Sisältö

1 Johdanto **1**

1.1 Todennäköisyys ja tilastotiede 1
1.2 Havaitut frekvenssit ja empiiriset jakaumat 1
1.3 Todennäköisyysmallit 3
1.3.1 Satunnaiskoe 3
1.3.2 Otosavaruudet, tapahtumat ja joukko-operaatiot 4
1.3.3 Todennäköisyys 8
1.3.4 Äärettömät otosavaruudet 11
1.3.5 Todennäköisyyden tulkinnat 12
1.4 Ehdollinen todennäköisyys 14
1.4.1 Ehdollisen todennäköisyyden frekvenssitulkinta 15
1.4.2 Kertolaskusääntö 15
1.4.3 Riippumattomuus 15
1.5 Odotetut frekvenssit 15
Yhteenveto 16
Harjoituksia 17

2 Todennäköisyys, satunnaismuuttuja ja perustuloksia **21**

2.1 Todennäköisyyden ominaisuuksia 21
2.2 Symmetriaan perustuva todennäköisyys 24
2.3 Aksiomaattinen lähestymistapa 25
2.3.1 Äärellinen additiivisuus 25
2.3.2 Todennäköisyyden yleiset aksioomat 25
2.4 Kombinatoriikkaa 27
2.4.1 Summa- ja tuloperiaate 27
2.4.2 Valinta järjestyksessä 27
2.4.3 Osajoukon valinta 28
2.4.4 Otanta palauttaen, kun järjestystä ei oteta huomioon . 30
2.4.5 Kombinatoriikan merkintöjä ja identiteettejä 31
2.4.6 Binomilause, hypergeometrinen identiteetti ja multinomilause 32
2.5 Satunnaismuuttuja 33
2.6 Satunnaismuuttujan jakauma 35
2.6.1 Kertymäfunktio 37
2.6.2 Satunnaismuuttujan tiheysfunktio 40

2.7 Otanta palauttamatta 44
2.7.1 Hypergeometrinen jakauma 45
2.7.2 Tarkistusotanta teollisuudessa 46
2.8 Otanta palauttaen 46
2.9 Binomijakauma 48
2.9.1 Binomijakauma hypergeometrisen jakauman likiarvona 49
Yhteenveto 50
Harjoituksia 52

3 Ehdollinen todennäköisyys ja riippumattomuus 57
3.1 Ehdollinen todennäköisyys 57
3.1.1 Tulosääntö, kokonaistodennäköisyys ja Bayesin kaava 58
3.1.2 Riippumattomuus 60
3.1.3 Joukko-oppi ja todennäköisyys 64
3.2 Ehdolliset jakaumat 64
3.3 Yleinen tulokaava ja Bayesin lause 65
3.3.1 Yleinen tulokaava 65
3.3.2 Bayesin lause 68
3.3.3 Peräkkäisotanta 71
3.3.4 Useiden tapahtumien unionin todennäköisyys 72
Yhteenveto 74
Harjoituksia 75

4 Satunnaismuuttujien tunnusluvut ja riippumattomuus 77
4.1 Odotusarvo, varianssi ja kovarianssi 77
4.1.1 Odotusarvo 77
4.1.2 Ehdollinen odotusarvo 84
4.1.3 Varianssi 85
4.1.4 Kovarianssi ja korrelaatio 87
4.2 Satunnaismuuttujan funktio 88
4.3 Satunnaismuuttujien identtisyys 89
4.4 Satunnaismuuttujien riippumattomuus 90
4.4.1 Kaksi satunnaismuuttujaa 91
4.4.2 Useita satunnaismuuttujia 93
4.5 Suurten lukujen laki 93
4.6 Generoivat funktiot ja momentit 96
4.6.1 Momentit 96
4.6.2 Momenttifunktio 96
4.6.3 Todennäköisyydet generoiva funktio (tgf) 99
4.7 Kokeiden yhdistäminen ja tulomallit 100
Yhteenveto 102
Harjoituksia 104

5 Diskreettejä yksiulotteisia jakaumia **107**
5.1 Diskreetti satunnaismuuttuja 107
5.2 Bernoullin kokeet ja binomijakauma 109
5.2.1 Jakauman symmetria 114
5.3 Odotusaikojen jakaumat 115
5.3.1 Odotusajat Bernoullin kokeissa 115
5.3.2 Geometrinen jakauma ja negatiivinen binomijakauma . 118
5.3.3 Odotusajat peräkkäisotannassa 121
5.3.4 Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma 123
5.3.5 Tasajakauma 125
5.4 Poissonin jakauma 125
5.5 Poissonin prosessi 131
5.5.1 Laskuriprosessi 131
5.5.2 Poissonin prosessin määritely 132
5.5.3 Satunnaistapahtumat tila-avaruudessa 134
5.5.4 Symmetrinen jakauma 135
Yhteenveto 136
Harjoituksia 138

6 Jatkuvat jakaumat **141**
6.1 Jatkuvat satunnaismuuttujat 141
6.2 Tasajakauma ja eksponenttijakauma 148
6.2.1 Tasajakauma 148
6.2.2 Eksponenttijakauma 150
6.2.3 Elinaikajakauma 152
6.3 Gammajakauma ja χ^2-jakauma 153
6.4 Normaalijakauma 155
6.4.1 Standardimuotoinen normaalijakauma 155
6.4.2 Yleinen normaalijakauma 157
6.5 Muuttujien vaihto 160
6.5.1 Muunnos kertymäfunktio avulla 160
6.5.2 Muunnos tiheysfunktion avulla 161
6.5.3 Normaalimuuttujan muunnokset 164
6.6 Satunnaismuuttujan odotusarvo 165
6.6.1 Momentifunktio ja momentit 167
Yhteenveto 169
Harjoituksia 171

7 Moniuloteiset jakaumat **175**
7.1 Kaksiulotteiset jakaumat 175
7.1.1 Reunajakaumat ja ehdolliset jakaumat 180
7.1.2 Ehdollisen odotusarvon ominaisuuksia 186
7.1.3 Hierarkkiset mallit ja yhdistetyt jakaumat 189
7.1.4 Kaksiulotteinen Bernoullin jakauma 191

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 $x_1,\ x_2,\ \ldots,\ x_n$ muodostavat havaintoaineiston. Lukujen $x_1,\ x_2,\ \ldots,\ x_n$ (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.

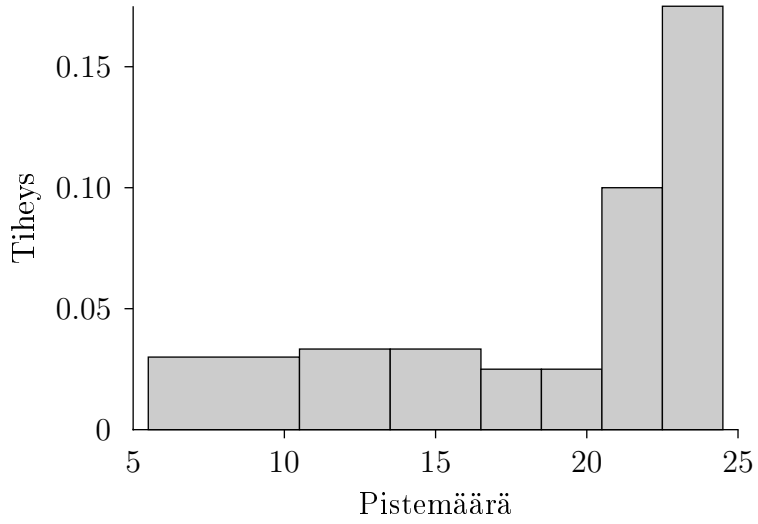
Esimerkki 1.2 Kurssin 1. välikokeen pistemäärät $x_i, 1 \leq i \leq 20$ olivat

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 kruuna (R) ja klaava (L), joten otosavaruus $\Omega = \{\mathrm{L}, \mathrm{R}\}$ ja $|\Omega| = 2$.

Heitetään lanttia, kunnes saadaan ensimmäinen kruunu. Silloin otosavaruus

$$\Omega = \{\mathrm{R}, \mathrm{LR}, \mathrm{LLR}, \mathrm{LLLR}, \ldots\}$$

ja $|\Omega| = \infty$. Jos tapahtuma A on 'enintään kaksi klaavaa ennen 1. kruunaa', niin $A = \{\mathrm{R}, \mathrm{LR}, \mathrm{LLR}\}$. □

Esimerkki 1.4 Tarkastellaan laitteen kestoa. Jokainen positiivinen reaaliluku voidaan tulkita kestoajaksi. Silloin

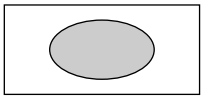
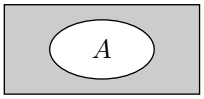
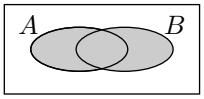
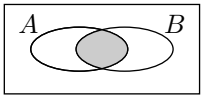
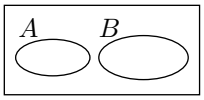
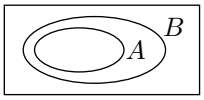
$$\Omega = \{\, \omega \in \mathbb{R} \mid \omega > 0 \,\}.$$

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

1.3.2 Otosavaruudet, tapahtumat ja 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

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	$A \subset B$	

(Ω:sta) satunnaisesti pisteen. Jokainen suorakaiteen piste on alkeistapaus. Jokainen suorakaiteen osa-alue on tapahtuma.

Taulukossa 1.1 on esitetty joukko-opilliset operaatiot *komplementti, yhdiste* ja *leikkaus.* Nämä operaatiot toteuttavat monia käyttökelpoisia ominaisuuksia, kuten esimerkiksi

$$(A^c)^c = A, \quad A \cup A^c = \Omega, \quad A \cap A^c = \emptyset.$$

Yksinkertaista, mutta todennäköisyyslaskennassa hyödyllistä relaatiota $(A^c)^c = A$ kutsutaan *kaksinkertaisen komplementin säännöksi.* Keskeisiä joukko-opin laskusääntöjä ovat *vaihdantalait*

$$A \cup B = B \cup A, \quad A \cap B = B \cap A$$

liitäntälait

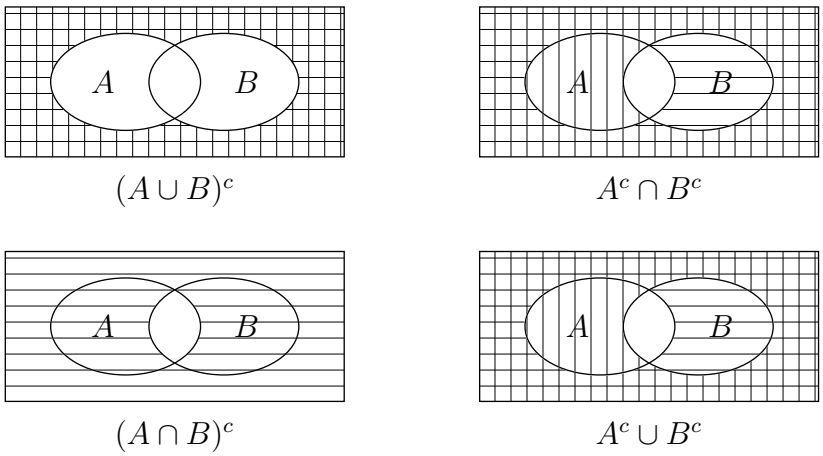
$$A \cup (B \cup C) = (A \cup B) \cup C, \quad A \cap (B \cap C) = A \cap (B \cap C)$$

osittelulait

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{ja}$$

De Morganin lait

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



Kuvio 1.2. De Morganin lait.

Huomaa, että

$$A \cap (B \cup C) \neq (A \cap B) \cup C,$$

paitsi erikoistapauksissa. Lauseke $A \cap B \cup C$ ei ole siis hyvin määritelty, vaan tarvitaan sulut osoittamaan, kummasta tapahtumasta on kyse.

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

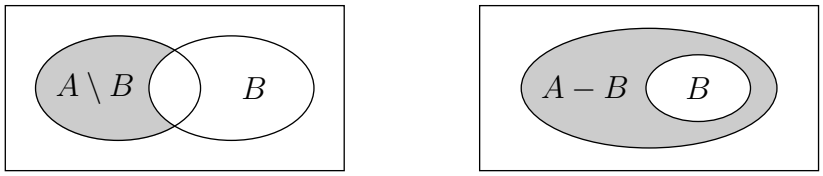
$$A \setminus B = A \cap B^c = \{\, \omega \mid \omega \in A \text{ ja } \omega \notin B \,\}.$$

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

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

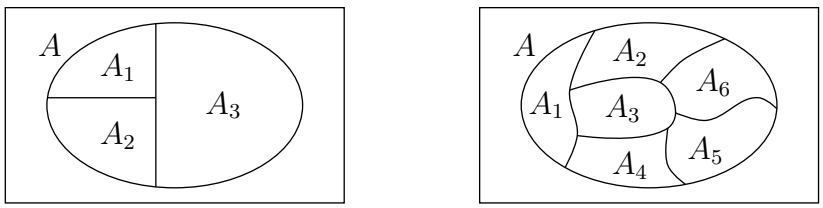
ja

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



Kuvio 1.3. Joukkojen erotus.

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



Kuvio 1.4. Joukon A osituksia.

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

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

Jos $A_1, A_2, \ldots, A_n$ on jono tapahtumia, niiden unioni on

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cdots \cup A_n$$

ja leikkaus

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cdots \cap A_n.$$

Kun tapahtumien jono $\{A_n\}, n \geq 1$ on ääretön, jonon tapahtumien unioni ja leikkaus voidaan määritellä äärellisten unionien ja leikkausten raja-arvona:

$$\bigcup_{n=1}^{\infty} A_n = \lim_{m\to\infty} \bigcup_{n=1}^{m} A_n, \quad \bigcap_{n=1}^{\infty} A_n = \lim_{m\to\infty} \bigcap_{n=1}^{m} A_n.$$

Kun tapahtumien jono on $\{A_n\}, n = 1, 2, \ldots$ on äärellinen tai ääretön, voimme merkitä myös

$$\begin{aligned} \bigcup_{n} A_n &= \{\omega \,|\, \omega \in A_n \text{ ainakin yhdellä } n\text{:n arvolla}\}, \\ \bigcap_{n} A_n &= \{\omega \,|\, \omega \in A_n \text{ kaikilla } n\text{:n arvoilla}\}. \end{aligned}$$

Huomaa, että $\bigcup_{n=1}^{m} A_n$ ei vähene ja $\bigcap_{n=1}^{m} A_n$ ei kasva, kun m kasvaa. Jono $\{\bigcup_{n=1}^{m} A_n\}, m = 1, 2, \ldots$ paisuu kohti joukkoa $\bigcup_{n=1}^{\infty} A_n$ ja jono $\{\bigcap_{n=1}^{m} A_n\}, m = 1, 2, \ldots$ kutistuu kohti joukkoa $\bigcap_{n=1}^{\infty} A_n$. Ne ovat monotonisia jonoja. Jonoa $\{B_n\}, n = 1, 2, \ldots$ sanotaan *monotoniseksi*, jos $B_1 \subset B_2 \subset \ldots$ (*kasvava*) tai $B_1 \supset B_2 \supset \ldots$ (*vähenevä*). Monotonisille jonoille voidaan määritellä raja-arvo seuraavasti:

$$\lim_{n\to\infty} B_n = \bigcup_{n=1}^{\infty} B_n, \quad \text{kun } \{B_n\}, n \geq 1 \text{ kasvava},$$

ja

$$\lim_{n\to\infty} B_n = \bigcap_{n=1}^{\infty} B_n, \quad \text{kun } \{B_n\}, n \geq 1 \text{ vähenevä}.$$

Osittelulait ja De Morganin lait voidaan yleistää ilmeisellä tavalla koskemaan äärellisiä ja äärettömiä joukkojen jonoja. Esimerkiksi

$$\begin{aligned} B \cap (\bigcup_{n} A_n) &= \bigcup_{n} (B \cap A_n), \\ (\bigcap_{n} A_n)^c &= \bigcup_{n} A_n^c. \end{aligned}$$

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* $\{\omega\}$ *todennäköisyys.* Tapahtuman A eli Ω:n osajoukon todennäköisyys määritellään lukuna

$$P(A) = \sum_{\omega\in A} P(\{\omega\}). \tag{1.3.1}$$

Näin funktio P voidaan laajentaa joukkofunktioksi, joka liittää jokaiseen tapahtumaan $A \subset \Omega$ luvun $0 \leq P(A) \leq 1$. 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 $\{\, p_i \mid p_i = P(\{\omega_i\}),\ 1 \leq i \leq n \,\}$ määrittelee todennäköisyysjakauman Ω:ssa.

