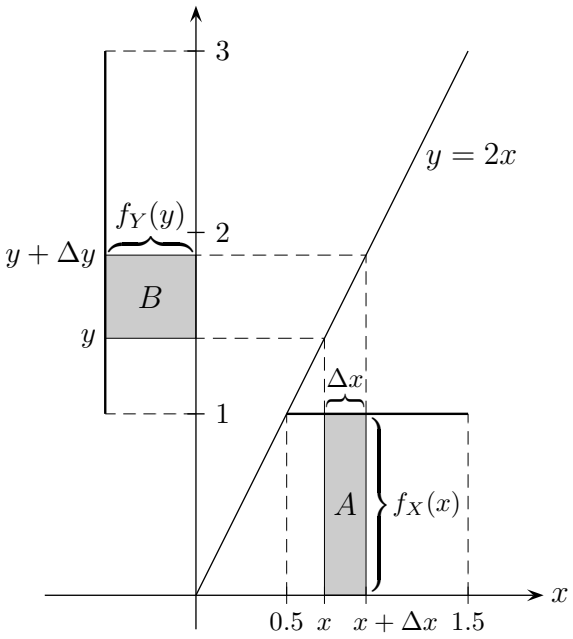
ja

$$\frac{\mathrm{d}y}{\mathrm{d}x} = g'(y) = \frac{1}{a}.$$

Esimerkki 5.14 Oletetaan, että $X \sim \mathrm{Tas}(0.5, 1.5)$ ja $Y = 2X$. Mitä jakaumaa Y noudattaa?

Kuviossa 5.8 on alueen A pinta-ala

$$P[X \in (x, x + \Delta x)] = f_X(x) \cdot \Delta x = \Delta x$$



Kuvio 5.8. Tasajakaumaa Tas$(0.5, 1.5)$ noudattavan satunnaismuuttujan X lineaarinen muunnos.

ja alueen B pinta-ala

$$P[Y \in (y, y + \Delta y)] = f_Y(y) \cdot \Delta y.$$

Tapahtumat $X \in (x, x + \Delta x)$ ja $Y \in (y, y + \Delta y)$ sattuvat täsmälleen samanaikaisesti, joten

$$P[X \in (x, x + \Delta x)] = P[Y \in (y, y + \Delta y)]. \tag{5.5.1}$$

Koska $y = 2x$ ja $y + \Delta y = 2(x + \Delta x)$, niin $\Delta y = 2\Delta x$ ja identiteetistä (5.5.1) seuraa, että $f_Y(y) = \frac{1}{2}$. Koska $0.5 < x < 1.5$, niin $1 < y < 3$. Näin siis $Y \sim$ Tas$(1, 3)$:

$$f_Y(y) = \begin{cases} \frac{1}{2}, & 1 < y < 3; \\ 0, & \text{muualla.} \end{cases}$$

□

Olkoon X satunnaismuuttuja, jonka arvoavaruus on S_X. Silloin satunnaismuuttujan $Y = h(X)$ arvoavaruus S_Y määräytyy siten, että

$$X \in S_X \Leftrightarrow Y \in S_Y.$$

Seuraavassa lauseessa esitettävässä menetelmässä oletetaan, että funktio $y = h(x)$ on tarkasteltavalla arvoalueella kääntyvä. Silloin on olemassa sellainen funktio $x = g(y)$, että

$$y = h(x) \Leftrightarrow x = g(y).$$

Lause 5.5 *Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on $f_X(x)$ ja arvoavaruus S_X. Olkoon $Y = h(X)$ sellainen funktio, että sillä on käänteisfunktio $x = g(y)$ ja käänteisfunktion derivaatta $g'(y)$ on olemassa kaikilla $y \in S_Y$, missä S_Y on Y:n arvoavaruus. Silloin Y:n tiheysfunktio on*

$$f_Y(y) = f_X\big(g(y)\big)|g'(y)|, \qquad y \in S_Y.$$

Todistus. Oletuksen mukaan $g(y)$ on derivoituva, joten se on jatkuva. Koska h ja g ovat kääntyviä, niin h ja g ovat molemmat joko kasvavia tai väheneviä. Oletetaan h ja g ovat väheneviä. Silloin

$$F_Y(y) = P(Y \le y) = P(h(X) \le y) = P(X \ge g(y)) = 1 - F_X\big(g(y)\big).$$

Derivoidaan $1 - F_X[g(y)]$ ketjusäännön avulla, jolloin saadaan

$$\begin{aligned} f_Y(y) = F_Y'(y) &= -F_X'\big(g(y)\big)g'(y) \\ &= -f_X\big(g(y)\big)g'(y) = f_X\big(g(y)\big)|g'(y)|. \end{aligned}$$

Viimeinen yhtäsuuruus seuraa siitä, että $g'(y)$ on negatiivinen, koska g on vähevä.

Jos h ja g ovat kasvavia, niin todistus on melkein samanlainen ja se jätetään harjoitustehtäväksi. □

Esimerkki 5.15 Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on $f_X(x) = \mathrm{e}^{-x}$ ja $S_X = \{\, x \mid x > 0 \,\}$. Olkoon $Y = X^{1/2}$, joten $X = Y^2 = g(Y)$ ja $S_Y = S_X$. Koska $g'(y) = 2y$, niin

$$f_Y(y) = f_X(y^2)|2y| = 2y\mathrm{e}^{-y^2}, \qquad y > 0.$$

Tarkastellaan vielä satunnaismuuttujaa $V = \mathrm{e}^{-X}$. Silloin $X = -\log(V)$. Merkitään nyt $-\log(V) = \tilde{g}(V)$. Silloin $S_V = [0, 1]$ ja $\tilde{g}'(v) = -1/v$. Siksi

$$f_V(v) = f_X[-\log(v)]|\frac{-1}{v}| = \frac{v}{v} = 1,$$

joten V noudattaa tasajakaumaa välillä $[0, 1]$. □

Mikäli muunnosfunktiolla h ei ole käänteisfunktiota X:n arvoavaruudessa S_X, niin Lauseen 5.5 muunnosmenetelmää ei voi suoraan soveltaa. Jos kuitenkin on olemassa sellainen S_X:n ositus yhteispisteettömiin osaväleihin A_1, A_2, …, A_m, että

$$S_X = A_1 \cup A_2 \cup \cdots \cup A_m \tag{5.5.2}$$

ja h on kääntyvä jokaisella osavälillä, voidaan muunnos tehdä jokaisella osavälillä erikseen. Sitä varten määritellään funktiot

$$h(x) = \begin{cases} h_i(x), & \text{kun } x \in A_i; \\ 0 & \text{muualla.} \end{cases}$$

Silloin $h(x)$ voidaan kirjoittaa muodossa $h(x) = \sum_{i=1}^{m} h_i(x)$, missä jokainen $h_i(x)$ on kääntyvä välillä A_i. Olkoot funktioiden h_i käänteisfunktiot vastaavasti g_i, $i = 1, 2, \ldots, m$. Satunnaismuuttujan $Y = h(X)$ tiheysfunktio voidaan nyt esittää Lauseen 5.5 avulla muodossa

$$f_Y(y) = \sum_{i=1}^{m} f_X\big(g_i(y)\big)|g_i'(y)|. \qquad y \in S_Y. \tag{5.5.3}$$

Huomattakoon, että joskus tarvitaan äärellisen osituksen (5.5.2) sijasta ositus, jossa jakovälejä A_1, A_2, ... on ääretön määrä ($m = \infty$).

5.5.3 Normaalimuuttujan muunnokset

Jos $X \sim \mathrm{N}(0,1)$, niin X:n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}, \qquad -\infty < x < \infty,$$

joka on standardimuotoisen normaalijakauman tiheysfunktio. Johdetaan nyt satunnaismuuttujan $U = X^2$ jakauma. Muunnosfunktio $u = h(x) = x^2$ ei ole kääntyvä, koska $x = \pm\sqrt{u}$ ei ole funktio. Siksi esitämme arvoavaruuden $S_X = \{\, -\infty < x < \infty \,\}$ ositettuna muodossa

$$S_X = (-\infty, 0] \cup (0, \infty).$$

Silloin funktiolla $h(x)$ on välillä $(-\infty, 0]$ käänteisfunktio $g_1(u) = -\sqrt{u}$ ja välillä $(0, \infty)$ käänteisfunktio $g_2(u) = \sqrt{u}$. Nyt siis kaavan (5.5.3) mukaan U:n tiheysfunktio on

$$f_U(u) = f_X(-\sqrt{u})\frac{1}{2\sqrt{u}} + f_X(\sqrt{u})\frac{1}{2\sqrt{u}} = \frac{1}{\sqrt{2\pi u}} \mathrm{e}^{-u/2}, \tag{5.5.4}$$

kun $u \in (-\infty, \infty)$. U noudattaa χ^2-jakaumaa vapausastein 1. Käsittelemme tilastotieteessä tärkeää χ^2-jakaumaa vielä jatkossa tarkemmin.

Lause 5.6 *Jos $X \sim \mathrm{N}(\mu, \sigma^2)$, $\sigma^2 > 0$, niin silloin*

$$\frac{(X-\mu)^2}{\sigma^2} \sim \mathrm{Khi2}(1).$$

Todistus. Koska $X \sim \mathrm{N}(\mu, \sigma^2)$, niin määritelmän mukaan $\frac{X-\mu}{\sigma} = Z \sim \mathrm{N}(0,1)$. Edellä näytettiin, että $Z^2 \sim \mathrm{Khi2}(1)$. Näin on lause todistettu. □

Lause 5.7 *Jos Z_i:t ovat riippumattomat ja $Z_i \sim \mathrm{N}(0,1)$, $i = 1, 2, \ldots, n$, niin*

$$Z_1^2 + Z_2^2 + \cdots Z_n^2 \sim \mathrm{Khi2}(n).$$

Jos tehdään otos normaalijakaumasta $\mathrm{N}(0,1)$, niin Lauseen 5.7 mukaan havaintojen neliösumma noudattaa Khi2-jakaumaa vapausastein n, missä n on otoskoko.

Seuraus 5.1 *Jos X_i:t ovat riippumattomat ja $X_i \sim \mathrm{N}(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$, niin*

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \mathrm{Khi2}(n).$$

Jos vastaavasti tehdään n:n suuruinen otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, niin Seurauslauseen 5.1 mukaan standardoitujen havaintojen neliösumma noudattaa Khi2-jakaumaa vapausastein n.

Lause 5.8 *Olkoot X_1 ja X_2 riippumattomat ja $X_i \sim \mathrm{Khi2}(n_i)$, $i = 1, 2$. Silloin*

$$X_1 + X_2 \sim \mathrm{Khi2}(n_1 + n_2).$$

5.6 Satunnaismuuttujan funktion odotusarvo

Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio $f(x)$ on määritelty arvoavaruudessa S. Olkoon $h(X)$ satunnaismuuttujan X reaaliarvoinen funktio, joka siis määrittelee uuden satunnaimuuttujan.

Määritelmä 5.3 Jos X on jatkuva satunnaismuuttuja, niin satunnaismuuttujan $h(X)$ odotusarvo on

$$E[h(X)] = \int_S h(x) f(x)\,\mathrm{d}x, \tag{5.6.1}$$

mikäli $E\big(|h(X)|\big) < \infty$. Jos $E\big(|h(X)|\big) = \infty$, niin sanomme, että $E[h(X)]$ ei ole olemassa.

Huomautus 5.1 Odotusarvon $E[h(X)]$ olemassaolo tarkoittaa siis sitä, että funktion $|h(X)|$ odotusarvo on äärellinen. Jos X noudattaa esimerkiksi eksponenttijakaumaa keskiarvolla 1, niin $f(x) = \mathrm{e}^{-x}$ ja $S = [0, \infty)$. Silloin X:n odotusarvo on

$$\begin{aligned} E(X) &= \int_0^\infty x\mathrm{e}^{-x}\,\mathrm{d}x \\ &= \int_0^\infty (-x\mathrm{e}^{-x}) + \int_0^\infty \mathrm{e}^{-x}\,\mathrm{d}x \qquad \text{(osittaisintegrointi)} \\ &= \int_0^\infty \mathrm{e}^{-x}\,\mathrm{d}x = 1, \end{aligned}$$

joten odotusarvo on olemassa. Hyvin usein odotusarvot ovat epäoleellisia integraaleja, niin kuin tässäkin esimerkissä.

Jos $h(X)$ integroituu itseisesti, eli

$$\int_S |h(x)|$$

on äärellisenä olemassa, niin $E[h(X)]$ on olemassa. Funktio $V = h(X)$ on satunnaismuuttuja, jonka tiheysfunktio $g(v)$ on määritelty arvoavaruudessa $S_V = \{\, v \mid v = h(x),\ x \in S \,\}$. Silloin

$$E[h(X)] = E(V) = \int_{S_V} v g(v).$$

Esimerkki 5.16 Tarkastellaan nyt *Cauchyn jakaumaa* noudattavaa satunnaismuuttujaa X, jonka tiheysfunktio on

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty. \tag{5.6.2}$$

Kaava (5.6.2) todellakin määrittelee tiheysfunktion, koska

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}\,\mathrm{d}x = \frac{2}{\pi} \int_0^{\infty} \arctan(x) = \frac{2}{\pi} \cdot \frac{\pi}{2} = 1.$$

Osoitamme nyt, että $E(|X|) = \infty$, mistä seuraa, että *Cauchyn jakaumalla ei ole keskiarvoa.* Symmetrian nojalla voidaan kirjoittaa

$$E(|X|) = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}\,\mathrm{d}x = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x.$$

Jokaista reaalilukua $M > 0$ kohti saadaan

$$\int_0^M \frac{x}{(1+x^2)}\,\mathrm{d}x = \int_0^M \frac{\log(1+x^2)}{2} = \frac{\log(1+M^2)}{2}.$$

Tästä seuraa, että

$$E(|X|) = \lim_{M\to\infty} \frac{2}{\pi} \int_0^M \frac{x}{1+x^2}\,\mathrm{d}x = \frac{1}{\pi} \lim_{M\to\infty} \log(1+M^2) = \infty,$$

joten $E(X)$ ei ole olemassa. □

Taulukko 5.1. Tärkeitä odotusarvoja.

$h(x)$	$E[h(X)]$	Merkintä	Nimitys
x	$E(X)$	μ	odotusarvo
x^r	$E(X^r)$	α_r	r. momentti
$x^{(r)}$	$E[X^{(r)}]$	g_r	r. tekijämomentti
$(x-\mu)^2$	$E[(X-\mu)^2]$	σ^2	varianssi
$(x-\mu)^r$	$E[(X-\mu)^r]$	μ_r	r. keskusmomentti

5.6.1 Momentifunktio ja momentit

Kun $h(X) = X^r$, niin $E[h(X)] = E(X^r)$ on X:n r. momentti. Jatkuvien satunnaismuuttujien momentit määritellään vastaavasti kuin diskreettien satunnaismuuttujien momentit. Summalausekkeet vain korvataan integraaleilla. Taulukossa 5.1 esitetään yhteenveto eri momenteista

Momenttifunktio määriteltiin 3. luvussa (Määritelmä 3.12) ja jatkuville satunnaismuuttujille alaluvussa 5.1 [ks. identiteetti (5.1.5)]. Jatkuvan satunnaismuuttujan X momentifunktio on

$$M(t) = E(\mathrm{e}^{tX}) = \int_S \mathrm{e}^{tx} f(x)\,\mathrm{d}x, \qquad t \in A,$$

missä $f(x)$ on X:n tiheysfunktio ja A sellainen t:n arvojen joukko, että $M(t)$ on äärellinen kaikilla $t \in A$. Koska $M(0) = 1$, niin $0 \in A$. Sanomme, että $M(t)$ on olemassa, jos $(-a, a) \subset A$ jollakin $a > 0$. Momenttifunktion perusominaisuudet esitettiin Pykälässä 3.5.2.

Esimerkki 5.17 Huomautuksessa 5.1 laskettiin odotusarvo $E(X)$, kun $X \sim$Exp(1). Silloin X:n tiheysfunktio on $f(x) = \mathrm{e}^{-x} \geq 0$ välillä $S = [0, \infty)$ ja $f(x) = 0$ muualla. Kaikki momentit $E(X^r)$ voidaan määrittää osittaisintegroinnilla, mutta käytetämme nyt momenttifunktiota, joka on

$$M(t) = E(\mathrm{e}^{tX}) = \int_0^\infty \mathrm{e}^{tx}\mathrm{e}^{-x}\,\mathrm{d}x = \frac{1}{1-t}, \qquad t < 1.$$

Derivoimalla $M(t)$ toistuvasti r kertaa saadaan $M^{(r)}(t) = \frac{r!}{(1-t)^{k+1}}$. Siksi

$$E(X^r) = M^{(r)}(0) = r!,$$

joten

$$\mu = E(X) = 1, \qquad E(X^2) = 2, \qquad \sigma^2 = E(X^2) - \mu^2 = 1.$$

□

Erityisesti keskiarvo μ, varianssi σ^2 ja hajonta $\sigma = \sqrt{\mathrm{Var}(X)}$ ovat tavallisimmat tunnusluvut, joilla jakaumaa luonnehditaan. Jakauman yksityiskohtaisemmassa tarkastelussa voidaan käyttää myös korkeampia momentteja, mikäli ne ovat olemassa.

Vinous ja huipukkuus

Satunnaismuuttujan 1. momentti μ määrittää jakauman sijainnin. Keskistetyn muuttujan $X - \mu$ toinen momentti (keskusmomentti) on varianssi σ^2 ja se mittaa todennäköisyysmassan hajaantumista. Normeeratun muuttujan $(X - \mu)/\sigma$ kolmas ja neljäs momentti luonnehtivat jakauman muotoa.

Jakauman *vinouskerroin,* josta käytetään merkintää γ_1, määritellään seuraavasti:

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}, \tag{5.6.3}$$

missä μ_3 on jakauman 3. keskusmomentti ja $\sigma = \sqrt{\mathrm{Var}(X)}$ on hajonta. Olkoon X:n tiheysfunktio $f(x)$. Silloin X:n jakauma on *symmetrinen pisteen* a suhteen, jos

$$f(a - x) = f[-(a - x)]$$

kaikilla x:n arvoilla. Jos $E(X)$ on olemassa, niin silloin $E(X) = a$. Symmetrisen jakauman vinouskerroin on nolla. Jos jakaumalla on pitkä häntä oikealle, kuten Poissonin jakaumalla ja geometrisella jakaumalla, niin jakauma on positiivisesti vino ja $\gamma_1 > 0$. Jos jakaumalla on pitkä häntä vasemmalle, niin $\gamma_1 < 0$. Jakaumalla on tietysti oltava 3. momentti, jotta vinouskerroin voidaan laskea. Huomaa, että Cauchyn jakauma, jonka tiheysfunktio on

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty,$$

on symmetrinen pisteen $a = 0$ suhteen, mutta 0 ei ole jakauman keskiarvo, koska jakaumalla ei ole keskiarvoa (ks. Esimerkki 5.16). Cauchyn jakauman vinouskerrointa ei voida laskea, vaikka määritelmän nojalla voimme todeta jakauman olevan symmetrinen.

Huipukkuuskerrointa merkitään γ_2 ja se määritellään normeeratun muuttujan 4. momentin avulla seuraavasti:

$$\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} - 3, \tag{5.6.4}$$

missä μ_4 on X:n 4. keskusmomentti. Normaalijakauman huipukkuus on 0. Jos jakaumalla on paksummat hännät kuin normaalijakaumalla, niin silloin $\gamma_2 > 0$. Jos hännät ovat ohuemmat kuin normaalijakaumalla, niin $\gamma_2 < 0$.

5.7 Kaksiulotteiset jakaumat

Tarkastellaan nyt kahden jatkuvan satunnaismuuttujan yhteisjakaumaa. Yleistys usean muuttujan tapaukseen on sen jälkeen suoraviivainen.

Määritelmä 5.4 Olkoot X ja Y samassa otosavaruudessa määritellyt jatkuvat satunnaismuuttujat. Olkoon kaksiulotteisen jatkuvan satunnaismuuttujan (X, Y) arvoavaruus S. Funktio $f(x, y)$ on (X, Y):n tiheysfunktio (X:n ja Y:n yhteisjakauman tiheysfunktio), jos sillä on seuraavat ominaisuudet:

1. $f(x, y) \geq 0 \qquad$ kaikilla $\quad (x, y) \in \mathbb{R}^2$,
2. $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1 \qquad$ ja
3. $$P[(X, Y) \in A] = \iint\limits_{(x,y)\in A} f(x, y)\,\mathrm{d}x\,\mathrm{d}y,$$

   missä $(X, Y) \in A$ on tasossa määritelty tapahtuma.

Esimerkki 5.18 Olkoon X:n ja Y:n yhteisjakauman tiheysfunktio

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \qquad -1 < x < 1, \quad -1 < y < 1.$$

Määritellään $A = \{\, (x, y) \mid 0 < x < 1,\ 0 < y < x \,\}$. Todennäköisyys, että $(X, Y) \in A$, on

$$\begin{aligned} P[(X, Y) \in A] &= \int_0^1 \int_0^x \frac{3}{2}x^2(1 - y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 \frac{3}{2}x^2 \Big/_0^x \left(y - \frac{y^2}{2}\right)\mathrm{d}x \\ &= \int_0^1 \frac{3}{2}\left(x^3 - \frac{x^4}{2}\right)\mathrm{d}x = \frac{3}{2} \Big/_0^1 \left(\frac{x^4}{4} - \frac{x^5}{10}\right) = \frac{9}{40}. \end{aligned}$$

□

5.7.1 Reunajakauma ja ehdollinen jakauma

Kaksiulotteista satunnaismuuttujaa (X, Y) kutsutaan kaksiulotteiseksi satunnaisvektoriksi. Silloin X ja Y ovat tietysti (yksiulotteisia) satunnaismuuttujia. X:n *reunajakauman tiheysfunktio,* jota merkitään $f_X(x)$, on pelkästään X:n tiheysfunktio, jossa Y:tä ei oteta huomioon. Satunnaismuuttujan X *ehdollinen tiheysfunktio* ehdolla $Y = y$ on on X:n tiheysfunktio, kun Y:n arvo tunnetaan. X:n ehdollista tiheysfunktiota ehdolla $Y = y$ merkitään $f_X(x \mid Y = y)$ tai lyhyesti $f_X(x \mid y)$.

Määritelmä 5.5 Olkoon $f(x, y)$ jatkuvan satunnaisvektorin (X, Y) tiheysfunktio ja S sen arvoavaruus. Silloin satunnaismuuttujat X ja Y ovat jatkuvia ja niiden reunajakaumien tiheysfunktiot ovat

$$f_X(x) = \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y, \quad x \in S_X; \qquad f_Y(y) = \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x, \quad y \in S_Y,$$

missä S_X on X:n ja S_Y on Y:n arvoavaruus. Satunnaismuuttujat X *ja* Y *ovat riippumattomat* jos ja vain jos

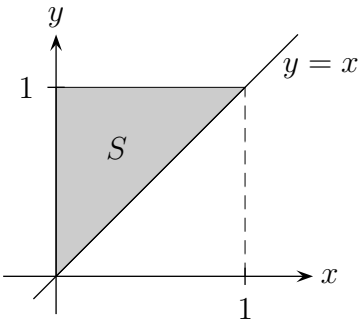
$$f(x,y) = f_X(x) f_Y(y) \qquad \text{kaikilla } x \in S_X \text{ ja } y \in S_Y; \tag{5.7.1}$$

muutoin X *ja* Y *riippuvat toisistaan.*

Esimerkki 5.19 Olkoon X:n ja Y:n yhteisjakauman tiheysfunktio

$$f(x,y) = 2, \qquad 0 \le x \le y \le 1,$$

muualla $f(x,y) = 0$. Satunnaisvektorin (X,Y) arvoavaruus on $S = \{\,(x,y) \mid 0 \le x \le y \le 1\,\}$.



$$S = \{\,(x,y) \mid 0 \le x \le y \le 1\,\}$$

Kuvio 5.9. Tasajakauman $f(x,y) = 2$ määrittelyalue S.

Silloin esimerkiksi todennäköisyys

$$P\left(0 \le X \le \frac{1}{2},\ 0 \le Y \le \frac{1}{2}\right) = P\left(0 \le X \le Y,\ 0 \le Y \le \frac{1}{2}\right)$$
$$= \int_0^{1/2} \int_0^{y} 2 \,\mathrm{d}y \,\mathrm{d}x = \int_0^{1/2} 2y \,\mathrm{d}y = \frac{1}{4}.$$

Reunajakaumien tiheysfunktiot ovat

$$f_X(x) = \int_x^1 2 \,\mathrm{d}y = 2(1-x), \qquad 0 \le x \le 1,$$

ja

$$f_Y(y) = \int_0^y 2 \,\mathrm{d}x = 2y, \qquad 0 \le y \le 1.$$

Lasketaan vielä X:n ja Y:n odotusarvot sekä Y:n 2. momentti.

$$E(X) = \int_0^1 \int_x^1 2x \,\mathrm{d}y \,\mathrm{d}x = \int_0^1 2x(1-x) \,\mathrm{d}x = \frac{1}{3},$$

$$E(Y) = \int_0^1 \int_0^y 2y \,\mathrm{d}x \,\mathrm{d}y = \int_0^1 2y^2 \,\mathrm{d}y = \frac{2}{3},$$

$$E(Y^2) = \int_0^1 \int_0^y 2y^2 \,\mathrm{d}x \,\mathrm{d}y = \int_0^1 2y^3 \,\mathrm{d}y = \frac{1}{2}.$$

Odotusarvot $E(X)$, $E(Y)$ ja $E(Y)$ voidaan laskea joko suoraan reunajakaumasta tai sitten yhteisjakaumasta. □

Nähdään helposti, että Esimerkissä 5.19 satunnaismuuttujat X ja Y eivät ole riippumattomat, koska

$$f_X(x) f_Y(y) = 2(1-x)2y \neq f(x,y) = 2, \qquad (x,y) \in S.$$

Sen sijaan voidaan osoittaa, että Esimerkissä 5.18 satunnaismuuttujat X ja Y ovat riippumattomat.

Jatkuvan satunnaismuuttujan ehdollinen tiheysfunktio määritellään seuraavasti:

Määritelmä 5.6 Jos jatkuvan satunnaisvektorin (X,Y) tiheysfunktio on $f(x,y)$ ja arvoavaruus S, niin *X:n ehdollinen tiheysfunktio ehdolla $Y = y$* on

$$f_X(x \mid y) = \frac{f(x,y)}{f_Y(y)}, \qquad (x,y) \in S$$

ja *Y:n ehdollinen tiheysfunktio ehdolla $X = x$* on

$$f_Y(y \mid x) = \frac{f(x,y)}{f_X(x)}, \qquad (x,y) \in S.$$

Huomattakoon, että Määritelmässä 5.6 oletetaan, että $f_Y(y) > 0$ ja $f_X(x) > 0$.

Esimerkki 5.20 Olkoot satunnaismuuttujat X ja Y samat kuin Esimerkissä 5.19 Silloin

$$\begin{aligned} f(x,y) &= 2, \qquad 0 \leq x \leq y \leq 1, \\ f_X(x) &= 2(1-x), \qquad 0 \leq x \leq 1, \\ f_Y(y) &= 2y, \qquad 0 \leq y \leq 1. \end{aligned}$$

Määritetään nyt Y:n ehdollisen jakauman tiheysfunktio, kun $X = x$ on annettu. Määritelmän 5.6 mukaan

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \qquad x \leq y \leq 1, \quad 0 \leq x \leq 1.$$

Y:n ehdollinen odotusarvo ehdolla $X = x$ on

$$E(Y \mid x) = \int_x^1 y \frac{1}{1-x}\,\mathrm{d}y = \int_x^1 \frac{y^2}{2(1-x)} = \frac{1+x}{2}, \qquad 0 \le x \le 1.$$

Samalla tavalla voidaan osoittaa, että

$$E(X \mid y) = \frac{y}{2}, \qquad 0 \le y \le 1.$$

Suoraan määritelmän perusteella Y:n ehdollinen varianssi ehdolla $X = x$ on

$$\begin{aligned} E\big([Y - E(Y \mid x)]^2 \mid x\big) &= \int_x^1 \left(y - \frac{1+x}{2}\right)^2 \frac{1}{1-x}\,\mathrm{d}y \\ &= \int_x^1 \frac{1}{3(1-x)} \left(y - \frac{1+x}{2}\right)^3 \\ &= \frac{(1-x)^2}{12}. \end{aligned}$$

Jos $U \sim \mathrm{Tas}(a,b)$, niin $E(U) = \frac{a+b}{2}$ ja $\mathrm{Var}(U) = \frac{(b-a)^2}{12}$. Koska Y:n ehdollinen jakauma ehdolla $X = x$ on $\mathrm{Tas}(x,1)$, niin olisimme voineet tasajakauman ominaisuuksien perusteella suoraan todeta, että

$$E(Y \mid x) = \frac{x+1}{2} \qquad \text{ja} \qquad \mathrm{Var}(Y \mid x) = \frac{(1-x)^2}{12}.$$

Lasketaan vielä ehdollinen todennäköisyys

$$P(3/4 \le Y \le 7/8 \mid X = 1/4) = \int_{3/4}^{7/8} f(y \mid 1/4)\,\mathrm{d}y = \int_{3/4}^{7/8} \frac{1}{3/4}\,\mathrm{d}y = \frac{1}{6}.$$

□

Havaitsimme edellisessä esimerkissä, että Y:n ehdollinen odotusarvo on x:n lineaarinen funktio:

$$E(Y \mid x) = \frac{1}{2} + \frac{1}{2}x, \qquad 0 \le x \le 1.$$

Jos $E(Y \mid x)$ on lineaarinen, niin pitää yleisesti paikkansa, että

$$E(Y \mid x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

missä $\rho = \mathrm{Cor}(X,Y)$ on X:n ja Y:n välinen korrelaatio, σ_X on X:n hajonta ja σ_Y on Y:n hajonta. Jos $E(X \mid y)$ on lineaarinen, niin

$$E(X \mid y) = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y).$$

Ehdollisten odotusarvojen $E(Y \mid x)$ ja $E(X \mid y)$ yhtälöissä kertoimien $\rho\frac{\sigma_Y}{\sigma_X}$ ja $\rho\frac{\sigma_X}{\sigma_Y}$ tulo on ρ^2. Esimerkissä 5.20 näiden kertoimien tulo on $\rho^2 = \frac{1}{4}$. Siksi $\rho = \frac{1}{2}$, koska molemmat kertoimet ovat positiiviset. Näiden kertoimien suhde on σ_Y^2/σ_X^2 ja esimerkissä tämä suhde on 1. Tästä voimme päätellä, että Esimerkissä 5.20 $\sigma_X^2 = \sigma_Y^2$.

Satunnaismuuttujien X ja Y riippumattomuuden tarkistaminen suoraan relaation (5.7.1) perusteella edellyttää reunajakaumien tiheysfunktioiden $f_X(x)$ ja $f_Y(y)$ tuntemista. Seuraava apulause tekee riippumattomuuden tarkistamisen jonkin verran helpommaksi, koska siinä ei edellytetä reunajakaumien tuntemista.

Apulause 5.1 *Olkoon (X, Y) kaksiulotteinen satunnaisvektori, jonka yhteisjakauman tiheysfunktio on $f(x, y)$. Silloin satunnaismuuttujat X ja Y ovat riippumattomat, jos ja vain jos on olemassa sellaiset funktiot $g(x)$ ja $h(y)$, että*

$$f(x, y) = g(x)h(y) \qquad \textit{kaikilla } x \in \mathbb{R} \textit{ ja kaikilla } y \in \mathbb{R},$$

missä g riippuu vain x:stä ja h vain y:stä.

Kertymäfunktio

Kaksiulotteinen jakauma voidaan täydellisesti luonnehtia kertymäfunktionsa avulla. Satunnaisvektorin (X, Y) yhteisjakauman kertymäfunktio $F(x, y)$ määritellään relaatiolla

$$F(x, y) = P(X \leq x,\ Y \leq y)$$

missä $(x, y) \in \mathbb{R}^2$. Tiheysfunktion avulla lausuttuna kertymäfunktio on

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\,\mathrm{d}s\,\mathrm{d}t.$$

Integraalilaskennan peruslause kahden muuttujan tapauksessa sanoo, että

$$\frac{\partial^2 F(x, y)}{\partial x\,\partial y} = f(x, y) \tag{5.7.2}$$

kaikissa $f(x, y)$:n jatkuvuuspisteissä. Relaatio (5.7.2) on hyödyllinen silloin, kun kertymäfunktio tunnetaan ja halutaan johtaa tiheysfunktio. Silloin tiheysfunktio $f(x, y)$ saadaan derivoimalla $F(x, y)$ sekä x:n että y:n suhteen eli laskemalla osittaisderivaatta $\frac{\partial^2 F(x,y)}{\partial x\,\partial y}$.

Esimerkki 5.21 Olkoon X:n ja Y:n yhteisjakauman kertymäfunktio

$$F(x, y) = \begin{cases} xy, & 0 \leq x \leq 1 \text{ ja } 0 \leq y \leq 1; \\ y, & x > 1,\ 0 \leq y \leq 1; \\ x, & y > 1,\ 0 \leq x \leq 1; \\ 1, & x > 1 \text{ ja } y > 1; \\ 0, & x < 0 \text{ tai } y < 0. \end{cases}$$

Laskemalla osittaisderivaatta $\frac{\partial^2 F(x,y)}{\partial x\,\partial y}$ saadaan

$$f(x,y) = \begin{cases} 1, & 0 \le x \le 1,\ 0 \le y \le 1; \\ 0 & \text{muualla.} \end{cases}$$

Satunnaisvektori (X,Y) noudattaa siis kaksiulotteista tasajakaumaa $\mathrm{Tas}[(0,1)\times(0,1)]$. Todennäköisyys voidaan lausua kertymäfunktion avulla seuraavasti:

$$\begin{aligned} &P(x_1 \le X \le x_2,\ y_1 \le Y \le y_2) \\ &\qquad = \int_{x_1}^{x_2}\int_{y_1}^{y_2} \mathrm{d}y\,\mathrm{d}x = \Big/_{x_1}^{x_2}\Big/_{y_1}^{y_2} xy = (x_2 - x_1)(y_2 - y_1) \\ &\qquad = x_2y_2 - x_2y_1 - x_1y_2 + x_1y_1 \\ &\qquad = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1). \end{aligned}$$

Yleisesti pitää paikkansa, että

$$\begin{aligned} &P(x_1 \le X \le x_2,\ y_1 \le Y \le y_2) \\ &\qquad = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1). \end{aligned}$$

Kahden muuttujan tasajakauman $\mathrm{Tas}[(0,1)\times(0,1)]$ tapauksessa todennäköisyys $P\big(\frac{1}{4} \le X \le \frac{1}{2},\ \frac{1}{2} \le Y \le \frac{3}{4}\big)$ on

$$\begin{aligned} &F\left(\frac{1}{2},\frac{3}{4}\right) - F\left(\frac{1}{2},\frac{1}{2}\right) - F\left(\frac{1}{4},\frac{3}{4}\right) + F\left(\frac{1}{4},\frac{1}{2}\right) \\ &\qquad = \frac{1}{2}\cdot\frac{3}{4} - \frac{1}{2}\cdot\frac{1}{2} - \frac{1}{4}\cdot\frac{3}{4} + \frac{1}{4}\cdot\frac{1}{2} = \frac{1}{16}. \end{aligned}$$

□

5.7.2 Yhteisjakauman momenttifunktio

Kaksiulotteisen diskreetin satunnaisvektorin momenttifunktio määriteltiin alaluvussa 4.7.2. Jatkuvien satunnaismuuttujien X_1 ja X_2 yhteisjakauman eli jatkuvan satunnaisvektorin (X_1,X_2) jakauman momenttifunktio määritellään samalla tavalla kuin diskreetissä tapauksessa. Olkoon (X_1,X_2) jatkuva satunnaisvektori ja $t_1X_1 + t_2X_2$ satunnaismuuttujien X_1 ja X_2 lineaarinen yhdiste, missä $t_1,t_2 \in \mathbb{R}$. Satunnaisvektorin (X_1,X_2) jakauman momenttifunktio on

$$M(t_1,t_2) = E\big(\mathrm{e}^{t_1X_1+t_2X_2}\big).$$

Jatkuvien satunnaismuuttujien tapauksessa odotusarvon lauseke on muotoa

$$E\big(\mathrm{e}^{t_1X+t_2X_2}\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{t_1x_1+t_2x_2} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2.$$

Merkitään

$$M_i(t_1,t_2) = \frac{\partial M(t_1,t_2)}{\partial t_i},$$
$$M_{ii}(t_1,t_2) = \frac{\partial^2 M(t_1,t_2)}{\partial t_i^2},$$
$$M_{ij}(t_1,t_2) = \frac{\partial^2 M(t_1,t_2)}{\partial t_i\,\partial t_j},$$

missä $M_i(t_1,t_2)$ on M:n osittaisderivaatta t_i:n suhteen, $M_{ii}(t_1,t_2)$ on M:n 2. osittaisderivaatta t_i:n suhteen ja $M_{ij}(t_1,t_2)$ on osittaisderivaatta t_i:n ja t_j:n suhteen ($i=1,2$; $j=1,2$). Esitämme nyt seuraavassa lauseessa, miten momenttifunktio generoi satunnaisvektorin momentit.

Lause 5.9 *Oletetaan, että satunnaisvektorilla (X_1,X_2) on momenttifunktio. Silloin $E(X_i)$, $E(X_i^2)$ ja $E(X_iX_j)$ ovat äärelliset ja*

$$E(X_i) = M_i(0,0), \qquad E(X_i^2) = M_{ii}(0,0), \qquad E(X_iX_j) = M_{ij}(0,0)$$

kaikilla $i=1,2$ ja $j=1,2$.

Esimerkiksi X_1:n odotusarvo saadaan derivoimalla ensin momenttifunktio t_1:n suhteen ja sijoittamalla sitten derivaatan lausekkeeseen $t_1=0$ ja $t_2=0$. Sekamomentti $E(X_1X_2)$ saadaan määrittämällä toisen kertaluvun osittaisderivaatta $M_{12}(t_1,t_2)$ (derivoidaan momenttifunktio t_2:n ja t_1:n suhteen) ja laskemalla osottaisderivaatan arvo $M_{ij}(0,0)$ pisteessä $(t_1,t_2)=(0,0)$.

Esimerkki 5.22 Jos Z_1 ja Z_2 ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa, niin (Z_1,Z_2) noudattaa kaksiulotteista standardimuotoista normaalijakaumaa. (Z_1,Z_2):n momenttifunktio on

$$\begin{aligned} M(t_1,t_2) &= E\big(\mathrm{e}^{t_1Z_1+t_2Z_2}\big) = E(\mathrm{e}^{t_1Z_1})\,E(\mathrm{e}^{t_2Z_2}) \\ &= \mathrm{e}^{t_1^2/2}\mathrm{e}^{t_2^2/2} = \mathrm{e}^{(t_1^2+t_2^2)/2}. \end{aligned}$$

Tässä tapauksessa $M_1(t_1,t_2) = t_1\mathrm{e}^{(t_1^2+t_2^2)/2}$, joten $E(X_1) = M_1(0,0) = 0$. Vastaavasti $M_{11} = \mathrm{e}^{(t_1^2+t_2^2)/2} + t_1\mathrm{e}^{(t_1^2+t_2^2)/2}$ ja $E(X_1^2) = M_{11}(0,0) = 1$. □

Huomattakoon, että myös satunnaisvektoreiden tapauksessa pätee momenttifunktioden yksikäsitteisyyttä koskeva lause (vrt. Lause 3.12). Jos siis satunnaisvektoreilla (X_1,X_2) ja (Y_1,Y_2) on sama momenttifunktio, niin niillä on sama jakauma. Reunajakaumien momenttifunktiot saadaan kätevästi yhteisjakuman momenttifunktiosta.

Lause 5.10 *Oletetaan, että satunnaisvektorin (X,Y) momenttifunktio on $M(s,t)$ sekä X:n ja Y:n momenttifunktiot vastaavasti $M_X(s)$ ja $M_Y(t)$.*

1. *Silloin*
$$M_X(s) = M(s,0) \qquad ja \qquad M_Y(t) = M(0,t).$$
2. *X ja Y ovat riippumattomat jos ja vain jos*
$$M(s,t) = M_X(s)M_Y(t).$$

5.8 Kahden muuttujan normaalijakauma

5.8.1 Standardimuoto

Oletetaan, että satunnaismuuttujat Z ja V ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa. Silloin Z:n ja V:n riippumattomuuden nojalla satunnaisvektorin (Z, V) yhteisjakauman tiheysfunktio on

$$f_{Z,V}(z, v) = f(z)f(v) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-v^2/2} = \frac{1}{2\pi}\mathrm{e}^{-(z^2+v^2)/2}. \tag{5.8.1}$$

Sanomme, että satunnaisvektori (Z, V) noudattaa kaksiulotteista standardimuotoista normaalijakaumaa ja funktio (5.8.1) on tämän jakauman tiheysfunktio. Merkitään $(Z, V) \sim \mathrm{N}_2(\mathbf{0}, \boldsymbol{I})$, missä $\mathbf{0}$ on 2×1-nollavektori eli $\mathbf{0} = (0, 0)^\mathsf{T}$. Merkintä $(0, 0)^\mathsf{T}$ tarkoittaa vektorin $(0, 0)$ transponointia, joka muuntaa vaakavektorin $(0, 0)$ pystytoriksi. Matriisi

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

on 2×2-identiteettimatriisi. Yhteisjakauman reunajakaumien keskiarvot ovat $E(Z) = E(V) = 0$ ja varianssit $\mathrm{Var}(Z) = \mathrm{Var}(V) = 1$ sekä $\mathrm{Cov}(Z, V) = 0$. Satunnaisvektorin (Z, V) odotusarvovektori on $[E(Z), E(V)]^\mathsf{T} = \mathbf{0}$ ja kovarianssimatriisi

$$\begin{pmatrix} \mathrm{Var}(Z) & \mathrm{Cov}(Z, V) \\ \mathrm{Cov}(V, Z) & \mathrm{Var}(V) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Huomaa, että aina $\mathrm{Cov}(Z, V) = \mathrm{Cov}(V, Z)$, joten kovarianssimatriisi on symmetrinen. Voidaan merkitä myös $(Z, V) \sim \mathrm{N}_2(0, 0; 1, 1, 0)$, missä odotusarvot, varianssit ja korrelaatio on annettu sulkeissa.

5.8.2 Korreloivat muuttujat

Oletetaan, että $X \sim \mathrm{N}(0, 1)$ ja $Z \sim \mathrm{N}(0, 1)$ ovat riippumattomat. Niiden avulla voidaan konstruoida normaalijakaumaa noudattava satunnaismuuttuja Y siten, että X ja Y korreloivat. Kiertämällä x-akselia kulman θ verran vastapäivään saadaan y-akseli (Kuvio 5.10). Projisoidaan satunnaispiste (X, Z) y-akselille ja merkitään tätä projektiota Y:llä. On helppo todeta geometrisen päättelyn avulla (Kuvio 5.10), että

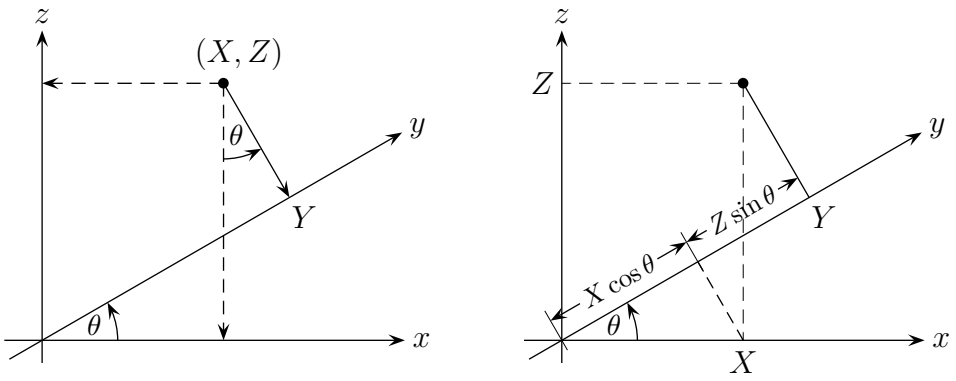
$$Y = X\cos\theta + Z\sin\theta.$$

Satunnaismuuttuja Y saadaan siis X:n ja Z:n lineaarisena muunoksena. Tästä seuraa, että

$$E(Y) = \cos\theta \cdot E(X) + \sin\theta \cdot E(Z) = 0$$

ja

$$\mathrm{Var}(Y) = \cos^2\theta \cdot \mathrm{Var}(X) + \sin^2\theta \cdot \mathrm{Var}(Z) = 1,$$



Kuvio 5.10.

koska $E(X) = E(Z) = 0$ ja $\mathrm{Var}(X) = \mathrm{Var}(Z) = 1$. Lauseen 5.4 mukaan $Y \sim \mathrm{N}(0,1)$. Satunnaisuuttujien X ja Y 2. kertaluvun sekamomentti on

$$\begin{aligned} E(XY) &= E[X(X\cos\theta + Z\sin\theta)] \\ &= \cos\theta \cdot E(X^2) + \sin\theta \cdot E(XZ) \\ &= \cos\theta. \end{aligned}$$

Viimeinen yhtäsuuruus seuraa siitä, että $E(X^2) = 1$ ja $E(XZ) = E(X)\,E(Z) = 0$. Satunnaisuuttujien X ja Y välinen korrelaatio $\mathrm{Cor}(X,Y) = E(X,Y)$, koska $E(X) = E(Y) = 0$ ja $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

5.9 Satunnaisvektoreiden muunnokset

Oletetaan, että jatkuvien satunnaismuuttujien X ja Y yhteisjakauman tiheysfunktio on f. Olkoon

$$U = h_1(X,Y); \qquad V = h_2(X,Y) \tag{5.9.1}$$

sellainen satunnaisvektorin (X,Y) muunnos, että sillä on käänteismuunnos. Silloin mitä tahansa satunnaisvektorin (U,V) arvoa $(u,v) \in \mathbb{R}^2$ vastaa yksikäsitteinen satunnaisvektorin (X,Y) arvo $(x,y) \in \mathbb{R}^2$. Voimme silloin määritellä käänteiskuvauksen

$$x = g_1(u,v); \qquad y = g_2(u,v).$$

Vektoreiden $(u,v) \in \mathbb{R}^2$ ja $(x,y) \in \mathbb{R}^2$ välillä on yksi-yksinen vastaavuus. Oletetamme lisäksi, että funktioilla g_1 ja g_2 on jatkuvat osittaisderivaatat. Yksiulotteisen muunnoksen tapauksessa laskettavaa derivaattaa g' vastaa satunnaisvektorien muunnoksen *Jacobin determinantti,* joka on funktioiden g_1 ja g_2 osittaisderivaattojen matriisin determinantti. Jacobin determinanttia kutsutaan muunnoksen *Jakobiaaniksi.*

Muunnoksen (5.9.1) Jakobiaani on

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}, \tag{5.9.2}$$

missä

$$\frac{\partial x}{\partial u} = \frac{\partial g_1(u,v)}{\partial u}, \qquad \frac{\partial x}{\partial v} = \frac{\partial g_1(u,v)}{\partial v},$$
$$\frac{\partial y}{\partial u} = \frac{\partial g_2(u,v)}{\partial u}, \qquad \frac{\partial y}{\partial v} = \frac{\partial g_2(u,v)}{\partial v}.$$

Jacobin determinanttia merkitään $J = \frac{\partial(x,y)}{\partial(u,v)}$. Oletetaan, että $J \neq 0$, kun $f(x,y) > 0$. Satunnaisvektorin (U,V) yhteisjakauman tiheysfunktio on

$$f_{U,V}(u,v) = f\big(g_1(u,v), g_2(u,v)\big)|J|. \tag{5.9.3}$$

Satunnaisvektorin (U,V) arvoavaruus $S_{U,V}$ saadaan tarkastelemalla kuvausta (5.9.1), joka kuvaa satunnaisvektorin (X,Y) arvojoukon $S_{X,Y}$ kuvajoukoksi $S_{U,V}$.

Esimerkki 5.23 Olkoot X ja Y jatkuvat satunnaismuuttujat, joiden yhteisjakauman tiheysfunktio on f. Määritellään satunnaismuuttujat U ja V siten, että

$$U = X + Y; \qquad V = X - Y. \tag{5.9.4}$$

Johdetaan nyt (U,V):n jakauman tiheysfunktio. Muunnoksen (5.9.4) käänteismuunnos on

$$x = g_1(u,v) = \frac{u+v}{2}, \qquad y = g_2(u,v) = \frac{u-v}{2},$$

ja muunnoksen Jakobiaani on

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Satunnaisvektorin (U,V) yhteisjakauman tiheysfunktio on yhtälön (5.9.3) nojalla

$$f_{U,V}(u,v) = \frac{1}{2} f\left(\frac{u+v}{2}, \frac{u-v}{2}\right). \tag{5.9.5}$$

Jos esimerkiksi X ja Y ovat riippumattomat ja noudattavat tasajakaumaa $\mathrm{Tas}(0,1)$, niin (X,Y):n yhteisjakauman tiheysfunktio $f(x,y) = 1$, kun $x \in [0,1]$ ja $y \in [0,1]$. Silloin (U,V):n tiheysfunktio on

$$f_{U,V}(u,v) = \begin{cases} \frac{1}{2}, & 0 \leq u+v \leq 2,\ 0 \leq u-v \leq 2. \\ 0 & \text{muualla.} \end{cases}$$

□

Yleinen muunnos

Tarkasteltavalla muunnoksella ei tietenkään aina ole käänteismuunnosta. Jos muunnos (5.9.1) ei ole yksi-yksinen eli bijektio, niin sillä ei ole käänteismuunnosta. Jos kuitenkin on olemassa sellainen arvoavaruuden $S_{X,Y}$ ositus yhteispisteettömiin (x, y)-tason osaväleihin A_1, A_2, ..., A_m, että

$$S_{X,Y} = A_1 \cup A_2 \cup \cdots \cup A_m \tag{5.9.6}$$

ja muunnoksella

$$u = h_1(x, y), \qquad v = h_2(x, y)$$

on käänteismuunnos

$$x = g_1(x, y), \qquad y = g_2(x, y)$$

jokaisella osavälillä A_i, $i = 1, 2, \ldots, m$, niin kaavaa (5.9.3) voidaan soveltaa kullakin osavälillä erikseen vastaavalla tavalla kuin yhden muuttujan tapauksessa. Määritellään funktiot

$$h_{ki}(x, y) = \begin{cases} h_k(x, y), & \text{kun } (x, y) \in A_i; \\ 0 & \text{muualla,} \end{cases}$$

kun $k = 1, 2$. Silloin $h_1(x, y) = \sum_{i=1}^{m} h_{1i}(x, y)$ ja $h_2(x, y) = \sum_{i=1}^{m} h_{2i}(x, y)$. Jokaisella muunnoksella

$$u = h_{1i}(x, y), \qquad v = h_{2i}(x, y)$$

on käänteismuunnos välillä A_i, $i = 1, 2, \ldots, m$. Merkitään näitä käänteismuunnoksia

$$x = g_{1i}(u, v), \qquad y = g_{2i}(u, v).$$

Satunnaisvektorin (U, V) tiheysfunktio voidaan silloin esittää kaavan (5.9.3) avulla seuraavasti

$$f_{U,V}(u, v) = \sum_{i=1}^{m} f_{X,Y}\big(g_{1i}(u, v), g_{2i}(u, v)\big)|J_i|, \tag{5.9.7}$$

missä J_i on muunnoksen $x = g_{1i}(u, v)$, $y = g_{2i}(u, v)$ Jakobiaani.

