

# Matemaattinen tilastotiede

Erkki Liski

Matematiikan, Tilastotieteen ja Filosofian Laitos  
Tampereen Yliopisto



# Alkusanat

Tämä moniste perustuu vuosina 2002-2004 pitämiini matemaattisen tilastotieteen luentoihin Tampereen yliopistossa. Aineopintotasoinen matemaattisen tilastotieteen opintojakso voidaan suorittaa heti 2. opiskeluvuoden syksyllä tilastotieteen ja analyysin perusopintojen jälkeen. Moniste on tarkoitettu luentoihin liittyväksi oheismateriaaliksi ja on siis viimeistelemätön kehitysversio. Opetuksessa on kuitenkin hyödyllistä, että voidaan viitata tiettyyn perusmateriaaliin.

Lukuihin on liitetty jonkin verran harjoituksia, mutta kurssin aikana niitä tulee jatkuvasti lisää. Asioiden omaksuminen onnistuu parhaiten tekemällä harjoituksia ja ne antavat myös viitteen siitä, millaisia asioita edellytetään hallittavan kurssin väli- ja loppukokeissa.

Jarmo Niemelää kiitän erittäin asiantuntevasta avusta monisteen teknisessä toteuttamisessa  $\text{\LaTeX}$ illa.

Tampereella 9. syyskuuta 2005

Erkki Liski



# Sisältö

<b>Alkusanat</b>	<b>iv</b>
<b>1 Johdanto</b>	<b>1</b>
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 Joukko-operaatiot . . . . .	4
1.3.3 Todennäköisyys . . . . .	7
1.3.4 Äärettömät otosavaruudet . . . . .	8
1.3.5 Todennäköisyyden tulkinnat . . . . .	9
1.4 Ehdollinen todennäköisyys . . . . .	10
1.4.1 Ehdollisen todennäköisyyden frekvenssitulkinta . . . . .	11
1.4.2 Kertolaskusääntö . . . . .	12
1.4.3 Riippumattomuus . . . . .	12
1.5 Odotetut frekvenssit . . . . .	12
Yhteenvedo . . . . .	12
Harjoituksia . . . . .	14
<b>2 Todennäköisyyslaskenta ja kombinatoriikka</b>	<b>19</b>
2.1 Todennäköisyyden ominaisuuksia . . . . .	19
2.2 Symmetriaan perustuva todennäköisyys . . . . .	22
2.3 Aksiomaattinen lähestymistapa . . . . .	23
2.4 Kombinatoriikkaa . . . . .	23
2.4.1 Summa- ja tuloperiaate . . . . .	23
2.4.2 Valinta järjestyksessä . . . . .	23
2.4.3 Osajoukon valinta . . . . .	24
2.4.4 Otanta palauttaen, kun järjestystä ei oteta huomioon . . . . .	26
2.4.5 Kombinatoriikan merkintöjä ja identiteettejä . . . . .	27
2.4.6 Binomilause, hypergeometrinen identiteetti ja multinomilause . . . . .	28
2.4.7 Gammafunktio . . . . .	29
2.5 Satunnaismuuttuja . . . . .	30
2.5.1 Satunnaismuuttujan jakauma . . . . .	32
2.5.2 Kertymäfunktio . . . . .	33

2.5.3	Diskreetin satunnaismuuttujan todennäköisyysfunktio . . . . .	36
2.5.4	Diskreetti tasajakauma . . . . .	37
2.6	Otanta palauttamatta . . . . .	38
2.6.1	Hypergeometrinen jakauma . . . . .	39
2.6.2	Tarkistusotanta teollisuudessa . . . . .	40
2.7	Otanta palauttaen . . . . .	40
2.8	Binomijakauma . . . . .	42
2.8.1	Binomijakauma hypergeometrisen jakauman likiarvona . . . . .	43
2.9	Todennäköisyyden yleiset aksioomat . . . . .	44
	Yhteenvedo . . . . .	45
	Harjoituksia . . . . .	47
<b>3</b>	<b>Satunnaismuuttajat, ehdollistaminen ja riippumattomuus</b>	<b>51</b>
3.1	Ehdollinen todennäköisyys . . . . .	51
3.1.1	Todennäköisyyksien tulosääntö . . . . .	52
3.1.2	Riippumattomuus . . . . .	54
3.1.3	Joukko-oppi ja todennäköisyys . . . . .	58
3.2	Ehdolliset jakaumat . . . . .	58
3.3	Satunnaismuuttajien ominaisuuksia . . . . .	59
3.3.1	Diskreetin satunnaismuuttujan odotusarvo . . . . .	59
3.3.2	Ehdollisen jakauman odotusarvo . . . . .	65
3.3.3	Satunnaismuuttujan varianssi . . . . .	67
3.3.4	Kovarianssi ja korrelaatio . . . . .	69
3.3.5	Satunnaismuuttujan funktion jakauma . . . . .	70
3.3.6	Identtisesti jakautuneet satunnaismuuttajat . . . . .	71
3.3.7	Satunnaismuuttajien riippumattomuus . . . . .	72
3.3.8	Useiden satunnaismuuttajien riippumattomuus . . . . .	73
3.4	Suurten lukujen laki . . . . .	74
3.5	Generoivat funktiot ja momentit . . . . .	77
3.5.1	Momentit . . . . .	77
3.5.2	Momenttifunktio . . . . .	77
3.5.3	Todennäköisyydet generoiva funktio (tgf) . . . . .	80
3.6	Kokeiden yhdistäminen ja tulomallit . . . . .	81
3.6.1	Yleinen tulokaava . . . . .	83
3.7	Bayesin lause . . . . .	85
3.7.1	Peräkkäisotanta . . . . .	88
3.8	Usean tapahtuman unionin todennäköisyys . . . . .	90
	Yhteenvedo . . . . .	92
	Harjoituksia . . . . .	95
<b>4</b>	<b>Diskreetit jakaumat</b>	<b>99</b>
4.1	Diskreetti satunnaismuuttaja . . . . .	99
4.2	Bernoullin kokeet ja binomijakauma . . . . .	101
4.3	Odotusaikojen jakaumat . . . . .	106
4.3.1	Odotusajat Bernoullin kokeissa . . . . .	107

4.3.2	Geometrinen jakauma ja negatiivinen binomijakauma . . . . .	109
4.3.3	Odotusajat peräkkäisotannassa . . . . .	112
4.3.4	Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma . . . . .	114
4.3.5	Tasajakauma . . . . .	116
4.4	Poissonin jakauma . . . . .	116
4.5	Poissonin prosessi . . . . .	122
4.5.1	Laskuriprosessi . . . . .	122
4.5.2	Poissonin prosessin määrittely . . . . .	123
4.5.3	Satunnaistapahtumat tila-avaruudessa . . . . .	125
4.6	Kaksiulotteiset jakaumat . . . . .	126
4.6.1	Reunajakauma ja ehdollinen jakauma . . . . .	128
4.6.2	Satunnaismuuttujien funktion jakauma . . . . .	131
4.6.3	Ehdollinen odotusarvo . . . . .	131
4.6.4	Symmetrinen jakauma . . . . .	133
4.6.5	Kaksiulotteinen Bernoullin jakauma . . . . .	134
4.7	Satunnaismuuttujien funktion odotusarvo . . . . .	136
4.7.1	Momentit . . . . .	136
4.7.2	Satunnaisvektorin momenttifunktio . . . . .	137
4.8	Riippumattomat satunnaismuuttujat . . . . .	138
4.8.1	Riippumattomat kokeet . . . . .	138
4.8.2	Samoin jakautuneet riippumattomat (SJR) satunnaismuuttujat . . . . .	139
4.8.3	Riippumattomien satunnaismuuttujien funktio . . . . .	139
4.9	Multinomijakauma ja moniulotteinen hypergeometrinen jakauma . . . . .	140
	Yhteenvedo . . . . .	142
	Harjoituksia . . . . .	144
<b>5</b>	<b>Jatkuvat jakaumat</b>	<b>151</b>
5.1	Jatkuvat satunnaismuuttujat . . . . .	151
5.2	Tasajakauma ja eksponenttijakauma . . . . .	160
5.2.1	Tasajakauma . . . . .	160
5.2.2	Eksponenttijakauma . . . . .	161
5.2.3	Elinaikajakauma . . . . .	163
5.3	Gammajakauma ja $\chi^2$ -jakauma . . . . .	164
5.4	Normaalijakauma . . . . .	167
5.4.1	Standardimuotoinen normaalijakauma . . . . .	167
5.4.2	Yleinen normaalijakauma . . . . .	169
5.5	Muuttujien vaihto . . . . .	172
5.5.1	Muunnos kertymäfunktio avulla . . . . .	172
5.5.2	Muunnos tiheysfunktion avulla . . . . .	173
5.5.3	Normaalimuuttujan muunnokset . . . . .	176
5.6	Satunnaismuuttujan odotusarvo . . . . .	177
5.6.1	Momenttifunktio ja momentit . . . . .	179

5.7	Kaksiulotteiset jakaumat . . . . .	180
5.7.1	Reunajakauma ja ehdollinen jakauma . . . . .	181
5.7.2	Yhteisjakauman momenttifunktio . . . . .	186
5.8	Kahden muuttujan normaalijakauma . . . . .	188
5.8.1	Standardimuoto . . . . .	188
5.8.2	Korreloivat muuttujat . . . . .	188
5.9	Satunnaisvektoreiden muunnokset . . . . .	189
5.9.1	Yleinen kahden muuttujan normaalijakauma . . . . .	193
5.9.2	Studentin $t$ -jakauma, $F$ -jakauma ja beta-jakauma . . .	195
5.9.3	Hierarkkiset mallit ja yhdistetyt jakaumat . . . . .	199
	Yhteenvedo . . . . .	202
	Harjoituksia . . . . .	206
<b>6</b>	<b>Otantajakaumien teoria</b>	<b>213</b>
6.1	Riippumattomat satunnaismuuttujat . . . . .	213
6.2	Riippumattomien satunnaismuuttujien summan jakauma . . .	214
6.3	Normaalijakaumaan liittyvät jakaumat . . . . .	218
6.4	Järjestyssuureet . . . . .	221
6.4.1	Maksimi ja minimi . . . . .	221
6.4.2	Järjestyssuureen $X_{(k)}$ jakauma . . . . .	222
6.5	Keskeinen rajaväittäjä . . . . .	223
6.6	Jakaumien likiarvot normaalijakauman avulla . . . . .	225
6.7	$t$ -jakauma ja $F$ -jakauma . . . . .	226
6.8	Momenttifunktion rajafunktiot . . . . .	227
6.9	Suppenemiskäsitteet . . . . .	227
6.10	Estimaattorit . . . . .	230
6.10.1	Estimaattoreiden ominaisuuksia . . . . .	230
6.10.2	Delta-menetelmä . . . . .	231
<b>7</b>	<b>Uskottavuusfunktioon perustuva estimointi</b>	<b>233</b>
7.1	Tilastolliset mallit . . . . .	233
7.2	Estimoinnista . . . . .	234
7.3	Suurimman uskottavuuden menetelmä . . . . .	235
7.4	Pistefunktio ja informaatiofunktio . . . . .	237
7.4.1	Frekvenssitaulukkoon liittyvä uskottavuusfunktio . . .	238
7.5	Riippumattomien otosten yhdistäminen . . . . .	240
7.6	Normitettu uskottavuusfunktio . . . . .	242
7.6.1	Uskottavuusvälit ja uskottavuusalueet . . . . .	243
7.7	Uskottavuus jatkuvilla malleilla . . . . .	244
7.8	Invarianssi . . . . .	246
7.9	Normaalijakaumaan perustuva likiarvo . . . . .	247
7.10	Moniparametriset uskottavuusfunktiot . . . . .	248
7.10.1	Pistevektori ja informaatiomatriisi . . . . .	249
7.10.2	Normitettu uskottavuus ja korkeuskäyrät . . . . .	251
7.10.3	Profiliuskottavuus . . . . .	252



7.10.4	Normaalijakaumaan perustuva likiarvo . . . . .	253
7.11	Normaalinen regressiomalli . . . . .	254
7.11.1	Ehdollinen normaalimalli . . . . .	254
7.11.2	Kahden muuttujan normaalimalli . . . . .	254
7.11.3	Yksinkertainen lineaarinen regressio . . . . .	255
7.12	Probit- ja logitmallit: Esimerkki . . . . .	256
7.12.1	Probitmalli . . . . .	256
7.12.2	Logistinen malli . . . . .	256
7.12.3	Suurimman uskottavuuden estimointi . . . . .	257
7.13	Uskottavuussyhtälön ratkaiseminen numeerisesti . . . . .	258
	Harjoituksia . . . . .	259
<b>8</b>	<b>Frekvenssitulkinta uskottavuuspäätelyssä</b>	<b>265</b>
8.1	Otantajakaumat . . . . .	265
8.2	Peitetodennäköisyys . . . . .	268
8.2.1	Peitetodennäköisyyden arviointi uskottavuustestisuu- reen avulla . . . . .	270
8.3	$\chi^2$ -likiarvo . . . . .	271
8.3.1	Uskottavuustestisuureen jakauma . . . . .	271
8.3.2	Pistefunktion jakauma . . . . .	273
8.4	Luottamusvälit . . . . .	276
8.4.1	Normaalijakaumaan perustuva likiarvo . . . . .	278
8.4.2	Napasuureet . . . . .	281
8.5	Kahden parametrin mallit . . . . .	283
8.6	Odotettu informaatio ja kokeiden suunnittelu . . . . .	284
8.6.1	Johdanto . . . . .	284
8.6.2	Pistefunktion ja informaatiofunktion ominaisuuksia . . . . .	285
8.6.3	Cramérin ja Raon alaraja . . . . .	287
8.6.4	Suurimman uskottavuuden estimaattorin ominaisuuksia . . . . .	288
8.6.5	Uskottavuusfunktioon liittyviä testisuureita . . . . .	289
<b>9</b>	<b>Hypoteesien testaus</b>	<b>293</b>
9.1	Yleisiä näkökohtia . . . . .	293
9.1.1	Testisuureet ja $p$ -arvot . . . . .	294
9.2	Uskottavuussuhdetestit: Yksinkertaiset hypoteesit . . . . .	295
9.2.1	Yksi parametri . . . . .	295
9.2.2	Useita parametreja . . . . .	297
9.3	Uskottavuussuhdetestit: Yhdistetyt hypoteesit . . . . .	298
9.3.1	$p$ -arvon määrittäminen . . . . .	299
9.3.2	Kaksi parametria, joista toista testataan . . . . .	299
9.3.3	Homogeenisuuden testaus . . . . .	301
9.3.4	Binomitodennäköisyyksien testaaminen . . . . .	305
9.3.5	Multinomitodennäköisyyksien testaaminen . . . . .	307
9.3.6	Riippumattomuuden testaus kontingenssitaulukkoissa . . . . .	309
9.4	Testiteoriaa . . . . .	312

9.4.1	Ongelman määrittely . . . . .	312
9.4.2	Testin voimakkuus . . . . .	313
9.4.3	Testien konstruointi . . . . .	315

# Luku 1

## Johdanto

### 1.1 Todennäköisyys ja tilastotiede

Tämä kurssi käsittelee sekä todennäköisyyslaskentaa että tilastotiedettä. Uhapelurien ongelmat inspiroivat todennäköisyyslaskennan uranuurtajien ajattelua, mutta nykyisin todennäköisyyslaskennan sovellusalue on erittäin monipuolinen ja jatkuvasti laajeneva. Tilastotieteessä laaditaan satunnaisilmiölle 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 satunnaiskoikeista 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

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

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 aksiomeista. Itse asiassa (1.2.1) voidaan perustella *vahvan suurten lukujen lain* avulla. Se on tilastotieteen kannalta yksi todennäköisyyslaskennan tärkeimpiä lauseita.

Olkoon  $x_1, x_2, \dots, x_n$  jokin lukujono. Tavallisesti nämä luvut  $x_1, x_2, \dots, x_n$  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  $x_1, x_2, \dots, x_n$  muodostavat havaintoaineiston. Lukujen  $x_1, x_2, \dots, x_n$  (havaintoaineiston) *empiirinen kertymäfunktio* (ekf) reaali-lukuakselilla  $(-\infty, \infty)$  on

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

missä  $-\infty < a < \infty$  ja  $|\cdot|$  on joukon alkioden lukumäärä.

Lukujen  $x_1, x_2, \dots, x_n$  *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  $\{x_1, x_2, \dots, x_n\}$ :

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

**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.  $\square$

Empiirisen jakauman kuvaajana käytetään tavallisesti histogrammia. Histogrammin piirtäminen aloitetaan valitsemalla ensin *jakopisteet*  $b_1 < b_2 < \dots < b_m$  siten, että kaikki luvut  $x_i$  sisältyvät avoimelle välille  $(b_1, b_m)$  ja mikään jakopiste ei ole mittaluku. Jakopisteet määrittelevät  $m - 1$  osaväliä  $(b_j, b_{j+1})$ ,  $1 \leq j \leq m - 1$ . 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

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

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  $\Delta x$ . Jokainen mittaluku on silloin muotoa *kokonaisluku*  $\cdot \Delta x$ . Kahden

mittaluvun pienin mahdollinen erotus on  $\Delta x$ . Jakopisteet valitaan siten, että ne ovat muotoa

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

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.

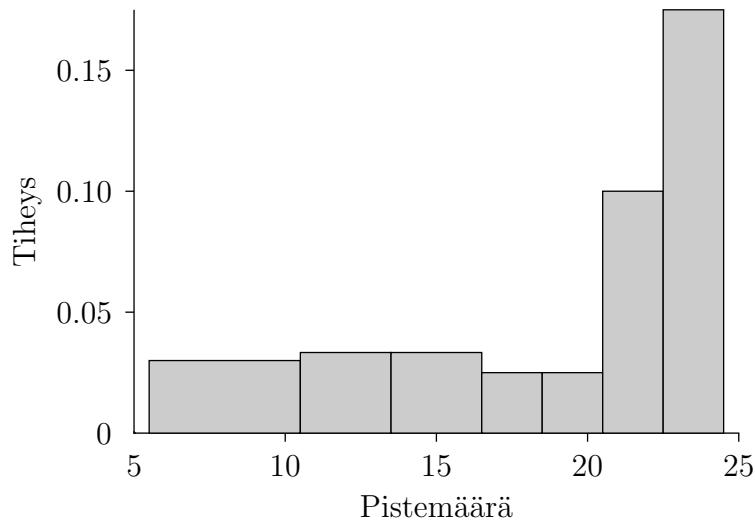
**Esimerkki 1.2** Kurssin 1. välikokeen pistemäärät  $x_i, 1 \leq i \leq 20$  olivat

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  $\Omega$ :lla. Satunnaiskokeen yksittäistä mahdollista tulosta kutsutaan *alkeistapaukseksi* (satunnaiskokeeseen liittyvän otosavaruuden  $\Omega$  yksi piste). Jos otosavaruus on äärellinen, merkitään

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

missä alkeistapaukset ovat  $\omega_1, \omega_2, \dots, \omega_n$  ja  $\Omega$ :n alkeistapausten lukumäärä  $|\Omega| = n$ . Otosavaruus voi olla myös ääretön.

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

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

Heitetään lanttia, kunnes saadaan ensimmäinen klaava. Silloin otosavaruus

$$\Omega = \{L, RL, RRL, RRRL, \dots\}$$

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

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


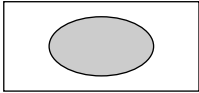

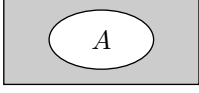
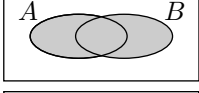
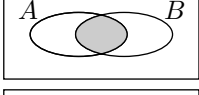
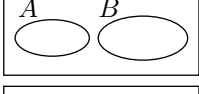

$$\Omega = \{\omega \in \mathbb{R} \mid \omega > 0\}.$$

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

### 1.3.2 Joukko-operaatiot

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

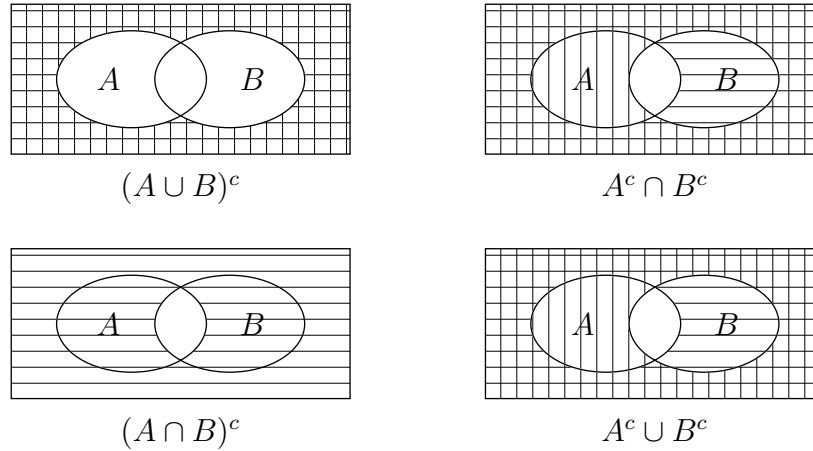
**Taulukko 1.1.** Joukko-opillisen ja todennäköisyyslaskennan terminologian vastaavuus.

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

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

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

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



**Kuvio 1.2.** De Morganin lait.

*Kaksinkertaisen komplementin sääntö*

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

on myös usein käyttökelpoinen.

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

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

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

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

ja

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

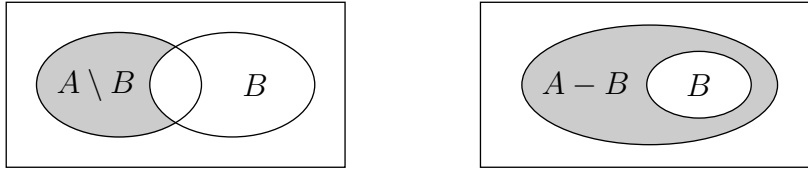
Sanomme, että tapahtumat  $A_1, A_2, \dots, A_m$  muodostavat tapahtuman  $A$  *osituksen* (tai jaon), jos  $A = A_1 \cup A_2 \cup \dots \cup A_m$  ja tapahtumat  $A_1, A_2, \dots, A_m$  ovat toisensa poissulkevat ( $A_i \cap A_j = \emptyset$ , kun  $i \neq j$ ). Esimerkiksi  $A, A^c$  muodostaa otosavaruuden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

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

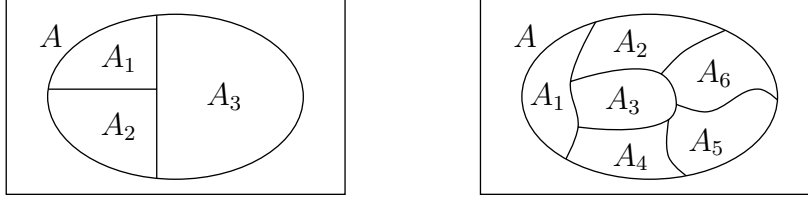
Jos  $A_1, A_2, A_3$  on  $A$ :n jako, niin

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





Kuvio 1.3. Joukkojen erotus.



Kuvio 1.4. Joukon A osituksia.

### 1.3.3 Todennäköisyys

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

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

$$P: \Omega \rightarrow [0, 1],$$

jolla on seuraavat ominaisuudet:

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

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

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

Näin funktio  $P$  voidaan laajentaa joukkofunktioksi, joka liittyy jokaiseen tapahtumaan  $A \subset \Omega$  luvun  $0 \leq P(A) \leq 1$ . Koska todennäköisyys on joukkofunktio, pitäisi alkeistapahtuman todennäköisyyttä oikeastaan merkitä  $P(\{\omega\})$ , mutta käytämme kuitenkin yleensä lyhyempää merkintää  $P(\omega)$ . Ominaisuuksiensa nojalla todennäköisyyttä kutsutaan yleisessä teoriassa todennäköisyysmitaksi. Jos  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , 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  $P(\emptyset)$  nolaksi. Satunnaiskokeen todennäköisyysmalli määritellään antamalla kokeen otosavaruus  $\Omega$  ja siihen liittyvä funktio  $P$ , joka toteuttaa Määritelmän 1.1 ehdot. *Todennäköisyysmalli* on siis pari  $(\Omega, P)$ .

Määritelmän mukaan  $P(\emptyset) = 0$ . Mahdoton tapahtuma  $\emptyset$  on varman tapahtuman  $\Omega$  komplementti eli  $\Omega^c = \emptyset$ . Tapahtuman  $A$  komplementti on joukko, johon kuuluvat kaikki ne alkeistapaukset, jotka eivät kuulu joukkoon  $A$ . Koska jokainen alkeistapaus  $\omega$  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*  $\Omega$ :ssa. Jos  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , niin voimme esittää todennäköisyysjakauman muodossa

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

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  $\Omega$ :ssa.

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

$$P(i) = \frac{1}{6}, \quad i = 1, \dots, 6.$$

Tapahtuman 'silmäluku pariton' todennäköisyys on

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

□

### 1.3.4 Äärettömät otosavaruudet

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

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

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

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

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

### 1.3.5 Todennäköisyyden tulkinnat

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

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

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

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

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

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

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

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

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

□

Ukkapelurit 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 voitosi 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 vedonlyönnissä 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 kokonaisipanoksesta tulee olla  $P(A)$ .

Eivät ainoastaan tapahtumien mahdollisuudet vaan myös mahdollisuuksien suhteet ovat keskeisiä pelitilanteiden analysoinnissa. Ne ovat tärkeitä käsitteitä myös esimerkiksi frekvenssiaineistojen analyysissä 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

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

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

## 1.4 Ehdollinen todennäköisyys

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

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

Kun tarkastellaan tapahtuman  $A$  ehdollista todennäköisyyttä  $P(A | B)$ , 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  $P(A | B)$  kaavalla

$$(1.4.1) \quad P(A | B) = \frac{P(A \cap B)}{P(B)}.$$

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

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

Ikä	20	45	65
Elossa	98040	95662	84483

Tässä voidaan ajatella, että alkuperäinen otosavaruus  $\Omega$  on 100000 tyttölästä. 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

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

□

### 1.4.1 Ehdollisen todennäköisyyden frekvenssitulkinta

Olkoot  $A$  ja  $B$  jotkut satunnaiskokeen  $\mathcal{E}$  otosavaruuteen  $\Omega$  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

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

kun toistojen lukumäärä  $n$  on suuri.

### 1.4.2 Kertolaskusääntö

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

$$(1.4.3) \quad P(A \cap B) = P(B) P(A | B)$$

tapahtuman  $A \cap B$  todennäköisyyden laskemiseksi.