Esimerkki 1.5 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

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

□

Olemme nyt määritelleet todennäköisyysmallin (Ω, P) äärellisessä otosavaruudessa siten, että jokaisen tapahtuman todennäköisyys voitiin määritellä. Olemme kiinnostuneita Ω :n tapahtumien todennäkäisyyksistä. Tapahtumista johdetaan uusia tapahtumia joukko-opin operaatioilla.

Määritelmä 1.2 Otosavaruuden Ω osajoukkojen kokoelma $\mathcal{A}$ on *algebra*, jos seuraavat kolme ehtoa toteutuvat:

a_1. $\Omega \in \mathcal{A}$.

a_2. Jos $A \in \Omega$, niin $A^c \in \Omega$.

a_3. Jos $A, B \in \Omega$, niin $A \cup B \in \Omega$.

Todennäköisyyslaskennassa tarkasteltavat joukkokokoelmat (tapahtumien kokoelmat) muodostavat aina algebran. Esimerkkejä joukkoalgebroista ovat:

(a) Suppein madollinen algebra $\{\Omega, \emptyset\}$, johon kuuluvat vain otosavaruus Ω ja tyhjä joukko $\emptyset$.

(b) Tapahtuman A generoima algebra $\{A, A^c, \Omega, \emptyset\}$.

(c) Otosavaruuden Ω kaikkien osajoukkojen kokoelma $\{A | A \subset \Omega\}$, joka sisältää myös tyhjän joukon $\emptyset$.

Todettakoon, että kaikki mainitut joukkoalgebrat liittyvät johonkin otosavaruuden Ω ositukseen. Olkoon

$$\mathcal{D} = \{D_1, \ldots, D_n\}$$

otosavaruuden ositus. Silloin $D_1 + \cdots + D_n = \Omega$. Jos esimerkiksi $\Omega = \{\omega_1, \omega_2, \omega_3\}$, niin $\{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$ on Ω:n ositus, koska $\Omega = \{\omega_1\} + \{\omega_2\} + \{\omega_3\}$. Tämän osituksen avulla voidaan määritellä 5 eri ositusta: $\mathcal{D}_1 = \{\omega_1, \omega_2, \omega_3\}$,

$\mathcal{D}_2 = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$, $\mathcal{D}_3 = \{\{\omega_1, \omega_3\}, \{\omega_2\}\}$, $\mathcal{D}_4 = \{\{\omega_2, \omega_3\}, \{\omega_1\}\}$ $\mathcal{D}_5 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$. Jos muodostetaan osituksen $\mathcal{D}$ joukkojen kaikki unionit, niin saadaan joukkokokoelma, joka on algebra. Mukaan otetaan myös $\emptyset$, joka aina voidaan ajatella olevan osituksessa. Syntyvää joukkokokoelmaa sanotaan osituksen $\mathcal{D}$ indusoimaksi joukkokokoelmaksi $\alpha(\mathcal{D})$. Myös käänteinen tulos pitää paikkansa. Jos $\mathcal{A}$ on äärellisen otosavaruuden Ω osajoukkojen muodostama algebra, niin on olemassa sellainen yksikäsitteinen Ω:n ositus $\mathcal{D}$, että $\mathcal{A}$ on osituksen $\mathcal{D}$ indusoima algebra eli $\mathcal{A} = \alpha(\mathcal{D})$.

Esimerkki 1.6 (*a*) Tarkastellaan otosavaruuden $\Omega = \{\omega_1, \omega_2, \omega_3\}$ ositusta $\mathcal{D}_2 = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$. Merkitään $A = \{\omega_1, \omega_2\}$, joten $A^c = \{\omega_3\}$. Silloin osituksen $\mathcal{D}_2$ indusoima algebra on itse asiassa joukon A indusoima algebra.
(*b*) Olkoon otosavaruus $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_3\}$ ja sen ositus $\mathcal{D} = \{\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4\}\}$. Osituksen $\mathcal{D}_2$ indusoima algebra on $\{\Omega, \emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_3, \omega_4\}\}$. Jos merkitään $D_1 = \{\omega_1, \omega_2\}, D_2 = \{\omega_3\}$ ja $D_3 = \{\omega_4\}$, niin osituksen $\mathcal{D} = \{D_1, D_2, D_3\}$ indusoima algebra saadaan muodostamalla joukkojen D_1, D_2 ja D_3 kaikki mahdolliset unionit. Esimerkiksi $D_1 \cup D_3 = \{\omega_1, \omega_2, , \omega_4\}$ ja $D_2 \cup D_3 = \{\omega_3, \omega_4\}$. □

Kun satunnaiskokeelle määritellään todennäköisyysmalli, kiinnitetään ensin otosavaruus $\Omega = \{\omega_1, \ldots, \omega_n\}$. Sen jälkeen valitaan jokin sellainen osajoukkojen kokoelma $\mathcal{A}$, joka muodostaa algebran. Kokoelman $\mathcal{A}$ alkiot ovat tapahtumia. Kun Ω on äärellinen, valitaan joukkoalgebraksi $\mathcal{A}$ tavallisesti Ω:n kaikkien osajoukkojen kokoelma. Sitten jokaiseen alkeistapaukseen $\omega_i \in \Omega,\ 1 \leq i \leq n$ liitetään Määritelmän 1.1 mukaisesti epänegatiivinen paino. Tapahtuman $A \in \mathcal{A}$ todennäköisyys $P(A)$ määritellään kaavan (1.3.1) mukaisesti lukuna

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

Sanomme, että kolmikko

$$(\Omega, \mathcal{A}, P)$$

määrittelee *todennäköisyysmallin*, tai *todennäköisyysavaruuden*. Jos äärellisen otosavaruuden yhteydessä ei erikseen mainita joukkoalgebraa $\mathcal{A}$, tarkoitetaan Ω:n kaikkien osajoukkojen muodostamaa algebraa.

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 todennäköisyysfunktiofunktio 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 $P(\{\omega_i\}) = p_i$. Tapahtuman $A \in \Omega$ todennäköisyys on

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}), \tag{1.3.2}$$

mutta nyt nyt summa voi olla ääretön. Jos Ω ei ole numeroituva (eli on ylinumeroituva), niin Määritelmä 1.1 ei sovellu tapahtumien todennäköisyyden määrittelemiseen, vaan tarvitaan uusia käsitteitä. Niihin palataan myöhemmin.

Esimerkki 1.7 Esimerkissä 1.3 tarkasteltiin satunnaiskoetta, jossa heitetään lanttia, kunnes saadaan ensimmäinen klaava. Silloin otosavaruus

$$\Omega = \{\mathrm{L}, \mathrm{LR}, \mathrm{LLR}, \mathrm{LLLR}, \ldots\},$$

missä alkeistapaus

$$\omega_i = \underbrace{\mathrm{LL}\ldots\mathrm{L}}_{i-1}\mathrm{R}.$$

Jos kruunan todennäköisyys $P(\{R\}) = p$ ja klaavan todennäköisyys $P(\{L\}) = q$ $(p + q = 1)$, niin $P(\{\omega_i\}) = q^{i-1}p$. Silloin

$$\sum_{i=1}^{\infty} P(\{\omega_i\}) = \sum_{i=1}^{\infty} q^{i-1}p = \frac{p}{1-q} = 1.$$

□

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

$$\mathrm{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}. \tag{1.3.3}$$

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

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

Esimerkki 1.8 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 A mahdollisuus on 1 vastaan 10 ja lyöt euron vetoa tapahtuman puolesta, niin A:n 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

$$\theta(A,B) = \frac{\text{odds}(A)}{\text{odds}(B)} = \frac{P(A)/[1-P(A)]}{P(B)/[1-P(B)]}. \tag{1.3.4}$$

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 1.9 Heitetään harhatonta noppaa kuten Esimerkissä 1.5. 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 1.10 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-vuotiaaksi eläneet ovat eläneet myös 20-vuotiaaksi. 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
$$F_n(a) = \frac{1}{n}|\{\, i : 1 \le i \le n,\, x_i \le a \,\}|,$$
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. $P(\omega) \ge 0$ 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. 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.

2. 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.

3. 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?

4. 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.

5. 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?

6. 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.

7. (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`)?

8. (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.

9. 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?

10. 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).

11. 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.

Luku 2

Todennäköisyys ja satunnaismuuttuja

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*

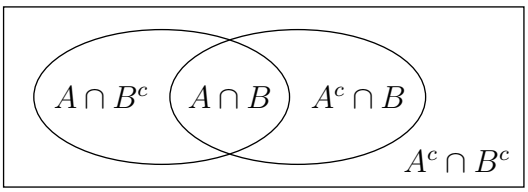
$$P(A) = P(A \cap B) + P(A \cap B^c).$$

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

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

Seurauslauseen 2.1 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

Todennäköisyys luonnehditaan otosavaruuden Ω tietyssä osajoukkojen kokoelmassa määriteltynä funktiona. Tässä alaluvussa käsitelemme hieman todennäköisyyden aksiomatiikkaa. Esitämme todennäköisyyden määritelmän, josta sen ominaisuudet voidaan johtaa. Esimerkiksi Lauseessa 2.1 esitetyt todennäköisyyttä koskevat tulokset seuraavat suoraan Määritelmässä 2.1 esitetyistä aksioomeista. Määritelmään 2.1 perustuvissa todistuksissa ei tarvita oletusta, että Ω on numeroituva (äärellinen tai numeroituvasti ääretön).

2.3.1 Äärellinen additiivisuus

Kun otosavaruus on äärellinen, voidaan todennäköisyyksiä tarkastella otosavaruuden Ω osajoukkojen muodostamassa algebrassa. Määritelmän 2.1 toisessa kohdassa esitetty todennäköisyyden yksinkertainen äärellinen additiivisuus riittää äärellisissä otosavaruuksissa (ominaisuuksien 1 ja 3 lisäksi) todennäköisyyden määrittelemiseen.

Määritelmä 2.1 Olkoon $\mathcal{A}$ otosavaruuden Ω osajoukkojen muodostama algebra. Todennäköisyys on kuvaus $P : \mathcal{A} \to [0, 1]$, joka toteuttaa seuraavat kolme aksioomaa:

1. $P(A) \geq 0$ kaikilla $A \in \mathcal{A}$.
2. Jos $A, B \in \mathcal{A}$ ja $A \cap B = \emptyset$, niin $P(A \cup B) = P(A) + P(B)$.
3. $P(\Omega) = 1$.

Jos $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$ on numeroituvasti ääretön, alkeistapahtuman $\{\omega_i\}$ todennäköisyys $P(\{\omega_i\}) = p_i \geq 0,\ i \geq 1$ ja ja $P(\Omega) = \sum_{i=1}^{\infty} p_i = 1$, niin Määritelmän 2.1 aksiomeista seuraa, että Ω:n kaikkien osajoukkojen $A \subset \Omega$ todennäköisyys saadaan kaavalla (1.3.2) eli $P(A) = \sum_{\omega_i \in A} P(\{\omega_i\})$. Silloin kokoelma $\mathcal{A}$ on Ω:n kaikkien osajoukkojen muodostama algebra. Huomattakoon, että kaikkien alkeistapahtumien todennäköisyys ei voi olla sama, jos Ω on ääretön. Yleisessä tapauksessa tarvitaan vahvempia oletuksia todennäköisyyden määrittelemiseksi, mutta numeroituvan Ω:n tapauksessa Määritelmä 2.1 on riittävä.

2.3.2 Todennäköisyyden yleiset aksioomat

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

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

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 aksioomasta 2.1 (2.) seuraa

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

Ylinumeroituvasti äärettömille otosavaruuksilla eivät Määritelmässä 2.1 esitetyt aksioomat riitä. Tarvitaan yleisemmät aksioomat. Yleisessä teoriassa Ω:n *kaikki osajoukot eivät ole tapahtumia.* Määrittelimme edellä todennäköisyyden Ω:n joukkokokoelmassa, joka muodosti algebran. Yleisessä teoriassa niiden Ω:n osajoukkojen, jotka ovat tapahtumia, täytyy muodostaa ns. *σ-algebra.*

Määritelmä 2.2 Kokoelma $\mathcal{F}$ on otosavaruuden Ω osajoukkojen muodostama σ-algebra, jos seuraavat ehdot toteutuvat:

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 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Tapahtuman A todennäköisyys $P(A)$ on reaaliluku, jonka on oltava yksikäsitteisesti määritelty, kun tapahtum $A \in \mathcal{F}$ on annettu.

Määritelmä 2.3 Kuvaus $P : \mathcal{F} \to \mathbb{R}$ on todennäköisyys, 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 joukot $A_i \in \mathcal{F}, i = 1, 2, \ldots$ ovat parittain pistevieraat ($A_i \cap A_j$, kun $i \neq j$), niin
$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Näistä aksioomista voidaan tietysti johtaa samat lauseet kuin edellä numeroituvien otosavaruuksien tapauksessa. Kolmikko $(\Omega, \mathcal{F}, P)$ on *todennäköisyysavaruus,* missä Ω on ei-tyhjä otosavaruus, $\mathcal{F}$ on σ-algebra ja P: $\mathcal{F} \to [0, 1]$ on todennäköisyys(mitta). Tämän todennäköisyyden aksiomatisoinnin esitti venäläinen matemaatikko A. N. Kolmogorov (1903–87) vuonna 1929.

Äärellisen additiivisuuden nojalla

$$P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \tag{2.3.1}$$

kaikilla $n \geq 1$. Koska yhtälön (2.3.1) vasen puoli on korkeintaan 1 kaikilla $n \geq 1$, niin oikealla puolella oleva sarja suppenee. Silloin

$$\lim_{n\to\infty} P(\sum_{i=1}^{n} A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{\infty} P(A_i). \tag{2.3.2}$$

On huomattava, että tästä ei seuraa Määritelmässä 2.3 esitetty σ-additiivisuus

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

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 $n_1 n_2$.

Esimerkki 2.3 Tarkastellaan seuraavia kysymyksiä:

1. Sinulla on kolme paitaa, neljät housut, kolmet kengät ja kymmenet sukat. Montako asukokonaisuutta näistä voit muodostaa?