Satunnaismuuttujien funktion jakauma

Usein tarkasteltavana on vain yksi satunnaismuuttujien X ja Y funktio $U = h_1(X, Y)$. Funktio $h_1(X, Y)$ voi olla esimerkiksi muotoa $X+Y$, XY, X^2+Y^2 jne. Esitettyä muunnostekniikkaa voidaan edelleen käyttää, jos löydetään sellainen apumuuttuja $V = h_2(X, Y)$, että muunnoksella

$$U = h_1(X, Y), \qquad V = h_2(X, Y)$$

on käänteismuunnos. Muunnoskaavan (5.9.3) avulla saadaan sitten satunnaisvektorin (U, V) yhteisjakauman tiheysfunktio, josta voidaan määrittää U:n reunajakauman tiheysfunktio. Jos esimerkiksi $U = h_1(X, Y) = X + Y$, niin voidaan valita apumuuttuja $V = h_2(X, Y) = X - Y$. Silloin muunnoksella $u = x + y$, $v = x - y$ on käänteismuunnos

$$x = g_1(u, v) = \frac{u + v}{2}, \qquad y = g_2(u, v) = \frac{u - v}{2}$$

ja (U, V):n tiheysfunktio saadaan kaavalla (5.9.3), niinkuin Esimerkissä 5.23 osoitettiin. Kun tästä (U, V):n tiheysfunktiosta "integroidaan v pois", saadaan U:n tiheysfunktio. Jos ollaan kiinnostuneita esimerkiksi satunnaismuuttujan $U = XY$ tiheysfunktiosta, voidaan valita apumuuttuja $V = X$, sillä muunnoksella $u = xy$, $v = x$ on käänteismuunnos. Sitten sovelletaan jälleen edellä kuvattua tekniikkaa. Huomaa, että apumuuttujan valinta ei ole yksikäsitteinen, vaan useilla eri valinnoilla voidaan päästä haluttuun tulokseen.

Tässä yhteydessä on syytä palauttaa mieleen Lause 3.6. Siinä osoitettiin riippumattomille satunnaismuuttujille X ja Y seuraava tulos: Jos $g(X)$ ei riipu Y:stä ja $h(Y)$ ei riipu X:stä, niin silloin satunnaismuuttujat $g(X)$ ja $h(Y)$ ovat riippumattomat. Lause todistettiin diskreettien satunnaismuuttujien tapauksessa, mutta se pitää paikkansa myös jatkuville muuttujille.

Esimerkki 5.24 Jos X ja Y ovat riippumattomat ja noudattavat standardimuotoista normaalijakumaa, niin mitä jakaumaa noudattaa $X + Y$? Määritelään ensin apumuuttuja $V = X - Y$. Kuten esimerkissä 5.23 osoitettiin, muunnoksen $u = x + y$, $v = x - y$ käänteismuunnos on

$$x = g_1(u, v) = \frac{u + v}{2}, \qquad y = g_2(u, v) = \frac{u - v}{2},$$

ja muunnoksen Jakobiaani $J = -\frac{1}{2}$. Yhtälön (5.9.5) perusteella (U, V):n tiheysfunktio on

$$\begin{aligned} f_{U,V}(u, v) &= \frac{1}{4\pi} \mathrm{e}^{-[(u+v)^2/8 + (u-v)^2/8]} \\ &= \frac{1}{4\pi} \mathrm{e}^{-(u^2+v^2)/4} = \frac{1}{\sqrt{4\pi}} \mathrm{e}^{-u^2/4} \frac{1}{\sqrt{4\pi}} \mathrm{e}^{-v^2/4} \\ &= f_U(u) f_V(v). \end{aligned} \tag{5.9.8}$$

Näemme, että $f_U(u)$ ja $f_V(v)$ ovat kumpikin normaalijakauman $\mathrm{N}(0, 2)$ tiheysfunktioita. Siksi

$$\int_{-\infty}^{\infty} f_{U,V}(u, v)\, \mathrm{d}v = f_U(u) \int_{-\infty}^{\infty} f_V(v)\, \mathrm{d}v = f_U(u),$$

joten $U = X + Y \sim \mathrm{N}(0, 2)$. Identiteetistä (5.9.8) seuraa, että $X + Y$ ja $X - Y$ ovat riippumattomat. Havaitsemme myös, että $X - Y \sim \mathrm{N}(0, 2)$. Itse

asiassa voidaan todistaa seuraava tulos: Jos X ja Y ovat riippumattomat ja noudattavat samaa jakaumaa F, niin $X+Y$ ja $X-Y$ ovat riippumattomat jos ja vain jos F on normaalijakauma. □

Esimerkki 5.25 Olkoot X ja Y jatkuvat satunnaismuuttujat, joiden yhteisjakauman tiheysfunktio on f. Määritellään lineaarinen muunnos

$$\begin{aligned} u &= ax + by, \\ v &= cx + dy. \end{aligned} \tag{5.9.9}$$

Ratkaisemalla yhtälöryhmästä (5.9.9) x ja y saadaan käänteismuunnos

$$\begin{aligned} x &= \frac{du - bv}{D}, \\ y &= \frac{av - cu}{D}, \end{aligned}$$

missä $D = ad - bc$. Käänteismuunnos on olemassa, jos $D \neq 0$. Muunnoksen Jakobiaani on

$$J = \begin{vmatrix} \frac{d}{D} & \frac{-b}{D} \\ \frac{-c}{D} & \frac{a}{D} \end{vmatrix} = \frac{ad - bc}{D^2} = \frac{1}{D}.$$

Satunnaisvektorin (U, V) yhteisjakauman tiheysfunktio on yhtälön (5.9.3) nojalla

$$f_{U,V}(u, v) = \frac{1}{|D|} f[(du - bv)/D, (av - cu)/D]. \tag{5.9.10}$$

Esimerkin 5.23 yhtälö (5.9.5) on yhtälön (5.9.10) erikoistapaus. Kun $a = b = c = 1$ ja $d = -1$ sijoitetaan yhtälöön (5.9.10), saadaan yhtälö (5.9.5). Lineaarinen muunnos

$$\begin{aligned} u &= ax + by + e, \\ v &= cx + dy + f. \end{aligned}$$

voidaan palauttaa muunnokseen (5.9.9) merkitsemällä $u^* = u - e$ ja $v^* = v - f$, jolloin

$$\begin{aligned} u^* &= ax + by, \\ v^* &= cx + dy. \end{aligned}$$

Yhtälöstä (5.9.10) saadaan sitten (U^*, V^*):n tiheysfunktio. □

5.9.1 Yleinen kahden muuttujan normaalijakauma

Standardimuotoinen kaksiulotteinen normaalijakauma määriteltiin alaluvussa 5.8.1. Yleinen kaksiulotteinen normaalijakauma voidaan määritellä standardimuotoisen normaalijakauman avulla vastaavasti kuin yhden muuttujan tapauksessa. Olkoon (X, Y) sellainen satunnaisvektori, että $E(X) = \mu_1$,

$E(Y) = \mu_2$, $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{Var}(Y) = \sigma_2^2$ ja $\mathrm{Cov}(X,Y) = \sigma_{12}$. Silloin satunnaisvektorin (X,Y) keskiarvovektori on $\boldsymbol{\mu} = (\mu_1, \mu_2)^\mathsf{T}$ ja kovarianssimatriisi

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Satunnaismuuttujien X ja Y välinen korrelaatiokerroin on $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$.

Määritelmä 5.7 Satunnaisvektori (X,Y) noudattaa normaalijakaumaa, jonka keskiarvovektori on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$, jos se voidaan lausua muodossa

$$\begin{aligned} X - \mu_1 &= \sigma_1 \sqrt{1-\rho^2} Z_1 + \rho \sigma_1 Z_2, \\ Y - \mu_2 &= \sigma_2 Z_2, \end{aligned} \tag{5.9.11}$$

missä Z_1 ja Z_2 ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa $\mathrm{N}(0,1)$ sekä $\sigma_1 > 0$, $\sigma_2 > 0$ ja $|\rho| \leq 1$.

Merkitsemme $(X,Y) \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kun (X,Y) noudattaa normaalijakaumaa, jonka keskiarvovektori on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$. Normaalijakaumaa $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ noudattava satunnaisvektori (X,Y) saadaan aina riippumattomista standardoiduista normaalimuuttujista lineaarisella muunnoksella (5.9.11)

Yleisen normaalijakauman tiheysfunktio saadaan suoraan Esimerkissä 5.25 esitetyllä tekniikalla. Merkitään $x - \mu_1 = u$ ja $y - \mu_2 = v$. Koska nyt lineaarisessa muunnoksessa (5.9.9) $c = 0$ ja $D = ad$, saa yhtälö (5.9.10) muodon

$$f_{U,V}(u,v) = \frac{1}{ad} f\left(\frac{du - bv}{ad}, \frac{v}{d} \right).$$

Koska f on kahden muuttujan standardimuotoisen normaalijakauman tiheysfunktio ja muunnoksessa (5.9.11) $a = \sigma_1 \sqrt{1-\rho^2}$, $b = \rho \sigma_1$ ja $d = \sigma_2$, niin

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_1^2 \sigma_2^2 (1-\rho^2)} (\sigma_2 u - \rho \sigma_1 v)^2 + \frac{v^2}{\sigma_2^2} \right] \\ &= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{u^2}{\sigma_1^2} - 2\rho \frac{u}{\sigma_1} \frac{v}{\sigma_2} + \frac{v^2}{\sigma_2^2} \right) \right]. \end{aligned}$$

Kun edellä johdettuun tiheysfunktioon sijoitetaan $u = x - \mu_1$ ja $v = y - \mu_2$, saadaan (X,Y):n tiheysfunktio

$$\begin{aligned} f_{X,Y}(x,y) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\Bigg[-\frac{1}{2(1-\rho^2)} \Bigg(&\left(\frac{x-\mu_1}{\sigma_1} \right)^2 \\ &- 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \Bigg) \Bigg]. \end{aligned} \tag{5.9.12}$$

Seuraavassa lauseessa esitetään kaksiulotteisen normaalijakauman keskeiset ominaisuudet.

Lause 5.11 *Oletaan, että satunnaisvektori (X, Y) noudattaa kaksiulotteista normaalijakaumaa $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, missä $\boldsymbol{\mu} = (\mu_1, \mu_2)^\mathsf{T} = [E(X), E(Y)]^\mathsf{T}$ ja*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{pmatrix}$$

ja $\rho = \mathrm{Cor}(X, Y) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$. Silloin pitävät paikkansa seuraavat ominaisuudet:

1. *$X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ ja $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$,*
2. *X ja Y ovat riippumattomat jos ja vain jos $\rho = 0$.*
3. *X:n ja Y:n ehdolliset jakaumat:*

$$X \mid y \sim \mathrm{N}\bigg(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y - \mu_2),\, \sigma_1^2(1 - \rho^2)\bigg),$$

eli X noudattaa normaalijakaumaa ehdolla, että $Y = y$ on annettu. Vastaavasti

$$Y \mid x \sim \mathrm{N}\bigg(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1),\, \sigma_2^2(1 - \rho^2)\bigg).$$

4. *Satunnaisvektorin (X, Y) momenttifunktio on*

$$M(s,t) = \exp\bigg(\mu_1 s + \mu_2 t + \frac{\sigma_1^2 s^2 + \sigma_2^2 t^2 + \rho\sigma_1\sigma_2 st}{2}\bigg).$$

Lause 5.12 *Satunnaisvektori (X, Y) noudattaa kaksiulotteista normaalijakaumaa, jos ja vain jos satunnaismuuttujien X ja Y lineaarikombinaatiot $aX+bY$ noudattavat yksiulotteista normaalijakaumaa kaikilla $a \in \mathbb{R}$ ja $b \in \mathbb{R}$.*

5.9.2 Studentin t-jakauma, F-jakauma ja beta-jakauma

Oletetaan, että satunnaismuuttujat Z ja U ovat riippumattomat, $Z \sim \mathrm{N}(0,1)$ ja U noudattaa χ^2-jakaumaa vapausastein r eli $U \sim \mathrm{Khi2}(r)$. Tarkastellaan nyt satunnaismuuttujan

$$T = \frac{Z}{\sqrt{U/r}} \tag{5.9.13}$$

jakaumaa. Tätä muotoa olevalla satunnaismuuttujalla on erittäin keskeinen rooli tilastollisessa päättelyssä. Satunnaismuuttuja (5.9.13) noudattaa ns. *Studentin t-jakaumaa* (tai lyhyesti t-jakaumaa) vapausastein r. Jakauma on nimetty englantilaisen tilastotieteilijän W. S. Gossetin mukaan. Gosset esitti tämän jakauman *Biometrikassa* vuonna 1908 nimimerkillä "Student". Kun T noudattaa Studentin t-jakaumaa vapausastein r, merkitään $T \sim \mathrm{t}(r)$.

Olkoon $X_1, X_2, \ldots, X_n$ on otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Silloin siis satunnaismuuttujat $X_1, X_2, \ldots, X_n$ ovat toisistaan riippumattomat ja $X_i \sim \mathrm{N}(\mu, \sigma^2)$, $1 \le i \le n$. Otoksesta laskettu suure

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \tag{5.9.14}$$

eli otossuure noudattaa t-jakaumaa vapausastein $n-1$, missä $\overline{X}$ on otoskeskiarvo ja

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

on otosvarianssi. Lausekkeen (5.9.14) osoittaja $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ noudattaa normaalijakaumaa $\mathrm{N}(0,1)$ ja satunnaismuuttuja S^2/σ^2 noudattaa jakaumaa $\mathrm{Khi2}(n-1)$. Koska $\overline{X}$ ja S^2 ovat toisistaan riippumattomat, niin lausekkeen (5.9.14) osoittaja ja nimittäjä ovat toisistaan riippumattomat.

Satunnaismuuttujan (5.9.13) eli Studentin t-jakauman tiheysfunktio vapausastein r on

$$f_T(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \frac{1}{\sqrt{r\pi}} \frac{1}{(1 + t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty. \tag{5.9.15}$$

Huomaa, että vapausastein $r = 1$ tiheysfunktiosta (5.9.15) tulee Cauchyn jakauman tiheysfunktio.

Tiheysfunktio (5.9.15) saadaan suoraviivaisesti edellä esitetyllä muunnostekniikalla. Lähdetään liikkeelle satunnaisvektorin (Z, U) yhteisjakaumasta. Koska Z ja U ovat riippumattomat, niin

$$f_{Z,U}(z,u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} u^{r/2-1} e^{-u/2}, \ -\infty < z < \infty, \ 0 < v < \infty.$$

Tehdään muunnos

$$t = \frac{z}{\sqrt{u/r}}, \qquad w = u,$$

jonka käänteismuunnos on

$$z = t\sqrt{w/r}, \qquad u = w.$$

Muunnoksen jakobiaani on $\sqrt{w/r}$. Sen jälkeen voidaan soveltaa muunnoskaavaa (5.9.3). T:n tiheysfunktio saadaan (T, W):n yhteisjakauman tiheysfunktiosta integroimalla w:n yli:

$$f_T(t) = \int_0^\infty f_{Z,U}\left[t(w/r)^{1/2}, w\right] (w/r)^{1/2} \, \mathrm{d}w.$$

Laskennan lopputuloksena on t-jakauman tiheysfunktio (5.9.15). Yksityiskohdat jätetään lukijan tehtäväksi.

Studentin t-jakaumalla ei ole momenttifunktiota, koska sillä ei ole kaikkien kertalukujen momentteja. Jos $T_r \sim \mathrm{t}(r)$, niin silloin T_r:llä on ainoastaan $r-1$ ensimmäistä momenttia. Esimerkiksi jakaumalla $\mathrm{t}(1)$ ei ole keskiarvoa ja jakaumalla $\mathrm{t}(2)$ ei ole varianssia. Voidaan laskemalla osoittaa, että

$$E(T_r) = 0, \quad \text{jos } r > 1, \qquad \mathrm{Var}(T_r) = \frac{r}{r-2}, \quad \text{jos } r > 2. \tag{5.9.16}$$

Toinen tilastollisessa päättelyssä keskeinen jakauma, *Snedecorin F-jakauma* tai vain lyhyesti F-jakauma, voidaan johtaa t-jakauman tapaan. Jos $X_1, X_2, \ldots, X_n$ on otos normaalijakaumasta $\mathrm{N}(\mu_1, \sigma_1^2)$ ja $X_1, X_2, \ldots, X_m$ otos normaalijakaumasta $\mathrm{N}(\mu_2, \sigma_2^2)$, niin silloin satunnaismuuttuja

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{5.9.17}$$

noudattaa F-jakaumaa vapausastein $n-1$ ja $m-1$. Lausekkeessa (5.9.17) S_1^2 ja S_2^2 ovat otosvarianssit. S_1^2/σ_1^2 ja S_2^2/σ_2^2 ovat riippumattomat ja

$$(n-1)S_1^2/\sigma_1^2 \sim \mathrm{Khi2}(n-1), \qquad (m-1)S_2^2/\sigma_2^2 \sim \mathrm{Khi2}(m-1).$$

F-jakauma määritellään seuraavasti: Jos $X \sim \mathrm{Khi2}(r)$ ja $Y \sim \mathrm{Khi2}(s)$ ovat riippumattomat, niin satunnaismuuttuja

$$F = \frac{X/r}{Y/s} \tag{5.9.18}$$

noudattaa F-jakaumaa vapausastein r ja s. Silloin merkitään $F \sim \mathrm{F}(r,s)$. F-jakauman tiheysfunktio voidaan johtaa vastaavalla tavalla kuin t-jakauman tiheysfunktio. Määritellään muunnos

$$U = X + Y, \qquad V = \frac{X/r}{Y/s},$$

missä F-suuretta (5.9.18) on merkitty kirjaimella V. Muunnoksen käänteismuunnos on

$$X = \frac{r}{s}\Big(1 + \frac{r}{s}V\Big)^{-1}UV, \qquad Y = \Big(1 + \frac{r}{s}V\Big)^{-1}U.$$

Koska X ja Y ovat riippumattomat, niin

$$f_{X,Y}(X,Y) = k_r k_s x^{(r/2)-1} y^{(s/2)-1} \mathrm{e}^{-(x+y)/2}, \qquad 0 < x < \infty, \quad 0 < y < \infty,$$

missä $k_r = 1/\Gamma\big(\frac{r}{2}\big)2^{r/2}$ ja $k_s = 1/\Gamma\big(\frac{s}{2}\big)2^{s/2}$. Kun lasketaan Jakobiaani ja tehdään sijoitukset, saadaan satunnaisvektorin (U,V) tiheysfunktio $f_{U,V}(u,v)$ kaavan (5.9.3) mukaisesti. V:n reunajakauman tiheysfuntio on sitten F-jakauman tiheysfuntio:

$$f_F(v) = \frac{\Gamma\big(\frac{r+s}{2}\big)}{\Gamma\big(\frac{r}{2}\big)\Gamma\big(\frac{s}{2}\big)}\Big(\frac{r}{s}\Big)^{r/2}\frac{v^{(r/2)-1}}{\big[1+\frac{r}{s}v\big]^{(r+s)/2}}, \qquad 0 < v < \infty. \tag{5.9.19}$$

F-jakauman keskiarvo ja varianssi ovat

$$E(F) = \frac{s}{s-2}, \qquad s > 2 \tag{5.9.20}$$

$$\mathrm{Var}(F) = \frac{2s^2(r+s-2)}{r(s-2)^2(s-4)}, \qquad s > 4. \tag{5.9.21}$$

Beta-jakauma

Olkoot $X \sim \mathrm{Gamma}(\alpha, \theta)$ ja $Y \sim \mathrm{Gamma}(\beta, \theta)$ riippumattomat gamma-jakaumaa noudattavat satunnaismuuttujat. Silloin satunnaisvektorin (X, Y) tiheysfunktio on

$$f(x,y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} x^{\alpha-1} y^{\beta-1} \exp\Big(-\frac{x+y}{\theta}\Big), \qquad 0 < x, \quad y < \infty.$$

Tarkastellaan muunnosta

$$U = \frac{X}{X+Y}, \qquad V = X + Y,$$

jonka käänteismuunnos on

$$X = UV, \qquad Y = V - UV.$$

Muunnoksen Jakobiaani on

$$\begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1-u) + uv = v.$$

Satunnaisvektorin (U, V) tiheysfunktio on kaavan (5.9.3) mukaan

$$f_{U,V}(u,v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (uv)^{\alpha-1} (v - uv)^{\beta-1} \mathrm{e}^{-v\theta} v,$$

missä $0 < u < 1$ ja $0 < v < \infty$. Kun tästä (U, V):n yhteisjakauman tiheysfunktiosta määritetään U:n reunajakuman tiheysfunktio, saadaan

$$f_U(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \qquad 0 < u < 1. \tag{5.9.22}$$

Sanomme, että U noudattaa betajakaumaa parametrein α ja β. Silloin merkitsemme $U \sim \mathrm{Beta}(\alpha, \beta)$. Koska (5.9.22) on tiheysfunktio, niin

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,\mathrm{d}u.$$

Funktiota

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{5.9.23}$$

kutsutaan *betafunktioksi.*

Betajakauma on yksi niitä harvoja nimettyjä jakaumia, joiden koko todennäköisyysmassa on äärellisellä välillä, eli välillä $(0, 1)$. Betajakauman momentit on helppo laskea tiheysfunktion erityisominaisuuksien avulla. Kun $n > -\alpha$, niin

$$E(X^n) = \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + n)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)}. \tag{5.9.24}$$

Sijoittamalla $n = 1$ ja $n = 2$ kaavaan (5.9.24) saadaan 1. ja 2. momentti ja niiden avulla varianssi:

$$E(X) = \frac{\alpha}{\alpha + \beta} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{5.9.25}$$

Esimerkki 5.26 Oletetaan, että on suoritettavana $n+m$ työtä. Töiden suorittamiseen tarvittavat ajat noudattavat toisistaan riippumatta eksponettijakaumaa keskiarvolla $\theta > 0$ (ts. gammajakaumaa parametrein $\alpha = 1$ ja θ.) Oletetaan, että kaksi eri työntekijää tekee nämä työt siten, että työntekijä A tekee työt $1, 2, \ldots, n$ ja B tekee työt $n+1, n+2, \ldots, n+m$. Jos merkitään X:llä työntekijän A käyttämää aikaa ja Y:llä työntekijän B käyttämää aikaa, niin toisistaan riippumatta $X \sim \mathrm{Gamma}(n, \theta)$ ja $Y \sim \mathrm{Gamma}(m, \theta)$. Silloin $(n + m)$:n työn vaatima kokonaisaika $X + Y$ noudattaa gammajakaumaa $\mathrm{Gamma}(n + m, \theta)$. Työntekijän A käyttämä suhteellinen osuus $X/(X + Y)$ kokonaisajasta noudattaa betajakaumaa $\mathrm{Beta}(n, m)$. □

5.9.3 Hierarkkiset mallit ja yhdistetyt jakaumat

Tarkastellaan aluksi esimerkkinä tietyn laitteen, esimerkiksi kopiokoneen, rikkoontumista. Oletetaan, että rikkoontumisten lukumäärä X vuodessa noudattaa Poissonin jakaumaa parametrilla λ. Kun laite on rikkoontunut,sen korjaamiseen tarvittava aika noudattaa eksponettijakaumaa keskiarvolla $\theta > 0$. Olkoon Y_i aika, joka tarvitaan i. rikkoontumisen korjaamiseen. Oletetaan lisäksi, että korjausajat eri kerroilla ovat riippumattomat. Jos vuodessa sattuu $X = x$ rikkoontumista, niin kokonaiskorjausaika on

$$Y = Y_1 + Y_2 + \cdots + Y_x.$$

Silloin

$$E(Y \mid x) = E(Y_1) + E(Y_2) + \cdots + E(Y_x) = x\theta$$

ja

$$\mathrm{Var}(Y \mid x) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_x) = x\theta^2.$$

Edellä lasketut keskiarvo ja varianssi ovat siis ehdollisia ehdolla $X = x$. Ehdollinen keskiarvo ja varianssi

$$\mu(x) = E(Y \mid x) = x\theta \qquad \text{ja} \qquad \sigma^2(x) = \mathrm{Var}(Y \mid x) = x\theta^2$$

ovat x:n funktioita. Koska $X \sim \text{Poi}(\lambda)$ on satunnaismuuttuja, niin myös

$$\mu(X) = E(Y \mid X) = X\theta \qquad \text{ja} \qquad \sigma^2(X) = \text{Var}(Y \mid X) = X\theta^2$$

ovat satunnaismuuttujia. Silloin

$$\begin{aligned} E[\mu(X)] &= E[E(Y \mid X)] = \theta\, E(X) = \theta\lambda, \\ E[\sigma^2(X)] &= E[\text{Var}(Y \mid X)] = \theta^2\, E(X) = \theta^2\lambda, \\ \text{Var}[\mu(X)] &= \text{Var}[E(Y \mid X)] = \theta\, \text{Var}(X) = \theta\lambda. \end{aligned}$$

Määritimme siis edellä kaksinkertaisen odotusarvon $E[E(Y \mid X)]$, missä sisimmäinen odotusarvo on otettu Y:n suhteen ja ulompi X:n suhteen.

Esitämme nyt kaksinkertaisia odotusarvoja koskevan lauseen, joka usein helpottaa huomattavasti odotusarvojen laskemista.

Lause 5.13 *Olkoot X ja Y mitkä tahansa kaksi satunnaismuuttujaa. Silloin*

$$E(Y) = E[E(Y \mid X)], \tag{5.9.26}$$

jos odotusarvot ovat olemassa.

Todistus. Olkoon $f(x, y)$ satunnaisvektorin (X, Y) tiheysfunktio. Määritelmän mukaan

$$E(Y) = \iint y f(x, y)\, \mathrm{d}y\, \mathrm{d}x = \int \left[\int y f(y \mid x)\, \mathrm{d}y \right] f_X(x)\, \mathrm{d}x, \tag{5.9.27}$$

missä $f(y \mid x)$ on Y:n ehdollisen jakauman tiheysfunktio ehdolla $X = x$ ja $f_X(x)$ on X:n reunajakauman tiheysfunktio. Integrointi tehdään yli (X, Y):n arvoavaruuden. Koska sisempi integraali luasekkeessa (5.9.27) on ehdollinen odotusarvo $E(Y \mid x)$, niin odotusarvo (5.9.27) voidaan kirjoittaa muodossa

$$E(Y) = \int E(Y \mid x) f_X(x)\, \mathrm{d}x = E[E(Y \mid X)],$$

niin kuin lauseessa väitetään. □

Voimme soveltaa nyt Lausetta 5.13 laitteen rikkoontumisten vaatimaa korjausaikaa koskevaan esimerkkiin. Kokonaiskorjausaika vuodessa on Y ja sen keskiarvo on Lauseen 5.13 mukaan

$$\begin{aligned} E(Y) &= E[E(Y \mid X)] \\ &= E(\theta X) = \theta\, E(X) = \theta\lambda. \end{aligned}$$

Tälläinen sovellus on selvintä ajatella kaksitasoisena hierarkisena mallina, missä 1. vaiheessa sattuvat rikkoontumiset Poissonin jakauman mukaan ja sitten 2. vaiheessa tarvittavat korjausajat jakaantuvat eksponenttijakauman

mukaan. Huomaa, että kokonaiskorjausajan Y keskiarvoparametri on nyt satunnaismuuttuja θX, missä X noudattaa Poissonin jakaumaa. Siksi Y:n jakaumaa on perusteltua kutsua yhdistetyksi jakaumaksi, koska siinä yhdistyvät parametrin Poissonin jakauma ja korjausajan eksponenttijakauama. Odotusarvoa $E(Y)$ laskettaessa on yksinkertaisinta laskea ensin ehdollinen odotusarvo $E(Y \mid X)$ ja sitten näiden ehdollisten odotusarvojen odotusarvo. Silloin laskennassa ei tarvita Y:n reunajakaummaa. Lopputulos on kuitenkin Lauseen 5.13 mukaan $E(Y)$.

Myös varianssi voidaan laskea ehdollisen jakauman varianssin avulla.

Lause 5.14 *Olkoot X ja Y mitkä tahansa kaksi satunnaismuuttujaa. Silloin*

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[E(Y \mid X)], \tag{5.9.28}$$

jos odotusarvot ovat olemassa.

Todistus. Varianssin määritelmän mukaan

$$\mathrm{Var}(Y) = E[Y - E(Y)]^2 = E\big([Y - E(Y \mid X) + E(Y \mid X) - E(Y)]^2\big),$$

missä viimeinen yhtäsuuruus on saatu lisäämällä ja vähentämällä $E(Y \mid X)$. Korottamalla lauseke neliöön ja ottamalla odotusarvot saadaan

$$\begin{aligned}\mathrm{Var}(Y) = {} & E\big([Y - E(Y \mid X)]^2\big) + E\big([(Y \mid X) - E(Y)]^2\big) \\ & + 2\,E\big([Y - E(Y \mid X)]\,E[(Y \mid X) - E(Y)]\big).\end{aligned} \tag{5.9.29}$$

Lausekkeen (5.9.29) viimeinen termi on nolla, mikä voidaan osoittaa laskemalla merkitytyt odotusarvot. Yhtälön (5.9.29) oikean puolen ensimmäinen termi voidaan kirjoittaa muodossa:

$$\begin{aligned}E\big([Y - E(Y \mid X)]^2\big) &= E\big(E[Y - E(Y \mid X)]^2 \mid X\big) \\ &= E[\mathrm{Var}(Y \mid X)]\end{aligned}$$

ja toinen termi muodossa

$$E\big([E(Y \mid X) - E(Y)]^2\big) = \mathrm{Var}[E(Y \mid X)],$$

joten identiteetti (5.9.28) pitää paikkansa. □

Jatkuvat jakaumat: Yhteenveto

- Satunnaismuuttuja X on (absoluuttisesti) jatkuva, jos X:llä on tiheysfunktio $f(x) \geq 0$, joka toteuttaa seuraavat ehdot:

  1. $f(x) > 0$, kun $x \in S$,
  2. $\int\limits_S f(x)\,\mathrm{d}x = 1$,
  3. $P(X \in A) = \int\limits_A f(x)\,\mathrm{d}x$ on tapahtuman $\{X \in A\}$ todennäköisyys.

  Satunnaismuuttujan X kertymäfunktio on

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

  ja momenttifunktio

$$M(t) = E(\mathrm{e}^{tX}) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\,\mathrm{d}x.$$

- Tasajakauma $X \sim \mathrm{Tas}(a, b)$. Tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{kun } x \in [a, b]; \\ 0 & \text{muualla} \end{cases} \quad ja$$

$$E(X) = \frac{a+b}{2} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

  Momenttifunktio

$$M(t) = \begin{cases} \dfrac{\mathrm{e}^{tb} - \mathrm{e}^{ta}}{t(b-a)}, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

- Eksponenttijakauma $X \sim \mathrm{Exp}(\theta)$, $\theta > 0$ ja $x \geq 0$,

$$f(x) = \frac{1}{\theta}\mathrm{e}^{-x/\theta} \qquad \text{ja} \qquad F(x) = 1 - \mathrm{e}^{-x/\theta}.$$

  Silloin

$$E(X) = \theta \qquad \text{ja} \qquad \mathrm{Var}(X) = \theta^2.$$

  Momenttifunktio

$$M(t) = \frac{1}{1-\theta t}, \qquad t < \frac{1}{\theta}.$$

- Gammajakauma $X \sim \text{Gamma}(\alpha, \beta)$, $\alpha > 0$ ja $\beta > 0$. Silloin

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta}, \qquad 0 < x < \infty,$$

$$E(X^c) = \frac{\Gamma(\alpha + c)\beta^c}{\Gamma(\alpha)} \qquad \text{kaikilla } c > -\alpha.$$

  Erityisesti

$$E(X) = \alpha\beta \qquad \text{ja} \qquad \text{Var}(X) = \alpha\beta^2.$$

  Momenttifunktio

$$M(t) = \left(\frac{1}{1 - \beta t}\right)^\alpha, \qquad t < \frac{1}{\beta}.$$

- χ^2-jakauma, $X \sim \text{Khi2}(r)$. χ^2-jakauma saadaan, kun gammajakaumassa valitaan $\alpha = \frac{r}{2}$ ja $\beta = 2$, missä r on positiivinen kokonaisluku. Silloin

$$E(X) = r, \qquad \text{Var}(X) = 2r \qquad \text{ja} \qquad M(t) = (1 - 2t)^{-r/2}, \quad t < \frac{1}{2}.$$

  Jos $X_i \sim \text{Khi2}(r_i)$, $i = 1, 2$, ovat riippumattomat, niin $X_1 + X_2 \sim \text{Khi2}(r_1 + r_2)$.

- Normaalijakauma $X \sim \text{N}(\mu, \sigma^2)$. Silloin

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty,$$

$$E(X) = \mu, \qquad \text{Var}(X) = \sigma^2 \qquad \text{ja} \qquad M(t) = \mathrm{e}^{\mu t + (\sigma^2 t^2)/2}.$$

- Jos $Z \sim \text{N}(0, 1)$, niin $Z^2 \sim \text{Khi2}(1)$.
  Jos $Z_i \sim \text{N}(0, 1)$, $i = 1, 2$, ovat riippumattomat, niin $Z_1^2 + Z_2^2 \sim \text{Khi2}(2)$.

- Studentin t-jakauma, $T \sim t(r)$.

$$T = \frac{Z}{\sqrt{U/r}} \sim t(r),$$

  jos $Z \sim \text{N}(0, 1)$ ja $U \sim \text{Khi2}(r)$ ovat riippumattomat. Silloin

$$E(T) = 0, \quad r > 1 \qquad \text{ja} \qquad \text{Var}(T) = \frac{r}{r - 2}, \quad r > 2.$$

- F-jakauma, $F \sim \text{F}(r, s)$.

$$F = \frac{X/r}{Y/s} \sim F(r, s),$$

  jos $X \sim \text{Khi2}(r)$ ja $Y \sim \text{Khi2}(s)$ ovat riippumattomat. Silloin

$$E(F) = \frac{s}{s - 2}, \quad s > 2; \qquad \text{Var}(F) = \frac{2s^2(r + s - 2)}{r(s - 2)^2(s - 4)}, \quad s > 4.$$

- Beta-jakauma, $X \sim \text{Beta}(\alpha, \beta)$; $\alpha > 0$, $\beta > 0$.

  Tiheysfunktio

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1.$$

$$E(X) = \frac{\alpha}{\alpha+\beta} \qquad \text{ja} \qquad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Muuttujien vaihto

Satunnaismuuttujan $Y = h(X)$ tiheysfunktio, kun X:n tiheysfunktio on $f_X(x)$:

$$f_Y(y) = f_X\big[g(y)\big]\,|g'(y)|, \qquad y \in S_Y,$$

missä $g(y)$ on $h(x)$:n käänteisfunktio.

Kaksiulotteiset jakaumat

- Jatkuvien satunnaismuuttujien X ja Y muodostaman satunnaisvektorin (X, Y) tiheysfunktio toteuttaa ehdot

  1. $f(x, y) \geq 0$ kaikilla (x, y),
  2. $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$,
  3. $P[(X, Y) \in A] = \iint\limits_{(x,y)\in A} f(x, y)\,\mathrm{d}x\,\mathrm{d}y$.

  Reunajakaumien tiheysfunktiot

$$f_X(x) = \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y, \quad x \in S_X; \qquad f_Y(y) = \int\limits_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x, \quad y \in S_Y$$

  Ehdolliset tiheysfunktiot

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)} \qquad \text{ja} \qquad f_Y(y \mid x) = \frac{f(x, y)}{f_X(x)},$$

  missä $f_X(x) > 0$ ja $f_Y(y) > 0$.

  Kertymäfunktio

$$F(x, y) = \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} f(s, t)\,\mathrm{d}s\,\mathrm{d}t.$$

- Normaalijakauma $(X, Y) \sim \mathrm{N}_2\big(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\big)$.

  Tiheysfunktio

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\,\exp\left[-\frac{1}{2(1-\rho^2)}Q(x,y)\right],$$

  missä

$$Q(x,y) = \left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2.$$

  Reunajakaumat:

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2) \qquad \text{ja} \qquad Y \sim \mathrm{N}(\mu_2, \sigma_2^2).$$

  Ehdolliset jakaumat:

  X:n jakauma ehdolla $Y = y$ on

$$X \mid y \sim \mathrm{N}\Big[\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y-\mu_2),\ \sigma_1^2(1-\rho^2)\Big],$$

  Y:n jakauma ehdolla $X = x$ on

$$Y \mid x \sim \mathrm{N}\Big[\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1),\ \sigma_2^2(1-\rho^2)\Big].$$

Muuttujien vaihto

Olkoon

$$U = h_1(X,Y); \qquad V = h_2(X,Y)$$

muunnos, jolla on käänteismuunnos $x = g_1(u,v), \quad y = g_2(u,v)$. Satunnaisvektorin (U,V) yhteisjakauman tiheysfunktio on

$$f_{U,V}(u,v) = f\big(g_1(u,v), g_2(u,v)\big)|J|,$$

missä $f(x,y)$ on satunnaismuuttujien X ja Y yhteisjakauman tiheysfunktio ja

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}.$$

Ehdolliset odotusarvot ja varianssit

- Ehdollinen odotusarvo

$$E(Y) = E[E(Y \mid X)]$$

- Ehdollinen varianssi

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[E(Y \mid X)].$$

Harjoituksia

1. Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio

$$f(x) = \begin{cases} \frac{1}{2}, & 0 \le x < \frac{1}{3}; \\ 1, & \frac{1}{3} \le x < \frac{2}{3}; \\ c, & \frac{2}{3} \le x \le 1; \\ 0 & \text{muualla.} \end{cases}$$

(a) Laske c.

(b) Määritä X:n kertymäfunktio.

(c) Piirrä X:n tiheysfunktio ja kertymäfunktio.

2. Olkoon jatkuvan satunnaismuuttujan X tiheysfunktio $f(x) = 2(1-x)$, kun $0 \le x \le 1$ ja $f(x) = 0$ muualla.

(a) Piirrä X:n tiheysfunktio.

(b) Määritä ja piirrä X:n kertymäfunktio.

(c) Laske (i) $P(0 \le X \le 1/2)$, (ii) $P(1/4 \le X \le 3/4)$, (iii) $P(X = 3/4)$ ja (iv) $P(X \ge 3/4)$.

3. Olkoon $f(x)$ jatkuvan satunnaismuuttujan X tiheysfunktio. (i) Määritä jokaisesta alla määritellystä funktiosta $f(x)$ vakio c siten, että $f(x)$ on tiheysfunktio. (ii) Määritä kertymäfunktio $F(x) = P(X \le x)$ ja (iii) hahmottele tiheysfunktion $f(x)$ ja kertymäfunktion $F(x)$ kuvaajat.

(a) $f(x) = x^3/4, \quad 0 < x < c$;

(b) $f(x) = (3/16)x^2, \quad -c < x < c$;

(c) $f(x) = c/\sqrt{x}, \quad 0 < x < 1$. Onko $f(x)$ rajoitettu?

4. Olkoon X:n tiheysfunktio $f(x) = c/x^2$, $1 < x < \infty$.

(a) Määritä c:n arvo siten, että $f(x)$ on tiheysfunktio.

(b) Osoita, että $E(X)$ ei ole äärellinen.

5. Olkoon $X \sim \text{Khi2}(12)$. Määritä vakiot a ja b siten, että

$$P(a < X < b) = 0.90 \qquad \text{ja} \qquad P(X < a) = 0.05.$$

6. Olkoon $X \sim \text{Khi2}(23)$.

(a) Laske $P(14.85 < X < 32.01)$.

(b) Määritä a ja b siten, että $P(a < X < b) = 0.95$ ja $P(X < a) = 0.025$.

(c) X:n keskiarvo ja varianssi.

7. Olkoon X:n momenttifunktio $M(t) = (1-2t)^{-12}$, $t < 1/2$. Laske

(a) $E(X)$,

(b) $\mathrm{Var}(X)$ ja

(c) $P(15.66 < X < 42.98)$.

8. Olkoon $X \sim \mathrm{N}(7, 4)$. Laske todennäköisyys $P[15.364 \leq (X-7)^2 \leq 20.096]$ (Vihje: Lause 5.6).

9. Oletetaan, että X ja Y ovat riippumattomat ja $X \sim \mathrm{Khi2}(8)$ ja $Y \sim \mathrm{Khi2}(12)$.

(a) Laske

$$P(1.646 < X \leq 20.09), \qquad P(Y > 6.304), \qquad P(X+Y = 19.34).$$

(b) Määritä b, c ja d siten, että

$$P(X \leq b) = 0.9, \qquad P(Y > c) = 0.9, \qquad P(X+Y > d) = 0.05.$$

(Vihje: Lause 5.8)

10. Olkoot Z_1, Z_2 ja Z_3 riippumattomat ja ne noudattavat $\mathrm{N}(0,1)$-jakaumaa. Määritellään satunnaismuuttujat

$$\bar{Z} = \frac{1}{3}(Z_1 + Z_2 + Z_3) \qquad \text{ja} \qquad U = Z_1^2 + Z_2^2 + Z_3^2.$$

Määritä vakiot a, b siten, että

$$P(|\bar{Z}| \leq a) = 0.95; \qquad P(U > b) = 0.025.$$

(Vihje: Voit olettaa, että $\bar{Z}$ noudattaa normaalijakaumaa. Ks. myös Lause 5.7).

11. Määritä Esimerkissä 5.13 X:n ja Y:n jakaumien (reunajakaumien) tiheysfunktiot. Totea, että X ja Y ovat riippumattomat.

12. Totea laskemalla, että Esimerkissä 5.20 $E(X \mid y) = \frac{y}{2}$, $0 \leq y \leq 1$.

13. Olkoon jatkuvien satunnaismuuttujien X ja Y yhteisjakauman kertymäfunktio

$$F(x,y) = \begin{cases} kxy(x+y), & \text{kun } 0 < x < 1 \text{ ja } 0 < y < 1; \\ 0 & \text{muualla.} \end{cases}$$

(a) Laske vakion k arvo ja määritä X:n ja Y:n yhteisjakauman tiheysfunktio.

(b) Määritä X:n reunajakauman ja ehdollisen jakauman tiheysfunktio.

(c) Laske todennäköisyydet

$$P(X < 0.5,\ Y < 0.5); \qquad P(X < 0.5); \qquad P(X < 0.5 \mid Y < 0.5).$$

14. Määritä a, b, c, d siten, että

(a) $P(a \leq F_{8,12} \leq b) = 0.8$; $\quad P(c \leq F_{6,6} \leq d) = 0.98$.

(b) $P(|t_9| \geq a) = 0.05$; $\quad P(|t_{20}| > b) = 0.95$.

(c) $P(F_{1,12} \geq b) = 0.05$; $\quad P(F_{1,12} \leq c) = 0.2$.

15. Olkoot Y ja Z sellaiset riippumattomat satunnaismuuttujat, että $Z \sim \mathrm{N}(0,1)$ ja $Y \sim \chi^2_4$. Määritä a, b siten, että

$$P(Z \leq a\sqrt{Y}) = 0.975; \qquad P(Y + Z^2 \leq b) = 0.975.$$

16. Oletetaan, että $X \sim t(\nu)$. Osoita muuttujien vaihtotekniikalla, että $X^2 \sim \mathrm{F}_{1,\nu}$. (Huomaa, että muunnos ei ole monotoninen).

17. Olkoon $X \sim$ Khi2(12).

(a) Määritä vakiot a ja b siten, että

$$P(a < X < b) = 0.90 \qquad \text{ja} \qquad P(X < a) = 0.05.$$

(b) Olkoon $X \sim \mathrm{N}(1,4)$ ja $P(a < X < b) = 0.50$. Määritä a ja b siten, että $b - a$ on mahdollisimman pieni.

18. Olkoot Z_1, Z_2 ja Z_3 riippumattomat ja $Z_i \sim \mathrm{N}(0,1)$-jakaumaa, $i = 1,2,3$. Määritellään satunnaismuuttujat $X_i = iZ_i + i$ ja

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3).$$

Määritä vakio a siten, että $P(|\bar{X}| \leq a) = 0.95$. (Vihje: Voit olettaa, että $\bar{X}$ noudattaa normaalijakaumaa.)

19. Olkoon $X \sim$ Khi2(23).

(a) Laske $P(14.85 < X < 32.01)$.

(b) Määritä a ja b siten, että $P(a < X < b) = 0.95$ ja $P(X < a) = 0.025$.

(c) Laske X:n keskiarvo ja varianssi.

20. Oletetaan, että $X \sim \mathrm{Gamma}(\alpha, \beta)$. Osoita gammajakauman momenttifunktion avulla, että $E(X) = \alpha\beta$ ja $\mathrm{Var}(X) = \alpha\beta^2$.

21. Sillalle saapuu autoja Poissonin prosessin mukaan keskimäärin 15 autoa 10:ssä minuutissa. Laske todennäköisyys, että siltamaksujen kerääjä joutuu odottamaan annetusta hetkestä alkaen kahdeksaa autoa (eli kahdeksatta autoa) ainakin puoli tuntia.

22. Oletetaan, että X noudattaa gammajakaumaa Gamma$(3, 2)$. Määritä satunnaismuuttujan $Y = \sqrt{X}$ tiheysfunktio.

23. Logistista jakaumaa noudattavan satunnaismuuttujan X tiheysfunktio on

$$f(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}, \qquad -\infty < x < \infty.$$

Osoita, että satunnaismuuttuja $Y = \frac{1}{1+\mathrm{e}^{-X}}$ noudattaa tasajakaumaa Tas$(0, 1)$.

24. Oletetaan, että $X \sim$ Tas$(-1, 3)$. Määritä satunnaismuuttujan $Y = X^2$ jakauma.

25. Olkoon momenttifunktio $M(t) = (1-2t)^{-12}$, $t < \frac{1}{2}$. Määritä $E(X)$, Var(X) ja $P(15 < X \leq 42)$.

26. Oletetaan, että matkustusaika kotoa töihin noudattaa normaalijakaumaa, jonka keskiarvo on 40 minuuttia ja hajonta 7 minuuttia. Jos haluat 95 %:n todelläköisyydellä olla työpaikalla klo 8:00, niin milloin viimeistään on lähdettävä kotoa?

27. Olkoon X:n momenttifunktio $M(t) = (1-2t)^{-12}$, $t < 1/2$. Laske

(a) $E(X)$,

(b) Var(X) ja

(c) $P(15.66 < X < 42.98)$.

28. Oletetaan, että $X \sim$ Gamma$(3, 1.5)$. Laske

(a) $P(X > 5)$,

(b) jakauman moodi (tiheysfunktion maksimi) sekä

(c) $E(Y)$ ja Var(Y), kun $Y = \frac{1}{X}$. (Ks. Lause 5.1.)

29. Tarkastellaan Poissonin prosessia, jonka intensiteetti on λ. Olkoon W odotusaika, kunnes α tapahtumaa sattuu. Silloin W:n kertymäfunktio on

$$F(w) = 1 - \sum_{x=0}^{\alpha-1} \mathrm{e}^{-\lambda w} \frac{(\lambda w)^x}{x!}.$$

Osoita, että tiheysfunktio $f(w)$ on

$$f(w) = \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!}\, \mathrm{e}^{-\lambda w}.$$

30. Satunnaismuuttujan Z tiheysfunktio on $f_Z(z)$. Olkoon $Y = aZ+b$, missä a ja b ovat annettuja vakioita.

(a) Osoita, että Y:n tiheysfunktio on

$$f_Y(y) = \frac{1}{|a|} f_Z(\frac{y-b}{a}).$$

(b) Esitä Y:n tiheysfunktio, kun $a = 2$, $b = 1$ ja $Z \sim \mathrm{N}(0, 1)$.

31. Olkoon X:n ja Y:n yhteisjakauman tiheysfunktio (Esimerkki 5.18)

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \qquad -1 < x < 1, \quad -1 < y < 1.$$

Määritä X:n ja Y:n jakaumien (reunajakaumien) tiheysfunktiot. Totea, että X ja Y ovat riippumattomat.

32. Oletetaan, että riippumattomat satunnaismuuttujat X ja Y noudattavat tasajakaumaa $\mathrm{Tas}(0, 1)$. Laske todennäköisyydet

(a) $P(|X - Y| \leq \frac{1}{2})$ ja
(b) $P(|\frac{X}{Y} - 1| \leq \frac{1}{2})$.

33. Olkoot X ja Y riippumattomat standardoidut normaalimuuttujat. Määritellään muuttujat

$$U = \frac{X + Y}{\sqrt{2}}; \qquad V = \frac{X - Y}{\sqrt{2}}.$$

Kirjoita U:n ja V:n yhteisjakauman tiheysfunktio. Näytä, että U ja V ovat riippumattomat.

34. Satunnaisvektori (X, Y) noudattaa kaksiulotteista normaalijakaumaa parametrein $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 4$, $\sigma_2^2 = 9$ ja $\rho = 0.8$.

(a) Kirjoita (X, Y):n yhteisjakauman tiheysfunktion lauseke,
(b) $E(Y \mid x)$ ja
(c) $\mathrm{Var}(Y \mid x)$.