### 1.4.3 Riippumattomuus

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

$$(1.4.4) \quad P(A \cap B) = P(A) P(B).$$

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

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

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

## 1.5 Odotetut frekvenssit

Kokeen  $\mathcal{E}$  todennäköisyysmalli  $(\Omega, P)$  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 suoritettussa 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: Yhteenvedo

- Empiirinen kertymäfunktio. Lukujen  $x_1, x_2, \dots, x_n$  *empiirinen kertymäfunktio* on

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

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

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

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

- Otosavaruus  $\Omega$  on satunnaiskokeen (tai satunnaisilmiön) mahdollisten tulosten (alkeistapausten  $\omega$ ) joukko. Satunnaiskokeessa voi sattua yksi ja vain yksi alkeistapaus.
- Tapahtuma on otosavaruuden  $\Omega$  osajoukko.

$A$ ja $B$ tapahtumia	$A \subset \Omega$ ja $B \subset \Omega$
$\Omega$	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	$A = A_1 \cup A_2 \cup \dots \cup A_m$ ja $A_i \cap A_j = \emptyset, i \neq j$

- De Morganin lait

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

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

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

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

- Tapahtuman  $A$  mahdollisuus

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

- Vedonlyöntisuhde

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

- $A$ :n todennäköisyys ehdolla  $B$

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

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

## Harjoituksia

1. Liitteessä 1 (ja tiedostossa `mtt/datat/hsarjat200.dat`) on kolme 200:n heiton sarjaa, joista yksi on tuotettu heittämällä harhatonta lanttia (200 riippumatonta toistokoetta, jossa kruunun  $tn = 1/2$ ). Muut sarjat poikkeavat selvästi (?) ”oikeasta” rahanheittokokeen tuloksesta. Koeta päätellä tai arvata, mikä on se aito rahanheiton tulos (Vrt. Mustonen: SURVO MM, Opetusohjelmat/Todennäköisyyksien laskentaa). Laske jokaisesta sarjasta kruunujen lkm. Onko Liitteen 1 tulosten perusteella uskottavaa, että sarjat on saatu harhattomalla rahalla (kruunu = 1 ja klaava = 0).
2. Aineistossa `kaivos_onn.dat` on aikajärjestyksessä pahojen (yli 10 kuollutta) peräkkäisten kaivosonnettomuuksien väliajat (päivinä) ajanjaksoilta 6. 12. 1875 – 29. 5. 1951. Piirrä väliaikojen frekvenssihistogramma koko aineistosta ja erilliset histogrammat 56:sta ensimmäisestä ja 53:sta viimeisestä havainnosta. Kommentoi eroja ja yhtäläisyyksiä.
3. Oletetaan, että histogrammassa kahden vierekkäisen suorakaiteen kannan leveydet ovat  $k_1$  ja  $k_2$  sekä korkeudet  $h_1$  ja  $h_2$ . Yhdistetään suorakaiteet yhdeksi suorakaiteeksi. Esitä uuden suorakaiteen korkeuden  $h$  lauseke ja osoita, että  $h$  on korkeuksien  $h_1$  ja  $h_2$  välissä.
4. Heitä harhatonta noppaa (R-ohjelma) 60, 120, 240, 480, 960 ja 2000 kertaa ja laske eri silmälukujen suhteelliset frekvenssit eri heittosarjoissa. Piirrä myös suhteellisten frekvenssien histogrammat. Miten heittojen lkm:n  $n$  kasvattaminen vaikuttaa suhteellisiin frekvensseihin?
5. Henkilöille X, Y, Z ja W on kullekin osoitettu kirje. Jokaiselle kirjeelle on varattu osoitteella varustettu kirjekuori. Kirjeet pannaan satunnaisesti kirjekuoriin.
  - (a) Mikä on tämän kokeen 24 alkeistapahtuman otosavaruus.
  - (b) Luettele seuraaviin tapahtumiin liittyvät alkaistapahtumat.
    - A: ”X:n kirje menee oikeaan kuoreen”;
    - B: ”Mikään kirje ei mene oikeaan kuoreen”;
    - C: ”Täsmälleen kaksi kirjettä menee oikeaan kuoreen”;
    - D: ”Täsmälleen kolme kirjettä menee oikeaan kuoreen”;



- (c) Laske edellisessä kohdassa mainittujen tapahtumien todennäköisyydet, jos oletetaan, että kaikki alkeistapaukset ovat yhtä todennäköisiä. Määritä tapahtumien  $A$ ,  $C$  ja  $D$  mahdollisuudet tapahtumaa  $B$  vastaan.
6. Kaksi joukkuetta pelaa paras seitsemästä sarjaa. Se joukkue voittaa, joka on ensiksi voittanut neljä peliä. Mikä on kokeen otosavaruus? Jos joukkueet ovat tasavahvoja (ja pelien tulokset toisistaan riippumattomia), niin mitkä ovat eri alkeistapahtumien todennäköisyydet? Mikä on todennäköisyys, että voittoon tarvitaan 7 peliä?
7. Tarkastellaan sellaista noppaa, että  $p_1 = p_2 = p_3 = p_4 = p$  ja  $p_5 = p_6 = q$ . Kirjoitetaan tn  $p$  muodossa  $p = \frac{1}{6} + \theta$ .
- (a) Lausu  $q$   $\theta$ :n avulla.
- (b) Heitetään noppaa  $n$  kertaa ja saadaan silmälukujen 1, 2, 3, 4, 5, 6 lukumääräksi  $n_1, n_2, n_3, n_4, n_5, n_6$ . Miten estimoisit  $\theta$ :n arvon?
- (c) Heitettiin noppaa 30, 120, 600 ja 1200. Silmälukujen frekvenssit olivat.

n	Silmäluvut					
	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  $\theta$ :n,  $p$ :n ja  $q$ :n estimaatit.

8. (a) Mikä on tn-malli, kun heitetään samanaikaisesti kolmea harhatonta lanttia.
- (b) Määritä tn saada  $x$  kruunua.
- (c) Heitettiin kolmea lanttia 80 kertaa ja saatiin seuraavat kruunujen lukumäärät.

```

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

```

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

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

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

- (b) Heitä kahta noppaa 100 kertaa. Vertaa tuloksen havaittuja frekvenssejä odotettuihin frekvensseihin.
10. Vuoden 2003 jääkiekon pudotuspelijoukkueet olivat HPK (1/3), Jokerit (1/2), Kärpät (1/3), Espoon BLUES (1/6), Tappara (1/3), JYP (1/7), HIFK (1/6) ja TPS (1/9). Eräällä työpaikalla järjestettiin ennen pudotuspelien alkua vuoden mestaria koskeva vedonlyönti käyttäen suluissa ilmoitettuja voiton mahdollisuuksia. Jos veikkasit esimerkiksi Tapparaa mestariksi, niin voitit panoksesi kolminkertaisena.
- (a) Laske annettujen voiton mahdollisuuksien (payoff odds) avulla joukkueiden voiton todennäköisyydet kaavalla (1.3.2). Laske todennäköisyyksien summa  $S$ .
- (b) Skaalaa edellisessä kohdassa lasketut ”todennäköisyydet” jakamalla ne summalla  $S$ . Miksi skaalaus on tarpeellinen?
- (c) Oleta, että skaalatut todennäköisyydet ovat ”oikeita”. Laske odotettu voittonsi, jos veikkasit Tapparaa [voitto  $\times P(A)$  + panoksesi  $\times (1 - P(A))$ ]. Toteuttaako veikkaus reilun pelin säännön?
11. Eräässä kyselyssä tutkittiin suhtautumista lailliseen aborttiin ja saatiin oheisessa taulukossa esitetyt tulokset.

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

Käytä todennäköisyyksien estimaatteina suhteellisia frekvenssejä.

- (a) Laske todennäköisyys, että (i) nainen (ii) mies suhtautuu aborttiin positiivisesti (tarkasteltavassa otosavaruudessa).
- (b) Laske mahdollisuudet (odds), että (i) nainen (ii) mies suhtautuu aborttiin positiivisesti.
- (c) Laske mahdollisuuksien suhde (odds ratio, vedonlyöntisuhde).
12. Esimerkissä 1.2 (luennot) on annettu erään kurssin 1. välikokeen piste-määrät.
- (a) Laske empiirisen kertymäfunktion (ekf) arvo pisteessä 15.3.
- (b) Lausu empiirisen jakauman arvo  $P_{20}(18.5, 20.5)$  ekf:n avulla.
- (c) Laske histogrammissa luokkaa  $[18.5, 20.5]$  kuvaavan pylvään korkeus.

**Liite 1**

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

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

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



# Luku 2

## Todennäköisyyslaskenta ja kombinatoriikka

Tässä luvussa käsitellään lähinnä vain äärellisiä ja numeroituvasti äärettömiä otosavaruuksia  $\Omega$ . Lopuksi esitetään todennäköisyyden aksioomat, jotka soveltuvat myös silloin, kun  $\Omega$  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  $\Omega$  on numeroituva otosavaruus ja  $P$  on  $\Omega$ :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.  $\square$

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

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

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

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

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

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

**Lause 2.3** *Olkkoon kokoelma  $A_1, A_2, \dots, A_n$  otosavaruuden  $\Omega$  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*

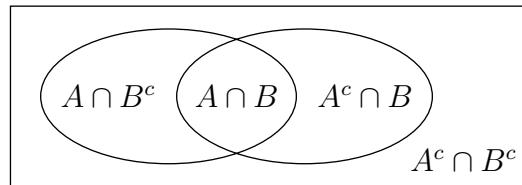
$$P(A) = P(A \cap B) + P(A \cap B^c).$$

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*

$$(2.1.1) \quad P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**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

$$(2.1.2) \quad P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B).$$

Lauseen 2.2 mukaan vastaavasti

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

joten

$$(2.1.3) \quad P(A) + P(B) = P(A \cap B^c) + P(A^c \cap B) + 2P(A \cap B).$$

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.  $\square$

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  $\Omega$ :n osajoukkoja (tapahtumia), niin*

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

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

$$(2.1.5) \quad P(A \cap B) \geq P(A) + P(B) - 1.$$

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.  $\square$

## 2.2 Symmetriaan perustuva todennäköisyys

Jos äärellisen otosavaruuden  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$  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

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

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  $p_i = \frac{1}{6}$ ,  $i = 1, \dots, 6$  otosavaruudessa  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . Jos heitetään kahta noppaa, voidaan symmetrisiksi alkeistapauksiksi valita järjestetyt parit

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

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  $P(\{(i, j)\}) = \frac{1}{36}$  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 Méré'n ongelma) ja niinpä hän kääntyi ranskalaisen filosofin ja matemaatikon Pascalin puoleen (n. 1650). De Méré'n ongelman uskotaan antaneen alkusysäyksen kuuluisaan Pascalin ja Fermatin väliseen kirjeenvaihtoon, joka johti todennäköisyyslaskennan syntyyn.  $\square$

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



## 2.3 Aksiomaattinen lähestymistapa

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

**Määritelmä 2.1** Todennäköisyys(mitta)  $P$  on otosavaruudessa  $\Omega$  määritelty  $\Omega$ :n osajoukkojen kuvaus eli funktio, joka toteuttaa seuraavat kolme aksiomaa:

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

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

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

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

- (iii) Koko otosavaruuden  $\Omega$  kuva on 1 eli

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

## 2.4 Kombinatoriikkaa

### 2.4.1 Summa- ja tuloperiaate

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

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

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

### 2.4.2 Valinta järjestyksessä

Tarkastellaan ensin *valintaa palauttaen*. Olkoon perusjoukon  $\Omega$  alkioden lukumäärä  $|\Omega| = n$  ja voimme siis ajatella, että alkiot on numeroitu 1:stä  $n$ :ään. Valitaan  $\Omega$ :sta peräkkäin  $r$  alkioita ja jokainen valittu alkio palautetaan takaisin  $\Omega$ :aan ennen seuraavaa valintaa. Valinnan tuloksena saatua järjestettyä jonoa kutsutaan *järjestetyksi  $r$ -otokseksi*  $(a_1, a_2, \dots, a_r)$ , 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  $\Omega$ :sta valittujen järjestettyjen  $r$ -otosten lukumäärä on  $n^r$ .

Tarkastellaan nyt *valintaa palauttamatta*. Jos  $\Omega$ :sta valitaan järjestyksessä  $r$  alkioita ( $r \leq n$ ) palauttamatta, saadaan järjestetty  $r$ -otos, jossa sama alkio voi esiintyä vain kerran. Tällaista järjestettyä  $r$ -otosta kutsutaan  $\Omega$ :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  $\Omega$ :sta on

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

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

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

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

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

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

Silloin todennäköisyys, että ainakin kahdella sama syntymäpäivä on

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

□

### 2.4.3 Osajoukon valinta

Kun  $\Omega$ :sta valitaan  $r$  alkioita ( $r \leq n$ ) palauttamatta, saadaan  $\Omega$ :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  $r$ -permutaatioiden lukumäärä jokaista  $r$ -kombinaatiota kohti on  $r!$  ja  $r$ -permutaatioiden 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 jälkeen  $c$   $\binom{4}{3}$  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  $n$  alkioita, joista  $n_1$  kuuluu 1. ryhmään,  $n_2$  2. ryhmään ja lopulta  $n_k$  alkioita  $k$ . ryhmään, joten  $n = n_1 + n_2 + \dots + n_k$ . Joukosta valitaan peräkkäin palauttamatta alkioita kunnes kaikki on valittu. Kuinka monta tunnistettavasti erilaista alkiojonoa voidaan saada? Nyt ajatellaan, että kunkin ryhmän alkiot ovat keskenään samanlaisia. Emme voi siis tunnistaa erilaisia ryhmän sisäisiä järjestyksiä. Vastaus saadaan samalla tavalla kuin edellisessä esimerkissä, joten valintojen lukumäärä on

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

Tätä lauseketta sanotaan *multinomikerroimeksi* ja sitä merkitään

$$(2.4.1) \quad \frac{n!}{n_1! n_2! \dots n_k!} = \binom{n}{n_1 \ n_2 \ \dots \ n_k}.$$

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

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

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

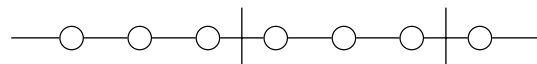
Jokaisen valinnan jälkeen pistetään merkki ”\*” kyseisen pallon kohdalle. Kaikkien valintojen jälkeen meillä on 3 ( $r$ ) merkkiä. Huomaa, että  $r$  voi olla suurempi kuin  $k$ , vaikka esimerkissä  $r = k = 3$ . Taulukon 2.1 viimeisellä sarakkeella pallojen valinta on esitetty merkkien ”\*” ja ”|”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

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

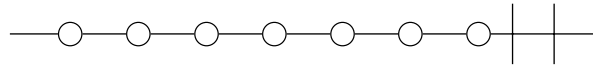
**Esimerkki 2.4** Tarkastellaan yhtälöä

$$(2.4.2) \quad x_1 + x_2 + \cdots + x_k = r,$$

missä  $k$  ja  $r$  ovat annettuja positiivisia kokonaislukuja ja muuttujat  $x_1, x_2, \dots, x_k$  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  $\binom{k+r-1}{k-1}$ .  $\square$

### 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:

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

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

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

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

**Esimerkki 2.5** Kertoman ja binomikertoimen laskuesimerkkejä:

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

$\square$

Määritelmien perusteella on suoraviivaista todeta, että

$$(2.4.5) \quad \binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r},$$

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

Jos  $s$  on epänegatiivinen kokonaisluku, niin silloin

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

*Stirlingin kaava*

$$(2.4.7) \quad n! \approx \sqrt{2\pi n} \cdot n^n e^{-n}.$$

antaa kertomalle hyvän likiarvon.

## 2.4.6 Binomilause, hypergeometrinen identiteetti ja multinomilause

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

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

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

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

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

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

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

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

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

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

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

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

## 2.4.7 Gammafunktio

Gammafunktio  $\Gamma$  määritellään integraalina

$$(2.4.12) \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

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

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

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

josta seuraa relaatio

$$(2.4.13) \quad \Gamma(\alpha + 1) = \alpha \Gamma(\alpha).$$

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

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

Koska

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

niin positiivisilla kokonaisluvuilla

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

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

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

kun  $n$  on positiivinen kokonaisluku.

## 2.5 Satunnaismuuttuja

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

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

Satunnaismuuttujia merkitään isoilla kirjaimilla  $X, Y, Z, \dots$ . Voimme kirjoittaa

$$X: \Omega \rightarrow \mathbb{R},$$

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

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

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

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

**Esimerkki 2.7** Mieli-pidekyselyssä 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  $\{0, 1, \dots, 100\}$ , jonka alkoiden lukumäärä on 101. Alkeistapaus  $\omega$  on 100:n pituinen tyyppiä "KEVV-VE...EV" oleva jono. Satunnaismuuttujan  $X$  arvo  $X(\omega)$  on alkeistapauksesta  $\omega$  laskettu kannattajien lukumäärä, esimerkiksi 36. □

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



**Taulukko 2.2.** Joitakin satunnaismuuttujia ja niiden arvoalueet.

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

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

**Huomautus 2.1** Jos  $X$  ja  $Y$  ovat satunnaismuuttujia, niin

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

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

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

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

*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  $\omega$ . Olkoon  $Y$  kruunujen lukumäärä 3:ssa lantin heitossa (Taulukko 2.2). Silloin esimerkiksi

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

ovat satunnaismuuttujia.

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

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

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

**Määritelmä 2.3 (Indikaattorifunktio)** Olkoon  $A$  tapahtuma otosavaruudessa  $\Omega$ . 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

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

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

$$(2.5.1) \quad X = X_1 + X_2 + \cdots + X_n$$

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

### 2.5.1 Satunnaismuuttujan jakauma

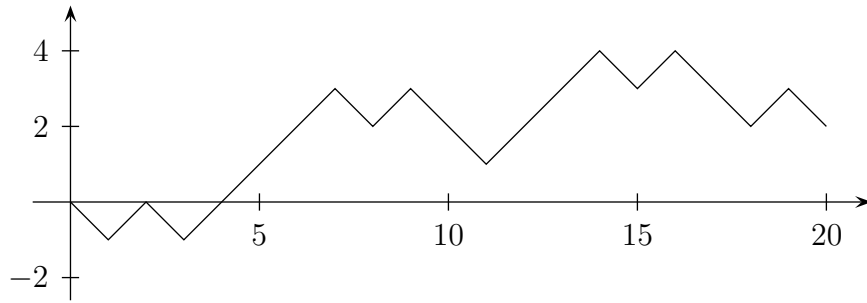
Olkoon  $Y$  kruunujen lukumäärä 3:ssa lantin heitossa (Esimerkki 2.6). Silloin satunnaismuuttujaa koskevat väittämät, kuten 'täsmälleen yksi kruunu'  $\equiv$  " $Y = 1$ " tai 'korkeintaan 2 kruunua'  $\equiv$  " $Y \leq 2$ " määrittelevät tapahtuman. Tapahtumat voidaan silloin kirjoittaa muodossa "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

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

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

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

LRLRRRRRLRLLRRRLRLLRL



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

Pekka voittaa 2 euroa.

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

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

missä

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

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

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

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.  $\square$

## 2.5.2 Kertymäfunktio

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

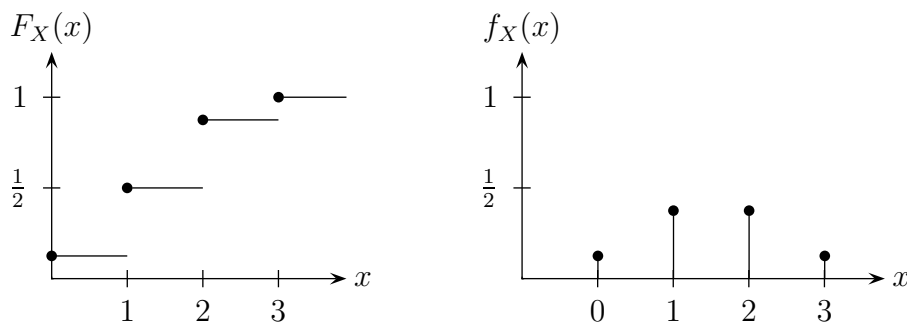
missä kaikki alkeistapaukset ovat yhtä todennäköisiä. Silloin

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

Satunnaismuuttujan  $X$  kertymäfunktio on siis

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

□



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

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

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

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

### 2.5.3 Diskreetin satunnaismuuttujan todennäköisyysfunktio

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

$$(2.5.4) \quad f_X(x) = P(X = x).$$

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

$$f_X(x): S_X \rightarrow [0, 1].$$

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

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

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

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

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

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

todennäköisyysfunktio. Nyt

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

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

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

Satunnaismuuttujan  $X$  todennäköisyysfunktio on siis

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

□

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

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

joka on porrasmuoto. Porrasfunktion  $F(x)$  'hyppäykset' ovat pisteissä  $x_1, x_2, \dots$  ja hyppäysten suuruudet ovat  $f_X(x_1), f_X(x_2), \dots$ . Jos esimerkiksi

$$f(x) = \frac{1}{2^x}, \quad x = 1, 2, \dots,$$

niin esimerkiksi

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

## 2.5.4 Diskreetti tasajakauma

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

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

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

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

## 2.6 Otanta palauttamatta

Tarkastellaan nyt koetta, jossa valitaan  $n$  alkiota  $N$ :n alkion joukosta ( $n \leq N$ ), jota kutsutaan *populaatioksi*. Valintaprosessia kutsutaan *otannaksi*. Halutaan esimerkiksi tietää ennen presidentin vaaleja, mikä on ehdokkaiden kannatus. Kannatuksesta voi saada tietoa tiedustelemalla äänestäjiltä, ketä he aikovat äänestää. On käytännössä mahdotonta haastatella kaikkia potentiaalisia äänestäjiä. Siksi tehdään otos, eli valitaan vain osa mahdollisista äänestäjistä ja haastatellaan heidät. Populaation muodostavat siis äänioikeutetut 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

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

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

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



nämä  $n$  kerttia ovat tietyssä järjestyksessä (esimerkiksi pakan päällä), on  $(N - n)!$  kappaletta. Todennäköisyys saada  $n$  kerttia 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 ( $\spadesuit 1, \heartsuit 2, \clubsuit 3, \diamondsuit 4, \spadesuit 5$ )? 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 ( $\spadesuit 1, \heartsuit 2, \clubsuit 3, \diamondsuit 4, \spadesuit 5$ ) todennäköisyys on  $1/311875200$ .  $\square$

### 2.6.1 Hypergeometrinen jakauma

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

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

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

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

## 2.6.2 Tarkistusotanta teollisuudessa

Mikään teollisuusprosessi ei ole täydellinen, siksi myös virheellisiä tuotteita on odotettavissa. Yrityksillä on käytössä erilaisia laadunvarmistusjärjestelmiä, jotta voitaisiin pitää yllä riittävän hyvää laatua. Virheelliset tuotteet olisi havaittava ja poistettava, jotta ne eivät joutuisi asiakkaalle saakka. Tietysti voitaisiin tarkistaa jokainen tuote riittävän tarkasti. Täydellinen tarkistus ei ole käytännössä yleensä 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, kasvatamalla 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.

## 2.7 Otanta palauttaen

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

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

Oletetaan esimerkiksi, että valitaan kolme numeroa palauttaen numeroista 0, 1, 2, ..., 9. Silloin voidaan saada  $10^3 = 1000$  yhtä mahdollista järjestettyä jonoa 000, 001, 002, ..., 999. Osajoukko  $\{1, 2, 3\}$  voidaan valita  $3! = 6$  tavalla, joten otoksen  $\{1, 2, 3\}$  todennäköisyys on 0.006. Otos  $\{1, 1, 3\}$  voidaan saada 3:lla tavalla, koska järjestetyt jonot (1, 1, 3), (1, 3, 1) ja (3, 1, 1) sisältävät samat alkioita. 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,  $i = 1, 2, \dots, N$ . Koska valinnan (kokeen) tulos on varmasti yksi ja vain yksi tapahtumista  $A_1, A_2, \dots, A_N$ , niin  $\Omega = A_1 \cup A_2 \cup \dots \cup A_N$  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 ( $i = 1, 2, \dots, N$ ). Silloin  $\sum_{i=1}^n n_i = n$  ja erilaisten järjestettyjen otosten lukumäärä on tuloksen (2.4.1) mukaan

$$\binom{n}{n_1 \ n_2 \ \dots \ n_N} = \frac{n!}{n_1! n_2! \dots n_N!}.$$

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

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

**Esimerkki 2.12** Valitaan populaatiosta  $\{A_1, A_2, A_3\}$  ( $N = 3$ ) 5 kertaa ( $n = 5$ ) alkio palauttaen. Silloin  $A_1 A_1 A_3 A_1 A_3$  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  $A_1 A_1 A_3 A_1 A_3$  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 alkioita ( $n \leq N$ ) missä tahansa järjestyksessä? Kyseinen tapahtuma sattuu täsmälleen silloin, kun  $X_1 = X_2 = \dots = X_n = 1$  ja  $X_{n+1} = \dots = X_N = 0$ . Tämän tapahtuman todennäköisyys on siis

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

□

### Otoksen kaikki alkiot erilaisia

Sellaisia järjestettyjä otoksia, joissa mikään alkio ei toistu, on

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

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

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

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

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

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

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

## 2.8 Binomijakauma

Oletetaan, että populaatiossa on kahdenlaisia alkioita:  $a$  kappaletta tyyppiä  $A$  ja  $b$  kappaletta tyyppiä  $B$ . Valitaan populaatiosta  $n$  alkioita palauttaen. Mikä on todennäköisyys, että otokseen tulee  $x$  alkioita tyyppiä  $A$  ja  $n-x$  alkioita tyyppiä  $B$ ? Voimme käyttää vastavaa urnamallia kuin hypergeometrisen jakauman yhteydessä. Uurnassa on  $a$  valkoista ja  $b$  mustaa palloa. Valitaan urnasta satunnaisesti palauttaen  $n$  palloa. Mikä on todennäköisyys, että otokseen tulee  $x$  valkoista palloa? Koska otanta tehdään palauttaen, urnan 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 satunnaisuuttuja  $X$  tyyppiä  $A$  olevien alkioiden lukumäärä otoksessa. Silloin  $X$ :n todennäköisyysfunktio on

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

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

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

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

### 2.8.1 Binomijakauma hypergeometrisen jakauman likiarvona

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

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

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

missä  $p = a/(a + b)$ .

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

## 2.9 Todennäköisyyden yleiset aksioomat

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

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

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

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

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

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

1.  $0 \leq P(A) \leq 1$ .
2.  $P(\emptyset) = 0$  ja  $P(\Omega) = 1$ .
3. Jos  $A_1, A_2, A_3, \dots$  on parittain pistevieraitten  $\Omega$ :n osajoukkojen jono, niin

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

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

**Määritelmä 2.6** Kokoelma  $\mathcal{F}$  on otosavaruuden  $\Omega$  osajoukkojen muodostama  $\sigma$ -algebra, jos

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

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

## Todennäköisyyslaskenta ja kombinatoriikka: Yhteenveto

### Todennäköisyyden ominaisuuksia

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

### Symmetriaan perustuva todennäköisyys

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

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

### Kombinatoriikkaa

- Järjestettyjen  $n$ -otosten lukumäärä, kun perusjoukon koko on  $N$ :
  - 1) valinta paluttaen:  $N^n$ ,
  - 2) valinta palauttamatta:  $N^{(n)} = N(N-1)(N-2)\cdots(N-n+1), \quad 0 \leq n \leq N,$   
 $N^{(N)} = N!.$
- Otokset (palauttamatta) eli  $n$ -kombinaatiot

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

- Multinomikerroin

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

- 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  $X: \Omega \rightarrow \mathbb{R}$ .
- $X$ :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

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

ovat satunnaismuuttujia, missä  $a$  on reaalivakio.