2. Esitä kaikki 2-numeroiset luvut, jotka numeroista $\{1, 5, 6, 9\}$ voidaan muodostaa?

3. Moneenko eri järjestykseen 10 kirjaa voidaan hyllyssä asettaa?

Kaikkiin esimerkissä esitettyihin kysymyksiin saadaan vastaus tuloperiaatteen avulla. □

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 2.4 (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, mutta erilaisia kuin miden ryhmien alkiot. Emme voi siis tunnistaa erilaisia ryhmän sisäisiä järjestyksiä. Vastaus saadaan samalla tavalla kuin edellisessä 10-kirjaimisia sanoja koskevassa esimerkissä. 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	$\odot$	$\ominus$	$\oplus$	
$\odot\odot\odot$	***			***||
$\odot\odot\ominus$	**	*		**|*|
$\odot\odot\oplus$	**		*	**||*
$\odot\ominus\ominus$	*	**		*|**|
$\odot\ominus\oplus$	*	*	*	*|*|*
$\odot\oplus\oplus$	*		**	*||**
$\ominus\ominus\ominus$		***		|***|
$\ominus\ominus\oplus$		**	*	|**|*
$\ominus\oplus\oplus$		*	**	|*|**
$\oplus\oplus\oplus$			***	||***

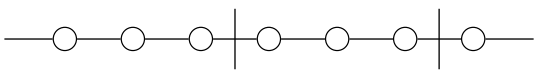
Jokaisen valinnan jälkeen pistetään merkki "$*$" kyseisen pallon kohdalle. Kaikkien valintojen jälkeen meillä on 3 (r) merkkiä. Huomaa, että r voi olla suurempi kuin k, vaikka esimerkissä $r = k = 3$. Taulukon 2.1 viimeisellä sarakkeella pallojen valinta on esitetty merkkien "$*$" ja "$|$" jonona ilmeisellä 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

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

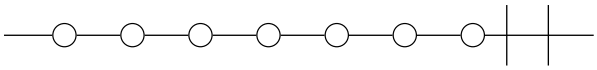
Esimerkki 2.5 Tarkastellaan yhtälöä

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

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

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

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 2.6 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

Binomikertoimiin liittyy monia tärkeitä identiteettejä. Tässä alaluvussa esitettävät kolme identiteettiä ovat sovellusten kannalta keskeisiä.

Lause 2.6 (Binomilause) *Olkoot a ja b reaalilukuja sekä n epänegatiivinen kokonaisluku. Silloin*

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} \tag{2.4.8}$$

Kertoimia $\binom{n}{r}$ kutsutaan binomikertoimiksi.

Lause 2.7 (Hypergeometrinen identiteetti) *Olkoot a ja b reaalilukuja ja n positiivinen kokonaisluku. Silloin*

$$\sum_{r=0}^{n} \binom{a}{r}\binom{b}{n-r} = \binom{a+b}{n}. \tag{2.4.9}$$

Lause 2.8 (Multinomilause) *Olkoon annettu positiivinen kokonaisluku n ja reaaliluvut t_1, t_2, …, t_k. Silloin*

$$(t_1 + t_2 + \cdots + t_k)^n = \sum \sum \cdots \sum_{n_1+\cdots+n_k=n} \binom{n}{n_1\ n_2\ \ldots\ n_k} t_1^{n_1} t_2^{n_2} \cdots t_k^{n_k}, \tag{2.4.10}$$

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

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 2.4 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, ... arvoja merkitään pienillä kirjaimilla x, y, z, ... 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*. Jatkossa määritellään myös jatkuva satunnaismuuttuja X, kun X saa kaikki arvot jollain välillä $I \subset \mathbb{R}$. On myös muita satunnaismuuttujia, mutta käsittelemme vain diskreettejä ja jatkuvia satunnaismuuttujia.

Esimerkki 2.7 Heitetään harhatonta lanttia 3 kertaa. Satunnaismuuttuja Y on 'kruunien lukumäärä'. Merkitään R = 'kruuna' ja L = 'klaava'. Silloin

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

Silloin esimerkiksi $Y(\text{RRL}) = Y(\text{RLR}) = 2$. Olkoon A_r tapahtuma "kruunien lukumäärä r". Merkintä $(Y = 2)$ tarkoittaa tapahtumaa $A_2 = \{\text{RRL}, \text{RLR}, \text{LRR}\}$ ja $P(Y = 2) = P(A_2) = 3/8$. Vastaavasti $(Y \leq 1)$ on tapahtuma $A_0 \cup A_1 = \{\text{RLL}, \text{LRL}, \text{LLR}, \text{LLL}\}$ ja $P(Y \leq 1) = 1/2$. Satunnaismuuttuja Y arvojoukko on $S = \{0, 1, 2, 3\}$. □

Yleisesti merkintä $(X \in B)$, $B \subset \mathbb{R}$ tarkoittaa sellaista tapahtumaa $A \subset \Omega$, että $A = \{\omega \in \Omega | X(\omega) \in B\}$. Tätä tapahtumaa merkitään myös $X^{-1}(B) = A$. Esimerkissä 2.7 tapahtuma $(Y \in [0, 1])$ on $A = \{\text{RLL}, \text{LRL}, \text{LLR}, \text{LLL}\}$, koska $Y(\omega) \in [0, 1]$, kun $\omega \in A$ ja $Y(\omega) \notin [0, 1]$, kun $\omega \in A^c$. Siis $Y^{-1}([0, 1]) = A$.

Esimerkki 2.8 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.

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. kruuna	$\{1,2,3\ldots\}$
W	Korttipakasta satunnaisesti valitun kortin maa	$\{\spadesuit,\heartsuit,\clubsuit,\diamondsuit\}$

Määritelmässä 2.4 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 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 kruunien 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 2.5 (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 'kruuna k. heitossa' ($1 \leq k \leq n$) indikaattorifunktio. Silloin satunnaismuuttuja

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

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

2.6 Satunnaismuuttujan jakauma

Satunnaismuuttujassa olemme kiinnostuneita erityisesti sen jakaumasta. Haluamme siis tietää, millä todennäköisyydellä X:n arvot kuuluvat mihin tahansa reaaliakselin $\mathbb{R}$:n osajoukkoon B. Kun tapahtumaa $\{\omega \in \Omega | X(\omega) \in B\}$ merkitään $(X \in B)$, ovat mielenkiinnon kohteena todennäkäisyydet

$$P(X \in B) = P(\{\omega \in \Omega | X(\omega) \in B\}), B \subset \Omega. \tag{2.6.1}$$

Merkintä $P(X \in B)$ osoittaa, että tapahtuma on määritelty satunnaismuuttujan X avulla. Koska B voi olla mikä tahansa reaaliakselin $\mathbb{R}$:n osajoukko, todennäköisyydet $P(X \in B)$ määrittelevät *satunnaismuuttujan jakauman.*

Määritelmä 2.6 (Todennäköisyysjakauma) Määritellään satunnaismuuttujan X todennäköisyysjakauma (lyhyesti jakauma) ja kertymäfunktio.

(i) Satunnaismuuttujan X *todennäköisyysjakauma* on joukkofunktio P_X, joka liittää jokaiseen $\mathbb{R}$:n osajoukkoon B arvon relaation (2.6.1) mukaisesti. Joukkoon B liitettyä arvoa merkitään $P_X(B)$.

(ii) Kun valitaan $B = (-\infty, x]$, eli B on puoliavoin väli, niin identiteetin (2.6.1) nojalla saadaan

$$P(X \in (-\infty, x]) = P(\{\omega \in \Omega | X(\omega) \leq x\}) = P(X \leq x). \tag{2.6.2}$$

Relaatio (6.1.4) määrittelee pistefunktion $F_X(x) = P(X \leq x)$. Funktiota F_X kutsutaan X:n *kertymäfunktioksi.*

Huomautus 2.2 Satunnaismuuttujan X jakauma on siis $\mathbb{R}$:n osajoukoille määritelty joukkofunktio. Se on itse asiassa todennäköisyysfunktio. Jos $P_X(B)$ tunnetaan kaikilla $B \subset \mathbb{R}$, niin tunnetaan myös kertymäfunktion arvot $F_X(x)$ kaikilla $x \in \mathbb{R}$. Kertymäfunktion määrittelemiseksi täytyy itse asiassa tuntea vain kaikkien puoliavoimien välien $B = (-\infty, x]$ todennäköisyydet $P_X((-\infty, x])$. Yllättävää kyllä, myös käänteinen tulos pitää paikkansa. Jos kertymäfunktio F on annettu, niin on olemassa siihen liittyvä yksikäsitteinen todennäköisyysfunktio, joka on määritelty tietyssä $\mathbb{R}$:n osajoukkojen luokassa. Tähän kertymäfunktion ja todennäköisyyden vastaavuuteen perustuu kertymäfunktion keskeinen asema todennäköisyyslaskennassa.

Esimerkki 2.9 Olkoon Y kruunien lukumäärä kolmen lantin heitossa (Esimerkki 2.7). Silloin satunnaismuuttujaa koskevat väittämät, kuten 'täsmälleen yksi kruuna' $\equiv$ "$Y = 1$" tai 'korkeintaan 2 kruunaa' $\equiv$ "$Y \leq 2$" määrittelevät tapahtuman. Tapahtumat voidaan silloin kirjoittaa muodossa "$Y \in B$". Jos $B = (-\infty, 2]$, niin B määrittelee otosavaruudessa Ω tapahtuman

$$\{\omega \in \Omega | Y(\omega) \in B\}) = \{\mathrm{RRL}, \mathrm{RLR}, \mathrm{RLL}, \mathrm{LRR}, \mathrm{LRL}, \mathrm{LLR}, \mathrm{LLL}\}$$

ja $P_Y(B) = P(\{\omega \in \Omega | Y(\omega) \in B\}) = 7/8$. Jatkossa tulemme pääsääntöisesti tarkastelemaan *satunnaismuuttujien määrittämiä tapahtumia.* □

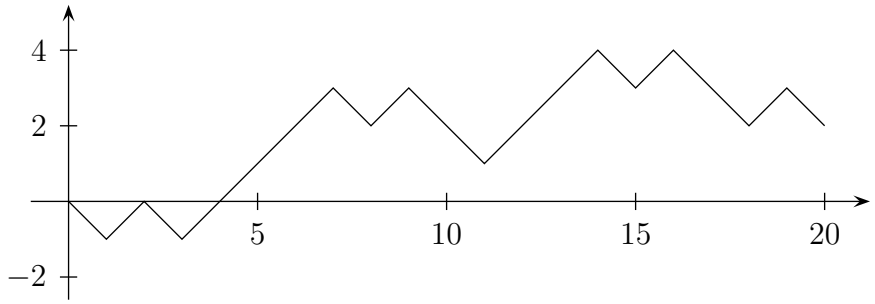
Esimerkki 2.10 (Satunnaiskävely, Random Walk) Pekka ja Paavo pelaavat "kruunaa ja klaavaa". Tässä pelissä heitetään peräkkäin lanttia n kertaa – tässä esimerkissä $n = 20$. Aina kun tulee kruuna (R), Pekka voittaa euron Paavolta. Kun tulee klaava (L), Pekka häviää euron Paavolle. Kuviossa 2.2 esitetyn pelin tulos ($n = 20$) on

$$\mathrm{L\,R\,L\,R\,R\,R\,R\,L\,R\,L\,L\,R\,R\,R\,L\,R\,L\,L\,R\,L.}$$

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:

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



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

missä

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

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. □

2.6.1 Kertymäfunktio

Satunnaismuuttuja X kertymäfunktio määriteltiin (ks. Määritelmä 2.6 (*ii*)) X:n jakauman avulla. Jos $x_1 \leq x_2$, niin $\{X \leq x_1\} \subset \{X \leq x_2\}$ ja todennäköisyyden monotonisuusominaisuuden perusteella [Lause (2.1), kohta 3)] $P(X \leq x_1) \leq P(X \leq x_2)$, joten kertymäfunktio $F(x)$ on *kasvava* (ei vähenevä). Seuraavassa lauseessa esitetään kertymäfunktion ominaisuudet.

Lause 2.9 *Satunnaismuuttujan X kertymäfunktiolla F on seuraavat ominaisuudet:*

1. $0 \leq F(x) \leq 1, \quad x \in \mathbb{R}$.
2. *$F(x)$ on x:n kasvava (ei vähenevä) 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).*

4. $\lim_{x \to -\infty} F(x) = 0$ *ja* $\lim_{x \to \infty} F(x) = 1$.

Satunnaismuuttujaa X tarkasteltaessa joudutaan usein laskemaan muotoa $P(X = a)$, $P(X < b)$, $P(X \leq b), P(X > a)$, $P(X \geq a)$, $P(a \leq X \leq b)$, $P(a < X \leq b)$, $P(a \leq X < b)$ and $P(a < X < b)$ olevia todennäköisyyksiä, missä $a \leq b$ ovat reaalilukuja. Huomattakoon, että P on X:n jakaumaan liittuvä todennäköisyysfunktio, jota on merkitty aikaisemmin P_X. Jätämme jatkossa alaindeksin pois, jos tilanne on yhteydestä selvä. Esimerkiksi

$$\begin{aligned} (a < X \leq b) &= \{\, \omega \mid a < X(\omega) \leq b \,\} \\ &= \{\, \omega \mid X(\omega) \leq b \,\} - \{\, \omega \mid X(\omega) \leq a \,\}, \end{aligned} \tag{2.6.3}$$

koska $\{\, \omega \mid X(\omega) \leq b \,\} = \{\, \omega \mid X(\omega) \leq a \,\} + \{\, \omega \mid a < X(\omega) \leq b \,\}$. Silloin

$$P(\{\, \omega \mid a < X(\omega) \leq b \,\}) = F_X(b) - F_X(a),$$

koska

$$\begin{aligned} P(\{\, \omega \mid X(\omega) &\leq b \,\} - \{\, \omega \mid X(\omega) \leq a \,\}) \\ &= P(\{\, \omega \mid X(\omega) \leq b \,\}) - P(\{\, \omega \mid X(\omega) \leq a \,\}) \\ &= P_X(X \leq b) - P_X(X \leq a) = F_X(b) - F_X(a). \end{aligned}$$

Olkoon X henkilön ikä vuosissa ja T elinaika, kun Ω on suomalaisten joukko. Silloin $\{X = 20\} = \{\, \omega \mid X(\omega) = 20 \,\} = \{\, \omega \mid 20 \leq T(\omega) < 21 \,\})$ on 20-vuotiaiden suomalaisten joukko.