(Ks. Lause 5.11.)

35. Erääseen naisten kunto-ohjelmaan osallistuneilta mitattiin kehon rasvaprosentti X ennen ohjelman alkua ja rasvaprosenttin muutos Y ohjelman lopussa. Oletetaan, että muuttujien yhteisjakauma on normaalinen ja $\mu_X = 24.5$, $\sigma_X = 4.8$, $\mu_Y = -0.2$, $\sigma_Y = 3.0$ ja $\rho_{XY} = -0.32$. Laske

(a) $P(1.3 \leq Y \leq 5.8)$,
(b) $E(Y \mid X = x)$,

(c) $\mathrm{Var}(Y \mid X = x)$,

(d) $P(1.3 \leq Y \leq 5.8 \mid X = 18)$.

(Ks. Lause 5.11.)

36. Oletetaan, että satunnaisvektori (X, Y) noudattaa standardimuotoista normaalijakaumaa, missä $\mathrm{Cor}(X, Y) = 0.6$. Laske $P(X-Y < 1,\ X+Y > 2)$.

37. Olkoot X ja Y riippumattomat satunnaismuuttujat, joiden tiheysfunktiot ovat

$$f(x) = \mathrm{e}^{-x}, \qquad f(y) = \mathrm{e}^{-y}, \qquad 0 < x < \infty, \quad 0 < y < \infty.$$

Esitä satunnaisvektorin (U, V) yhteisjakauman tiheysfunktio, kun $U = X - Y$ ja $V = X + Y$.

Luku 6

Otantajakaumien teoria

6.1 Riippumattomat satunnaismuuttujat

Muistamme edellisistä luvuista, että satunnaismuuttujat X_1 ja X_2 ovat riippumattomat (määritelmät 4.6 ja 5.5), jos

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \qquad \text{kaikilla } x_1 \in S_1,\ x_2 \in S_2,$$

missä $f(x_1, x_2)$ on X_1:n ja X_2:n yhteisjakauman tiheysfunktio, $f_1(x_1)$ on X_1:n ja $f_2(x_2)$ on X_2:n tiheysfunktio. Määritelmä yleistyy suoraviivaisesti usean satunnaismuuttujan tapaukseen. Satunnaismuuttujat $X_1,\ X_2,\ \ldots,\ X_n$ *ovat riippumattomat,* jos

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n).$$

Jos satunnaismuuttujat $X_1,\ X_2,\ \ldots,\ X_n$ *ovat riippumattomat,* niin myös niiden funktiot $u_1(X_1),\ u_2(X_2),\ \ldots,\ u_n(X_n)$ ovat riippumattomat, mikäli kukin funktio $u_i,\ i = 1, 2, \ldots, n$ riippuu vain satunnaismuuttujasta X_i eikä siis satunnaismuuttujista $X_j,\ j \neq i$. Silloin erityisesti Lauseen 3.10 mukaan

$$E[u_1(X_1)u_2(X_2)\cdots u_n(X_n)] = E[u(X_1)]\,E[u_2(X_2)]\cdots E[u_n(X_n)].$$

Jos riippumattomat satunnaismuuttujat $X_1,\ X_2,\ \ldots,\ X_n$ noudattavat samaa jakaumaa (RSJ), jonka kertymäfunktio on $F(x)$, niin sanomme, että $X_1, X_2, \ldots, X_n$ on n:n kokoinen otos jakaumasta F. Kertymäfunktio edustaa populaatiota, josta otos tehdään.

Esimerkki 6.1 Olkoon $X_1, X_2, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Silloin $X_i \sim \mathrm{N}(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ ja $X_1,\ X_2,\ \ldots,\ X_n$ ovat riippumattomat.

Otoksen $X_1, X_2, \ldots, X_n$ yhteisjakauman tiheysfunktio on siis

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-[1/(2\sigma^2)](x_i-\mu)^2} \\ &= \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \mathrm{e}^{-[1/(2\sigma^2)]\sum_{i=1}^{n}(x_i-\mu)^2}, \end{aligned}$$

missä

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-[1/(2\sigma^2)](x_i-\mu)^2}, \qquad i = 1, 2, \ldots, n.$$

□

Lause 6.1 (Apulauseen 5.1 yleistys) *Satunnaisvektorit* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ *ja* $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ *ovat riippumattomat jos ja vain jos on olemassa sellaiset funktiot* $g(\boldsymbol{X})$ *ja* $h(\boldsymbol{Y})$*, että*

$$f(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x})h(\boldsymbol{y})$$

kaikilla $\boldsymbol{x}$*:n ja* $\boldsymbol{y}$*:n arvoilla, missä* g *ei riipu* $\boldsymbol{y}$*:stä ja* h *ei riipu* $\boldsymbol{x}$*:stä.*

6.2 Riippumattomien satunnaismuuttujien summan jakauma

Tilastollisissa sovelluksissa tarkastellaan tavallisesti erilaisia satunnaismuuttujien funktioita. Otoksesta $X_1, X_2, \ldots, X_n$ laskettua reaali-, tai vektoriarvoista suuretta $T(X_1, X_2, \ldots, X_n)$ sanotaan otoksen *tunnusluvuksi* (*statistics*). Kaksi tärkeää otoksen tunnuslukua ovat otoskeskiarvo $\overline{X}$ ja otosvarianssi S^2. Esimerkiksi $T_1(X_1, X_2, \ldots, X_n) = \overline{X}$ on reaaliarvoinen ja $T_2(X_1, X_2, \ldots, X_n) = (\overline{X}, S^2)$ on vektoriarvoinen.

Lause 6.2 *Olkoon* $X_1, X_2, \ldots, X_n$ *otos jakaumasta, jonka tiheysfunktio on* $f(x)$*. Silloin satunnaismuuttujien* X_1*,* X_2*, . . . ,* X_n *yhteisjakauman tiheysfunktio on* $f_1(x_1)f_2(x_2)\cdots f_n(x_n)$*. Jos* $g(y)$ *on satunnaismuuttujan* $Y = u(X_1, X_2, \ldots, X_n)$ *tiheysfunktio, niin*

$$\begin{aligned} E(Y) &= \int_{S_y} y g(y)\,\mathrm{d}y \\ &= \int_S \int_S \cdots \int_S u(x_1, x_2, \ldots, x_n) f_1(x_1) f_2(x_2) \cdots f_n(x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_n, \end{aligned}$$

mikäli odotusarvo on olemassa. Diskreettejä satunnaismuuttujia koskeva vastaava tulos saadaan korvaamalla integraalit summalausekkeilla. Satunnaismuuttujien X_1*,* X_2*, . . . ,* X_n *arvoalue on* S *ja* Y*:n arvoalue on* S_y*.*

Esimerkki 6.2 Heitetään kahta noppaa. Olkoon 1. nopan silmäluku X_1 ja 2. nopan silmäluku X_2. Määritetään nyt silmälukujen summan $Y = X_1 + X_2$ todennäköisyysfunktio $g(y)$. Tarkastellaan ensin yksittäisen arvon, esimerkiksi $y = 4$, todennäköisyyden $g(4)$ laskemista. Tapahtuma $\{Y = 4\}$ voi sattua kolmella toisensa poissulkevalla tavalla: $\{X_1 = 1,\ X_2 = 3\}$, $\{X_1 = 2,\ X_2 = 2\}$ ja $\{X_1 = 3,\ X_2 = 1\}$. Siksi

$$\begin{aligned} g(4) &= P(Y = 4) \\ &= P(X_1 = 1,\ X_2 = 3) + P(X_1 = 2,\ X_2 = 2) + P(X_1 = 3,\ X_2 = 1) \\ &= \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36}. \end{aligned}$$

Jatkamalla samalla periaatteella saadaan todennäköisyysfunktio $g(y)$:

y	2	3	4	5	6	7	8	9	10	11	12
$g(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

□

Yleisesti Esimerkin 6.2 todennäköisyydet voidaan laskea ns. *konvoluutiokaavalla*

$$g(y) = P(Y = y) = \sum_{k=1}^{y-1} f(k) f(y - k),$$

missä

$$f(k) = \frac{1}{6}, \qquad k = 1, 2, 3, 4, 5, 6$$

on nopan silmäluvun todennäköisyysfunktio.

Toinen tapa johtaa $g(y)$, on käyttää momenttifunktiota. Nopan silmäluvun momenttifunktio on

$$M_X(t) = E\big(\mathrm{e}^{tX}\big) = \frac{1}{6}\mathrm{e}^{t} + \frac{1}{6}\mathrm{e}^{2t} + \frac{1}{6}\mathrm{e}^{3t} + \frac{1}{6}\mathrm{e}^{4t} + \frac{1}{6}\mathrm{e}^{5t} + \frac{1}{6}\mathrm{e}^{6t}.$$

Koska silmäluvut ovat riippumattomat, niin Y:n momenttifunktio on

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) = [M_X(t)]^2.$$

Koska e^{kt}:n kerroin $M_Y(t)$:n lausekkeessa on todennäköisyys $P(Y = k)$, $k = 2, 3, \ldots, 12$, ne muodostavat Y:n todennäköisyysfunktion.

Lause 6.3 *Olkoot riippumattomien satunnaismuuttujien X_1, X_2, …, X_n odotusarvot μ_1, μ_2, …, μ_n ja varianssit σ_1^2, σ_2^2, …, σ_n^2. Silloin satunnaismuuttujan $Y = \sum_{i=1}^{n} a_i X_i$ odotusarvo ja varianssi ovat*

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \qquad \textit{ja} \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2,$$

missä a_1, a_2, …, a_n ovat annettuja vakioita.

Todistus. Koska odotusarvo on lineaarinen operaattori, niin

$$E(Y) = E\Big(\sum_{i=1}^{n} a_i X_i\Big) = \sum_{i=1}^{n} E(a_i X_i)$$
$$= \sum_{i=1}^{n} a_i\, E(X_i) = \sum_{i=1}^{n} a_i \mu_i.$$

Vastaavasti

$$\sigma^2 = E[(Y-\mu_Y)^2] = E\bigg[\Big(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i\mu_i\Big)^2\bigg]$$
$$= E\Big[\sum_{i=1}^{n} a_i(X_i-\mu_i)\Big]^2 = E\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (X_i-\mu_i)(X_j-\mu_j)\Big]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j\, E[(X_i-\mu_i)(X_j-\mu_j)] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma_{ij},$$

missä

$$\sigma_{ij} = E[(X_i-\mu_i)(X_j-\mu_j)].$$

Koska X_i ja X_j ovat riippumattomat, niin $\sigma_{ij} = 0$, kun $i \neq j$. Tästä seuraa, että

$$\sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

□

Esimerkki 6.3 Olkoot X_1 ja X_2 riippumattomat satunnaismuuttujat, joiden odotusarvot ovat $\mu_1 = -4$ ja $\mu_2 = 3$ sekä varianssit vastaavasti $\sigma_1^2 = 4$ ja $\sigma_2^2 = 9$. Silloin satunnaismuuttujan $Y = 3X_1 - 2X_2$ odotusarvo ja varianssi ovat

$$\mu_Y = 3\cdot(-4) + (-2)\cdot 3 = -18$$

ja

$$\sigma_Y^2 = 3^2\cdot 4 + (-2)^2\cdot 9 = 72.$$

□

Esimerkki 6.4 Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta, jonka odotusarvo on μ ja varianssi σ^2. Silloin otoskeskiarvon

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

odotusarvo ja varianssi ovat

$$\mu_{\overline{X}} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)\mu = \mu \qquad \text{ja} \qquad \sigma_{\overline{X}}^2 = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n}.$$

Otosvarianssi on muotoa

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big),$$

joten

$$E(S^2) = \frac{1}{n-1}\Big[\sum_{i=1}^{n} E(X_i^2) - n\,E(\overline{X}^2)\Big].$$

Koska $E(X_i^2) = \sigma^2 + \mu^2$ ja $E(\overline{X}^2) = \sigma^2/n + \mu^2$, niin laskemalla on helppo todeta, että $E(S^2) = \sigma^2$. Olemme siis osoittaneet, että $\overline{X}$ on μ:n ja S^2 on σ^2:n harhaton estimaattori. □

Lause 6.4 *Jos X_1, X_2, …, X_n ovat riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat $M_{X_i}(t)$, $i = 1, 2, \ldots, n$, niin satunnaismuuttujan $Y = \sum_{i=1}^{n} a_i X_i$ momenttifunktio on*

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t).$$

Todistus. Satunnaismuuttujan Y momenttifunktio on

$$\begin{aligned} M_Y(t) &= E\big(\mathrm{e}^{tY}\big) = E\big(\mathrm{e}^{t(a_1X_1 + a_2X_2 + \cdots + a_nX_n)}\big) \\ &= E\big(\mathrm{e}^{a_1tX_1} \cdot \mathrm{e}^{a_2tX_2} \cdots \mathrm{e}^{a_ntX_n}\big) \\ &= E\big(\mathrm{e}^{a_1tX_1}\big)\,E\big(\mathrm{e}^{a_2tX_2}\big) \cdots E\big(\mathrm{e}^{a_ntX_n}\big), \end{aligned}$$

koska satunnaismuuttujat $\mathrm{e}^{a_itX_i}$ ovat keskenään riippumattomat. Momenttifunktion määritelmän mukaan

$$E\big(\mathrm{e}^{tX_i}\big) = M_{X_i}(t),$$

joten

$$E\big(\mathrm{e}^{a_itX_i}\big) = M_{X_i}(a_it).$$

Siksi

$$M_Y(t) = M_{X_1}(a_1t)M_{X_2}(a_2t)\cdots M_{X_n}(a_nt) = \prod_{i=1}^{n} M_{X_i}(a_it).$$

□

Esimerkki 6.5 Olkoon $X_1, X_2, \ldots, X_n$ otos Bernoullin jakaumasta $\mathrm{Ber}\big(\frac{1}{3}\big)$. Silloin

$$M(t) = \tfrac{2}{3} + \tfrac{1}{3}\mathrm{e}^t.$$

Jos $Y = X_1 + X_2 + \cdots + X_n$, niin

$$M_Y(t) = \prod_{i=1}^{n}\big(\tfrac{2}{3} + \tfrac{1}{3}\mathrm{e}^t\big) = \big(\tfrac{2}{3} + \tfrac{1}{3}\mathrm{e}^t\big)^n.$$

Tästä näemme, että $Y \sim \mathrm{Bin}\big(n, \frac{1}{3}\big)$. □

Seuraus 6.1 *Jos $X_1, X_2, \ldots, X_n$ on otos jakaumasta, jonka momenttifunktio on $M(t)$, niin*

1. *satunnaismuuttujan $Y = \sum_{i=1}^n X_i$ momenttifunktio on*

$$M_Y(t) = \prod_{i=1}^n M(t) = [M(t)]^n.$$

2. *otoskeskiarvon $\overline{X} = \sum_{i=1}^n (1/n)X_i$ momenttifunktio on*

$$M_{\overline{X}}(t) = \prod_{i=1}^n M\left(\frac{t}{n}\right) = \left[M\left(\frac{t}{n}\right)\right]^n.$$

Esimerkki 6.6 Olkoon X_1, X_2, X_3 otos eksponenttijakaumasta, jonka odotusarvo on θ. Eksponenttijakauman momenttifunktio on $M(t) = 1/(1-\theta t)$, $t < 1/\theta$. Silloin summan $Y = X_1 + X_2 + X_3$ momenttifunktio on

$$M_Y(t) = [1/(1-\theta t)]^3 = (1-\theta t)^{-3}, \qquad t < \frac{1}{\theta},$$

mikä on gammajakauman Gamma$(3, \theta)$ momenttifunktio, joten $Y \sim$ Gamma$(3, \theta)$. Toisaalta $\overline{X}$:n momenttifunktio on

$$M_{\overline{X}}(t) = \left[\left(1 - \frac{\theta t}{3}\right)^{-1}\right]^3 = \left(1 - \frac{\theta t}{3}\right)^{-3}, \qquad \text{kun } t < \frac{3}{t}.$$

Otoskeskiarvo $\overline{X}$ noudattaa siis gammajakaumaa Gamma$(3, \theta/3)$. □

6.3 Normaalijakaumaan liittyvät jakaumat

Lause 6.5 *Jos $X_1, X_2, \ldots, X_n$ on otos normaalijakaumasta* N(μ, σ^2)*, niin otoskeskiarvon $\overline{X} = \sum_{i=1}^n (1/n)X_i$ jakauma on* N$(\mu, \sigma^2/n)$.

Todistus. Koska $X_i \sim$ N(μ, σ^2), niin

$$M_{X_i}(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Seurauslauseen 6.1 mukaan

$$\begin{aligned} M_{\overline{X}}(t) &= \left[\exp\left(\mu \cdot \frac{t}{n} + \frac{\sigma^2 (t/n)^2}{2}\right)\right]^2 \\ &= \exp\left(\mu t + \frac{(\sigma^2/n) t^2}{2}\right), \end{aligned}$$

joka on normaalijakauman N$(\mu, \sigma^2/n)$ momenttifunktio. Koska momenttifunktio määrittää yksikäsitteisesti satunnaismuuttujan jakauman, niin $\overline{X} \sim$ N$(\mu, \sigma^2/n)$. □

Lause 6.6 *Olkoot $X_i \sim \chi^2(r_i)$, $i = 1, 2, \ldots, n$. Jos X_1, X_2, …, X_n ovat riippumattomat, niin satunnaismuuttujan $Y = X_1 + X_2 + \cdots + X_n$ jakauma on $\chi^2(r_1 + r_2 + \cdots + r_k)$.*

Todistus. Satunnaismuuttujan Y momenttifunktio voidaan kirjoittaa muodossa

$$\begin{aligned} M_Y(t) &= E\left[\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}\right] \\ &= E\left(\mathrm{e}^{tX_1}\right) \cdot E\left(\mathrm{e}^{tX_2}\right) \cdots E\left(\mathrm{e}^{tX_n}\right) \\ &= \prod_{i=1}^{n} M_{X_i}(t). \end{aligned}$$

Koska

$$M_{X_i}(t) = (1-2t)^{-r_i/2}; \qquad t < \frac{1}{2},$$

niin

$$M_Y(t) = \prod_{i=1}^{n}(1-2t)^{-r_i/2} = (1-2t)^{-(r_1+r_2+\cdots+r_n)/2}, \qquad t < \frac{1}{2}$$

on χ^2-jakauman $\chi^2(r_1 + r_2 + \cdots + r_n)$ momenttifunktio. Tästä seuraa, että $Y \sim \chi^2(r_1 + r_2 + \cdots + r_n)$. □

Lause 6.7 *Olkoon $Z_1, Z_2, \ldots, Z_n$ otos standardimuotoisesta normaalijakaumasta $\mathrm{N}(0,1)$. Silloin $W = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ noudattaa jakaumaa $\chi^2(n)$.*

Todistus. Koska $Z_i \sim \chi^2(1)$, $i = 1, 2, \ldots, n$ ja Z_1^2, Z_2^2, …, Z_n^2 ovat keskenään riippumattomat, niin tulos seuraa Lauseesta 6.6. □

Seuraus 6.2 *Olkoot X_1, X_2, …, X_n riippumattomat ja $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$. Silloin satunnaismuuttuja*

$$W = \sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\sigma_i^2}$$

noudattaa jakaumaa $\chi^2(n)$.

Lause 6.8 *Olkoon $X_1, X_2, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, $\overline{X} = \sum_{i=1}^{n}(1/n)X_i$ on otoskeskiarvo ja $S^2 = [1/(n-1)]\sum_{i=1}^{n}(X_i - \overline{X})^2$ on otosvarianssi. Silloin*

1. *$\overline{X}$ ja S^2 ovat riippumattomat satunnaismuuttujat,*
2. *$\overline{X}$ noudattaa normaalijakaumaa $\mathrm{N}(\mu, \sigma^2/n)$,*
3. *$(n-1)S^2/\sigma^2$ noudattaa χ^2-jakaumaa $\chi^2(n-1)$.*

Todistus. Kohdan 1 todistus sivuutetaan tässä yhteydessä. Kohta 2 on lause 6.5. Todistetaan nyt väite, että

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Laskemalla voidaan todeta, että

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \sum_{i=1}^{n}\left[\frac{(X_i-\overline{X})+(\overline{X}-\mu)}{\sigma}\right]^2 \\ &= \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 + \frac{n(\overline{X}-\mu)^2}{\sigma^2} \\ &= \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.\end{aligned}$$

Koska $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$, niin $Z^2 \sim \chi^2(1)$. Vastaavasti Seurauslauseen 6.2 mukaan $W = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$.

Koska S^2 ja Z^2 ovat kohdan 1 mukaan riippumattomat, niin

$$\begin{aligned}E\big(\mathrm{e}^{tW}\big) &= E\big(\mathrm{e}^{t[(n-1)S^2/\sigma^2+Z^2]}\big) = E\big(\mathrm{e}^{t(n-1)S^2/\sigma^2}\cdot\mathrm{e}^{tZ^2}\big) \\ &= E\big(\mathrm{e}^{t(n-1)S^2/\sigma^2}\big)\,E\big(\mathrm{e}^{tZ^2}\big).\end{aligned}$$

Koska $W \sim \chi^2(n)$ ja $Z^2 \sim \mathrm{N}(0,1)$, niin

$$(1-2t)^{-n/2} = E\big[\mathrm{e}^{t(n-1)S/\sigma^2}\big]\cdot(1-2t)^{-1/2}.$$

Tästä seuraa, että

$$E\big[\mathrm{e}^{t(n-1)S/\sigma^2}\big] = (1-2t)^{-(n-1)/2}; \qquad t < \frac{1}{2},$$

joka on jakauman $\chi^2(n-1)$ momenttifunktio. Näin on lauseen väite 3 todistettu. □

Lause 6.9 *Olkoot X_1, X_2, …, X_n keskenään riippumattomat normaalijakaumaa noudattavat satunnaismuuttujat, joiden odotusarvot ovat μ_1, μ_2, …, μ_n ja varianssit σ_1^2, σ_2^2, …, σ_n^2. Silloin lineaarikombinaatio*

$$Y = \sum_{i=1}^{n} a_i X_i$$

noudattaa normaalijakaumaa

$$\mathrm{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Todistus. Tulos saadaan soveltamalla Lausetta 6.4 normaalijakaumaan. □

6.4 Järjestyssuureet

Otoksen suurin ja pienin arvo sekä keskimääräinen arvo, mediaani, ovat tärkeitä otossuureiden arvojen järjestykseen perustuvia tunnuslukuja. Olkoon $X_1, X_2, \ldots, X_n$ otos. Merkitään otoksen pienintä arvoa $X_{(1)}$ seuraavaksi pienintä $X_{(2)}$ ja niin edelleen, joten

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Tämä indeksointi tarkoittaa sitä, että otosarvot pannaan kasvavaan järjestykseen. Jos otos on esimerkiksi 5.0, 3.1, 2.7, 6.1, 5.3, niin järjestetty otos on 2.7, 3.1, 5.0, 5.3, 6.1. Nyt siis esimerkiksi $X_1 = 5.0$, $X_{(1)} = 2.7$ ja $X_{(3)} = 5.0$ on mediaani ja $X_3 = 2.7$. Nyt siis

$$X_{(1)} = \min(X_1, \ldots, X_n)$$

ja

$$X_{(n)} = \max(X_1, \ldots, X_n).$$

Tunnusluku $X_{(k)}$ on otoksen k. järjestystunnusluku.

6.4.1 Maksimi ja minimi

Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta, jonka kertymäfunktio on $F(x)$. Maksimin kertymäfunktio on

$$\begin{aligned} F_{(n)}(x) &= P(X_{(n)} \leq x) \\ &= P(X_1 \leq x,\ X_2 \leq x,\ \ldots,\ X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x)\cdots P(X_n \leq x), \end{aligned}$$

koska X_1, X_2, …, X_n ovat riippumattomat. Kertymäfunktion määritelmän mukaan $P(X_i \leq x) = F(x)$, joten

$$F_{(n)}(x) = [F(x)]^n.$$

Minimin kertymäfunktio on

$$\begin{aligned} F_{(1)}(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x,\ X_2 > x,\ \ldots,\ X_n > x) \\ &= 1 - P(X_1 > x)P(X_2 > x)\cdots P(X_n > x) \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

Esimerkki 6.7 Olkoon $X_1, X_2, \ldots, X_n$ otos eksponenttijakaumasta $\mathrm{Exp}(\lambda)$. Määritetään minimin $X_{(1)}$ jakauma. Eksponenttijakauman $\mathrm{Exp}(\lambda)$ kertymäfunktio on

$$F(x) = \begin{cases} 0, & \text{kun } x < 0; \\ 1 - \mathrm{e}^{\lambda x}, & \text{kun } x \geq 0. \end{cases}$$

Silloin minimin kertymäfunktio on

$$F_{(1)}(x) = \begin{cases} 0, & \text{kun } x < 0; \\ 1 - \mathrm{e}^{-n\lambda x}, & \text{kun } x \geq 0. \end{cases}$$

Minimi noudattaa siis eksponenttijakaumaa $\mathrm{Exp}(n\lambda)$. □

Jos otos on jatkuvasta jakaumasta, jonka tiheysfunktio on $f(x)$, saadaan $X_{(1)}$:n ja $X_{(n)}$:n jakaumien tiheysfunktiot derivoimalla kertymäfunktiot $F_{(n)}(x)$ ja $F_{(1)}(x)$. Nyt siis maksimin tiheysfunktio on

$$f_{(n)}(x) = \frac{\mathrm{d}}{\mathrm{d}x}[F(x)]^n = n[F(x)]^{n-1} f(x)$$

ja minimin tiheysfunktio on

$$f_{(1)}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\big(1 - [1 - F(x)]^n\big) = n[1 - F(x)]^{n-1} f(x).$$

6.4.2 Järjestyssuureen $X_{(k)}$ jakauma

Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta, jonka kertymäfunktio on $F(x)$. Johdetaan nyt järjestystunnusluvun $X_{(k)}$, $1 < k < n$, jakauma. Jos $\{X_{(k)} \leq x\}$, niin silloin ainakin k otosarvoa on pienempiä tai korkeintaan yhtä suuria kuin x. Tarkastellaan yksinkertaisuuden vuoksi tapausta $n = 3$. Johdetaan mediaanin $X_{(2)}$ jakauma. Tapahtuma $\{X_{(2)} \leq x\}$ toteutuu täsmälleen silloin, kun $\{X_1 \leq x,\ X_2 \leq x\}$ tai $\{X_1 \leq x,\ X_3 \leq x\}$ tai $\{X_2 \leq x,\ X_3 \leq x\}$ tai $\{X_1 \leq x,\ X_2 \leq x,\ X_3 \leq x\}$. Koska

$$P(X_i \leq x,\ X_j \leq x) = [F(x)]^2[1 - F(x)], \qquad i \neq j$$

ja

$$P(X_1 \leq x,\ X_2 \leq x,\ X_3 \leq x) = [F(x)]^3,$$

niin

$$\begin{aligned} F_{(2)}(x) = P(X_{(2)} \leq x) &= 3[F(x)]^2[1 - F(x)] + [F(x)]^3 \\ &= \sum_{i=2}^{3} \binom{3}{i} [F(x)]^i [1 - F(x)]^{3-i}. \end{aligned} \tag{6.4.1}$$

Yleisessä tapauksessa vastaava kaava voidaan johtaa samalla periaatteella. Emme kuitenkaan käsittele yleisen kaavan johtoa sen tarkemmin, toteamme vain, että $X_{(k)}$:n kertymäfunktio on

$$F_{(k)}(x) = P(X_{(k)} \leq x) = \sum_{i=k}^{n} \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i}. \tag{6.4.2}$$

Jos otos on jatkuvasta jakaumasta, saadaan vastaava tiheysfunktio derivoimalla kertymäfunktio. Esitetään ensin $X_{(2)}$:n tiheysfunktio, kun $n = 3$. Kun kertymäfunktio (6.4.1) derivoidaan, saadaan

$$\begin{aligned} f_{(2)}(x) = F'_{(2)}(x) &= 3 \cdot 2F(x)f(x)[1 - F(x)] - 3[F(x)]^2 f(x) + 3[F(x)]^2 f(x) \\ &= 3!F(x)[1 - F(x)]f(x). \end{aligned}$$

Derivoimalla lauseke (6.4.2) saadaan satunnaismuuttujan $X_{(k)}$ tiheysfunktio yleisessä tapauksessa ($1 \leq k \leq n$):

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1 - F(x)]^{n-k}f(x). \tag{6.4.3}$$

Esimerkki 6.8 Olkoon X_1, X_2, X_3, X_4, X_5 otos jakaumasta, jonka tiheysfunktio on $f(x) = 2x$, $0 < x < 1$. Jakauman kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ x^2, & 0 \leq x \leq 1; \\ 1, & x > 1. \end{cases}$$

Silloin mediaanin tiheysfunktio on lausekkeen (6.4.3) nojalla

$$f_{(3)}(x) = \frac{5!}{2!\,2!}x^4(1 - x^2)^2 \cdot 2x = 60x^5(1 - x^2)^2, \qquad 0 < x < 1.$$

Vastaavasti minimin tiheysfunktio on

$$f_{(1)}(x) = 10x(1 - x^2)^4, \qquad 0 < x < 1$$

ja maksimin tiheysfunktio on

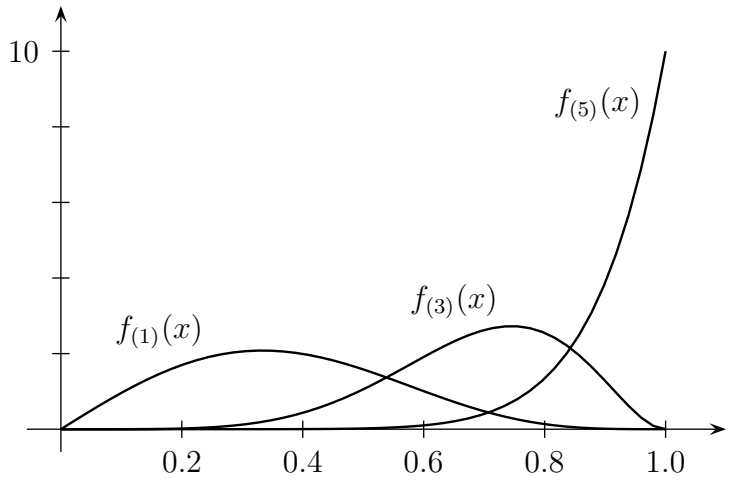
$$f_{(5)}(x) = 10x^9, \qquad 0 < x < 1.$$

□

6.5 Keskeinen rajaväittämä

Olemme havainneet, että otossuureen jakauma riippuu tavallisesti otoskoosta n. Jos $X_1, X_2, \ldots, X_n$ on otos Bernoullin jakaumasta Ber(p), niin $X = X_1 + X_2 + \cdots + X_n \sim \mathrm{Bin}(n, p)$. Satunnaismuuttujan X jakauma riippuu siis otoskoosta n. Jos $X_1, X_2, \ldots, X_n$ on otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, niin otoskeskiarvon $\overline{X}$ jakauma $\mathrm{N}(\mu, \sigma^2/n)$ riippuu n:stä.

Olkoon $(X_i; i \geq 1) = X_1, X_2, X_3, \ldots$ satunnaismuuttujien jono, missä satunnaismuuttujien X_n, $n = 1, 2, 3, \ldots$ jakauma riippuu n:stä. Merkitään satunnaismuuttujan X_n kertymäfunktiota $F_n(x)$, joka siis riippuu n:stä. Seuraavassa määritellään satunnaismuuttujien jonon $(X_i; i \geq 1)$ *suppeneminen jakaumamielessä.*



Kuvio 6.1. Minimin, maksimin ja mediaanin tiheysfunktio, kun otos on jakaumasta $f(x) = 2x$, $0 < x < 1$.

Määritelmä 6.1 Satunnaismuuttujien $X_1, X_2, X_3, \ldots$ jono *suppenee jakaumaltaan* kohti satunnaismuuttujaa X, jos $\lim_{n\to\infty} F_n(x) = F(x)$ kaikissa pisteissä x, joissa $F(x)$ on jatkuva.

Kun jono $\{X_n\}$ suppenee jakaumaltaan kohti satunnaismuuttujaa X, merkitään $X_n \xrightarrow{d} X$. Momenttifunktioiden yhteydessä esitettiin momenttifunktion ja ja jakauman (kertymäfunktion) yksikäsitteistä vastaavuutta koskeva Lause 3.12. Samassa yhteydessä esitettiin myös momenttifunktioiden suppenemista koskeva Lause 3.15, jota voidaan soveltaa raja-jakaumien määrittämiseen.

Merkitään satunnaismuuttujien X_1, X_2, $\ldots$, X_n summaa ja keskiarvoa seuraavasti:

$$S_n = \sum_{i=1}^{n} X_i \qquad \text{ja} \qquad \overline{X}_n = \frac{S_n}{n}$$

Lause 6.10 (Keskeinen rajaväittämä) *Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta, jonka keskiarvo on μ ja varianssi σ^2. Merkitään*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Silloin Z_n:n jakauma lähenee normaalijakaumaa $\mathrm{N}(0,1)$*, kun $n \to \infty$.*

Keskeisen rajaväittämän mukaan riippumattomien satunnaismuuttujien summa noudattaa likimain normaalijakaumaa, kun n on suuri. Merkitsemme

$$Z_n \simeq \mathrm{N}(0,1),$$

kun n on suuri. Merkki $\simeq$ tarkoittaa ”noudattaa likimain jakaumaa”. Käytännössä keskeisen rajaväittämän avulla voidaan arvioida Z_n:n jakaumaa, kun

n on riittävän suuri. Silloin

$$P(Z_n \leq z) \approx \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}\,\mathrm{d}x = \Phi(z),$$

missä $\Phi(z)$ on normitetun normaalijakauman kertymäfunktio. Voimme merkitä saman asian myös seuraavasti:

$$P(Z_n \leq z) \to \Phi(z),$$

kun $n \to \infty$.

Esimerkki 6.9 Olkoon $X_1, X_2, \ldots, X_{15}$ otos jakaumasta, jonka tiheysfunktio on $f(x) = \left(\frac{3}{2}\right)x^2$, $-1 < x < 1$. Jakauman odotusarvo $\mu = 0$ ja varianssi $\sigma^2 = 3/5$. Esimerkiksi todennäköisyys $P(\overline{X} \leq 0.15)$ voidaan laskea johtamalla ensin $\overline{X}$:n jakauma ja määrittämällä siitä kysytty todennäköisyys. Keskeisen rajaväittämän avulla saadaan tämän todennäköisyyden tarkka arvio ilman tietoa $\overline{X}$:n tarkasta jakaumasta:

$$\begin{aligned} P(\overline{X} \leq 0.15) &= P\left(\frac{\overline{X} - 0}{\sqrt{3/5}\,/\sqrt{15}} \leq \frac{0.15 - 0}{\sqrt{3/5}\,/\sqrt{15}}\right) \\ &= P(Z_{15} \leq 0.75) \\ &\approx \Phi(0.75) = 0.7734. \end{aligned}$$

Arvion tarkkuudesta keskeinen rajaväittämä ei kuitenkaan anna käsitystä. □

6.6 Jakaumien likiarvot normaalijakauman avulla

Olkoon $X_1, X_2, \ldots, X_n$ otos Bernoullin jakaumasta $\mathrm{Ber}(p)$. Silloin $S_n = X_1 + X_2 + \cdots + X_n$ noudattaa binomijakaumaa $\mathrm{Bin}(n, p)$. Keskeisen rajaväittämän mukaan

$$Z_n = \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

noudattaa likimain normaalijakaumaa $\mathrm{N}(0, 1)$, kun n on suuri. Tuloksen mukaan binomijakauma lähenee normaalijakaumaa, kun n kasvaa. Peukalosääntönä voidaan pitää, että n on riittävän suuri, kun $np \geq 5$ ja $n(1-p) \geq 5$. Mitä enemmän p poikkeaa 0.5:stä, sitä suurempi n tarvitaan.

Esimerkki 6.10 Oletetaan, että $X \sim \mathrm{Bin}(10, 0.5)$. Lasketaan todennäköisyys $P(3 \leq X < 6)$. Voidaan kirjoittaa

$$P(3 \leq X < 6) = P(2.5 \leq X \leq 5.5).$$

Arvioidaan nyt jälkimmäistä todennäköisyyttä keskeisen rajaväittämän nojalla normaalijakauman avulla. Silloin

$$P(2.5 \leq X \leq 5.5) = P\left(\frac{2.5-5}{\sqrt{10/4}} \leq \frac{X-5}{\sqrt{10/4}} \leq \frac{5.5-5}{\sqrt{10/4}}\right)$$
$$\approx \Phi(0.316) - \Phi(-1.581) = 0.5670.$$

Tarkka todennäköisyys binomijakauman avulla on $P(3 \leq X < 6) = 0.5683$. □

6.7 t-jakauma ja F-jakauma

Oletetaan, että $X_1, X_2, \ldots, X_n$ on otos jakaumasta $\mathrm{N}(\mu, \sigma^2)$, jonka varianssi σ^2 tunnetaan. Tarkastellaan lauseketta

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}},$$

joka tunnetaan t-testisuureena. Tiedämme, että

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

ja Lauseen 6.8 mukaan

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Lisäksi Lauseen 6.8 mukaan Z ja U ovat riippumattomat. Tällainen satunnaismuuttuja noudattaa t-jakaumaa vapausastein $r = n-1$. Alaluvussa 5.9.2 esitettiin t-jakauman tiheysfunktio.

Usein halutaan verrata kahden normaalijakauman $\mathrm{N}(\mu_1, \sigma_1^2)$ ja $\mathrm{N}(\mu_2, \sigma_2^2)$ variansseja. Teemme n_1:n kokoisen otoksen jakaumasta $\mathrm{N}(\mu_1, \sigma_1^2)$ ja n_2:n kokoisen otoksen jakaumasta $\mathrm{N}(\mu_2, \sigma_2^2)$. Oletetaan, että otokset ovat toisistaan riippumattomat. Olkoot S_1^2 ja S_2^2 näistä eri otoksista lasketut otosvarianssit. Lauseen 6.8 mukaan

$$U = (n_1-1)\frac{S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \qquad \text{ja} \qquad V = (n_2-1)\frac{S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

Koska otokset ovat keskenään riippumattomat, niin satunnaismuuttujat U ja V ovat riippumattomat. Varianssien yhtäsuuruutta voidaan testatata tarkastelemalla suhdetta

$$F = \frac{U/r_1}{V/r_2}, \tag{6.7.1}$$

missä $r_1 = n_1-1$ ja $r_2 = n_2-1$. Alaluvussa 5.9.2 osoitettiin, että suhde (6.7.1) noudattaa F-jakaumaa vapausastein r_1 ja r_2.

6.8 Momenttifunktion rajafunktiot

Tarkastelemme nyt satunnaismuuttujan (usein otoksen tunnusluku) jakauman riippuvuutta otoskoosta n. Otoskeskiarvo ja otosvarianssi ovat tavallisimmat otoksesta lasketut tunnusluvut. Oletetaan esimerkiksi, että $X_n \sim \text{Bin}(n, p)$. Eri n:n arvoilla saamme eri binomijakauman. Miten jakauma muuttuu n:n kasvaessa? Olemme keskeisen rajaväittämän avulla jo osoittaneet, että $\text{Bin}(n, p)$ lähenee normaalijakaumaa, kun n kasvaa.

Voimme tutkia $\text{Bin}(n, p)$:n rajajakaumaa myös ehdolla, että jakauman odotusarvo np pidetään vakiona λ. Jos $np = \lambda$ on vakio ja $n \to \infty$, niin $p \to 0$. Satunnaismuuttujan $X_n \sim \text{Bin}(n, p)$ momenttifunktio on

$$M_n(t) = (1 - p + p\mathrm{e}^t)^n.$$

Koska $p = \lambda/n$, niin

$$M_n(t) = \left[1 - \frac{\lambda}{n} + \frac{\lambda}{n}\mathrm{e}^t\right]^n = \left[1 + \frac{\lambda(\mathrm{e}^t - 1)}{n}\right]^n.$$

Käyttäen hyväksi analyysin tulosta

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = \mathrm{e}^a,$$

saadaan

$$\lim_{n\to\infty} M_n(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)} = M(t),$$

joka on olemassa kaikilla $t \in \mathbb{R}$. Koska

$$M(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}$$

on Poissonin jakauman $\text{Poi}(\lambda)$ momenttifunktio, niin Lauseen 3.15 mukaan X_n:n jakauma lähestyy siis Poissonin jakaumaa $\text{Poi}(\lambda)$, kun $n \to \infty$.

6.9 Suppenemiskäsitteet

Olemme edellä jo useaan otteeseen tutustuneet suppenemiseen jakaumamielessä. Käsite määriteltiin alaluvussa 4.3.4 (Määritelmä 4.2). Satunnaismuuttujien jono $(X_n, n \geq 1) = (X_1, X_2, \ldots)$ suppenee jakaumaltaan kohti satunnaismuuttujaa X ($X_n \xrightarrow{d} X$, kun $n \to \infty$), jos

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

kaikssa pisteissä x, joissa $F_X(x)$ on jatkuva. Tässä yhteydessä on myös syytä muistaa, että momenttifunktioiden jonon suppenemisesta seuraa vastaavien jakaumien suppeneminen jakaumamielessä.

Esimerkki 6.11 Olkoon $\{X_n\}$ sellainen satunnaismuuttujien jono, että

$$p_n(x) = P(X_n = x) = \begin{cases} 1, & \text{kun } x = 2 + \frac{1}{n}; \\ 0, & \text{kun } x \neq 2 + \frac{1}{n}. \end{cases}$$

Huomaa, että $p_n(2) = 0$ kaikilla n. Tästä seuraa, että $p_n(x) \to p(x)$, missä $p(x) = 0$ kaikilla x. Satunnaismuuttujan X_n kertymäfunktio on muotoa

$$F_n(x) = \begin{cases} 0, & \text{kun } x < 2 + \frac{1}{n}; \\ 1, & \text{kun } x \geq 2 + \frac{1}{n}. \end{cases}$$

Kun $n \to \infty$, niin $F_n(x) \to F(x)$, missä

$$F(x) = \begin{cases} 0, & x < 2; \\ 1, & x \geq 2. \end{cases}$$

$F(x)$ on pisteeseen $x = 2$ degeneroituneen jakauman kertymäfunktio eli $P(X = 2) = 1$. Todennäköisyysfunktioiden $p_n(x)$ jono ei kuitenkaan suppene kohti tämän jakauman todennäköisyysfunktiota. □

Olkoon $\{X_n\}$ jono satunnaismuuttujia, joiden odotusarvo on μ ja varianssi σ^2. Silloin keskeisen rajaväittämän mukaan

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

missä $Z \sim \mathrm{N}(0, 1)$. Huomattakoon, että Z_n:n jakaumat ovat usein diskreettejä, mutta silti rajajakauma on normaalijakauma. Kun n on riittävän suuri, niin

$$P\left(a \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a).$$

Jos esimerkiksi $X_n \sim \mathrm{Bin}(n, p)$, niin silloin keskeisen rajaväittämän mukaan

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} Z,$$

missä $Z \sim \mathrm{N}(0, 1)$. Tätä tulosta kutsutaan *De Moivren ja Laplacen* lauseeksi.

Osoitimme alavuvussa 3.4 Tšebyševin epäyhtälön avulla, että otoskeskiarvo $\overline{X}$ on hyvä populaation keskiarvon tunnusluku. Tarkastelu ei perustunut suppenemiseen jakaumamielessä vaan ns. stokastiseen suppenemiseen.

Määritelmä 6.2 Satunnaismuuttujien jono $\{X_n\}$ suppenee stokastisesti kohti satunnaismuuttujaa X, jos kaikilla $\varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

tai yhtäpitävästi

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Stokastista suppenemista sanotaan myös suppenemiseksi todennäköisyyden mielessä ja merkitään $X_n \xrightarrow{P} X$. Usein tarkastellaan tilannetta, että satunnaismuuttuja, jota lähestytään, on vakio. Tällainen tilanne on heikossa suurten lukujen laissa (Lause 3.11, HSLL). Esitetty heikon suurten lukujan lain todistus oli sillä tavalla yleinen, että se on pätevä myös jatkuville satunnaismuuttujille. HSLL sanoo, että otoskeskiarvo suppenee stokastisesti kohti populaation keskiarvoa, kun otoskoko kasvaa.

Olkoon $\{X_n\}$ sellaisten satunnaismuuttujien jono, että $E(X_n) = \mu$ ja $\mathrm{Var}(X_n) = \sigma^2$. Heikon suurten lukujen lain mukaan

$$\overline{X}_n \xrightarrow{P} \mu,$$

missä $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n$. Lause todistettiin Tšebyševin epäyhtälön avulla.

Esimerkki 6.12 Olkoon $\{X_n\}$ jono sellaisia diskreettejä satunnaismuuttujia, että

$$P(X_n = 1) = \frac{1}{n} \qquad \text{ja} \qquad P(X_n = 0) = 1 - \frac{1}{n}.$$

Silloin

$$P(|X_n| > \varepsilon) = \begin{cases} P(X_n = 1) = \dfrac{1}{n}, & \text{kun } 0 < \varepsilon < 1; \\ 0, & \text{kun } \varepsilon \geq 1. \end{cases}$$

Tästä nähdään, että $P(|X_n| > \varepsilon) \to 0$, kun $n \to \infty$. Voimme siis sanoa, että $X_n \xrightarrow{P} 0$. □

Esimerkki 6.13 (Otosvarianssin tarkentuvuus) Olkoon $\{X_n\}$ sellainen satunnaismuuttujien jono, että $E(X_n) = \mu$ ja $\mathrm{Var}(X_n) = \sigma^2 < \infty$. Otosvarianssi on

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Tiedämme, että $E(S_n^2) = \sigma^2$. Tšebyševin epäyhtälön mukaan

$$P(|S_n^2 - \sigma^2| \geq \varepsilon) \leq \frac{E(S_n^2 - \sigma^2)}{\varepsilon^2} = \frac{\mathrm{Var}(S_n^2)}{\varepsilon^2}.$$

Jos nyt $\mathrm{Var}(S_n^2) \to 0$, kun $n \to \infty$, niin $\lim_{n\to\infty} P(|S_n^2 - \sigma^2| \geq \varepsilon) = 0$ ja $(S_n^2, n \geq 1)$ suppenee stokastisesti kohti populaation varianssia. □

Tässä yhteydessä on tietysti luonnollista kysyä, miten stokastinen suppeneminen ja suppeneminen jakaumamielessä suhteutuvat toisiinsa. Voidaan osoittaa, että stokastinen suppeneminen implikoi suppenemisen jakaumamielessä. Jos siis $X_n \xrightarrow{P} X$, niin $X_n \xrightarrow{d} X$. Jos jono $\{X_n\}$ suppenee kohti vakiota μ, niin silloin $X_n \xrightarrow{P} \mu$ jos ja vain jos $X_n \xrightarrow{d} \mu$.

Rajoitumme tässä esityksessä kahteen edellä esitettyyn suppenemiskäsitteeseen: stokastiseen suppenemiseen ja suppenemiseen jakaumamielessä. Esitämme kuitenkin vielä ns. melkein varman (m.v.) suppenemisen.

Määritelmä 6.3 Jono $\{X_n\}$ suppenee melkein varmasti kohti satunnaismuuttujaa X, jos

$$P\big(\lim_{n\to\infty} |X_n - X| < \varepsilon\big) = 1.$$

Näennäisesti määritelmä muistuttaa stokastisen suppenemisen määritelmää, vaikka käsitteet ovat sisällöllisesti erilaisia.

6.10 Estimaattorit

6.10.1 Estimaattoreiden ominaisuuksia

Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta, jonka tiheysfunktio on $f(x, \theta)$. Jos haluamme estimoida jakauman tunnuslukua θ jollakin otoksen tunnusluvulla, merkitsemme usein tätä otoksen tunnuslukua $\hat{\theta}$. On siis muistettava, että $\hat{\theta}$ on otoksen funktio ja täydellisempi merkintä olisi $\hat{\theta}(X_1, X_2, \ldots, X_n)$. Havaitusta otoksesta $x_1, x_2, \ldots, x_n$ laskettua estimaattorin arvoa $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ sanotaan estimaatiksi. Olemme edellä jo tarkastelleet useita estimaattoreita. Tavanomaisia odotusarvon μ ja varianssin σ^2 estimaattoreita ovat otoskeskiarvo $\overline{X}$ ja otosvarianssi S^2, eli $\hat{\mu} = \overline{X}$ ja $\hat{\sigma}^2 = S^2$.

Määritelmä 6.4 Estimaattori $\hat{\theta}$ on parametrin *harhaton estimaattori*, jos $E(\hat{\theta}) = \theta$ kaikilla θ:n arvoilla. Muutoin $\hat{\theta}$ on harhainen ja $\hat{\theta}$:n harha on $\text{harha}(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Olemme jo aikaisemmin osoittaneet, että $\hat{\mu} = \overline{X}$ ja $\hat{\sigma}^2 = S^2$ ovat harhattomia estimaattoreita.

Eräs intuitiivisesti hyväksyttävä estimaattorille asetettava vaatimus on, että se antaa 'tarkempia' estimaatteja kun otoskoko kasvaa. Tarkan estimaattorin arvot osuvat suurella todennäköisyydellä lähelle parametrin θ oikeata arvoa. *Tarkentvuvuus* sisältää tämän ajatuksen.

Määritelmä 6.5 Tunnusluku $\hat{\theta}$ on parametrin θ *tarkentuva estimaattori*, jos $\hat{\theta} \xrightarrow{P} \theta$, kun otoskoko n kasvaa rajatta.

Selvempi olisi merkitä $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n)$, missä $(\hat{\theta}_n; n \geq 1)$ on satunnaismuuttujien jono. Jos $\hat{\theta}_n$ on θ:n tarkentuva estimaattori, niin jono $(\hat{\theta}_n; n \geq 1)$ suppenee stokastisesti kohti parametrin arvoa θ.

Määritelmä 6.6 Estimaattorin $\hat{\theta}$ keskineliövirhe (MSE = Mean Square Error) on

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

Määritelmästä seuraa suoraviivaisesti, että

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{harha}(\hat{\theta})]^2.$$

Voidaan osoittaa, että $\hat{\theta}$ on θ:n tarkentuva estimaattori, jos $\text{MSE}(\hat{\theta}) \to 0$ otoskoon n kasvaessa rajatta.