- Diskreetin satunnaismuuttujan  $X$  jakauma

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

- $X$ :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

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

- Binomijakauman todennäköisyysfunktio

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



## Kolmogorovin aksioomat

Olkoon  $\mathcal{F}$  jokin  $\Omega$ :n osajoukkojen muodostama  $\sigma$ -algebra. Kuvaus  $P: \mathcal{F} \rightarrow \mathbb{R}$  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  $A_i \in \mathcal{F}$  ( $i = 1, 2, \dots$ ) 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 binomilauseetta.)

(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 binomilauseetta.)

3. (a) Valitaan satunnaisesti 2 lukua luvuista  $1, 2, \dots, 39$  palauttamatta. Millä todennäköisyydellä saadaan peräkkäiset luvut?
- (b) Valitaan 7 lukua luvuista  $1, 2, \dots, 39$  palauttamatta (Lotto). Millä todennäköisyydellä saadaan peräkkäiset luvut?
- (c) Valitaan 2 lukua luvuista  $1, 2, \dots, n$  palauttamatta. Millä todennäköisyydellä saadaan peräkkäiset luvut?
4. Kahdestatoista verinäytteestä 4 oli positiivisia ja 8 negatiivisia. Sekaanuksen takia näytteet unohtuivat merkittömättä, 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 satunnaisesti valittua potilasta sai tutkittavaa uutta hoitoa (hoitoryhmä) ja loput kuuluivat vertailuryhmään. Mikä on todennäköisyys, että hoitoryhmään tulee
- ainakin yksi kumpaakin sukupuolta?
  - 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
- täsmälleen 2 viallista?
  - 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 yhdessäkin laatikossa ei ole enempää kuin yksi pallo?
- (b) Eräessä 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
- pallot ovat samanvärisiä;
  - pallot ovat erivärisiä.

Laske vastaavat todennäköisyydet, kun otanta on palauttamatta.

9. Eräessä 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
- 51 henkilöä kannattaa  $A$ :ta?
  - yli puolet kannattaa  $A$ :ta?
  - 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-alaan yhtä suureen kaistaleeseen  $A$ ,  $B$  ja  $C$ . Lisäksi oletetaan, että kaistaleiden yksikköhinnat ovat toisiinsa suhteessa  $1 : 2 : 3$ . Jokaisen (mittaisen) osa-alueen  $M$  suhteellinen hinta verrattuna koko maa-alueen hintaan saadaan kaavalla

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

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  $\Omega$  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).
- Millä todennäköisyydellä RW saa negatiivisen pistemäärän?
  - Mikä on RW:n pistemäärän todennäköisyysjakauma?
  - 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.
- Millä todennäköisyydellä 25 omenan satunnaisotoksessa (otanta palauttamatta) on 2 pilaantunutta?
  - Millä todennäköisyydellä 25 omenan satunnaisotoksessa on korkeintaan 2 pilaantunutta?
  - 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 mukanasasi. 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 kokeilemälläsi 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  $r_1, r_2, \dots, r_n$ ,  $r_i \geq 0$ ,  $i = 1, 2, \dots, n$  ja

$$(2.9.1) \quad r_1 + r_2 + \dots + r_n = r.$$

- (a) Mikä on erilaisten varausjakaumien (Yhtälön (2.9.1) ratkaisujen lukumäärä)?
- (b) Mikä on todennäköisyys, että tietyllä prosessorilla on  $k$ ,  $0 \leq k \leq r$  työtä?
- (c) Oletetaan, että annetut  $n$  lukua  $r_1, r_2, \dots, r_n$  toteuttavat yhtälön (2.9.1) ja kaikki mahdolliset  $n^r$  töiden sijoittelua prosessoreille ovat yhtä mahdollisia. Mikä on todennäköisyys saada varausluvut  $r_1, r_2, \dots, r_n$ ?

# Luku 3

## Satunnaismuuttujat, ehdollistaminen ja riippumattomuus

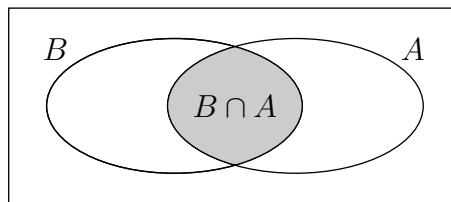
Tässä luvussa käsitellään satunnaismuuttujien ominaisuuksia ja täydennetään todennäköisyyslaskennan tietoja. Erityisesti satunnaismuuttujien odotusarvo on keskeinen käsite. Satunnaismuuttujien tarkastelussa rajoitutaan diskreettiin tapaukseen, mutta vastaavat tulokset pitävät paikkansa myös jatkuville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

### 3.1 Ehdollinen todennäköisyys

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

$$(3.1.1) \quad P(B | A) = \frac{P(B \cap A)}{P(A)}.$$

Lauseke  $P(B | A)$  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  $P(B | A)$  on siis alueen  $B \cap A$  pinta-alan suhteellinen osuus  $A$ :n pinta-alasta.

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

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

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

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

Voimme nyt helposti todeta, että

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

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

### 3.1.1 Todennäköisyyksien tulosääntö

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

$$(3.1.2) \quad P(A \cap B) = P(A) P(B | A),$$

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

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

joten saamme *kokonaistodennäköisyyden* kaavan

$$(3.1.3) \quad P(B) = P(A) P(B | A) + P(A^c) P(B | A^c).$$

Ehdollisen todennäköisyyden määritelmän mukaan

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

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*

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

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

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

$$P(A \cap B \cap C) = P(A) P(B | A) P(C | A \cap B).$$

Tulokaavan, kokonaistodennäköisyyden ja Bayesin kaavan yleistykset käsitellään luvun loppupuolella.

**Esimerkki 3.2** Suuri teollisuuskonserni valmistaa kännyköitä kolmessa eri maassa, jotka ovat nimeltään Fahru, Russo ja Swedla. Ostat kännykän, mut-

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

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

ta et tiedä, missä se on valmistettu. Olkoon  $V$  tapahtuma, että tuote on viallinen.  $F$  on tapahtuma, että tuote on valmistettu Fahrussa. Vastaavasti  $R$  ja  $S$  viittaavat valmistusmaihin Russo ja Swedla. Lasketaan todennäköisyydet (a)  $P(F | S^c)$ , (b)  $P(V | S^c)$ , (c)  $P(V)$ , (d)  $P(F | V)$ . Oletetaan, että kaikki valmistetut 6000000 kännykkää ovat yhtä todennäköisiä.

**Ratkaisu.**

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

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

Kohdat (c) ja (d) jätetään harjoitustehtäviksi.  $\square$

**Esimerkki 3.3 (Väärä positiivinen)** Oletetaan, että eräs verinäytteen 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 testitulosta on positiivinen, mikä on todennäköisyys, että hän sairastaa tautia  $A$ ?

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

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

Soveltamalla Bayesin kaavaa (3.7.4) saadaan

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

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

### 3.1.2 Riippumattomuus

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

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

Tämä kysymys johtaa riippumattomuuden määritelmään.

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

$$(3.1.4) \quad P(A \cap B) = P(A) P(B)$$

Jos tapahtumat  $A$  ja  $B$  ovat riippumattomat, niin silloin identiteetit

$$P(A | B) = P(A) \quad \text{ja} \quad P(B | A) = P(B)$$

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

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



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

ovat riippumattomat.

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

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

joten  $A$  ja  $B^c$  ovat riippumattomat. Muut kohdat todistetaan vastaavalla tavalla.  $\square$

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

1. Tapahtuma  $B$ : Kohdunkaulassa on epänormaaleja soluja, mutta ne *ei*vät 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  $P(D) = d$ .

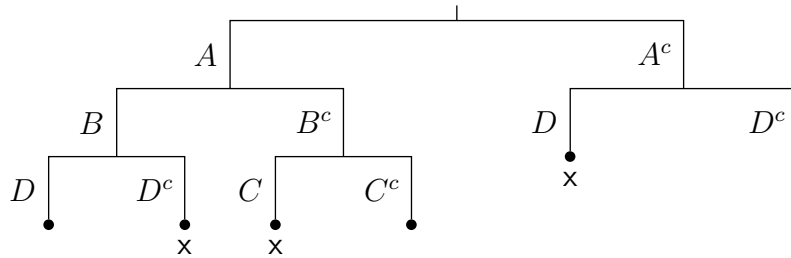
Oletetaan, että kaikki mainitut otanta- ja määrittämisvirheet ovat toisistaan riippumattomat. Jos satunnaisesti valitulle 30–65-vuotiaalle naiselle tehdään Papa-koe, niin

- (a) millä todennäköisyydellä koe antaa väärän tuloksen?
- (b) Jos testituloksella osoitetaan 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)

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

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

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

Virhetodennäköisyyksien 1 ja 3 mukaan

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

ja vastaavasti virheiden 1 ja 2 seurauksena

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

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

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

ja  $P(A \cap B \cap C) = P(A) P(B) P(C)$ .

**Esimerkki 3.5** Keskinäinen riippumattomuus ei seuraa parittaisesta riippumattomuudesta. Olkoon  $\Omega$  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.  $\square$

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

Tapahtumien keskinäinen riippumattomuus vaatii toteutuakseen varsin voimakkaita ehtoja.

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

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

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

### 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  $\Omega$  ositus, joten  $A \cup A^c = \Omega$  ja  $P(A \cup A^c) = P(\Omega) = 1$ . Tapahtuma ” $A$  tai  $A^c$ ” sattuu varmasti. Lauseen 2.1 (kohta 4) perusteella tiedämme tuloksen  $P(A \cup A^c) = P(A) + P(A^c)$ , josta seuraa erittäin käyttökelpoinen sääntö (Lause 2.1, kohta 5)

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

*De Morganin sääntö*

$$(3.1.5) \quad (A \cap B)^c = A^c \cup B^c$$

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  $\Omega$  määritelty satunnaisuuttaja 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 mukailen.

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

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

Ehdollisen todennäköisyyden määritelmän mukaan

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

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

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

Määritellään nyt funktio

$$(3.2.3) \quad f(x | A) = P(B_x | A) = P(X = x | A),$$

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

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

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

ja

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

Siksi  $X$ :n ehdollinen todennäköisyysfunktio ehdolla  $A$  on

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

□

## 3.3 Satunnaismuuttujien ominaisuuksia

### 3.3.1 Diskreetin satunnaismuuttujan odotusarvo

Numeroituvassa otosavaruudessa  $\Omega$  määritellyn satunnaismuuttujan  $X$  odotusarvo on

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

jos

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

Jos ehto (3.3.2) toteutuu, sarja (3.3.1) suppenee itseisesti. Tässä tapauksessa sanomme, että satunnaismuuttujalla  $X$  on odotusarvo. Muutoin satunnaismuuttujalla ei ole odotusarvoa. Jos  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$  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  $A_1, A_2, \dots$  sellainen otosavaruuden jako

$$\Omega = \bigcup_i A_i,$$

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

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

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

$$(3.3.3) \quad E(X) = \sum_i P(A_i)x_i = \sum_i p_i x_i.$$

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

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

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

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

$$(3.3.4) \quad E(X^r) = \sum_i p_i x_i^r.$$

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

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

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

jos summa suppenee itseisesti.

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

**Lause 3.2** *Oletetaan, että otosavaruudessa  $\Omega$  määritellyllä diskreeteillä satunnaisuuttujilla  $X$  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 satunnaisuuttujilla  $h(X)$ ,  $h_1(X)$  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  $E(X + Y) = E(X) + E(Y)$ :

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

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

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.  $\square$

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  $P(X_k = 1) = p$ ,  $k = 1, 2, \dots, n$ . Nyt satunnaismuuttuja

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

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

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

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



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

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

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

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

$Y$ :n arvojen 1 ja 0 todennäköisyydet ovat

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

Siksi

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

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

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

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

Määritellään nyt satunnaismuuttuja

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

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

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

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

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

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

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

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

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

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

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

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

Vastaavasti

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

□

**Esimerkki 3.11** Hypergeometrinen jakauma esiteltiin tarkasteltaessa otantaa palauttamatta (alaluku 2.6.1). Esimerkiksi tarkistusotannassa tuotteet luokitellaan viallisiksi tai hyväksyttäväiksi. Olkoon tuote-erässä  $N$  tuotetta, joista viallisia  $a$  ja hyväksyttäviä  $N - a$  kappaletta. Tehdään  $n$ :n alkion satunnaisosotus palauttamatta. Viallisten lukumäärä  $X$  otoksessa noudattaa hypergeometrista jakaamaa 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

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

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

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

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

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

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

Lasketaan nyt hypergeometrisen jakauman odotusarvo

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

Identiteetin (2.4.5) nojalla saadaan

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

ja

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

joten

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

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

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

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

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

□

### 3.3.2 Ehdollisen jakauman odotusarvo

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

$$(3.3.7) \quad E(X | A) = \sum_x x f(x | A).$$

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

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

□

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

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

$$E(X) = P(A)E(X | A) + P(A^c)E(X | A^c).$$

**Todistus.** Seurauslauseen 2.1 mukaan

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

ja ehdollisen todennäköisyyden määritelmän nojalla

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

ja

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

Tästä seuraa, että

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

Siksi

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

niinkuin väitettiin. □

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

$$E(X) = \sum_i P(A_i)E(X | A_i).$$

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  $\Omega$ :n ositus, niin

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

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

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

### 3.3.3 Satunnaismuuttujan varianssi

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

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

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

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

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

joten  $E(XY)$  on olemassa. □

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

$$(3.3.9) \quad [E(XY)]^2 \leq E(X^2) E(Y^2).$$

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

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

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

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

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

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

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

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

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

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

Merkintöjen  $\mu_X$  ja  $\sigma_X^2$  sijasta käytämme tavallisesti lyhyempiä versioita  $\mu$  ja  $\sigma^2$ , 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

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

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

$$(3.3.12) \quad \text{Var}(cX) = c^2 \text{Var}(X), \quad E(X^2) = \mu^2 + \text{Var}(X).$$

**Esimerkki 3.13** Lasketaan diskreettiä tasajakaumaa  $\text{Tasd}(1, N)$  noudattavan satunnaismuuttujan varianssi. Esimerkin 3.10 mukaan

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

Soveltamalla kaavaa (3.3.11) saadaan

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

□

### 3.3.4 Kovarianssi ja korrelaatio

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

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

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

Kovarianssin avulla voidaan sitten määritellä korrelaatiokerroin.

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

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

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

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

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

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

**Todistus.** Todistetaan 1. kohta. Määritelmän mukaan

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

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

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

Kaava (3.3.15) voidaan todistaa induktiolla. □

### 3.3.5 Satunnaismuuttujan funktion jakauma

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

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

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

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

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

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

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

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

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

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

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

$$h: S_X \rightarrow S_Y.$$

Määritellään *joukon*  $A$  *alkukuva* kuvauksessa  $h$  seuraavasti:

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

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

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

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



### 3.3.6 Identtisesti jakautuneet satunnaismuuttujat

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

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

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

Satunnaismuuttujan jakauma voidaan luonnehtia kertymäfunktion avulla.

**Lause 3.5** Seuraavat kaksi väitettä ovat yhtäpitävät:

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

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

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

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

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

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

ja

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

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  $p = \frac{1}{2}$ .  $\square$

### 3.3.7 Satunnaismuuttujien riippumattomuus

Määrittelimme tapahtumien riippumattomuuden alaluvussa 3.1.2. Tarkastelemme nyt satunnaismuuttujien riippumattomuutta.

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

$$(3.3.17) \quad P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

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

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

$$(3.3.18) \quad P(X = x, Y = y) = P(X = x) P(Y = y) = f_X(x) f_Y(y),$$

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

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

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

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

joten  $U$  ja  $V$  ovat riippumattomat. □

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

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

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

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

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

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

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

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

Jos  $I_A$  ja  $I_B$  ovat riippumattomat, niin

$$(3.3.19) \quad P(I_A = x, I_B = y) = P(I_A = x) P(I_B = y)$$

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

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

Lisäksi saadaan identiteetit

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

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

### 3.3.8 Useiden satunnaisuuttujen riippumattomuus

Satunnaisuuttujat  $X_1, \dots, X_n$  ovat riippumattomat, jos

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

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

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

joten identiteetin (3.3.20) mukaan

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

### 3.4 Suurten lukujen laki

#### Riippumattomat, samoin jakautuneet satunnaismuuttujat (rsj).

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

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

Jos siis yhden satunnaismuuttujan  $X_k$  odotusarvo on  $\mu$  ja varianssi  $\sigma^2$ , silloin niiden kaikkien kaikkien odotusarvo on  $\mu$  ja varianssi  $\sigma^2$ .

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

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

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

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

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

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

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

Markovin epäyhtälön avulla on helppo todistaa erittäin käyttökelpoinen Tšebyševin epäyhtälö.

**Lause 3.8 (Tšebyševin epäyhtälö)** *Olkoon  $X$  satunnaismuuttuja, jonka keskiarvo on  $\mu$  ja varianssi  $\sigma^2$ . Silloin*

$$(3.4.1) \quad P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \quad \text{kaikilla } \varepsilon > 0.$$

**Todistus.** Määritellään satunnaismuuttuja  $Y = h(X) = (X - \mu)^2$  ja valitaan  $a = \varepsilon^2 > 0$ . Koska  $Y \geq 0$  ja  $E(Y) = \sigma^2$ , seuraa Tšebyševin epäyhtälö (3.4.1) suoraan Markovin epäyhtälöstä.  $\square$

**Lause 3.9** *Oletetaan, että otosavaruudessa  $\Omega$  määritellyllä diskreeteillä satunnaismuuttujilla  $X$  ja  $Y$  on odotusarvo ja  $a \in \mathbb{R}$  on vakio. Silloin*

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

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

2.  $E[h(X)] = \sum_x h(x) f_X(x) = \sum_x h(x) P(X = x)$
3. Jos  $h_1(x) \geq h_2(x)$  kaikilla  $x$ , niin  $E[h_1(X)] \geq E[h_2(X)]$ .

**Lause 3.10 (Tulon odotusarvo, riippumattomat SM:t)** *Olkoot satunnaismuuttujat  $X$  ja  $Y$  riippumattomat.*

1. Jos  $E(X)$  ja  $E(Y)$  ovat olemassa, niin  $E(XY) = E(X)E(Y)$ .

*Olkoot satunnaismuuttujat  $X_1, X_2, \dots, X_n$  riippumattomat.*

2. Jos satunnaismuuttujilla  $X_1, X_2, \dots, X_n$  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

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy P(X = x, Y = y) \\ &= \sum_x \sum_y xy P(X = x) P(Y = y) \quad [X \text{ ja } Y \text{ riippumattomat}] \\ &= \left[ \sum_x x P(X = x) \right] \left[ \sum_y y P(Y = y) \right] \\ &= E(X) E(Y). \end{aligned}$$

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.  $\square$

**Apulause 3.3 (Summan varianssi, riippumattomat SM:t)** *Oletetaan, että  $X_1, X_2, \dots, X_n$  ovat riippumattomat ja niillä on varianssi. Silloin*

$$\text{Cov}(X_i, X_j) = 0, \quad i \neq j,$$

ja

$$\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n).$$

**Todistus.** Jos  $i \neq j$ , niin

$$\begin{aligned} \text{Cov}(X_i, X_j) &= E(X_i X_j) - E(X_i) E(X_j) \\ &= E(X_i) E(X_j) - E(X_i) E(X_j) = 0, \end{aligned}$$

koska  $X_i$ :n ja  $X_j$ :n riippumattomuuden nojalla  $E(X_i X_j) = E(X_i) E(X_j) = 0$ . Summan varianssin  $\text{Var}(\sum_{i=1}^n X_i)$  lauseke seuraa nyt suoraan Apulauseesta 3.2.  $\square$

**Apulause 3.4 (Otoskeskiarvon odotusarvo ja varianssi)** *Olko  $X_1, X_2, \dots, X_n$  RSJ satunnaismuuttujat, joiden keskiarvo on  $\mu$  ja varianssi  $\sigma^2$ . Määritellään satunnaismuuttujat*

$$S_n = X_1 + X_2 + \dots + X_n, \quad \bar{X}_n = \frac{S_n}{n}.$$

Silloin

$$E(S_n) = n\mu, \quad \text{Var}(S_n) = n\sigma^2, \quad E(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Voimme nyt todistaa Tšebyševin epäyhtälön avulla ns. *heikon suurten lukujen lain* (HSSL).

**Lause 3.11 (Heikko suurten lukujen laki (HSSL))** *Olko  $X_1, X_2, \dots, X_n$  ääretön RSJ satunnaismuuttujien jono, jossa jokaisen satunnaismuuttujan keskiarvo on  $\mu$  ja varianssi  $\sigma^2$ . Olko  $S_n = X_1 + X_2 + \dots + X_n$  ja*

$$\bar{X}_n = \frac{S_n}{n}.$$

Silloin jokaisella  $\varepsilon > 0$ ,

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0, \quad \text{kun } n \rightarrow \infty.$$

**Todistus.** Apulauseen 3.4 ja Tšebyševin epäyhtälön mukaan

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Kun  $n \rightarrow \infty$ , niin  $\sigma^2/(n\varepsilon^2) \rightarrow 0$ , joten

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0.$$

Näin on lause todistettu.  $\square$

Heikko suurten lukujen laki sanoo, että otoskeskiarvo lähenee todennäköisyyden mielessä todellista keskiarvoa, kun otoskoko kasvaa.

## 3.5 Generoivat funktiot ja momentit

### 3.5.1 Momentit

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

**Määritelmä 3.11** Olkoon  $r$  positiivinen kokonaisluku. Jos odotusarvo

$$\alpha_r = E(X^r)$$

on olemassa, se on satunnaismuuttujan  $X$  (tai  $X$ :n jakauman)  $r$ . momentti. Vastaavasti  $X$ :n  $r$ . keskusmomentti on

$$\mu_r = E[(X - \mu)^r],$$

missä  $\mu = E(X) = \alpha_1$ .

Momenttia  $\alpha_r$  kutsutaan joskus myös *origomomentiksi*. Jakauman keskiarvo on siis 1. origomomentti ja varianssi 2. keskusmomentti. Satunnaismuuttujan  $X$  *tekijämomentit*  $g_r$ ,  $r = 1, 2, \dots$  määritellään seuraavasti:

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

Ensimmäiset kaksi tekijämomenttia ovat

$$g_1 = E(X) = \alpha_1 = \mu,$$

$$g_2 = E[X(X-1)] = E(X^2 - X) = E(X^2) - E(X) = \alpha_2 - \mu.$$

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

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

### 3.5.2 Momenttifunktio

Esittelemme nyt uuden todennäköisyysjakaumaan liittyvän funktion, *momentteja generoivan funktion*, 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 (mikäli se on olemassa).

**Määritelmä 3.12** Olkoon  $X$  diskreetti satunnaismuuttuja, jonka todennäköisyysfunktio on  $f(x)$  ja arvoavaruus  $S$ . Silloin reaaliarvoisen  $t$  funktion

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

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

$$E(e^{tX}) = \sum_{x \in S} e^{tx} f(x)$$

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

Määritelmän perusteella on selvää, että

$$M(0) = E(e^{0 \cdot X}) = \sum_{x \in S} f(x) = 1.$$

Olkoon  $S = \{x_1, x_2, \dots\}$ . Silloin

$$M_X(t) = e^{tx_1} f(x_1) + e^{tx_2} f(x_2) + \dots,$$

missä  $e^{tx_k}$ :n kertoimet

$$f(x_k) = P(X = x_k), \quad k = 1, 2, \dots$$

ovat todennäköisyyksiä. Olkoon  $f(x)$  satunnaismuuttujan  $X$  todennäköisyysfunktio,  $g(y)$  satunnaismuuttujan  $Y$  todennäköisyysfunktio ja  $S = \{a_1, a_2, \dots\}$   $X$ :n ja  $Y$ :n yhteinen arvoavaruus. Jos

$$M_X(t) = M_Y(t), \quad \text{kaikilla } t, -h < t < h,$$

niin matemaattisen analyysin teorian nojalla

$$f(a_k) = g(a_k), \quad k = 1, 2, \dots$$

Jos siis kahdella satunnaismuuttujalla on sama momenttifunktio, niin niillä täytyy olla sama jakauma. Olkoon  $F_X(u)$   $X$ :n ja  $F_Y(u)$   $Y$ :n kertymäfunktio. Esitetään nyt momenttifunktion yksikäsitteisyyttä koskeva tulos lauseen muodossa.

**Lause 3.12** *Olkoot satunnaismuuttujien  $X$  ja  $Y$  momenttifunktiot  $M_X(t)$  ja  $M_Y(t)$ . Jos  $M_X(t) = M_Y(t)$  kaikilla  $t$  jossain nollan ympäristössä, niin  $F_X(u) = F_Y(u)$  kaikilla  $u$ :n arvoilla eli  $X$ :llä ja  $Y$ :llä on sama jakauma.*

**Esimerkki 3.17** Jos  $X \sim \text{Ber}(p)$ , niin

$$M(t) = E(e^{tX}) = e^{t \cdot 1} p + e^{t \cdot 0} q = e^t p + q,$$

missä  $q = 1 - p$ . □

**Lause 3.13** *Olkoot  $X$  ja  $Y$  riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat  $M_X(t)$  ja  $M_Y(t)$ . Silloin satunnaismuuttujan  $Z = X + Y$  momenttifunktio on*

$$(3.5.1) \quad M_Z(t) = M_X(t)M_Y(t).$$

**Todistus.** Koska  $e^{tX}$  on pelkästään  $x$ :n ( $X$ :n arvojen) funktio ja  $e^{tY}$  pelkästään  $y$ :n funktio, niin Lauseen 3.6 mukaan  $e^{tX}$  ja  $e^{tY}$  ovat riippumattomat. Väite

$$E(e^{tZ}) = E[e^{t(X+Y)}] = E[e^{tX} e^{tY}] = E(e^{tX}) E(e^{tY})$$

seuraa sitten suoraan Lauseesta 3.10. □



Usean satunnaismuuttujan tapauksessa on voimassa vastaava tulos.