Jos X:n arvojoukko $S_X = \{x_1, x_2, \ldots, x_n, \ldots, \}$ on numeroituva, määritellään

$$p_n = P(X = x_n), \quad x_n \in S_X,\ n = 1, 2, \ldots.$$

Jos $x \notin S_X$, niin $P(X = x) = 0$. Jos tunnemme kaikki todennäköisyydet p_n, on ilmeistä, että voimme määrittää satunnaismuuttujan X jakauman. Silloin todennäköisyys

$$P(X \in B) = \sum_{x_n \in B} p_n$$

kaikilla $B \subset S_X$. Jokaista reaalilukua x kohti

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

ja kaikilla $a < b$

$$\begin{aligned} P(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \sum_{a < x_n \leq b} p_n. \end{aligned}$$

Esimerkki 2.11 Esimerkissä 2.7 heitettiin harhatonta lanttia 3 kertaa. Satunnaismuuttuja X on 'kruunien lukumäärä' ja X:n arvojoukko $S = \{0, 1, 2, 3\}$. Nyt

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

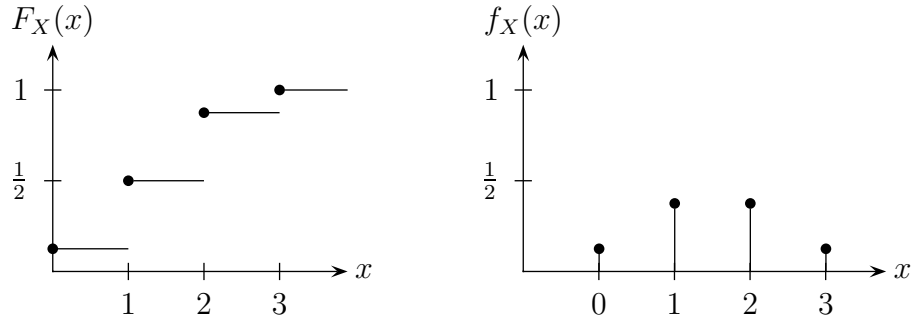
missä kaikki alkeistapaukset ovat yhtä todennäköisiä. Silloin

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

Satunnaismuuttujan X kertymäfunktio on siis

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

□

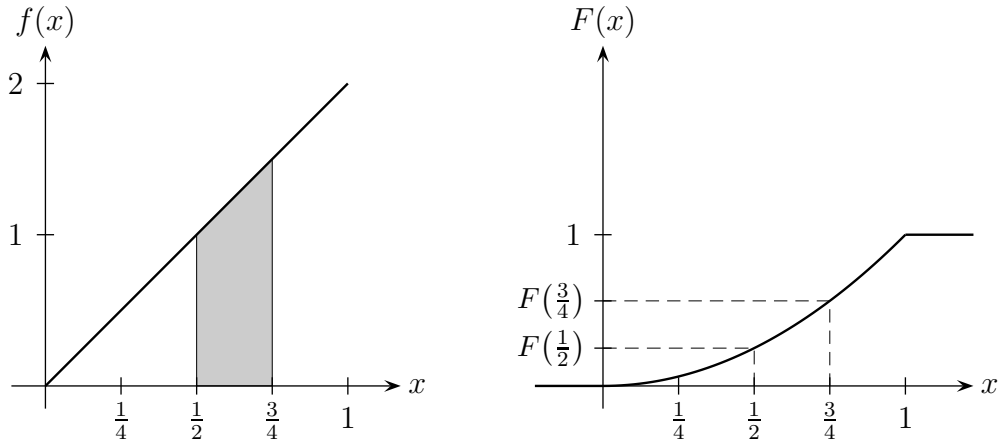


Kuvio 2.3. Satunnaismuuttujan X kertymäfunktion $F_X(x)$ ja todennäköisyysfunktion $f_X(x)$ kuvaajat.

Esimerkin 2.11 kertymäfunktio on porrasfunktio. Pitää yleisestikin paikkansa, että diskreetin satunnaismuuttujan kertymäfunktio on porrasfunktio. Voimme sanoa, että satunnaismuuttuja on diskreetti satunnaismuuttuja, jos sen kertymäfunktio on porrasfunktio.

Esimerkki 2.12 Olkoon X satunnaismuuttuja, jonka kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ x^2, & 0 \leq x < 1; \\ 1, & 1 \leq x. \end{cases}$$



Kuvio 2.4. Jatkuvan satunnaismuuttujan X tiheysfunktio $f(x) = 2x$ ja kertymäfunktio $F(x) = x^2$.

Kertymäfunktion avulla voidaan laskea todennäköisyyksiä. Esimerkiksi todennäköisyys

$$P\left(\frac{1}{2} < X \leq \frac{3}{4}\right) = F\left(\frac{3}{4}\right) - F\left(\frac{1}{2}\right) = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$

ja

$$P\left(\frac{3}{4} < X \leq \frac{3}{2}\right) = F\left(\frac{3}{2}\right) - F\left(\frac{3}{4}\right) = 1 - \left(\frac{3}{4}\right)^2 = \frac{7}{16}.$$

□

Satunnaismuuttujan X jatkuvuus voidaan määritellä kertymäfunktion F_X jatkuvuuden avulla. Esimerkissä 2.12 käsitellyn satunnaismuuttujan kertymäfunktio on jatkuva.

Määritelmä 2.7 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.

2.6.2 Satunnaismuuttujan tiheysfunktio

Satunnaismuuttujaan X liittyvien todennäköisyyksien laskennassa on monesti käyttökelpoisin on X:n tiheysfunktio. Määritellään diskreetin ja jatkuvan satunnaismuuttujan tiheysfunktiot erikseen.

Määritelmä 2.8 Olkoon X diskreetti satunnaismuuttuja, jolla on numeroituva määrä arvoja $x_i,\ i \geq 1$, joiden todennäköisyydet ovat $P(X = x_i),\ i \geq 1$. Määritellään funktio f_X seuraavasti:

$$f_X(x) = \begin{cases} P(X = x_i), & \text{kun } x = x_i,\ i \geq 1 \\ 0, & \text{muutoin.} \end{cases} \tag{2.6.4}$$

Funtiota f_X kutsutaan satunnaismuuttujan X *tiheysfunktioksi*. Diskreetin satunnaismuuttujan tiheysfunktiota sanotaan myös todennäköisyysfunktioksi.

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

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

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

Lauseessa 2.10 esitetyt todennäköisyysfunktion ominaisuudet seuraavat suoraan Määritelmästä 2.8.

Lause 2.10 *Relaatiolla (2.6.4) mmääritellyllä diskreetin satunnaismuuttujan X todennäköisyysfunktiolla f_X on seuraavat ominaisuudet:*

1. $f_X(x) \geq 0$ *kaikilla* $x \in \mathbb{R}$.
2. *Mille tahansa* $B \subset \mathbb{R}$, $P(X \in B) = \sum_{x_i \in B} f_X(x_i)$.
3. *Jos F_X on X:n kertymäfunktio, niin*
$$F_X(x) = \sum_{x_i \leq x} f_X(x_i), \text{ ja } \sum_{x_i \in \mathbb{R}} f_X(x_i) = 1.$$
4. *Oletetaan, että* $x_i < x_{i+1},\ i \geq 1$. *Silloin*
$$f_X(x_{i+1}) = F_X(x_{i+1}) - F_X(x_i),\ i \geq 1,\ f_X(x_1) = F_X(x_1).$$

Lauseen 2.10 kohdissa 3 ja 4 on esitetty kertymäfunktion ja todennäköisyysfunktion välinen yhteys. Diskreetin satunnaismuuttujan kertymäfunktio on porrasfunktio. Porrasfuntion $F(x)$ 'hyppäykset' $F_X(x_{i+1}) - F_X(x_i)$ 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

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

Esimerkki 2.13 Jatketaan esimerkiä 2.11, jossa heitettiin harhatonta lanttia 3 kertaa. Satunnaismuuttuja X on 'kruunien 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

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

□

Esimerkki 2.14 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 $X \sim \mathrm{Tasd}(x_1, x_2, \ldots, x_N)$. Hyvin usein X:n arvojoukko on $S = \{1, 2, \ldots, N\}$, jolloin merkitään $X \sim \mathrm{Tasd}(1, 2, \ldots, N)$. Esimerkiksi nopanheitossa silmäluvun X arvojoukko on $S = \{1, 2, 3, 4, 5, 6\}$ ja todennäköisyysfunktio

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

□

Määritelmä 2.9 Olkoon X jatkuva satunnaismuuttuja. Oletetaan, että on olemassa sellainen funktio f_X, että

$$f_X(x) \geq 0 \text{ kaikilla } x \in \mathbb{R}, \text{ ja } P(X \in B) = \int_B f_X(x)\,\mathrm{d}x, \ B \subset \mathbb{R}. \tag{2.6.5}$$

Funtiota f_X kutsutaan jatkuvan satunnaismuuttujan X *tiheysfunktioksi.*

Määritelmästä 2.9 ja integraalilaskennan tuloksista saadaan Lauseessa 2.11 esitetyt X:n tiheysfunktion ominaisuudet.

Lause 2.11 *Jatkuvan satunnaismuuttujan X tiheysfunktiolla (2.6.5) on seuraavat ominaisuudet:*

1. $F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$ *kaikilla* $x \in \mathbb{R}$.

2. $\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

3. $F_X'(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = f_X(x)$ *kaikissa pisteissä* $x \in \mathbb{R}$, *joissa* $f_X(x)$ *on jatkuva.*

Lauseessa 2.11 kohdassa 1 valitaan $B = (-\infty, x]$ ja 2. kohdassa $B = \mathbb{R}$. Lauseessa 2.10 todennäköisyyksien laskeminen palautuu yhteenlaskuun ja lauseessa 2.11 integrointiin. Integrointialue B on väli tai muodostuu korkeintaan äärellisestä määrästä välejä. Integraalit ovat analyysista tuttuja Riemanin integraaleja.

Huomautus 2.3 Jos F on kertymäfunktio, voidaan aina konstruoida sellainen satunnaismuuttuja X, että $F_X = F$. Milloin annettu funktio f on tiheysfunktio? Edellä esitettyjen tiheysfunktion ominaisuuksien nojalla täytyy olla (1) $f(x) \geq 0$ kaikilla $x \in \mathbb{R}$ ja (2) $\sum_{x_i} f(x_i) = 1$, kun X on diskreetti ja $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$, kun X on jatkuva. Jatkuvien satunnaismuuttujien tapauksessa yhden pisteen todennäköisyys on $P(X = x) = 0$ kaikilla $x \in \mathbb{R}$.

Esimerkki 2.15 Olkoon X jatkuva satunnaismuuttuja, jonka tiheysfunktio on $f(x) = 2x$, kun $0 < x < 1$. Tämän satunnaismuuttujan X kertymäfunktiota

$$F(x) = \begin{cases} 0, & x < 0; \\ x^2, & 0 \leq x < 1; \\ 1, & 1 \leq x. \end{cases}$$

tarkasteltiin Esimerkissä 2.12. Nyt voidaan todeta, että

$$F(x) = \int_0^x 2t\,\mathrm{d}t = x^2, \qquad \text{kun } 0 \leq x < 1.$$

Kun kertymäfunktio on annettu, niin tiheysfunktio saadaan derivoimalla kertymäfunktio:

$$F'(x) = \frac{\mathrm{d}}{\mathrm{d}x}x^2 = 2x, \qquad 0 \leq x < 1.$$

Todennäköisyyksiä voidaan laskea tiheysfunktion avulla integroimalla. Esimerkiksi $P\left(\frac{1}{2} \leq X \leq \frac{3}{4}\right)$ saadaan 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 myös kertymäfunktion avulla. □

2.7 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 2.16 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 $1/311875200$. □

2.7.1 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

$$f(x) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}, \qquad x = 0, 1, 2, \ldots \tag{2.7.1}$$

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

2.7.2 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:

Hyväksy erä, jos $x \leq h$, 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 virheellisten lukumäärän v jakauma yli tuote-erien.

2.8 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 $000, 001, 002, \ldots, 999$. 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 2.17 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

$$P(\text{'Sama ei toistu'}) = \frac{N^{(n)}}{N^n} = \frac{N!}{(N-n)!\, N^n}. \tag{2.8.1}$$

On selvää, että todennäköisyys (2.8.1) on 0, jos $n > N$. Huomaa, että $N^{(n)} = 0$, jos $n > N$. Syntymäpäiväongelmassa (Esimerkki 2.4) $N = 365$ ja $n = r$.

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

$$\frac{N!}{(N-n)!\, N^n} \approx \left(\frac{N}{N-n}\right)^{N-n+1/2} \mathrm{e}^{-n}. \tag{2.8.2}$$

Kun $N \to \infty$ ja n on kiinnitetty, niin lauseke (2.8.2) 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.

2.9 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

$$f(x) = \binom{n}{x}p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{2.9.1}$$

Funktio (2.9.1) on *binomijakauman* todennäköisyysfunktio. Kun X noudattaa binomijakaumaa, merkitsemme $X \sim \mathrm{Bin}(n, p)$.

2.9.1 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 (2.9.1) on hypergeometrisen jakauman (2.7.1) hyvä likiarvo. Otanta palauttamatta voidaan luonnehtia hypergeometrisen jakauman avulla ja otanta palauttaen binomijakauman avulla.

Lause 2.12 *Jos $a+b \gg n$, niin*

$$\frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} \approx \binom{n}{x}p^x(1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, \tag{2.9.2}$$

missä $p = a/(a+b)$.

Koska binomitodennäköisyydet on helpompi laskea kuin hypergeometriset todennäköisyydet, voidaan relaatiota (2.9.2) käyttää laskennassa hyväksi, kun $a + b \gg n$. Tosin nykyisillä ohjelmilla on helppo laskea tarkat todennäköisyydet suoraan hypergeometrisesta jakaumasta, vaikka $a + b$ on suuri.

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

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

- 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 $X\colon \Omega \to \mathbb{R}$.
- X:n arvojoukko S: $X(\omega) \in S \subset \mathbb{R}$.
- 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: $F(x) = P(X \leq x)$.
- 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
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

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, ..., r_n, $r_i \geq 0$, $i = 1, 2, \ldots, n$ ja

$$r_1 + r_2 + \cdots + r_n = r. \tag{2.9.3}$$

(a) Mikä on erilaisten varausjakaumien (Yhtälön (2.9.3) 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, ..., r_n toteuttavat yhtälön (2.9.3) ja kaikki mahdolliset n^r töiden sijoittelua prosessoreille ovat yhtä mahdollisia. Mikä on todennäköisyys saada varausluvut r_1, r_2, ..., r_n?

Luku 3

Ehdollinen todennäköisyys ja riippumattomuus

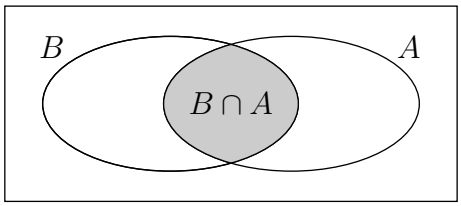
Tässä luvussa täydennetään ja laajennetaan edellisissä luvuissa esitettyjä todennäköisyyslaskennan tietoja. Ehdollistaminen ja riippumattomuus ovat todennäköisyyslaskennan ja tilastotieteen keskeisiä käsitteitä.

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 saada 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 Tulosääntö, kokonaistodennäköisyys ja Bayesin kaava

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.

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

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

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. Viallisten %-osuus eri maissa on annettu oheisessa taulukossa. 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.

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 (3.3.6) 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

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

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

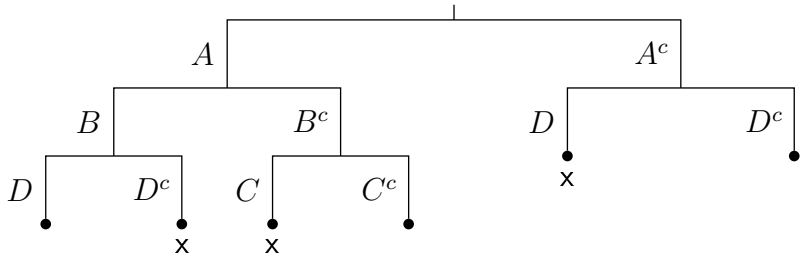
(a) millä todennäköisyydellä koe antaa väärän tuloksen?