6.10.2 Delta-menetelmä

Määritelmä 6.7 Funktion $g(x)$ r. asteen *Taylorin polynomi* pisteessä a on

$$T_r(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots + \frac{g^{(r)}(a)}{r!}(x-a)^r, \tag{6.10.1}$$

missä $g^{(r)}(x) = \frac{\mathrm{d}^r}{\mathrm{d}x^r} g(x)$ on funktion $g(x)$ r. derivaatta.

Taylorin lauseen mukaan

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0, \tag{6.10.2}$$

jos $g^{(r)}(a)$ on olemassa. Funktio $g(x)$ voidaan lausua pisteen $x = a$ ympäristössä muodossa

$$g(x) = T_r(x) + R_{r+1}(x),$$

missä $R_{r+1}(x) = g(x) - T_r(x)$ on jäännöstermi, joka siis toteuttaa ehdon 6.10.2.

Oletetaan, että X on satunnaismuuttuja, jonka odotusarvo on $E(X) = \mu \neq 0$. Jos estimoidaan funktiota $g(\mu)$, niin sen Taylorin polynomiin perustuva 1. kertaluvun likiarvo pisteessä μ on

$$g(X) = g(\mu) + g'(\mu)(X - \mu). \tag{6.10.3}$$

Jos käytetään $g(\mu)$:n estimaattorina funktiota $g(X)$, niin

$$E[g(X)] \approx g(\mu)$$

ja

$$\operatorname{Var}[g(X)] = [g'(\mu)]^2 \operatorname{Var}(X).$$

Esimerkki 6.14 Tarkastellaan odotusarvon $E(X) = \mu \neq 0$ funktion $g(\mu) = 1/\mu$ estimointia. Olkoon estimaattorina $1/X$. Silloin edellisen mukaan

$$E\left(\frac{1}{X}\right) \approx \frac{1}{\mu}$$

ja

$$\operatorname{Var}\left(\frac{1}{X}\right) \approx \left(\frac{1}{\mu}\right)^4 \operatorname{Var}(X).$$

□

Lause 6.11 (Delta-menetelmä) *Olkoon $\{X_n\}$ sellainen satunnaismuuttujien jono, että $\sqrt{n}(X_n - \theta)$ lähenee jakaumamielessä normaalijakaumaa $\mathrm{N}(0, \sigma^2)$. Oletetaan, että annetulla funktiolla g on määrätyllä arvolla θ derivaatta $g'(\theta) \neq 0$. Silloin*

$$\sqrt{n}\,[g(X_n) - g(\theta)] \to \mathrm{N}\big(0, \sigma^2[g'(\theta)]^2\big)$$

jakaumamielessä.

Esimerkki 6.15 Olkoon $X_1, \ldots, X_n$ otos jakaumasta $\mathrm{Ber}(p)$. Onnistumisen todennäköisyyden p estimaattori on tavallisesti $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Onnistumisen mahdollisuus (odds) $p/(1-p)$ on vedonlyönnissä ja biostatiikassa tavanomainen parametri. Voimme käyttää $p/(1-p)$:n estimaattorina $\hat{p}$:n funktiota $\hat{p}/(1-\hat{p})$. Mitä voimme sanoa tämän estimaattorin ominaisuuksista? Nyt estimoidaan siis funktiota $g(p) = p/(1-p)$. Koska $g'(p) = 1/(1-p)^2$, niin lausekkeen 6.10.3 mukaan

$$\begin{aligned}\mathrm{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx [g'(p)]^2 \,\mathrm{Var}(\hat{p}) \\ &= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}.\end{aligned}$$

□

Luku 7

Uskottavuusfunktioon perustuva estimointi

7.1 Tilastolliset mallit

Tilastollinen malli on joukko havaintojen yhteisjakaumaa koskevia oletuksia. Havaintojen vaihtelu jaetaan systemaattiseen osaan ja satunnaisosaan:

$$\text{havainnot} = \text{systemaattinen osa} + \text{satunnaisosa}.$$

Havaintojen oletetaan olevan peräisin jostain jakaumaperheestä ja tavallisimmin ns. parametrisesta jakaumaperheestä. Systemaattinen osa on esimerkiksi havaintojen X_1, X_2, ..., X_n odotusarvoja $E(X_i)$, $1 \leq i \leq n$, koskeva oletus, joka lausutaan regressiofunktiona. Tavallisesti odotusarvo riippuu joistain selittävistä muuttujista (eli kovariaatista).

Esimerkki 7.1 Oletetaan, että $Y_1, Y_2, \ldots, Y_n$ on otos normaalijakaumasta $\mathrm{N}(\mu_i, \sigma^2)$, missä $E(Y_i) = \mu_i$. Oletetaan lisäksi, että $E(Y_i)$ riippuu lineaarisesti selittävästä muuttujasta x, joten

$$\mu_i = \alpha + \beta x_i, \qquad 1 \leq i \leq n.$$

Malli voidaan kirjoittaa myös muodossa

$$Y_i = \mu_i + \varepsilon_i,$$

missä $\varepsilon_i = Y_i - E(Y_i)$. Virhetermi ε_i noudattaa normaalijakaumaa $\mathrm{N}(0, \sigma^2)$. □

Esimerkki 7.2 Tarkastellaan auto-onnettomuuksien vakavuusastetta, kun selittäjänä on kuljettajan ikä. Usein väitetään, että nuoret kuljettajat aiheuttavat keskimääräistä enemmän vakavia onnettomuuksia.

Oletetaan, että onnettomuuksien lukumäärä kuukaudessa noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda)$. Tarkastellaan neljää onnettomuustyyppiä, jotka on

Taulukko 7.1. Vakavien onnettomuuksien lukumäärä alueella A tammikuussa vuonna 2000.

Yli 21-vuotiaat		Alle 21-vuotiaat	
Kuolemaan johtaneet	Muut	Kuolemaan johtaneet	Muut
Y_1	Y_2	Y_3	Y_4
11	62	4	7

määritelty kuljettajan iän ja onnettomuuden vakavuusasteen mukaan. Eri tyyppisten onnettomuuksien lukumäärien $Y_i, 1 \leq i \leq 4$, oletetaan noudattavan toisistaan riippumatta Poissonin jakaumaa $\text{Poi}(\lambda_i)$. Oheisessa taulukossa on annettu eräs aineisto. Silloin esimerkiksi alle 21-vuotiaiden kuolemaan johtaneiden onnettomuuksien lukumäärä $Y_3 \sim \text{Poi}(\lambda_3)$. Parametrit λ_1, λ_2, λ_3 ja λ_4 ovat satunnaismuuttujien Y_1, Y_2, Y_3 ja Y_4 odotusarvoja. Odotusarvo λ_i kertoo onnettomuusasteen i. kategoriassa. Vastaavasti esimerkiksi yli 21-vuotiaiden onnettomuusaste on $\lambda_1 + \lambda_2$ ja alle 21-vuotiaiden $\lambda_3 + \lambda_4$. Merkitään $\theta_1 = \lambda_1 + \lambda_2$ ja $\theta_2 = \lambda_3 + \lambda_4$. Näin todennäköisyys, että yli 21-vuotias aiheuttaa kohtalokkaan onnettomuuden, on

$$\pi_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

ja alle 21-vuotiaan todennäköisyys aiheuttaa kohtalokas onnettomuus on

$$\pi_2 = \frac{\lambda_3}{\lambda_3 + \lambda_4}.$$

Nelikko $(\theta_1, \pi_1, \theta_2, \pi_2)$ muodostaa uuden parametrisoinnin, joka saattaa olla tulkinnallisesti selkeämpi ja mielenkiintoisempi kuin alkuperäinen. □

7.2 Estimoinnista

Tarkastelemme tässä luvussa satunnaismuuttujia, joiden todennäköisyysfunktion (tai tiheysfunktion) funktionaalinen muoto tunnetaan, mutta jakauma riippuu jostain tuntemattomasta parametrista θ. Oletetaan, että parametrin θ mahdolliset arvot kuuluvat johonkin annettuun joukkoon Θ, jota kutsutaan *parametriavaruudeksi*. Tiedetään esimerkiksi, että jonkin tuotteen elinaika X noudattaa eksponenttijakaumaa

$$f(x;\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}, \qquad 0 < x < \infty,$$

missä $\theta \in \Theta = \{\, \theta \mid 0 < \theta < \infty \,\}$. Parametriavaruus Θ on siis positiivisten reaalilukujen joukko. Haluamme valita funktioperheestä

$$\mathcal{F} = \{\, f(x;\theta),\ \theta \in \Theta \,\}$$

yhden tiheysfunktion, joka esittää parhaiten tuotteen elinaikaa. Valitaan siis yksi parametrin θ arvo eli parametrin θ *piste-estimaatti*, joka määrittää jakauman.

Parametrin arvo arvioidaan eli estimoidaan havaintojen perusteella. Teemme jakaumasta havainnon $X = x$ ja estimoimme parametrin θ arvon havainnon x perusteella. Parametrin θ estimoimiseen käytettävää otosfunktiota $T(X)$ kutsutaan parametrin θ *estimaattoriksi* ja estimaattorin $T(X)$ arvoa $t = T(x)$ kutsutaan parametrin θ *estimaatiksi*. Estimaattori pyritään valitsemaan siten, että se antaa hyviä arvioita parametrista θ.

Esimerkki 7.3 Estimoidaan ehdokkaan A kannattajien suhteellinen osuus π eräässä suuressa kaupungissa. Valitaan kaupungin äänioikeutetuista satunnaisesti n henkilöä, joilta tiedustellaan, kannattavatko he ehdokasta A. Olkoon X ehdokkaan A kannattajien lukumäärä otoksessa. Koska populaation koko on suuri verrattuna otoskokoon n, voidaan olettaa, että $X \sim \text{Bin}(n, \pi)$, missä π on todennäköisyys, että satunnaisesti valittu henkilö kannattaa A:ta. Binomijakaumaa noudattavan satunnaismuuttujan X todennäköisyysfunktio on muotoa

$$f(x; n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \qquad x = 0, 1, \ldots, n; \quad 0 \leq \pi \leq 1.$$

Binomijakauman parametriavaruus on $\Theta = \{\, \pi \mid 0 \leq \pi \leq 1 \,\}$. Tehtävänämme on määrittää π:n estimaattori $T(X)$ siten, että havaitun arvon $X = x$ perusteella saadaan hyvä π:n piste-estimaatti $T(x)$. Havainnon $X = x$ todennäköisyys on

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}. \tag{7.2.1}$$

Eräs tapa määrittää π:n estimaatti on tarkastella todennäköisyyttä $P(X = x)$ parametrin π funktiona ja etsiä sellainen π:n arvo, että havainnon x todennäköisyys saavuttaa maksiminsa. Voidaan osoittaa, että havainnon $X = x$ todennäköisyys maksimoituu, kun $\pi = x/n$. Tätä estimaattia kutsutaan π:n *suurimman uskottavuuden estimaatiksi* ja sitä merkitään

$$\hat{\pi} = \frac{x}{n}.$$

□

7.3 Suurimman uskottavuuden menetelmä

Kun todennäköisyyttä (7.2.1) tarkastellaan parametrin π funktiona, huomaamme, että tekijä $\binom{n}{x}$ ei riipu parametrista π. Siksi määrittelemmekin *uskottavuusfunktion* seuraavasti:

$$L(\theta) = c \cdot f(x; \theta), \tag{7.3.1}$$

missä $f(x;\theta)$ on todennäköisyysfunktio ja $f(x;\theta)$ on siis havainnon $X = x$ todennäköisyys. Vakio c ei riipu parametrista θ, vaikkakin se voi riippua havainnosta x. Vakio c pyritään valitsemaan siten, että $L(\theta)$:lle saadaan yksinkertainen lauseke. Uskottavuusfunktioon perustuvat päätelmät eivät riipu vakion c valinnasta.

Tavallisesti uskottavuusfunktio tulee olemaan useiden tekijöiden tulo ja mm. siitä syystä on osoittautunut käteväksi työskennellä uskottavuusfunktion logaritmin avulla. *Logaritmoitu uskottavuusfunktio* $l(\theta)$ on uskottavuusfunktion luonnollinen logaritmi eli

$$l(\theta) = \log L(\theta). \tag{7.3.2}$$

Esityksestä (7.3.1) seuraa, että

$$l(\theta) = \log c + \log f(x;\theta),$$

missä vakio c ei riipu θ:sta. Jatkossakin kaikki logaritmit ovat luonnollisia logaritmeja, ellei toisin mainita.

Suurimman uskottavuuden estimaatti (SUE) $\hat{\theta}$ on se parametrin θ arvo, joka maksimoi havainnon x todennäköisyyden $f(x;\theta)$. Sama arvo $\hat{\theta}$ maksimoi myös funktiot $L(\theta)$ ja $l(\theta)$. Suurimman uskottavuuden estimaatti $\hat{\theta}$ on siis uskottavuusfunktion ja logaritmoidun uskottavuusfunktion maksimikohta. Tavallisesti tarkastellaan logaritmoitua uskottavuusfunktiota, koska se on usein matemaattisesti yksinkertaisempi kuin uskottavuusfunktio. Logaritmoidulla uskottavuusfunktiolla on myös teoreettisesti merkittävä tilastollinen tulkinta.

Esimerkki 7.4 Tarkastellaan edelleen Esimerkkiä 7.3, jossa havaintojen todennäköisyysfunktio on $f(x;\theta) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$. Kun uskottavuusfunktiossa (7.3.1) valitaan vakion arvoksi $c = 1\big/\binom{n}{x}$, saadaan esitysmuoto

$$L(\pi) = \pi^x(1-\pi)^{n-x}, \qquad 0 \le \pi \le 1.$$

Tässä uskottavuusfunktion esityksessä ei ole ”turhia” vakiotekijöitä. Tätä uskottavuusfunktion esitysmuotoa kutsutaan myös *uskottavuusfunktion ytimeksi*. Yleensä esitämme uskottavuusfunktion tässä *ydinmuodossa*. Logaritmoitu uskottavuusfunktio on

$$l(\pi) = x\log\pi + (n-x)\log(1-\pi), \qquad 0 < \pi < 1.$$

Parametrin π suurimman uskottavuuden estimaatti on siis se π:n arvo, joka maksimoi funktion $l(\pi)$. Huomattakoon, että $l(\pi)$ ei ole määritelty välin $[0,1]$ päätepisteissä, mutta $L(\pi)$ on. □

7.4 Pistefunktio ja informaatiofunktio

Estimaatin $\hat{\theta}$ laskemiseksi on siis määritettävä funktion $l(\theta)$ maksimikohta parametriavaruudessa Θ. Jos funktio on riittävän säännöllinen, voidaan $\hat{\theta}$ määrittää derivointikeinolla. Funktion $l(\theta)$ 1. derivaattaa kutsutaan *Fisherin pistefunktioksi* tai lyhyesti vain pistefunktioksi ja se määritellään seuraavasti:

$$u(\theta; x) = \frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta; x). \tag{7.4.1}$$

Merkitsemme myös lyhyesti $u(\theta) = l'(\theta)$. Informaatiofunktio $\mathcal{I}(\theta)$ on logaritmoidun uskottavuusfunktion $l(\theta)$ 2. derivaatta miinus-merkkisenä:

$$\mathcal{I}(\theta) = -l''(\theta) = -u'(\theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} l(\theta). \tag{7.4.2}$$

Huomattakoon, että funktiot $u(\theta)$ ja $\mathcal{I}(\theta)$ eivät riipu uskottavuusfunktioon (7.3.1) liittyvän vakion c valinnasta.

Tavallisesti parametriavaruus Θ on jokin reaalilukuväli ja $l(\theta)$:n 1. ja 2. derivaatta ovat olemassa kaikissa Θ:n sisäpisteissä. Jos nyt $\hat{\theta}$ on Θ:n sisäpiste, niin $l'(\hat{\theta}) = 0$ ja $l''(\hat{\theta}) < 0$. Näiden ehtojen vallitessa on siis

$$u(\hat{\theta}) = 0 \qquad \text{ja} \qquad \mathcal{I}(\hat{\theta}) > 0. \tag{7.4.3}$$

Suurimman uskottavuuden estimaatti $\hat{\theta}$ voidaan siis määrittää ratkaisemalla *uskottavuusyhtälö*

$$u(\theta) = 0.$$

Esimerkiksi $\mathcal{I}(\theta)$:n merkkiä tarkastelemalla on vielä tutkittava, onko kyseessä maksimi. Jos edellä mainitut säännöllisyysoletukset eivät pidä paikkaansa, ei $\hat{\theta}$:a välttämättä saada uskottavuusyhtälön ratkaisuna.

Esimerkki 7.5 Binomijakauman tapauksessa (Esimerkki 7.4)

$$u(\pi) = \frac{x}{\pi} - \frac{n-x}{1-\pi}, \qquad \mathcal{I}(\pi) = \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2}, \qquad 0 < \pi < 1.$$

Kun $1 \leq x \leq n-1$, niin uskottavuusyhtälöllä $u(\pi) = 0$ on yksikäsitteinen ratkaisu $\pi = x/n$. Koska $\mathcal{I}(\pi) > 0$ pisteessä $\pi = x/n$, niin funktiolla $l(\pi)$ [ja funktiolla $L(\pi)$] on maksimi pisteessä $\pi = x/n$. Koska $L(0) = L(1) = 0$, niin $\hat{\pi} = x/n$ on globaali maksimi ja siksi π:n suurimman uskottavuuden estimaatti. Kun $x = 0$, niin uskottavuusyhtälöllä ei ole ratkaisua, mutta uskottavuusfunktio on silloin

$$L(\pi) = (1-\pi)^n, \qquad 0 \leq \pi \leq 1.$$

Nähdään helposti, että $L(0) = \max_{\pi} L(\pi)$, joten $\hat{\pi} = 0$. Vastaavasti kun $x = n$, niin $\hat{\pi} = 1$. Näin siis kaava $\hat{\pi} = x/n$ pätee kaikilla havaintoarvoilla, vaikka kaikkia estimaatteja ei saada uskottavuusyhtälön ratkaisuna. □

Esimerkki 7.6 Oletetaan, että puhelinvaihteeseen päivän aikana tulevien "väärien" puheluiden lukumäärä noudattaa Poissonin jakaumaa, jonka odotusarvo on μ. Oletetaan, että jakauma on kaikkina päivinä sama ja eri päivien havainnot ovat toisistaan riippumattomat. Olkoot x_1, x_2, ..., x_n eri päivinä havaitut virhepuheluiden lukumäärät. Havainnon x_i todennäköisyys on

$$f(x_i;\mu) = \frac{1}{x_i!}\mu^{x_i}\mathrm{e}^{-\mu}, \qquad x_i = 0, 1, 2, \ldots.$$

Koska eri päivinä havaittujen virhepuheluiden lukumäärät ovat toisistaan riippumattomat, niin otoksen $x_1, x_2, \ldots, x_n$ todennäköisyys on

$$\begin{aligned} f(x_1, x_2, \ldots, x_n;\mu) &= f(x_1;\mu)f(x_2;\mu)\cdots f(x_n;\mu) \\ &= \prod_{i=1}^{n} \frac{1}{x_i!}\mu^{x_i}\mathrm{e}^{-\mu} = \frac{1}{x_1!x_2!\ldots x_n!}\mu^{\sum x_i}\mathrm{e}^{-n\mu}. \end{aligned}$$

Valitaan uskottavuusfunktiossa suhteellisuusvakioksi $c = x_1!x_2!\cdots x_n!$, jolloin uskottavuusfunktion ydin on

$$L(\mu) = \mu^{\sum x_i}\mathrm{e}^{-n\mu}, \qquad 0 \le \mu < \infty$$

ja vastaava logaritmoitu uskottavuusfunktio on

$$l(\mu) = \sum x_i \log(\mu) - n\mu.$$

Pistefunktio ja informaatiofunktio ovat vastaavasti

$$u(\mu) = \frac{1}{\mu}\sum x_i - n \qquad \text{ja} \qquad \mathcal{I}(\mu) = \frac{\sum x_i}{\mu^2}.$$

Funktiot $u(\mu)$ ja $\mathcal{I}(\mu)$ eivät riipu vakion c valinnasta.

Jos $\sum x_i > 0$, niin uskottavuusyhtälöllä $u(\mu) = 0$ on yksikäsitteinen ratkaisu $\mu = \sum x_i/n = \bar{x}$. Koska $\mathcal{I}(\mu) > 0$ pisteessä $\mu = \bar{x}$, niin $\bar{x}$ on maksimikohta. Se on myös globaali maksimi, koska $L(0) = 0$ ja $L(\mu) \to 0$, kun $\mu \to \infty$. Jos $\sum x_i = 0$, niin uskottavuusyhtälöllä $u(\mu) = 0$ ei ole ratkaisua, mutta uskottavuusfunktio saavuttaa maksiminsa pisteessä $\mu = 0$. Näin siis $\hat{\mu} = \bar{x}$ kaikilla havaintoarvoilla. Havaintojen x_1, x_2, ..., x_n todennäköisyys maksimoituu, kun jakauman tuntematon parametri μ estimoidaan otoskeskiarvolla $\bar{x}$. □

7.4.1 Frekvenssitaulukkoon liittyvä uskottavuusfunktio

Usein tarkastellaan koetta, jossa on useita toisensa poissulkevia tulosvaihtoehtoja. Toistetaan n kertaa koe, jolla on k toisensa poissulkevaa tulosvaihtoehtoa. Oletetaan lisäksi, että koetoistot ovat toisistaan riippumattomat. Silloin koetulokset voidaan esittää seuraavana frekvenssitaulukkona:

Tulos	T_1	T_2	...	T_k	Yhteensä
Havaittu frekvenssi	n_1	n_2	...	n_k	n
Teoreettinen frekvenssi	$n_1\pi_1$	$n_2\pi_2$	...	$n_k\pi_k$	n

Esimerkki 7.7 Tuotantoprosessissa syntyvä tuote luokitellaan yhteen ja vain yhteen seuraavista kategorioista: ensiluokkainen (T_1), sekunda (T_2) tai viallinen (T_3). Olkoot ensiluokkaisen, sekundan ja viallisen tuotteen todennäköisyydet vastaavasti π_1, π_2 ja $\pi_3 = 1 - \pi_1 - \pi_2$. Testataan n tuotetta. Olkoon X_1 ensiluokkaisten lukumäärä, X_2 sekundan lukumäärä ja X_3 viallisten lukumäärä tuote-erässä. Koska $n_1 + n_2 + n_3 = n$ ja $\pi_1 + \pi_2 + \pi_3 = 1$, niin todennäköisyys saada n_1 ensiluokkaista, n_2 sekundatuotetta ja n_3 viallista on

$$\begin{aligned} f(n_1, n_2, n_3; \pi_1, \pi_2) &= \frac{n!}{n_1!n_2!n_3!}\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3} \\ &= \frac{n!}{n_1!n_2!(n-n_1-n_2)!}\pi_1^{n_1}\pi_2^{n_2}(1-\pi_1-\pi_2)^{n-n_1-n_2}. \end{aligned}$$

Parametrien π_1 ja π_2 uskottavuusfunktioksi saadaan

$$L(\pi_1, \pi_2) = \pi_1^{n_1}\pi_2^{n_2}(1-\pi_1-\pi_2)^{n-n_1-n_2},$$

kun suhteellisuusvakioksi valitaan $c = \frac{n_1!n_2!(n-n_1-n_2)!}{n!}$. Logaritmoitu uskottavuusfunktio on vastaavasti

$$l(\pi_1, \pi_2) = n_1\log(\pi_1) + n_2\log(\pi_2) + (n-n_1-n_2)\log(1-\pi_1-\pi_2).$$

Tässä tapauksessa saadaan siis kahden tuntemattoman parametrin uskottavuusfunktio. Mallissa on alunperin kolme tuntematonta parametria π_1, π_2 ja π_3, mutta niitä sitoo reunahto $\pi_1 + \pi_2 + \pi_3 = 1$, joten mallissa on vain 2 *vapaata parametria*. Uskottavuusfunktio on lausuttavissa minkä tahansa parameteista π_1, π_2 ja π_3 valitun kahden parametrin avulla. Maksimoimalla logaritmoitu uskottavuusfunktio saadaan parametrien π_1, π_2 ja π_3 suurimman uskottavuuden estimaateiksi

$$\hat{\pi}_i = \frac{n_i}{n}, \qquad i = 1, 2, 3.$$

□

Todennäköisyys, että kokeessa saadaan edellä esitetyn frekvenssitaulukon mukaiset havainnot, on

$$f(n_1, n_2, \ldots, n_k; \pi_1, \pi_2, \ldots, \pi_k) = \frac{n!}{n_1!n_2!\cdots n_k!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k},$$

joten uskottavuusfunktion ydin on

$$L(\pi_1, \pi_2, \ldots, \pi_k) = \pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k} \tag{7.4.4}$$

ja logaritmoitu uskottavuusfunktio

$$l(\pi_1, \pi_2, \ldots, \pi_k) = n_1 \log(\pi_1) + n_2 \log(\pi_2) + \cdots + n_k \log(\pi_k), \tag{7.4.5}$$

kun suhteellisuusvakioksi valitaan $c = \frac{n_1! n_2! \cdots n_k!}{n!}$. Koska parametreja π_1, π_2, $\ldots$, π_k sitoo ehto $\pi_1 + \pi_2 + \cdots + \pi_k = 1$, niin uskottavuusfunktiossa (7.4.4) on vain $k - 1$ vapaata parametria. Parametrien π_1, π_2, $\ldots$, π_k arvot $\hat{\pi}_1$, $\hat{\pi}_2$, $\ldots$, $\hat{\pi}_k$, jotka maksimoivat uskottavuusfunktion (7.4.4), ovat parametrien suurimman uskottavuuden estimaatit. Niille pätee siis epäyhtälö

$$L(\pi_1, \pi_2, \ldots, \pi_k) \leq L(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k), \qquad \text{kaikilla } (\pi_1, \pi_2, \ldots, \pi_k) \in \Theta,$$

missä

$$\Theta = \left\{ (\pi_1, \pi_2, \ldots, \pi_k) \;\middle|\; 0 \leq \pi_i,\ i = 1, 2, \ldots, k \ \text{ ja } \sum_{i=1}^{k} \pi_i = 1 \right\}.$$

Näiden estimaattien avulla saadaan sitten odotettujen frekvenssien suurimman uskottavuuden estimaatit $n\hat{\pi}_1$, $n\hat{\pi}_2$, $\ldots$, $n\hat{\pi}_k$.

Seuraavassa apulauseessa esitetään multinomijakauman parametrien suurimman uskottavuuden estimaatit, jotka ovat erittäin käyttökelpoisia analysoitaessa frekvenssitaulukoita.

Apulause 7.1 *Olkoon* $l(\pi_1, \pi_2, \ldots, \pi_k) = \sum_{i=1}^{k} n_i \log(\pi_i)$. *Jos* $n_i > 0$ *ja* $\pi_i > 0$, $i = 1, 2, \ldots, k$ *sekä* $\pi_1 + \pi_2 + \cdots \pi_k = 1$, *niin funktio* $l(\pi_1, \pi_2, \ldots, \pi_k)$ *saavuttaa maksiminsa pisteessä* $(\pi_1, \pi_2, \ldots, \pi_k) = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k)$, *missä* $\hat{\pi}_i = n_i/n$.

7.5 Riippumattomien otosten yhdistäminen

Tehdään kaksi toisistaan riippumatonta koetta (tai otosta), jotka antavat informaatiota samasta parametrista θ. Havainnon $X = x$ todennäköisyys 1. kokeessa on $f_1(x; \theta)$ ja havainnon $Y = y$ todennäköisyys 2. kokeessa on $f_2(y; \theta)$. Vastaavasti 1. kokeeseen perustuva uskottavuusfunktio on

$$L_1(\theta) = c_1 f_1(x; \theta)$$

ja 2. kokeeseen perustuva uskottavuusfunktio on

$$L_2(\theta) = c_2 f_2(y; \theta),$$

missä c_1 ja c_2 ovat joitain positiivisia vakioita. Olkoon satunnaismuuttujien X ja Y yhteisjakauman todennäköisyysfunktio $f(x, y)$. Yhdistettyyn kokeeseen perustuva uskottavuusfunktio on

$$L(\theta) = cf(x, y; \theta),$$

missä c on jokin positiivinen vakio. Koska X ja Y ovat riippumattomat (kokeet ovat riippumattomat), niin $f(x,y) = f_1(x)f_2(y)$. Yhdistettyyn kokeeseen perustuva uskottavuusfunktio voidaan siis kirjoittaa muodossa

$$L(\theta) = c^* L_1(\theta)\, L_2(\theta),$$

missä $c^* = \frac{c}{c_1 c_2}$.

Jos nyt valitaan $c^* = 1$, niin yhdistettyyn kokeeseen perustuva uskottavuusfunktio on

$$L(\theta) = L_1(\theta)\, L_2(\theta). \tag{7.5.1}$$

Ottamalla uskottavuusfunktiosta (7.5.1) puolittain logaritmit saadaan vastaava logaritmoitu uskottavuusfunktio

$$l(\theta) = l_1(\theta) + l_2(\theta). \tag{7.5.2}$$

Kahdesta riippumattomasta kokeesta saatava parametria θ koskeva informaatio voidaan siis yhdistää siten, että kerrotaan yksittäisiin kokeisiin liittyvät uskottavuusfunktiot keskenään ja vastaavasti logaritmoidut uskottavuusfunktiot lasketaan yhteen. On helppo nähdä, että useamman kuin kahden riippumattoman kokeen antamat tulokset voidaan yhdistää vastaavasti kertomalla kokeisiin liittyvät uskottavuusfunktiot ja laskemalla yhteen logaritmoidut uskottavuusfunktiot. Suoraan pistefunktion määritelmästä seuraa, että yhdistetyn kokeen perusteella saatu θ:n pistefunktio on

$$u(\theta) = u_1(\theta) + u_2(\theta), \tag{7.5.3}$$

missä $u_1(\theta)$ ja $u_2(\theta)$ ovat 1. ja 2. kokeen perusteella määritetyt pistefunktiot. Yhtälöstä (7.5.3) ja identiteetistä (7.4.2) seuraa, että yhdistettyyn kokeeseen perustuva informaatiofunktio on

$$\mathcal{I}(\theta) = \mathcal{I}_1(\theta) + \mathcal{I}_2(\theta), \tag{7.5.4}$$

missä $\mathcal{I}_1(\theta)$ ja $\mathcal{I}_2(\theta)$ ovat yksittäisistä kokeista saadut informaatiofunktiot.

Esimerkki 7.8 Jatketaan esimerkin 7.3 käsittelyä. Oletetaan, että kaksi eri tutkijaa tekevät samaan aikaan toisistaan riippumatta ehdokkaan A kannattusta mittaavan haastattelututkimuksen siten, että 1. tutkija haastatteli n_1 ja 2. tutkija n_2 äänioikeutettua. Merkitään $n = n_1 + n_2$. Havaittiin, että 1. tutkijan otoksessa oli x_1 ja 2. tutkijan otoksessa x_2 ehdokkaan A kannattajaa.

Nyt 1. otokseen perustuva logaritmoitu uskottavuusfunktio on

$$l_1(\pi) = x_1 \log(\pi) + (n_1 - x_1)\log(1-\pi)$$

ja 2. otokseen perustuva logaritmoitu uskottavuusfunktio on vastaavasti

$$l_2(\pi) = x_2 \log(\pi) + (n_2 - x_2)\log(1-\pi).$$

Otoksista lasketut suurimman uskottavuuden estimaatit ovat vastaavasti

$$\hat{\pi}_1 = \frac{x_1}{n_1} \qquad \text{ja} \qquad \hat{\pi}_2 = \frac{x_2}{n_2}.$$

Koska populaatio on suuri, otokset voidaan olettaa riippumattomiksi. Silloin tuloksen (7.5.2) mukaan yhdistettyyn otokseen perustuva logaritmoitu uskottavuusfunktio on

$$\begin{aligned} l(\pi) &= l_1(\pi) + l_2(\pi) \\ &= (x_1 + x_2)\log(\pi) + (n_1 + n_2 - x_1 - x_2)\log(1-\pi) \\ &= (x_1 + x_2)\log(\pi) + (n - x_1 - x_2)\log(1-\pi). \end{aligned} \tag{7.5.5}$$

Logaritmoidusta uskottavuusfunktiosta (7.5.5) laskettu π:n suurimman uskottavuuden estimaatti on

$$\hat{\pi} = \frac{x_1 + x_2}{n} = \frac{x_1}{n} + \frac{x_2}{n} = \frac{n_1}{n}\hat{\pi}_1 + \frac{n_2}{n}\hat{\pi}_2.$$

□

7.6 Normitettu uskottavuusfunktio

Parametrin θ *normitettu uskottavuusfunktio* $R(\theta)$ saadaan jakamalla uskottavuusfunktio $L(\theta)$ uskottavuusfunktion maksimiarvolla:

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}. \tag{7.6.1}$$

Koska c ei riipu θ:sta lausekkeessa $L(\theta) = cf(x;\theta)$, niin

$$R(\theta) = \frac{cf(x;\theta)}{cf(x;\hat{\theta})} = \frac{f(x;\theta)}{f(x;\hat{\theta})}.$$

Huomaa, että $0 \leq R(\theta) \leq 1$, sillä $L(\theta) \leq L(\hat{\theta})$ kaikilla θ:n arvoilla.

Logaritminen normitettu uskottavuusfunktio $\log R(\theta)$ on normitetun uskottavuusfunktion logaritmi, joten se voidaan lausua logaritmoidun uskottavuusfunktion $l(\theta)$ avulla seuraavasti:

$$r(\theta) = l(\theta) - l(\hat{\theta}). \tag{7.6.2}$$

Olkoon θ_1 jokin parametrin θ arvo. Silloin

$$R(\theta_1) = \frac{L(\theta_1)}{L(\hat{\theta})} = \frac{f(x;\theta_1)}{f(x;\hat{\theta})},$$

missä $f(x;\theta_1)$ on havainnon x todennäköisyys, kun $\theta = \theta_1$. Havainnon x suurin mahdollinen todennäköisyys on $f(x;\hat{\theta})$, jos oletetaan todennäköisyysfunktio $f(x;\theta)$. Jos esimerkiksi $R(\theta_1) = 0.05$, niin parametrin arvo $\theta = \theta_1$

on hyvin epäuskottava, sillä parametrin arvolla $\theta = \hat{\theta}$ havainto on 20 kertaa todennäköisempi kuin parametrin arvolla $\theta = \theta_1$. Normitetun uskottavuusfunktion avulla voidaan asettaa parametrin arvot uskottavuutensa mukaan järjestykseen.

Yleisemmin normitettu uskottavuusfunktio voidaan määritellä suhteena

$$R(\theta) = \frac{L(\theta)}{\sup\limits_{\theta} L(\theta)},$$

koska $\hat{\theta}$ ei ole aina olemassa. Koska $L(\theta) = cf(x, \theta)$ ja $f(x, \theta) \leq 1$, niin supremum on äärellinen. Normitettu uskottavuusfunktio on siis aina olemassa tämän yleisemmän määritelmän mielessä. Useimmiten tarkastelemme sellaisia tilanteita, joissa $\hat{\theta}$ on olemassa ja yksikäsitteinen.

7.6.1 Uskottavuusvälit ja uskottavuusalueet

Parametrien arvojen uskottavuutta voidaan nyt verrata normitetun uskottavuusfunktion avulla suurimman uskottavuuden estimaatin uskottavuuteen. Parametrin θ $100p$ %:n uskottavuusalue $\mathrm{ua}(\theta, 100p)$ muodostuu niistä θ:n arvoista, jotka toteuttavat epäyhtälön $R(\theta) \geq p$:

$$\mathrm{ua}(\theta, 100p) = \{\, \theta \mid R(\theta) \geq p \,\}.$$

Tavallisesti uskottavuusalue $\mathrm{ua}(\theta, 100p)$ on reaalilukuväli, jolloin uskottavuusaluetta kutsutaan uskottavuusväliksi (uv). Tavallisesti tarkastellaan 50 %:n, 10 %:n ja 1 %:n uskottavuusvälejä.

Uskottavuusalue on usein kätevämpää määrittää normitetun uskottavuusfunktion logaritmin $r(\theta)$ avulla. Esimerkiksi θ:n 50 %:n uskottavuusväli saadaan määrittämällä epäyhtälön

$$R(\theta) \geq 0.5$$

tai vastaavasti epäyhtälön

$$r(\theta) \geq \log(0.5) = -0.69$$

toteuttavien θ:n arvojen joukko. Parametria θ koskeva informaatio voidaan kuvata havainnollisesti $R(\theta)$:n tai $r(\theta)$:n kuvaajan avulla. Kun funktio $r(\theta)$ on matemaattisesti riittävän yksinkertainen, kuten useimmissa esimerkeissämme, saadaan parametrista riittävän tarkka yleiskuva esittämällä suurimman uskottavuuden estimaatti $\hat{\theta}$ ja pari kolme uskottavuusväliä. Niiden avulla voidaan jo hahmotella $r(\theta)$:n kuvaaja.

Esimerkki 7.9 Tuotantolijalla testattiin 100:n tuotteen satunnaisotos. Virheellisten tuotteiden lukumäärä X otoksessa noudattaa binomijakaumaa

$\mathrm{Bin}(100, \pi)$. Otoksessa on kaksi viallista tuotetta ($X = 2$). Logaritmoitu uskottavuusfunktio on silloin

$$l(\theta) = 2\log(\theta) + 98\log(1-\theta)$$

ja $\hat{\theta} = 0.02$. Logaritmoidun uskottavuusfunktion maksimi

$$l(\hat{\theta}) = \log(0.02) + 98\log(0.98) = -9.80.$$

Logaritmoitu suhteellinen uskottavuusfunktio on siis

$$r(\theta) = l(\theta) - l(\hat{\theta}) = 2\log(\theta) + 98\log(1-\theta) + 9.80.$$

Esimerkiksi 50 % uskottavuusalue on

$$\{\, \theta \mid r(\theta) - \log(0.5) \geq 0 \,\} = \{\, \theta \mid 2\log(\theta) + 98\log(1-\theta) + 9.80 + 0.69 \geq 0 \,\}.$$

Uskottavuusalue on väli. □

7.7 Uskottavuus jatkuvissa malleissa

Kun X on jatkuva satunnaismuuttuja, niin tiheysfunktion arvo $f(x;\theta)$ ei ole todennäköisyys $P(X = x)$. Itse asiassa $P(X = x) = 0$ kaikilla x, kun X on jatkuva satunnaismuuttuja. Koska mittaustarkkuus on aina äärellinen, sovelluksissa tapahtuma $\{X = x\}$ tarkoittaa, että x kuuluu johonkin mittaustarkkuuden määrittämään väliin. Tapahtumien $\{X = x\}$ sijasta tarkastellaan siis tapahtumia $\{a < X \leq b\}$, missä $a < b$. Silloin tapahtuman $\{X = x\}$ todennäköisyys on siis muotoa

$$P(a < X \leq b) = \int_a^b f(x;\theta)\,\mathrm{d}x = F(b) - F(a).$$

Olkoon $X_1, X_2, \ldots, X_n$ otos jakaumasta $F(x)$. Havaitaan arvot $X_1 = x_1$, $X_2 = x_2$, …, $X_n = x_n$, joihin liittyvät mittaustarkkuudet ovat (a_i, b_i), $1 \leq i \leq n$. Otoksen todennäköisyys on

$$\begin{aligned} P(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n) &= \prod_{i=1}^{n} P(a_i < X_i \leq b_i) \\ &= \prod_{i=1}^{n} [F(b_i) - F(a_i)]. \end{aligned} \tag{7.7.1}$$

Jos väli $\Delta_i = b_i - a_i$ on lyhyt ja F on kohtuullisen sileä, niin $F(b_i) \approx F(a_i)$. Silloin voidaan käyttää likiarvoa

$$P(a_i < X_i \leq b_i) = F(b_i) - F(a_i) \approx f(x_i;\theta)\Delta_i. \tag{7.7.2}$$

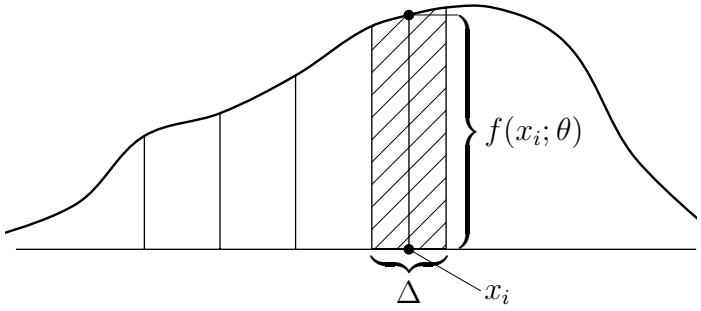
Kun likiarvoa (7.7.2) sovelletaan todennäköisyyden (7.7.1) laskemiseen, saadaan

$$P(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n) \approx \prod_{i=1}^{n} f(x_i;\theta)\Delta_i = \Big(\prod_{i=1}^{n} \Delta_i\Big) \prod_{i=1}^{n} f(x_i;\theta).$$

Koska pituudet Δ_i eivät riipu parametrista θ, niin uskottavuusfunktio on vakio kertaa tiheysfunktion arvojen tulo, eli

$$L(\theta) = c \prod_{i=1}^{n} f(x_i;\theta), \tag{7.7.3}$$

missä c on mikä tahansa sopivasti valittu positiivinen vakio. Usein $\Delta_i = \Delta$, $1 \le i \le n$, jolloin $c = \Delta^n$.



Esimerkki 7.10 Olkoon $x_1, x_2, \ldots, x_n$ havaittu otos eksponenttijakaumasta $\mathrm{Exp}(\theta)$. Valitaan $c = 1$ identiteetissä (7.7.3). Silloin

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left(-\frac{x_i}{\theta}\right) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum x_i\right).$$

Logaritmoitu uskottavuusfunktio on

$$l(\theta) = -n \log \theta - \frac{1}{\theta} \sum x_i,$$

pistefunktio

$$u(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum x_i$$

ja informaatiofunktio

$$\mathcal{I}(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum x_i.$$

Kun ratkaistaan yhtälö $u(\theta) = 0$, saadaan ratkaisu $\hat{\theta} = \frac{1}{n} \sum x_i = \overline{x}$. Koska

$$\mathcal{I}(\hat{\theta}) = -\frac{n}{\hat{\theta}^2} + \frac{2n\hat{\theta}}{\hat{\theta}^3} = \frac{n}{\hat{\theta}^2} > 0,$$

niin ratkaisu $\hat{\theta}$ on uskottavuusfunktion maksimikohta. Logaritmoidun uskottavuusfunktion maksimi on $l(\hat{\theta}) = -n \log \hat{\theta} - n$. Logaritminen normitettu uskottavuusfunktio on

$$r(\theta) = l(\theta) - l(\hat{\theta}) = -n\left(\frac{\hat{\theta}}{\theta} - 1 - \log \frac{\hat{\theta}}{\theta}\right).$$

□

7.8 Invarianssi

Oletetaan esimerkiksi, että erään elektronisen komponentin elinikä noudattaa eksponenttijakaumaa $\mathrm{Exp}(\theta)$, jolloin sen tiheysfunktio on

$$f(x) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}, \qquad 0 < x < \infty. \tag{7.8.1}$$

Jokaista parametrin θ arvoa vastaa yksi jakauma. Olemme esittäneet eksponenttijakauman tiheysfunktion myös muodossa

$$f(x) = \lambda \mathrm{e}^{-\lambda x}, \qquad 0 < x < \infty, \tag{7.8.2}$$

missä $\lambda = 1/\theta$. Tavallisesti parametrisointi valitaan siten, että parametri esittää jotain tärkeää jakauman ominaisuutta tai siten, että jakauman matemaattinen esitystapa saadaan yksinkertaiseksi. Parametrisoinnissa (7.8.1) θ on jakauman keskiarvo.

Jos esimerkiksi $\theta = 2$, niin $\lambda = \frac{1}{2}$. Jokaista θ:n valintaa vastaa yksikäsitteinen jakauma parametrisoinnissa (7.8.2) ja kääntäen. Uskottavuusmenetelmällä on se miellyttävä piirre, että menetelmä on invariantti bijektiivisten (yksi-yksisten) parametrimuunnosten suhteen. Olkoon $\theta = g(\lambda)$ bijektiivinen parametrimuunnos ja $L(\theta)$ parametrin θ uskottavuusfunktio. Kun θ:n paikalle sijoitetaan $\theta = g(\lambda)$, saadaan λ:n uskottavuusfunktio, sanokaamme $\tilde{L}(\lambda)$. Silloin nämä uskottavuusfunktiot riippuvat toisistaan siten, että

$$\tilde{L}(\lambda) = L(\theta),$$

missä $\theta = g(\lambda)$. Siksi molemmilla funktioilla on sama maksimiarvo ja

$$\tilde{R}(\lambda) = R(\theta),$$

missä $\theta = g(\lambda)$. Jos $\lambda_1 = \lambda$ ja $\theta_1 = g(\lambda_1)$, niin silloin arvoilla λ_1 ja θ_1 on sama suhteellinen uskottavuus. Suhteelliset uskottavuudet eivät riipu valitusta parametrisoinnista. Jos $\hat{\lambda}$ on λ:n suurimman uskottavuuden estimaatti, niin $\hat{\theta} = g(\hat{\lambda})$ on θ:n suurimman uskottavuuden estimaatti.

7.9 Normaalijakaumaan perustuva likiarvo

Olkoon $l(\theta)$ parametriavaruudessa Θ määritelty jatkuvan parametrin θ logaritmoitu uskottavuusfunktio. Vastaavasti θ:n pistefunktio on $u(\theta) = l'(\theta)$ ja informaatiofunktio $\mathcal{I}(\theta) = -l''(\theta)$. Oletetaan, että $\hat{\theta}$ on olemassa jossain Θ:n sisäpisteessä ja että $l(\theta)$:lla on pisteessä $\theta = \hat{\theta}$ Taylorin sarjakehitelmä:

$$l(\theta) = l(\hat{\theta}) + \frac{l'(\hat{\theta})}{1!}(\theta - \hat{\theta}) + \frac{l''(\hat{\theta})}{2!}(\theta - \hat{\theta})^2 + \frac{l'''(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \cdots .$$

Koska $l'(\hat{\theta}) = 0$ ja $r(\theta) = l(\theta) - l(\hat{\theta})$, saadaan

$$r(\theta) = -\frac{1}{2!}(\theta - \hat{\theta})^2 \mathcal{I}(\hat{\theta}) + \frac{l'''(\hat{\theta})}{3!}(\theta - \hat{\theta})^3 + \cdots . \tag{7.9.1}$$

Funktion $r(\theta)$ kehitelmään (7.9.1) perustuva likiarvo määritellään seuraavasti:

$$r_N(\theta) = -\tfrac{1}{2}(\theta - \hat{\theta})^2 \mathcal{I}(\hat{\theta}). \tag{7.9.2}$$

Jos $|\theta - \hat{\theta}|$ on pieni, niin lausekkeessa (7.9.1) kolmannen ja sitä korkeamman asteen termit ovat pieniä ja silloin $r(\theta) \approx r_N(\theta)$. Kun otoskoko kasvaa, on $|\theta - \hat{\theta}|$ suurella todennäköisyydellä pieni ja $r_N(\theta)$ on silloin $r(\theta)$:n hyvä likiarvo.

Parametrin θ $100p$ %:n uskottavuusväli on sellaisten θ:n arvojen joukko, että $r(\theta) \geq \log p$. Kun $r(\theta)$ korvataan likiarvollaan, saadaan epäyhtälöstä $r_N(\theta) \geq \log p$ parametrin θ $100p$ %:n uskottavuusvälin likiarvo

$$\left[\hat{\theta} \pm \sqrt{-\frac{2 \log p}{\mathcal{I}(\hat{\theta})}} \right]. \tag{7.9.3}$$

Esimerkki 7.11 Olkoon $x_1, x_2, \ldots, x_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, jossa μ on tuntematon ja σ^2 annettu. Silloin

$$\begin{aligned} L(\mu) &= c \cdot \prod_{i=1}^{n} f(x_i; \mu) = c \cdot \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\ &= \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right], \end{aligned}$$

kun valitaan $c = (2\pi\sigma^2)^{n/2}$. Logaritmoitu uskottavuusfunktio on

$$l(\mu) = -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2,$$

pistefunktio

$$u(\mu) = \frac{1}{\sigma^2} \sum (x_i - \mu)$$

ja informaatiofunktio

$$\mathcal{I}(\mu) = \frac{n}{\sigma^2}.$$

Ratkaisemalla yhtälö $u(\mu) = 0$ saadaan $\hat{\mu} = \overline{x}$ ja logaritminen normitettu uskottavuusfunktio on

$$r(\mu) = l(\mu) - l(\hat{\mu}) = -\frac{1}{2\sigma^2}\sum(x_i - \mu)^2 + \frac{1}{2\sigma^2}\sum(x_i - \overline{x})^2.$$

Sieventämällä lauseketta saadaan

$$r(\mu) = -\tfrac{1}{2}(\mu - \hat{\mu})^2\,\mathcal{I}(\hat{\mu}).$$

Normaalijakauman tapauksessa $r_N(\mu) = r(\mu)$ kaikilla μ:n arvoilla. Siksi likiarvoa (7.9.2) kutsutaan normaalijakaumaan perustuvaksi likiarvoksi. Uskottavuusväli voidaan kirjoittaa muodossa

$$\hat{\mu} \pm \sigma\sqrt{\frac{-2\log p}{n}}.$$

□

Esimerkki 7.12 Olkoon $x_1, x_2, \ldots, x_n$ otos eksponenttijakaumasta $\mathrm{Exp}(\theta)$. Esimerkissä 7.10 johdettiin θ:n uskottavuusfunktio, pistefunktio ja informaatiofunktio. Logaritminen normitettu uskottavuusfunktio on

$$r(\theta) = -n\left(\frac{\hat{\theta}}{\theta} - 1 - \log\frac{\hat{\theta}}{\theta}\right).$$

Tämän funktion likiarvo (7.9.2) on

$$r_N(\theta) = -\frac{1}{2}(\theta - \hat{\theta})^2\,\mathcal{I}(\hat{\theta}) = -\frac{n}{2}\left(\frac{\theta}{\hat{\theta}} - 1\right)^2.$$

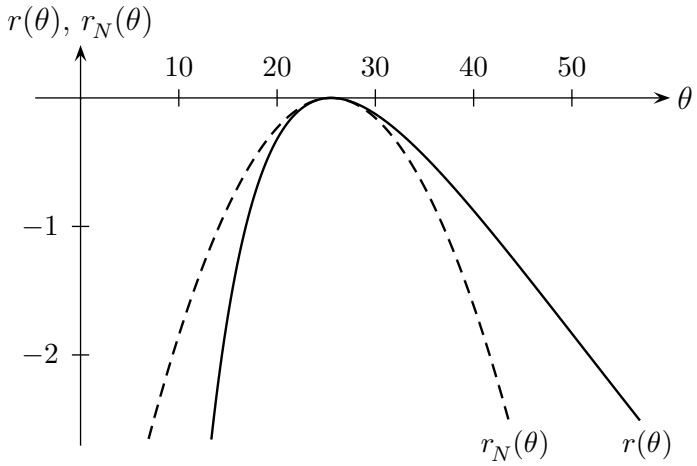
Jos esimerkiksi $n = 10$ ja $\hat{\theta} = 25.5$, saadaan

$$r(\theta) = -10\left(\frac{25.5}{\theta} - 1 - \log\frac{25.5}{\theta}\right) \qquad \text{ja} \qquad r_N(\theta) = -5\left(\frac{\theta}{25.5} - 1\right)^2.$$

Likiarvon $r_N(\theta)$ ja funktion $r(\theta)$ yhteensopivuus ei ole tässä tapauksessa hyvä, koska $r(\theta)$ on voimakkaasti epäsymmetrinen (Kuvio 7.1). □

7.10 Moniparametriset uskottavuusfunktiot

Seuraavassa tarkastellaan parametrien estimointia uskottavuusmenetelmällä, kun todennäköisyysmallissa on kaksi kaksi tai useampia tuntemattomia parametreja.