**Seuraus 3.1** *Olkoot  $X_1, X_2, \dots, X_n$  riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat  $M_{X_i}(t)$ ,  $i = 1, 2, \dots, n$ . Silloin summan*

$$S_n = X_1 + X_2 + \dots + X_n$$

*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

$$(3.5.2) \quad M(t) = \sum_{x \in S} e^{tx} f(x)$$

derivoidaan puolittain, voidaan oikea puoli derivoida termeittäin ja yhtäsuuruus säilyy. Derivoimalla lauseke (3.5.2) puolittain muuttujan  $t$  suhteen saadaan

$$M(t)' = \sum_{x \in S} x e^{tx} f(x),$$

$$M(t)'' = \sum_{x \in S} x^2 e^{tx} f(x)$$

ja jokaisella positiivisella kokonaisluvulla  $r$

$$M(t)^{(r)} = \sum_{x \in S} x^r e^{tx} f(x).$$

Sijoittamalla  $t = 0$  saadaan

$$M(0)' = \sum_{x \in S} x f(x) = E(X),$$

$$M(0)'' = \sum_{x \in S} x^2 f(x) = E(X^2)$$

ja yleisesti

$$M(0)^{(r)} = \sum_{x \in S} x^r f(x) = E(X^r).$$

Erityisesti

$$\mu = M(0)' \quad \text{ja} \quad \sigma^2 = M(0)'' - [M(0)']^2.$$

**Lause 3.14** *Olkoon  $M_X(t)$  satunnaismuuttujan  $X$  momenttifunktio ja  $Y = aX + b$ , missä  $a$  ja  $b$  ovat annettuja reaaliarvoisia vakioita. Silloin  $M_Y(t) = e^{bt} M_X(at)$ .*

**Lause 3.15 (Momenttifunktioiden suppeneminen)** Olkoon  $X_1, X_2, X_3, \dots$  satunnaismuuttujien jono, jossa jokaisella  $X_n$ :llä on momenttifunktio  $M_{X_n}(t)$ ,  $n = 1, 2, 3, \dots$ . Oletetaan lisäksi, että

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

kaikilla  $t$ :n arvoilla jossain nollan ympäristössä  $(-h, h)$ , kun  $n \rightarrow \infty$ . Jos  $M_X(t)$  on momenttifunktio, niin silloin on olemassa yksikäsitteinen kertymäfunktio  $F_X(x)$ , jonka momenttifunktio on  $M_X(t)$  ja

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

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

Satunnaismuuttujien momenttifunktioiden suppenemisestä seuraa siis satunnaismuuttujien kertymäfunktioiden suppeneminen.

### 3.5.3 Todennäköisyydet generoiva funktio (tgf)

Diskreetin satunnaismuuttujan  $X$  todennäköisyydet generoiva funktio (tgf)  $G(t)$  määritellään seuraavasti:

$$G(t) = E(t^X) = \sum_{i=1}^{\infty} f(x_i)t^{x_i}.$$

Nähdään helposti, että  $G(1) = \sum_{i=1}^{\infty} f(x_i) = 1$ . Sarja suppenee ainakin silloin, kun  $|t| < 1$ . Kun sarja derivoidaan termeittäin, saadaan

$$G'(t) = \sum_{i=1}^{\infty} x_i f(x_i)t^{x_i-1}.$$

Jos  $G(t)$  on olemassa jollain välillä  $(-h-1, h+1)$ ,  $h > 0$ , niin

$$G'(1) = E(X)$$

ja yleisesti

$$G^{(r)}(1) = E(X^{(r)}) = E[X(X-1)\cdots(X-r+1)]$$

kaikilla positiivisilla kokonaisluvuilla  $r$ . Todennäköisyydet generoiva funktio liittyy läheisesti momenttifunktioon, sillä

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

### 3.6 Kokeiden yhdistäminen ja tulomallit

Tarkastellaan nyt satunnaiskokeita  $\mathcal{E}_1$  ja  $\mathcal{E}_2$ , joiden otosavaruudet ovat vastavasti  $\Omega_1$  ja  $\Omega_2$ . Olkoot satunnaiskokeisiin liittyvät todennäköisyysjakaumat  $\{p_i\}$  ja  $\{q_j\}$   $i = 1, 2, \dots$ . Tarkastelemme seuraavassa vain numeroituvia otosavaruuksia. Yhdistetään kokeet siten, että tehdään kokeet  $\mathcal{E}_1$  ja  $\mathcal{E}_2$ . Merkitään yhdistettyä koetta  $\mathcal{E}_1 \times \mathcal{E}_2$ . Yhdistetyn kokeen tulos esitetään järjestettynä parina  $(\omega_i, \omega_j)$ , 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  $\Omega_1$  ja  $\Omega_2$  *kartesainen 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*

$$(3.6.1) \quad P(\omega_i, \omega_j) = p_i q_j$$

kaikilla  $\omega_i \in \Omega_1$  ja  $\omega_j \in \Omega_2$ , missä  $p_i = p(\omega_i)$  on  $\omega_i$ :n todennäköisyys  $\Omega_1$ :ssä ja  $q_j = p(\omega_j)$  on  $\omega_j$ :n todennäköisyys  $\Omega_2$ :ssä. Selvästikin  $P(\omega_i, \omega_j) \geq 0$  kaikilla  $(\omega_i, \omega_j) \in \Omega_1 \times \Omega_2$ . Koska  $\sum_{\omega_i \in \Omega_1} p_i = \sum_{\omega_j \in \Omega_2} q_j = 1$ , niin

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

Identiteetti (3.6.1) siis määrittelee todennäköisyysjakauman  $\Omega_1 \times \Omega_2$ :ssa. Sitä kutsutaan yhdistetyn kokeen  $\mathcal{E}_1 \times \mathcal{E}_2$  *tulomalliksi*.

#### Riippumattomat toistot

Tulomallin tärkeä erikoistapaus saadaan toistamalla  $n$  kertaa koe  $\mathcal{E}$ , jonka otosavaruus on  $\Omega$ . Tällaista koetta sanotaan *toistokokeeksi* ja sitä merkitään  $\mathcal{E}^n$ . Yhdistetyn kokeen otosavaruus on  $\Omega \times \Omega \times \dots \times \Omega$ , jonka alkeistapaukset ovat muotoa  $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n)$ , missä  $\omega_i$  on  $i$ . toiston tulos. Olkoon  $p(\omega)$  satunnaiskokeeseen  $\mathcal{E}$  liittyvässä otosavaruudessa  $\Omega$  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  $\Omega = \{O, E\}$ . Satunnaismuuttuja  $X$  noudattaa *Bernoullin jakaumaa*, kun

$$(3.6.2) \quad X = \begin{cases} 1, & \text{todennäköisyydellä } P(O) = p; \\ 0, & \text{todennäköisyydellä } 1 - p, \end{cases}$$

missä  $0 \leq p \leq 1$ . Myös satunnaismuuttujan arvoa  $X = 1$  kutsutaan onnistumiseksi ja  $p$ :tä onnistumistodennäköisyydeksi. Vastaavasti arvoa  $X = 0$  kutsutaan epäonnistumiseksi. Huomaa, että  $X$  on 'onnistumisen' indikaattorifunktio. Bernoullin kokeen riippumattomat toistot muodostavat *Bernoullin toistokokeen*.

**Esimerkki 3.18** Esimerkissä 2.6 heitetään harhatonta lanttia 3 kertaa. Yhdessä lantin heitossa otosavaruus  $\Omega = \{R, L\}$ . 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 riippumattonta Bernoullin koetta. Tähän yhdistettyyn kokeeseen liittyvä otosavaruus on  $S \times S \times S = \{(s_1, s_2, s_3) \mid s_i \in S\} = \{111, 110, 101, 100, 011, 010, 001, 000\}$ .  $\square$

Kun toistetaan Bernoullin koe  $n$  kertaa (riippumattomat toistot), ovat kokeen mahdolliset tulokset  $n$ :n pituisia  $1$ :n ja  $0$ :n muodostamia jonoja. Tyyppillinen jono on muotoa  $111011000 \dots 110$ , jonka todennäköisyys on

$$ppp(1-p)p(1-p)(1-p)ppp \dots pp(1-p) = p^k(1-p)^{n-k},$$

missä  $k$  on onnistumisten lukumäärä ja  $n - k$  epäonnistumisten lukumäärä. Erilaisten mahdollisten jonojen lukumäärä on  $2^n$ .

*Binomijakauma* voidaan määritellä Bernoullin toistokokeen avulla. Olkoon  $X_1, X_2, \dots, X_n$  samaa Bernoullin jakaumaa noudattavien riippumattomien satunnaismuuttujien jono, missä  $P(X_i = 1) = p$  ja  $P(X_i = 0) = 1 - p = q$ ,  $i = 1, 2, \dots, n$ . Silloin  $E(X_i) = p$  ja  $\text{Var}(X_i) = pq$ . Onnistumisten lukumäärä  $n$ :ssä riippumattomassa Bernoullin kokeessa on

$$X = X_1 + X_2 + \dots + X_n.$$

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

$$\bar{X}_n = \frac{S_n}{n}.$$

Se on onnistumisten suhteellinen frekvenssi  $n$ :ssä riippumattomassa Bernoullin kokeessa, esimerkiksi kruunujen suhteellinen frekvenssi lantin heitossa. Apulauseen 3.4 mukaan  $E(\bar{X}_n) = p$  ja  $\text{Var}(\bar{X}_n) = pq/n$ . HSSL:n mukaan

$$P(|\bar{X}_n - p| > \varepsilon) \rightarrow 0$$

kaikilla  $\varepsilon > 0$ , kun  $n$  kasvaa. Kruunujen suhteellinen frekvenssi lähenee  $p$ :tä todennäköisyyden mielessä, kun heittojen määrä kasvaa. Bernoulli todisti tämän tuloksen 1713. Tulosta kutsutaan hänen mukaansa *Bernoullin suurten lukujen laiksi*. Ensimmäisessä luvussa tarkasteltiin suhteellisen frekvenssin raja-arvoa todennäköisyyden tulkintana ja eräänlaisena perusteluna todennäköisyydelle. Nyt näemme, että tämä suhteellisen frekvenssin raja-arvotulos on yksi todennäköisyyslaskennan perustuloksista.

### 3.6.1 Yleinen tulokaava

Yleensä todennäköisyysongelmat koskevat useita tapahtumia tai satunnaismuuttujia, joiden keskinäisiä riippuvuuksia tarkastellaan. Tietyssä mielessä kaikki todennäköisyydet ovat ehdollisia, mutta tavallisesti selvänä pidetyt ehdot jätetään mainitsematta. Rahanheitossa mainitsemme vain vaihtoehdot 'kruunu' ja 'klaava', vaikka lantti voi jäädä myös reunalleen. Presidenttiehdokkaasta tulee presidentti vain sillä ehdolla, että säilyy hengissä vaalikampanjan ajan. Valitsemistodennäköisyyttä laskettaessa ei hengissäpysymisen todennäköisyyttä tavallisesti oteta huomioon.

Seuraavassa esitetään yleinen tulokaava. Huomaa, että jatkossa leikkausta  $A_1 \cap A_2$  merkitään kaavojen yksinkertaistamiseksi lyhyesti  $A_1 A_2$ .

**Väittäjä 3.1 (Tulokaava)** *Olkoot  $A_1, A_2, \dots, A_n$  mitä tahansa tapahtumia. Silloin*

$$(3.6.3) \quad P(A_1 A_2 \cdots A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 A_2) \cdots \\ \cdot P(A_n | A_1 A_2 \cdots A_{n-1}),$$

jos  $P(A_1 A_2 \cdots A_{n-1}) > 0$ .

**Todistus.** Jos  $P(A_1 A_2 \cdots A_{n-1}) > 0$ , niin kaavassa (3.6.3) esitetyt ehdolliset todennäköisyydet ovat hyvin määritellyt, koska

$$P(A_1) \geq P(A_1 A_2) \geq \cdots \geq P(A_1 A_2 \cdots A_{n-1}) > 0.$$

Kun yhtälön (3.6.3) oikea puoli kirjoitetaan auki ehdollisen todennäköisyyden kaavaa (3.1.1) soveltaen, saadaan

$$\frac{P(A_1)}{P(\Omega)} \cdot \frac{P(A_1 A_2)}{P(A_1)} \cdot \frac{P(A_1 A_2 A_3)}{P(A_1 A_2)} \cdots \frac{P(A_1 A_2 \cdots A_n)}{P(A_1 A_2 \cdots A_{n-1})},$$

joka supistuu todennäköisyydeksi  $P(A_1 A_2 \cdots A_n)$ . □

Kutsomme kaavaa (3.6.3) tapahtumien yhdisteen *yleiseksi tulokaavaksi*. Jos  $A_1, A_2, \dots, A_n$  ovat keskenään riippumattomat, niin saadaan

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

Oletetaan, että satunnaismuuttujien  $X_1, X_2, \dots, X_n, \dots$  arvoalueet  $S_i$  ovat numeroituvia. Määritellään tapahtumat

$$A_i = \{X_i = x_i\}, \quad i = 1, 2, \dots,$$

missä  $x_i \in S_i$ . Silloin voimme kirjoittaa kertolaskukaavan (3.6.3) avulla

$$(3.6.4) \quad \begin{aligned} &P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1) P(X_2 = x_2 \mid X_1 = x_1) P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2) \cdots \\ &\quad \cdot P(X_n = x_n \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}). \end{aligned}$$

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

**Esimerkki 3.19 (Syntymäpäiväongelma uudelleen)** Olemme jo aikaisemmin implisiittisesti soveltaneet yleistä tulokaavaa (3.6.3). Tarkastellaan uudelleen Esimerkin 2.3 syntymäpäiväongelmaa. Kutsuilla on  $r$  henkilöä. Millä todennäköisyydellä ainakin kahdella henkilöllä on sama syntymäpäivä? Käytössämme on osanottajalista, johon syntymäpäivät on merkitty (karkausvuotta ei oteta huomioon). Käydään listaa läpi alusta lähtien järjestyksessä niin pitkälle, kunnes löydetään syntymäpäivä, joka jo oli listalla aikaisemmin. Silloin etsintä lopetetaan siihen ja todetaan, että ainakin kahdella vieraalla on sama syntymäpäivä. Jos lista päästään läpi löytämättä toistoa, kellä 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

$$A_r^c = B_2 \cup B_3 \cup \dots \cup B_r$$

on tapahtuma, että ainakin kahdella on sama syntymäpäivä. Koska tapahtumat  $B_2, B_3, \dots, B_r$  ovat toisensa poissulkevat, niin

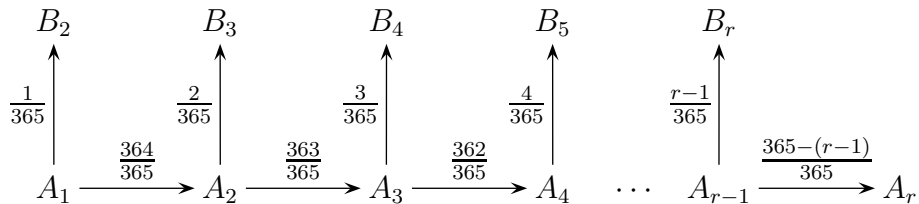
$$P(A_r^c) = P(B_2) + P(B_3) + \dots + P(B_r),$$

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 vierailloilla on eri syntymäpäivä. Sitä ennen ovat sattuneet  $A_2, A_3, \dots, A_{r-1}$ . Esimerkiksi  $A_2$  on tapahtuma, että tarkistusprosessi ei pysähdy 2. vieraaseen, vaan hänellä on eri syntymäpäivä kuin 1. vieraalla. Todennäköisyys

$$P(A_2) = \frac{364}{365} = 1 - \frac{1}{365} = 1 - P(B_2).$$

koska valittavana on 364 päivää, jotka poikkeavat 1. vieraan syntymäpäiväpäivästä. Jos  $j$ :n ensimmäisen syntymäpäivän joukossa ei ole samoja, niin ei myöskään  $i$ :n ensimmäisen, jos  $i < j$ , jolloin  $A_i \subset A_j$ . Tästä seuraa, että  $A_2 A_3 \cdots A_j = A_j$  ja

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

Soveltamalla tapahtumien yhdisteen tulokaavaa saadaan

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

□

## 3.7 Bayesin lause

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

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

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

**Lause 3.16** *Olkoon*

$$\Omega = \sum_i H_i$$

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

$$(3.7.1) \quad P(T) = \sum_i P(H_i) P(T | H_i).$$

**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 2.5) perusteella seuraa kaava

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

Kun kaavaan sijoitetaan

$$P(H_i T) = P(H_i) P(T | H_i),$$

saadaan (3.7.1). □

Jos kaavassa (3.7.1) jokin  $P(H_i) = 0$ , vastaava summan termi on 0, vaikka  $P(T | H_i)$  ei olekaan määritelty. Kaavaa (3.7.1) kutsutaan *kokonaistodennäköisyyden kaavaksi*.

Olkoot  $X$  ja  $Y$  kokonaislukuarvoiset satunnaismuuttujat ja  $k$  jokin kokonaisluku. Soveltamalla kaavaa (3.7.1) tapahtumiin

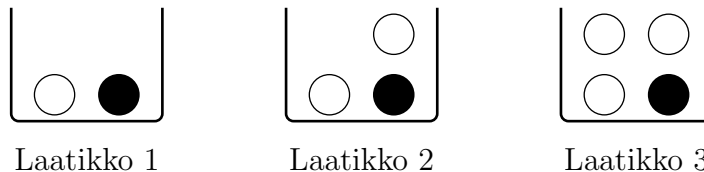
$$H_i = \{X = i\}, \quad T = \{Y = k\}$$

saadaan

$$(3.7.2) \quad P(Y = k) = \sum_i P(X = i) P(Y = k | X = i),$$

missä summa käy yli kaikkien kokonaislukujen. Jos  $P(X = i) = 0$ , niin vastaava yhteenlaskettava summassa on 0. Kaava on helppo yleistää mille tahansa satunnaismuuttujalle  $X$ , jonka arvojoukko  $S_X$  on numeroituva.  $Y$  voi olla jokin yleisempi satunnaismuuttuja, ei välttämättä kokonaislukuarvoinen, ja tapahtuma  $T = \{Y = k\}$  voidaan korvata vaikkapa tapahtumalla  $T = \{Y > a\}$ ,  $a \in \mathbb{R}$ .

**Esimerkki 3.20 (Mikä laatikko?)** Meillä on 3 samanlaista laatikkoa. Laatikossa  $i$  on  $i$  valkoista palloa ja yksi musta,  $i = 1, 2, 3$ . Tilanne on siis oheisen kuvion kaltainen



Laatikko 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  $\lceil \frac{3}{2} \rceil = 2$  ja valitaan Laatikko

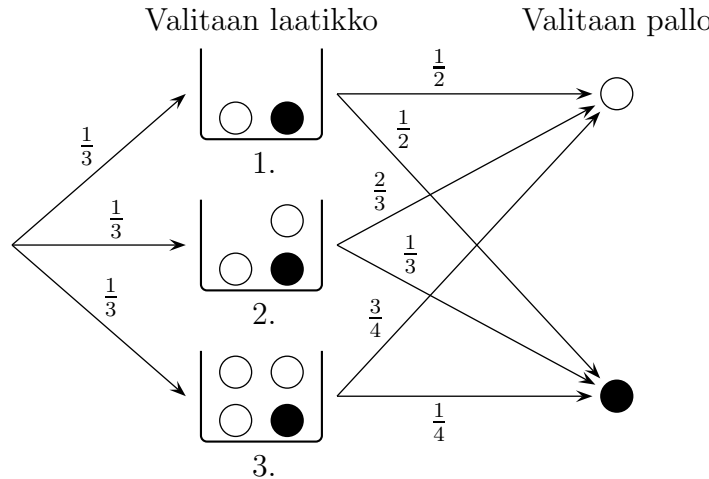


2. Kun valitun pallon väri on tiedossa, arvataan, mistä laatikosta pallo on valittu.

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

$$(3.7.3) \quad P(H_i | T) = \frac{P(H_i T)}{P(T)}, \quad i = 1, 2, 3.$$

Seuraavassa kuviossa on esitetty havainnollisesti tilanteeseen liittyvät todennäköisyydet.



Kaavassa (3.7.3) osoittaja on

$$P(H_i T) = P(H_i) P(T | H_i) = \frac{1}{3} \cdot \frac{i}{i+1}, \quad i = 1, 2, 3.$$

Koska  $\sum_{i=1}^3 H_i T = T$  ja  $T_1$ ,  $T_2$  ja  $T_3$  muodostavat  $T$ :n osituksen, niin yhteenlaskulauseen perusteella

$$P(T) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{3}{4} = \frac{23}{36}.$$

Kaavasta (3.7.3) saadaan

$$P(H_i | T) = \frac{\frac{1}{3} \cdot \frac{i}{i+1}}{\frac{23}{36}} = \frac{12}{23} \cdot \frac{i}{i+1}, \quad i = 1, 2, 3.$$

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.  $\square$

**Väittämä 3.2 (Bayesin lause)** *Olkoon  $H_1, H_2, \dots, H_k$  otosavaruuden  $\Omega$  ositus ja  $T$  sellainen tapahtuma, että  $P(T) > 0$ . Silloin*

$$(3.7.4) \quad P(H_i | T) = \frac{P(H_i) P(T | H_i)}{\sum_{j=1}^k P(H_j) P(T | H_j)}.$$

**Todistus.** Todennäköisyyksien tulokaavan nojalla saadaan

$$P(H_i T) = P(H_i) P(T | H_i) = P(T) P(H_i | T),$$

mistä seuraa

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

Väittämän 3.16 mukaan  $P(T) = \sum_j P(H_j) P(T | H_j)$ , joten kaava (3.7.4) on todistettu.  $\square$

Kaavaa (3.7.4) kutsutaan Bayesin säännöksi. Tapahtumat  $H_1, H_2, \dots, H_k$  voidaan usein ajatella *hypoteeseiksi*, joista täsmälleen yksi on tosi.  $T$  on taas jokin tunnettu tieto satunnaiskokeen tuloksesta: tiedämme, että tapahtuma  $T$  on sattunut. Todennäköisyydet  $P(H_i)$ ,  $i = 1, 2, \dots, k$  ovat hypoteeseja koskevia ns. *prioritodennäköisyyksiä*, jotka voivat kuvastaa uskoa tai luottamusta kyseisiin hypoteeseihin. Ehdollista todennäköisyyttä  $P(H_i | T)$  kutsutaan hypoteesin  $H_i$  *posterioritodennäköisyydeksi* tai posterioriluottamukseksi hypoteesiin  $H_i$ . Tapahtuman  $T$  todennäköisyys  $P(T | H_i)$  ehdolla, että hypoteesi  $H_i$  on tosi, on tapahtuman  $T$  *uskottavuus* (likelihood) ehdolla  $H_i$ .

### 3.7.1 Peräkkäisotanta

Populaatiossa on  $N$  henkilöä, joista  $Np$  ( $0 \leq p \leq 1$ ) 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  $\{X_1, X_2, \dots, X_n\}$  tai lyhyesti  $\{X_i | 1 \leq i \leq n\}$ . 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

$$(3.7.5) \quad P(A_2) = P(A_1) P(A_2 | A_1) + P(A_1^c) P(A_2 | A_1^c).$$

Helposti nähdään, että

$$P(A_1) = \frac{Np}{N} = p \quad \text{ja} \quad P(A_1^c) = \frac{N - Np}{N} = 1 - p.$$

Toisaalta

$$P(A_2 | A_1) = \frac{Np - 1}{N - 1} \quad \text{ja} \quad P(A_2 | A_1^c) = \frac{Np}{N - 1},$$

koska 1. haastatellun jälkeen jäljellä on  $N - 1$  haastateltavaa, joiden joukossa on  $Np - 1$   $B$ :n kannattajaa, jos 1. haastateltava oli  $B$ :n kannattaja. Jos 1. haastateltava ei ollut  $B$ :n kannattaja, niin jäljellä on vielä  $Np$   $B$ :n kannattajaa. Kun nämä todennäköisyydet sijoitetaan kaavaan (3.7.5), saadaan

$$P(A_2) = p \frac{Np - 1}{N - 1} + (1 - p) \frac{Np}{N - 1} = p.$$

Näin olemme osoittaneet, että  $P(A_1) = P(A_2)$ . Mutta tämä tulos pitää paikkansa yleisesti:

$$(3.7.6) \quad P(A_i) = p, \quad i = 1, 2, \dots, n; \quad 1 \leq n \leq N.$$

Näytämme nyt, että tämä yleinen tulos pitää paikkansa. Voimme ajatella, että  $B$ :n kannattajat on numeroitu  $1, 2, \dots, Np$  ja muut  $Np + 1, Np + 2, \dots, N$ . Kysymyksessä on otanta palauttamatta, kun järjestys otetaan huomioon. Tarkastellaan tapahtumaa  $A_{i+1}$ , että  $(i + 1)$ . haastateltava on  $B$ :n kannattaja. Kaikkien  $(i + 1)$ :n kokoisten järjestettyjen jonojen (otosten) lukumäärä on  $N^{(i+1)}$ . Sellaisia jonoja, joissa  $(i + 1)$ . alkio on 1 ( $B$ :n kannattaja) on  $Np(N - 1)^{(i)}$  kappaletta, koska  $B$ :n kannattaja voidaan valita  $Np$  tavalla ja loput  $i$  otosalkiota  $(N - 1)^{(i)}$  tavalla. Tuloperiaatteen mukaan suotuisia otoksia on siis  $Np(N - 1)^{(i)}$  kappaletta. Tästä seuraa, että

$$\begin{aligned} P(A_{i+1}) &= \frac{Np(N - 1)^{(i)}}{N^{(i+1)}} = \frac{pN(N - 1) \cdots (N - 1 - i + 1)}{N^{(i+1)}} \\ &= \frac{pN^{(i+1)}}{N^{(i+1)}} = p. \end{aligned}$$

Olemme näin todistaneet tuloksen (3.7.6)

Määritellään nyt satunnaismuuttuja

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

joka on  $B$ :n kannattajien lukumäärä otoksessa. Tiedämme aikaisempien tarkastelujen perusteella, että  $X$  noudattaa hypergeometrista jakaumaa  $H\text{Geo}(n, N, p)$ . Johdimme Esimerkissä 3.11 hypergeometrisen jakauman odotusarvon. Nyt tämä odotusarvo on helppo laskea satunnaismuuttujan  $X$  avulla, koska

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

koska

$$E(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p, \quad i = 1, 2, \dots, n.$$

Jos satunnaismuuttuja  $X/n$  valitaan  $p$ :n estimaattoriksi, voimme todeta, että

$$E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p.$$

Sanomme, että  $X/n$  on *harhaton estimaattori*.

### 3.8 Usean tapahtuman unionin todennäköisyys

Lauseessa 2.5 esitettiin kolmen tapahtuman  $A_1$ ,  $A_2$  ja  $A_3$  unionin todennäköisyyden  $P(A_1 \cup A_2 \cup A_3)$  lauseke. Yleistetään nyt tämä tulos  $n$ :n tapahtuman  $A_1, A_2, \dots, A_n$  unionin tapaukseen.