(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 $A_{i_1}, \ldots, A_{i_k}$ $(1 < k \leq n)$ toteuttaa ehdon

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

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, sillä $A \cup A^c = \Omega$ ja . Tapahtuma "A tai A^c" sattuu varmasti eli $P(A \cup A^c) = P(A) + P(A^c) = 1$. Tästä 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

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

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 Jos X:n arvojoukko on $S_X = \{1, 2, \ldots, N\}$ ja $P(X = i) = 1/N$ kaikilla $i \in S_X$, niin sanomme, että X noudattaa diskreettiä tasajakaumaa Tasd$(1, N)$. 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 Yleinen tulokaava ja Bayesin lause

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.

3.3.1 Yleinen tulokaava

Tässä alaluvussa 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*

$$P(A_1A_2\cdots A_n) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1A_2)\cdots \cdot P(A_n \mid A_1A_2\cdots A_{n-1}), \tag{3.3.1}$$

jos $P(A_1A_2\cdots A_{n-1}) > 0$.

Todistus. Jos $P(A_1A_2\cdots A_{n-1}) > 0$, niin kaavassa (3.3.1) 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 (3.3.1) 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 (3.3.1) 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, ..., X_n, ... 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 (3.3.1) avulla

$$\begin{aligned} &P(X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n) \\ &\quad = P(X_1 = x_1)\,P(X_2 = x_2 \mid X_1 = x_1)\,P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2)\cdots \\ &\qquad\qquad \cdot P(X_n = x_n \mid X_1 = x_1,\, \ldots,\, X_{n-1} = x_{n-1}). \end{aligned} \tag{3.3.2}$$

$P(X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n)$ on satunnaismuuttujien X_1, X_2, ..., X_n *yhteistodennäköisyys,* joka on lausuttu peräkkäisten ehdollisten todennäköisyyksien avulla.

Esimerkki 3.8 (Syntymäpäiväongelma uudelleen) Olemme jo aikaisemmin implisiittisesti soveltaneet yleistä tulokaavaa (3.3.1). Tarkastellaan uudelleen Esimerkin 2.4 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, …, B_r ovat toisensa poissulkevat, niin

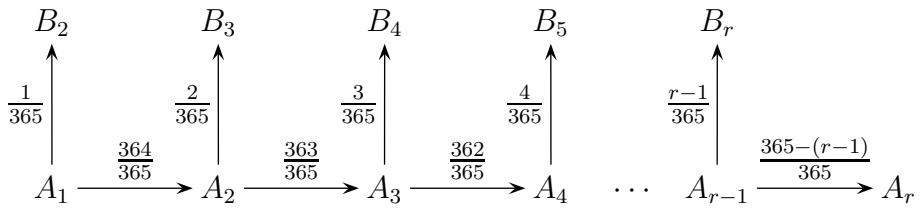
$$P(A_r^c) = P(B_2) + P(B_3) + \cdots + P(B_r),$$

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 syntymäpäivästä. 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 $A_2 A_3 \cdots A_j = A_j$ ja

$$P(A_{j+1} \mid A_2 A_3 \cdots A_j) = P(A_{j+1} \mid A_j) = \frac{365 - j}{365} = 1 - \frac{j}{365}.$$

Soveltamalla tapahtumien yhdisteen tulokaavaa saadaan

$$\begin{aligned} P(A_r) &= P(A_2 A_3 A_4 \cdots A_r) \\ &= P(A_2)\, P(A_3 \mid A_2)\, P(A_4 \mid A_2 A_3) \cdots P(A_r \mid A_2 \cdots A_{r-1}) \\ &= P(A_2)\, P(A_3 \mid A_2)\, P(A_4 \mid A_3) \cdots P(A_r \mid A_{r-1}) \\ &= \frac{364}{365} \cdot \frac{362}{365} \cdot \frac{362}{365} \cdots \frac{365 - r + 1}{365} = \frac{365^{(r)}}{365^r}. \end{aligned}$$

□

3.3.2 Bayesin lause

Pastori Thomas Bayesin (1763) mukaan nimetty lause seuraa suoraan ehdollisen todennäköisyyden määritelmästä. Bayesilainen lähestymistapa tilastotieteeseen perustuu tähän lauseeseen. Olkoot $H_1, H_2, \ldots, H_k$ sellaiset tapahtumat, että

$$H_i H_j = \emptyset \quad (i \neq j) \qquad \text{ja} \qquad \sum_{i=1}^{k} H_i = \Omega.$$

Nyt siis tapahtumajoukko $H_1, H_2, \ldots, H_k$ muodostaa otosavaruuden Ω osituksen. Tämä tarkoittaa sitä, että yksi ja vain yksi tapahtumista $H_1, H_2, \ldots, H_k$ sattuu, kun tehdään satunnaiskoe $\mathcal{E}$, jonka otosavaruus on Ω. Oletamme lisäksi, että $P(H_i) > 0$ kaikilla $i = 1, 2, \ldots, k$.

Lause 3.2 *Olkoon*

$$\Omega = \sum_i H_i$$

jokin otosavaruuden ositus. Silloin minkä tahansa tapahtuman $T \subset \Omega$ todennäköisyys voidaan lausua muodossa

$$P(T) = \sum_i P(H_i)\, P(T \mid H_i). \tag{3.3.3}$$

Todistus. Joukko-opin sääntöjen nojalla saadaan

$$T = \Omega T = \left(\sum_i H_i\right) T = \sum_i H_i T,$$

josta todennäköisyyden P additiivisuuden (Määritelmä 2.3) perusteella seuraa kaava

$$P(T) = P\left(\sum_i H_i T\right) = \sum_i P(H_i T).$$

Kun kaavaan sijoitetaan

$$P(H_i T) = P(H_i)\, P(T \mid H_i),$$

saadaan (3.3.3). □

Jos kaavassa (3.3.3) jokin $P(H_i) = 0$, vastaava summan termi on 0, vaikka $P(T \mid H_i)$ ei olekaan määritelty. Kaavaa (3.3.3) kutsutaan *kokonaistodennäköisyyden kaavaksi.*

Olkoot X ja Y kokonaislukuarvoiset satunnaismuuttujat ja k jokin kokonaisluku. Soveltamalla kaavaa (3.3.3) tapahtumiin

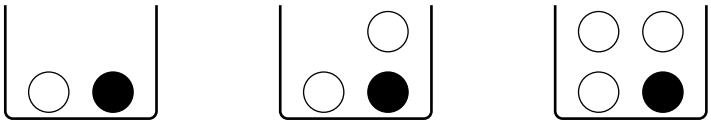
$$H_i = \{X = i\}, \qquad T = \{Y = k\}$$

saadaan

$$P(Y = k) = \sum_i P(X = i)\, P(Y = k \mid X = i), \tag{3.3.4}$$

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 3.9 (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



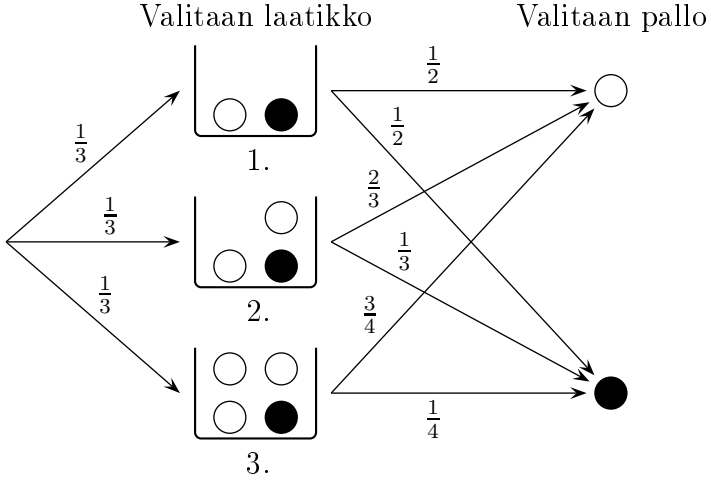
Laatikko 1 Laatikko 2 Laatikko 3

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

$$P(H_i \mid T) = \frac{P(H_i T)}{P(T)}, \qquad i = 1, 2, 3. \tag{3.3.5}$$

Seuraavassa kuviossa on esitetty havainnollisesti tilanteeseen liittyvät todennäköisyydet.



Kaavassa (3.3.5) osoittaja on

$$P(H_iT) = P(H_i)\,P(T \mid H_i) = \frac{1}{3} \cdot \frac{i}{i+1}, \qquad i = 1, 2, 3.$$

Koska $\sum_{i=1}^{3} H_iT = T$ 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 (3.3.5) 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*

$$P(H_i \mid T) = \frac{P(H_i)\,P(T \mid H_i)}{\sum_{j=1}^{k} P(H_j)\,P(T \mid H_j)}. \tag{3.3.6}$$

Todistus. Todennäköisyyksien tulokaavan nojalla sadaan

$$P(H_iT) = P(H_i)\,P(T \mid H_i) = P(T)\,P(H_i \mid T),$$

mistä seuraa

$$P(H_i \mid T) = \frac{P(H_i)\,P(T \mid H_i)}{P(T)}.$$

Väittämän 3.2 mukaan $P(T) = \sum_j P(H_j)\,P(T \mid H_j)$, joten kaava (3.3.6) on todistettu. □

Kaavaa (3.3.6) kutsutaan Bayesin säännöksi. Tapahtumat H_1, H_2, …, 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.

3.3.3 Peräkkäisotanta

Populaatiossa on N henkilöä, joista Np $(0 \leq p \leq 1)$ 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 $1 \leq i \leq n$ ja $1 \leq n \leq N$. Tarkastellaan siis satunnaismuuttujien jonoa $\{X_1, X_2, \ldots, X_n\}$ tai lyhyesti $\{\, X_i \mid 1 \leq i \leq n \,\}$. 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

$$P(A_2) = P(A_1)\,P(A_2 \mid A_1) + P(A_1^c)\,P(A_2 \mid A_1^c). \tag{3.3.7}$$

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 (3.3.7), 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:

$$P(A_i) = p, \qquad i = 1, 2, \ldots, n; \quad 1 \leq n \leq N. \tag{3.3.8}$$

Näytämme nyt, että tämä yleinen tulos pitää paikkansa. Voimme ajatella, että B:n kannattajat on numeroitu $1, 2, \ldots, Np$ ja muut $Np+1, Np+2, \ldots, N$. 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 (3.3.8)

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 4.4 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*.

3.3.4 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 3.3 *Samassa otosavaruudessa määriteltyjen tapahtumien A_1, A_2, …, A_n unionin $\bigcup_{i=1}^{n} A_i$ todennäköisyys on*

$$\begin{aligned} (3.3.9) \quad P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i) - \sum_{j>i}^{n} P(A_iA_j) + \sum_{k>j>i}^{n} P(A_iA_jA_k) \\ - + \cdots + (-1)^{n-1}\,P(A_1A_2\cdots A_n). \end{aligned}$$

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 $A_1^c A_2^c \cdots A_n^c$ indikaattorifunktio on $\prod_{i=1}^n (1 - \alpha_i)$. Koska $\bigcup_{i=1}^n A_i = (A_1^c A_2^c \cdots A_n^c)^c$, niin sen indikaattorifunktio on

$$\begin{aligned} (3.3.10) \qquad I_{\bigcup A_i} &= 1 - \prod_{i=1}^n (1 - \alpha_i) \\ &= \sum_{i=1}^n \alpha_i - \sum_{j>i}^n \alpha_i \alpha_j + \sum_{k>j>i}^n \alpha_i \alpha_k \alpha_k \\ &\qquad - + \cdots + (-1)^{n-1} \alpha_1 \alpha_2 \cdots \alpha_n. \end{aligned}$$

Kun nyt yhtälössä (3.3.10) otetaan odotusarvot puolittain ja käytetään hyväksi odotusarvon lineaarisuutta, saadaan tulos (3.3.9). Huomaa, että indikaattorifunktion I_A odotusarvo $E(I_A) = P(A)$ on vastaavan tapahtuman todennäköisyys. Silloin $E(I_{\bigcup A_i}) = P\big(\bigcup_{i=1}^n A_i\big)$, $E(\alpha_i) = P(A_i)$, $E(\alpha_i \alpha_j) = P(A_i A_j)$, $\ldots$, $E(\alpha_1 \alpha_2 \cdots \alpha_n) = P(A_1 A_2 \cdots A_n)$. □

Esimerkki 3.10 (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, $\ldots$, 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

$$(3.3.11) \qquad P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Jos kiinnitetään i-pari ja j-pari $(i \neq j)$, niin loput $(n-2)$ korttia voidaan permutoida $(n-2)!$ tavalla. Silloin

$$(3.3.12) \qquad P(A_i A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

Vastaavalla tavalla voidaan laskea todennäköisyys, että saadaan i-pari, j-pari ja k-pari $(i \neq j \neq k)$:

$$(3.3.13) \qquad P(A_i A_j A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}$$

ja yleisesti

$$P(A_{i_1}A_{i_2}\dots A_{i_m}) = \frac{(n-m)!}{n!} = \frac{1}{n(n-1)\cdots(n-m+1)}, \qquad 1 \le m \le n.$$

Todennäköisyys, että saadaan ainakin yksi pari on siis Lauseen 3.3 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. □

Ehdollinen todennäköisyys 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
$$\begin{aligned} P(A_1A_2A_3\cdots A_{n-1}A_n) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1A_2)\cdots \\ \cdot P(A_n \mid A_1A_2\cdots A_{n-1}). \end{aligned}$$
- Riippumattomuus. A ja B ovat riippumattomat, jos $P(AB) = P(A)\,P(B)$.
- $P(A_1$ tai A_2 tai $A_3)$
$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) \\ - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3). \end{aligned}$$

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. 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?

2. 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 virhe sattui jo ensimmäisessä siirrossa?

3. Tarkastellaan kaksilapsisia perheitä. Oletetaan pojat ja tytöt yhtä todennäköisiksi ja 2. lapsen sukupuoli on riippumaton 1. lapsen sukupuolesta. Tarkastellaan neljää tapahtumaa:

$$\begin{aligned} A &= \text{1. lapsi on poika,} \\ B &= \text{lapset ovat eri sukupuolta,} \\ C &= \text{1. lapsi on tyttö,} \\ D &= \text{2. lapsi on poika.} \end{aligned}$$

(a) Mitkä tapahtumaparit $\{A, B\}$, $\{A, C\}$, $\{B, C\}$ ovat keskenään riippumattomat?

(b) Ovatko tapahtumat A, B ja D keskenään eli täydellisesti riippumattomat?

4. 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?

5. 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?

Luku 4

Satunnaismuuttujien tunnusluvut ja riippumattomuus

Tässä luvussa käsitellään satunnaismuuttujien ominaisuuksia. Erityisesti satunnaismuuttujien odotusarvo on keskeinen käsite. Tarkastelujen painopiste on diskreetteissä satunnaismuuttujissa ja kaikkia vastaavia tuloksia ei toisteta jatkuvien satunnaismuuttujien tapauksessa. Tulosten todistaminen ja soveltaminen on yleensä huomattavasti yksinkertaisempaa diskreettien satunnaismuuttujien yhteydessä.