Kuvio 7.1. Logaritminen normitettu uskottavuusfunktio ja sen normaalijakaumaan perustuva likiarvo.

7.10.1 Pistevektori ja informaatiomatriisi

Oletetaan, että mallissa on kaksi tuntematonta parametria α ja β. Parametrien α ja β yhteinen uskottavuusfunktio on

$$L(\alpha, \beta) = c \cdot f(\boldsymbol{x}; \alpha, \beta), \tag{7.10.1}$$

missä $f(\boldsymbol{x}; \alpha, \beta)$ on havaintojen $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ yhteisjakauman todennäköisyysfunktio tai tiheysfunktio ja c on parametreista riippumaton vakio. Logaritmoitua (luonnollinen logaritmi) uskottavuusfunktiota merkitään $l(\alpha, \beta)$.

Parametriparin (α, β) suurimman uskottavuuden estimaatti $(\hat{\alpha}, \hat{\beta})$ on sellainen arvopari, joka maksimoi funktiot $L(\alpha, \beta)$ ja $l(\alpha, \beta)$. Kahden parametrin tapauksessa pistefunktio on vektoriarvoinen, eli

$$\boldsymbol{u}(\alpha, \beta) = \begin{pmatrix} u_1(\alpha, \beta) \\ u_2(\alpha, \beta) \end{pmatrix},$$

missä $u_1(\alpha, \beta) = \frac{\partial l}{\partial \alpha}$ ja $u_2(\alpha, \partial\beta) = \frac{\partial l}{\beta}$. Estimaatit $(\hat{\alpha}, \hat{\beta})$ löydetään tavallisesti ratkaisemalla yhtälö $\boldsymbol{u}(\alpha, \beta) = \boldsymbol{0}$, eli yhtälöpari

$$u_1(\alpha, \beta) = 0 \qquad \text{ja} \qquad u_2(\alpha, \beta) = 0. \tag{7.10.2}$$

Kahden parametrin tapauksessa informaatiofunktio $\mathcal{I}(\alpha, \beta)$ on matriisi

$$\mathcal{I}(\alpha, \beta) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 l}{\partial \alpha^2} & -\frac{\partial^2 l}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l}{\partial \beta \partial \alpha} & -\frac{\partial^2 l}{\partial \beta^2} \end{pmatrix}, \tag{7.10.3}$$

missä i_{jk}:t ovat siis α:n ja β:n funktiota. Olkoon informaatiomatriisin $\mathcal{I}(\alpha, \beta)$ arvo pisteessä $(\hat{\alpha}, \hat{\beta})$

$$\mathcal{I}(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} \hat{\imath}_{11} & \hat{\imath}_{12} \\ \hat{\imath}_{21} & \hat{\imath}_{22} \end{pmatrix}.$$

Paikallisessa maksimissa $\mathcal{I}(\hat{\alpha}, \hat{\beta})$ on positiivisesti definiitti. Silloin $\mathcal{I}(\hat{\alpha}, \hat{\beta})$:n alkiot toteuttavat ehdot $\hat{\imath}_{11} > 0$, $\hat{\imath}_{22} > 0$ ja $\hat{\imath}_{11}\hat{\imath}_{22} - \hat{\imath}_{12}^2 > 0$.

Esimerkki 7.13 Mitataan kaksi kappaletta erikseen ja yhdessä, joten saadaan kolme mittaustulosta X_1, X_2 ja X_3. Kappaleiden painot μ_1 ja μ_2 ovat tuntemattomia. Tiedetään kokemuksesta, että mittaukset ovat toisistaan riippumattomat ja noudattavat normaalijakaumaa, jonka varianssi on yksi. Muuttujat X_1, X_2 ja X_3 ovat siis keskenään riippumattomat ja

$$X_1 \sim \mathrm{N}(\mu_1, 1); \qquad X_2 \sim \mathrm{N}(\mu_2, 1); \qquad X_3 \sim \mathrm{N}(\mu_1 + \mu_2, 1).$$

Havaintojen X_1, X_2, X_3 yhteisjakauman tiheysfunktio on

$$f(x_1, x_2, x_3) = \left(\frac{1}{\sqrt{2\pi}}\right)^3 \exp\left\{-\tfrac{1}{2}\left[(x_1-\mu_1)^2 + (x_2-\mu_2)^2 + (x_3-\mu_1-\mu_2)^2\right]\right\}.$$

Uskottavuusfunktion ydin on

$$L(\mu_1, \mu_2) = \exp\left\{-\tfrac{1}{2}\left[(x_1-\mu_1)^2 + (x_2-\mu_2)^2 + (x_3-\mu_1-\mu_2)^2\right]\right\}$$

ja logaritmoitu uskottavuusfunktio on

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}\left[(x_1-\mu_1)^2 + (x_2-\mu_2)^2 + (x_3-\mu_1-\mu_2)^2\right].$$

Pistefunktion komponentit ovat

$$u_1(\mu_1, \mu_2) = \frac{\partial\, l}{\partial \mu_1} = (x_1-\mu_1) + (x_3-\mu_1-\mu_2);$$

$$u_2(\mu_1, \mu_2) = \frac{\partial\, l}{\partial \mu_2} = (x_2-\mu_2) + (x_3-\mu_1-\mu_2).$$

Toiset derivaatat ovat

$$\frac{\partial^2 l}{\partial \mu_1^2} = -2, \qquad \frac{\partial^2 l}{\partial \mu_2^2} = -2 \qquad \text{ja} \qquad \frac{\partial^2 l}{\partial \mu_1 \partial \mu_2} = -1,$$

joten informaatiomatriisi on

$$\mathcal{I}(\mu_1, \mu_2) = -\begin{pmatrix} -2 & -1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Tässä tapauksessa $\mathcal{I}(\mu_1, \mu_2)$ ei riipu parametreista. Suurimman uskottavuuden estimaattien määrittämiseksi ratkaistaan yhtälöpari

$$u_1(\mu_1, \mu_2) = 0 \qquad \text{ja} \qquad u_2(\mu_1, \mu_2) = 0.$$

Silloin ensimmäisestä yhtälöstä saadaan

$$\hat{\mu}_1(\mu_2) = \tfrac{1}{2}(x_1 + x_3 - \mu_2),$$

joka on μ_1:n suurimman uskottavuuden estimaatti, jos μ_2 tunnetaan. Kun $\mu_1 = \frac{1}{2}(x_1 + x_3 - \mu_2)$ sijoitetaan toiseen yhtälöön ja ratkaistaan μ_2, saadaan

$$\hat{\mu}_2 = \tfrac{1}{3}(2x_2 + x_3 - x_1) = \tfrac{2}{3}x_2 + \tfrac{1}{3}(x_3 - x_1).$$

Tästä saadaan

$$\hat{\mu}_1 = \tfrac{1}{2}(x_1 + x_3 - \hat{\mu}_2) = \tfrac{1}{3}(2x_1 + x_3 - x_2) = \tfrac{2}{3}x_1 + \tfrac{1}{3}(x_3 - x_2).$$

□

7.10.2 Normitettu uskottavuus ja korkeuskäyrät

Parametrin α ja β *yhteinen normitettu uskottavuusfunktio* (YNUF) määritellään suhteena

$$R(\alpha,\beta)=\frac{L(\alpha,\beta)}{L(\hat\alpha,\hat\beta)} \tag{7.10.4}$$

Huomaa, että $0\le R(\alpha,\beta)\le 1$ ja $R(\hat\alpha,\hat\beta)=1$. Logaritmoitu yhteinen normitettu uskottavuusfunktio r on R:n logaritmi, kuten yhden parametrinkin tapauksessa:

$$r(\alpha,\beta)=\log R(\alpha,\beta)=l(\alpha,\beta)-l(\hat\alpha,\hat\beta).$$

Parametrin arvojen (α_0,β_0) normitettu uskottavuus (NU) on

$$R(\alpha_0,\beta_0)=\frac{L(\alpha_0,\beta_0)}{L(\hat\alpha,\hat\beta)}.$$

Jos $R(\alpha_0,\beta_0)\approx 0$, niin arvopari (α_0,β_0) ei ole uskottava, koska on muita sellaisia arvopareja, joilla havainnot ovat paljon todennäköisempiä. Parametrien arvot voidaan asettaa $R(\alpha,\beta)$:n avulla uskottavuuden suhteen järjestykseen.

$100p$ %:n *uskottavuusalue* on epäyhtälön $R(\alpha,\beta)\ge p$ toteuttavien parametrin arvojen joukko:

$$\mathrm{ua}_p(\alpha,\beta)=\{\,(\alpha,\beta)\mid R(\alpha,\beta)\ge p\,\}.$$

Käyrä $R(\alpha,\beta)=p$ on alueen $\mathrm{ua}_p(\alpha,\beta)$ reuna ja sitä kutsutaan *uskottavuuskäyräksi*.

Esimerkki 7.14 Esimerkissä 7.13 parametrien μ_1 ja μ_2 logaritmoitu uskottavuusfunktio oli

$$l(\mu_1,\mu_2)=-\tfrac12\big[(x_1-\mu_1)^2+(x_2-\mu_2)^2+(x_3-\mu_1-\mu_2)^2\big].$$

Oletetaan, että saatiin mittaustulokset $x_1=15.6$, $x_2=29.3$ ja $x_3=45.8$. Silloin $\hat\mu_1=15.9$ ja $\hat\mu_2=29.6$, joten

$$l(\hat\mu_1,\hat\mu_2)=-\tfrac12(0.3^2+0.3^2+0.3^2)=-0.135.$$

Siksi μ_1:n ja μ_2:n yhteinen logaritmoitu normitettu uskottavuusfunktio on

$$r(\hat\mu_1,\hat\mu_2)=l(\mu_1,\mu_2)+0.135.$$

Esimerkiksi 10 %:n uskottavuuskäyrä on

$$-\tfrac12\big[(15.6-\mu_1)^2+(29.3-\mu_2)^2+(45.8-\mu_1-\mu_2)^2\big]+0.135=\log 0.1.$$

Yhtälön kuvaaja on $(\hat\mu_1,\hat\mu_2)$-keskinen ellipsi. 10 %:n uskottavuusalue on tämän ellipsin rajaama alue. □

7.10.3 Profiiliuskottavuus

Vaikka mallissa on kaksi tuntematonta parametria (α, β), saattaa olla, että olemme kiinnostuneita vain toisesta parametrista, sanokaamme β:sta. Parametria α pidetään kiusaparametrina. Parametrien yhteinen normitettu uskottavuusfunktio asettaa arvoparit (α, β) niiden uskottavuuden mukaiseen järjestykseen. Haluamme nyt mallit vain β:aa koskevan informaation mukaiseen järjestykseen.

Olkoon $\hat{\alpha}_\beta$ parametrin α arvo, joka maksimoi uskottavuusfunktion annetulla β:n arvolla. Määritellään *uskottavuuden profiilifunktio* β:n suhteen seuraavasti:

$$L^*(\beta) = \max_\alpha L(\alpha, \beta) = L(\hat{\alpha}_\beta, \beta). \tag{7.10.5}$$

Vastaavasti *logaritmoitu profiilifunktio* on

$$l^*(\beta) = \log L^*(\hat{\alpha}_\beta, \beta), \tag{7.10.6}$$

normitettu profiilifunktio

$$R^*(\beta) = \max_\alpha R(\alpha, \beta) = R(\hat{\alpha}_\beta, \beta) \tag{7.10.7}$$

ja *logaritmoitu normitettu profiilifunktio* on

$$r^*(\beta) = r(\hat{\alpha}_\beta, \beta) = l(\hat{\alpha}_\beta, \beta) - l(\hat{\alpha}, \hat{\beta}). \tag{7.10.8}$$

Koska $\hat{\alpha}_\beta$ on α:n suurimman uskottavuuden estimaatti, kun β on kiinnitetty, sen arvo saadaan usein ratkaisemalla yhtälö $u_1(\alpha, \beta)$.

Funtio $R^*(\beta)$ muistuttaa ominaisuuksiltaan vastaavaa normitettua yhden parametrin uskottavuusfunktiota. Esimerkiksi

$$0 \le R^*(\beta) \le 1 \qquad \text{ja} \qquad R^*(\hat{\beta}) = 1.$$

Funktioiden $R^*(\beta)$ tai $r^*(\beta)$ avulla voidaan määrittää β:n $100p$ %:n *suurimman uskottavuuden alue* (tai väli), joka on $\{\, \beta \mid R^*(\beta) \ge p \,\}$.

Esimerkki 7.15 Tarkastellaan Esimerkin 7.13 logaritmoitua uskottavuusfunktiota

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(x_1 - \mu_1)^2 - \tfrac{1}{2}(x_2 - \mu_2)^2 - \tfrac{1}{2}(x_3 - \mu_1 - \mu_2)^2.$$

Kuten Esimerkissä 7.13 todettiin, μ_1:n suurimman uskottavuuden estimaatti on

$$\hat{\mu}_1(\mu_2) = \tfrac{1}{2}(x_1 + x_3 - \mu_2)$$

ehdolla, että μ_2 on annettu. Logaritmoitu normitettu profiilifunktio on

$$r^*(\mu_2) = l[\hat{\mu}_1(\mu_2), \mu_2] - l(\hat{\mu}_1, \hat{\mu}_2),$$

josta sieventämällä saadaan

$$r^*(\mu_2) = -\tfrac{3}{4}(\mu_2 - \hat{\mu}_2)^2.$$

Jos $\hat{\mu}_2 = 29.6$, kuten Esimerkissä 7.14, niin μ_2:n suurimman uskottavuuden väli on $27.85 \le \mu_2 \le 31.35$. □

7.10.4 Normaalijakaumaan perustuva likiarvo

Alaluvussa 7.9 johdettiin logaritmoidun normitetun uskottavuusfunktion normaalijakaumaan perustuva likiarvo

$$r(\theta) \approx -\tfrac{1}{2}(\theta - \hat{\theta})^2 \mathcal{I}(\hat{\theta})$$

Taylorin sarjakehitelmän avulla. Kahden parametrin tapauksessa voidaan johtaa vastaavanlainen likiarvo:

$$r_N(\alpha, \beta) = -\tfrac{1}{2}(\alpha - \hat{\alpha})^2 \hat{\imath}_{11} - \tfrac{1}{2}(\beta - \hat{\beta})^2 \hat{\imath}_{22} - (\alpha - \hat{\alpha})(\beta - \hat{\beta})\hat{\imath}_{12}, \tag{7.10.9}$$

missä $\hat{\imath}_{kj} = i(\hat{\alpha}, \hat{\beta})$ ovat informaatiomatriisin alkioita. Merkitään $\boldsymbol{\theta} = (\alpha, \beta)^\mathsf{T}$, $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta})^\mathsf{T}$ ja

$$\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \hat{\imath}_{11} & \hat{\imath}_{12} \\ \hat{\imath}_{21} & \hat{\imath}_{12} \end{pmatrix},$$

jolloin likiarvo (7.10.9) voidaan kirjoittaa matriisimuodossa

$$r_N(\boldsymbol{\theta}) = -\tfrac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\mathsf{T} \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Kun $r_N(\alpha, \beta)$ derivoidaan α:n suhteen, saadaan pistefunktion $u_1(\alpha, \beta)$ likiarvo

$$u_1(\alpha, \beta) \approx (\alpha - \hat{\alpha})\hat{\imath}_{11} - (\beta - \hat{\beta})\hat{\imath}_{12},$$

joka on α:n ja β:n lineaarinen funktio. Asettamalla likiarvo nollaksi ja ratkaisemalla α saadaan

$$\hat{\alpha}(\beta) \approx \hat{\alpha} + (\hat{\beta} - \beta)\hat{\imath}_{12}/\hat{\imath}_{11}. \tag{7.10.10}$$

Kun ratkaisu sitten sijoitetaan α:n paikalle lausekkeeseen (7.10.9), saadaan β:n logarimoidun normitetun profiilifunktion likiarvo:

$$r^*(\beta) \approx -\tfrac{1}{2}(\beta - \hat{\beta})^2(\hat{\imath}_{22} - \hat{\imath}_{12}^2/\hat{\imath}_{11}). \tag{7.10.11}$$

Likiarvo (7.10.11) voidaan kirjoittaan myös muodossa

$$r^*(\beta) \approx -\tfrac{1}{2}(\beta - \hat{\beta})^2/\hat{\imath}^{22}, \tag{7.10.12}$$

missä $\hat{\imath}^{22}$ on informaatiomatriisin $\mathcal{I}(\hat{\alpha}, \hat{\beta})$ käänteismatriisin $(2, 2)$-alkio.

Esimerkki 7.16 Tarkastellaan Esimerkin 7.13 kaksiulotteista normaalijakaumaa, jossa parametrien μ_1 ja μ_2 logaritmoitu uskottavuusfunktio

$$l(\mu_1, \mu_2) = -\tfrac{1}{2}(x_1 - \mu_1)^2 - \tfrac{1}{2}(x_2 - \mu_2)^2 - \tfrac{1}{2}(x_3 - \mu_1 - \mu_2)^2$$

on μ_1:n ja μ_2:n 2. asteen polynomi. Tästä seuraa, että likiarvot (7.10.9) ja (7.10.11) pitävät tässä tapauksessa täsmällisesti paikkansa. Koska $\hat{\imath}_{11} = \hat{\imath}_{22} = 2$ ja $\hat{\imath}_{12} = 1$, niin

$$r_N(\mu_1, \mu_2) = -(\mu_1 - \hat{\mu}_1)^2 - (\mu_2 - \hat{\mu}_2)^2 - (\mu_1 - \hat{\mu}_1)(\mu_2 - \hat{\mu}_2).$$

Funktion $r(\mu_1, \mu_2)$ tasa-arvokäyrät (korkeuskäyrät) ovat ellipsejä. □

7.11 Normaalinen regressiomalli

7.11.1 Ehdollinen normaalimalli

Oletetaan, että satunnaismuuttujat $Y_1, \ldots, Y_n$ ovat riippumattomat ja

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2), \qquad 1 \le i \le n. \tag{7.11.1}$$

Havaittu aineisto muodostuu n:stä arvoparista $(x_1, y_1), \ldots, (x_n, y_n)$. Ennustemuuttujan x arvot $x_1, \ldots, x_n$ ajatellaan tunnetuiksi vakioiksi. Populaation regressiofunktio on siis

$$E(Y \mid x) = \alpha + \beta x$$

ja kaikilla satunnaismuuttujilla Y_i on sama varianssi σ^2. Malli (7.11.1) voidaan myös lausua muodossa

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad 1 \le i \le n,$$

missä $\varepsilon_1, \ldots, \varepsilon_n$ ovat riippumattomat ja $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $1 \le i \le n$.

Havaintojen $Y_1, \ldots, Y_n$ yhteisjakauman tiheysfunktio on

$$\begin{aligned} f(y_1, \ldots, y_n \mid \alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} f_1(y_i \mid \alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]. \end{aligned}$$

Tästä nähdään, että uskottavuusfunktion ydin on

$$L(\alpha, \beta, \sigma^2) = (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]$$

ja logaritmoitu uskottavuusfunktio on

$$l(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

7.11.2 Kahden muuttujan normaalimalli

Oletetaan, että havaintoparit $(x_1, y_1), \ldots, (x_n, y_n)$ ovat satunnaisvektorien $(X_1, Y_1), \ldots, (X_n, Y_n)$ arvoja. Satunnaisvektorit ovat riippumattomat ja noudattavat kaksiulotteista normaalijakaumaa

$$(X_i, Y_i) \sim \mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho), \qquad 1 \le i \le n.$$

Otoksen $(X_1, Y_1), \ldots, (X_n, Y_n)$ yhteisjakauman tiheysfunktio on kahden muuttujan normaalijakaumien tiheysfunktioiden tulo.

Koska regressiossa ajatellaan X selittäjäksi ja Y vasteeksi, niin tarkastelemme vastemuuttujan käyttäytymistä x:n funktiona. Haluamme siis estimoida ehdollisen odotusarvon

$$E(Y \mid x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \left(\mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x\right) + \rho\frac{\sigma_y}{\sigma_x}x.$$

Kahden muuttujan normaalimallista seuraa, että populaation regressiofunktio on x:n lineaarinen funktio. Tässä tapauksessa

$$E(Y \mid x) = \alpha + \beta x,$$

missä $\beta = \rho(\sigma_x/\sigma_y)$ ja $\alpha = \mu_y - \rho(\sigma_x/\sigma_y)\mu_x$. Tässä mallissa Y:n ehdollinen varianssi ehdolla $X = x$ ei riipu x:n arvosta

$$\mathrm{Var}(Y \mid x) = \sigma_y^2(1 - \rho^2).$$

7.11.3 Yksinkertainen lineaarinen regressio

Lineaarisessa regressiossa oletetaan, että vastemuuttuja riippuu lineaarisesti selittäjistä. Malli on muotoa

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad 1 \le i \le n,$$

missä Y_i on havaittava satunnaismuuttuja ja ε_i on virhetermi, α ja β ovat tuntemattomia vakioita ja $x_1, \ldots, x_n$ ovat tuntemattomia koevakioita. Oletamme, että $E(\varepsilon_i) = 0$, joten

$$E(Y_i) = \alpha + \beta x_i. \tag{7.11.2}$$

Itse asiassa (7.11.2) on ehdollinen odotusarvo

$$E(Y_i \mid x_i) = \alpha + \beta x_i.$$

Tässä oletetaan, että regressiofunktio on lineaarinen. Pienimmän neliösumman keinolla saadaan parametrien estimaateiksi

$$\hat{\beta} = \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = \frac{s_y}{s_x} \cdot r$$
$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

Voidaan osoittaa, että $\hat{\alpha}$ ja $\hat{\beta}$ ovat minimivarianssisia kaikkien α:n ja β:n lineaaristen harhattomien estimaattorien joukossa.

7.12 Probit- ja logitmallit: Esimerkki

Tutkitaan lääkeannoksen vaikutusta kokeilemalla eri annosmääriä. Olkoot kokeiltavat annokset $x_1, x_2, \ldots, x_k$. Annokset ilmoitetaan usein vaikuttavan aineen pitoisuuden logaritmina, joten annos $x \to -\infty$, kun pitoisuus lähenee nollaa. Oletetaan, että jokainen yksilö joko reagoi annokseen x tai ei reagoi, joten vastemuuttuja on dikotominen. Esimerkiksi uusi lääke parantaa potilaan tilaa (positiivinen vaste) tai ei paranna (ei vastetta). Kun hyönteismyrkkyä kokeillaan hyönteisiin, hyönteiset joko reagoivat myrkkyyn (kuolevat) tai eivät reagoi (jäävät henkiin).

Olkoon $\pi(x)$ todennäköisyys, että yksilö reagoi annokseen x. Oletamme, että $\pi(x)$ on x:n ei-vähenevä funktio. Lisäksi oletamme, että $\pi(x) \to 0$, kun $x \to -\infty$ ja $\pi(x) \to 1$, kun $x \to \infty$. Ajattelemme siis, että yksilö ei reagoi, kun annos on riittävän pieni. Toisaalta taas yksilö reagoi, kun annos on riittävän suuri. Tähän tulokseen voidaan päätyä myös kuvittelemalla, että yksilöillä on erilainen lääkkeen sietokyky. Olkoon X minimiannos, joka saa satunnaisesti valitun yksilön reagoimaan. Annos x aiheuttaa reaktion silloin ja vain silloin, kun yksilön sietokyky on korkeintaan x. Silloin todennäköisyys $\pi(x)$ saada vaste annokseen x on

$$\pi(x) = P(X \leq x) = F(x),$$

missä F on satunnaismuuttujan X kertymäfunktio.

7.12.1 Probitmalli

Oletetaan, että sietokyky $X \sim \mathrm{N}(\mu, \sigma^2)$. Silloin

$$\begin{aligned} \pi(x) = P(X \leq x) &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(\alpha + \beta x), \end{aligned} \tag{7.12.1}$$

missä $\alpha = -\mu/\sigma$, $\beta = 1/\sigma$ ja Φ on standardoidun normaalijakauman kertymäfunktio. Yhtälö voidaan myös lausua muodossa

$$\Phi^{-1}(\pi) = \alpha + \beta x, \tag{7.12.2}$$

missä Φ^{-1} on standardimuotoisen normaalijakauman kertymäfunktion käänteisfunktio. Mallia (7.12.2) kutsutaan probit malliksi.

7.12.2 Logistinen malli

Logistisen jakauman kertymäfunktio

$$G(z) = 1 - \frac{1}{1 + \mathrm{e}^z}, \qquad -\infty < z < \infty$$

on muodoltaan hyvin lähellä normaalijakaumaa. Historiallisesti logistisen jakauman suosio regressiosovelluksissa perustunee osittain siihen, että sen kertymäfunktion avulla todennäköisyydet voidaan laskea helposti ilman numeerista integrointia.

Kun yhtälössä (7.12.1) korvataan Φ logistisen jakauman kertymäfunktiolla G saadaan

$$\pi(x) = G(\alpha + \beta x) = 1 - \frac{1}{1 + \mathrm{e}^{\alpha+\beta x}}. \tag{7.12.3}$$

Kun ratkaistaan yhtälö $G(z) = \pi$, saadaan $z = \log\left(\frac{\pi}{1-\pi}\right)$ ja siksi malli voidaan kirjoittaa muodossa

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x, \tag{7.12.4}$$

missä log on luonnollinen logaritmi. Mallia (7.12.4) kutsutaan logistiseksi malliksi ja muunnosta $\log\left(\frac{\pi}{1-\pi}\right)$ todennäköisyyden π *logistiseksi muunnokseksi*.

7.12.3 Suurimman uskottavuuden estimointi

Oletetaan, että n_i yksilöä saa annoksen x_i ja Y_i yksilöä reagoi. Silloin $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, $i = 1, 2, \ldots, k$ ja

$$\pi_i = \frac{\mathrm{e}^{\alpha+\beta x_i}}{1 + \mathrm{e}^{\alpha+\beta x_i}}.$$

Otoksen $Y_1, Y_2, \ldots, Y_k$ yhteisjakauman todennäköisyysfunktio on

$$f(y_1, y_2, \ldots, y_k) = \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

Uskottavuusfunktion ydin on

$$L(\alpha, \beta) = \prod_{i=1}^{k} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} = \prod_{i=1}^{k} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} (1 - \pi_i)^{n_i},$$

ja logaritmoitu uskottavuusfunktio

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^{k} \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + n_i \log(1 - \pi_i) \right] \\ &= \sum_{i=1}^{k} \left[y_i(\alpha + \beta x_i) + n_i \log(1 - \pi_i) \right]. \end{aligned}$$

Koska $\pi_i = \mathrm{e}^{\alpha+\beta x_i}/(1+\mathrm{e}^{\alpha+\beta x_i})$, niin

$$\frac{\partial \pi_i}{\alpha} = \frac{\mathrm{e}^{\alpha+\beta x_i}}{(1+\mathrm{e}^{\alpha+\beta x_i})^2} = \pi_i(1-\pi_i);$$

$$\frac{\partial \pi_i}{\beta} = \frac{\mathrm{e}^{\alpha+\beta x_i} x_i}{(1+\mathrm{e}^{\alpha+\beta x_i})^2} = \pi_i(1-\pi_i)x_i.$$

Derivoimalla logaritmoitu uskottavuusfunktio $l(\alpha, \beta)$ saadaan osittaisderivaatat

$$\begin{aligned}\frac{\partial\, l}{\partial \alpha} &= \sum_{i=1}^{k}\left(y_i - \frac{n_i}{1-\pi_i}\cdot\frac{\partial \pi_i}{\partial \alpha}\right)\\ &= \sum_{i=1}^{k}\left[y_i - \frac{n_i}{1-\pi_i}\cdot \pi_i(1-\pi_i)\right] = \sum_{i=1}^{k}(y_i - n_i\pi_i)\end{aligned}$$

ja

$$\begin{aligned}\frac{\partial\, l}{\partial \beta} &= \sum_{i=1}^{k}\left(x_i y_i - \frac{n_i}{1-\pi_i}\cdot\frac{\partial \pi_i}{\partial \beta}\right)\\ &= \sum_{i=1}^{k}\left[x_i y_i - \frac{n_i}{1-\pi_i}\cdot \pi_i(1-\pi_i)x_i\right] = \sum_{i=1}^{k}(y_i - n_i\pi_i)x_i.\end{aligned}$$

Tästä seuraa, että pistefunktiot voidaan kirjoittaa muodossa

$$u_1(\alpha,\beta) = \frac{\partial\, l}{\alpha} = \sum_{i=1}^{k}(y_i - \mu_i);$$

$$u_2(\alpha,\beta) = \frac{\partial\, l}{\beta} = \sum_{i=1}^{k}(y_i - \mu_i)x_i,$$

missä $E(y_i) = \mu_i = n_i\pi_i$.

Suurimman uskottavuuden estimaatit saadaan ratkaisemalla yhtälöpari

$$u_1(\alpha,\beta) = 0 \qquad \text{ja} \qquad u_2(\alpha,\beta) = 0.$$

Nämä yhtälöt ovat α:n ja β:n suhteen epälineaarisia ja ne on ratkaistava jollakin numeerisella iteratiivisella menetelmällä.

7.13 Uskottavuusyhtälön ratkaiseminen numeerisesti

Yhden parametrin tapauksessa uskottavuusyhtälö on muotoa

$$u(\theta) = 0,$$

missä $u(\theta) = l'(\theta)$. Newtonin menetelmä tarjoaa erään numeerisen menetelmän uskottavuusyhtälön ratkaisemiseksi. Olkoon θ_0 jokin parametrin θ alkuarvo. Silloin päivityskaavalla

$$\theta_1 = \theta_0 + \frac{u(\theta_0)}{\mathcal{I}(\theta_0)}$$

saadaan uusi arvio, kun $u(\theta_0)$ on pistefunktion arvo ja $\mathcal{I}(\theta_0)$ informaatiofunktion arvo pisteessä $\theta = \theta_0$. Näin saadaan arvojono $\theta_1, \theta_2, \theta_3, \ldots$, jota voidaan jatkaa, kunnes peräkkäisten arvojen erotuksen itseisarvo $|\theta_i - \theta_{i-1}|$ on riittävän pieni, jolloin $u(\theta_i) \approx 0$.

Kahden tai useamman parametrin tapauksessa voidaan käyttää Newtonin ja Raphsonin menetelmää, joka on Newtonin menetelmän yleistys. Olkoon $\boldsymbol{\theta} = (\alpha, \beta)^\mathsf{T}$ parametrivektori. Silloin saadaan päivityskaava

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{u}(\boldsymbol{\theta}_0),$$

missä $\boldsymbol{u}(\boldsymbol{\theta}_0) = [u_1(\alpha_0, \beta_0),\ u_2(\alpha_0, \beta_0)]^\mathsf{T}$ on alkuarvolla $\boldsymbol{\theta}_0 = (\alpha_0, \beta_0)^\mathsf{T}$ laskettu pistevektorin arvo ja $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$ on pisteessä $\boldsymbol{\theta}_0$ laskettu informaatiomatriisin arvo.

Harjoituksia

1. Oletetaan, että puustotauti A on levinnyt satunnaisesti ja tasaisesti yli laajan metsäalueen ja sairastuneita puita on keskimäärin λ kappaletta hehtaaria kohden. Kymmenellä satunnaisesti valitulla neljän hehtaarin koeaalalla havaittujen sairaiden puiden lukumäärät olivat: 0, 1, 3, 0, 0, 2, 2, 0, 1, 1. Määritä λ:n suurimman uskottavuuden estimaatti (Vihje: Poissonin jakauma).

2. Bakteerien lukumäärä X joesta otetussa vesinäytteessä noudattaa Poissonin jakaumaa, jonka keskiarvo on μ. Kun analysoitiin n näytettä, saatiin tulokseksi seuraava aineisto:

Bakteerien lukumäärä	0	1	2	...	Yhteensä
Havaittu frekvenssi	f_0	f_1	f_2	...	n

(a) Määritä μ:n uskottavuusfunktio ja logaritmoitu uskottavuusfunktio.

(b) Määritä μ:n suurimman uskottavuuden estimaatti.

3. Tarkastellaan koetta, jonka tarkoituksena on määrittää verytyypin A suhteellinen osuus θ eräässä suuressa populaatiossa. Tarkastellaan seuraavia kahta tilannetta:

(a) Valitaan henkilöitä satunnaisesti (yksitellen), kunnes löydetään 10 verityyppiä A. Havaittiin, että oli testattava 100 henkilöä.

(b) Valitaan 100:n hengen otos, josta löydettiin 10 A-tyypin henkilöä.

Lausu kummassakin tapauksessa θ:n uskottavuusfunktio ja osoita, että funktiot ovat suhteellisuuskerrointa vaille samat (ts. saadaan samat suurimman uskottavuuden estimaatit).

4. Erään geneettisen teorian mukaan verityyppien MM, NM ja NN suhteelliset osuudet suuressa populaatiossa ovat θ^2, $2\theta(1-\theta)$ ja $(1-\theta)^2$, missä $0 \leq \theta \leq 1$ on tuntematon parametri.

(a) Valitaan populaatiosta n:n alkion otos, jossa eri verityyppien lukumäärät olivat x_1, x_2 ja x_3. Esitä $\hat{\theta}$:n lauseke.

(b) Olkoon $n = 100$ ja havaitut frekvenssit 32, 46 ja 22. Laske $\hat{\theta}$ ja odotettujen frekvenssien estimaattit.

5. Heitetään n kertaa sellaista yleistettyä noppaa (suorakulmainen särmiö), että 4 tahoa ovat keskenään yhtä todennäköiset ($p_1 = p_2 = p_3 = p_4 = p$) ja kaksi muuta tahoa (vastakkaiset) ovat vastaavasti keskenään yhtä todennäköiset ($p_5 = p_6 = q$). Oletetaan, että i. taho sattuu x_i kertaa ($i = 1, 2, \ldots, 6$), missä $\sum x_i = n$.

(a) Osoita, että $\hat{\theta} = (3t - 2n)/12n$, kun $p = 1/6 + \theta$ ja $t = x_1 + x_2 + x_3 + x_4$.

(b) Oletetaan, että havaitut frekvenssit ovat 11, 15, 13, 15, 22, 24. Laske odotettujen frekvenssien estimaatit, kun noppaa koskevat oletukset (malli) oletetaan oikeiksi.

6. Hatussa on N arpalippua, jotka on numeroitu peräkkäin $1, 2, \ldots, N$. Hatusta valittiin satunnaisotannalla palauttaen 8 arpalippua, joiden järjestysnumerot olivat 137, 24, 86, 33, 92, 129, 17, 111. Osoita, että $\hat{N} = 137$.

7. Haluttiin selvittää tautiin A tartunnan saaneiden suhteellinen osuus θ eräässä populaatiossa. Analysoitiin nk:lta satunnaisesti valitulta henkilöltä saadut verinäytteet. Ajan säästämiseksi yhdistettiin aina k näytettä yhdeksi näytteeksi, jolloin saatiin n yhdistettyä näytettä. Yhdistetty näyte on negatiivinen, jos kellään näytteeseen kuuluvalla k:lla henkilöllä ei ole tartuntaa, muutoin näyte on positiivinen. Yhdistetyistä näytteistä on x negatiivista ja $n-x$ positiivista. Määritä suurimman uskottavuuden estimaattori $\hat{\theta}$.

8. Oletetaan, että $\hat{\theta}$ on estimaattien $\hat{\theta}_1$ ja $\hat{\theta}_2$ painotettu keskiarvo. Silloin siis

$$\hat{\theta} = a\hat{\theta}_1 + (1-a)\hat{\theta}_2,$$

missä $0 \leq a \leq 1$.

(a) Osoita, että $\hat{\theta}$:n arvo on $\hat{\theta}_1$:n ja $\hat{\theta}_2$:n välissä.

(b) Olkoon

$$\hat{\theta} = a_1\hat{\theta}_1 + \cdots + a_n\hat{\theta}_n,$$

missä a_i:t ovat positiivisia ja $\sum a_i = 1$. Osoita, että $\hat{\theta}$ on pienimmän ja suurimman $\hat{\theta}_i$:n välissä.

9. Piirrä 1. tehtävän tapauksessa logaritminen normitettu uskottavuusfunktio ja sen perusteella 50 %:n ja 10 %:n uskottavuusvälit.

10. Määritä tehtävässä 1.5 (yleistetty noppa) arvon $\theta = 0$ (symmetrinen noppa) suhteellinen uskottavuus ja logaritmisen normitetun uskottavuusfunktion arvo pisteessä $\theta = 0$. Onko tämä parametrin arvo uskottava?

11. Perheiden tuloja X mitataan sellaisella asteikolla, että arvo $X = 1$ on minimipalkka. Oletetaan, että tulojakauman tiheysfunktio on

$$f(x) = \frac{\theta}{x^{\theta+1}}, \qquad \text{kun} \quad x \geq 1$$

ja $\theta > 0$. Satunnaisesti valittujen 10 perheen tulot olivat:

$$1.02,\ 1.41,\ 1.75,\ 2.31,\ 3.42,\ 4.31,\ 9.21,\ 17.4,\ 38.6,\ 392.8$$

Määritä θ:n suurimman uskottavuuden estimaatti ja 10 %:n uskottavuusväli.

12. Kun tietyn metallin pitoisuus liuoksessa mitataan, noudattaa mittauksessa syntyvä virhe normaalijakaumaa $\mathrm{N}(0, \sigma^2)$. Jos oikea pitoisuus on μ, niin havaittu pitoisuus $X \sim \mathrm{N}(\mu, \sigma^2)$. Varianssi σ^2 tunnetaan aikaisemmasta kokemuksesta.

(a) Olkoot $x_1, x_2, \ldots, x_n$ tuntemattoman pitoisuuden μ riippumattomat mittaukset. Osoita, että $\hat{\mu} = \overline{x}$ ja logaritminen normitettu uskottavuusfunktio on

$$r(\mu) = -\frac{n}{2\sigma^2}(\overline{x} - \mu)^2, \qquad -\infty < \mu < \infty.$$

(Vihje: Osoita $\sum(x_i - \mu)^2 = \sum(x_i - \bar{x}_i)^2 + n(\overline{x}_i - \mu)^2$.)

(b) Laimennetaan alkuperäisen liuoksen pitoisuus puoleen niin, että pitoisuus on $\mu/2$. Tehdään lisämittaukset $y_1, y_2, \ldots, y_m$. Määritä kaikkiin $n + m$ mittaukseen perustuva μ:n suurimman uskottavuuden estimaatti.

13. Kimmoisesta materiaalista valmistettuja näytekappaleita testattiin iskemällä niitä toistuvasti vasaralla vakiovoimalla, kunnes havaittiin murtuma. Jos näytekappale kestää iskun vaurioitumatta todennäköisyydellä θ,

niin murtumiseen tarvittavien iskujen lukumäärä X noudattaa geometrista jakaumaa

$$f(x) = \theta^{x-1}(1-\theta); \qquad x = 1, 2, \ldots.$$

Kun testattiin 200 näytettä, saatiin seuraavat tulokset:

Iskujen lkm	1	2	3	Ainakin 4	Yhteensä
Näytteiden lkm	112	36	22	30	200

Laske θ:n suurimman uskottavuuden estimaatti ja odotetut frekvenssit.

14. Tuotteen kestonaika noudattaa eksponettijakaumaa siten, että kestoaika on ainakin $c > 0$:

$$f(x) = \mathrm{e}^{c-x} \qquad \text{kun } x \geq c.$$

Olkoon $x_1, \ldots, x_n$ otos tästä jakaumasta.

(a) Osoita, että pienin havainto $x_{(1)}$ on c:n suurimman uskottavuuden estimaatti. Määritä c:n normitettu uskottavuusfunktio.

(b) Määritä c:n $100p$ %:n uskottavuusvälin lauseke.

15. Olkoon $x_1, \ldots, x_n$ otos tasajakaumasta $\mathrm{Tas}(\theta, 2\theta)$. Määritä θ:n normitettu uskottavuusfunktio.

16. Tehdään riippumattomat mittaukset $x_1, \ldots, x_n$ aikayksikön välein. Mittaukset $i = 1, 2, \ldots, \theta$ noudattavat normaalijakaumaa $\mathrm{N}(0, 1)$. Jakauman keskiarvossa sattuu yhden yksikön siirtymä hetken θ jälkeen. Siksi mittaukset $i = \theta + 1, \theta + 2, \ldots, n$ noudattavat normaalijakaumaa $\mathrm{N}(1, 1)$.

(a) Osoita, että θ:n uskottavuusfunktion ydin on

$$L(\theta) = \exp\Big[\sum_{i=1}^{\theta}\big(x_i - \tfrac{1}{2}\big)\Big].$$

(b) Saatiin seuraavat 20 peräkkäistä havintoa:

$$\begin{array}{rrrrrrr} -1.26, & -0.16, & -0.64, & 0.56, & -1.82, & -0.76, & -2.08, \\ -0.58, & 0.14, & 0.94, & -0.58, & 0.78, & 1.80, & 0.58, \\ 0.02, & 0.86, & 2.30, & 1.80, & 0.84, & -0.18. & \end{array}$$

Hahmottele logaritmisen normitetun uskottavuusfunktion $r(\theta)$ kuvaaja. Minkä θ:n arvojen suhteellinen uskottavuus on yli 10 %?

17. Komponentin elinaika X noudattaan eksponenttijakaumaa $\mathrm{Exp}(\theta)$. Olkoon γ elinajan mediaani.

(a) Osoita, että $\gamma = \theta \log 2$.

(b) Määritä γ:n 10 %:n uskottavuusväli.

18. Määritä tehtävien 3.2 ja 3.3 tapauksissa likimääräiset 10 %:n uskottavuusvälit ja tutki näissä tapauksissa tämän normaalijakaumaan perustuvan likiarvon tarkkuutta funktion $r(\theta)$ avulla.

19. Olkoon $x_1, \ldots, x_n$ otos Poissonin jakumasta Poi(μ). Tarkastellaan parametrimuunnosta $\mu = \lambda^a$, missä $a \neq 0$.

(a) Määritä logaritmoitu uskottavuusfunktio $l(\lambda)$. Osoita, että $l(\lambda)$:n Taylorin sarjakehitelmässä ($\hat{\lambda}$:n ympäristössä) 3. asteen termi on 0, kun $a = 3$.

(b) Määritä parametrin $\lambda = \mu^{1/3}$:n likimääräinen 10 %:n uskottavuusväli ja muunna se μ:n uskottavuusväliksi.

20. Kasvitaudin leviämistä tutkivassa kokeessa on istutettu kuusia jonoon. Lasketaan kahden peräkkäisten sairastuneiden kuusten väliin jäävien terveiden kuusten lukumäärät X. Saatiin seuraava aineisto:

Terveiden puiden lkm	0	1	2	3	Ainakin 4	Yhteensä
Havaittu frekvenssi	50	23	14	8	5	100

Jos tauti ei ole tarttuva, terveiden lukumäärän X pitäisi noudattaa geometrista jakaumaa

$$f(x) = \theta^x(1-\theta); \qquad x = 0, 1, 2, \ldots,$$

missä $0 < \theta < 1$.

(a) Määritä θ:n 10 %:n uskottavuusväli, jos malli oletetaan oikeaksi.

(b) Estimoi mallin avulla odotetut frekvenssit. Sopiiko malli hyvin aineistoon?

21. Hernekasvit luokitellaan niiden tuottamien herneiden muodon (pyöreä tai särmikäs) ja värin (vihreä tai keltainen) mukaan. Geneettisen teorian mukaan eri tyyppien PV, PK, SV ja SK todennäköisyydet ovat $\alpha\beta$, $\alpha(1-\beta)$, $(1-\alpha)\beta$ ja $(1-\alpha)(1-\beta)$ ja eri kasvit ovat riippumattomia toisistaan. Seuraavassa on eri tyyppien jakauma 500 kasvi otoksessa:

Hernetyyppi	PV	PK	SV	SK
Havaittu frekvenssi	276	104	94	26

Laske parametrien suurimman uskottavuuden estimaatit. Estimoi odotet frekvenssit, kun malli oletetaan oikeaksi.

22. Olkoon $x_1, \ldots, x_n$ otos normaalijakumasta N(μ, σ^2), missä μ ja σ^2 ovat tuntemattomia parametreja.

(a) Osoita, että $\hat{\mu} = \overline{x}$ ja $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$.

(b) Osoita, että

$$\sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2 = \sum x_i^2 - \Big(\sum x_i\Big)^2/n.$$

23. Oletetaan, että riippumattomilla satunnaismuuttujilla $X_1, \ldots, X_n$ on sama odotusarvo, eri varianssit:

$$X_i \sim \mathrm{N}(\mu, \sigma^2/a_i) \qquad \text{kun } i = 1, 2, \ldots, n,$$

missä $a_1, a_2, \ldots, a_n$ ovat positiivisia vakioita.

(a) Osoita, että μ:n suurimman uskottavuuden estimaatti $\hat{\mu}(\sigma)$, kun σ annettu, on sama kaikilla σ:n arvoilla.

(b) Johda $\hat{\mu}$:n ja $\hat{\sigma}$:n lausekkeet.

(c) Osoita, että parametrivektorin (μ, σ) pisteessä $(\hat{\mu}, \hat{\sigma})$ laskettu informaatiomatriisi $\mathcal{I}(\hat{\mu}, \hat{\sigma})$ on positiivisesti definiitti.

24. Johda Tehtävässä 21 normitetun uskottavuusfunktion maksimi. Määritä parametrin α 10 %:n uskottavuusväli.

25. Riippumattomat satunnaismuuttujat X ja Y noudattavat Poissonin jakaumaa siten, että $E(X) = \mu$ ja $E(Y) = \lambda\mu$.

(a) Johda normitetun uskottavuusfunktion $R(\lambda)$ maksimi.

(b) Eräässä valtiossa tehtiin 47 murhaa kuolemanrangaistuksen poistumista edellisenä vuonna. Kuolemanrangaistuksen poistumisen jälkeisenä vuonna tehtiin 57 murhaa. Määritä λ:n 10 %:n uskottavuusväli, jos havainnot riippumattomien Poissonin satunnaismuuttujien arvoja ja muuttujien odotusarvot ovat μ ja $\lambda\mu$. Onko uskottavaa, että murhatiheys ei ole muuttunut?

Luku 8

Frekvenssitulkinta uskottavuuspäättelyssä

Tilastollisia menetelmiä arviodaan tavallisesti siten, että tarkastellaan menetelmän käyttäytymistä koetta toistettaessa. Koetta ei tosiasiassa käytännössa useimmiten toisteta, vaan se vain ajatellaan toistettavaksi niissä samoissa olosuhteissa, joiden vallitessa havaittu aineisto saatiin.

8.1 Otantajakaumat

Jos koetta toistetaan, saadaan eri toistoilla todennäköisesti eri havainnot. Uskottavuusfunktio $L(\theta)$ riippuu saadusta havainnosta, jolloin koetta toistettaessa saadaan aina eri uskottavuusfunktio. Jos halutaan eksplisiittisesti korostaa uskottavuusfunktion riippuvuutta havainnoista, merkitään $L(\theta; x)$. Jatkossakin käytämme useimmiten lyhyempää merkintää $L(\theta)$. Koetta toistettaessa saadaan tietysti aina myös erilainen suurimman uskottavuuden estimaatti $\hat{\theta}$ tai vaikkapa erilaiset 5 %:n uskottavuusvälin päätepisteet ja ne ovat tällaisessa asetelmassa siis satunnaismuuttujia. Kun koetta toistetaan annetulla θ:n arvolla, saadaan estimaattorin *otantajakaumia*.

Yksi tärkeimpiä uskottavuusfunktioon liittyvistä tunnusluvuista on *uskottavuustestisuure*

$$D(\theta; x) = -2 \log \frac{L(\theta; x)}{L(\hat{\theta}; x)} = -2[l(\theta; x) - l(\hat{\theta}; x)]$$

tai lyhyemmin

$$D(\theta) = -2\, r(\theta) = -2[l(\theta) - l(\hat{\theta})]. \tag{8.1.1}$$

Testisuure $D(\theta)$ noudattaa normaalijakauman tapauksessa χ^2-jakaumaa. Ja yleisemminkin, D noudattaa asymptoottisesti χ^2-jakaumaa.