**Lause 3.17** *Samassa otosavaruudessa määriteltyjen tapahtumien  $A_1, A_2, \dots, A_n$  unionin  $\bigcup_{i=1}^n A_i$  todennäköisyys on*

$$(3.8.1) \quad P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{j>i}^n P(A_i A_j) + \sum_{k>j>i}^n P(A_i A_j A_k) \\ - + \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n).$$

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

$$(3.8.2) \quad I_{\bigcup_{i=1}^n A_i} = 1 - \prod_{i=1}^n (1 - \alpha_i) \\ = \sum_{i=1}^n \alpha_i - \sum_{j>i}^n \alpha_i \alpha_j + \sum_{k>j>i}^n \alpha_i \alpha_j \alpha_k \\ - + \dots + (-1)^{n-1} \alpha_1 \alpha_2 \dots \alpha_n.$$

Kun nyt yhtälössä (3.8.2) otetaan odotusarvot puolittain ja käytetään hyväksi odotusarvon lineaarisuutta, saadaan tulos (3.8.1). Huomaa, että indikaattorifunktion  $I_A$  odotusarvo  $E(I_A) = P(A)$  on vastaavan tapahtuman todennäköisyys. Silloin  $E(I_{\bigcup_{i=1}^n A_i}) = P(\bigcup_{i=1}^n A_i)$ ,  $E(\alpha_i) = P(A_i)$ ,  $E(\alpha_i \alpha_j) = P(A_i A_j)$ ,  $\dots$ ,  $E(\alpha_1 \alpha_2 \dots \alpha_n) = P(A_1 A_2 \dots A_n)$ .  $\square$

**Esimerkki 3.21 (Yhteensopivuusongelma)** Meillä on kaksi  $n$ :n kortin korttipakkaa, joiden kortit on numeroitu juoksevasti 1:stä  $n$ :ään. Asetetaan 1. pakan kortit pöydälle riviin numerojärjestyksessä 1, 2,  $\dots$ ,  $n$ . 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

$$(3.8.3) \quad P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

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

$$(3.8.4) \quad P(A_i A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

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

$$(3.8.5) \quad P(A_i A_j A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}$$

ja yleisesti

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

Todennäköisyys, että saadaan ainakin yksi pari on siis Lauseen 3.17 mukaan

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \binom{n}{1} \frac{1}{n} - \binom{n}{2} \frac{1}{n(n-1)} + \binom{n}{3} \frac{1}{n(n-1)(n-2)} \\ &\quad - + \dots + (-1)^{n-1} \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - + \dots + (-1)^{n-1} \frac{1}{n!}. \end{aligned}$$

Huomaa, että

$$1 - \frac{1}{2!} + \frac{1}{3!} - + \dots + (-1)^{n-1} \frac{1}{n!} + \dots = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i!} = 1 - e^{-1} = 0.632\dots$$

Kun siis  $n$  on suuri, niin

$$P\left(\bigcup_{i=1}^n A_i\right) \approx 1 - e^{-1} = 0.632\dots$$

Suurilla  $n$ :n arvoilla todennäköisyys saada ainakin yksi pari on hyvin lähellä lukua 0.632.  $\square$

## Satunnaisuuttujat, ehdollistaminen ja riippumattomuus: Yhteenveto

### Todennäköisyys

- Ehdollinen todennäköisyys

$$P(B | A) = \frac{P(AB)}{P(A)}, \quad P(A) \neq 0.$$

- Tulosääntö  $P(AB) = P(A)P(B | A)$ .
- Yleinen tulokaava

$$P(A_1 A_2 A_3 \cdots A_{n-1} A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 A_2) \cdots \\ \cdot P(A_n | A_1 A_2 \cdots A_{n-1}).$$

- Riippumattomuus.  $A$  ja  $B$  ovat riippumattomat, jos  $P(AB) = P(A)P(B)$ .
- $P(A_1$  tai  $A_2$  tai  $A_3)$

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

### Satunnaisuuttujat

- 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

$$E(X + Y) = E(X) + E(Y) \quad \text{ja} \quad E(cX) = cE(X), \quad \text{missä } c \text{ on vakio.}$$

- Varianssi

$$\text{Var}(X) = E(X - \mu)^2 = E(X^2) - \mu^2, \quad \mu = E(X).$$

- Lineaarinen muunnos  $cX + b$

$$E(cX + b) = cE(X) + b, \quad \text{Var}(cX + b) = c^2 \text{Var}(X), \quad b \text{ ja } c \text{ vakioita.}$$

- Cauchyn ja Schwarzin epäyhtälö

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

- Kovarianssi

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y,$$

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

- Summat

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y),$$

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

- 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)  $\text{Cov}(X, Y) = 0$ ,
- 4)  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

- Markovin epäyhtälö

$$P(X \geq a) \leq \frac{E(X)}{a}, \quad \text{missä } X \geq 0 \text{ ja } a > 0.$$

- Tšebyševin epäyhtälö

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2},$$

missä  $\varepsilon > 0$ ,  $\mu = E(X)$ ,  $\sigma^2 = \text{Var}(X)$ .

- Otoskeskiarvo

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n),$$

$$E(\bar{X}_n) = \mu \quad \text{ja} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n},$$

jos  $E(X_i) = \mu$  ja  $\text{Var}(X_i) = \sigma^2$ ,  $i = 1, 2, \dots, n$ .

- Suurten lukujen laki (heikko):  $P(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0$ , kun  $n \rightarrow \infty$  ja  $X_1, X_2, \dots, X_n$  ovat riippumattomat ja noudattavat samaa jakaumaa.

## Generoivat funktiot ja momentit

- Satunnaismuuttujan momentit

$$\begin{aligned} X\text{:n } r\text{-momentti} & \quad \alpha_r = E(X^r), \\ r\text{-keskusmomentti} & \quad \mu_r = E(X - \mu)^r, \\ r\text{-tekijämomentti} & \quad g_r = E[X^{(r)}] = E[X(X-1)\cdots(X-r+1)]. \end{aligned}$$

- Momenttifunktio

$$M_X(t) = E(e^{tX}); \quad t \in (-a, a), \quad a > 0.$$

- 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

$$G(t) = E(t^X), \quad X \text{ on diskreetti.}$$

- $r$ -tekijämomentti.  $G$ :n  $r$ -derivaatta pisteessä  $t = 1$  on  $X$ :n  $r$ -tekijämomentti:  $G^{(r)}(1) = E[X^{(r)}]$ .
- $G(t)$  vs.  $M(t)$ :  $G(e^t) = E(e^{tX}) = M(t)$ .

## Bayesin lause

- Kokonaistodennäköisyys

$$P(T) = \sum_{i=1}^k P(H_i) P(T | H_i),$$

missä  $T \subset \Omega$  ja  $H_1, H_2, \dots, H_k$  on  $\Omega$ :n ositus.



- Bayesin kaava

$$P(H_i | T) = \frac{P(H_i) P(T | H_i)}{\sum_{j=1}^k P(H_j) P(T | H_j)}.$$

- Prioritodennäköisyydet  $P(H_i)$ .
- Posterioritodennäköisyydet  $P(H_i | T)$ ,  $i = 1, 2, \dots, n$ .
- Uskottavuus.  $P(T | H_i)$  on tapahtuman  $T$  uskottavuus ehdolla, että  $H_i$  on tosi.

## Harjoituksia

1. Oletetaan, että  $P(X = 0) = 1 - P(X = 1)$  ja  $E(X) = 3 \text{Var}(X)$ . Laske  $P(X = 0)$ .
2. Olkoon satunnaismuuttujan  $X$  todennäköisyysfunktio

$$f(x) = \frac{(|x| + 1)^2}{9}, \quad x = -1, 0, 1.$$

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  $g(b) = E[(X - b)^2] = E(X^2) - 2bE(X) + b^2$ .)
4. Olkoon  $\Omega = \{\omega_1, \omega_2, \omega_3\}$  ja  $P(\omega_1) = P(\omega_2) = P(\omega_3) = \frac{1}{3}$ . Määritellään satunnaismuuttujat  $X$ ,  $Y$  ja  $Z$  seuraavasti:

$$\begin{array}{lll} X(\omega_1) = 1, & X(\omega_2) = 2, & X(\omega_3) = 3, \\ Y(\omega_1) = 2, & Y(\omega_2) = 3, & Y(\omega_3) = 1, \\ Z(\omega_1) = 3, & Z(\omega_2) = 1, & Z(\omega_3) = 2. \end{array}$$

- (a) Osoita, että satunnaismuuttujilla  $X$ ,  $Y$  ja  $Z$  on sama todennäköisyysjakauma.
  - (b) Määritä satunnaismuuttujien  $X + Y$ ,  $Y + Z$ ,  $X + Z$  ja
  - (c) satunnaismuuttujien  $\sqrt{(X^2 + Y^2)Z}$  ja  $Z/|X - Y|$  todennäköisyysjakauma.
5. Populaatiossa on  $M$  miestä ja  $N$  naista. Miehistä on  $m$  ja naisista  $n$  tupakoitsijaa. Populaatiosta valitaan satunnaisesti yksi.  $A$  on tapahtuma, että valittu on mies ja  $B$  tapahtuma, että on valittu tupakoitsija. Mitkä ehdot lukumäärien  $M$ ,  $N$ ,  $m$  ja  $n$  on toteutettava, jotta  $A$  ja  $B$  ovat toisistaan riippumattomat?

6. Tarkastellaan Esimerkin 2.8 tilannetta, jossa Pekka ja Paavo pelaavat ”kruunua ja klaavaa” (satunnaiskävely,  $n = 20$ ).
- Mikä on todennäköisyys, että Pekka on 5 heiton jälkeen voitolla yhden euron, 10 heiton jälkeen 2 euroa, 20 heiton jälkeen 2 euroa?
  - Mikä on Pekan voiton odotusarvo 20 heiton sarjassa?
  - Jos Pekka on 5. heiton jälkeen voitolla euron, mikä on Pekan voiton odotusarvo 20. heiton jälkeen?
7. Tarkastellaan Tehtävän 6 peliä simuloimalla ( $n = 20$ ).
- Mikä on todennäköisin voittosumma? Epätodennäköisin voittosumma? Hahmottele voittosumman todennäköisyysjakauma.
  - Kuinka usein Pekka on voitolla pelin aikana? Hahmottele tämän satunnaismuuttujan todennäköisyysjakauma.
8. Oletetaan, että  $X \sim \text{Tasd}(1, N)$  noudattaa diskreettiä tasajakaumaa.
- Jos  $E(x) = 6$ , niin mitä on  $\text{Var}(X)$ ?
  - Olkoon  $X \sim \text{Tasd}(3, 8)$ . Laske  $E(X)$  ja  $\text{Var}(X)$ .
9. Suuressa tehtaassa sattuu 5:n päivän jakson aikana 3 onnettomuutta. Oletetaan, että kaikki mahdolliset  $5^3$  erilaista 3:n onnettomuuden sijoitumista 5:n päivän jaksolle ovat yhtä todennäköisiä.
- Olkoon  $Y = \{\text{Onnettomuuspäivien lukumäärä jakson aikana}\}$  ja  $X$  niiden päivien lukumäärä, jolloin onnettomuuksia ei satu.
- Määritä satunnaismuuttujan  $X = 5 - Y$  todennäköisyysfunktio.
  - Laske  $E(X)$  ja  $\text{Var}(X)$ .
10. Lennolla Havannasta Helsinkiin laukkuni eivät olleet perillä Helsingissä samaan aikaan kuin minä. Laukkuja on reitillä siirretty koneesta toiseen 3 kertaa ja todennäköisyydet, että siirtoa ei ole tehty ajoissa tai oikein, ovat siirtojärjestyksessä 0.4, 0.2 ja 0.1. Mikä on todennäköisyys, että moka sattui jo ensimmäisessä siirrossa?
11. Olkoot  $X$  ja  $Y$  riippumattomat kokonaislukuarvoiset satunnaismuuttujat, joilla on sama todennäköisyysfunktio  $f_X(n) = f_Y(n) = p_n$ ,  $n \geq 1$ . Laske todennäköisyydet  $P(X = Y)$  ja  $P(X \leq Y)$ .
12. Tarkastellaan kaksilapsisia perheitä. Oletetaan pojat ja tytöt yhtä todennäköisiksi ja 2. lapsen sukupuoli on riippumaton 1. lapsen sukupuolesta.

Tarkastellaan neljää tapahtumaa:

$A = 1$ . lapsi on poika,

$B =$  lapset ovat eri sukupuolta,

$C = 1$ . lapsi on tyttö,

$D = 2$ . lapsi on poika.

- (a) Mitkä tapahtumaparit  $\{A, B\}$ ,  $\{A, C\}$ ,  $\{B, C\}$  ovat keskenään riippumattomat?
- (b) Ovatko tapahtumat  $A$ ,  $B$  ja  $D$  keskenään eli täydellisesti riippumattomat?
- 13.**  $A$ ,  $B$  ja  $C$  ampuvat maaliin 20 laukausta. Yhden laukauksen osumistodennäköisyys on  $A$ :lla 0.4,  $B$ :llä 0.3,  $C$ :llä 0.1 ja laukaukset ovat toisistaan riippumattomat. Olkoot  $X_A$ ,  $X_B$  ja  $X_C$  vastaavasti  $A$ :n,  $B$ :n ja  $C$ :n osumien lukumäärät ja  $X$  osumien kokonaismäärä.
- (a) Määrittele  $X$  riippumattomien satunnaismuuttujien summana ja laske sen avulla  $X$ :n odotusarvo ja varianssi.
- (b) Määritä Tšebyševin epäyhtälön avulla väli, jolle osumien kokonaismäärä osuu vähintään todennäköisyydellä  $\frac{8}{9}$ .
- 14.** Liukuhihnalta tulevat pullot ovat vikaantuneita, toisistaan riippumatta, todennäköisyydellä 0.2. Hihnalta tulevat pullot tarkistetaan, vikaantuneet poistetaan ja loput pakataan 12 pullon laatikoihin.
- (a) Millä todennäköisyydellä on tutkittava täsmälleen 17 pulloa, kunnes laatikko saadaan täyteen?
- (b) Ainakin 17 pulloa, kunnes laatikko saadaan täyteen?
- 15.** Lääkärillä oli oheisessa taulukossa esitetty uuden hoidon vaikutusta koskeva potilasaineisto:

	Asuu kaupungissa		Asuu maaseudulla	
	Saanut hoidon	Ei hoitoa	Saanut hoidon	Ei hoitoa
Elossa	1000	50	95	5000
Kuollut	9000	950	5	5000

Tarkastellaan tapahtumia  $A =$  'potilas elossa',  $B =$  'saanut hoidon' ja  $C =$  'asuu kaupungissa'. Estimoi tarvittavat todennäköisyydet taulukon frekvenssien avulla ja laske

- (a)  $P(A | B)$  ja  $P(A | B^c)$  sekä
- (b)  $P(A | BC)$ ,  $P(A | B^cC)$ ,  $P(A | BC^c)$  ja  $P(A | B^cC^c)$ .

(c) Oliko hoidosta apua?

- 16.** Olkoot  $X$  ja  $Y$  sellaiset satunnaismuuttujat, että  $E(X) = \mu_X$ ,  $E(Y) = \mu_Y$ ,  $\text{Var}(X) = \sigma_X^2$ ,  $\text{Var}(Y) = \sigma_Y^2$  ja  $\rho = \text{Cor}(X, Y)$ . Käytetään satunnaismuuttujan  $Y$  arvioimiseen regressioennustetta  $\hat{Y} = \alpha + \beta X$ , missä  $\alpha$  ja  $\beta$  ovat vakioita. Ennusteen keskineliövirhe määritellään

$$\text{MSE}(\hat{Y}) = E([Y - (\alpha + \beta X)]^2).$$

(a) Osoita laskemalla, että

$$\text{MSE}(\hat{Y}) = [\mu_Y - (\alpha + \beta\mu_X)]^2 + \text{Var}(Y - \beta X).$$

(b) Valitse edellisessä  $\text{MSE}(\hat{Y})$ :n lausekkeessa  $\alpha = \mu_Y - \beta\mu_X$  ja näytä, että silloin

$$\text{MSE}(\hat{Y}) = (\beta\sigma_X - \rho\sigma_Y)^2 + \sigma_Y^2(1 - \rho^2).$$

(c) Päätele nyt, että  $\text{MSE}(\hat{Y})$  saavuttaa miniminsä  $\sigma_Y^2(1 - \rho^2)$ , kun  $\alpha = \mu_Y - \beta\mu_X$  ja  $\beta = \rho\sigma_Y/\sigma_X$ .

- 17.** Olkoon  $X$  sellainen diskreetti satunnaismuuttuja, että sen todennäköisyysfunktio on  $P(X = x_i) = p_i$ ,  $i \geq 1$  ja 2. momentti  $E(X^2) = \sum_i p_i x_i^2$  on olemassa. Olkoon  $A = \{i \mid |x_i| \geq \varepsilon\}$ , missä  $\varepsilon > 0$ .

(a) Osoita, että

$$P(|X| \geq \varepsilon) = \sum_{i \in A} p_i \text{ ja } E(X^2) \geq \sum_{i \in A} p_i x_i^2,$$

(b)  $\sum_{i \in A} p_i x_i^2 \geq \sum_{i \in A} p_i \varepsilon^2$

(c) ja lopuksi  $P(|X| \geq \varepsilon) \leq E(X^2)/\varepsilon^2$ .

# Luku 4

## Diskreetit jakaumat

Diskreetti satunnaismuuttuja määriteltiin alaluvussa 2.5. Olemme jo edellisissä luvuissa käsitelleet hypergeometrista jakaumaa (alaluku 2.6.1), binomijakaumaa (alaluvut 2.8 ja 3.6) ja sen erikoistapauksena Bernoullin jakaumaa sekä diskreettiä tasajakaumaa (alaluku 2.5.4), jotka kaikki ovat esimerkkejä *diskreeteistä jakaumista*.

### 4.1 Diskreetti satunnaismuuttuja

**Määritelmä 4.1** Otosavaruudessa  $\Omega$  määritelty satunnaismuuttuja  $X$  on diskreetti, jos sen arvojoukko  $S \subset \mathbb{R}$  on numeroituva ja  $P(X \in S) = 1$ . Joukon  $S$  pisteillä on positiivinen todennäköisyys ja ne ovat  $X$ :n kertymäfunktion  $F$  *hyppypisteitä* ja näiden pisteiden todennäköisyydet ovat  $F$ :n hyppyjä.

Määritellään nyt yksinkertainen *hyppyfunktio*  $\varepsilon(x)$  seuraavasti:

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

Olkoon  $X$ :n arvoalue  $S = \{1, 2, 3, \dots\}$  ja  $P(x = i) = p_i$ ,  $i \geq 1$ . Silloin  $X$ :n kertymäfunktio  $F(X)$  voidaan kirjoittaa muodossa

$$(4.1.1) \quad F(x) = \sum_{i=1}^{\infty} p_i \varepsilon(x - i).$$

Vaikka usein tarkastelemme vain kokonaislukuarvoisia satunnaismuuttujia, se ei ole teoreettiselta kannalta oleellinen rajoitus. Olkoon  $S^* = \{x_1, x_2, x_3, \dots\}$  diskreetin satunnaismuuttujan arvojoukko. Silloin joukkojen  $S$  ja  $S^*$  välillä on bijektiivinen vastaavuus  $g(x_i) = i$  ja  $P(X = x_i) = P(g(X) = i)$ , joten voimme aina tarvittaessa siirtyä tarkastelemaan vastaavaa kokonaislukuarvoista satunnaismuuttujaa.

**Esimerkki 4.1** Yksinkertaisin satunnaismuuttuja  $X$  on sellainen, jonka arvoalue  $S = \{c\}$  on yksi piste, jolloin  $P(X = c) = 1$ . Silloin  $X$ :n kertymäfunktio on

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

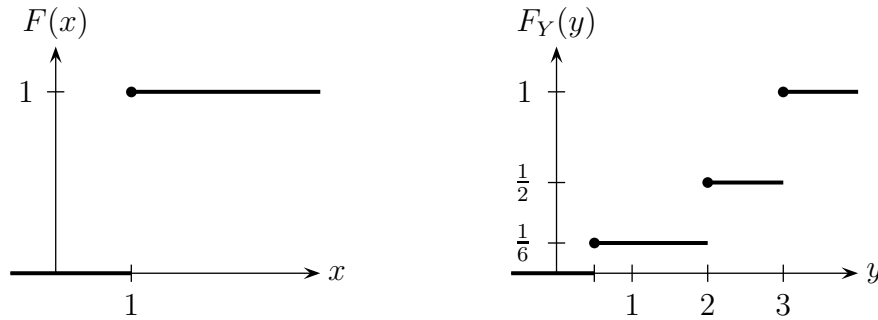
Olkoon  $Y$ :n todennäköisyysfunktio

$$P(Y = \frac{1}{2}) = \frac{1}{6}, \quad P(Y = 2) = \frac{1}{3} \quad \text{ja} \quad P(Y = 3) = \frac{1}{2}.$$

Silloin  $Y$ :n kertymäfunktio on

$$F_Y(y) = \frac{1}{6} \varepsilon(y - \frac{1}{2}) + \frac{1}{3} \varepsilon(y - 2) + \frac{1}{2} \varepsilon(y - 3).$$

□



**Kuvio 4.1.** Funktioiden  $F(x) = \varepsilon(x - 1)$  ja  $F_Y(y)$  kuvaajat.

**Esimerkki 4.2** Hatussa on  $N$  arpalippua, jotka on numeroitu juoksevasti ykkösestä lähtien. Valitaan hatusta arpa satunnaisesti palauttaen  $n$  kertaa ja merkitään valittujen arpojen numerot muistiin. Olkoon  $X$  suurin valittujen arpojen numeroista. Silloin  $P(X \leq r) = (r/N)^n$  ja

$$\begin{aligned} P(X = r) &= P(X \leq r) - P(X \leq r - 1) \\ &= \left(\frac{r}{N}\right)^n - \left(\frac{r-1}{N}\right)^n. \end{aligned}$$

Määritelmän mukaan  $X$ :n odotusarvo on

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

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

□

## 4.2 Bernoullin kokeet ja binomijakauma

Alaluvussa 2.8 binomijakauma esiteltiin tarkastelemalla otantaa palauttaen ja alaluvussa 3.6 binomijakauma liitettiin Bernoullin kokeisiin. Bernoullin koe on satunnaiskoe, jolla on täsmälleen kaksi toisensa poissulkevaa tulostulosta (onnistuminen ja epäonnistuminen — lyhyesti O ja E). Esimerkiksi mielipidetiedustelussa henkilö kannattaa tai ei kannata ehdokasta, laatu- ja virhetiedustelussa tuote on virheetön tai viallinen, hoidon tuloksena potilas paranee tai ei parane.

Satunnaismuuttuja  $X$  noudattaa *Bernoullin jakaumaa*, kun

$$(4.2.1) \quad X = \begin{cases} 1 & \text{todennäköisyydellä } p, \\ 0 & \text{todennäköisyydellä } 1 - p, \end{cases}$$

missä  $0 \leq p \leq 1$ . 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

$$E(X) = p \quad \text{ja} \quad \text{Var}(X) = pq,$$

sillä

$$E(X) = p \cdot 1 + q \cdot 0 = p, \quad E(X^2) = p \cdot 1^2 + q \cdot 0^2 = p$$

ja

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p) = pq.$$

Merkitsemme  $X \sim \text{Ber}(p)$ , kun  $X$  noudattaa Bernoullin jakaumaa, jonka odotusarvo on  $p$ .

Jos  $X \sim \text{Ber}(p)$ , 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  $\text{Ber}(p)$  momenttifunktio on

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

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

**Esimerkki 4.3 (Sabharwal 1969).** Olkoon  $n$ :n Bernoullin kokeen jonossa  $X_1, X_2, \dots, X_n$  onnistumistodennäköisyys  $P(O) = p$  ja vastaavasti  $P(E) = 1 - p$  ( $E$  = epäonnistuminen). Olkoon  $Y_n$  tapahtuman OE (osajono) esiintymisten lukumäärä koejonossa. Mikä on tällaisten osajonojen lukumäärän odotusarvo  $E(Y_n)$ ? Määritellään ensin uusi satunnaismuuttuja

$$Z_i = h(X_i, X_{i+1}) = \begin{cases} 1, & \text{jos } X_i = O \text{ ja } X_{i+1} = E; \\ 0 & \text{muulloin,} \end{cases}$$

kun  $i = 1, 2, \dots, n - 1$ . 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  $n$  riippumattonta 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  $X_1, X_2, \dots, X_n$ , missä  $P(X_i = 1) = p$  ja  $P(X_i = 0) = q$ ,  $i = 1, 2, \dots, n$ . Kun koe on tehty, tulos voisi olla esimerkiksi 111011000...110. Tällaisen tuloksen todennäköisyys (ennen koetta) olisi

$$ppp(1-p)p(1-p)(1-p)ppp \cdots pp(1-p) = p^k(1-p)^{n-k},$$

missä  $k$  on onnistumisten lukumäärä ja  $n - k$  epäonnistumisten lukumäärä. Olkoon  $X$  onnistumisten lukumäärä  $n$ :ssä riippumattomassa Bernoullin kokeessa. Alaluvussa 3.6 totesimme, että  $X$  noudattaa binomijakaumaa parametrein  $n$  ja  $p$ . Silloin merkitään  $X \sim \text{Bin}(n, p)$ . Binomijakauman todennäköisyysfunktio on

$$(4.2.2) \quad f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n.$$

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



**Lause 4.1** Tehdään  $n$  riippumatonta Bernoullin koetta, joissa jokaisessa onnistumistodennäköisyys on  $p$ . Olkoon  $X$  onnistumisten lukumäärä. Silloin

$$X \sim \text{Bin}(n, p).$$

**Todistus.** Koska  $X$  on onnistumisten lukumäärä  $n$ :ssä riippumattomassa Bernoullin kokeessa, niin  $X = X_1 + X_2 + \dots + X_n$ , missä  $X_i \sim \text{Ber}(p) = \text{Bin}(1, p)$ ,  $i = 1, 2, \dots, n$  ovat riippumattomat ja noudattavat samaa Bernoullin jakaumaa. Merkitään nyt  $X = S_n$  ja

$$S_n = X_1 + X_2 + \dots + X_n = S_{n-1} + X_n.$$

Todistamme väitteen induktiolla.

Kun  $n = 1$ , niin oletuksen mukaan  $X = X_1 \sim \text{Ber}(p) = \text{Bin}(1, p)$ , joten väite pitää paikkansa tapauksessa  $n = 1$ . Teemme nyt induktio-oletuksen  $S_{n-1} \sim \text{Bin}(n-1, p)$  ja näytämme, että  $S_n \sim \text{Bin}(n, p)$ .