4.1 Odotusarvo, varianssi ja kovarianssi

4.1.1 Odotusarvo

Numeroituvassa otosavaruudessa Ω määritellyn satunnaismuuttujan X *odotusarvo* on

$$E(X) = \sum_{\omega \in \Omega} X(\omega)\, P(\{\omega\}), \tag{4.1.1}$$

jos

$$\sum_{\omega \in \Omega} |X(\omega)\, P(\{\omega\})| < \infty. \tag{4.1.2}$$

Jos ehto (4.1.2) toteutuu, sarja (4.1.1) 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

$$E(X) = \sum_i P(A_i)x_i = \sum_i p_i x_i. \tag{4.1.3}$$

Tämä kaava saadaan ryhmittelemällä alkeistapaukset kaavassa (4.1.1) osajoukkoihin A_i ja summaamalla sitten yli indeksin i.

Kaavasta (4.1.1) 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*

$$E(X^r) = \sum_i p_i x_i^r. \tag{4.1.4}$$

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

Määritelmä 4.1 (Odotusarvo) Jos X on *diskreetti satunnaismuuttuja*, jonka arvojoukko on S ja todennäköisyysfunktio f, niin X:n odotusarvo μ_X on

$$\mu_X = E(X) = \sum_{x_i\in S} x_i f(x_i) = \sum_{x_i\in S} x_i\, P(X = x_i),$$

jos summa suppenee itseisesti.

Jos X on *jatkuva satunnaismuuttuja*, jonka tiheysfunktio on f, niin X:n odotusarvo on

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \text{ mikäli integraali on olemassa.}$$

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 4.1 *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 $E(X+Y) = E(X) + E(Y)$:

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

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 4.1 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 4.1 (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 4.2 Indikaattorifunktio (Määritelmä 2.5) 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 4.1, 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 4.3 Satunnaismuuttuja X noudattaa diskreettiä tasajakaumaa Tasd$(1, N)$, kun $P(X = i) = \frac{1}{N}$, $i = 1, 2, \ldots, N$ (ks. alaluku 2.14). 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 4.4 Hypergeometrinen jakauma esiteltiin tarkasteltaessa otantaa palauttamatta (alaluku 2.7.1). 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

$$P(X = x; N, n, p) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, \ldots, n, \tag{4.1.5}$$

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 (4.1.5):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 (2.4.9), 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

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

□

Esimerkki 4.5 Alaluvussa 3.3.3 tarkasteltiin peräkkäisotantaa äärellisestä populaatiosta. 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.). 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 $1 \le n \le N$.

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 4.4 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.* □

4.1.2 Ehdollinen 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

$$E(X \mid A) = \sum_x x f(x \mid A). \tag{4.1.6}$$

Esimerkki 4.6 Oletetaan, että $X \sim \text{Tasd}(1, N)$ 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 4.2 *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

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

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

4.1.3 Varianssi

Varianssin laskemiseksi tarvitaan funktion $h(X) = X^2$ odotusarvo (Vertaa Lauseen 4.1 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 4.1 *Oletetaan, että satunnaismuuttujilla X ja Y on 2. momentti ja $c \in \mathbb{R}$ on vakio. Silloin odotusarvot*

$$E[(cX)^2], \quad E[(X+Y)^2], \quad E(X), \quad E(Y) \quad ja \quad E(XY) \tag{4.1.7}$$

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 4.1 (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 4.1 (kohta 3) mukaan

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

joten $E(XY)$ on olemassa.

□

Lause 4.3 (Cauchyn ja Schwarzin epäyhtälö) *Jos satunnaismuuttujilla X ja Y on 2. momentti, niin*

$$[E(XY)]^2 \le E(X^2)\,E(Y^2). \tag{4.1.8}$$

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 4.1 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 (4.1.8) pitää triviaalisti paikkansa. □

Yhtäsuuruus (4.1.8):ssä vallitsee silloin, kun $aX=-bY$ (todennäköisyydellä 1). Silloin $Y=-\frac{a}{b}X$, jos $b \neq 0$. Epäyhtälössä (4.1.8) pätee siis yhtäsuuruus, kun X ja Y ovat lineaarisesti riippuvia. Epäyhtälö (4.1.8) voidaan lausua myös muodossa

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

Määritelmä 4.2 (Varianssi) Jos satunnaismuuttujalla X on 2. momentti $E(X^2)$, niin sillä on odotusarvo μ_X ja X:n varianssi on

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X-\mu_X)^2]. \tag{4.1.9}$$

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

$$\sigma^2 = \mathrm{Var}(X) = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2. \tag{4.1.10}$$

satunnaismuuttujan X hajonta $\sigma_X = \sqrt{\mathrm{Var}(X)}$. Odotusarvon määritelmästä ja identiteetistä (4.1.10) saamme erittäin käyttökelpoisen tuloksen:

$$\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X), \qquad E(X^2) = \mu^2 + \mathrm{Var}(X). \tag{4.1.11}$$

Esimerkki 4.7 Lasketaan diskreettiä tasajakaumaa $\mathrm{Tasd}(1, N)$ noudattavan satunnaismuuttujan varianssi. Esimerkin 4.3 mukaan

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

Soveltamalla kaavaa (4.1.10) saadaan

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

□

4.1.4 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 4.1 nojalla.

Määritelmä 4.3 (Kovarianssi) Satunnaismuuttujien X ja Y *kovarianssi* σ_{XY} määritellään odotusarvona

$$\begin{aligned} \sigma_{XY} &= \mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY) - \mu_X\mu_Y. \end{aligned} \tag{4.1.12}$$

Kovarianssin avulla voidaan sitten määritellä korrelaatiokerroin.

Määritelmä 4.4 (Korrelaatiokerroin) Satunnaismuuttujien X ja Y *korrelaatiokerroin*

$$\rho_{XY} = \mathrm{Cor}(X,Y) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \tag{4.1.13}$$

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

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

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathrm{Cov}(X_i, X_j). \tag{4.1.14}$$

Todistus. Todistetaan 1. kohta. Määritelmän mukaan

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

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

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

Kaava (4.1.14) voidaan todistaa induktiolla. □

4.2 Satunnaismuuttujan funktio

Lauseen 4.1 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 4.8 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:

$$h^{-1}(A) = \{\, x \in S_X \mid h(x) \in A \,\}. \tag{4.2.1}$$

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 $h^{-1}(y) = x$.

4.3 Satunnaismuuttujien identtisyys

Määritelmä 4.5 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 4.9 Esimerkissä 2.7 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 4.4 *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 4.10 Heitetään harhatonta lanttia 4 kertaa. Olkoon kruunun todennäköisyys p. X ja Y on määritelty samoin kuin Esimerkissä 4.9. 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

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

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 $p = \frac{1}{2}$. □

4.4 Satunnaismuuttujien riippumattomuus

Määrittelimme tapahtumien riippumattomuuden alaluvussa 3.1.2. Tarkastelemme nyt satunnaismuuttujien riippumattomuutta.

4.4.1 Kaksi satunnaismuuttujaa

Määritelmä 4.6 (Satunnaismuuttujien riippumattomuus) Satunnaismuuttujat X ja Y ovat riippumattomat jos

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B) \tag{4.4.1}$$

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 $B \subset \mathbb{R}$.

Jos X ja Y ovat diskreettejä, niin riippumattomuuden määritelmän nojalla

$$P(X = x, Y = y) = P(X = x)\, P(Y = y) = f_X(x) f_Y(y) \tag{4.4.2}$$

kaikilla $x, y \in \mathbb{R}$, missä $f_X(x)$ on X:n ja $f_Y(y)$ on Y:n todennäköisyysfunktio. Diskreettien satunnaismuuttujien X ja Y yhteisjakauman todennäköisyysfunktio määritellään:

$$P(X = x, Y = y) = f_{X,Y}(x, y) \tag{4.4.3}$$

kaikilla $x, y \in \mathbb{R}$. Huomattakoon, että $f_{X,Y}(x, y) > 0$ täsmälleen silloin, kun $(x, y) \in S_X \times S_Y$ ja muutoin $f_{X,Y}(x, y) = 0$. Diskreetit satunnaismuuttujat X ja Y ovat riippumattomat silloin ja vain silloin kun

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{4.4.4}$$

kaikilla $x, y \in \mathbb{R}$.

Lause 4.5 *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 4.6 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 (4.4.1) 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 (4.4.4) voidaan myös pitää diskreettien satunnaismuuttujien X ja Y riippumattomuuden määritelmänä, sillä siitä seuraa identiteetti (4.4.1). Jos valitaan kaksi mielivaltaista numeroituvaa joukkoa $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$ sekä oletetaan (4.4.4), saadaan

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x_i \in A} \sum_{y_j \in B} P(X = x_i, Y = y_j) \\ &= \sum_{x_i \in A} \sum_{y_j \in B} P(X = x_i)\, P(Y = y_j) \qquad [(4.4.4)] \\ &= \sum_{x_i \in A} P(X = x_i) \sum_{y_j \in B} P(Y = y_j) \\ &= P(X \in A)\, P(Y \in B). \end{aligned}$$

Näin olemme todenneet, että ehdot (4.4.1) ja (4.4.4) 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

$$P(I_A = x, I_B = y) = P(I_A = x)\, P(I_B = y) \tag{4.4.5}$$

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.

4.4.2 Useita satunnaismuuttujia

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

$$(4.4.6)\quad P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)\, P(X_2 \in A_2) \cdots P(X_n \in A_n)$$

kaikilla (sopivasti valituilla) joukoilla $A_i \subset \mathbb{R}$, $1 \le i \le n$. Jos $X_1, \ldots, X_n$ ovat diskreettejä, niin (4.4.6) pitää paikkansa kaikille joukoille $A_i \subset \mathbb{R}$, $1 \le i \le n$. Yleisessä tapauksessa on A_i:t ($1 \le i \le n$) 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 [$1 \le k \le n$ 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 (4.4.6) mukaan

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

4.5 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

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

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 4.6 (Markovin epäyhtälö) *Olkoon $X \ge 0$ epänegatiivinen satunnaismuuttuja. Silloin*

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

Todistus. Olkoon I_A joukon $A = \{\, \omega \mid X(\omega) \ge a \,\}$ indikaattorifunktio [ks. (2.5)]. Koska sekä indikaattorifunktio että X ovat epänegatiiviset ja $I_A + I_{A^c} = 1$, niin

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

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 4.1, 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 4.7 (Tšebyševin epäyhtälö) *Olkoon X satunnaismuuttuja, jonka keskiarvo on μ ja varianssi σ^2. Silloin*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \qquad \textit{kaikilla } \varepsilon > 0. \tag{4.5.1}$$

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 (4.5.1) suoraan Markovin epäyhtälöstä. □

Lause 4.8 (Riippumattomat satunnaismuuttujat, tulon odotusarvo) *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 4.3 (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 4.2. □

Apulause 4.4 (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 4.9 (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 4.4 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.

4.6 Generoivat funktiot ja momentit

4.6.1 Momentit

Eräs tapa luonnehtia satunnaismuuttujan jakaumaa, on laskea jakauman momentit. Ne määritellään odotusarvon avulla.

Määritelmä 4.7 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

$$\begin{aligned} g_1 &= E(X) = \alpha_1 = \mu, \\ g_2 &= E[X(X-1)] = E(X^2 - X) = E(X^2) - E(X) = \alpha_2 - \mu. \end{aligned}$$

Koska $\sigma^2 = \alpha_2 - \mu^2$, niin

$$\sigma^2 = g_2 + \mu - \mu^2.$$

4.6.2 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 4.8 Olkoon X satunnaismuuttuja, jonka tiheysfunktio on $f(x)$. Reaalimuuttujan t funktio

$$M(t) = E(\mathrm{e}^{tX})$$

on satunnaismuuttujan X (tai X:n jakauman) momenttifunktio (mf), jos odotusarvo

$$E(\mathrm{e}^{tX}) = \begin{cases} \sum_i \mathrm{e}^{tx_i} f(x_i) & \text{diskreetti satunnaismuuttuja} \\ \int_{-\infty}^{\infty} \mathrm{e}^{tx} f(x)\,\mathrm{d}x, & \text{jatkuva satunnaismuuttuja} \end{cases}$$

on olemassa jollain avoimella välillä $-a < t < a$, missä $a > 0$.

Määritelmän perusteella on selvää, että

$$M(0) = \sum_i f(x_i) = 1, \text{kun } X \text{ diskreetti ja } M(0) = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1,$$

kun X on jatkuva. 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 4.10 *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 4.11 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 4.11 *Olkoot X ja Y riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat $M_X(t)$ ja $M_Y(t)$. Silloin satunnaismuuttujan $Z = X + Y$ momenttifunktio on*

$$M_Z(t) = M_X(t) M_Y(t). \tag{4.6.1}$$

Todistus. Koska e^{tX} on pelkästään x:n (X:n arvojen) funktio ja e^{tY} pelkästään y:n funktio, niin Lauseen 4.5 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 4.8. □

Usean satunnaismuuttujan tapauksessa on voimassa vastaava tulos.

Seuraus 4.1 *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

$$M(t) = \sum_{x\in S} \mathrm{e}^{tx} f(x) \tag{4.6.2}$$

derivoidaan puolittain, voidaan oikea puoli derivoida termeittäin ja yhtäsuuruus säilyy. Derivoimalla lauseke (4.6.2) 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 4.12 *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 4.13 (Lévyn jatkuvuuslause) *Olkoon $X_1, X_2, \ldots$ jono satunnaismuuttujia, joiden kertymäfunktiot ovat $F_{X_1}, F_{X_2}, \ldots$ ja vastaavasti momenttifunktiot $M_{X_1}(t), M_{X_2}(t), \ldots$. Olkoon X satunnaismuuttuja, jonka kertymäfunktio on F_X ja momenttifunktio $M_X(t)$. Jos n:n kasvaessa rajatta*

$$M_{X_n}(t) \to M_X(t)$$

kaikilla t:n arvoilla jossain nollan ympäristössä $(-h, h)$, $h > 0$, niin silloin

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

kaikissa pisteissä x, joissa $F_X(x)$ on jatkuva.

Satunnaismuuttujien momenttifunktioiden suppenemisesta seuraa siis satunnaismuuttujien kertymäfunktioiden suppeneminen. Tällöin sanomme, että satunnaismuuttujat $X_1, X_2, \ldots$ suppenevat jakaumamielessä kohti satunnaismuuttujaa X.

4.6.3 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

$$G(\mathrm{e}^t) = E(\mathrm{e}^{tX}) = M(t).$$

4.7 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

$$P(\omega_i, \omega_j) = p_i q_j \tag{4.7.1}$$

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

$$\sum_{(\omega_i, \omega_j) \in \Omega_1 \times \Omega_2} P(\omega_i, \omega_j) = \sum_{\omega_i \in \Omega_1} \sum_{\omega_j \in \Omega_2} p_i q_j = \Bigg(\sum_{\omega_i \in \Omega_1} p_i \Bigg) \Bigg(\sum_{\omega_j \in \Omega_2} q_j \Bigg) = 1.$$

Identiteetti (4.7.1) 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

$$X = \begin{cases} 1, & \text{todennäköisyydellä } P(\mathrm{O}) = p; \\ 0, & \text{todennäköisyydellä } 1 - p, \end{cases} \tag{4.7.2}$$

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 4.12 Esimerkissä 2.7 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 4.4 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.