Esimerkki 8.1 Valitaan suuresta tuotepopulaatiosta satunnaisesti 10 tuotetta. Tarkoituksena on estimoida viallisten tuotteiden suhteellinen osuus

θ tuotepopulaatiossa. Viallisten lukumäärä X noudattaa binomijakaumaa $\mathrm{Bin}(n, \theta)$. Parametrin θ logaritmoitu uskottavuusfunktio on

$$l(\theta) = x \log(\theta) + (10 - x) \log(1 - \theta), \qquad 0 \leq \theta \leq 1.$$

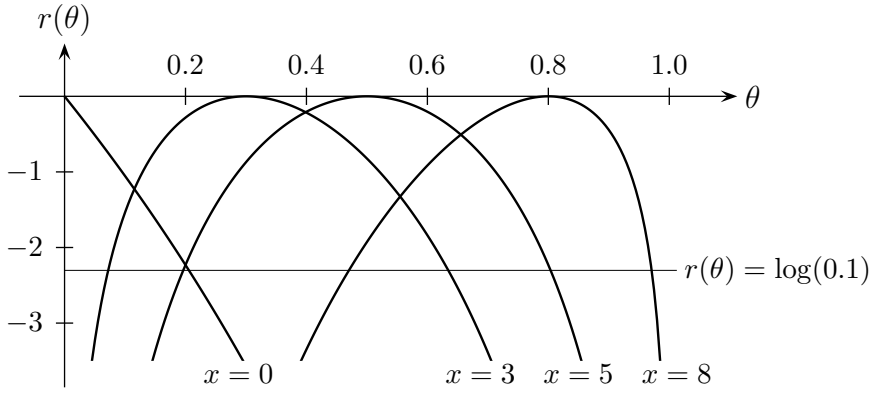
Riippuen x:n arvosta ($x = 0, 1, \ldots, 10$) on mahdollista saada 11 erilaista uskottavuusfunktiota. Suurimman uskottavuuden estimaatti $\hat{\theta} = x/10$ ja logaritmoitu normitettu uskottavuusfunktio on $r(\theta) = l(\theta) - l(\hat{\theta})$. Jos esimerkiksi $x = 3$, niin $\hat{\theta} = 0.3$, $l(\hat{\theta}) = -6.109$ ja

$$r(\theta) = 3 \log(\theta) + 7 \log(1 - \theta) + 6.109, \qquad 0 \leq \theta \leq 1.$$

Suurimman uskottavuuden estimaatti $\hat{\theta} = x/10$ on funktion $r(\theta)$ maksimikohta ja eri x:n arvoilla saadaan siis eri maksimikohta. Mahdollisten maksimikohtien ($\hat{\theta}$:n mahdollisten arvojen) lukumäärä on 11 ($x = 0, 1, \ldots, 10$). Näiden arvojen todennäköisyydet saadaan binomijakaumasta $\mathrm{Bin}(10, \theta)$ ja nämä todennäköisyydet riippuvat θ:n arvosta. Jokaista annettua $\theta = \theta_0$ kohti saadaan määritettyä $\hat{\theta}$:n jakauma. Vastaavalla tavalla voidaan tarkastella esimerkiksi suureen

$$D(\theta) = -2 \log(\theta) = -2[l(\theta) - l(\hat{\theta})].$$

jakaumaa. Jokaista annettua $\theta = \theta_0$ kohti $D(\theta_0)$:llä on 11 mahdollista arvoa, jotka vastaavat X:n mahdollisia arvoja. Näiden arvojen todennäköisyydet saadaan jälleen X:n jakaumasta $\mathrm{Bin}(10, \theta_0)$. Näin voimme määrittää D:n otantajakauman.



Kuvio 8.1. Binomijakauman logaritmoitu normitettu uskottavuusfunktio $r(\theta) = l(\theta) - l(\hat{\theta})$, kun $n = 10$ ja $x = 0, 3, 5$ ja 8

Uskottavuusväli $\mathrm{uv}(x)$ on pelkästään havainnon x funktio. Se voidaan määrittää heti, kun havaintoarvo x on annettu. Koska havainnon x todennäköisyys riippuu parametrin θ arvosta, voidaan välin $\mathrm{uv}(x)$ todennäköisyys laskea, kun θ:lle on annettu jokin arvo θ_0. □

Esimerkki 8.2 Oletetaan, että $X \sim \mathrm{N}(\theta, 1)$. Tästä jakaumasta on saatu yksi havainto $X = x$, jolloin θ:n logaritmoitu uskottavuusfunktio on

$$l(\theta) = -\tfrac{1}{2}(x-\theta)^2, \qquad -\infty < \theta < \infty.$$

Silloin $\hat{\theta} = x$ ja $l(\hat{\theta}) = 0$. Nyt siis logaritmoitu normitettu uskottavuusfunktio on

$$r(\theta) = -\tfrac{1}{2}(x-\theta)^2, \qquad -\infty < \theta < \infty.$$

Ratkaisemalla epäyhtälö $r(\theta) \geq \log p$ saadaan θ:n $100p$ %:n uskottavuusväli

$$x - \sqrt{-2\log p} \leq \theta \leq x + \sqrt{-2\log p}.$$

Jos koetta toistetaan, kun $\theta = \theta_0$, niin $X \sim \mathrm{N}(\theta_0, 1)$. Silloin vaihtelevat myös $\hat{\theta}$:n arvot, eli se on satunnaismuuttuja. Edellä jo totesimme, että $\hat{\theta} = X$, joten

$$\hat{\theta} \sim \mathrm{N}(\theta_0, 1).$$

Vastaavasti uskottavuusvälin päätepisteet $R_a = X - \sqrt{-2\log p}$ ja $R_y = X + \sqrt{-2\log p}$ ovat satunnaismuuttujia ja

$$R_a \sim \mathrm{N}\big(\theta_0 - \sqrt{-2\log p},\, 1\big); \qquad R_y \sim \mathrm{N}\big(\theta_0 + \sqrt{-2\log p},\, 1\big).$$

Uskottavuustestisuure

$$D = -2\, r(\theta_0) = (X - \theta_0)^2,$$

on tässä satunnaismuuttuja. Koska

$$Z = X - \theta_0 \sim \mathrm{N}(0, 1),$$

niin

$$D = Z^2 \sim \chi^2(1).$$

Uskottavuussuure noudattaa siis χ^2-jakaumaa vapausastein 1. □

Esimerkki 8.3 Oletetaan, että havainnot $x_1, x_2, \ldots, x_n$ ovat otos normaalijakaumasta $\mathrm{N}(\mu, \sigma_0^2)$, missä σ_0^2 oletetaan tunnetuksi. Odotusarvon suurimman uskottavuuden estimaatti on $\hat{\mu} = \overline{x}$ ja logaritmoitu normitettu uskottavuusfunktio on

$$r(\mu) = l(\mu) - l(\hat{\mu}) = -\frac{n}{2\sigma_0^2}(\overline{x} - \mu)^2, \qquad -\infty < \mu < \infty.$$

Silloin parametrin μ $100p$ %:n uskottavuusväli on

$$\mathrm{uv}(\overline{x}) = \Big\{ \mu \;\Big|\; -\frac{n}{2\sigma_0^2}(\overline{x} - \mu)^2 \geq \log p \Big\}.$$

Tästä saadaan uskottavuusväliksi

$$\mathrm{uv}(\overline{x}) = \left[\overline{x} - \frac{c\sigma_0}{\sqrt{n}},\, \overline{x} + \frac{c\sigma_0}{\sqrt{n}}\right],$$

missä $c = \sqrt{-2 \log p}$.

Koetta toisttaessa annetulla $\mu = \mu_0$ estimaattori $\hat{\mu} = \overline{x}$ on satunnaismuuttuja

$$\hat{\mu} \sim \mathrm{N}\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

Silloin uskottavuusväli on satunnaisväli

$$\mathrm{uv}(\overline{X}) = \left[\overline{X} - \frac{c\sigma_0}{\sqrt{n}}, \overline{X} + \frac{c\sigma_0}{\sqrt{n}}\right].$$

Uskottavuustestisuure on

$$D = -2\, r(\mu_0) = \frac{n}{\sigma_0^2}(\overline{X} - \mu_0)^2,$$

joka noudattaa χ^2-jakaumaa vapaustein 1. □

8.2 Peitetodennäköisyys

Ajatellaan, että koetta toistetaan annetulla parametrin arvolla $\theta = \theta_0$ aivan kuten alaluvussa 8.1. Määritetään esimerkiksi 10 %:n uskottavuusväli epäyhtälöstä $r(\theta) \geq \log 0.1$. Koetta toistettaessa välin alaraja R_a ja yläraja R_y vaihtelevat. Välin päätepisteet R_a ja R_y ovat satunnaismuuttujia. Niiden otantajakaumat voidaan johtaa havaintojen jakauman avulla ja ne riippuvat yleensä parametrin arvosta θ_0. Väli $[R_a, R_y]$ myös vaihtelee toistosta toiseen ja väli joko sisältää parametrin arvo θ_0 tai ei sisällä. On toivottavaa, että väli sisältää θ_0:n, eli peittää θ_0:n, mahdollisimman usein.

Satunnaisvälin $[R_a, R_y]$ *peitetodennäköisyys* on todennäköisyys, että väli $[R_a, R_y]$ sisältää parametrin arvon θ_0:

$$\mathrm{pt}(\theta_0) = P(R_a \leq \theta_0 \leq R_y \mid \theta = \theta_0). \tag{8.2.1}$$

Huomaa, että (8.2.1) ei ole ehdollinen todennäköisyys. Merkintä korostaa vain sitä, että peitetodennäköisyyttä laskettaessa käytetään parametrin arvoa θ_0. Peitetodennäköisyys riippuu siis tavallisesti tuntemattoman parametrin θ arvosta.

Esimerkki 8.4 Ajatellaan, että toistetaan esimerkin 8.1 koetta parametrin arvolla θ_0. Lasketaan jokaisella toistolla 10 %:n uskottavuusväli. Kaikki mahdolliset uskottavuusvälit saadaan määritelmän mukaan

$$[a(x), y(x)] = \mathrm{uv}(x) = \{\, \theta \mid r(\theta; x) \geq \log 0.1 \,\}, \tag{8.2.2}$$

missä $x = 0, 1, \ldots, 10$. Haluamme nyt tietää, kuinka usein $\theta = \theta_0$ kuuluu 10 %:n uskottavuusvälille. Olkoon $\theta = 0.1$. Määritämme kaikki 11 mahdollista 10 %:n uskottavuusväliä. Voimme todeta, että $R_a \leq 0.1 \leq R_y$, kun $0 \leq X \leq 3$. Siksi 10 %:n uskottavuusvälin peitetodennäköisyys

$$\mathrm{pt}(0.1) = P(R_a \leq 0.1 \leq R_y \mid \theta = 0.1) = P(0 \leq X \leq 3 \mid \theta = 0.1),$$

kun koetta toistetaan parametrin arvolla $\theta = 0.1$. Binomijakaumasta $\mathrm{Bin}(10, 0.1)$ saadaan peitetodennäköisyys

$$\mathrm{pt}(0.1) = 0.349 + 0.387 + 0.194 + 0.057 = 0.987.$$

Ennen koetta ei tunneta X:n arvoa, joten myös uskottavuusväli on satunnaisväli

$$[R_a, R_y] = \mathrm{uv}(X) = \{\, \theta \mid r(\theta; X) \geq \log 0.1 \,\}, \tag{8.2.3}$$

jonka mahdolliset arvot ovat (8.2.2):n määrittämät 11 väliä. Esimerkiksi todennäköisyys

$$P[\mathrm{uv}(1)] = P(X = 1 \mid \theta) = 10\theta(1 - \theta)^9,$$

riippuu tuntemattomasta θ:sta. Kun $\theta = 0.1$, niin $P[\mathrm{uv}(1)] = 0.387$.

Koska binomijakauma on diskreetti, peitetodennäköisyys $\mathrm{pt}(\theta)$ on θ:n epäjatkuva funktio. Tarkastellaan nyt satunnaisväliä (8.2.3). Silloin

$$\begin{aligned}\mathrm{pt}(\theta_0) &= P[\theta_0 \in \mathrm{uv}(X) \mid \theta = \theta_0] \\ &= \sum_{x:\, \theta_0 \in \mathrm{uv}(x)} P(X = x \mid \theta = \theta_0) = \sum_{x:\, \theta_0 \in \mathrm{uv}(x)} \binom{10}{x} \theta_0^x (1 - \theta_0)^{10-x}.\end{aligned}$$

Jos esimerkiksi $\theta = 0.206$, niin $0.206 \in \mathrm{uv}(x)$, kun $x = 0, 1, 2, 3, 4$ ja 5. Silloin saadaan peitetodennäköisyydeksi

$$pt(0.206) = \sum_{x=0}^{5} \binom{10}{x} 0.1^x \cdot 0.9^{10-x}.$$

□

Esimerkki 8.5 Esimerkissä 8.2 θ:n $100p$ %:n uskottavuusväli on muotoa

$$X - c \leq \theta \leq X + c,$$

missä $c = \sqrt{-2 \log p}$. Määritelmän (8.2.1) mukaan tämän välin peitetodennäköisyys parametrin arvolla $\theta = \theta_0$ on

$$\begin{aligned}\mathrm{pt}(\theta_0) &= P(X - c \leq \theta_0 \leq X + c \mid \theta = \theta_0) \\ &= P(-c \leq X - \theta_0 \leq c \mid \theta = \theta_0).\end{aligned}$$

Koska oletuksen mukaan $X \sim \mathrm{N}(\theta_0, 1)$, niin

$$Z = X - \theta_0 \sim \mathrm{N}(0, 1).$$

Tästä seuraa, että

$$\mathrm{pt}(\theta_0) = P(-c \leq Z \leq c).$$

Jos esimerkiksi $p = 0.1$, niin $c = \sqrt{-2\log 0.1} = 2.146$. Silloin 10 %:n uskottavuusvälin peitetodennäköisyys on

$$\text{pt} = P(-2.146 \leq Z \leq 2.146) = 0.968.$$

Vastaavasti 10 %:n uskottavuusväli on $X \pm 2.146$. Kun koetta toistetaan kiinnitetyllä parametrin arvolla $\theta = \theta_0$, niin 96.8 %:a uskottavuusväleistä peittää parametrin todellisen (kiinnitetyn) arvon. Toisaalta tiedämme esimerkiksi, että

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Valitsemalla siis $c = 1.96$ saadaan peitetodennäköisyys 0.95. Ratkaisemalla p yhtälöstä $\sqrt{-2\log p} = 1.96$ saadaan

$$p = \exp\left(\frac{1.96^2}{2}\right) = 0.147.$$

Havaitsemme, että $X \pm 1.96$ on 14.7 %:n uskottavuusväli. □

8.2.1 Peitetodennäköisyyden arviointi uskottavuustestisuureen avulla

Peitetodennäköisyydet voidaan määrittää uskottavuustestisuureen D otantajakauman avulla. Parametrin θ $100p$ %:n uskottavuusväli on epäyhtälön $r(\theta) \geq \log p$, toteuttavien parametrin arvojen joukko. Parametrin arvo $\theta = \theta_0$ kuuluu uskottavuusvälille, jos ja vain jos $r(\theta_0) \geq \log p$. Koska $D = -2\,r(\theta_0)$, niin θ_0 on $100p$ %:n uskottavuusvälillä jos ja vain jos $D \leq -2\log p$. Siksi $100p$ %:n uskottavuusvälin peitetodennäköisyys on

$$\text{pt}(\theta_0) = P(D \leq -2\log p \mid \theta = \theta_0). \tag{8.2.4}$$

Esimerkki 8.6 (jatkoa Esimerkkiin 8.2) Esimerkin 8.2 uskottavuussuureen D otantajakuma on $\chi^2(1)$ parametrin arvosta riippumatta. Siksi $100p$ %:n uskottavuusvälin peitetodennäköisyys on

$$\text{pt} = P(D \leq -2\log p).$$

Koska $Z^2 \sim \chi^2(1)$, niin jokaista $d > 0$ kohti

$$P\big(\chi^2(1) \leq d\big) = P\big(Z^2 \leq d\big) = P\big(-\sqrt{d} \leq Z \leq \sqrt{d}\big),$$

missä $Z \sim \text{N}(0,1)$. Tästä seuraa, että

$$\text{pt} = P\big(-\sqrt{-2\log p} \leq Z \leq \sqrt{-2\log p}\big).$$

Sama tulos saatiin Esimerkissä 8.2 uskottavuusvälien avulla. □

8.3 χ^2-likiarvo

8.3.1 Uskottavuustestisuureen jakauma

Alaluvussa 8.2 osoitimme, että peitetodennäöisyydet voidaan määrittää uskottavuussuureen $D = -2\,r(\theta)$ avulla. Yleensä D:n jakauma riippuu parametrin $\theta = \theta_0$ arvosta. Normaalijakuman tapauksessa totesimme, että $D \sim \chi^2(1)$ kaikilla θ:n arvoilla. Usein D:n tarkka otantajakauma on hankala johtaa. Onneksi tarkan jakauman hyvä likiarvo on useimmiten helppo määrittää.

Osoittautuu, että D:n jakauma on usein varsin lähellä jakaumaa $\chi^2(1)$:

$$D = -2\,r(\theta_0) \approx \chi^2(1). \tag{8.3.1}$$

Normaalijakaumaa koskeva tarkka tulos on siis monissa tilanteissa hyvä likiarvo. Kun meillä on otos $X_1, X_2, \ldots, X_n$ jostain jakaumasta (n identtisesti jakautunutta riippumatonta satunnaismuuttujaa), voidaan tulos (8.3.1) perustella keskeisen rajaväittämän avulla. Voidaan näyttää, että melko lievien oletusten vallitessa

$$\lim_{n\to\infty} P(D \le d \mid \theta = \theta_0) = P\big[\chi^2(1) \le d\big], \tag{8.3.2}$$

kaikilla $d > 0$.

Tärkein tuloksen (8.3.2) todistamiseksi vaadittava sännöllisyysehto on se, että havaintojen X_i, $i = 1, 2, \ldots, n$ vaihtelualue ei saa riippua parametrista θ. Parametrin arvon θ_0 tulee olla lisäksi parametriavaruuden sisäpiste. Jos θ_0 on lähellä reunaa, saatetaan tarvita varsin suuri havaintojen määrä n, jotta D:n jakauma on riittävän lähellä χ^2-jakaumaa. Uskottavuussuureen rajakauma on $\chi^2(1)$ kaikissa parametriavaruuden sisäpisteissä θ_0, mutta annetun tarkkuuden saavuttamiseksi tarvittava otoskoko voi riippua θ_0:sta.

Esimerkki 8.7 Esimerkissä 8.1 valittiin 10 tuotetta satunnaisesti suuresta tuotepopulaatiosta. Otoksesta estimoitiin viallisten tuotteiden suhteellinen osuus θ. Kun parametrille annetaan jokin arvo, esimerkiksi $\theta = 0.1$, saadaan D:n tarkka otantajakauma binomijakauman $\mathrm{Bin}(10, 0.1)$ avulla. $\chi^2(1)$-jakaumasta saadaan todennäköisyys $P\big(\chi^2(1) \le 2.707\big) = 0.90$. Lasketaan sitten tarkka todennäköisyys $P(D \le 2.707)$. Koska $D \le 2.707$ sattuu täsmälleen silloin, kun $X \le 2$, niin

$$P(D \le 2.707) = P(X \le 2) = 0.349 + 0.387 + 0.194 = 0.930.$$

Vastaavasti $\chi^2(1)$-jakauman 95 %:n ja 99 %:n pisteet ovat 3.841 ja 6.635. D:n tarkan jakauman mukaan vastaavat todennäköisyydet ovat

$$P(D \le 3.841) = P(X \le 3) = 0.987 \qquad \text{ja}$$
$$P(D \le 6.635) = P(X \le 4) = 0.998.$$

□

Esimerkki 8.8 Olkoon $X_1, X_2, \ldots, X_n$ otos Poissonin jakaumasta, jonka keskiarvo on μ. Silloin μ:n logaritmoitu uskottavuusfunktio on

$$l(\mu) = s \log \mu - n\mu,$$

missä $\mu > 0$ ja $s = \sum x_i$. Parametrin μ suurimman uskottavuuden estimaatti on $\hat{\mu} = \sum x_i/n$ ja logaritmoitu normitettu uskottavuusfunktio on

$$\begin{aligned} r(\mu) = l(\mu) - l(\hat{\mu}) &= s \log \frac{\mu}{\hat{\mu}} - n\mu + n\hat{\mu} \\ &= -s \log\left(\frac{s}{n\mu}\right) - n\mu + s. \end{aligned}$$

Ajatellaan, että koetta toistetaan, kun $\mu = \mu_0$. Silloin havaintojen summa $S = \sum X_i$ vaihtelee otoksesta toiseen ja se on siis satunnaismuuttuja. Poissonin jakauman yhteenlaskuominaisuuden mukaan $S \sim \text{Poi}(n\mu_0)$.

Uskottavuustestisuure on

$$D = -2\, r(\mu_0) = 2\left[S \log \frac{S}{n\mu_0} + n\mu_0 - S\right].$$

D on diskreetti satunnaismuuttuja, jonka arvot määräytyvät satunnaismuuttujan S arvojen $S = 0, 1, 2, \ldots$ perusteella. Arvojen todennäköisyydet saadaan Poissonin todennäköisyyksien

$$P(S = s) = \frac{(n\mu_0)^s \mathrm{e}^{-s}}{s!}, \qquad s = 0, 1, 2, \ldots$$

avulla. Kun esimerkiksi $\mu_0 = 0.5$ ja $n = 20$, niin $n\mu_0 = 10$ ja

$$D = 2\left[S \log \frac{S}{10} + 10 - S\right].$$

Termin $S \log \frac{S}{10}$ arvoksi annetaan 0, kun $S = 0$. Jos $D \sim \chi^2(1)$, niin $P(D \leq 2.706) = 0.9$. Tarkka todennäköisyys $P(D \leq 2.706)$ saadaan Poissonin jakauman avulla. Laskemalla todetaan, että $D \leq 2.706$, jos ja vain jos $6 \leq S \leq 15$. Koska $S \sim \text{Poi}(10)$, niin

$$P(D \leq 2.706) = \sum_{s=6}^{15} 10^s \mathrm{e}^{-10}/s! = 0.884.$$

□

Esimerkki 8.9 Erään elektronisen komponentin kestoaika X noudattaa eksponenttijakaumaa, jonko odotusarvo on $E(X) = \theta$. Kun testattiin n:n komponentin kesto, saatiin n:n kestoajan muodostama aineisto $x_1, x_2, \ldots, x_n$. Parametrin θ suurimman uskottavuuden estimaatti $\hat{\theta} = s/n$, missä $s = \sum x_i$.

Eksponenttijakauman logaritmoitu uskottavuusfunktio on $l(\theta) = -n \log \theta - s/\theta$, joten logaritmoitu normitettu uskottavuusfunktio on

$$r(\theta) = -n\left(\frac{\hat{\theta}}{\theta} - 1 - \log \frac{\hat{\theta}}{\theta}\right) = -n\left(\frac{s}{n\theta} - 1 - \log \frac{s}{n\theta}\right).$$

Ajatellaan nyt, että koetta toistetaan annetulla parametrin arvolla $\theta = \theta_0$. Kokonaiskestoaika $S = \sum X_i$ on jatkuva satunnaismuuttuja. Siksi myös uskottavuussuure

$$D = -2\, r(\theta_0) = 2n\left(\frac{S}{n\theta_0} - 1 - \log \frac{S}{n\theta_0}\right)$$

on jatkuva satunnaismuuttuja. Huomaa, että

$$D = 2n(Y - 1 - \log Y),$$

missä $Y = S/(n\theta_0)$. Siksi

$$P(D \le d) = P\left(Y - 1 - \log Y \le \frac{d}{2n}\right).$$

Tarkastellaan funktiota

$$g(y) = y - 1 - \log y, \qquad y > 0.$$

Koska $g'(y) = 1 - 1/y$ ja $g''(y) = 1/y^2$, niin funktiolla $g(y)$ on pisteessä $y = 1$ yksikäsitteinen minimi. Siksi jokaista $d > 0$ kohti on olemassa sellaiset ehdon $y_1 < 1 < y_2$ toteuttavat y:n arvot, että $g(y) \le d/(2n)$, jos ja vain jos $y_1 \le y \le y_2$. Näiden arvojen löytämiseksi ratkaistaan yhtälö $g(y) - d/2n = 0$. Silloin saadaan tulos

$$\begin{aligned} P(D \le d) &= P(y_1 \le Y \le y_2) = P\left(y_1 \le \frac{S}{n\theta_0} \le y_2\right) \\ &= P\left(2ny_1 \le \frac{2S}{\theta_0} \le 2ny_2\right). \end{aligned}$$

Satunnaismuuttuja $2S/\theta_0$ noudattaa χ^2-jakaumaa vapausastein $2n$, joten todennäköisyys $P(D \le d)$ voidaan laskea χ^2-jakauman avulla. □

8.3.2 Pistefunktion jakauma

Seuraavassa perustelemme, miksi tulos (8.3.2) pitää paikkansa. Tämä perustelu ei ole tarkka muodollinen todistus. Olkoon $X_1, X_2, \ldots, X_n$ otos jostain jakaumasta eli X_1, X_2, …, X_n ovat riippumattomat ja noudattavat samaa jakaumaa. Alaluvussa 7.5 tarkasteltiin riippumattomien otosten yhdistämistä. Koska X_1, …, X_n ovat riippumattomat, niin havaintojen $X_1 = x_1$, …, $X_n = x_n$ perusteella määritetty uskottavuusfunktio on muotoa

$$L(\theta; x_1, x_2, \ldots, x_n) = L(\theta; x_1)\, L(\theta; x_2) \cdots L(\theta; x_n).$$

Vastaavasti logaritmoitu uskottavuusfunktio on

$$l(\theta; x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} l(\theta; x_i),$$

missä $l(\theta; x_i) = \log L(\theta; x_i)$, $1 \leq i \leq n$. Suoraan pistefunktion määritelmästä seuraa, että

$$u(\theta; x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} u(\theta; x_i), \tag{8.3.3}$$

missä $u(\theta; x_i) = l'(\theta; x_i)$.

Kun pistefunktion $u(\theta; x_1, x_2, \ldots, x_n)$ esityksessä (8.3.3) annetaan parametrille jokin arvo $\theta = \theta_0$ ja tarkastellaan havaintoja satunnaismuuttujina, saadaan satunnaismuuttuja

$$u(\theta_0; X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} u(\theta_0; X_i). \tag{8.3.4}$$

Myös satunnaismuuttujaa (8.3.4) kutsutaan pistefunktioksi. Jatkossa pistefunktiota (8.3.4) merkitään usein lyhyesti $u(\theta_0; X_1, X_2, \ldots, X_n) = U(\theta_0)$.

Vastaavasti voidaan myös informaatiofunktiota tarkastella satunnaismuuttujana. Silloin saamme

$$\mathcal{I}(\theta_0; X_1, X_2, \ldots, X_n) = -\sum_{j=1}^{n} u'(\theta_0; X_j) = \sum_{j=1}^{n} i(\theta_0, X_j), \tag{8.3.5}$$

missä $u'(\theta_0; X_j)$ on pistefunktion $u(\theta; X_j)$ derivaatan arvo pisteessä $\theta = \theta_0$ ja $i(\theta_0, X_j)$ on havaintoon X_j perustuva informaatio. Informaation (8.3.5) odotusarvo

$$I(\theta_0) = E[\mathcal{I}(\theta_0; X_1, X_2, \ldots, X_n)] \tag{8.3.6}$$

on *odotettu informaatio*. Osoitamme alaluvussa 8.6, että $U(\theta_0)$:n odotusarvo on 0 ja varianssi on odotettu informaatio $I(\theta_0)$ Keskeisen rajaväittämän mukaan

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \approx \mathrm{N}(0, 1),$$

kun n on riittävän suuri.

Esimerkki 8.10 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\theta, 1)$. Silloin

$$l(\theta) = -\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2$$

ja $l'(\theta) = \sum_{i=1}^{n}(X_i - \theta)$. Kun $\theta = \theta_0$, niin

$$l'(\theta_0) = u(\theta_0; X_1, \ldots, X_n) = \sum_{i=1}^{n}(X_i - \theta_0).$$

Koska $l''(\theta) = -n$, niin $\mathcal{I}(\theta) = I(\theta) = n$. Nyt siis

$$\frac{u(\theta_0; X_1, \ldots, X_n)}{\sqrt{I(\theta_0)}} = \frac{\sum_{i=1}^{n}(X_i - \theta_0)}{\sqrt{n}} = \frac{\overline{X} - \theta_0}{1/\sqrt{n}} \sim \mathrm{N}(0,1).$$

□

Esimerkki 8.11 Olkoon $X_1, \ldots, X_n$ otos Poissonin jakaumasta Poi(μ). Silloin

$$l(\mu) = \Big(\sum_{i=1}^{n} X_i\Big)\log\mu - n\mu, \quad l'(\mu) = \frac{\sum X_i}{\mu} - n \quad \text{ja} \quad l''(\mu) = -\frac{\sum X_i}{\mu^2},$$

joten $\mathcal{I}(\mu) = \sum X_i/\mu^2$. Odotettu informaatio on siis

$$I(\mu) = E\left(\frac{\sum X_i}{\mu^2}\right) = \frac{n\mu}{\mu^2} = \frac{n}{\mu}.$$

Kun oletetaan $\mu = \mu_0$, niin pistefunktio on

$$U(\mu_0; X_1, \ldots, X_n) = \frac{\sum X_i - n\mu_0}{\mu} = \frac{n}{\mu}(\overline{X} - \mu_0).$$

Tästä seuraa, että

$$\frac{U(\mu_0)}{\sqrt{I(\mu_0)}} = \frac{(n/\mu)(\overline{X} - \mu_0)}{\sqrt{n/\mu_0}} = \frac{\overline{X} - \mu_0}{\sqrt{\mu_0/n}}.$$

Koska $\mathrm{Var}(\overline{X}) = \mu_0/n$, niin keskeisen rajaväittämän mukaan

$$\frac{\overline{X} - \mu_0}{\sqrt{\mu_0/n}} \xrightarrow{d} \mathrm{N}(0,1), \qquad \text{kun } n \to \infty.$$

□

Esimerkki 8.12 Kun $X_1, \ldots, X_n$ on otos eksponenttijakaumasta Exp(θ), niin

$$l(\theta) = -n\log\theta - \frac{\sum X_i}{\theta}$$

ja

$$l'(\theta) = -\frac{n}{\theta} + \frac{\sum X_i}{\theta^2} = \frac{1}{\theta^2}\Big(\sum X_i - n\theta\Big) = \frac{n}{\theta^2}(\overline{X} - \theta).$$

Koska $l''(\theta) = n/\theta^2 - \big(2\sum X_i\big)/\theta^3$, niin

$$I(\theta) = E\left(\frac{2\sum X_i}{\theta^3} - \frac{n}{\theta^2}\right) = \frac{2n\theta}{\theta^3} - \frac{n}{\theta^2} = \frac{n}{\theta^2}.$$

Tästä seuraa, että

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} = \frac{(n/\theta_0^2)(\overline{X} - \theta_0)}{\sqrt{n/\theta_0^2}} = \frac{\overline{X} - \theta_0}{\sqrt{\theta_0^2/n}},$$

joka keskeisen rajaväittämän mukaan lähenee normaalijakaumaa $\mathrm{N}(0,1)$, kun n kasvaa. □

Suurimman uskottavuuden estimaatti on yhtälön $U(\hat{\theta}) = 0$ ratkaisu. Edellä esitetyn tuloksen perusteella voidaan osoittaa, että $\hat{\theta} \to \theta_0$ todennäköisyyden mielessä, kun $n \to \infty$. Voidaan lisäksi osoittaa, että $\mathcal{I}(\hat{\theta})/I(\theta_0) \to 1$ todennäköisyyden mielessä, kun $n \to \infty$. Siksi

$$\frac{U(\theta_0)}{\sqrt{\mathcal{I}(\hat{\theta})}} \approx \mathrm{N}(0,1), \tag{8.3.7}$$

kun n on riittävän suuri.

Koska $\hat{\theta}$ on lähellä arvoa θ_0, kun n on suuri, niin likiarvon (7.9.2) mukaan

$$r(\theta_0) \approx -\tfrac{1}{2}(\hat{\theta} - \theta_0)^2\, \mathcal{I}(\hat{\theta}).$$

Derivoimalla θ_0:n suhteen (θ_0 tässä muuttuja) saadaan

$$U(\theta_0) \approx (\hat{\theta} - \theta_0)\, \mathcal{I}(\hat{\theta}).$$

Silloin tuloksen (8.3.7) mukaan

$$(\hat{\theta} - \theta_0)\sqrt{\mathcal{I}(\hat{\theta})} \approx \mathrm{N}(0,1). \tag{8.3.8}$$

Tästä saadaan uskottavuussuureen D likiarvo

$$D = -2\, r(\theta_0) \approx (\hat{\theta} - \theta_0)^2\, \mathcal{I}(\hat{\theta}).$$

Koska (8.3.8):n mukaan $(\hat{\theta} - \theta_0)\sqrt{\mathcal{I}(\hat{\theta})}$ noudattaa likimain normaalijakaumaa $\mathrm{N}(0,1)$, niin D noudattaa likimain χ^2-jakaumaa vapausastein 1, kun n on suuri.

8.4 Luottamusvälit

Satunnaisväliä $[R_a, R_y]$ kutsutaan parametrin θ *luottamusväliksi* (LV), jos sen peitetodennäköisyys

$$\mathrm{pt}(\theta_0) = P(R_a \le \theta_0 \le R_y \mid \theta = \theta_0)$$

ei riipu parametrin arvosta θ_0. Luottamusvälin peitetodennäköisyyttä sanotaan *luottamuskertoimeksi*. Esimerkiksi väli $[R_a, R_y]$ on θ:n 95 %:n luottamusväli, jos

$$P(R_a \leq \theta_0 \leq R_y \mid \theta = \theta_0) = 0.95$$

kaikilla mahdollisilla θ_0:n arvoilla. Kun koetta toistetaan kiinniteyllä parametrin arvolla $\theta = \theta_0$, niin luottamusväli peittää θ_0:n 95 %:ssa kaikista toistoista.

Esimerkissä 8.5 $100p$ %:n uskottavuusvälin peitetodennäköisyys oli sama parametrin arvosta riippumatta, joten uskottavuusväli on luottamusväli. Sen sijaan Esimerkissä 8.4 $100p$ %:n uskottavuusvälin peitetodennäköisyys riippuu parametrin todellisesta arvosta θ_0, joten siinä tapauksessa uskottavuusväli ei ole luottamusväli. Usein on kuitenkin mahdollista konstruoida likimääräinen luottamusväli siten, että $\mathrm{pt}(\theta_0)$ on likimain vakio tarkasteltavalla parametrin arvoalueella. χ^2-likiarvon ansiosta uskottavuusvälit ovat useimmissa sovelluksissa likimääräisiä luottamusvälejä. Jos tulosta (8.3.1) voidaan soveltaa, niin

$$\mathrm{pt} \approx P\big[\chi^2(1) \leq -2 \log p\big].$$

On väärin ajatella, että esimerkiksi jokin annettu 95 %:n luottamusväli $[a, y]$ peittää todellisen parametrin arvon θ_0 todennäköisyydellä 0.95. Voi esimerkiksi sattua, että $[a, y]$ peittää kaikki mahdolliset parametrin arvot, jolloin väli peittää θ_0:n todennäköisyydellä 1. Siitä huolimatta väli on 95 %:n luottamusväli. Peitetodennäköisyys on teoreettinen tunnusluku, joka liittyy ajateltuun koetoistojen sarjaan. Peitetodennäköisyys ei ole yksittäisen välin ominaisuus, vaan se on välien määrittämiseen käytetyn menetelmän ominaisuus.

Luottamusväli voidaan muodostaa uskottavuusfunktion avulla. Jos esimerkiksi tarvitaan θ:n 95 %:n luottamusväli, määritetään $100p$ %:n uskottavuusväli valitsemalla p siten, että välin peitetodennäköisyys on 0.95. Jos tarvitaan esimerkiksi 95 %:n luottamusväli, määritetään $100p$ %:n uskottavuusväli valitsemalla p siten, että välin peitetodennäköisyys on 0.95.

Esimerkki 8.13 Esimerkissä 8.5

$$Z = X - \theta_0 \sim \mathrm{N}(0, 1).$$

Koska $P(-1.96 \leq Z \leq 1.96) = 0.95$, niin

$$P(X - 1.96 \leq \theta_0 \leq X + 1.96) = P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Välin $[X - 1.96,\ X + 1.96]$ peitetodennäköisyys on 0.95 kaikilla θ_0:n arvoilla ja se on siksi 95 %:n luottamusväli. Se on myös uskottavuusväli. Esimerkissä 8.2 nähtiin, että $p \cdot 100$ %:n uskottavuusväli on muotoa

$$\big[X - \sqrt{-2 \log p},\ X + \sqrt{-2 \log p}\,\big].$$

Kun p määritellään siten, että saadaan 95 % luottamusväli, saadaan $p = 0.147$. Tässä tapauksessa luottamusväli voidaan muodostaa monella muullakin tavalla. Esimerkiksi

$$P(-2.376 \leq Z \leq 1.751) = 0.95,$$

joten välin $[X - 2.376,\ X + 1.751]$ peitetodennäköisyys on 0.95 kaikilla θ_0 ja se on siis luottamusväli. Se ei kuitenkaan ole uskottavuusväli. □

8.4.1 Normaalijakaumaan perustuva likiarvo

On suositeltavaa muodostaa luottamusväli uskottavuusvälin avulla. Kun muodostettaan esimerkiksi 95 %:n luottamusväli, määritetään ensin $100p$ %:n uskottavuusväli siten, että välin peitetodennäköisyys on 0.95. Jos likiarvo (8.3.1) on pätevä, niin silloin uskottavuusväli on likimääräinen (tai tarkka) luottamusväli.

Joskus luottamusväli voidaan muodostaa laskennallisesti helpommin normaalijakaumaan perustuvan likiarvon avulla. Likiarvon (7.9.2) mukaan

$$\left[\hat{\theta} \pm \frac{c}{\sqrt{\mathcal{I}(\hat{\theta})}}\right]$$

on θ:n likimääräinen uskottavuusväli, missä $c = \sqrt{-2\log p}$. Välin peitetodennäköisyys on

$$\begin{aligned}\mathrm{pt}(\theta_0) &= P\left(\hat{\theta} - \frac{c}{\sqrt{\mathcal{I}(\hat{\theta})}} \leq \theta_0 \leq \hat{\theta} + \frac{c}{\sqrt{\mathcal{I}(\hat{\theta})}} \;\middle|\; \theta = \theta_0\right) \\ &= P\left(-c \leq (\hat{\theta} - \theta_0)\sqrt{\mathcal{I}(\hat{\theta})} \leq c \;\middle|\; \theta = \theta_0\right).\end{aligned}$$

Tuloksen (8.3.8) mukaan

$$\mathrm{pt}(\theta_0) \approx P(-c \leq Z \leq c),$$

missä $Z \sim \mathrm{N}(0,1)$. Koska $P(-1.96 \leq Z \leq 1.96) = 0.95$, niin väli

$$\left[\hat{\theta} - \frac{1.96}{\sqrt{\mathcal{I}(\hat{\theta})}},\ \hat{\theta} + \frac{1.96}{\sqrt{\mathcal{I}(\hat{\theta})}}\right] \tag{8.4.1}$$

on θ:n likimääräinen 95 %:n luottamusväli.

Esimerkki 8.14 Oletetaan, että jakaumasta $\mathrm{Bin}(100, \theta)$ on saatu havainto $x = 17$. Lasketaan θ:n likimääräinen 95 %:n luottamusväli. Nyt θ:n suurimman uskottavuuden estimaatti on $\hat{\theta} = x/n = 0.17$ ja

$$r(\theta) = 17\log\theta + 83\log(1-\theta) + 45.581, \qquad 0 < \theta < 1.$$

Laskemalla voidaan todeta, että $r(\theta) \geq \log 0.147$, kun $0.105 \leq \theta \leq 0.251$. Tämä on θ:n 14.7 %:n uskottavuusväli ja likimain 95 %:n luottamusväli.

Informaatiofunktio on

$$\mathcal{I}(\theta) = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}, \qquad 0 < \theta < 1.$$

Sijoittamalla informaatiofunktioon $\theta = \hat{\theta}$ saadaan

$$\mathcal{I}(\hat{\theta}) = \frac{x}{\hat{\theta}^2} + \frac{n-x}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}} + \frac{n}{1-\hat{\theta}} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

Nyt (8.4.1):n mukaan

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.17 \pm 0.0736$$

on θ:n likimääräinen 95 %:n luottamusväli. Väli ei kuitenkaan ole uskottavuusväli, sillä välin alarajan 0.096 suhteellinen uskottavuus $R(0.096) = 0.072$ on paljon pienempi kuin ylärajan 0.244 suhteellinen uskottavuus $R(0.244) = 0.200$. □

Esimerkki 8.15 Oletetaan, että $X \sim \text{Bin}(3, \theta)$. Silloin

$$\begin{aligned} l(\theta) &= x \log\theta + (3-x)\log(1-\theta) \\ &= 3[\hat{\theta}\log\theta + (1-\hat{\theta})\log(1-\theta)], \end{aligned}$$

missä $\hat{\theta} = x/3$ ja $0 \leq \theta \leq 1$. Silloin

$$l(\hat{\theta}) = 3[\hat{\theta}\log\hat{\theta} + (1-\hat{\theta})\log(1-\hat{\theta})],$$

missä $\hat{\theta}$ voi saada arvot 0, $\frac{1}{3}$, $\frac{2}{3}$ ja 1. Parametrin 10 %:n uskottavuusväli on

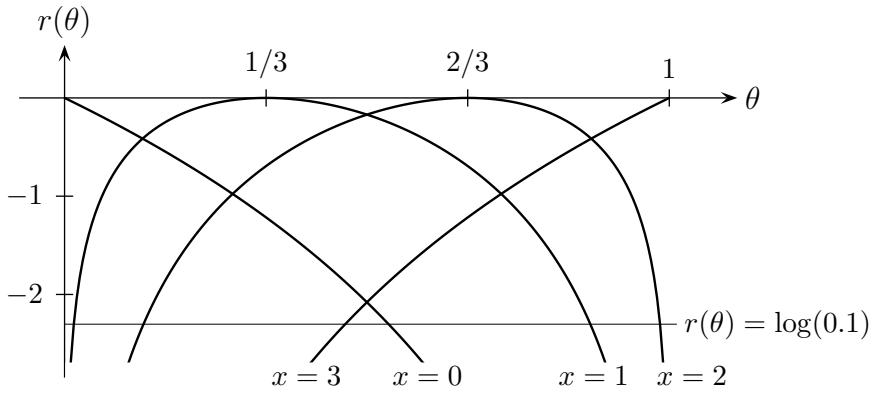
$$\text{uv}(x; 10\,\%) = \{\, \theta \mid l(\theta) - l(\hat{\theta}) \geq \log 0.1 \,\}.$$

Jos esimerkiksi $x = 1$, niin

$$\begin{aligned} \text{uv}(x; 10\,\%) &= \{\, \theta \mid \log\theta + 2\log(1-\theta) \geq -0.39 \,\} \\ &= [0.015, 0.869]. \end{aligned}$$

Eri x:n arvoilla saadaan seuraavat θ:n 10 %:n luottamusvälit

$$\begin{aligned} \text{uv}(0) &= [0, 0.536], \\ \text{uv}(1) &= [0.015, 0.869], \\ \text{uv}(2) &= [0.131, 0.985], \\ \text{uv}(3) &= [0.464, 1]. \end{aligned}$$



Kuvio 8.2. Logaritmoitu normitettu uskottavuusfunktio $r(\theta) = l(\theta) - l(\hat{\theta})$, kun $x = 0, 1, 2$ ja 3.

Todennäköisyys, että väli peittää parametrin todellisen arvon, riippuu nyt parametrin arvosta.

Normaalijakaumaan perustuvan likiarvon avulla johdettu tavanomainen 95 %:n luottamusväli on muotoa

$$l_V(\hat{\theta}) = \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{3}}.$$

Koska $\hat{\theta}$:n mahdolliset arvot ovat 0, $\frac{1}{3}$, $\frac{2}{3}$ ja 1, niin mahdolliset luottamusvälit ovat

$$0, \quad [0.061, 0.605], \quad [0.395, 0.939] \quad \text{ja} \quad 1.$$

Havaintoarvoilla $X = 0$ ja $X = 3$ väli degeneroituu yhdeksi pisteeksi. Kun $0 < \theta < 0.061$ tai $0.939 < \theta < 1$, on luottamusvälin peitetodennäköisyys 0 eli $P[\theta \in l_V(\hat{\theta})] = 0$! □

Esimerkki 8.16 Oletetaan, että eksponenttijakaumasta $\mathrm{Exp}(\theta)$ on saatu 10 riippumatonta havaintoa. Oletetaan, että $\hat{\theta} = 28.8$. Pyrimme nyt määrittämään 95 % luottamusvälin likiarvon.

Logaritminen normitettu uskottavuusfunktio on

$$r(\theta) = -n\left(\frac{\hat{\theta}}{\theta} - 1 - \log\frac{\hat{\theta}}{\theta}\right) = -10\left(\frac{28.8}{\theta} - 1 - \log\frac{28.8}{\theta}\right),$$

kun $\theta > 0$. Määritetään epäyhtälöstä $r(\theta) \geq \log 0.147$ parametrin 14.7 %:n uskottavuusväli, joka on $[16.42, 57.47]$. Tämä on 95 %:n luottamusvälin likiarvo. Luottamusvälin tarkka peitetodennäköisyys on 0.948.

Vaihtoehtoisesti voimme käyttää likiarvoa (8.4.1). Koska

$$\mathcal{I}(\theta) = -\frac{n}{\theta^2} + \frac{2n\hat{\theta}}{\theta^3} \qquad \text{ja} \qquad \mathcal{I}(\hat{\theta}) = \frac{n}{\theta^2},$$

niin $\sqrt{\mathcal{I}(\hat{\theta})} = \sqrt{n}/\hat{\theta} = 0.1098$ ja

$$\left[\hat{\theta} - \frac{1.96}{\sqrt{\mathcal{I}(\hat{\theta})}},\, \hat{\theta} + \frac{1.96}{\sqrt{\mathcal{I}(\hat{\theta})}}\right] = [10.95, 46.65]. \tag{8.4.2}$$

Likiarvo

$$(\hat{\theta} - \theta_0)\sqrt{\mathcal{I}(\hat{\theta})} \approx \mathrm{N}(0,1) \tag{8.4.3}$$

ei ole tässä tapauksessa kovin tarkka, sillä välin (8.4.2) oikea peitetodennäköisyys on 0.9035. Väli (8.4.2) on symmetrinen $\hat{\theta}$:n suhteen, kun taas $r(\theta)$:n antamat uskottavuusvälit eivät ole symmetrisiä $\hat{\theta}$:n suhteen. Likiarvo (8.4.3) paranee, jos tehdään parametrimuunnos, joka tekee funktiosta $r(\theta)$ mahdollisimman symmetrisen $\hat{\theta}$:n suhteen. □

8.4.2 Napasuureet

Olkoon $X_1, \ldots, X_n$ otos tasajakaumasta $\mathrm{Tas}(0, \theta)$ ja olkoon $Y = X_{(n)}$ havaintojen maksimi. Muodostetaan tuntemattomalle parametrille θ väliestimaatti. Tarkastellaan kahta vaihtoehtoa:

$$\begin{aligned} &[Y, aY], && 1 < a; \\ &[Y, Y + b], && 0 < b, \end{aligned}$$

missä a ja b ovat annettuja vakioita. Ensimmäisen välin peitetodennäköisyys on

$$P(\theta \in [Y, aY]) = P(Y \le \theta \le aY) = P\left(\frac{1}{a} \le \frac{Y}{\theta} \le 1\right).$$

Koska Y:n tiheysfunktio on $f_Y(y) = ny^{n-1}/\theta^n$, $0 \le y \le \theta$, niin satunnaismuuttujan $T = Y/\theta$ tiheysfunktio on $f_T(t) = nt^{n-1}$, $0 \le t \le 1$. Siksi peitetodennäköisyys on

$$P\left(\frac{1}{a} \le T \le 1\right) = \int_{1/a}^{1} nt^{n-1}\,\mathrm{d}t = 1 - \left(\frac{1}{a}\right)^n.$$

Peitetodennäköisyys ei riipu θ:sta ja siksi välin $[Y, aY]$ luottamustaso on $1 - (1/a)^n$. Toisen välin peitetodennäköisyys on

$$\begin{aligned} P(\theta \in [Y, Y + b]) &= P(Y \le \theta \le Y + b) \\ &= P\left(1 - \frac{b}{\theta} \le T \le 1\right) \\ &= \int_{1-b/\theta}^{1} nt^{n-1}\,\mathrm{d}t = 1 - \left(1 - \frac{b}{\theta}\right)^n. \end{aligned}$$

Tässä tapauksessa peitetodennäköisyys riippuu parametrista θ. Väli $[Y, Y+b]$ ei siis ole edellä esitetyn määritelmän mielessä luottamusväli.

Sanomme satunnaismuuttujaa $T(X_1, \ldots, X_n; \theta)$ *napasuureeksi* (*pivotal quantity* tai *pivot*), jos $T(X_1, \ldots, X_n; \theta)$:n jakauma ei riipu parametrista θ. Napasuureen avulla voidaan siis konstruoida luottamusvälejä.

Esimerkki 8.17 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma_0^2)$, missä σ_0^2 on tunnettu. Silloin $(\overline{X} - \mu)/(\sigma_0/\sqrt{n})$ on napasuure, jonka avulla voidaan määrittää μ:n luottamusväli. Jokaista $0 < \alpha < 1$ kohti voidaan löytää standardoidun normaalimuuttujan Z avulla sellainen luku $z_{\alpha/2}$, että

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Jos esimerkiksi $1 - \alpha = 0.95$, niin $z_{\alpha/2} = z_{0.025} = 1.96$. Tästä saadaan μ:n luottamusväli

$$\left\{ \mu \;\middle|\; \overline{X} - z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \frac{\sigma_0}{\sqrt{n}} \right\},$$

jonka luottamustaso on $1 - \alpha$. Sanomme myös, että väli on $(1 - \alpha)100$ %:n luottamusväli.