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

$$\begin{aligned} P(S_{n-1} + X_n = k) &= P(S_{n-1} = k) P(X_n = 0) + P(S_{n-1} = k-1) P(X_n = 1) \\ &= \binom{n-1}{k} p^k (1-p)^{n-1-k} (1-p) + \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} p \\ &= \binom{n-1}{k} p^k (1-p)^{n-k} + \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= \left[ \binom{n-1}{k} + \binom{n-1}{k-1} \right] p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned}$$

missä viimeinen yhtäsuuruus seuraa siitä, että  $\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$  [Pascalin kolmio]. Näin on lause todistettu.  $\square$

**Esimerkki 4.4** Erään kasvin siementen itämistodennäköisyydeksi on ilmoitettu 0.8. Siemenen itäminen on tässä ”onnistuminen” ja itämistodennäköisyys on onnistumistodennäköisyys. Jos kylvetään 10 siementä ja siementen itämistapahtumat ovat toisistaan riippumattomat, niin kylvöä voidaan pitää

kymmenenä riippumattomana Bernoullin kokeena, joissa onnistumistodennäköisyys on 0.8. Silloin itävien siementen lukumäärä  $X \sim \text{Bin}(10, 0.8)$ , eli

$$f(x) = \binom{10}{x} 0.8^x \cdot 0.2^{10-x}, \quad x = 0, 1, \dots, 10.$$

Mikä on todennäköisyys, että vähemmän kuin 9 jyvää itää? Todennäköisyys

$$\begin{aligned} P(X < 9) &= P(X \leq 8) = 1 - \sum_{k=9}^{10} P(X = k) \\ &= 1 - 10 \cdot 0.8^9 \cdot 0.2 - 0.8^{10} = 0.6242. \end{aligned}$$

□

Laskemme usein muotoa  $P(X \leq x)$  olevia todennäköisyyksiä, kuten edellisessä esimerkissä. Todennäköisyydet  $P(X \leq x)$  määrittelevät jakauman kertymäfunktion

$$F(x) = P(X \leq x).$$

Kertymäfunktio määriteltiin alaluvussa 2.5.2. Binomijakauman kertymäfunktion arvot pisteissä  $x = 0, 1, \dots, n$  ovat

$$F(x) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k}.$$

**Lause 4.2** Jos  $X \sim \text{Bin}(n, p)$ , niin

1.  $X$ :n todennäköisyysfunktio  $f(x)$  on

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

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 hyppöfunktio;

3.  $X$ :n odotusarvo, varianssi ja momenttifunktio ovat

$$\begin{aligned} \mu &= E(X) = np, & \text{Var}(X) &= np(1-p), \\ M(t) &= E(e^{tX}) = (1-p + pe^t)^n, & -\infty < t < \infty. \end{aligned}$$

**Todistus.** 1. Binomijakauman todennäköisyysfunktio johdettiin Lauseen 4.1 todistuksessa.

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

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

ja

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

3. Momenttifunktio on

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= E(e^{t(X_1+X_2+\dots+X_n)}) = E(e^{tX_1+tX_2+\dots+tX_n}) \\ &= E(e^{tX_1}e^{tX_2}\dots e^{tX_n}) \\ &= E(e^{tX_1})E(e^{tX_2})\dots E(e^{tX_n}), \end{aligned}$$

missä viimeinen yhtäsuuruus seuraa lauseista 3.6 ja 3.10. Koska  $X_i$  ja  $X_j$  ( $i \neq j$ ) ovat riippumattomat, niin  $e^{tX_i}$  ja  $e^{tX_j}$  ovat riippumattomat (Lause 3.6) ja riippumattomien satunnaismuuttujien  $e^{tX_1}, e^{tX_2}, \dots, e^{tX_n}$  tulon odotusarvo on yksittäisten tulon tekijöiden odotusarvojen tulo (Lause 3.10). Koska

$$M_{X_i}(t) = E(e^{tX_i}) = 1 - p + pe^t, \quad i = 1, 2, \dots, n,$$

niin

$$M(t) = (1 - p + pe^t)^n \quad \text{kaikilla } t \in \mathbb{R}.$$

Momenttifunktio itse asiassa määrittelee yksikäsitteisesti todennäköisyysfunktion (Lause 3.12). Näytämme kuitenkin vielä eksplisiittisesti, että binomitodennäköisyydet määrittelevät todennäköisyysfunktion. Koska Binomilauseen 2.6 perusteella

$$[p + (1-p)]^n = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = 1$$

kaikilla  $p \in [0, 1]$ , niin todennäköisyydet  $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$  määrittelevät todennäköisyysfunktion kaikilla  $p \in [0, 1]$  ja  $n \geq 1$ . Huomaa myös, että

$$M(0) = (1 - p + pe^0)^n = [p + (1-p)]^n.$$

□

**Seuraus 4.1** Jos  $X_1 \sim \text{Bin}(n_1, p)$  ja  $X_2 \sim \text{Bin}(n_2, p)$  ovat riippumattomat, niin  $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$ .

**Todistus.** Koska Lauseen 4.2 mukaan  $X_1$ :n momenttifunktio on  $(1 - p + pe^t)^{n_1}$  ja  $X_2$ :n momenttifunktio on  $(1 - p + pe^t)^{n_2}$ , niin satunnaismuuttujan  $X_1 + X_2$  momenttifunktio on Lauseen 3.13 mukaan  $(1 - p + pe^t)^{n_1+n_2}$ . Mutta Lauseen 4.2 perusteella  $(1 - p + pe^t)^{n_1+n_2}$  on binomijakuman  $\text{Bin}(n_1 + n_2, p)$  momenttifunktio. Tästä seuraa momenttifunktion yksikäsitteisyyden (Lause 3.12) nojalla, että  $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$ .  $\square$

Seurauslauseen 4.1 todistuksessa on käytetty esimerkin vuoksi yleistä momenttifunktiootekniikkaa. Tässä tapauksessa tulos saadaan kuitenkin helposti turvautumatta noin voimakkaisiin menetelmiin. Koska  $X_1$  esittää onnistumisten lukumäärää  $n_1$ :ssä Bernoullin kokeessa ja  $X_2$  onnistumisten lukumäärää  $n_2$ :ssa kokeessa, missä  $p$  on jokaisen kokeen onnistumistodennäköisyys, niin riippumattomien satunnaismuuttujien  $X_1$  ja  $X_2$  summa  $X_1 + X_2$  esittää onnistumisen lukumäärää  $(n_1 + n_2)$ :ssa kokeessa. Tämän perusteella saadaan tulos  $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$ . Analyttisesti tulos voidaan tarkistaa laskemalla lauseke

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{i=0}^{n_1} P(X_1 = i, X_2 = k - i) \\ &= \sum_{i=0}^{n_1} P(X_1 = i) P(X_2 = k - i) \\ &= \sum_{i=0}^{n_1} \binom{n_1}{i} p^i (1-p)^{n_1-i} \binom{n_2}{k-i} p^{k-i} (1-p)^{n_2-k+i}, \end{aligned}$$

missä  $\binom{n_2}{k-i} = 0$  kaikilla  $k - i > n_2$ . Tästä seuraa

$$P(X_1 + X_2 = k) = p^k (1-p)^{n_1+n_2-k} \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2}{k-i}.$$

Soveltamalla hypergeometrista identiteettiä (ks. Lause 2.8)

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

saadaan kaivattu tulos.

### 4.3 Odotusaikojen jakaumat

Monissa sovelluksissa on kiinnostuksen kohteena odotusaika siihen hetkeen, että jokin tietty tapahtuma sattuu. Tässä alaluvussa käsitellään Bernoullin kokeisiin ja yksinkertaiseen satunnaistantaan liittyviä odotusaikatehtäviä.

### 4.3.1 Odotusajat Bernoullin kokeissa

Tarkastellaan riippumattomien samaa Bernoullin jakaumaa noudattavien satunnaismuuttujien jonoa  $X_1, X_2, \dots, X_n$ , missä  $X_i \sim \text{Ber}(p)$ . Määritellään satunnaismuuttujat  $S_n$  ja  $W_r$  seuraavasti:

$$S_n = X_1 + X_2 + \dots + X_n,$$

$$W_r = r\text{:ään onnistumiseen tarvittavien yritysten määrä.}$$

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  $r, r+1, r+2, \dots$ . Tiedämme, että  $S_n \sim \text{Bin}(n, p)$ , mutta mikä on  $W_r$ :n jakauma?

**Esimerkki 4.5** Heitetään harhatonta lanttia, kunnes saadaan kruunu (R). Olkoon  $W_1$  tarvittavien heittojen lukumäärä. Tapahtuma  $\{W_1 = x\}$  sattuu vain silloin, kun  $(x-1)$ :llä ensimmäisellä heitolla on saatu pelkkiä klaavoja (L) ja  $x$ . heitolla saadaan kruunu:

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

Tästä seuraa, että

$$P(W_1 = x) = \frac{1}{2^x}, \quad x = 1, 2, \dots$$

Satunnaismuuttujan  $W_1$  odotusarvo on määritelmän mukaan

$$(4.3.1) \quad E(W_1) = \sum_{x=1}^{\infty} \frac{x}{2^x}.$$

Tiedämme, että

$$(4.3.2) \quad \sum_{x=0}^{\infty} p^x = 1 + p + p^2 + p^3 + \dots = \frac{1}{1-p}, \quad \text{kun } |p| < 1.$$

Kun derivoimme sarjan (4.3.2) termeittäin, saamme

$$(4.3.3) \quad 0 + 1 + 2p + 3p^2 + \dots = \sum_{x=0}^{\infty} (x+1)p^x = \frac{1}{(1-p)^2}, \quad \text{kun } |p| < 1.$$

Koska sarjan (4.3.2) suppenemissäde on 1, suppenee derivointiopeeraation tuloksena saatu sarja (4.3.3) arvoilla  $|p| < 1$ . Sijoittamalla  $p = \frac{1}{2}$  sarjaan (4.3.3) saadaan

$$\sum_{x=0}^{\infty} (x+1) \left(\frac{1}{2}\right)^x = 4,$$

joka voidaan esittää muodossa

$$\sum_{x=0}^{\infty} x \left(\frac{1}{2}\right)^x + \sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} x \left(\frac{1}{2}\right)^x + 2 = 4,$$

missä summa  $\sum_{x=0}^{\infty} \left(\frac{1}{2}\right)^x = 2$  saadaan kaavasta (4.3.2). Nyt siis odotusarvo (4.3.1) on 2.

Jos kruunun todennäköisyys on  $p$ , niin silloin

$$P(W_1 = x) = \underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ kertaa}} p = (1-p)^{x-1} p$$

ja

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

missä sarjan summa saadaan (4.3.3):n avulla. Satunnaismuuttuja  $W_1$  on siis kruunun tai yleisemmin 'onnistumisen' odotusaika. Jakaumaa

$$(4.3.4) \quad P(W_1 = x) = (1-p)^{x-1} p, \quad x = 1, 2, \dots$$

kutsutaan *geometriseksi jakaumaksi*. Todennäköisyydet (4.3.4) todellakin määrittelevät jakauman, koska

$$\sum_{x=1}^{\infty} P(W_1 = x) = \sum_{x=1}^{\infty} (1-p)^{x-1} p = p \cdot \sum_{x=0}^{\infty} (1-p)^x = p \cdot \frac{1}{p} = 1.$$

□

Tapahtuma  $\{W_r = x\}$  sattuu, kun  $(x-1)$ :ssä ensimmäisessä kokeessa on saatu  $r-1$  onnistumista ja  $x$ . kokeessa saadaan onnistuminen:

$$\begin{array}{l} \underbrace{\text{OOEOE}\dots\text{EO}} \\ \left. \begin{array}{l} x-1 \text{ koetta,} \\ r-1 \text{ onnistumista,} \\ \text{kokeiden järjestys} \\ \text{mielivaltainen} \end{array} \right\} \begin{array}{l} x. \text{ koe,} \\ r. \text{ onnistuminen} \end{array} \end{array}$$

Nyt siis  $\{W_r = x\} = \{S_{x-1} = r-1, X_x = 1\}$ . Koska  $X_i$ :t ( $i = 1, 2, \dots, x$ ) ovat riippumattomat, niin myös  $S_{x-1}$  ja  $X_x$  ovat riippumattomat. Silloin

$$\begin{aligned} (4.3.5) \quad P(W_r = x) &= P(S_{x-1} = r-1) P(X_x = 1) \\ &= \binom{x-1}{r-1} p^{r-1} (1-p)^{x-r} p = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \end{aligned}$$

koska  $S_{x-1} \sim \text{Bin}(x-1, p)$ . Todennäköisyydet (4.3.5) määrittelevät ns. *negatiivisen binomijakauman*. Soveltamalla identiteettiä [ks. (2.4.5)]

$$\frac{r}{x} \binom{x}{r} = \binom{x-1}{r-1}$$

saadaan

$$P(W_r = x) = \frac{r}{x} P(S_x = r).$$

Toinen usein käyttökelpoinen identiteetti on

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

### 4.3.2 Geometrinen jakauma ja negatiivinen binomijakauma

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

$$(4.3.6) \quad P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, r+2, \dots$$

Merkitsemme silloin

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

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

$$(4.3.7) \quad f(x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \dots$$

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

**Lause 4.3** *Oletetaan, että  $X \sim \text{NBin}(r, p)$ .*

1. *Funktio (4.3.6) on negatiivisen binomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvuilla  $r$  ja kaikilla  $0 < p < 1$  ja*

2.

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2},$$

$$M(t) = E(e^{tX}) = \frac{(pe^t)^r}{[1 - (1-p)e^t]^r}, \quad t < -\log(1-p).$$

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

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

Binomisarja  $\sum_{y=0}^{\infty} \binom{-r}{y} [-(1-p)e^t]^y$  suppenee (Lause 2.7), kun  $(1-p)e^t < 1$ , joka on yhtäpitävä epäyhtälön  $t < -\log(1-p)$  kanssa.

Koska  $M(0) = 1$  kaikilla positiivisilla kokonaisluvuilla  $r$  ( $r \in \mathbb{N}$ ) ja kaikilla  $0 < p < 1$ , niin (4.3.6) on todennäköisyysfunktio kaikilla  $r \in \mathbb{N}$  ja kaikilla  $0 < p < 1$ . Odotusarvo ja varianssi saadaan laskemalla ensin  $M(t)$ :n 1. ja 2. derivaatta ja niiden avulla

$$E(X) = M'(0) \quad \text{ja} \quad \text{Var}(X) = M''(0) - [M'(0)]^2.$$

□

**Seuraus 4.2** Jos  $X \sim \text{Geo}(p)$ , niin  $X \sim \text{NBin}(1, p)$  ja

1. funktio (4.3.7) on geometrisen jakauman todennäköisyysfunktio kaikilla  $0 < p < 1$  ja

2.

$$\begin{aligned}
 E(X) &= \frac{1}{p}, & \text{Var}(X) &= \frac{1-p}{p^2}, \\
 M(t) = E(e^{tX}) &= \frac{pe^t}{[1 - (1-p)e^t]}, & t &< -\log(1-p).
 \end{aligned}$$

Olkoon  $Y$  epäonnistumisten lukumäärä Bernoullin toistokokeessa, ennen kuin saadaan  $r$ . onnistuminen. Koska  $r$ . onnistumiseen tarvittavien yritysten määrä  $W_r \sim \text{NBin}(r, p)$ , niin

$$Y = W_r - r \quad \text{ja} \quad E(Y) = E(W_r) - r = \frac{r}{p} - r = \frac{r(1-p)}{p}.$$



$Y$ :n varianssi on tietysti sama kuin  $W_r$ :n varianssi. Nyt siis  $P(Y = y) = P(W_r = r + y)$  kaikilla  $y = 0, 1, 2, \dots$

Nimitys ”negatiivinen binomijakauma” on peräisin esitystavasta

$$1 = p^r \cdot p^{-r} = p^r [1 - (1 - p)]^{-r} = p^r \sum_{y=0}^{\infty} \binom{-r}{y} [-(1 - p)]^y,$$

mistä saadaan todennäköisyydet  $P(W_r = y + r)$ ,  $y = 0, 1, 2, \dots$ . Merkintä  $\binom{-r}{y}$  on määritelmänsä mukaan

$$\binom{-r}{y} = \frac{(-r)^{(y)}}{y!} = (-1)^y \binom{r + y - 1}{y},$$

missä  $r > 0$  ja  $y \geq 0$  ovat kokonaislukuja.

**Esimerkki 4.6** Geometrisella jakaumalla ja negatiivisella binomijakaumalla on tärkeä merkitys esimerkiksi jonoteoriassa. Oletetaan, että joukko asiakkaita jonottaa pääsyä palvelutiskille. Olkoon todennäköisyys  $p$ , että jokaisella pienellä aikavälillä tulee 1 uusi asiakas (0 uutta asiakasta todennäköisyydellä  $1 - p = q$ ). Silloin seuraavan asiakkaan odotusaika  $W \sim \text{Geo}(p)$ . Todennäköisyys  $P(W > k)$ , että seuraavan  $k$ :n aikayksikön aikana ei tule asiakasta, on

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

□

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  $k = 1, 2, \dots$ . 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:

$$(4.3.8) \quad P(W > k + s \mid W > k) = \frac{P(W > k + s)}{P(W > k)} = \frac{q^{k+s}}{q^k} = q^s.$$

Nyt siis todennäköisyys, että asiakkaan palveleminen kestää vielä  $s$  aikayksikköä, ei riipu siitä, kuinka kauan häntä on jo palveltu. Onneksi kuitenkin käytännössä palveluaika ei aina täysin noudata geometrista jakaumaa.

**Esimerkki 4.7 Banachin tulitikkuongelma.** Piippua polttelevalla matemaatikolla oli tapana pitää yksi tulitikkulaatikko oikeassa ja yksi vasemmassa taskussa. Joka kerta tikkua tarvitessaan hän valitsi taskun täysin satunnaisesti, joten kummankin taskun valintatodennäköisyys on  $\frac{1}{2}$ . Tarkastellaan tapahtumaa, että matemaatikko huomaa laatikon olevan tyhjä. Oletetaan, että kummassakin laatikossa oli alunperin  $N$  tikkua. Mikä on todennäköisyys, että toisessa laatikossa on täsmälleen  $k$  tikkua ( $k = 0, 1, \dots, N$ ) silloin, kun matemaatikko havaitsee toisen laatikon olevan tyhjä?

Olkoon  $A$  tapahtuma, että matemaatikko huomaa oikeanpuoleisen laatikon olevan tyhjä ja samalla vasemman taskun laatikossa on  $k$  tikkua. Tapahtuma voi sattua täsmälleen silloin, kun oikeanpuoleisen taskun laatikosta valitaan tikku ( $N+1$ ). kerran ja yhteensä valintoja on tehty  $N+1+N-k$  kappaletta. Teemme siis valintoja palauttamatta. Molemmissa laatikoissa on  $N$  tikkua, joten tapahtuma  $A$  on ekvivalentti tapahtuman  $\{W_{N+1} = N+1+N-k\}$  kanssa. Saamme kaavalla (4.3.6) todennäköisyydeksi

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

Koska myös todennäköisyys, että vasemmanpuoleinen laatikko huomataan tyhjäksi ja oikeanpuoleisessa on  $k$  tikkua, on  $P(W_{N+1} = N+1+N-k)$ , niin vastaus kysymykseen on

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

□

### 4.3.3 Odotusajat peräkkäisotannassa

Oletetaan, että populaatiossa on kahdenlaisia alkioita. Valitaan populaatiosta peräkkäisotos. Käytetään nyt apuna urnamallia. Olkoon urnassa  $a$  valkoista palloa ja  $b$  mustaa palloa eli yhteensä  $a+b=N$  palloa. Poimitaan satunnaisvalinnalla palloja urnasta 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.

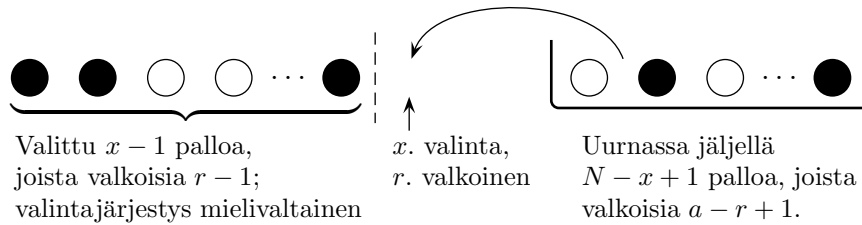
Jos otanta tehdään *palauttaen*, niin peräkkäiset nostot ovat riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on  $p = a/N$ . Tässä tapauksessa voidaan suoraan soveltaa edellä esitettyjä Bernoullin kokeita koskevia tuloksia.

Kun otanta tehdään *palauttamatta*, peräkkäiset nostot eivät ole riippumattomia, koska valkoisten pallojen suhteellinen osuus uurnassa riippuu siitä, mitä sieltä on jo valittu. Alaluvussa 2.6.1 osoitimme, että  $S_n$  noudattaa hypergeometrista jakaumaa, kun otanta tehdään palauttamatta (ks. myös alaluku 3.7.1). Silloin

$$(4.3.9) \quad P(S_n = x) = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}},$$

kun  $x = 0, 1, \dots, n$ . 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  $S_{x-1} \sim \text{HGeo}(x - 1, N, a/x)$  [ks. Esimerkki 3.11 ja (3.3.6)] ja  $X_x = 1$ , kun valitaan valkoinen pallo  $x$ . nostossa. Tästä seuraa, että

$$(4.3.10) \quad \begin{aligned} P(W_r = x) &= P(S_{x-1} = r - 1, X_x = 1) \\ &= P(S_{x-1} = r - 1) P(X_x = 1 \mid S_x = r - 1) \\ &= \frac{\binom{a}{r-1} \binom{N-a}{x-r}}{\binom{N}{x-1}} \cdot \frac{a - r + 1}{N - x + 1}, \end{aligned}$$

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

Todennäköisyys (4.3.10) voidaan kirjoittaa lausekkeena

$$(4.3.11) \quad P(W_r = x) = \binom{x-1}{r-1} \frac{\binom{N-x}{a-r}}{\binom{N}{a}},$$

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  $S_x \sim \text{HGeo}(x, N, a/N)$ . Vastaavanlainen tulos saatiin otannassa palauttaen. Samoin on jälleen helppo nähdä, että

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

Merkitään  $W_r \sim \text{NHGeo}(r, N, p)$ , missä  $p = a/N$ .

### 4.3.4 Hypergeometrinen jakauma ja negatiivinen hypergeometrinen jakauma

Olemme esitelleet hypergeometrisen jakauman tarkastelemalla otantaa palauttamatta (alaluku 2.6.1). Jakauman avulla voidaan siis ratkaista otantaan liittyviä todennäköisyystehtäviä. Hypergeometrisen jakauman momenttifunktiolla  $M(t)$  ei ole olemassa siistiä lauseketta, vaikka se tietysti voidaan lausua määritelmänsä mukaan äärellisenä summana, koska satunnaismuuttujan arvojoukko on äärellinen. Hypergeometrisen jakauman odotusarvon ja varianssin laskeminen ei myöskään ole aivan helppo tehtävä.

Olemme merkinneet populaation alkioden 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 tyyppiä 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  $\text{HGeo}(n, N, p)$ , niin  $X$ :n todennäköisyysfunktio on

$$(4.3.12) \quad f(x) = P(X = x) = \frac{\binom{Np}{x} \binom{N-Np}{n-x}}{\binom{N}{n}}, \quad x \in S.$$

Huomattakoon, että todennäköisyys (4.3.12) on määritelty myös arvoilla  $x \notin S$ , mutta silloin  $f(x) = 0$ .

**Lause 4.4** *Oletetaan, että  $X \sim \text{HGeo}(n, N, p)$ . Silloin*

$$E(X) = np \quad \text{ja} \quad \text{Var}(X) = \frac{N-n}{N-1} np(1-p).$$

**Todistus.** Hypergeometrisen jakauman odotusarvo laskettiin esimerkissä 3.11 ja alaluvussa 3.7.1. Varianssi voidaan laskea vastaavalla tavalla.  $\square$

**Lause 4.5** *Oletetaan, että  $Y \sim \text{NHGeo}(r, N, p)$ . Silloin*

$$E(Y) = r \cdot \frac{N+1}{Np+1} \quad \text{ja} \quad \text{Var}(Y) = \frac{rN(N+1)(1-p)(Np+1-r)}{(Np+1)^2(Np+2)}.$$

Mainitsimme jo alaluvussa 2.8.1, että binomijakaumaa voidaan käyttää hypergeometrisen jakauman likiarvona, kun  $N$  on suuri. Erityisesti, kun  $N$  on ääretön tai hyvin suuri (verrattuna otoskokoon), on yhdentekevää, käytetäänkö otantaa palauttaen vai palauttamatta. Oletetaan nyt, että

$$X_N \sim \text{HGeo}(n, N, p) \quad \text{ja} \quad X \sim \text{Bin}(n, p).$$

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

$$X_N \xrightarrow{d} X, \quad \text{kun} \quad N \rightarrow \infty.$$

Koska  $X \sim \text{Bin}(n, p)$ , niin

$$X_N \xrightarrow{d} \text{Bin}(n, p),$$

eli  $X_N$ :n jakauma lähestyy binomijakaumaa, jonka parametrit ovat  $n$  ja  $p$ . Sanomme myös, että  $X_N$ :n jakauma suppenee kohti  $X$ :n jakaumaa  $N$ :n kasvaessa. Kutsumme  $X$ :n jakaumaa  $X_N$ :n *asymptoottiseksi jakaumaksi*.

Lauseen 3.5 mukaan satunnaismuuttujilla on sama jakauma, jos niillä on sama kertymäfunktio. Voimme nyt tarkastella satunnaismuuttujien jonoa

$$\{X_N; N = 1, 2, \dots\} = X_1, X_2, \dots$$

ja vastaavaa kertymäfunktioiden jonoa

$$\{F_N; N = 1, 2, \dots\} = F_1, F_2, \dots,$$

missä  $F_N(x)$  on  $X_N$ :n kertymäfunktio.

**Määritelmä 4.2** Jono  $\{X_N; N = 1, 2, \dots\}$  suppenee jakaumaltaan kohti satunnaismuuttujaa  $X$ , jos

$$\lim_{N \rightarrow \infty} F_N(x) = F(x)$$

kaikissa pisteissä  $x \in \mathbb{R}$ , joissa  $X$ :n kertymäfunktio  $F(x)$  on jatkuva.

Diskreettien satunnaismuuttujien tapauksessa voidaan helposti todistaa tulos, joka osoittaa, että suppenemista jakaumamielessä voidaan tarkastella yhtä hyvin myös todennäköisyysfunktioiden avulla.