Satunnaismuuttujien tunnusluvut: Yhteenveto

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

$$[E(XY)]^2 \le E(X^2)\,E(Y^2).$$

- 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 $\quad \alpha_r = E(X^r)$,

  r. keskusmomentti $\quad \mu_r = E(X-\mu)^r$,

  r. tekijämomentti $\quad g_r = E[X^{(r)}] = E[X(X-1)\cdots(X-r+1)]$.

- 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. $M(t)$: $\quad G(\mathrm{e}^t) = E(\mathrm{e}^{tX}) = M(t)$.

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. Tarkastellaan Esimerkin 2.10 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?

6. Tarkastellaan Tehtävän 5 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.

7. 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 $\text{Var}(X)$.

8. 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 $\text{Var}(X)$.

9. 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 $P(X \leq Y)$.

10. 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 $\frac{8}{9}$.

11. 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 $\beta = \rho\sigma_Y/\sigma_X$.

12. 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 5

Diskreettejä yksiulotteisia jakaumia

Diskreetti satunnaismuuttuja määriteltiin alaluvussa 2.6.1. Olemme jo edellisissä luvuissa käsitelleet hypergeometrista jakaumaa (alaluku 2.7.1), binomijakaumaa (alaluvut 2.9 ja sen erikoistapauksena Bernoullin jakaumaa sekä diskreettiä tasajakaumaa (alaluku 2.14), jotka kaikki ovat esimerkkejä *diskreeteistä jakaumista.*

5.1 Diskreetti satunnaismuuttuja

Otosavaruudessa Ω määritellyn diskreetin satunnaismuuttujan X 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

$$F(x) = \sum_{i=1}^{\infty} p_i \, \varepsilon(x - i). \tag{5.1.1}$$

Vaikka usein tarkastelemme vain kokonaislukuarvoisia satunnaismuuttujia, se ei ole 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 $P(X = x_i) = P\big(g(X) = i\big)$, joten voimme aina tarvittaessa siirtyä tarkastelemaan vastaavaa kokonaislukuarvoista satunnaismuuttujaa.

Esimerkki 5.1 Yksinkertaisin satunnaismuuttuja X on sellainen, jonka arvoalue $S = \{c\}$ on yksi piste, jolloin $P(X = c) = 1$. Silloin X:n kertymäfunktio on

$$F(x) = \varepsilon(x - c) = \begin{cases} 1, & x \geq c; \\ 0, & x < c. \end{cases}$$

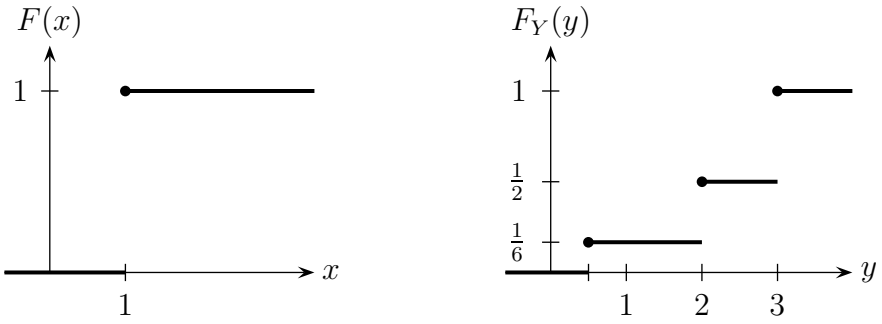
Olkoon Y:n todennäköisyysfunktio

$$P\left(Y = \tfrac{1}{2}\right) = \tfrac{1}{6}, \qquad P(Y = 2) = \tfrac{1}{3} \qquad \text{ja} \qquad P(Y = 3) = \tfrac{1}{2}.$$

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 5.1. Funktioiden $F(x) = \varepsilon(x - 1)$ ja $F_Y(y)$ kuvaajat.

Esimerkki 5.2 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

$$\begin{aligned} E(X) &= N^{-n} \sum_{r=1}^{N} [r^n - (r-1)^n] r \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n r\right] \\ &= N^{-n} \sum_{r=1}^{N} \left[r^{n+1} - (r-1)^n \big((r-1) + 1\big)\right] \end{aligned}$$

$$= N^{-n} \sum_{r=1}^{N} \big[r^{n+1} - (r-1)^{n+1} - (r-1)^n \big]$$

$$= N^{-n} \Big[N^{n+1} - \sum_{r=1}^{N} (r-1)^n \Big].$$

□

5.2 Bernoullin kokeet ja binomijakauma

Alaluvussa 2.9 binomijakauma esiteltiin tarkastelemalla otantaa palauttaen ja alaluvussa 4.7 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

$$X = \begin{cases} 1 & \text{todennäköisyydellä } p, \\ 0 & \text{todennäköisyydellä } 1-p, \end{cases} \tag{5.2.1}$$

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

$$\begin{aligned} M(t) = E(\mathrm{e}^{tX}) &= P(X=0)\mathrm{e}^{t\cdot 0} + P(X=1)\mathrm{e}^{t\cdot 1} \\ &= (1-p) + p\mathrm{e}^t = 1 + p(\mathrm{e}^t - 1), \end{aligned}$$

joka on määritelty kaikilla $t \in \mathbb{R}$.

Esimerkki 5.3 (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 $111011000\ldots110$. 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 4.7 totesimme, että X noudattaa binomijakaumaa parametrein n ja p. Silloin merkitään $X \sim \mathrm{Bin}(n, p)$. Binomijakauman todennäköisyysfunktio on

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n. \tag{5.2.2}$$

Esitetään nyt edellä mainittu binomijakauman luonnehdinta Bernoullin kokeiden avulla lauseen muodossa. Jatkossa oletetaan, että Bernoullin kokeet ovat toisistaan riippumattomat, vaikkei oletusta erikseen mainittaisikaan.

Lause 5.1 *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 5.4 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 2.6.1. 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 5.2 *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 5.1 todistuksessa.

2. Odotusarvo ja varianssi. Koska $X = X_1 + X_2 + \cdots + X_n$ on riippumattomien Bernoullin muuttujien $X_i \sim \mathrm{Ber}(p)$ summa, niin

$$\begin{aligned} E(X) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= p + p + \cdots + p = np \end{aligned}$$

ja

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p). \end{aligned}$$

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 4.5 ja 4.8. Koska X_i ja X_j ($i \neq j$) ovat riippumattomat, niin e^{tX_i} ja e^{tX_j} ovat riippumattomat (Lause 4.5) ja riippumattomien satunnaismuuttujien e^{tX_1}, e^{tX_2}, …, e^{tX_n} tulon odotusarvo on yksittäisten tulon tekijöiden odotusarvojen tulo (Lause 4.8). 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 4.10). 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 5.1 *Jos* $X_1 \sim \mathrm{Bin}(n_1, p)$ *ja* $X_2 \sim \mathrm{Bin}(n_2, p)$ *ovat riippumattomat, niin* $X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p)$.

Todistus. Koska Lauseen 5.2 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 4.11 mukaan $(1-p+pe^t)^{n_1+n_2}$. Mutta Lauseen 5.2 perusteella $(1-p+pe^t)^{n_1+n_2}$ on binomijakuman $\mathrm{Bin}(n_1+n_2,p)$ momenttifunktio. Tästä seuraa momenttifunktion yksikäsitteisyyden (Lause 4.10) nojalla, että $X_1+X_2 \sim \mathrm{Bin}(n_1+n_2,p)$. □

Seurauslauseen 5.1 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 $X_1+X_2 \sim \mathrm{Bin}(n_1+n_2,p)$. 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 2.7)

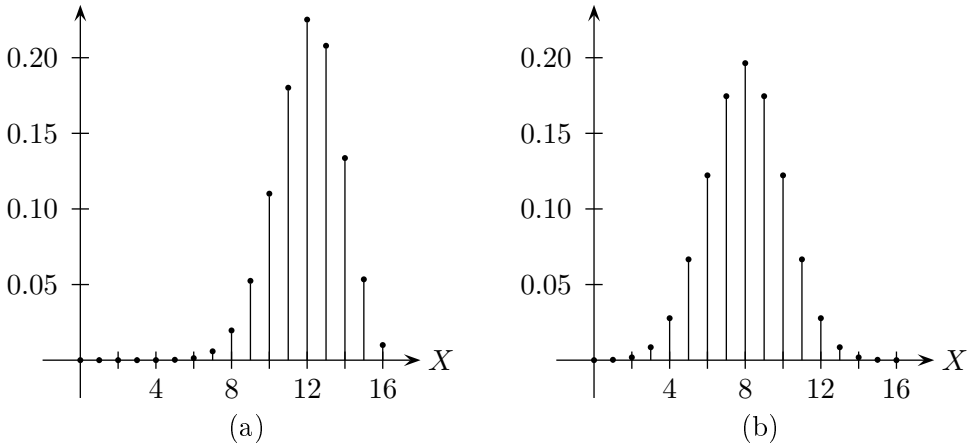
$$\binom{n_1+n_2}{k} = \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{k-i}$$

saadaan kaivattu tulos.

5.2.1 Jakauman symmetria

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



Kuvio 5.2. Binomijakauman kuvaajat, kun (a) $X \sim \text{Bin}(16, 0.75)$ (b) $X \sim \text{Bin}(16, 0.50)$.

on symmetrinen b:n suhteen jos ja vain jos $X - b$ on symmetrinen origon suhteen. Silloin

$$P(X \leq b - x) = P(X \geq b + x).$$

Esimerkiksi binomijakauma $\text{Bin}(16, 0.50)$ on symmetrinen pisteen 8 suhteen, mutta binomijakauma $\text{Bin}(16, 0.75)$ ei ole symmetrinen (Kuvio 5.4). Binomijakaumassa $\text{Bin}(16, 0.50)$ on

$$P(X = 8 + x) = P(X = 8 - x)$$

kaikilla x. Silloin jokaista $a \in \mathbb{R}$ kohti

$$P(X \leq 8 - a) = P(X \geq 8 + a).$$

5.3 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.

5.3.1 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 $S_n \sim \mathrm{Bin}(n,p)$, mutta mikä on W_r:n jakauma?

Esimerkki 5.5 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:

$$\underbrace{\mathrm{LLL}\ldots\mathrm{L}}_{x-1 \text{ kertaa}}\ \mathrm{R}.$$

Tästä seuraa, että

$$P(W_1 = x) = \frac{1}{2^x}, \qquad x = 1, 2, \ldots.$$

Satunnaismuuttujan W_1 odotusarvo on määritelmän mukaan

$$E(W_1) = \sum_{x=1}^{\infty} \frac{x}{2^x}. \tag{5.3.1}$$

Tiedämme, että

$$\sum_{x=0}^{\infty} p^x = 1 + p + p^2 + p^3 + \cdots = \frac{1}{1-p}, \qquad \text{kun } |p| < 1. \tag{5.3.2}$$

Kun derivoimme sarjan (5.3.2) termeittäin, saamme

$$0 + 1 + 2p + 3p^2 + \cdots = \sum_{x=0}^{\infty} (x+1)p^x = \frac{1}{(1-p)^2}, \qquad \text{kun } |p| < 1. \tag{5.3.3}$$

Koska sarjan (5.3.2) suppenemissäde on 1, suppenee derivointioperaation tuloksena saatu sarja (5.3.3) arvoilla $|p| < 1$. Sijoittamalla $p = \frac{1}{2}$ sarjaan (5.3.3) 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 (5.3.2). Nyt siis odotusarvo (5.3.1) 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

$$E(W_1) = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=0}^{\infty}(x+1)(1-p)^x$$
$$= p \cdot \frac{1}{[1-(1-p)]^2} = \frac{1}{p},$$

missä sarjan summa saadaan (5.3.3):n avulla. Satunnaismuuttuja W_1 on siis kruunun tai yleisemmin 'onnistumisen' odotusaika. Jakaumaa

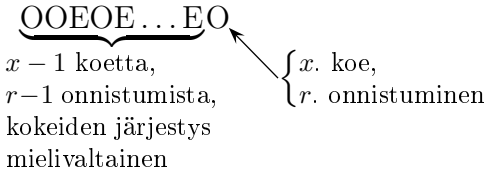
$$P(W_1 = x) = (1-p)^{x-1}p, \qquad x = 1, 2, \ldots \tag{5.3.4}$$

kutsutaan *geometriseksi jakaumaksi*. Todennäköisyydet (5.3.4) 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

$$P(W_r = x) = P(S_{x-1} = r-1)\,P(X_x = 1) \tag{5.3.5}$$
$$= \binom{x-1}{r-1}p^{r-1}(1-p)^{x-r}p = \binom{x-1}{r-1}p^r(1-p)^{x-r},$$

koska $S_{x-1} \sim \mathrm{Bin}(x-1, p)$. Todennäköisyydet (5.3.5) 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

$$P(W_r > x) = P(S_x < r).$$

5.3.2 Geometrinen jakauma ja negatiivinen binomijakauma

Sanomme, että satunnaismuuttuja X noudattaa *negatiivista binomijakaumaa* parametrein r ja p, jos

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, r+2, \ldots. \tag{5.3.6}$$

Merkitsemme silloin

$$X \sim \mathrm{NBin}(r, p).$$

Edellisessä pykälässä huomasimme, että odotusaika $W_r \sim \mathrm{NBin}(r, p)$. Kun $r = 1$, sanomme negatiivista binomijakaumaa *geometriseksi jakaumaksi.* Geometrisen jakauman todennäköisyysfunktio on siis

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots. \tag{5.3.7}$$

Kun siis $X \sim \mathrm{NBin}(1, p)$, niin X:n noudattaa geometrista jakaumaa parametrilla p. Merkitsemme silloin $X \sim \mathrm{Geo}(p)$.

Lause 5.3 *Oletetaan, että* $X \sim \mathrm{NBin}(r, p)$.

1. *Funktio (5.3.6) on negatiivisen binomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla* r *ja kaikilla* $0 < p < 1$ *ja*

2. $$E(X) = \frac{r}{p}, \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2},$$
$$M(t) = E(\mathrm{e}^{tX}) = \frac{(p\mathrm{e}^t)^r}{[1-(1-p)\mathrm{e}^t]^r}, \qquad t < -\log(1-p).$$

Todistus. Johdamme ensin negatiivisen binomijakauman momenttifunktion suoraan määritelmän nojalla. Koska $M(t) = E(\mathrm{e}^{tX})$, niin momenttifunktio

on

$$\begin{aligned}
E(\mathrm{e}^{tX}) &= \sum_{x=r}^{\infty} \mathrm{e}^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= p^r \sum_{y=0}^{\infty} \mathrm{e}^{t(y+r)} \binom{r+y-1}{r-1} p^r (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} \binom{r+y-1}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \mathrm{e}^{ty} (-1)^y \binom{-r}{y} (1-p)^y \\
&= p^r \mathrm{e}^{tr} \sum_{y=0}^{\infty} \binom{-r}{y} \big[-(1-p)\mathrm{e}^t\big]^y \\
&= p^r \mathrm{e}^{tr} \big[1-(1-p)\mathrm{e}^t\big]^{-r} = \left[\frac{p\mathrm{e}^t}{1-(1-p)\mathrm{e}^t}\right]^r.
\end{aligned}$$

Binomisarja $\sum_{y=0}^{\infty} \binom{-r}{y} \big[-(1-p)\mathrm{e}^t\big]^y$ suppenee (Lause **??**), 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 (5.3.6) 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 5.2 *Jos* $X \sim \mathrm{Geo}(p)$*, niin* $X \sim \mathrm{NBin}(1,p)$ *ja*

1. *funktio (5.3.7) 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 5.6 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

$$\begin{aligned} P(W > k) = \sum_{j=k+1}^{\infty} q^{j-1} p &= q^k (p + qp + q^2 p + \cdots) \\ &= q^k = 1 - P(W \leq k). \end{aligned}$$

□

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:

$$P(W > k + s \mid W > k) = \frac{P(W > k + s)}{P(W > k)} = \frac{q^{k+s}}{q^k} = q^s. \tag{5.3.8}$$

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 5.7 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 (5.3.6) 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

$$2\,P(W_{N+1} = N + 1 + N - k) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}.$$

□

5.3.3 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

$S_n =$ valkoisten pallojen (onnistumisten) lukumäärä n:ssä ensimmäisessä nostossa;

$W_r = r$:n valkoisen pallon saamiseksi tarvittavien nostojen määrä.