Jos otos on normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, missä myös σ^2 on tuntematon, niin

$$T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

noudattaa t-jakaumaa vapausastein $n-1$. Otosvarianssi S^2 on σ^2:n harhaton estimaattori. Silloin $(1 - \alpha)100$ %:n luottamusväli saadaan valitsemalla t-jakauman avulla luku $t_{\alpha/2;\, n-1}$ siten, että

$$P\left(-t_{\alpha/2;\, n-1} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2;\, n-1}\right) = P(-t_{\alpha/2;\, n-1} \leq T_{n-1} \leq t_{\alpha/2;\, n-1}).$$

Silloin μ:n $(1 - \alpha)100$ %:n luottamusväli on

$$\left\{ \mu \;\middle|\; \overline{X} - t_{\alpha/2;\, n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2;\, n-1} \cdot \frac{S}{\sqrt{n}} \right\}.$$

□

Esimerkki 8.18 Tarkastellaan nyt normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ parametrin σ^2 väliestimointia. Koska

$$T = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

on napasuure, niin

$$\begin{aligned} 1 - \alpha &= P(a \leq T \leq b), \qquad a < b \\ &= P\left(\frac{a}{n\sigma^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{n\sigma^2}\right) = P\left(\frac{n\hat{\sigma}^2}{b} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{a}\right) \end{aligned}$$

on luottamusvälin $[n\hat{\sigma}^2/b,\, n\hat{\sigma}^2/a]$ luottamustaso. □

8.5 Kahden parametrin mallit

Oletetaan nyt, että todennäköisyysmallissa on kaksi tuntematonta parametria α ja β. Olkoon $r(\alpha, \beta)$ logaritminen normitettu uskottavuusfunktio. Parametrien (α, β) $100p$ %:n uskottavuusalue on

$$\mathrm{ua}(\alpha, \beta) = \{\, (\alpha, \beta) \mid r(\alpha, \beta) \geq \log p \,\}.$$

Alaluvussa 7.10.3 määriteltiin parametrin β logaritminen normitettu profiilifunktio

$$r^*(\beta) = \max_{\alpha} r(\alpha, \beta).$$

Parametrin β $100p$ %:n *profiiliväli* on

$$\mathrm{pv}(\beta) = \{\, \beta \mid r^*(\beta) \geq \log p \,\} \tag{8.5.1}$$

Ajatellaan ensin, että koetta toistetaan, kun parametriparin (α, β) arvoksi on kiinnitetty (α_0, β_0). Määritetään kokeen tuloksesta uskottavuustestisuure

$$D = -2\, r(\alpha_0, \beta_0).$$

Toisaalta tarkastellaan koetoistoja, kun vain $\beta = \beta_0$ on kiinnitetty. Tämän kokeen tuloksesta määritetään uskottavuustestisuure

$$D_2 = -2\, r^*(\beta_0).$$

Myöhemmin testisuuretta D tullaan käyttämään hypoteesin $(\alpha, \beta) = (\alpha_0, \beta_0)$ ja testisuuretta D_2 hypoteesin $\beta = \beta_0$ testaamiseen.

Suureiden D ja D_2 arvot vaihtelevat kokeesta toiseen havaintojen arvoista riippuen. Periaatteessa niiden tarkka jakauma voidaan johtaa todennäköisyysmallista, muttä käytännössä se voi olla vaikeaa. Siksi käytämme useimmiten likiarvoa. Alaluvussa 8.3 esitettyjen ehtojen vallitessa D ja D_2 noudattavat suurissa otoksissa likimain χ^2-jakaumaa:

$$D \approx \chi^2(2) \qquad \text{ja} \qquad D_2 \approx \chi^2(1).$$

Arvopari (α_0, β_0) kuuluu $100p$ %:n uskottavuusalueeseen, jos $r(\alpha_0, \beta_0) \geq \log p$. Tällöin $100p$ %:n uskottavuusalueen peitetodennäköisyys on

$$\begin{aligned} \mathrm{pt}(\alpha_0, \beta_0) &= P(D \leq -2 \log p \mid \alpha = \alpha_0,\ \beta = \beta_0) \\ &\approx P(\chi^2(2) \leq -2 \log p). \end{aligned}$$

Esimerkki 8.19 Olkoon havainto x jakaumasta $\mathrm{N}(\mu, 1)$ ja y jakaumasta $\mathrm{N}(\theta, 1)$. Lisäksi x ja y ovat toisistaan riippumattomat. Silloin $\hat{\mu} = x$, $\hat{\theta} = y$ ja

$$r(\mu, \theta) = -\tfrac{1}{2}(x - \mu)^2 - \tfrac{1}{2}(y - \theta)^2,$$

missä $-\infty < \mu < \infty$ ja $-\infty < \theta < \infty$.

Kun koetta toistetaan parametrin arvoilla $\mu = \mu_0$ ja $\theta = \theta_0$, niin

$$D = -2\,r(\mu_0, \theta_0) = (X - \mu_0)^2 + (Y - \theta)^2 = Z_1^2 + Z_2^2$$

ja

$$D_2 = -2\,r_{\max}(\theta_0) = (Y - \theta_0)^2 = Z_2^2.$$

Koska $X \sim \mathrm{N}(\mu_0, 1)$ ja $Y \sim \mathrm{N}(\theta_0, 1)$ ovat riippumattomat, niin $Z_1 = X - \mu_0$ ja $Z_2 = Y - \theta_0$ ovat riippumattomat ja noudattavat normaalijakaumaa $\mathrm{N}(0, 1)$. Siksi D ja D_2 ovat riippumatomat ja $D \sim \chi^2(2)$ ja $D_2 \sim \chi^2(1)$. □

8.6 Odotettu informaatio ja kokeiden suunnittelu

8.6.1 Johdanto

Toistaiseksi on yleensä oletettu, että koe on tehty (havainnot käytettävissä). Olemme tarkastelleet sitä, miten havainnoista saadaan parametria θ koskevaa informaatiota. Tilastollisia menetelmiä voidaan käyttää myös kokeiden suunnitteluvaiheessa päätettäessä siitä, mitä kokeita kannattaa tehdä ja mitä estimaattoreita käyttää.

Alaluvussa 8.4 esitettyjen tulosten mukaan

$$\hat{\theta} \pm \frac{c}{\sqrt{\mathcal{I}(\hat{\theta})}}$$

on θ:n likimääräinen uskottavuusväli ja luottamusväli. Suuri informaatiofunktion arvo $\mathcal{I}(\hat{\theta})$ tarkoittaa sitä, että saadaan lyhyt luottamusväli. Jos siis kokeen tarkoituksena on saada hyvä θ:n estimaatti, kannattaa yrittää valita koe siten, että saadaan mahdollisimman suuri $\mathcal{I}(\hat{\theta})$. Emme kuitenkaan voi laskea esimaattia $\hat{\theta}$ ja vastaavasi informaatiofunktion arvoa $\mathcal{I}(\hat{\theta})$ ennen koetta.

Esimerkki 8.20 Testataan n:n tuotteen kestoaika tarkistamalla, moniko niistä kestää tietyn ennalta asetetun testiajan t_0. Olkoon Y testiajan t_0 kestävien tuotteiden lukumäärä. Oletetaan, että kestoajat X_i noudattavat toisistaan riippumatta eksponenttijakaumaa $\mathrm{Exp}(\theta)$. Miten t_0 olisi valittava, että odotettu informaatio olisi mahdollisimman suuri?

Koska $X \sim \mathrm{Exp}(\theta)$, niin

$$\pi(\theta) = P(X > t_0) = \int_{t_0}^{\infty} \frac{1}{\theta} \mathrm{e}^{-x/\theta}\,\mathrm{d}x = \mathrm{e}^{-t_0/\theta}.$$

Testiajan t_0 kestävien tuotteiden lukumäärä Y noudattaa binomijakaumaa $\mathrm{Bin}(n, \pi)$, jonka logaritmoitu uskottavuusfunktio on

$$l(\theta) = Y \log \pi(\theta) + (n - Y) \log[1 - \pi(\theta)],$$

missä $\pi(\theta) = \mathrm{e}^{-t_0/\theta}$. Derivoimalla $l(\theta)$ kaksi kertaa saadaan

$$\mathcal{I}(\theta) = \left[\frac{Y}{\pi^2} + \frac{n-Y}{(1-\pi)^2}\right]\left(\frac{\mathrm{d}\pi}{\mathrm{d}\theta}\right)^2 - \left(\frac{Y}{\pi} - \frac{n-Y}{1-\pi}\right)\frac{\mathrm{d}^2\pi}{\mathrm{d}\theta^2}.$$

Koska $E(Y) = n\pi$, niin

$$\begin{aligned} I(\theta) = E[\mathcal{I}(\theta)] &= \frac{n}{\pi(1-\pi)}\left(\frac{\mathrm{d}\pi}{\mathrm{d}\theta}\right)^2 \\ &= \frac{n\pi t_0^2}{(1-\pi)\theta^4} = \frac{n}{\theta^2}\cdot\frac{\pi}{1-\pi}(\log\pi)^2. \end{aligned}$$

Voidaan osoittaa, että $I(\theta)$ saavuttaa maksiminsa, kun $\pi = 0.203$. Siksi kannattaa yrittää valita t_0 siten, että noin 20 % testattavista tuotteista kestää koko testijakson ajan. □

8.6.2 Pistefunktion ja informaatiofunktion ominaisuuksia

Oletetaan, että X noudattaa jakaumaa, jonka tiheysfunktio on $f(x;\theta)$. Fisherin pistefunktio on

$$u(\theta) = u(\theta; x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\, l(\theta; x),$$

missä $l(\theta; x) = \log f(x;\theta)$. Tarkastellaan nyt funktiota satunnaismuuttujana $U = u(\theta; X)$, kuten alaluvussa 8.3.

Esimerkki 8.21 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\mu, \sigma_0^2)$, missä σ_0^2 on tunnettu. Silloin μ:n logaritmoitu uskottavuusfunktio on

$$l(\mu) = -\frac{1}{2\sigma_0^2}\sum(X_i - \mu)^2$$

ja

$$U = \frac{1}{\sigma_0^2}\sum(X_i - \mu).$$

Silloin

$$E(U) = E\left[\frac{1}{\sigma_0^2}\sum(X_i - \mu)\right] = \frac{1}{\sigma_0^2}\sum E(X_i - \mu) = 0$$

ja

$$\begin{aligned} \mathrm{Var}(U) = \mathrm{Var}\left[\frac{1}{\sigma_0^2}\sum(X_i - \mu)\right] &= \frac{1}{\sigma_0^4}\,\mathrm{Var}\Big[\sum(X_i - \mu)\Big] \\ &= \frac{\sum \mathrm{Var}(X_i - \mu)}{\sigma_0^4} = \frac{n\sigma_0^2}{\sigma_0^4} = \frac{n}{\sigma_0^2}. \end{aligned}$$

Toisaalta $l''(\mu) = -n/\sigma_0^2$, joten $\mathcal{I}(\mu) = n/\sigma_0^2$ ei riipu havainnoista ja silloin myös $I(\mu) = n/\sigma_0^2$. Havaitsemme siis, että $\mathrm{Var}(U) = I(\mu)$. Seuraavassa näytetään, että vastaava tulos pitää paikkansa kaikille piste- ja informaatiofunktioille, kunhan riittävät säännöllisyysehdot ovat voimassa. □

Voidaan osoittaa yleisesti, että $E(U) = 0$:

$$\begin{aligned} E(U) &= E\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X;\theta)\right] \\ &= \int_S \left[\frac{1}{f(x;\theta)}\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}f(x;\theta)\right] f(x;\theta)\,\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_S f(x;\theta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}1 = 0, \end{aligned} \tag{8.6.1}$$

missä S on X:n arvoalue. Odotusarvon laskemisessa on oletettu, että derivoinnin ja integroinnin järjestystä voidaan vaihtaa. Koska $E(U) = 0$, niin

$$\mathrm{Var}(U) = E(U^2). \tag{8.6.2}$$

Tätä suuretta kutsutaan (Fisherin) *odotetuksi informaatioksi* ja merkitään

$$\mathrm{Var}(U) = I(\theta).$$

Toisaalta odotettu informaatio voidaa lausua muodossa

$$I(\theta) = -E\left[\frac{\mathrm{d}^2\, l(\theta;x)}{\mathrm{d}\theta^2}\right]. \tag{8.6.3}$$

Todetaan ensin, että

$$\begin{aligned} \frac{\mathrm{d}\, l}{\mathrm{d}\theta^2} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{1}{f(x;\theta)}\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}f(x;\theta)\right] \\ &= \frac{1}{f(x;\theta)}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(x;\theta) - \frac{1}{[f(x;\theta)]^2}\left[\frac{\mathrm{d}f(x;\theta)}{\mathrm{d}\theta}\right]^2 \\ &= \frac{1}{f(x;\theta)}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(x;\theta) - \left[\frac{\mathrm{d}\log f(x;\theta)}{\mathrm{d}\theta}\right]^2. \end{aligned}$$

Koska

$$\begin{aligned} E\left[\frac{1}{f(x;\theta)}\cdot\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(x;\theta)\right] &= \int_S \frac{1}{f(x;\theta)}\cdot\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(x;\theta)\cdot f(x;\theta)\,\mathrm{d}x \\ &= \int_S \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(x;\theta)\,\mathrm{d}x = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\int_S f(x;\theta)\,\mathrm{d}x = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}1 = 0, \end{aligned}$$

niin

$$-E\left[\frac{\mathrm{d}^2 l(\theta;x)}{\mathrm{d}\theta^2}\right] = E\left[\frac{\mathrm{d}\log f(x;\theta)}{\mathrm{d}\theta}\right]^2 = E(U^2).$$

Näin on identiteetti (8.6.3) todistettu. Odotusarvon laskemisessa oletettiin jälleen, että derivoinnin ja integroinnin järjestys voidaan vaihtaa.

Kuten alaluvussa 7.5 todettiin, riippumattomista kokeista saatu parametrin θ pistefunktio saadaan laskemalla yhteen riippumattomista kokeista määritetyt pistefunktiot. Informaatiofunktiolla on vastaava ominaisuus. Olkoot X_1 ja X_2 riippumattomat satunnaismuuttujat ja olkoot $U_1(\theta) = u(\theta, X_1)$ ja $U_2(\theta) = u(\theta, X_2)$ vastaavat pistefunktiot. Silloin yhdistetyn kokeen pistefunktio $U(\theta) = u(\theta; X_1, X_2)$ on

$$U(\theta) = U_1(\theta) + U_2(\theta).$$

Ominaisuuden (8.6.1) nojalla $E(U) = 0$ ja

$$\mathrm{Var}(U) = \mathrm{Var}(U_1) + \mathrm{Var}(U_2),$$

koska U_1 ja U_2 ovat riippumattomat. Koska $\mathrm{Var}(U) = I(\theta)$, niin odotettu informaatio on additiivinen:

$$I(\theta) = I_1(\theta) + I_2(\theta), \tag{8.6.4}$$

missä $\mathrm{Var}(U_i) = I_i(\theta)$, $i = 1, 2$.

Jos nyt $X_1, \ldots, X_n$ on otos jakaumasta, jonka tiheysfunktio on $f(x;\theta)$, niin

$$U(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{n} U_i(\theta),$$

missä $E(U_i) = 0$ ja U_i:t ovat riippumattomat, $1 \le i \le n$. Hajotelmaa (8.6.4) vastaavasti

$$I(\theta) = \sum_{j=1}^{n} \mathrm{Var}(U_j) = n\, i(\theta),$$

missä $i(\theta) = \mathrm{Var}(U_j)$, $i \le j \le n$.

8.6.3 Cramérin ja Raon alaraja

Cramérin ja Raon lauseen avulla voidaan määrittää harhattoman estimaattorin varianssin alaraja.

Lause 8.1 (Cramér ja Rao) *Olkoon $\hat{\theta}$ parametrin θ harhaton estimaattori. Silloin tiettyjen säännöllisyysehtojen vallitessa*

$$\mathrm{Var}(\hat{\theta}) \ge \frac{1}{I(\theta)}, \tag{8.6.5}$$

missä $I(\theta)$ on odotettu informaatio.

Todistus. Koska pistefunktion U odotusarvo on 0, niin

$$\begin{aligned}\mathrm{Cov}(\hat{\theta}, U) &= E[(\hat{\theta} - \theta)U] \\ &= E(\hat{\theta}U) - \theta\, E(U) = E(\hat{\theta}U).\end{aligned}$$

Toisaalta saadaan

$$\begin{aligned}E(\hat{\theta}U) &= \int \hat{\theta}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x;\theta)\right] f(x;\theta)\,\mathrm{d}x \\ &= \int \hat{\theta}\left[\frac{1}{f(x;\theta)}\frac{\mathrm{d}}{\mathrm{d}\theta} f(x;\theta)\right] f(x;\theta)\,\mathrm{d}x \\ &= \int \hat{\theta}\frac{\mathrm{d}}{\mathrm{d}\theta} f(x;\theta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\theta}\int \hat{\theta} f(x;\theta)\,\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta}\theta = 1.\end{aligned}$$

Koska

$$[\mathrm{Cov}(\hat{\theta}, U)]^2 \le \mathrm{Var}(\hat{\theta})\,\mathrm{Var}(U),$$

$\mathrm{Cov}(\hat{\theta}, U) = 1$ ja $\mathrm{Var}(U) = I(\theta)$, niin tästä seuraa Cramérin ja Raon epäyhtälö (8.6.5)

$$\mathrm{Var}(\hat{\theta}) \ge \frac{1}{I(\theta)}.$$

□

8.6.4 Suurimman uskottavuuden estimaattorin ominaisuuksia

Totesimme alaluvussa 8.3.2, että $U(\theta_0)$ noudattaa likimain normaalijakaumaa, kun n on suuri:

$$U(\theta_0) \xrightarrow{d} \mathrm{N}[\theta_0, I(\theta_0)], \qquad \text{kun } n \to \infty.$$

Seuraavassa esitetään melko luettelomaisesti suurimman uskottavuuden estimaattorin asymptoottiset ominaisuudet.

Tarkentuva. Suurimman uskottavuuden estimaattori on tarkentuva eräiden yleisten ehtojen vallitessa. Silloin siis $\lim_{n\to\infty} P(|\hat{\theta} - \theta_0| > \varepsilon) = 0$ kaikilla $\varepsilon > 0$, missä $\hat{\theta}$ on parametrin $\theta_0 \in \Theta$ suurimman uskottavuuden estimaattori.

Asymptoottisesti normaalinen. Eräiden yleisten ehtojen vallitessa

$$\hat{\theta} \xrightarrow{d} \mathrm{N}\left(\theta_0, \frac{1}{I(\theta_0)}\right), \qquad \text{kun } n \to \infty.$$

Kun $\theta = \theta_0$, niin $\hat{\theta}$:n jakauma lähenee normaalijakaumaa n:n kasvaessa.

Esimerkki 8.22 Olkoon $X_1, \ldots, X_n$ otos eksponenttijakaumasta $\text{Exp}(1/\beta)$. Silloin $f(x) = \beta \text{e}^{-\beta x}$ ja

$$l(\beta) = n \log \beta - \beta \sum x_i,$$

joten $l''(\beta) = -n/\beta^2$ ja $\mathcal{I}(\beta) = n/\beta^2$. Suurilla n:n arvoilla $\hat{\beta}$ noudattaa likimain normaalijakaumaa:

$$\hat{\beta} \approx \text{N}\left(\beta, \frac{\beta^2}{n}\right).$$

□

Invariantti parametrisoinnin suhteen. Jos $\hat{\theta}$ on θ:n suurimman uskottavuuden estimaattori, niin $g(\hat{\theta})$ on $g(\theta)$:n suurimman uskottavuuden estimaattori, missä g on θ:n monotoninen funktio.

Esimerkki 8.23 Olkoon $X_1, \ldots, X_n$ otos eksponenttijakaumasta $\text{Exp}(\theta)$. Parametrin θ suurimman uskottavuuden estimaattori on $\hat{\theta} = \overline{x}$. Lasketaan todennäköisyyden $P(X \geq t_0)$ suurimman uskottavuuden estimaattori. Eksponenttijakauman perusteella $P(X \geq t_0) = \text{e}^{-t_0/\theta} = \pi(\theta)$. Todennäköisyyden π suurimman uskottavuuden estimaattori on

$$\hat{\pi} = \pi(\hat{\theta}) = \text{e}^{-t_0/\overline{x}}.$$

□

Lause 8.2 *Oletetaan, että $\tilde{\theta}$ on θ:n harhaton estimaattori, joka saavuttaa Cramérin ja Raon alarajan. Jos θ:n suurimman uskottavuuden estimaattori $\hat{\theta}$ on yhtälön $\frac{\partial l}{\partial \theta} = 0$ ratkaisu, niin $\tilde{\theta} = \hat{\theta}$.*

Lauseen todistus sivuutetaan.

Suurimman uskottavuuden estimaattori voi olla harhainen, mutta se on asymptoottisesti harhaton. Jos $\hat{\theta}$ on harhaton, niin sillä on Cramérin ja Raon lauseessa esitetty minimivarianssi $1/I(\theta)$. Tällaista estimaattoria sanotaan tehokkaaksi. Jos $\hat{\theta}$ on harhainen, niin $\hat{\theta}$ on asymptoottisesti tehokas eli $\hat{\theta}$:n varianssi lähenee varianssin alarajaa.

8.6.5 Uskottavuusfunktioon liittyviä testisuureita

Alaluvussa 8.1 määriteltiin uskottavuussuhteen avulla *uskottavuustestisuure* $D(\theta; x) = -2 \log \frac{L(\theta;x)}{L(\hat{\theta};x)}$. Seuraavassa esitetään kaksi muuta keskeistä uskottavuusfunktioon perustuvaa testisuuretta. Nämä ovat Waldin testisuure ja Raon pistetestisuure.

Koska suurilla n:n arvoilla likimain $\hat{\theta} \sim \text{N}[\theta, I(\theta)^{-1}]$, niin $(\hat{\theta} - \theta) I(\theta)^{1/2} \sim \text{N}(0, 1)$. Silloin siis likimain

$$(\hat{\theta} - \theta)^2 I(\theta) \sim \chi^2(1).$$

Kun testataan hypoteesia H_0: $\theta = \theta_0$, niin H_0:n vallitessa testisuure

$$W = (\hat{\theta} - \theta_0)^2 I(\hat{\theta}) \tag{8.6.6}$$

noudattaa likimain χ^2-jakaumaa vapausastein 1.

Vastaava tulos pätee myös vektoriarvoisille parametreille. Jos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\mathsf{T}$ ja $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^\mathsf{T}$, niin $\hat{\boldsymbol{\theta}}$ noudattaa suurilla n:n arvoilla likimain k:n muuttujan normaalijakaumaa $\mathrm{N}_k[\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1}]$, missä k on vapaiden parametrien lukumäärä ja $I(\boldsymbol{\theta})^{-1}$ on Fisherin informaatiomatriisin (odotettu informaatio) käänteismatriisi. Silloin likimain

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\mathsf{T} I(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \chi^2(k).$$

Kun H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ ja $I(\boldsymbol{\theta})$ korvataan estimaatillaan $I(\hat{\boldsymbol{\theta}})$, niin likimain

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\mathsf{T} I(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2(k). \tag{8.6.7}$$

Testisuureita (8.6.6) ja (8.6.7) kutsutaan *Waldin testisuureiksi*.

Pistesuure $U(\theta) = l'(\theta; X_1, \ldots, X_n)$ noudattaa likimain normaalijakaumaa $\mathrm{N}[0, I(\theta)]$, kun n on suuri. Silloin likimain

$$\frac{U(\theta)}{\sqrt{I(\theta)}} \sim \mathrm{N}(0,1) \qquad \text{ja vastaavasti} \qquad \frac{U(\theta)^2}{I(\theta)} \sim \chi^2(1).$$

Kun testataan hypoteesia H_0: $\theta = \theta_0$, niin H_0:n vallitessa testisuure

$$S = \frac{U(\theta_0)^2}{I(\theta_0)} \tag{8.6.8}$$

noudattaa myös likimain χ^2-jakaumaa vapausastein 1.

Kun $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\mathsf{T}$, niin suurilla n:n arvoilla likimain

$$U(\boldsymbol{\theta})^\mathsf{T} I(\boldsymbol{\theta})^{-1} U(\boldsymbol{\theta}) \sim \chi^2(k),$$

missä $U(\boldsymbol{\theta})$ on pistevektori ja $I(\boldsymbol{\theta})^{-1}$ on informaatiomatriisin käänteismatriisi. Kun testataan hypoteesia H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, niin H_0:n vallitessa likimain

$$S = U(\boldsymbol{\theta}_0)^\mathsf{T} I(\boldsymbol{\theta}_0)^{-1} U(\boldsymbol{\theta}_0) \sim \chi^2(k). \tag{8.6.9}$$

Testisuuretta (8.6.8) sanotaan *Raon pistetestisuureeksi*. Jos hypoteesi on yhdistetty ja testisuure (8.6.9) riippuu tuntemattomasta parametrista, ne estimoidaan H_0:n vallitessa. Raon pistetestisuureen asymptoottinen jakauma on sama kuin vastaavan uskottavuustestisuureen.

Esimerkki 8.24 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\theta, \sigma^2)$, missä σ^2 on tunnettu. Testataan hypoteesia H_0: $\theta = \theta_0$, missä θ_0 on jokin annettu θ:n arvo. Silloin $\hat{\theta} = \overline{X}$ ja $I(\theta) = n/\sigma^2$, joten

$$W = (\hat{\theta} - \theta_0) I(\hat{\theta}) = \frac{n}{\sigma^2}(\overline{X} - \theta_0)^2.$$

Huomaa, että $I(\hat{\theta}) = I(\theta)$, koska $I(\theta)$ ei riipu parametrin arvosta. Koska $(\overline{X} - \theta_0)/(\sigma/\sqrt{n}) \sim \mathrm{N}(0,1)$, niin

$$W = \left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \right)^2 \sim \chi^2(1).$$

Tässä tapauksessa Waldin testisuure noudattaa täsmällisesti χ^2-jakaumaa.

Raon pistetestisuure on

$$S = \frac{U(\theta_0)^2}{I(\theta_0)} = \frac{\sigma^2}{n} \left[\frac{n(\overline{X} - \theta_0)}{\sigma^2} \right]^2 = \frac{n}{\sigma^2} (\overline{X} - \theta_0)^2,$$

koska $U(\theta_0) = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \theta_0) = \frac{n}{\sigma^2}(\overline{X} - \theta_0)$.

Uskottavuustestisuure on

$$\begin{aligned} D &= 2[l(\hat{\theta}) - l(\theta_0)] \\ &= -\frac{1}{\sigma^2} \sum (X_i - \overline{X})^2 + \frac{1}{\sigma^2} \sum (X_i - \theta_0)^2 = \frac{n}{\sigma^2} (\overline{X} - \theta_0)^2. \end{aligned}$$

Tässä tapauksessa siis kaikki kolme testisuuretta W, S ja D ovat identtiset. □

Luku 9

Hypoteesien testaus

9.1 Yleisiä näkökohtia

Hypoteesin H_0 testaus on menettely, jolla mitataan havaintojen hypoteesia vastaan osoittamaa todistusvoimaa. Oletamme, että H_0 on tosi, ja tarkistamme sitten, onko tämä oletus sopusoinnussa havaintojen kanssa. Menettely muistuttaa jossain määrin matematiikasta tuttua todistustekniikkaa vastaoletuksen avulla. Tilastollinen hypoteesi H_0 ei tavallisesti johda havaintojen ja hypoteesin väliseen loogiseen ristiriitaan, mutta havainnot saattavat olla äärimmäisen epätodennäköisiä, mikäli H_0 olisi tosi. Mitä epätodennäköisempiä saadut havainnot ovat H_0:n vallitessa, sitä vahvemmin niiden katsotaan osoittavan, että H_0 ei pidä paikkansa.

Esimerkki 9.1 Heitetään rahaa 100 kertaa. Kruunujen lukumäärä X noudattaa binomijakaumaa $\mathrm{Bin}(100, \theta)$. Testataan, onko raha harhaton. Testattava hypoteesi on siis H_0: $\theta = \frac{1}{2}$.

Jos H_0 pitää paikkansa, niin

$$f(x) = \binom{100}{x} \left(\frac{1}{2}\right)^{100}, \qquad x = 0, 1, \ldots, 100.$$

Tiedämme, että $E(X) = 50$ ja odotusarvon läheisyydessä olevat X:n arvot ovat kaikkein todennäköisimpiä. Voimme valita *testisuureeksi* esimerkiksi satunnaismuuttujan $T = |X - 50|$, joka mittaa, kuinka hyvin havainnot ja hypoteesi sopivat yhteen.

Jos havaitaan esimerkiksi $X = 30$, niin $T = |30 - 50| = 20$. Todennäköisyys, että saadaan näin poikkeava arvo, on

$$P(T \geq 20) = P(|X - 50| \geq 20) \approx 0.$$

On äärimmäisen epätodennäköistä, että harhattomalla rahalla saadaan näin ”poikkeava” tulos. Havainto tukee voimakkaasti käsitystä, että H_0 on väärä ja raha on harhainen. □

9.1.1 Testisuureet ja p-arvot

Merkitsevyyden testauksessa mahdolliset koetulokset (havainnot) on kyettävä asettamaan järjestykseen sen mukaan, miten hyvin ne sopivat yhteen hypoteesin kanssa. Tätä tarkoitusta varten valitaan jokin *testisuure* T, joka mittaa havaintojen ja hypoteesin H_0 yhteensopivuutta. Testisuureen T pieni arvo osoittaa havaintojen hyvää yhteensopivuutta hypoteesin kanssa ja T:n suuri arvo osoittaa huonoa yhteensopivuutta. Testisuure valitaan ennen kuin tutkitaan havaintoja. Valintaan vaikuttaa tietysti se, millaisia poikkeamia halutaan tunnistaa.

Kun havainnot on saatu, voidaan laskea testisuureen arvo $T = t$. Sitten voidaan laskea todennäköisyys, että saadaan ainakin näin poikkeava T:n arvo, jos H_0 on tosi. Tätä todennäköisyyttä

$$\alpha_h = P(T \geq t \mid H_0 \text{ tosi}) \tag{9.1.1}$$

kutsutaan *p-arvoksi* tai *havaituksi merkitsevyystasoksi*. Jos α_h on hyvin pieni, niin se osoittaa, että sellaista havaintoa ei saataisi juuri koskaan, jos H_0 on tosi. Silloin havainnot tukevat vahvasti oletusta, että H_0 ei pidä paikkaansa.

Esimerkki 9.2 Henkilölle jaetaan satunnaisessa järjestyksessä pöytään kuvapuoli alaspäin 4 korttia, jotka ovat eri maata (♠, ♡, ♣, ♢). Jokaisen kortin kohdalla henkilö arvaa, mitä maata kortti on. Henkilö väittää, että hänellä on sellaisia yliaistillisia kykyjä, että hän pystyy keskittymällä "näkemään" kortin ja pääsemään parempiin tuloksiin kuin pelkkä arvaaja. Ajatellaan, että henkilöllä on 4 kirjekuorta, jotka on merkitty symboleilla ♠, ♡, ♣ ja ♢. Kortit on pistettävä oikeisiin kirjekuoriin.

Olkoon X oikein sijoitettujen korttien lukumäärä. Jos henkilö on pelkkä arvaaja, niin silloin oikeiden lukumäärän X todennäköisyysfunktio on

X	0	1	2	4	Yhteensä
$f(x)$	$\frac{9}{24}$	$\frac{8}{24}$	$\frac{6}{24}$	$\frac{1}{24}$	1

Laskemalla voidaan todeta, että $E(X) = 1$ ja $\mathrm{Var}(X) = 1$. Testataan nyt nollahypoteesi H_0, että henkilö on arvaaja. Toistetaan koe 50 kertaa ja olkoon X_i oikein sijoitettujen korttien lukumäärä i. kokeessa, $i = 1, \ldots, 50$. Silloin jokainen satunnaismuuttuja X_i noudattaa edellä määriteltyä X:n jakaumaa. Satunnaismuuttujien summa

$$S = X_1 + X_2 + \cdots + X_{50}$$

noudattaa keskeisen rajaväittämän mukaan likimain normaalijakaumaa. Nyt $E(S) = 50 \cdot 1$ ja $\mathrm{Var}(S) = 50 \cdot 1$, joten

$$\frac{S - 50}{\sqrt{50}} \approx \mathrm{N}(0, 1).$$

Henkilö sai seuraavan tuloksen:

Oikein sijoitettujen lkm	0	1	2	4	Yhteensä
Havaittu frekvenssi	17	18	9	6	50

Olkoon H_0, että henkilö on arvaaja. Aineistosta laskettu S:n arvo on 60. Siksi

$$\alpha_h = P(S \geq 60) \approx P\left(Z \geq \frac{60-50}{\sqrt{50}}\right) = 0.079,$$

missä Z noudattaa normaalijakaumaa $N(0,1)$. Koska p-arvo ei ole kovin pieni, niin H_0 saa jäädä voimaan. Toisaalta p-arvo on niin pieni, että se ei anna vakkuuttavaa näyttöä H_0:n puolesta. Eräs tapa ratkaista ongelma, on kerätä lisää aineistoa, mikäli mahdollista. □

9.2 Uskottavuussuhdetestit: Yksinkertaiset hypoteesit

Tilastollinen malli on tiheysfunktioiden jouko

$$\mathcal{F} = \{\, f(\,\cdot\,;\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k \,\}, \tag{9.2.1}$$

missä jokainen annettu $\boldsymbol{\theta} \in \Theta$ määrittelee jonkin jakauman tiheysfunktion. Kun satunnaismuuttuja on diskreetti, tiheysfunktion paikalla on todennäköisyysfunktio. Seuraavassa ajatellaan, että testattava *tilastollinen hypoteesi* voidaan lausua tilastollisen mallin tuntemattomia parametreja koskevina oletuksina.

Hypoteesi H on *yksinkertainen*, jos se määrittää kaikkien tuntemattomien parametrien arvot niin, ettei malliin jää tuntemattomia parametreja. *Yhdistetty hypoteesi* ei sen sijaan taysin määritä tuntemattomien parametrien arvoja, mutta rajoittaa mahdollisten arvojen joukkoa.

9.2.1 Yksi parametri

Oletetaan aluksi, että mallissa on yksi tuntematon parametri θ. Testataan hypoteesia

$$H_0 \colon \theta = \theta_0, \tag{9.2.2}$$

missä θ_0 on jokin θ:n numeerinen arvo.

Olkoon $l(\theta)$ logaritmoitu uskottavuusfunktio ja $\hat{\theta}$ parametrin θ suurimman uskottavuuden estimaatti. Uskottavuustestisuure hypoteesin (9.2.2) testaamiseksi on

$$\begin{aligned} D &= -2\log\frac{L(\theta_0)}{L(\hat{\theta})} \\ &= 2[l(\hat{\theta} - l(\theta_0))] = -2\,r(\theta_0), \end{aligned}$$

missä $r(\theta)$ on normitetun uskottavuusfunktion logaritmi. Pieni D:n arvo viittaa siihen, että parametrin arvo $\theta = \theta_0$ on uskottava.

Havainnoista x_1, x_2,, …, x_n voidaan laskea testisuureen arvo $d = D(x_1, x_2, \ldots, x_n)$. Kun arvo d on saatu, voidaan laskea testisuureen arvoon liittyvä todennäköisyys

$$\alpha_h = P(D \geq d \mid H \text{ tosi}),$$

joka on siis *havaittu merkitsevyystaso* eli *p-arvo*. Huomaa, että todennäköisyyden lausekkeessa suure $D = D(X_1, X_2, \ldots, X_n)$ on satunnaismuuttuja. Todennäköisyyden α_h tarkan arvon laskeminen osoittautuu usein vaikeaksi, mutta sopivien ehtojen vallitessa likiarvo

$$\alpha_h = P(D \geq d \mid \theta = \theta_0) \approx P(\chi_1^2 \geq d) \tag{9.2.3}$$

on riittävän hyvä. Likiarvo saadaan χ^2-jakauman avulla, sillä suurilla n:n arvoilla D noudattaa likimain χ^2-jakaumaa vapausastein 1.

Esimerkki 9.3 Tehdään n riippumatonta Bernoullin koetta, joissa onnistumistodennäköisyys on θ. Onnistumisten lukumäärä X noudattaa binomijakaumaa $\text{Bin}(n, \theta)$. Testataan hypoteesia H_0: $\theta = \theta_0$, missä θ_0 on jokin numeerinen arvo.

Binomijakauman logaritmoitu uskottavuusfunktio on

$$l(\theta) = x \log \theta + (n - x) \log(1 - \theta),$$

missä $0 < \theta < 1$. Silloin

$$l(\hat{\theta}) = x \log \frac{x}{n} + (n - x) \log\left(1 - \frac{x}{n}\right),$$

missä $\hat{\theta} = x/n$. Jos H_0 on tosi, silloin logaritmoidun uskottavuusfunktion arvo on

$$l(\theta_0) = x \log \theta_0 + (n - x) \log(1 - \theta_0).$$

Jos esimerkiksi $\theta_0 = 0.5$, silloin $l(0.5) = n \log(0.5)$. Uskottavuustestisuure hypoteesin H_0: $\theta = \theta_0$ testaamiseksi on

$$\begin{aligned} D &= 2[l(\hat{\theta}) - l(\theta_0)] = -2\, r(\theta_0) \\ &= 2\left[x \log \frac{x}{n\theta_0} + (n - x) \log \frac{n - x}{n(1 - \theta_0)}\right] \\ &= 2\left[x \log \frac{\hat{\theta}}{\theta_0} + (n - x) \log \frac{1 - \hat{\theta}}{1 - \theta_0}\right]. \end{aligned}$$

Jos n on suuri, D noudattaa likimain χ^2-jakaumaa vapausastein 1.

Olkoon esimerkiksi $n = 100$ ja $\theta_0 = \frac{1}{4}$. Silloin uskottavuustestisuure hypoteesin H_0: $\theta = \frac{1}{4}$ testaamiseksi on

$$D = 2\left[x \log \frac{x}{25} + (n - x) \log \frac{100 - x}{75}\right].$$

Jos havaitaan $X = 43$, on D:n havaittu arvo

$$d = 86 \cdot \log \frac{43}{25} + 114 \cdot \log \frac{57}{75} = 15.35.$$

Testisuureen arvoon liittyvä p-arvo on

$$\alpha_h = P\big(D \geq 15.35 \,\big|\, \theta = \tfrac{1}{4}\big) \approx P(\chi^2_1 \geq 15.35) \approx 0.$$

Jos $\theta = \frac{1}{4}$, on erittäin epätodennäköistä, että havaitaan arvo $X = 43$. Havainnot tukevat vahvasti käsitystä, että $\theta \neq \frac{1}{4}$.

Jos esimerkiksi $n = 10$ ja $\theta_0 = \frac{1}{4}$, niin χ^2-jakaumaan perustuva likiarvo ei ole enää kovin tarkka. Oletetaan, että havaittiin $X = 5$. Silloin D:n havaittu arvo on

$$d = 10 \cdot \log \frac{5}{2.5} + 10 \cdot \log \frac{5}{7.5} = 2.88$$

ja tarkka p-arvo on syytä laskea binomijakauman $\mathrm{Bin}\big(10, \frac{1}{4}\big)$ avulla:

$$\alpha_h = P(D \geq 2.88) = P(X = 0) + P(X \geq 5) = 0.1344.$$

□

9.2.2 Useita parametreja

Oletetaan, että malli riippuu tuntemattomien parametrien vektorista $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Olkoon $\boldsymbol{\theta}_0$ tämän vektorin annettu arvo. Uskottavuussuure hypoteesin

$$H_0 \colon \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{9.2.4}$$

testaamiseksi on

$$D = 2[l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}_0)] = -2\, r(\boldsymbol{\theta}_0),$$

missä $\hat{\boldsymbol{\theta}}$ on parametrivektorin $\boldsymbol{\theta}$ suurimman uskottavuuden estimaatti ja $r(\boldsymbol{\theta})$ on parametrien θ_1, θ_2, …, θ_k yhteisen normeeratun uskottavuusfunktion logaritmi. Voidaan osoittaa vastaavin perustein kuin alaluvussa 8.3, että D noudattaa likimain χ^2-jakaumaa. χ^2-jakauman vapausasteiden lukumäärä on mallissa olevien toisistaan funktionaalisesti riippumattomien parametrien lukumäärä. Tästä seuraa, että

$$\alpha_h = P(D \geq d \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0) \approx P[\chi^2(k) \geq d],$$

missä k on mallin riippumattomien parametrien lukumäärä

Esimerkki 9.4 Eräässä pitkäaikaisessa sydäntauteja koskevassa tutkimuksessa seurattiin suurta joukkoa miehiä, joilla ei ollut mitään merkkiä aikaisemmista sydänongelmista. Seurantajakson aikana miehistä 63 kuoli yllättävään sydäninfarktiin. Seuraavassa taulukossa kuolemat on kirjattu viikon päivän mukaan.

Viikon päivä	ma	ti	ke	to	pe	la	su
Kuolemien lkm	22	7	6	13	5	4	6

Testataan hypoteesi, että kuoleman todennäköisyys on sama kaikkina päivinä. Silloin testattava hypoteesi on

$$H_0 \colon \theta_i = \frac{1}{7}, \qquad i = 1, \ldots, 7,$$

missä θ_i on kuoleman todennäköisyys viikon i. päivänä. Mallissa on 6 vapaata parametria, sillä parametreja $\theta_1, \ldots, \theta_7$ sitoo rajoite $\sum_{i=1}^{7} \theta_i = 1$. Olkoon f_i viikon i. päivänä kuolleiden lukumäärä. Koska $\hat{\theta}_i = f_i/63$, niin

$$l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{7} f_i \log \hat{\theta}_i = -110.995.$$

Nollahypoteesin vallitessa

$$l\left(\tfrac{1}{7}\right) = \sum_{i=1}^{7} f_i \log\left(\tfrac{1}{7}\right) = -63 \log 7 = -122.592.$$

Silloin uskottavuustestisuureen arvo on

$$D = 2[l(\hat{\boldsymbol{\theta}}) - l\left(\tfrac{1}{7}\right)] = 23.274.$$

Testisuureeseen liittyvä p-arvo on

$$\alpha_h \approx P(\chi_6^2 \geq 23.274) = 0.0007.$$

Havainnot eivät tue hypoteesia H_0. □

9.3 Uskottavuussuhdetestit: Yhdistetyt hypoteesit

Oletetaan, että todennäköisyysmalli riippuu tuntemattomista parametreista $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Hypoteesi H_0 koskee nyt parametrivektorin $\boldsymbol{\theta}$ alkioita. Hypoteesi H_0 määrittää mallin, joka saadaan asettamalla joitain rajoitteita alkuperäiseen malliin. Hypoteesin H_0 määrittämä malli on alkuperäistä mallia yksinkertaisempi ja siinä on $q < k$ funktionaalisesti riippumatonta parametria.

Yksinkertainen hypoteesi määrittää kaikille parametreille arvot, niin että H_0:n vallitessa $q = 0$. Yhdistetty hypoteesi ei poista kaikkia tuntemattomia parametreja ja silloin $q > 0$.

Olkoon $l(\boldsymbol{\theta})$ parametrivektorin $\boldsymbol{\theta}$ logaritmoitu uskottavuusfunktio perusmallissa. Silloin $l(\hat{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta})$ kaikilla $\boldsymbol{\theta}$:n arvoilla, kun $\hat{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}$:n suurimman uskottavuuden estimaatti.

Olkoon $\tilde{\boldsymbol{\theta}}$ parametrivektorin $\boldsymbol{\theta}$ suurimman uskottavuuden estimaatti, kun oletetaan H_0 todeksi. Koska $l(\hat{\boldsymbol{\theta}}) \geq l(\boldsymbol{\theta})$ kaikilla $\boldsymbol{\theta}$:n arvoilla, niin $l(\hat{\boldsymbol{\theta}}) \geq l(\tilde{\boldsymbol{\theta}})$. Rajoitteen H_0 vallitessa laskettu maksimiarvo $l(\tilde{\boldsymbol{\theta}})$ ei voi olla suurempi kuin ilman rajoitetta laskettu maksimiarvo $l(\hat{\boldsymbol{\theta}})$.

Uskottavuussuhdetestisuure hypoteesin H_0 testaamiseksi määritellään näiden kahden maksimin erotuksen avulla:

$$D = 2[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})]. \tag{9.3.1}$$

Testisuureen D nimi johtuu siitä, että se voidaan lausua uskottavuussuhteen $L(\hat{\boldsymbol{\theta}})/L(\tilde{\boldsymbol{\theta}})$ funktiona seuraavasti:

$$D = 2 \log \frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})}.$$

Koska $l(\hat{\boldsymbol{\theta}}) \geq l(\tilde{\boldsymbol{\theta}})$, niin D on epänegatiivinen. Kun hypoteesi H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ on yksinkertainen, niin vektori $\boldsymbol{\theta}_0$ on jokin annettu numeerinen vektori. Silloin H_0:n vallitessa vain arvo $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ on mahdollinen, joten $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ ja funktion $l(\boldsymbol{\theta})$ maksimi yksinkertaisen hypoteesin tapauksessa on $l(\boldsymbol{\theta}_0)$.

9.3.1 p-arvon määrittäminen

Okoon $D = d$ havaittu uskottavuustestisuureen arvo. Silloin testisuureen arvoon liittyvä p-arvo on

$$\alpha_h = P(d \geq d \mid H_0 \text{ tosi}).$$

Useimmiten α_h:n tarkan arvon laskeminen on hankalaa, mutta χ^2-jakauman perusteella saadaan tyydyttävä likiarvo. Melko yleisten oletusten vallitessa (vrt. alaluku 8.3) *D noudattaa χ^2-jakaumaa vapausastein $k - q$, kun H_0 on tosi*. Tämän likiarvon perusteella

$$\alpha_h \approx P(\chi^2_{k-q} \geq d).$$

Huomattakoon, että χ^2-jakauman vapausasteet $k - q$ saadaan perusmallin parametrien lukumäärän k ja H_0:n määrittämän mallin parametrien lukumäärän q erotuksena.

9.3.2 Kaksi parametria, joista toista testataan

Oletetaan, että $\boldsymbol{\theta} = (\theta_1, \theta_2)$, joten $k = 2$. Testataan hypoteesia

$$H_0 \colon \theta_2 = a,$$

missä a on annettu numeerinen arvo. Hypoteesin H_0 vallitessa mallissa on yksi tuntematon parametri θ_1, joten $q = 1$.

Olkoon $l(\theta_1, \theta_2)$ parametrien θ_1 ja θ_2 logaritmoitu uskottavuusfunktio. Tavallisesti parametrien θ_1 ja θ_2 suurimman uskottavuuden estimaatit $\hat{\theta}_1$ ja $\hat{\theta}_2$ saadaan ratkaisemalla samanaikaisesti uskottavuusyhtälöt

$$u_1(\theta_1, \theta_2) = 0 \qquad \text{ja} \qquad u_2(\theta_1, \theta_2) = 0,$$

missä $u_1(\theta_1, \theta_2) = \frac{\partial\, l(\theta_1,\theta_2)}{\partial \theta_1}$ ja $u_2(\theta_1, \theta_2) = \frac{\partial\, l(\theta_1,\theta_2)}{\partial \theta_2}$. Kun H_0:n mukaisesti $\theta_2 = a$, saadaan θ_1:n ehdollinen suurimman uskottavuuden estimaatti $\hat{\theta}_1(a)$ ratkaisemalla yhtälö

$$u_1(\theta_1, a) = \frac{\partial\, l(\theta_1, a)}{\partial \theta_1} = 0.$$

Logaritmoidun uskottavuusfunktion $l(\theta_1, \theta_2)$ maksimi on $l(\hat{\theta}_1, \hat{\theta}_2)$. Hypoteesin H_0 vallitessa logaritmoidun uskottavuusfunktion maksimi on $l[\hat{\theta}_1(a), a]$. Siksi uskottavuustestisuure hypoteesin H_0: $\theta_2 = a$ testaamiseksi on

$$D = 2\big[l(\hat{\theta}_1, \hat{\theta}_2) - l\big(\hat{\theta}_1(a), a\big)\big].$$

Uskottavuusfunktio $l(\theta_1, \theta_2)$ maksimoidaan ensin ilman rajoitteita ja sitten rajoitteen H_0: $\theta_2 = a$ vallitessa ja sitten lasketaan maksimien erotus. Huomattakoon, että

$$D = -2\, r^*(a),$$

missä $r^*(\theta_2)$ on θ_2:n logaritmoitu normitettu profiilifunktio. Kun H_0: $\theta_2 = a$ on tosi, niin D:n jakauman likiarvona voidaan käyttää χ^2-jakaumaa vapausastein $k - q = 2 - 1 = 1$. Vastaavasti p-arvon likiarvo on

$$\alpha_h = P(D \geq d \mid \theta_2 = a) \approx P(\chi_1^2 \geq d).$$

Esimerkki 9.5 Olkoon $x_1, \ldots, x_n$ otos Weibullin jakaumasta, jonka tiheysfunktio on

$$f(x) = \alpha\beta x^{\beta-1} \exp(-\alpha x^\beta), \qquad 0 < x < \infty, \tag{9.3.2}$$

missä $\alpha > 0$ ja $\beta > 0$ ovat tuntemattomia parametreja. Testataan esimerkiksi iskunvaimentimien kestoa kokeilemalla, montako standardivoimakkuudella annettua iskua ne kestävät tietyissä vakio-olosuhteissa. Tuote-erästä valittiin kokeeseen 25 iskunvaimenninta.

Kun $\beta = 1$, saadaan Weibullin jakaumasta (9.3.2) erikoistapauksena eksponenttijakauma. Testataan siksi hypoteesi H_0: $\beta = 1$. Mallin (9.3.2) logaritmoitu uskottavuusfunktio on

$$l(\alpha, \beta) = n \log \alpha + n \log \beta + (\beta - 1) \sum \log x_i - \alpha \sum x_i^\beta.$$

Pistefunktiot ovat

$$\frac{\partial\, l}{\partial \alpha} = \frac{n}{\alpha} - \sum x_i^\beta,$$

$$\frac{\partial\, l}{\partial \beta} = \frac{n}{\beta} + \sum \log x_i - \alpha \sum x_i^\beta \cdot \log x_i.$$

Yhtälön $\frac{\partial l}{\partial \alpha} = 0$ ratkaisu on $\hat{\alpha}(\beta) = n\big/\sum x_i^\beta$. Sijoitetaan tämä ratkaisu yhtälöön $\frac{\partial l}{\partial \beta} = u_2(\alpha, \beta) = 0$ ja ratkaistaan β. Saadaan

$$u_2[\hat{\alpha}(\beta), \beta] = \frac{n}{\beta} + \sum \log x_i - \frac{n \sum x_i^\beta \log x_i}{\sum x_i^\beta}.$$

Yhtälö $u_2[\hat{\alpha}(\beta), \beta] = 0$ voidaan ratkaista numeerisesti esimerkiksi Newtonin menetelmällä.