**Lause 4.6** *Olkoon  $\{X_N; N = 1, 2, \dots\}$  sellainen epänegatiivisten kokonaislukuarvoisten satunnaismuuttujien jono, että  $X_N$ :n todennäköisyysfunktio on  $f_N(k)$ ,  $N = 1, 2, \dots$ . Olkoon  $X$  epänegatiivinen kokonaislukuarvoinen satunnaismuuttuja, jonka todennäköisyysfunktio on  $f(k)$ . Silloin*

$$X_N \xrightarrow{d} X \Leftrightarrow \lim_{N \rightarrow \infty} f_N(k) = f(k)$$

*kaikilla epänegatiivisilla kokonaisluvuilla  $k$ .*

**Todistus.** Jätetään harjoitustehtäväksi.  $\square$

**Lause 4.7** Jos  $X_N \sim \text{HGeo}(n, N, p)$ , niin

$$X_N \xrightarrow{d} \text{Bin}(n, p), \quad \text{kun } N \rightarrow \infty.$$

**Todistus.** Käytetään lausetta 4.6 ja osoitetaan, että  $P(X_N = k) = f_N(k) \rightarrow f(k)$  kaikilla epänegatiivisilla kokonaisluvuilla  $k$ , kun  $N \rightarrow \infty$ . Yksityiskohdat jätetään lukijan pohdittavaksi.  $\square$

### 4.3.5 Tasajakauma

Diskreetti tasajakauma esiteltiin ensimmäisen kerran alaluvussa 2.5.4. Satunnaismuuttuja  $X$ , jonka arvoavaruus on  $S = \{1, 2, \dots, N\}$ , noudattaa diskreettiä tasajakaumaa, jos

$$P(X = x) = \frac{1}{N}, \quad k = 1, 2, \dots, N.$$

Silloin merkitään  $X \sim \text{Tasd}(1, 2, \dots, N)$ , missä  $N \geq 1$  on annettu positiivinen kokonaisluku. Jos  $X \sim \text{Tasd}(1, 2, \dots, N)$ , niin

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

## 4.4 Poissonin jakauma

Satunnaismuuttuja  $X$ , jonka todennäköisyysfunktio on

$$(4.4.1) \quad f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \dots$$

noudattaa *Poissonin jakaumaa* parametrilla  $\lambda > 0$ , joka on Poissonin jakauman odotusarvo. Silloin merkitään

$$X \sim \text{Poi}(\lambda).$$

Poissonin jakaumalla on runsaasti sovelluksia eri aloilla. Sitä voidaan käyttää myös binomijakauman  $\text{Bin}(n, p)$  likiarvona, kun  $n$  on suuri ja  $p$  pieni. Silloin siis pätee

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{e^{-np} (np)^x}{x!}.$$

**Lause 4.8** Olkoon  $X \sim \text{Poi}(\lambda)$ . Silloin

1. funktio (4.4.1) on Poissonin jakauman todennäköisyysfunktio kaikilla  $\lambda > 0$  ja

2.

$$\begin{aligned}\mu &= E(X) = \lambda, & \text{Var}(X) &= \lambda, \\ M(t) &= E(e^{tX}) = \exp(\lambda e^t - \lambda).\end{aligned}$$

**Todistus.** Sovelletaan eksponenttifunktion sarjakehitelmää

$$(4.4.2) \quad \exp(\lambda) = e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}.$$

1. Ensinnäkin  $f(x) \geq 0$  kaikilla  $x = 0, 1, 2, \dots$ , ja eksponenttifunktion sarjakehitelmän (4.4.2) perusteella

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^\lambda = 1.$$

2. Johdetaan ensin momenttifunktion  $M(t)$  lauseke:

$$\begin{aligned}M(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} \cdot \exp(\lambda e^t) = \exp(\lambda e^t - \lambda).\end{aligned}$$

Odotusarvo ja varianssi saadaan sitten laskemalla  $M(t)$ :n 1. ja 2. derivaatta ja soveltamalla identiteettejä

$$E(X) = M'(0) \quad \text{ja} \quad \text{Var}(X) = M''(0) - [M'(0)]^2.$$

□

Riippumattomien Poissonin jakaumaa noudattavien satunnaismuuttujien summa noudattaa myös Poissonin jakaumaa.

**Lause 4.9** *Olko*  $X_1, X_2, \dots, X_n$  *riippumattomat ja*  $X_i \sim \text{Poi}(\lambda_i)$ ,  $i = 1, 2, \dots, n$ . *Olkoon*  $Y = X_1 + X_2 + \dots + X_n$ . *Silloin*

$$Y \sim \text{Poi}(\lambda),$$

missä  $\lambda = \sum_{i=1}^n \lambda_i$ .

**Todistus.** Seurauslauseen 3.1 mukaan

$$\begin{aligned}M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n \exp(\lambda_i e^t - \lambda_i) = \exp[(e^t - 1)\lambda],\end{aligned}$$

missä  $\lambda = \sum_{i=1}^n \lambda_i$ . Lauseesta 3.12 seuraa sitten väite  $Y \sim \text{Poi}(\lambda)$ .

□

Jos riippumattomat  $X_1, X_2, \dots, X_n$  noudattavat samaa Poissonin jakaumaa  $\text{Poi}(\lambda)$ , niin Lauseen 4.9 mukaan niiden summa  $Y = X_1 + X_2 + \dots + X_n$  noudattaa Poissonin jakaumaa  $\text{Poi}(n\lambda)$ . Poissonin jakauma on hyvä binomijakauman  $\text{Bin}(n, p)$  likiarvo silloin, kun  $n$  on suuri ja  $p$  pieni.

Kun  $X \sim \text{Bin}(n, p)$ , niin binomitodennäköisyys on

$$(4.4.3) \quad f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Annetaan nyt  $p$ :n riippua  $n$ :stä ja merkitään lausekkeessa (4.4.3)  $p = p_n$ . Valitaan erityisesti

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

Tarkastellaan nyt binomijakaumien jonoa

$$\text{Bin}(1, p_1), \text{Bin}(2, p_2), \text{Bin}(3, p_3), \dots$$

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

$$(4.4.4) \quad P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}, \quad 0 \leq x \leq n.$$

Merkitään todennäköisyyttä (4.4.4) lyhyesti  $b_x(n)$

Kiinnitetään nyt  $x$  ja annetaan  $n$ :n kasvaa rajatta. Osoittautuu, että  $b_x(n)$  suppenee kaikilla  $x$ . Valitaan ensin  $x = 0$ . Silloin saamme

$$(4.4.5) \quad \lim_{n \rightarrow \infty} b_0(n) = \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Se on eräs keskeinen eksponenttifunktioon liittyvä kaava, joka pitäisi analyysin kurssien perusteella muistaa. Tulos (4.4.5) saadaan esimerkiksi Taylorin sarjan

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

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

$$(4.4.6) \quad \begin{aligned} \log\left(1 - \frac{\lambda}{n}\right)^n &= n \log\left(1 - \frac{\lambda}{n}\right) = n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \frac{\lambda^3}{3n^3} - \dots\right) \\ &= -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \dots \\ &= -\lambda - \frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \dots\right). \end{aligned}$$

Kun  $n \rightarrow \infty$ , niin  $\frac{1}{n}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{3n} + \dots\right) \rightarrow 0$  ja siksi  $\log\left(1 - \frac{\lambda}{n}\right)^n \rightarrow -\lambda$ .



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  $\frac{n-x}{n} \rightarrow 1$  ja  $1 - \frac{\lambda}{n} \rightarrow 1$ , kun  $n \rightarrow \infty$ . Tästä seuraa, että

$$(4.4.7) \quad \lim_{n \rightarrow \infty} \frac{b_{x+1}(n)}{b_x(n)} = \frac{\lambda}{x+1}.$$

Kun lähdetään tuloksesta (4.4.5) ja käytetään hyväksi raja-arvoa (4.4.7), saadaan

$$\begin{aligned} \lim_{n \rightarrow \infty} b_1(n) &= \frac{\lambda}{1} \lim_{n \rightarrow \infty} b_0(n) = \lambda e^{-\lambda}, \\ \lim_{n \rightarrow \infty} b_2(n) &= \frac{\lambda}{2} \lim_{n \rightarrow \infty} b_1(n) = \frac{\lambda^2}{1 \cdot 2} e^{-\lambda}, \\ &\vdots \\ \lim_{n \rightarrow \infty} b_x(n) &= \frac{\lambda}{x} \lim_{n \rightarrow \infty} b_{x-1}(n) = \frac{\lambda^x}{1 \cdot 2 \cdots x} e^{-\lambda}. \end{aligned}$$

Olemme siis näyttäneet, että

$$(4.4.8) \quad \lim_{n \rightarrow \infty} b_x(n) = \frac{\lambda^x}{x!} e^{-\lambda},$$

missä raja-arvo on  $P(X = x)$ , kun  $X \sim \text{Poi}(\lambda)$ . Tulos (4.4.8) tunnetaan *Poissonin raja-arvolakina*.

Satunnaismuuttujat noudattavat samaa jakaumaa, kun niillä on sama kertymäfunktio (Lause 3.5). Jos diskreetit satunnaismuuttujat noudattavat samaa jakaumaa, niin niillä on sama todennäköisyysfunktio. Jos satunnaismuuttujan  $X_n$  jakauma lähenee  $X$ :n jakaumaa  $n$ :n kasvaessa rajatta, niin  $X_n$ :n todennäköisyysfunktio lähenee  $X$ :n todennäköisyysfunktioita, mikäli jakaumat ovat diskreettejä (Lause 4.6). Vaikka edellä olemmekin johtaneet Poissonin raja-arvolain (4.4.8), esitetään tulos vielä *Poissonin lauseena*.

**Lause 4.10 (Poissonin lause)** *Olkoon  $X_n \sim \text{Bin}(n, p)$ . Silloin*

$$X_n \xrightarrow{d} \text{Poi}(\lambda),$$

*kun  $n \rightarrow \infty$  siten, että  $np = \lambda$ .*

**Todistus.** Koska  $np = \lambda$ , voimme merkitä  $p = \lambda/n$ . Todistus perustuu

Lauseeseen 4.6. Jos  $X_n \sim \text{Bin}(n, p)$ , niin

(4.4.9)

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

Kiinteällä  $x$ :n arvolla

$$\lim_{n \rightarrow \infty} \left[ \binom{n}{n} \binom{n-1}{n} \cdots \binom{n-x+1}{n} \right] = 1$$

ja

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1.$$

Näistä tuloksista yhdessä raja-arvon (4.4.5) kanssa seuraa

$$\lim_{n \rightarrow \infty} f_{X_n}(x) = \frac{e^{-\lambda} \lambda^x}{x!}.$$

Satunnaismuuttujan  $X_n$  jakauma lähestyy siis Poissonin jakaumaa  $\text{Poi}(\lambda)$ , kun  $n \rightarrow \infty$ .  $\square$

Poissonin jakaumaa sanotaan usein harvinaisten tapahtumien laiksi. Tämä luonnehdinta perustuu edellisessä lauseessa esitettyyn ominaisuuteen. Jos tehdään suuri määrä riippumattomia Bernoullin kokeita, joissa onnistumistodennäköisyys on hyvin pieni, niin silloin Lauseen 4.10 mukaan onnistumisten lukumäärä noudattaa likimain Poissonin jakaumaa. Esimerkiksi suuri määrä ihmisiä on päivittäin alttiina liikenneonnettomuuksille. Yksittäisen henkilön todennäköisyys (onnistumistodennäköisyys!) joutua onnettomuuteen on pieni, mutta onnettomuuksille alttiina olevien henkilöiden lukumäärä  $n$  on suuri. Silloin onnettomuuksien lukumäärä noudattaa likimain Poissonin jakaumaa.

**Lause 4.11** *Olkoot  $X$  ja  $Y$  sellaiset riippumattomat satunnaismuuttujat, että  $X \sim \text{Poi}(\lambda_1)$  ja  $Y \sim \text{Poi}(\lambda_2)$ . Silloin  $X$ :n ehdollinen jakauma ehdolla  $X + Y$  on binomijakauma.*

**Todistus.** Olkoot  $m$  ja  $n$  sellaiset epänegatiiviset kokonaisluvut, että  $m < n$ .

Silloin

$$\begin{aligned}
 P(X = m \mid X + Y = n) &= \frac{P(X = m, X + Y = n)}{P(X + Y = n)} \\
 &= \frac{P(X = m, Y = n - m)}{P(X + Y = n)} \\
 &= \frac{P(x = m) P(Y = n - m)}{P(X + Y = n)} \\
 &= \frac{e^{-\lambda_1} (\lambda_1^m / m!) e^{-\lambda_2} [\lambda_2^{n-m} / (n - m)!]}{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^n / n!} \\
 &= \binom{n}{m} \frac{\lambda_1^m \lambda_2^{n-m}}{(\lambda_1 + \lambda_2)^n} \\
 &= \binom{n}{m} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^m \left( 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{n-m}
 \end{aligned}$$

on binomitodennäköisyys kaikilla  $m = 0, 1, \dots, n$ . Näin on lause todistettu.  $\square$

Lauseella 4.11 on tärkeä merkitys esimerkiksi frekvenssiaineistojen analyysissä.

**Esimerkki 4.8** Tiedetään, että auto-onnettomuuksien lukumäärä aikayksikössä (esimerkiksi kuukaudessa) noudattaa Poissonin jakaumaa. Tarkastellaan eräällä tieosuudella lokakuussa sattuvien onnettomuuksien lukumäärää. Aikaisempien tilastojen perusteella voidaan olettaa, että auto-onnettomuuksien lukumäärä  $Z$  kyseisellä tieosuudella (kuukaudessa) noudattaa Poissonin jakaumaa  $\text{Poi}(\lambda)$ . Onnettomuudet luokitellaan mahdollisten henkilövahinkojen mukaan vakaviin ja lieviin (jokainen onnettomuus kuuluu toiseen näistä luokista). Vakavien onnettomuuksien lukumäärä  $X \sim \text{Poi}(\lambda_1)$  ja lievien lukumäärä  $Y \sim \text{Poi}(\lambda_2)$ . Lisäksi  $X$  ja  $Y$  ovat toisistaan riippumattomat. Koska  $Z = X + Y$ , niin  $E(Z) = E(X) + E(Y)$  eli  $\lambda = \lambda_1 + \lambda_2$ .

Tutkijat valitsivat poliisin tiedostoista satunnaisesti valitun kuukauden (vuonna 2003) onnettomuudet. He havaitsivat onnettomuuksien lukumääräksi 120 ( $n = 120$ ), mutta he eivät olleet vielä luokitelleet onnettomuuksia. Mitä jakaumaa noudattaa vakavien onnettomuuksien lukumäärä? Lauseen 4.11 perusteella

$$P(X = m \mid Z = 120) = \binom{120}{m} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^m \left( 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{120-m},$$

$m = 0, 1, \dots, 120$ . Vakavien onnettomuuksien lukumäärä noudattaa siis binomijakaumaa  $\text{Bin}(120, \frac{\lambda_1}{\lambda_1 + \lambda_2})$ . Aikaisempien onnettomuustilastojen perusteella voimme arvioida parametrit  $\lambda_1$  ja  $\lambda_2$ , 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  $\lambda_1$ .  $\square$

## 4.5 Poissonin prosessi

### 4.5.1 Laskuriprosessi

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

**Esimerkki 4.9** Seuraavassa luetellaan esimerkkejä laskuriprosesseista.

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 tahansa välillä sattuvien tapahtumien lukumäärän jakauma riippuu vain välin pituudesta. Jos  $N(t)$  on stationaarinen laskuriprosessi, niin satunnaismuuttujilla  $N(t_2) - N(t_1)$  ja  $N(t_2 + s) - N(t_1 + s)$  on sama jakauma kaikilla väleillä  $(t_1, t_2]$  ja  $(t_1 + s, t_2 + s]$ , missä  $t_2 > t_1$  ja  $s > 0$ .

### 4.5.2 Poissonin prosessin määrittely

Poissonin prosessi on yksi tärkeimpiä laskuriprosesseja. Se määritellään seuraavasti:

**Määritelmä 4.3** Laskuriprosessi  $\{N(t), t \geq 0\}$  on Poissonin prosessi, jonka intensiteetti on  $\lambda$  ( $\lambda > 0$ ), jos

1.  $N(0) = 0$ .
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella  $h$ :n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on  $\lambda h$ :

$$P[N(h+t) - N(t) = x] = e^{-\lambda h} \frac{(\lambda h)^x}{x!}, \quad x = 0, 1, \dots$$

kaikilla  $h, t \geq 0$ .

Laskuriprosessin osoittaminen Poissonin prosessiksi Määritelmän 4.3 avulla saattaa olla hankalaa. Ei ole mitään yksinkertaista keinoa tarkistaa esimerkiksi ehdon 3 pätevyyttä. Siksi esitetään vielä toinen määritelmä, jonka avulla voi olla helpompaa tunnistaa prosessi. Voidaan osoittaa, että määritelmät 4.3 ja 4.4 ovat yhtäpitävät.

**Määritelmä 4.4** Laskuriprosessi  $\{N(t), t \geq 0\}$  on Poissonin prosessi, jonka intensiteetti on  $\lambda$  ( $\lambda > 0$ ), jos

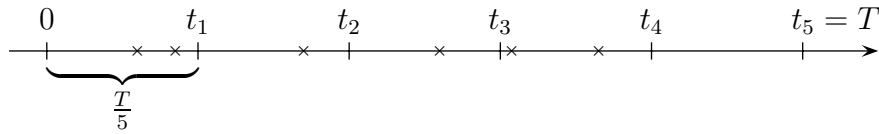
1.  $N(0) = 0$ .
2. Prosessin lisäykset ovat stationaariset ja riippumattomat.
3.  $P(N(t+h) - N(t) = 1) = \lambda h + o(h)$ .
4.  $P(N(t+h) - N(t) \geq 2) = o(h)$ .

Määritelmässä 4.4 käytetään merkintää  $o(h)$ . Sanomme, että funktio  $f(\cdot) = o(h)$ , jos

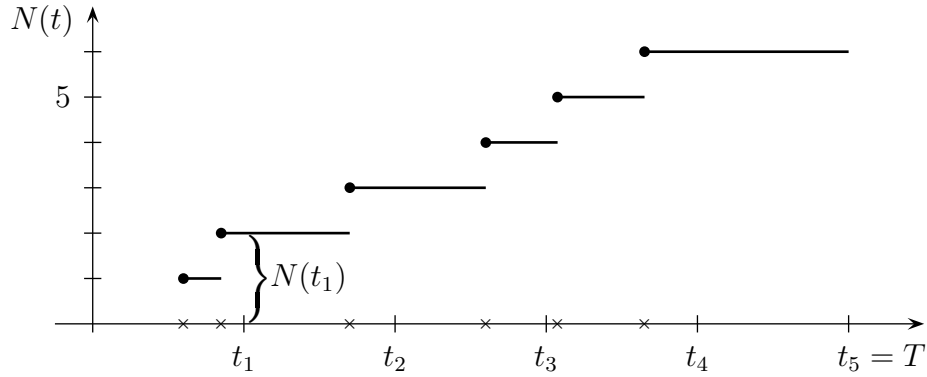
$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

**Esimerkki 4.10 Tieliikenneonnettomuudet.** Havainnoidaan esimerkiksi jollain tieosuudella sattuvien auto-onnettomuuksien lukumäärää. Onnettomuuksien määrä noudattaa tavallisesti varsin hyvin Poissonin prosessia.  $\square$

Tarkastellaan nyt hieman lähemmin Poissonin prosessin oletuksia. Oletetaan, että onnettomuuksien lukumäärä eräällä tieosuudella noudattaa aikavälillä  $(0, T)$  Poissonin prosessia, jonka intensiteetti on  $\lambda$ . Aikaväli voi olla esimerkiksi ruuhka-aika tietyssä perjantai-iltapäivänä klo 15–19 ja tieosuus jokin ulosmenotie. Oheisessa kuviossa on havaitut onnettomuudet merkitty aika-akselille.

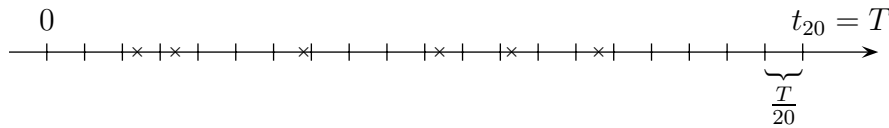


Tarkasteluväli  $(0, T]$  on jaettu viiteen yhtä pitkään osaväliin, joiden pituudet ovat  $T/5$ . Nyt esimerkiksi 1. osavälillä sattuneiden onnettomuuksien lukumäärä on  $N(t_1) - N(0) = N(t_1)$ , joka on siis hetkeen  $t$  mennessä sattuneiden onnettomuuksien lukumäärä. Kuvioon 4.2 on piirretty prosessin  $\{N(t), t \in (0, T]\}$  realisaatio, missä havaintoina ovat kyseiset onnettomuudet.



**Kuvio 4.2.** Poissonin prosessin  $\{N(t), t \in (0, T]\}$  erään realisaation kuvaaja.

Määritelmän 4.4 oletuksen 2 mukaan lisäykset  $N(t_1) - N(0)$ ,  $N(t_2) - N(t_1)$ ,  $N(t_3) - N(t_2)$ ,  $N(t_4) - N(t_3)$  ja  $N(t_5) - N(t_4)$  ovat riippumattomat ja noudattavat samaa jakaumaa. Määritelmän 4.4 oletukset 3 ja 4 tarkoittavat, että tapahtumat (onnettomuudet) sattuvat yksittäin ja samalla intensiteetillä koko tarkastelujakson ajan. Koska tapahtumat ovat erillisiä pisteitä, niin aina voidaan valita niin hienojakoinen välin ositus, että kullakin osavälillä on korkeintaan 1 tapahtuma. Jos tarkastelemassamme esimerkkitapauksessa valitaan osavälin pituudeksi  $T/20$ , sattuu tässä osituksessa kullekin osavälille korkeintaan 1 tapahtuma. Riippuen tietysti kulloisestakin havaintojaksosta, kuinka hienojakoinen ositus tarvitaan.



Todennäköisyys, että  $T/n$ :n pituiselle osavälille sattuu havainto, on Määritelmän 4.4 oletuksen 3 mukaan

$$P\left[N\left(t + \frac{T}{n}\right) - N(t) = 1\right] = \lambda \cdot \frac{T}{n} + o\left(\frac{T}{n}\right).$$

Vastaavasti todennäköisyys, että osavälillä sattuu enemmän kuin yksi havainto, on häviävän pieni, sillä Määritelmän 4.4 oletuksen 4 mukaan

$$P\left[N\left(t + \frac{T}{n}\right) - N(t) \geq 2\right] = o\left(\frac{T}{n}\right).$$

Voimme siis olettaa, että kullakin osavälillä sattuu vain 0 tai 1 tapahtumaa, kun  $n$  on riittävän suuri.

Määritellään nyt satunnaismuuttujat

$$X_i = N\left(\frac{iT}{n}\right) - N\left(\frac{(i-1)T}{n}\right), \quad i = 1, 2, \dots, n.$$

Muuttujia  $X_i$  voidaan käsitellä toisistaan riippumattomina Bernoullin jakaumaa noudattavina satunnaismuuttujina:

$$X_i \sim \text{Ber}\left(\frac{\lambda T}{n}\right), \quad i = 1, 2, \dots, n.$$

Koko välillä  $(0, T]$  havaittujen tapahtumien lukumäärä on

$$S_n = X_1 + X_2 + \dots + X_n,$$

joka noudattaa binomijakaumaa  $\text{Bin}(n, \frac{\lambda T}{n})$ . Koska  $E(S_n) = n \cdot \frac{\lambda T}{n} = \lambda T$  kaikilla  $n \in \mathbb{N}$ , niin  $E(S_n) = \lambda T$ , kun  $n \rightarrow \infty$ . Voimme siis soveltaa Poissonin lausetta (Lause 4.10), jonka mukaan  $S_n$  noudattaa Poissonin jakaumaa  $\text{Poi}(\lambda T)$ , kun  $n$  kasvaa rajatta. Näin esimerkiksi todennäköisyys, että välillä  $(0, T]$  sattuu  $x$  onnettomuutta, on

$$P(N(T) = x) = \frac{e^{-\lambda T} (\lambda T)^x}{x!}.$$

Todennäköisyys riippuu vain välin pituudesta  $T$  ja intensiteetistä  $\lambda > 0$ .

### 4.5.3 Satunnaistapahtumat tila-avaruudessa

Poissonin prosessilla mallinnetaan myös ilmiöitä, jotka tapahtuvat satunnaisesti tila-avaruudessa. Silloin Määritelmän 4.4 ehdot voidaan luonnehtia seuraavasti:

1. *Riippumattomuus*. Erillisillä alueilla sattuvien tapahtumien lukumäärät ovat riippumattomat.
2. *Yksittäisyys*. Todennäköisyys, että alueella sattuu enemmän kuin yksi tahtuma, on häviävän pieni.
3. *Homogeenisuus*. Tapahtumat sattuvat samalla intensiteetillä koko tarkasteltavalla alueella.

Tarkastellaan esimerkiksi Poissonin prosessia tasossa. Silloin todennäköisyys, että pinta-alaltaan  $A$ :n kokoisella alueella sattuu  $x$  tapahtumaa, on

$$f_A(x) = \frac{e^{-\lambda A}(\lambda A)^x}{x!}, \quad x = 0, 1, \dots,$$

missä  $\lambda$  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

$$f_V(x) = \frac{e^{-\lambda V}(\lambda V)^x}{x!}, \quad x = 0, 1, \dots,$$

missä  $\lambda$  on tapahtumien lukumäärän odotusarvo yhtä tilavuus-yksikköä kohti.

**Esimerkki 4.11** Leipomo valmistaa suuren erän pullataikinaa, josta tehdään rusinapullia. Leipuri haluaa, että ainakin 95 % pullista sisältää vähintään 2 rusinaa. Kuinka monta rusinaa pullaa kohti pitäisi sekoittaa taikinaan?

Olkoon pullan tilavuus  $V = 1$ . Kun rusinat sekoitetaan hyvin taikinaan, on kaikilla pullilla sama todennäköisyys sisältää rusinoita (homogeenisuus). Koska taikina on suuri, ovat eri pulliin sattuvien rusinoiden lukumäärät toisistaan riippumattomat. Todennäköisyys, että pieneen pullaan sattuu enemmän kuin yksi rusina, on hyvin pieni.