Jos ajatellaan, että nostoon menee yksi aikayksikkö, niin W_r on r:n valkoisen pallon saamiseksi tarvittava odotusaika.

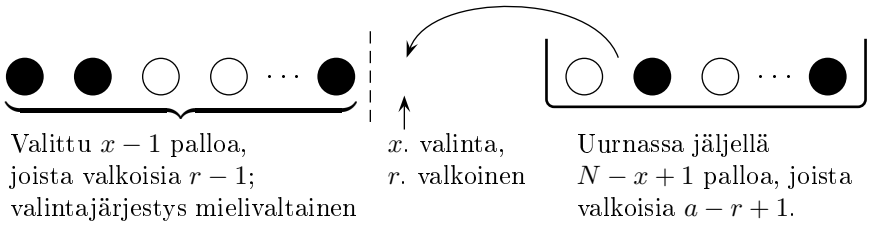
Jos otanta tehdään *palauttaen,* niin peräkkäiset nostot ovat riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on $p = a/N$. Tässä tapauksessa voidaan suoraan soveltaa edellä esitettyjä Bernoullin kokeita koskevia tuloksia.

Kun otanta tehdään *palauttamatta,* peräkkäiset nostot eivät ole riippumattomia, koska valkoisten pallojen suhteellinen osuus uurnassa riippuu siitä, mitä sieltä on jo valittu. Alaluvussa 2.7.1 osoitimme, että S_n noudattaa hypergeometristä jakaumaa, kun otanta tehdään palauttamatta (ks. myös alaluku 3.3.3). Silloin

$$P(S_n = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \tag{5.3.9}$$

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:



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 4.4 ja (4.1.5)] ja $X_x = 1$, kun valitaan valkoinen pallo x. nostossa. Tästä seuraa, että

$$\begin{aligned} P(W_r = x) &= P(S_{x-1} = r-1, X_x = 1) \\ &= P(S_{x-1} = r-1)\, P(X_x = 1 \mid S_x = r-1) \\ &= \frac{\binom{a}{r-1}\binom{N-a}{x-r}}{\binom{N}{x-1}} \cdot \frac{a-r+1}{N-x+1}, \end{aligned} \tag{5.3.10}$$

kun $x = r, r+1, \ldots, N$.

Todennäköisyys (5.3.10) voidaan kirjoittaa lausekkeena

$$P(W_r = x) = \binom{x-1}{r-1}\frac{\binom{N-x}{a-r}}{\binom{N}{a}}, \tag{5.3.11}$$

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 $p = a/N$.

5.3.4 Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma

Olemme esitelleet hypergeometrisen jakauman tarkastelemalla otantaa palauttamatta (alaluku 2.7.1). 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

$$f(x) = P(X = x) = \frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}, \qquad x \in S. \tag{5.3.12}$$

Huomattakoon, että todennäköisyys (5.3.12) on määritelty myös arvoilla $x \notin S$, mutta silloin $f(x) = 0$.

Lause 5.4 *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 4.4 ja alaluvussa 3.3.3. Varianssi voidaan laskea vastaavalla tavalla. □

Lause 5.5 *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 2.9.1, 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 4.4 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 5.1 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 5.6 *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 5.7 *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 5.6 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. □

5.3.5 Tasajakauma

Diskreetti tasajakauma esiteltiin ensimmäisen kerran alaluvussa 2.14. 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

$$E(X) = \frac{N+1}{2} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{(N+1)(N-1)}{12}.$$

5.4 Poissonin jakauma

Satunnaismuuttuja X, jonka todennäköisyysfunktio on

$$f(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots \tag{5.4.1}$$

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 5.8 *Olkoon* $X \sim \mathrm{Poi}(\lambda)$. *Silloin*

1. *funktio (5.4.1) 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

$$\exp(\lambda) = \mathrm{e}^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}. \tag{5.4.2}$$

1. Ensinnäkin $f(x) \geq 0$ kaikilla $x = 0, 1, 2, \ldots$, ja eksponenttifunktion sarjakehitelmän (5.4.2) 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 5.9 *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 4.1 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 4.10 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 5.9 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 $X \sim \mathrm{Bin}(n, p)$, niin binomitodennäköisyys on

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n. \tag{5.4.3}$$

Annetaan nyt p:n riippua n:stä ja merkitään lausekkeessa (5.4.3) $p = p_n$. Valitaan erityisesti

$$p_n = \frac{\lambda}{n}, \qquad n \geq 1.$$

Tarkastellaan nyt binomijakaumien jonoa

$$\mathrm{Bin}(1, p_1),\ \mathrm{Bin}(2, p_2),\ \mathrm{Bin}(3, p_3), \ldots$$

ja vastaavaa satunnaismuuttujien $X_1, X_2, X_3, \ldots$ jonoa, missä $X_n \sim \mathrm{Bin}(n, p_n)$, $n \geq 1$. Nyt siis

$$P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}, \qquad 0 \leq x \leq n. \tag{5.4.4}$$

Merkitään todennäköisyyttä (5.4.4) 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

$$\lim_{n\to\infty} b_0(n) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \mathrm{e}^{-\lambda}. \tag{5.4.5}$$

Se on eräs keskeinen eksponenttifunktioon liittyvä kaava, joka pitäisi analyysin kurssien perusteella muistaa. Tulos (5.4.5) saadaan esimerkiksi Taylorin sarjan

$$\log(1-p) = -\sum_{n=1}^{\infty} \frac{p^n}{n}$$

avulla, kun sijoitetaan $p = \frac{\lambda}{n}$:

$$\begin{aligned} \log\left(1 - \frac{\lambda}{n}\right)^n &= n \log\left(1 - \frac{\lambda}{n}\right) = n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \frac{\lambda^3}{3n^3} - \cdots\right) \\ &= -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \cdots \\ &= -\lambda - \frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \cdots\right). \end{aligned} \tag{5.4.6}$$

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

$$\lim_{n\to\infty} \frac{b_{x+1}(n)}{b_x(n)} = \frac{\lambda}{x+1}. \tag{5.4.7}$$

Kun lähdetään tuloksesta (5.4.5) ja käytetään hyväksi raja-arvoa (5.4.7), 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

$$\lim_{n\to\infty} b_x(n) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \tag{5.4.8}$$

missä raja-arvo on $P(X = x)$, kun $X \sim \mathrm{Poi}(\lambda)$. Tulos (5.4.8) tunnetaan *Poissonin raja-arvolakina.*

Satunnaismuuttujat noudattavat samaa jakaumaa, kun niillä on sama kertymäfunktio (Lause 4.4). 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 5.6). Vaikka edellä olemmekin johtaneet Poissonin raja-arvolain (5.4.8), esitetään tulos vielä *Poissonin lauseena.*

Lause 5.10 (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 5.6. Jos $X_n \sim \mathrm{Bin}(n,p)$, niin

(5.4.9)
$$\begin{aligned} f_{X_n}(x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \frac{n!}{(n-x)!\,n^x}\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \left[\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

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 (5.4.5) 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 5.10 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 5.11 *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 5.11 on tärkeä merkitys esimekiksi frekvenssiaineistojen analyysissa.

Esimerkki 5.8 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 5.11 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. □

5.5 Poissonin prosessi

5.5.1 Laskuriprosessi

Stokastinen prosessi $\{ N(t),\ t \geq 0 \}$ on *laskuriprosessi,* jos $N(t)$ on ajankohtaan t mennessä sattuneiden "tapahtumien" lukumäärä.

Esimerkki 5.9 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 $s > 0$.

5.5.2 Poissonin prosessin määritely

Poissonin prosessi on yksi tärkeimpiä laskuriprosesseja. Se määritellään seuraavasti:

Määritelmä 5.2 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 5.2 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 5.2 ja 5.3 ovat yhtäpitävät.

Määritelmä 5.3 Laskuriprosessi $\{ N(t),\ t \geq 0 \}$ on Poissonin prosessi, jonka intensiteetti on λ ($\lambda > 0$), jos

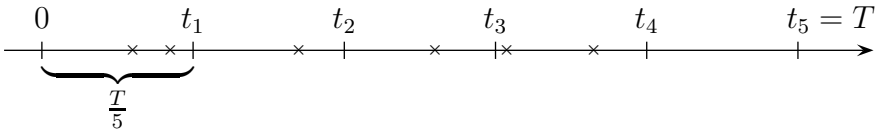
1. $N(0) = 0$.
2. Prosessin lisäykset ovat stationaariset ja riippumattomat.
3. $P\big(N(t+h) - N(t) = 1\big) = \lambda h + o(h)$.
4. $P\big(N(t+h) - N(t) \geq 2\big) = o(h)$.

Määritelmässä 5.3 käytetään merkintää $o(h)$. Sanomme, että funktio $f(\cdot) = o(h)$, jos

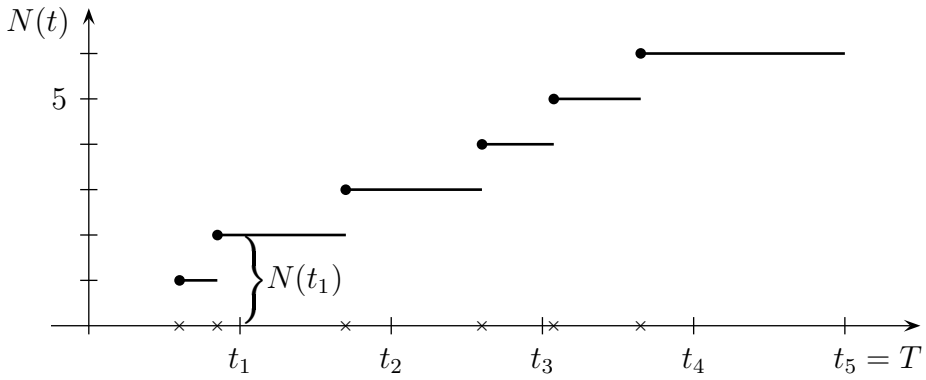
$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

Esimerkki 5.10 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.

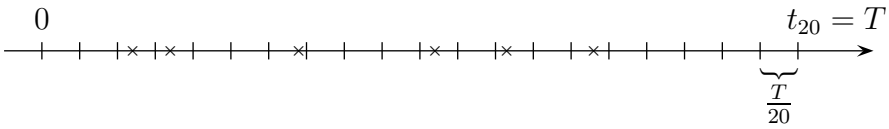


Tarkasteluväli $(0, T]$ on jaettu viiteen yhtä pitkään osaväliin, joiden pituudet ovat $T/5$. Nyt esimerkiksi 1. osavälillä sattuneiden onnettomuuksien lukumäärä on $N(t_1) - N(0) = N(t_1)$, joka on siis hetkeen t mennessä sattuneiden onnettomuuksien lukumäärä. Kuvioon 5.3 on piirretty prosessin $\{\, N(t),\ t \in (0, T] \,\}$ realisaatio, missä havaintoina ovat kyseiset onnettomuudet.



Kuvio 5.3. Poissonin prosessin $\{\, N(t),\ t \in (0, T] \,\}$ erään realisaation kuvaaja.

Määritelmän 5.3 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 5.3 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 5.3 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 5.3 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 5.10), 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 $\lambda>0$.

5.5.3 Satunnaistapahtumat tila-avaruudessa

Poissonin prosessilla mallinnetaan myös ilmiöitä, jotka tapahtuvat satunnaisesti tila-avaruudessa. Silloin Määritelmän 5.3 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 5.11 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

$$1 - \mathrm{e}^{-\lambda} - \mathrm{e}^{-\lambda}\lambda \geq 0.95.$$

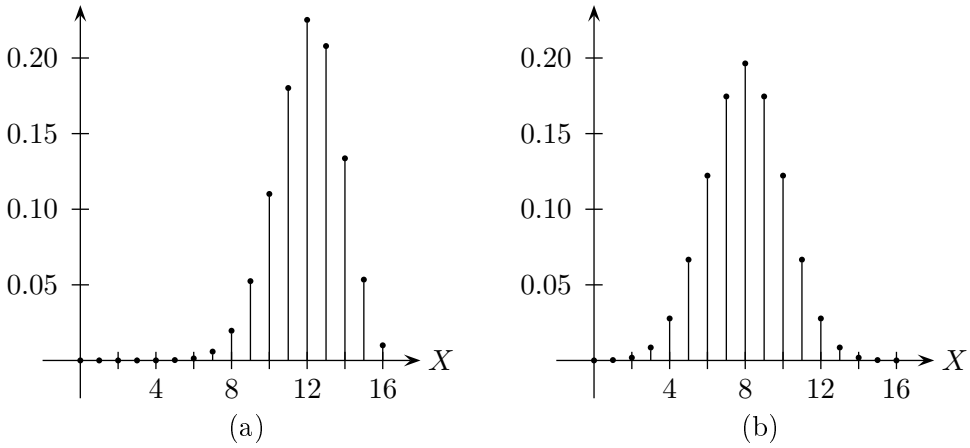
Epäyhtälö toteutuu, kun $\lambda \geq 4.74$, joten rusinoita on sekoitettava taikinaan 5 rusinaa pullaa kohti. □

5.5.4 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 5.4. 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 5.4). Binomijakaumassa $\text{Bin}(16, 0.50)$ on

$$P(X = 8 + x) = P(X = 8 - x)$$

kaikilla x. Silloin jokaista $a \in \mathbb{R}$ kohti

$$P(X \leq 8 - a) = P(X \geq 8 + a).$$

Diskreetit jakaumat: Yhteenveto

Bernoulli
$\text{Ber}(p)$

$$f(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1$$
$$E(X) = p, \qquad \text{Var}(X) = p(1-p)$$
$$M(t) = 1 - p + p\mathrm{e}^t$$

Binomi
$\text{Bin}(n, p)$

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$
$$E(X) = np, \qquad \text{Var}(X) = np(1-p)$$
$$M(t) = (1 - p + p\mathrm{e}^t)^n$$

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 $s, t \ge 0$.

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 $X \sim \mathrm{Ber}(p)$ 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 $\mathrm{Var}(X)$.

(c) Mitä on toistosten RRRR lukumäärän odotusarvo?

(Vihje: Katso Esimerkki 5.2.)

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 5.2). 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 $\{\,(x, y) \mid 0 \le x \le 1,\ 0 \le y \le 1\,\}$. Olkoon Z yksikköympyrään $\{\,(x, y) \mid x^2 + y^2 \le 1\,\}$ 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?