Aineistosta saatiin suurimman uskottavuuden estimaatit

$$\hat{\alpha} = 9.515 \cdot 10^{-5} \qquad \text{ja} \qquad \hat{\beta} = 2.1021.$$

Uskottavuusfunktion maksimi on

$$l(\hat{\alpha}, \hat{\beta}) = -113.691.$$

Kun H_0: $\beta = 1$ on tosi, niin α:n suurimman uskottavuuden estimaatti on

$$\hat{\alpha}(1) = \frac{n}{\sum x_i} = \frac{25}{1805} = 0.01385$$

ja uskottavuusfunktion maksimi

$$l\big(\hat{\alpha}(1), 1\big) = -121.433.$$

Uskottavuustestisuureen arvo testattaessa hypoteesia H_0: $\beta = 1$ on

$$D = 2\big[l(\hat{\alpha}, \hat{\beta}) - l\big(\hat{\alpha}(1), 1\big)\big] = 2(-113.691 + 121.433) = 15.48.$$

Testisuureen arvoon liittyvä p-arvo on

$$\alpha_h \approx P(\chi_1^2 \geq 15.48) < 0.001.$$

Havainnot tukevat vahvasti käsitystä, että H_0: $\beta = 1$ ei pidä paikkaansa. □

9.3.3 Homogeenisuuden testaus

Oletetaan, että tehdään k riippumatonta koetta. Oletetaan aluksi, että jokaisen kokeen tuloksista estimoidaan eri parametri θ_i, $i = 1, \ldots, k$. Olkoon $l_i(\theta)$ ja $\hat{\theta}_i$ kokeisiin $i = 1, \ldots, k$ liittyvät logaritmoidut uskottavuusfunktiot ja suurimman uskottavuuden estimaatit. Yhdistettyyn kokeeseen liittyvä uskottavuusfunktio on

$$l(\theta_1, \ldots, \theta_k) = l_1(\theta_1) + \cdots + l_k(\theta_k)$$

ja funktion maksimi on $\sum_{i=1}^k l(\hat{\theta}_i)$.

Testataan nyt hypoteesia

$$H_0 \colon \theta_1 = \cdots = \theta_k. \tag{9.3.3}$$

Hypoteesin H_0 vallitessa kaikissa kokeissa parametrin arvo on sama ja silloin logaritmoitu uskottavuusfunktio on

$$l(\theta_1, \ldots, \theta_k) = \sum l_i(\theta), \tag{9.3.4}$$

missä yhteistä parametrin arvoa on merkitty θ:lla. Merkitään yhdistetystä kokeesta estimoitua suurimman uskottavuuden estimaattia $\tilde{\theta}$:lla, jolloin funktion (9.3.4) maksimi on $\sum l_i(\tilde{\theta})$. Uskottavuustestisuure hypoteesin (9.3.3) testaamiseksi on

$$D = 2\sum_{i=1}^{k}\Big[l_i(\hat{\theta}_i) - l_i(\tilde{\theta})\Big] = -2\sum_{i=1}^{k} r_i(\tilde{\theta}),$$

missä $r_i(\theta_i) = l_i(\theta_i) - l_i(\hat{\theta}_i)$ on i. kokeeseen liittyvä logaritmoitu normitettu uskottavuusfunktio.

Jos D on suuri, niin silloin ei ole olemassa sellaista parametrien arvoa, joka olisi uskottava kaikissa kokeissa. Hypoteesin testauksessa on $k-1$ vapausastetta, koska H_0 pudottaa tuntemattomien parametrien lukumäärän k:sta parametrista yhteen parametriin. Jos havaitaan $D = d$, niin

$$\alpha_h \approx P(\chi^2_{k-1} \geq d).$$

Esimerkki 9.6 Olkoon $X_1, \ldots, X_n$ otos jakaumasta $\mathrm{Exp}(\theta_1)$ ja $Y_1 \ldots, Y_m$ otos jakaumasta $\mathrm{Exp}(\theta_2)$. Silloin

$$f(x_1, \ldots, x_n; \theta_1) = \prod_{i=1}^{n} \frac{1}{\theta_1} \mathrm{e}^{-x_i/\theta_1} = \frac{1}{\theta_1^n} \mathrm{e}^{-\sum x_i/\theta_1}$$

ja

$$f(y_1, \ldots, y_n; \theta_1) = \prod_{j=1}^{m} \frac{1}{\theta_2} \mathrm{e}^{-y_j/\theta_2} = \frac{1}{\theta_2^m} \mathrm{e}^{-\sum y_j/\theta_2}.$$

Parametrin θ_1 logaritmoitu uskottavuusfunktio on

$$l_1(\theta_1) = -n \log \theta_1 - n\overline{X}/\theta_1$$

ja θ_2:n logaritmoitu uskottavuusfunktio on

$$l_2(\theta_2) = -m \log \theta_2 - m\overline{Y}/\theta_2,$$

missä $n\overline{X} = \sum_{i=1}^{n} X_i$ ja $m\overline{Y} = \sum_{j=1}^{m} Y_j$. Koska $\hat{\theta}_1 = \overline{X}$ ja $\hat{\theta}_2 = \overline{Y}$, niin

$$l_1(\hat{\theta}_1) = -n \log \overline{X} - n$$

ja

$$l_2(\hat{\theta}_2) = -m \log \overline{Y} - m.$$

Testataan hypoteesia $H_0 : \theta_1 = \theta_2$. Jos H_0 on tosi, niin $\theta_1 = \theta_2$ ja silloin

$$X_1, \ldots, X_n, Y_1, \ldots, Y_m \sim \text{Exp}(\theta),$$

missä $\theta = \theta_1 = \theta_2$. Parametrin θ logaritmoitu uskottavuusfunktio on

$$l(\theta) = -(n+m)\log\theta - \frac{\sum X_i + \sum Y_j}{\theta}$$

ja θ:n suurimman uskottavuuden estimaatti on yhdistetyn otoksen keskiarvo

$$\tilde{\theta} = \frac{\sum X_i + \sum Y_j}{n+m} = \frac{n\overline{X} + m\overline{Y}}{n+m}.$$

Silloin

$$l(\tilde{\theta}) = -(n+m)\log\tilde{\theta} - (n+m)$$

ja

$$\begin{aligned} D &= 2[l_1(\hat{\theta}_1) + l_2(\hat{\theta}_2) - l(\tilde{\theta})] \\ &= 2[-n\log\overline{X} - m\log\overline{Y} - (n+m) + (n+m)\log\tilde{\theta} + (n+m)] \\ &= 2\left[n\log\frac{\tilde{\theta}}{\overline{X}} + m\log\frac{\tilde{\theta}}{\overline{Y}}\right], \end{aligned}$$

joka riippuu vain otoskeskiarvoista $\overline{X}$, $\overline{Y}$ ja otoskoosta n ja m. □

Esimerkki 9.7 Testataan, onko erään järven vedessä tiettyä bakteeria. Ei ole mahdollista laskea näytteessä olevien bakteerien määrää, vaan voidaan ainoastaan todeta, onko näyteputkessa bakteereita vai ei. Negatiivinen tulos osoittaa, että näyteputkessa ei ole bakteereita. Positiivinen tulos tarkoittaa, että näytteessä on ainakin yksi bakteeri.

Bakteerien lukumäärä X tilavuudessa V vettä noudattaa Poissonin jakaumaa $\text{Poi}(\mu V)$:

$$f(x) = \frac{(\mu V)^x \mathrm{e}^{-\mu V}}{x!}, \qquad x = 0, 1, \ldots.$$

Negatiivisen reaktion todennäköisyys on

$$\pi = f(0) = \mathrm{e}^{-\mu V}$$

ja positiivisen reaktion todennäköisyys on

$$1 - \pi = 1 - \mathrm{e}^{-\mu V}.$$

Kun testataan n vesinäytettä, noudattaa negatiivisten näytteiden lukumäärä Y binomijakaumaa $\text{Bin}(n, \pi)$:

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

Kun vesinäytteitä otetaan, on määritettävä yhteen näytteeseen tarvittavan vesimäärän tilavuus. Jos V on liian suuri, kaikkiin näytteisiin tulee bakteereita ja kaikki näytteet ovat positiivisia. Jos taas V on liian pieni, on vaara, että kaikki näytteet ovat negatiivisia. Eräs tapa suojautua tätä ongelmaa vastaan on valmistaa erikokoisia näytteitä.

Tehtiin kaksi riippumatonta koetta. Ensimmäisessä kokeessa oli 40 näyteputkea, joiden tilavuus $V = 10\,\mathrm{ml}$. Toisessa kokeessa tarkistettiin 40 vesinäytettä, joiden tilavuus oli 1 ml. Kun $V = 10\,\mathrm{ml}$, saatiin 28 negatiivista ja 12 positiivista tulosta. Kun näytteen koko $V = 1\,\mathrm{ml}$, saatiin 37 negatiivista ja 3 positiivista tulosta.

Ensimmäisessä kokeessa uskottavuusfunktio on

$$L_1(\mu_1) = \pi_1^{28}(1-\pi_1)^{12},$$

missä $\pi_1 = \mathrm{e}^{-10\mu_1}$. Parametrin μ_1 suurimman uskottavuuden estimaatti on $\hat{\mu}_1 = 0.0357$. Toisesta kokeesta saadaan uskottavuusfunktioksi

$$L_2(\mu_2) = \pi_2^{37}(1-\pi_2)^{3},$$

missä $\pi_2 = \mathrm{e}^{-\mu_2}$. Parametrin μ_2 suurimman uskottavuuden estimaatti on $\hat{\mu}_2 = 0.0780$. Kokeisiin liittyvät logaritmoidut uskottavuusfunktiot ovat vastaavasti

$$l_1(\mu_1) = 28\log\pi_1 + 12\log(1-\pi_1)$$

ja

$$l_2(\mu_2) = 37\log\pi_2 + 3\log(1-\pi_2).$$

Jos oletetaan, että bakteerien määrä millilitrassa on molemmissa kokeissa keskimäärin sama, voidaan kokeet yhdistää. Kaikkiin 80 näytteeseen perustuvan yhdistetyn kokeen parametrin μ logaritmoitu uskottavuusfunktio on

$$l(\mu) = l_1(\mu) + l_2(\mu).$$

Tästä funktiosta maksimoimalla saatu suurimman uskottavuuden estimaatti $\tilde{\mu} = 0.04005$. Ensimmäisestä kokeesta saadaan logaritmoitu normitettu uskottavuusfunktio

$$r_1(\mu) = -280\mu + 12\log(1-\mathrm{e}^{-10\mu}) + 24.43$$

ja toisesta kokeesta

$$r_2(\mu) = -37\mu + 3\log(1-\mathrm{e}^{-\mu}) + 10.66.$$

Kun testataan hypoteesia

$$H_0\colon \mu_1 = \mu_2,$$

on uskottavuustestisuureen arvo

$$D = 2[l_1(\hat{\mu}_1) + l_2(\hat{\mu}_2) - l_1(\tilde{\mu}) - l_2(\tilde{\mu})] = -2[r_1(\tilde{\mu}) + r_2(\tilde{\mu})] = 1.24.$$

Hypoteesin testauksessa on yksi vapausaste, joten

$$\alpha_h \approx P(\chi_1^2 \geq 1.24) > 0.25.$$

Havainnot ovat sopusoinnussa H_0:n kanssa. □

9.3.4 Binomitodennäköisyyksien testaaminen

Oletetaan, että verrataan erilaisten hoitojen vaikutusta. Olkoon vertailtavana k erilaista hoitomenetelmää. Ensimmäistä hoitoa annettiin n_1:lle potilaalle, joista Y_1 parani. Vastaavasti i. hoito annettiin n_i:lle potilaalle, joista Y_i parani. Tällaisen hoitokokeen tulokset ovat Taulukossa 9.1 esitettyä muotoa.

Taulukko 9.1. Hoitokokeiden tulokset parantuneiden lukumäärän mukaan.

Käsittely	1	2	...	k
Parantuneiden lkm	Y_1	Y_2	...	Y_k
Epäonnistumisten lkm	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_k - Y_k$
Potilaiden lkm	n_1	n_2	...	n_k

Oletamme, että parantuneiden lukumäärä Y_i i. hoidossa noudattaa binomijakaumaa $\mathrm{Bin}(n_i, \pi_i)$ ja että lukumäärät $Y_1, \ldots, Y_k$ ovat toisistaan riippumattomat. Hoitojen tehoa voidaan vertailla onnistumistodennäköisyyksien $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ avulla. Perusmallissa on k tuntematonta parametria ja logaritmoitu uskottavuusfunktio on

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{k} y_i \log \pi_i + \sum_{i=1}^{k} (n_i - y_i) \log(1 - \pi_i).$$

Parametrin π_i suurimman uskottavuuden estimaatti on $\hat{\pi}_i = y_i/n_i$ ja $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$. Logaritmoidun uskottavuusfunktion maksimi on

$$l(\hat{\boldsymbol{\pi}}) = \sum y_i \log \frac{y_i}{n_i} + \sum (n_i - y_i) \log\Big(1 - \frac{y_i}{n_i}\Big).$$

Testataan nyt hypoteesi, että hoidoilla ei ole eroa:

$$H_0 \colon \pi_1 = \pi_2 = \cdots = \pi_k. \tag{9.3.5}$$

Hypoteesin H_0 mukaisessa mallissa on yksi tuntematon parametri, koska onnistumistodennäköisyyksien yhteinen arvo on tuntematon. Kun H_0 on tosi on logaritmoitu uskottavuusfunktio

$$l(\pi) = \sum y_i \log \pi + \sum (n_i - y_i) \log(1 - \pi), \tag{9.3.6}$$

missä π on hoitojen yhteinen onnistumistodennäköisyys. Logaritmoidusta uskottavuusfunktiosta (9.3.6) määritetty suurimman uskottavuuden estimaatti on $\tilde{\pi} = y/n$ ja $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}, \ldots, \tilde{\pi})$, missä $y = \sum_{i=1}^{k} y_i$ ja $n = \sum_{i=1}^{k} n_i$. Logaritmoidun uskottavuusfunktion (9.3.6) maksimi on

$$l(\tilde{\boldsymbol{\pi}}) = \sum_{i=1}^{k} y_i \log \tilde{\pi} + \sum_{i=1}^{k} (n_i - y_i) \log(1 - \tilde{\pi}).$$

Uskottavuustestisuure hypoteesin (9.3.5) testaamiseksi on

$$(9.3.7)\quad D = 2[l(\hat{\boldsymbol{\pi}}) - l(\tilde{\boldsymbol{\pi}})] = 2\left[\sum_{i=1}^{k} y_i \log \frac{y_i}{n_i\tilde{\pi}} + \sum_{i=1}^{k}(n_i - y_i)\log\frac{n_i - y_i}{n_i(1-\tilde{\pi})}\right].$$

Huomattakoon, että onnistumisten lukumäärien odotusarvot ovat $n_i\pi$ ja epäonnistumisten lukumäärien odotusarvot $n_i(1-\pi)$, jos H_0 on tosi. Vastaavasti $n_i\tilde{\pi}$ ja $n_i(1-\tilde{\pi})$ ovat näiden odotettujen frekvenssien estimaatit ja $y_1, \ldots, y_k$ ovat havaitut frekvenssit. Testisuure (9.3.7) vertailee siis havaittuja ja H_0:n vallitessa estimoituja frekvenssejä yli kaikkien luokkien ($2k$ kappaletta).

Jos hoidot ovat lääkkeen eri annosmääriä $a_1, a_2, \ldots, a_k$, niin voitaisiin testata hypoteesia

$$(9.3.8)\qquad H\colon \pi_i = 1 - \frac{1}{1+\mathrm{e}^{\alpha+\beta a_i}}, \qquad i = 1, \ldots, k.$$

Tämä hypoteesi olettaa, että paranemistodennäköisyys riippuu annoksesta logistisen mallin mukaan. Hypoteesin (9.3.8) spesifioimassa mallissa on kaksi tuntematonta parametria. Kun hypoteesi (9.3.8) oletetaan oikeaksi ja kirjoitetaan logaritmoitu uskottavuusfunktio parametrien α ja β funktiona, voidaan määrittää niiden suurimman uskottavuuden estimaatit $\hat{\alpha}$ ja $\hat{\beta}$. Niiden avulla voidaan laskea sitten estimaatit

$$\tilde{\pi}_i = 1 - \frac{1}{1+\mathrm{e}^{\hat{\alpha}+\hat{\beta}a_i}}, \qquad i = 1, \ldots, k.$$

Koska $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_k)$, niin uskottavuustestisuure on

$$D = 2\left[\sum y_i \log \frac{y_i}{n_i\tilde{\pi}_i} + \sum (n_i - y_i)\log\frac{n_i - y_i}{n_i(1-\tilde{\pi}_i)}\right].$$

Kun havaitaan $D = d$, niin

$$\alpha_h \approx P(\chi^2_{k-q} \geq d).$$

Hypoteesin (9.3.5) tapauksessa $q = 1$ ja hypoteesin (9.3.8) tapauksessa $q = 2$. Likiarvon voidaan oletaa olevan kohtuullisen tarkka, mikäli odotettujen frekvenssien estimaatit ovat kohtuullisen suuria.

Esimerkki 9.8 Tutkittiin erään ruuan lisäaineen mahdollisesti syöpää aiheuttavaa vaikutusta rotilla siten, että 44 rotalle annettiin ainetta pieni annos ja 44 rotalle suuri annos. Myöhemmin tutkittiin mahdollisesti kehittyneet kasvaimet. Tulokset ovat Taulukossa 9.2.

Olkoot Y_1 ja Y_2 niiden rottien lukumäärät, joilla oli kasvain. Nyt $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, missä $n_1 = n_2 = 44$ ja π_1 on todennäköisyys saada kasvain pienellä annoksella ja π_2 todennäköisyys saada kasvain suurella annoksella. Testataan hypoteesia

$$H_0\colon \pi_1 = \pi_2.$$

Taulukko 9.2. Kasvainten lukumäärät koeaineistossa rotilla lisäaineen annoksen mukaan.

Käsittely	Pieni annos	Suuri annos
Kasvain	4	14
Ei kasvainta	40	30
Yhteensä	44	44

H_0:n vallitessa kasvaintodennäköisyyden suurimman uskottavuuden estimaatti on

$$\tilde{\pi} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{4 + 14}{44 + 44} = \frac{9}{44}.$$

Odotettujen frekvenssien estimaatit ovat

$$n_1\tilde{\pi} = n_2\tilde{\pi} = 9 \qquad \text{ja} \qquad n_1(1 - \tilde{\pi}) = n_2(1 - \tilde{\pi}) = 35.$$

Lausekkeen (9.3.7) mukaan uskottavuussuhdetestisuureen arvo on

$$D = 2\big(4 \log \tfrac{4}{9} + 14 \log \tfrac{14}{9} + 40 \log \tfrac{40}{35} + 30 \log \tfrac{30}{35}\big) = 7.32.$$

Koska nyt

$$\alpha_h \approx P(\chi^2_{2-1} \geq 7.32) < 0.01,$$

koetulokset tukevat vahvasti käsitystä, että H_0: $\pi_1 = \pi_2$ ei pidä paikkaansa. □

9.3.5 Multinomitodennäköisyyksien testaaminen

Tarkastellaan koetta, jolla on k toisensa poissulkevaa tulosvaihtoehtoa T_1, …, T_k. Kun tehdään n riippumatonta toistoa, saadaan aineisto f_1, …, f_k, missä f_i on niiden kokeiden lukumäärä, joissa saadaan tulokseksi T_i.

Taulukko 9.3. Polkupyöräonnettomuuksien lukumäärä eräässä kaupungissa vuonna 2003 viikon päivän mukaan.

Päivä	ma	ti	ke	to	pe	la	su	Yhteensä
Onnettomuuksien lkm	28	26	23	30	33	21	12	174

Taulukossa 9.3 on vuoden aikana sattuneet polkupyöräonnettomuudet luokiteltuna viikonpäivän mukaan. Tässä esimerkissä $k = 7$ ja $n = \sum_{i=1}^{k} f_i = 174$. Frekvenssien todennäköisyydet saadaan multinomijakaumasta

$$p(f_1, \ldots, f_k) = \frac{n!}{f_1! f_2! \cdots f_k!} \pi_1^{f_1} \pi_2^{f_2} \cdots \pi_k^{f_k},$$

missä π_i on tulosvaihtoehdon T_i todennäköisyys. Parametrin $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ logaritmoitu uskottavuusfunktio on

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{k} f_i \log \pi_i,$$

missä $\sum_{i=1}^{k} \pi_i = 1$. Parametrin π_i suurimman uskottavuuden estimaatti on $\hat{\pi}_i = f_i/n$. Siksi logaritmoidun uskottavuusfunktion maksimi on

$$l(\hat{\boldsymbol{\pi}}) = \sum_{i=1}^{k} f_i \log\left(\frac{f_i}{n}\right).$$

Hypoteesi määrittää parametreille (tai osalle niistä) numeeriset arvot tai esittää ne tuntemattomien parametrien funktiona. Parametrien $\pi_1, \ldots, \pi_k$ suurimman uskottavuuden estimaatit lasketaan H_0:n vallitessa ja merkitään estimaatteja $\tilde{\pi}_1, \ldots, \tilde{\pi}_k$. Logaritmoidun uskottavuusfunktion maksimi H_0:n vallitessa on

$$l(\tilde{\boldsymbol{\pi}}) = \sum_{i=1}^{k} f_i \log \tilde{\pi}_j.$$

Edellä esitettyjen yleisten periaatteiden mukaan uskottavuustestisuure H_0:n testaamiseksi on muotoa

$$D = 2[l(\hat{\boldsymbol{\pi}}) - l(\tilde{\boldsymbol{\pi}})] = 2\sum_{i=1}^{k} f_i \log \frac{f_i}{e_i}, \tag{9.3.9}$$

missä $e_i = n\tilde{\pi}_i$ on odotetun frekvenssin estimaatti H_0:n vallitessa.

Koska $\sum_{i=1}^{k} \pi_i = 1$, niin perusmallissa on $k-1$ vapaata parametria. Taulukon 9.3 aineistolla voitaisiin esimerkiksi testata hypoteesia

$$H_0\colon \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = \pi_7.$$

H_0:n mukaan kaikkina viikon päivinä on sama onnettomuustodennäköisyys $\pi_i = \frac{1}{7}, i = 1, \ldots, 7$. Silloin siis todennäköisyydet on täysin määrätty, joten mallissa ei H_0:n vallitessa ole vapaita parametreja. Jos havaitaan $D = d$, niin

$$\alpha_h \approx P(\chi^2_{k-1} \geq d) = P(\chi^2_6 \geq d).$$

Ehkä realistisempi hypoteesi olisi olettaa, että viikon arkipäivinä ja viikonloppuna on erilainen onnettomuustodennäköisyys:

$$H\colon\ \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_v \quad \text{ja} \quad \pi_6 = \pi_7 = \pi_s.$$

Silloin $5\,\pi_v + 2\,\pi_s = 1$, joten hypoteesin vallitessa mallissa on yksi vapaa parametri. Silloin siis $q = 1$ ja $k - 1 - q = 5$.

Luokilla, joissa $e_i \approx 0$ ja $f_i \geq 1$, on suuri vaikutus testisuureeseen D. Tavallisesti käytetään peukalosääntöä, että e_i:n tulisi olla vähintään 5, jotta

χ^2-jakaumaan perustuva likiarvo olisi riittävän hyvä. Toinen tavanomainen multinomitodennäköisyyksien testaamiseen käytettävä testisuure on Pearsonin testisuure

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}. \tag{9.3.10}$$

Myös X^2 noudattaa asymptoottisesti χ^2-jakaumaa samoin vapausastein kuin D. Suureisiin (9.3.9) ja (9.3.10) perustuvia testejä sanotaan usein yhteensopivuustesteiksi.

9.3.6 Riippumattomuuden testaus kontingenssitaulukoissa

Oletetaan nyt, että tilastoyksiköt luokitellaan muuttujien X ja Y suhteen toisensa poissulkeviin luokkiin. Olkoon X esimerkiksi hiusten väri ja Y silmien väri. Hiukset luokitellaan vaaleisiin (Va) tummiin (T) ja punaisiin (P) sekä silmät sinisiin (S), ruskeisiin (R) ja vihreisiin (Vi). Jos valitaan n henkilön otos, saadaan Taulukon 9.4 kaltainen aineisto. Tällaista kahden muuttujan

Taulukko 9.4. Hiusten ja silmien värin välinen ristiintaulukko.

		Silmien väri			
		S	R	Vi	Yhteensä
Hiusten väri	Va	f_{11}	f_{12}	f_{13}	$f_{1.}$
	T	f_{21}	f_{22}	f_{23}	$f_{2.}$
	P	f_{31}	f_{32}	f_{33}	$f_{3.}$
Yhteensä		$f_{.1}$	$f_{.2}$	$f_{.3}$	n

suhteen esitettyä frekvenssitaulukkoa kutsutaan kontingenssitaulukoksi tai ristiintaulukoksi. Taulukossa 9.4 hiusten väriä kutsutaan rivimuuttujaksi ja simien väriä sarakemuuttujaksi.

Taulukko 9.5. $I \times J$-kontingenssitaulukko.

	S_1	S_2	...	S_J	Yhteensä
R_1	f_{11}	f_{12}	...	f_{1J}	$f_{1.}$
R_2	f_{21}	f_{22}	...	f_{2J}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R_I	f_{I1}	f_{I2}	...	f_{IJ}	$f_{I.}$
Yhteensä	$f_{.1}$	$f_{.2}$	...	$f_{.J}$	n

Taulukossa 9.5 on kaikkiaan $k = IJ$ solua. Olkoon π_{ij} todennäköisyys, että havainto, osuu soluun (i, j). Kun koetta toistetaan n kertaa (tai tehdään

n:n alkion satunnaisotos), noudattavat frekvenssit f_{ij} multinomijakaumaa. Parametrivektorin $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \ldots, \pi_{IJ})$ logaritmoitu uskottavuusfunktio on

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log \pi_{ij},$$

missä $\sum\sum \pi_{ij} = 1$. Tässä tilanne on aivan samanlainen kuin multinomijakauman tapauksessa, paitsi että nyt käytetään kaksinkertaista indeksointia.

Todennäköisyyksiä π_{ij} koskevat hypoteesit H testaan uskottavuustestisuureella

$$D = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log \frac{f_{ij}}{e_{ij}},$$

missä $e_{ij} = n\tilde{\pi}_{ij}$ on luokan $R_i S_j$ odotetun frekvenssin estimaatti ja $\tilde{\pi}_{ij}$ on solutodennäköisyyden π_{ij} suurimman uskottavuuden estimaatti H_0:n vallitessa.

Riippumattomuushypoteesi

Esimerkki 9.9 Usein ollaan kiinnostuneita siitä, onko rivi- ja sarakemuuttujien välillä riippuvuutta vai ovatko muuttujat riippumattomat. Ovatko esimerkiksi silmien ja hiusten väri toisistaan riippumattomat? Riippumattomuushypoteesi voidaan lausua seuraavasti:

$$H_0\colon P(R_i S_j) = P(R_i)\, P(S_j) \qquad \text{kaikilla } i \text{ ja } j. \tag{9.3.11}$$

$P(R_i S_j) = \pi_{ij}$ on todennäköisyys, että tilastoyksikkö sattuu soluun (i, j). Vastaavasti

$$P(R_i) = \pi_{i1} + \cdots + \pi_{iJ} = \pi_{i\cdot}$$

on todennäköisyys, että tilastoyksikkö sattuu rivimuuttujan i. luokkaan. Todennäköisyys, että tilastoyksikkö kuuluu sarakemuuttujan j. luokkaan, on

$$P(S_j) = \pi_{1j} + \cdots + \pi_{Ij} = \pi_{\cdot j}.$$

Hypoteesi (9.3.11) voidaan lausua siis seuraavasti:

$$H_0\colon \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}\,; \qquad 1 \le i \le I, \quad 1 \le j \le J.$$

Silloin tuntemattomat parametrit ovat

$$\pi_{i\cdot} = P(R_i), \qquad 1 \le i \le I$$

ja

$$\pi_{\cdot j} = P(S_j), \qquad 1 \le j \le J.$$

Koska $\sum \pi_{i\cdot} = 1$ ja $\sum \pi_{\cdot j} = 1$, niin funktionaalisesti riippumattomien parametrien lukumäärä $q = (I-1) + (J-1)$. Siksi H_0:n testaamisessa vapausasteiden lukumäärä on

$$(k-1) - q = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1).$$

□

Koska H_0:n vallitessa $\pi_{ij} = \pi_{i.}\pi_{.j}$, niin logaritmoitu uskottavuusfunktio on

$$\begin{aligned}\sum_i \sum_j f_{ij} \log \pi_{ij} &= \sum_i \sum_j f_{ij}(\log \pi_{i.} + \log \pi_{.j}) \\ &= \sum_i \Big[\Big(\sum_j f_{ij} \Big) \log \pi_{i.} \Big] + \sum_j \Big[\Big(\sum_i f_{ij} \Big) \log \pi_{.j} \Big] \\ &= \sum_i f_{i.} \log \pi_{i.} + \sum_j f_{.j} \log \pi_{.j}. \qquad (9.3.12)\end{aligned}$$

Kun funktio (9.3.12) maksimoidaan rajoitteiden $\sum_i \pi_{i.} = 1$ ja $\sum_j \pi_{.j} = 1$ vallitessa, saadaan suurimman uskottavuuden estimaateiksi (H_0:n vallitessa)

$$\tilde{\pi}_{i.} = \frac{f_{i.}}{n} \qquad \text{ja} \qquad \tilde{\pi}_{.j} = \frac{f_{.j}}{n}.$$

Siksi odotettujen frekvenssien estimaatit H_0:n vallitessa ovat

$$e_{ij} = n\tilde{\pi}_{ij} = n\tilde{\pi}_{i.}\tilde{\pi}_{.j} = \frac{f_{i.}f_{.j}}{n}. \qquad (9.3.13)$$

Esimerkki 9.10 Tarkastellaan nyt taulukon 9.6 aineistoa, joka on saatu Yhdysvaltain vuoden 1991 yleisestä väestökyselystä (Agresti 1996, s. 31). Taulukossa on annettu havaitut frekvenssit ja sulkeissa odotettujen frekvenssien estimaatit. Kaavan (9.3.13) mukaan esimerkiksi

Taulukko 9.6. Puoluekanta sukupuolen mukaan.

	Puoluekanta			
Sukupuoli	Demokraatti	Riippumaton	Republikaani	Yhteensä
Naiset	279	73	235	577
	(261.4)	(70.7)	(244.9)	
Miehet	165	47	191	403
	(182.6)	(49.3)	(171.1)	
Yhteensä	444	120	416	980

$$e_{11} = \frac{577 \cdot 444}{980} = 261.4 \qquad \text{ja} \qquad e_{23} = \frac{403 \cdot 416}{980} = 171.1.$$

Uskottavuustestisuureen arvoksi saadaan 7.0026. Riippumattomuushypoteesin vallitessa D noudaattaa likimain χ^2-jakaumaa vapausastein $(I-1)(J-1) = (2-1)(3-1) = 2$, joten $\alpha_h \approx P(\chi^2_2 \geq 7.0026) = 0.0302$.

Pearsonin χ^2-suure

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

on tavallisimmin käytetty testisuure riippumattomuuden testaamisessa. Sekä D että Pearsonin χ^2-suure noudattavat H_0:n vallitessa asymptoottisesti samaa jakaumaa. Esimerkiksi taulukon 9.6 aineistosta $X^2 = 7.0095$ ja silloin $\alpha_h \approx P(\chi^2_2 \geq 7.0095) = 0.0301$. Tässä tapauksessa D ja Pearsonin χ^2-suure antavat melko tarkkaan saman tuloksen, koska $n = 980$ on riittävän suuri.

Sukupuolen ja puoluekannan välillä on tilastollisesti merkitsevä riippuvuus. Vertailemalla havaittuja frekvenssejä estimoituihin odotettuihin frekvensseihin saadaan käsitys tämän riippuvuuden luonteesta. Koska suurilla frekvenssien f_{ij} arvoilla myös erotus $f_{ij} - e_{ij}$ saa nollahypoteesin vallitessa itseisarvoltaan suurempia arvoja kuin pienillä frekvensseillä, ei raakaresiduaali $f_{ij} - e_{ij}$ ole riittävä poikkeamien mitta. Riippumattomuuden tarkastelussa käyttökelpoiset soluresiduaalit ovat muotoa

$$\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}(1 - \tilde{\pi}_{i.})(1 - \tilde{\pi}_{.j})}}.$$

□

9.4 Testiteoriaa

9.4.1 Ongelman määrittely

Tämä alaluku on lyhyt johdatus tilastolliseen testiteoriaan, joka perustuu testin voimakkuuden käsitteeseen. Testin voimakkuus luonnehtii testisuureen herkkyyttä tunnistaa vaihtoehtoinen hypoteesi. Nyt hypoteesille H_0 (nollahypoteesi) asetetaan eksplisiittisesti *vaihtoehtoinen hypoteesi* H_1, joka valitaan sen mukaan, millaisia poikkeamia H_0:sta halutaan tunnistaa. *Testimenettely* (testaus) on sääntö, joka havaintojen perusteella valitsee toisen ja vain toisen hypoteeseista H_0 ja H_1.

Teknisesti hypoteesit konstruoidaan siten, että tehdään *parametriavaruuden* Θ *jako* kahteen osajoukkoon Θ_0 ja Θ_1 siten, että $\Theta = \Theta_0 \cup \Theta_1$ ($\Theta_0 \cap \Theta_1 = \emptyset$). Nyt siis $\Theta_1 = \Theta \setminus \Theta_0 = \Theta_0^c$, joten Θ_1 on Θ_0:n komplementti Θ:ssa. Näitä vaihtoehtoja merkitään tavallisesti

$$H_0 \colon \theta \in \Theta_0 \qquad \text{vastaan} \qquad H_1 \colon \theta \in \Theta_1.$$

Jos $\Theta = \mathbb{R}$ on reaalilukujen joukko, niin tyypillisiä hypoteeseja ovat esimerkiksi

$$\begin{cases} H_0 \colon \theta = 0 \\ H_1 \colon \theta \neq 0 \end{cases} \qquad \begin{cases} H_0 \colon \theta \leq 1 \\ H_1 \colon \theta > 1. \end{cases}$$

Ensimmäisessä tapauksessa $\Theta_0 = \{0\}$ on yksi piste (H_0 yksinkertainen) ja $\Theta_1 = \{\, \theta \mid \theta \neq 0 \,\}$. Toisessa tapauksessa molemmat hypoteesit ovat yhdistettyjä: $\Theta_0 = \{\, \theta \mid \theta \leq 1 \,\}$, $\Theta_1 = \{\, \theta \mid \theta > 1 \,\}$.

Testimenettely on itse asiassa *otosavaruuden S jako*. Olkoon T valittu testisuure ja $x \in S$ jokin havainto. Testisuureen T arvon $T(x)$ perusteella valitaan joko H_0 tai H_1. Hylkäysalue $C \subset S$ määritellään siten, että H_0

hylätään kaikilla $x \in C$. H_0 hyväksytään, jos $x \notin C$. Hylkäysaluetta kutsutaan myös testin *kriittiseksi alueeksi*. Vastaavasti H_0:n hyväksymisalue on C:n komplementti C^c, joten $S = C \cup C^c$.

Testi ei ole kuitenkaan erehtymätön. Voi sattua, että $x \in C$, vaikka $\theta \in \Theta_0$. Silloin syntyy *I lajin virhe*: H_0 hylätään, vaikka se on tosi. Vastaavasti voi sattua, että $x \in C^c$, vaikka $\theta \in \Theta_1$. Tässä tapauksessa tehdään *II lajin virhe*: H_0 hyväksytään, vaikka se ei ole tosi. Tavoitteena on valita kriittinen alue siten, että virhetodennäköisyydet ovat mahdollisimman pienet

9.4.2 Testin voimakkuus

Testin keskeiset ominaisuudet voidaan luonnehtia *voimakkuusfunktiota*

$$\begin{aligned}\gamma(\theta) &= P(X \in C;\, \theta) \\ &= P(H_0 \text{ hylätään, kun parametrin arvo on } \theta)\end{aligned}$$

käyttäen. Testin *merkitsevyystaso* α on

$$\alpha = \sup_{\theta \in \Theta_0} \gamma(\theta),$$

joka on siis I lajin virheen maksimi H_0:n vallitessa.

Esimerkki 9.11 Olkoon X havainto normaalijakaumasta $\mathrm{N}(\theta, 1)$. Testataan havainnon $X = x$ avulla hypoteesit

$$\begin{aligned}H_0 &\colon \theta \leq 0, \\ H_1 &\colon \theta > 0.\end{aligned}$$

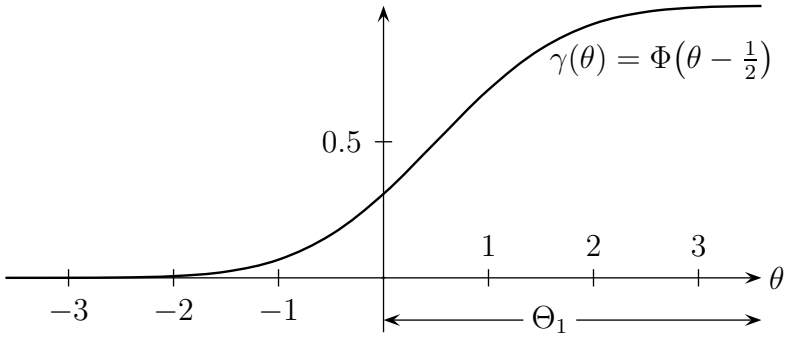
Olkoon kriittinen alue $C = \left\{ x \;\middle|\; x > \frac{1}{2} \right\}$. Silloin testin voimakkuusfunktio on

$$\begin{aligned}\gamma(\theta) &= P\big(X > \tfrac{1}{2};\, \theta\big) \\ &= P\big(X - \theta > \tfrac{1}{2} - \theta;\, \theta\big) \\ &= 1 - \Phi\big(\tfrac{1}{2} - \theta\big) = \Phi\big(\theta - \tfrac{1}{2}\big),\end{aligned}$$

missä $X - \theta \sim \mathrm{N}(0, 1)$. Tässä tapauksessa γ on θ:n aidosti kasvava funktio ja $\gamma(\theta) \to 1$, kun $\theta \to \infty$ ja ja $\gamma(\theta) \to 0$, kun $\theta \to -\infty$. Testin merkitsevyystaso

$$\begin{aligned}\alpha &= \max_{\theta \in \Theta_0} \Phi\big(\theta - \tfrac{1}{2}\big) \\ &= \Phi\big(-\tfrac{1}{2}\big) \approx 0.31.\end{aligned}$$

□



Kuvio 9.1. Voimakkuusfunktio $\gamma(\theta)$, kun $X \sim \mathrm{N}(\theta, 1)$ ja kriittinen alue $C = \{\, x \mid x > \frac{1}{2} \,\}$.

Asettamamme tavoitteen mukaisesti pitäisi voimakkuuden $\gamma(\theta)$ olla mahdollisimman suuri, kun $\theta \in \Theta_1$. Vastaavasti voimakkuuden $\gamma(\theta)$ tulisi olla mahdollisimman pieni, kun $\theta \in \Theta_0$. Nämä vaatimukset ovat kuitenkin keskenään ristiriitaiset. Kun kriittistä aluetta kasvatetaan, kasvaa I lajin virhe. Kun hyväksymisaluetta kasvatetaan, kasvaa II lajin virhe. Eräs keino ratkaista tämä pulma, on kiinnittää ensin merkitsevyystaso α ja valita sitten näistä saman merkitsevyystason testeistä se, joka on voimakkain alueella $\theta \in \Theta_1$.

Tämän klassisen lähestymistavan esittivät Jerzy Neyman ja Egon S. Pearson. Tämä Neymanin ja Pearsonin lähestymistapakaan ei välttämättä johda yksikäsitteiseen ratkaisuun. Olkoon kahdella testillä voimakkuusfunktiot $\gamma_1(\theta)$ ja $\gamma_2(\theta)$ ja sama merkitsevyystaso. Silloin ei välttämättä pidä paikkansa, että $\gamma_1(\theta) \geq \gamma_2(\theta)$ kaikilla $\theta \in \Theta_1$ tai $\gamma_2(\theta) \geq \gamma_1(\theta)$ kaikilla $\theta \in \Theta_1$. Kumpikaan testi ei ole välttämättä tasaisesti voimakkaampi kuin toinen. Kuitenkin useissa tärkeissä testitilanteissa voidaan löytää tällainen *tasaisesti voimakkain testi*.

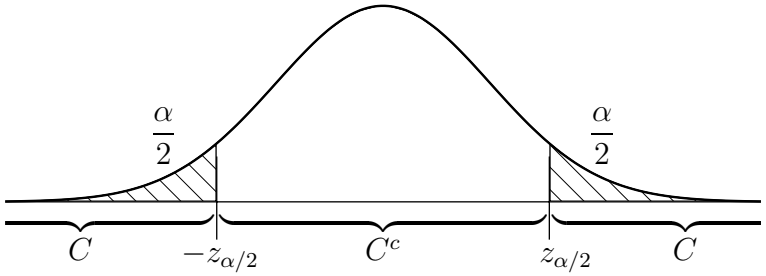
Esimerkki 9.12 Olkoon $X_1, \ldots, X_n$ otos normaalijakaumasta $\mathrm{N}(\theta, \sigma^2)$, missä σ^2 on tunnettu. Halutaan testata H_0: $\theta = \theta_0$ vastaan $\theta \neq \theta_0$. Nyt tiedetään, että $\overline{X} \sim \mathrm{N}(\theta, \sigma^2/n)$, joten testisuure

$$Z = \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1),$$

kun H_0 tosi. Hylätään H_0, jos

$$|Z| = \left| \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

Kriittinen alue on siis $C = \{\, Z \mid |Z| > z_{\alpha/2} \,\}$.



Testin merkitsevyystaso on

$$\alpha = P(H_0 \text{ hylätään; } \theta = \theta_0) = P(|Z| > z_{\alpha/2};\, \theta = \theta_0).$$

Seuraavassa merkitään $P(Z \in C;\, \theta) = P_\theta(Z \in C)$. Voimakkuusfunktio on

$$\begin{aligned}
\gamma(\theta) &= P(Z \in C;\, \theta) \\
&= P_\theta(|Z| > z_{\alpha/2}) = P_\theta(Z < -z_{\alpha/2}) + P_\theta(Z > z_{\alpha/2}) \\
&= P_\theta\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) + P_\theta\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > z_{\alpha/2}\right) \\
&= P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} + \frac{\theta - \theta_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}\right) + P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} + \frac{\theta - \theta_0}{\sigma/\sqrt{n}} > z_{\alpha/2}\right) \\
&= P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} < -z_{\alpha/2} - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) + P_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} > z_{\alpha/2} - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-z_{\alpha/2} - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(z_{\alpha/2} - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

Merkitsevyystaso on

$$\gamma(\theta_0) = \Phi(-z_{\alpha/2}) + 1 - \Phi(z_{\alpha/2}) = 2\,\Phi(-z_{\alpha/2}) = 2 \cdot \frac{\alpha}{2} = \alpha.$$

Voimme todeta, että $\gamma(\theta) \to 1$, kun $\theta \to \infty$ tai $\theta \to -\infty$. □

9.4.3 Testien konstruointi

Olkoon X havainto jakaumasta, jonka tiheysfunktio on $f(x;\theta)$. Testataan nyt vaihtoehtoisia yksinkertaisia hypoteeseja

$$H_0\colon \theta = \theta_0 \qquad \text{ja} \qquad H_1\colon \theta = \theta_1. \tag{9.4.1}$$

Kun havainto $X = x$ on saatu, voidaan laskea *uskottavuussuhde*

$$\lambda(x) = \frac{L(\theta_0; x)}{L(\theta_1; x)},$$

jonka valitsemme testisuureeksi. Kun $\lambda(x)$ on suuri, olemme taipuvaisia hyväksymään H_0:n. Jos taas $\lambda(x)$ on pieni, valitsemme H_1:n.

Neymanin ja Pearsonin lähestymistavan mukaan kiinnitetään ensin merkitsevyystaso α. Tämän perusteella voimme valita λ:n kriittisen arvon siten, että

$$P[\lambda(x) \leq \lambda_\alpha;\ \theta_0] = \alpha. \tag{9.4.2}$$

Testin hylkäysalue on siis

$$C = \{\, x \mid \lambda(x) \leq \lambda_\alpha \,\}. \tag{9.4.3}$$

Lause 9.1 (Neymanin ja Pearsonin apulause) *Kun testataan parametria θ koskevia hypoteeseja* (9.4.1)*, niin testi* (9.4.3) *on voimakkain kaikista merkitsevyystasoa α olevista testeistä. Merkitsevyystaso α on määritelty identiteetillä* (9.4.2)*.*

Todistus. Olkoon D mikä tahansa toinen kriittinen alue (toinen testi), jonka merkitsevyystaso on korkeintaan α. Silloin

$$\alpha = \int_C f(x;\theta_0)\,\mathrm{d}x \geq \int_D f(x;\theta_0)\,\mathrm{d}x,$$

joten

$$\int_{C\cap D^c} f(x;\theta_0)\,\mathrm{d}x \ \geq \int_{D\cap C^c} f(x;\theta_0)\,\mathrm{d}x. \tag{9.4.4}$$

Epäyhtälö (9.4.4) seuraa siitä, että

$$\int_{C\cap D^c} = \int_C - \int_{C\cap D} \qquad \text{ja} \qquad \int_{D\cap C^c} = \int_D - \int_{C\cap D}.$$

Jos $x \in C \cap D^c$, niin $x \in C$ ja silloin

$$f(x;\theta_1)\lambda_\alpha \geq f(x;\theta_0).$$

Jos taas $x \in D \cap C^c$, niin $x \in C^c$ ja

$$f(x;\theta_0) > f(x;\theta_1)\lambda_\alpha.$$

Siksi

$$\begin{aligned} \lambda_\alpha \int_{C\cap D^c} f(x;\theta_1)\,\mathrm{d}x \ &\geq \int_{C\cap D^c} f(x;\theta_0)\,\mathrm{d}x \\ &\geq \int_{D\cap C^c} f(x;\theta_0)\,\mathrm{d}x \ \geq\ \lambda_\alpha \int_{D\cap C^c} f(x;\theta_1)\,\mathrm{d}x, \end{aligned} \tag{9.4.5}$$

missä toiseksi viimeinen epäyhtälö on aito, elleivät C ja D ole samat. Kun epäyhtälö (9.4.5) jaetaan puolittain λ_α:lla ja epäyhtälön molemmille puolille lisätään integraali $\int_{C\cap D} f(x;\theta_1)\,\mathrm{d}x$, saadaan

$$\int_C f(x;\theta_1)\,\mathrm{d}x \geq \int_D f(x;\theta_1)\,\mathrm{d}x.$$

Koska määritelmän mukaan $\gamma_C(\theta_1) = \int_C f(x;\theta_1)\,\mathrm{d}x$ ja $\gamma_D(\theta_1) = \int_D f(x;\theta_1)\,\mathrm{d}x$, niin testi (9.4.3) on voimakkain kaikista testeistä, joiden merkitsevyystaso on korkeintaan α. □

Esimerkki 9.13 Olkoon $X_1,\ldots,X_n$ otos normaalijakaumasta $\mathrm{N}(\mu,1)$. Testataan kahta yksinkertaista hypoteesia H_0: $\mu = 1$ vastaan H_1: $\mu = 2$. Etsimme nyt parhaan kriittisen alueen C, kun testin merkitsevyystaso α on kiinnitetty. Määritämme ne otosavaruuden S pisteet, joissa

$$\lambda(x_1,\ldots,x_n) = \frac{L(1;x_1,\ldots,x_n)}{L(2;x_1,\ldots,x_n)} \leq \lambda_\alpha,$$

missä $P[\lambda(x_1,\ldots,x_n) \leq \lambda_\alpha] = \alpha$. Koska $L(1) = c\cdot\exp\bigl[-\frac{1}{2}\sum_{i=1}^n (x_i-1)^2\bigr]$ ja $L(2) = c\cdot\exp\bigl[-\frac{1}{2}\sum_{i=1}^n (x_i-2)^2\bigr]$, niin

$$\log\lambda(x) = -\frac{1}{2}\sum_{i=1}^n (x_i-1)^2 + \frac{1}{2}\sum_{i=1}^n (x_i-2)^2 \leq \log\lambda_\alpha$$

Tästä seuraa, että

$$\sum_{i=1}^n (x_i-2)^2 - \sum_{i=1}^n (x_i-1)^2 \leq 2\lambda_\alpha. \tag{9.4.6}$$

Koska $\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu)$, niin

$$\sum_{i=1}^n (x_i-2)^2 - \sum_{i=1}^n (x_i-1)^2 = n(\overline{x}-2)^2 - n(\overline{x}-1)^2 = n(-2\overline{x}+3).$$

Siksi epäyhtälö (9.4.6) voidaan kirjoittaa muodossa

$$\overline{x} \geq \frac{\lambda_\alpha}{n} + \frac{3}{2} = c_\alpha.$$

Koska H_0:n vallitessa $\overline{X} \sim \mathrm{N}(1, 1/\sqrt{n})$, niin määritetään c_α siten, että

$$P(\overline{X} \geq c_\alpha) = P[Z \geq \sqrt{n}(c_\alpha - 1)] = \alpha,$$

missä $Z = \sqrt{n}(\overline{X}-1) \sim \mathrm{N}(0,1)$ H_0:n vallitessa.

Oletetaan esimerkiksi, että $n = 16$ ja testin merkitsevyystaso $\alpha = 0.05$. Määritetään siis $c_{0.05}$ siten, että

$$P[Z \geq 4(c_{0.05} - 1)] = 0.05.$$

Silloin $4(c_{0.05} - 1) = 1.645$, joten

$$c_{0.05} = \frac{1.645}{4} + 1 = 1.41125.$$

Testin hylkäysalue merkitsevyystasolla $\alpha = 0.05$ on

$$C = \left\{ (x_1, \ldots, x_n) \;\middle|\; \overline{x} = \frac{1}{16} \sum_{i=1}^{16} x_i \geq 1.41125 \right\}.$$

□