Tässä tilanteessa on kyse Poissonin prosessista 3-ulotteisessa tila-avaruuksessa. Pullassa on  $x$  rusinaa todennäköisyydellä

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

ja ainakin 2 rusinaa todennäköisyydellä

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-\lambda} - e^{-\lambda}\lambda. \end{aligned}$$

Leipuri vaatii, että

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

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

## 4.6 Kaksiulotteiset jakaumat

Tilastollisissa sovelluksissa tarkastellaan tavallisesti useita muuttujia samanaikaisesti. Esimerkiksi haastattelututkimuksessa valitaan opiskelijoista satunnaisotos. Jokaiselta otokseen osuneelta kysytään useita kysymyksiä ja lisäksi saadaan haastateltavien taustatiedot kuten ikä, sukupuoli, asuinpaikka jne. Otosavaruudessa on siis määritelty useita muuttujia (kysymykset ja



taustamuuttujat). Tällainen asetelma mahdollistaa muuttujien välisten riippuvuuksien tarkastelun. Seuraavassa esitellään usean muuttujan jakaumiin liittyvää käsitteistöä. Ensin käsitellään kahden muuttujan tapaus yksityiskohtaisesti. Sen jälkeen on suoraviivaista yleistää tarkastelu usean muuttujan tapaukseen.

**Määritelmä 4.5** Olkoot  $X$  ja  $Y$  samassa otosavaruudessa määritellyt diskreetit satunnaismuuttujat ja olkoon kaksiulotteisen diskreetin satunnaismuuttujan  $(X, Y)$  arvoavaruus  $S$ . Tapahtuman " $\{X = x\}$  ja  $\{Y = y\}$  sattuvat" todennäköisyyttä merkitään  $P(X = x, Y = y) = f(x, y)$ . Funktio  $f(x, y)$  on  $(X, Y)$ :n todennäköisyysfunktio (tnf), jolla on seuraavat ominaisuudet:

1.  $0 \leq f(x, y) \leq 1$ ,
2.  $\sum_{(x,y) \in S} f(x, y) = 1$  ja
3.  $P[(X, Y) \in A] = \sum_{(x,y) \in A} f(x, y)$ , missä  $A \subset S$ .

Funktiota  $f(x, y)$  sanotaan myös  $X$ :n ja  $Y$ :n yhteisjakauman todennäköisyysfunktioiksi. Moniulotteista satunnaismuuttujaa kutsutaan satunnaisvektoriksi (SV).

**Esimerkki 4.12** Olkoon  $(X, Y)$  satunnaisvektori, jonka arvoavaruus on

$$S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$$

ja todennäköisyysfunktio

$$f(x, y) = c(x + 2y), \quad (x, y) \in S.$$

Todennäköisyysfunktion ominaisuuksista seuraa, että

$$\sum_{(x,y) \in S} c(x + 2y) = c(2 + 4 + 1 + 3 + 2) = 12c = 1,$$

joten  $c = \frac{1}{12}$ . Silloin esimerkiksi

$$P(X > Y) = f(1, 0) + f(2, 0) = \frac{3}{12}$$

ja

$$P(X \geq Y) = f(1, 0) + f(2, 0) + f(1, 1) = \frac{6}{12}.$$

□

### 4.6.1 Reunajakauma ja ehdollinen jakauma

Jos  $(X, Y)$  on kaksiulotteinen satunnaisvektori, niin  $X$  ja  $Y$  ovat satunnaismuuttujia. Satunnaismuuttujan  $X$  *reunajakauman todennäköisyysfunktio*, jota merkitään  $f_X(x)$ , on on  $X$ :n todennäköisyysfunktio, kun  $Y$ :tä ei oteta huomioon. Satunnaismuuttujan  $X$  *ehdollisen jakauman todennäköisyysfunktio*, jota merkitään  $f_1(x | y)$ , on on  $X$ :n todennäköisyysfunktio, kun  $Y$ :n arvo  $Y = y$  on kiinnitetty.

**Määritelmä 4.6** Olkoon diskreetin satunnaisvektorin  $(X, Y)$  todennäköisyysfunktio  $f(x, y)$  ja arvoavaruus  $S$ . Silloin satunnaismuuttujat  $X$  ja  $Y$  ovat diskreettejä ja niiden reunajakaumien todennäköisyysfunktiot ovat

$$f_X(x) = \sum_{y \in S_Y} f(x, y), \quad x \in S_X; \quad f_Y(y) = \sum_{x \in S_X} f(x, y), \quad y \in S_Y,$$

missä  $S_X$  on  $X$ :n ja  $S_Y$   $Y$ :n arvoavaruus. Satunnaismuuttujat  $X$  ja  $Y$  ovat *riippumattomat* jos ja vain jos

$$(4.6.1) \quad P(X = x, Y = y) = P(X = x)P(Y = y)$$

kaikilla  $x \in S_X$  ja  $y \in S_Y$ . Jos  $X$  ja  $Y$  eivät ole riippumattomia, niin ne ovat riippuvia. Todennäköisyysfunktion avulla ehto (4.6.1) voidaan lausua muodossa:

$$f(x, y) = f_X(x)f_Y(y) \quad \text{kaikilla } x \in S_X \text{ ja } y \in S_Y.$$

Merkitään  $A = \{X = x\}$  ja  $B = \{Y = y\}$ , missä  $(x, y) \in S$ . Silloin  $A \cap B = \{X = x, Y = y\}$ . Koska

$$P(A \cap B) = P(X = x, Y = y) = f(x, y)$$

ja

$$P(B) = P(Y = y) = f_Y(y) > 0 \quad (\text{koska } y \in S_Y),$$

niin

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{f(x, y)}{f_Y(y)}.$$

Siksi voimme määritellä ehdollisen todennäköisyysfunktion seuraavasti:

**Määritelmä 4.7** Jos diskreetin satunnaisvektorin  $(X, Y)$  todennäköisyysfunktio on  $f(x, y)$  ja arvoavaruus  $S$ , niin  $X$ :n *ehdollinen todennäköisyysfunktio ehdolla*  $Y = y$  on

$$f_1(x | y) = \frac{f(x, y)}{f_Y(y)}, \quad (x, y) \in S$$

ja  $Y$ :n *ehdollinen todennäköisyysfunktio ehdolla*  $X = x$  on

$$f_2(y | x) = \frac{f(x, y)}{f_X(x)}, \quad (x, y) \in S.$$

**Esimerkki 4.13** Esimerkissä 4.12 käsitellyn satunnaisvektorin  $(X, Y)$  todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \quad \text{kun } (x, y) \in S,$$

missä  $S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$ . Huomaa, että  $f(x, y) = 0$ , kun  $(x, y) \notin S$ .  $X$ :n ja  $Y$ :n reunajakaumien todennäköisyysfunktioit ovat

$$f_X(x) = \sum_{y=0}^2 f(x, y) \quad \text{ja} \quad f_Y(y) = \sum_{x=0}^2 f(x, y).$$

$X$ :n ehdollinen todennäköisyysfunktio ehdolla  $Y = y$  on

$$f(x | y) = \frac{f(x, y)}{f_Y(y)}, \quad x = 0, 1, 2.$$

Saadaan siis kolme  $X$ :n ehdollista todennäköisyysfunktioita:

$$f_1(x | 0) = \frac{x}{3}, \quad x = 0, 1, 2;$$

$$f_1(x | 1) = \frac{x + 2}{5}, \quad x = 0, 1;$$

$$f_1(x | 2) = 1, \quad x = 0.$$

Vastaavalla tavalla saadaan kolme  $Y$ :n ehdollista todennäköisyysfunktioita ehdolla  $X = x$ . Satunnaisuuttajat  $X$  ja  $Y$  eivät ole riippumattomat, koska esimerkiksi

$$f(0, 1) = \frac{1}{6} \neq f_X(0)f_Y(1) = \frac{1}{2} \cdot \frac{5}{12} = \frac{5}{24}.$$

$X$ :n ja  $Y$ :n riippuvuutta (vs. riippumattomuutta) on luontevaa tarkastella ehdollisen jakauman avulla.  $Y$ :n ehdollinen todennäköisyysfunktio ehdolla  $X = 1$  on

$$f_2(y | 1) = \frac{f(1, y)}{f_X(1)} = \frac{1 + 2y}{12} \bigg/ \frac{1}{3} = \frac{1 + 2y}{4}, \quad y = 0, 1.$$

Koska

$$f_2(y | 1) \neq f_Y(y),$$

voimme jälleen päätellä, että  $X$  ja  $Y$  eivät ole riippumattomat.  $\square$

Ehdollisen jakauman odotusarvoa kutsutaan jakauman *ehdolliseksi odotusarvoksi*.  $X$ :n ehdollinen odotusarvo ehdolla  $Y = y$  on

$$(4.6.2) \quad E(X | Y = y) = \sum_{x \in S_X} xf(x | y)$$

ja  $Y$ :n ehdollinen odotusarvo ehdolla  $X = x$  on

$$(4.6.3) \quad E(Y | X = x) = \sum_{y \in S_Y} yf(y | x).$$

Näitä odotusarvoja merkitään myös

$$E(X | y) = \mu_{X|y} \quad \text{ja} \quad E(Y | x) = \mu_{Y|x}.$$

Vastaavasti määritellään  $X$ :n ehdollinen varianssi ehdolla  $Y = y$  ja  $Y$ :n ehdollinen varianssi ehdolla  $X = x$ .  $Y$ :n ehdollinen varianssi ehdolla  $X = x$  on

$$(4.6.4) \quad \begin{aligned} \text{Var}(Y | x) &= E[(Y - \mu_{Y|x})^2 | x] \\ &= \sum_{y \in S_Y} (y - \mu_{Y|x})^2 f(y | x), \end{aligned}$$

jota merkitään myös  $\text{Var}(Y | x) = \sigma_{Y|x}^2$ . Samalla periaatteella voidaan ehdollisen jakauman avulla määritellä mikä tahansa jakauman ehdollinen tunnusluku, kuten esimerkiksi ehdolliset momentit tai ehdollinen mediaani.

Kun  $E(Y | x)$  lasketaan eri  $x$ :n arvoilla, riippuu tulos yleensä  $x$ :n arvosta. Jos halutaan korostaa  $E(Y | x)$ :n riippuvuutta  $x$ :stä, merkitään esimerkiksi  $E(Y | x) = g(x)$ . Silloin ehdollinen odotusarvo määrittelee funktion  $g(x)$ .

**Esimerkki 4.14** Esimerkissä 4.13 määritettiin ehdolliset todennäköisyysfunktiot  $f_1(x | 0)$ ,  $f_1(x | 1)$  ja  $f_1(x | 1)$ , kun  $(X, Y)$ :n yhteisjakauman todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \quad (x, y) \in S.$$

Lasketaan nyt ehdolliset odotusarvot  $E(X | y)$ ,  $y = 0, 1, 2$ :

$$E(X | 0) = 0 + 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3},$$

$$E(X | 1) = 0 \cdot \frac{2}{5} + 1 \cdot \frac{3}{5} + 2 \cdot 0 = \frac{3}{5},$$

$$E(X | 2) = 0 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 = 0.$$

Ehdolliset varianssit  $\text{Var}(X | y)$ ,  $y = 0, 1, 2$ , ovat vastaavasti:

$$\text{Var}(X | 0) = \left(1 - \frac{5}{3}\right)^2 \cdot \frac{1}{3} + \left(2 - \frac{5}{3}\right)^2 \cdot \frac{2}{3} = \frac{2}{9},$$

$$\text{Var}(X | 1) = \left(0 - \frac{3}{5}\right)^2 \cdot \frac{2}{5} + \left(1 - \frac{3}{5}\right)^2 \cdot \frac{3}{5} + \left(2 - \frac{3}{5}\right)^2 \cdot 0 = \frac{6}{25},$$

$$\text{Var}(X | 2) = (0 - 0)^2 \cdot 1 + (1 - 0)^2 \cdot 0 + (2 - 0)^2 \cdot 0 = 0.$$

□

### 4.6.2 Satunnaismuuttujien funktion jakauma

Usein tarvitaan satunnaismuuttujien  $X$  ja  $Y$  jonkin funktion  $h(X, Y)$  jakaumaa. Funktio  $h(X, Y)$  voi olla esimerkiksi muotoa  $X + Y$ ,  $XY$ ,  $X^2 + Y^2$  jne. Jos  $h$  on jokin reaaliarvoinen funktio  $h(x, y)$ , voimme määritellä uuden satunnaismuuttujan  $Z = h(X, Y)$ . Olkoon satunnaisvektorin  $(X, Y)$  arvoalue  $S$ . Merkitään

$$A_z = \{ (x, y) \in S \mid h(x, y) = z \}.$$

Silloin todennäköisyys  $P(Z = z)$  voidaan laskea seuraavasti:

$$P(Z = z) = \sum_{(x,y) \in A_z} f(x, y).$$

Lasketaan siis yhteen todennäköisyydet  $f(x, y)$  kaikkissa pisteissa  $(x, y)$ , jotka toteuttavat ehdon  $h(x, y) = z$ . Tällä tavalla voidaan johtaa  $Z$ :n todennäköisyysfunktio.

**Esimerkki 4.15** Oletetaan, että satunnaisvektorin  $(X, Y)$  (Esimerkki 4.12) todennäköisyysfunktio on

$$f(x, y) = \frac{x + 2y}{12}, \quad (x, y) \in S,$$

missä  $S = \{(0, 1), (0, 2), (1, 0), (1, 1), (2, 0)\}$ . Määritellään kokonaislukuarvoinen satunnaismuuttuja

$$Z = h(X, Y) = XY.$$

Silloin  $Z$ :n arvojen joukko on  $S_z = \{0, 1\}$ , ja vastaavasti

$$\begin{aligned} A_1 &= \{ (x, y) \mid xy = 1 \} = \{(1, 1)\}, \\ A_0 &= \{ (x, y) \mid xy = 0 \} = \{(0, 1), (0, 2), (1, 0), (2, 0)\}. \end{aligned}$$

Nyt siis

$$\begin{aligned} P(Z = 0) &= P((X, Y) \in A_0) = \frac{3}{4}, \\ P(Z = 1) &= P((X, Y) \in A_1) = \frac{1}{4}, \end{aligned}$$

joten  $Z \sim \text{Ber}(\frac{1}{4})$ . □

### 4.6.3 Ehdollinen odotusarvo

Ehdollinen odotusarvo esiteltiin jo aluvussa 3.3.2 (identiteetti 3.3.7). Aluvussa 4.6.1 ehdollinen odotusarvo luonnehdittiin ehdollisen jakauman odotusarvona (ks. (4.6.2) ja (4.6.4)). Määritelmän mukaan

$$E(X \mid Y = y) = \sum_x x f(x \mid y) = \sum_x x \frac{f(x, y)}{f_Y(y)}.$$

Huomaa, että  $E(X | Y = y)$  on  $y$ :n funktio, eli  $E(X | Y = y) = h(y)$ . Merkitään  $E(X | Y) = h(Y)$ , missä siis  $E(X | Y)$  on satunnaismuuttuja, joka saa arvoja  $E(X | Y = y)$ ,  $y \in S_Y$ . Voimme nyt laskea satunnaismuuttujan  $E(X | Y)$  odotusarvon, joka on  $E(X)$ . Monissa sovelluksissa odotusarvon laskeminen on luontevinta ehdollistamisen kautta.

**Lause 4.12** *Olko  $X$  ja  $Y$  mitkä tahansa kaksi satunnaismuuttujaa, joilla on odotusarvo. Silloin  $E[E(X | Y)] = E(X)$ .*

**Todistus.** Odotusarvon määritelmän mukaan

$$\begin{aligned} E[E(X | Y)] &= \sum_y E(X | Y = y) f_Y(y) \\ &= \sum_y \sum_x x \frac{f(x, y)}{f_Y(y)} f_Y(y) \\ &= \sum_y \sum_x x f(x, y) \\ &= \sum_x x \sum_y f(x, y) \\ &= \sum_x x f_X(x) = E(X), \end{aligned}$$

missä  $\sum_y f(x, y) = f_X(x)$  on  $X$ :n reuna-jakauma. □

Ehdollinen varianssi

$$\begin{aligned} \text{Var}(X | Y = y) &= E[(X - E(X | Y = y))^2 | Y = y] \\ &= E(X^2 | Y = y) - [E(X | Y = y)]^2 \end{aligned}$$

määriteltiin alaluvussa 4.6.1 (ks. identiteetti (4.6.4)). Ehdollinen varianssi  $\text{Var}(X | Y)$  on satunnaismuuttuja, joka saa arvoja  $\text{Var}(X | Y = y)$ . Koska

$$\text{Var}(X | Y) = E(X^2 | Y) - [E(X | Y)]^2,$$

niin

$$\begin{aligned} E[\text{Var}(X | Y)] &= E[E(X^2 | Y)] - E[E(X | Y)]^2 \\ (4.6.5) \quad &= E(X^2) - E[E(X | Y)]^2. \end{aligned}$$

Lauseen 4.12 mukaan  $E[E(X | Y)] = E(X)$  ja  $E[E(X^2 | Y)] = E(X^2)$ , joten

$$(4.6.6) \quad \text{Var}[E(X | Y)] = E[E(X | Y)]^2 - [E(X)]^2.$$

Laskemalla yhtälöt (4.6.5) ja (4.6.6) puolittain yhteen, saadaan seuraavassa lauseessa esitettävä tulos.

**Lause 4.13** *Mille tahansa satunnaismuuttujille  $X$  ja  $Y$  pitää paikkansa identiteetti*

$$\text{Var}(X) = E[\text{Var}(X | Y)] + \text{Var}[E(X | Y)],$$

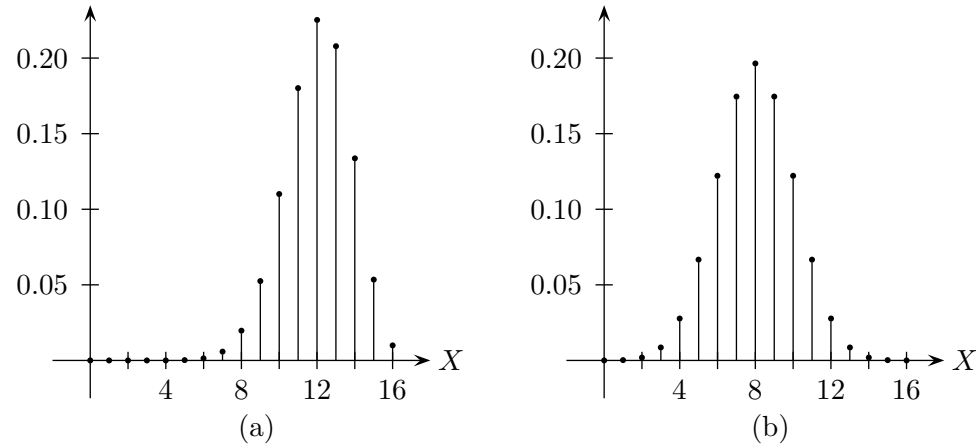
*jos odotusarvot ovat olemassa.*

### 4.6.4 Symmetrinen jakauma

Symmetriaan perustuvaa argumentointia voidaan usein hyödyntää todennäköisyyksien laskemisessa.

**Symmetria pisteen suhteen.** Jos  $P(X = b + x) = P(X = b - x)$  kaikilla  $x$ , niin  $X$ :n jakauma on *symmetrinen pisteen  $b$  suhteen*. Satunnaismuuttuja  $X$  on symmetrinen  $b$ :n suhteen jos ja vain jos  $X - b$  on symmetrinen origon suhteen. Silloin

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



**Kuvio 4.3.** Binomijakauman kuvaajat, kun (a)  $X \sim \text{Bin}(16, 0.75)$   
(b)  $X \sim \text{Bin}(16, 0.50)$ .

Esimerkiksi binomijakauma  $\text{Bin}(16, 0.50)$  on symmetrinen pisteen 8 suhteen, mutta binomijakauma  $\text{Bin}(16, 0.75)$  ei ole symmetrinen (Kuvio 4.3). Binomijakaumassa  $\text{Bin}(16, 0.50)$  on

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

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

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

**Esimerkki 4.16** Hatussa on 3 korttia, jotka on numeroitu yhdestä kolmeen. Valitaan hatusta peräkkäin satunnaisesti palauttamatta 2 korttia. Olkoon  $X$  ensiksi valitun kortin numero ja  $Y$  toisen kortin numero. Selvästikin

$$P(X = i) = f_X(i) = \frac{1}{3}, \quad i = 1, 2, 3.$$

On helppo havaita, että toisen valinnan tulos  $Y$  riippuu 1. valinnan tulokses-

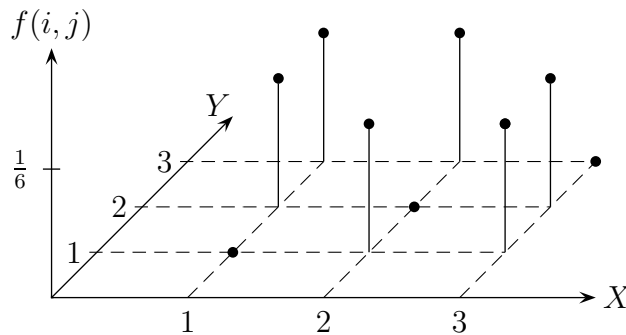
ta:

$$\begin{aligned} P(Y = 1 | X = 1) &= 0, & P(Y = i | X = 1) &= \frac{1}{2}, & i &= 2, 3; \\ P(Y = 2 | X = 2) &= 0, & P(Y = i | X = 2) &= \frac{1}{2}, & i &= 1, 3; \\ P(Y = 3 | X = 3) &= 0, & P(Y = i | X = 3) &= \frac{1}{2}, & i &= 1, 2. \end{aligned}$$

Koska  $P(X = i, Y = j) = P(X = i)P(Y = j | X = i)$ , niin  $X$ :n ja  $Y$ :n yhteisjakauman todennäköisyysfunktio on

$$(4.6.7) \quad f(i, j) = f_X(i)f_2(j | i) = \begin{cases} \frac{1}{6}, & \text{kun } i \neq j; \\ 0, & \text{kun } i = j; 1 \leq i \leq 3, 1 \leq j \leq 3. \end{cases}$$

Satunnaisvektorin  $(X, Y)$  arvojoukko  $S = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1)\}$ ,



**Kuvio 4.4.** Satunnaisvektorin  $(X, Y)$  yhteisjakauman todennäköisyysfunktio  $f(i, j)$ , kun  $X$  on 1. valinta ja  $Y$  on 2. valinta palauttamatta joukosta  $\{1, 2, 3\}$ .

$(3, 2)\}$ , sillä  $P(X = i, Y = j) > 0$  kaikilla  $(i, j) \in S$  ja  $P(X = i, Y = j) = 0$ , jos  $(i, j) \notin S$ .  $\square$

Jakauma (4.6.7) on esimerkki symmetrisestä 2-ulotteisesta jakaumasta. Diskreetin satunnaisvektorin  $(X, Y)$  jakauma on symmetrinen, jos sen todennäköisyysfunktio  $f(x, y)$  on symmetrinen funktio. Se tarkoittaa sitä, että

$$f(x, y) = f(y, x) \quad \text{kaikilla } (x, y) \in S,$$

missä  $S$  on  $(X, Y)$ :n arvoalue.

### 4.6.5 Kaksiulotteinen Bernoullin jakauma

Bernoullin jakaumaa noudattava satunnaismuuttuja  $X \sim \text{Ber}(p)$  on eräs yksinkertaisimpia ajateltavissa olevia satunnaismuuttujia. Sen todennäköisyysfunktio on

$$f_X(x) = p^x(1 - p)^{1-x}, \quad x \in \{0, 1\},$$



kun  $0 \leq p \leq 1$ . Bernoullin jakauma on binomijakauman erikoistapaus siten, että  $X \sim \text{Bin}(1, p)$ .

Kaksiulotteista Bernoullin jakaumaa noudattava satunnaismuuttuja  $(X, Y)$  voi saada arvot  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$ ,  $(1, 1)$ . Sen todennäköisyysfunktio on

$$(4.6.8) \quad f(x, y) = p_{xy}, \quad x \in \{0, 1\} \quad \text{ja} \quad y \in \{0, 1\},$$

missä  $p_{00} + p_{01} + p_{10} + p_{11} = 1$ . Todennäköisyysfunktio voidaan esittää myös muodossa

$$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy},$$

kun  $x \in \{0, 1\}$  ja  $y \in \{0, 1\}$ ; muualla  $f(x, y) = 0$ . Todennäköisyydet  $P(X = x, Y = y) = p_{xy}$  on esitetty Taulukossa 4.1

**Taulukko 4.1.** Kaksiulotteisen Bernoullin jakauman todennäköisyysfunktio.

$f(x, y)$	$y = 0$	$y = 1$	$f_X(x)$
$x = 0$	$p_{00}$	$p_{01}$	$1 - p_1$
$x = 1$	$p_{10}$	$p_{11}$	$p_1$
$f_Y(y)$	$1 - p_2$	$p_2$	$1$

”Reunatodennäköisyydet” määritellään  $p_{00} + p_{01} = p_1$  ja  $p_{00} + p_{10} = p_2$ . On helppo havaita, että  $X \sim \text{Ber}(p_1)$  ja  $Y \sim \text{Ber}(p_2)$ . Näiden reunajakaumien todennäköisyysfunktiot ovat siis

$$f_X(x) = p_1^x (1 - p_1)^{1-x}, \quad x \in \{0, 1\}$$

ja

$$f_Y(y) = p_2^y (1 - p_2)^{1-y}, \quad y \in \{0, 1\}.$$

Nyt esimerkiksi  $Y$ :n ehdollinen todennäköisyysfunktio ehdolla  $X = 1$  on

$$(4.6.9) \quad f_2(y | 1) = \frac{p_{1y}}{p_1}, \quad y \in \{0, 1\}$$

kun  $p_1 > 0$ . Satunnaismuuttujan  $Y$  ehdollinen jakauma ehdolla  $X = 1$  on siis  $\text{Ber}(p_{11}/p_1)$ . Satunnaismuuttujat  $X$  ja  $Y$  ovat riippumattomat täsmälleen silloin, kun  $P(X = x, Y = y) = P(X = x)P(Y = y)$  kaikilla  $x \in \{0, 1\}$  ja  $y \in \{0, 1\}$ .

Koska  $p_{00} + p_{01} + p_{10} + p_{11} = 1$ , niin kaksiulotteinen Bernoullin jakauma voidaan luonnehtia kolmella parametrilla. Jakauman kolme ”luonnollista” parametria ovat

$$\begin{aligned} p_1 &= E(X) = P(X = 1), \\ p_2 &= E(Y) = P(Y = 1), \\ p_{11} &= E(XY) = P(X = 1, Y = 1). \end{aligned}$$

Kun  $(X, Y)$  noudattaa kaksiulotteista Bernoullin jakaumaa parametrein  $p_1$ ,  $p_2$  ja  $p_{11}$ , niin merkitään  $(X, Y) \sim \text{Ber}(p_1, p_2, p_{11})$ .

## 4.7 Satunnaismuuttujien funktion odotusarvo

Olkoot  $X$  ja  $Y$  diskreetit satunnaismuuttujat, joiden yhteinen todennäköisyysfunktio  $f(x, y)$  on on määritelty arvoavaruudessa  $S$ . Olkoon  $h(X, Y)$  satunnaismuuttujien  $X$  ja  $Y$  reaaliarvoinen funktio. Silloin

$$E[h(X, Y)] = \sum_{(x,y) \in S} h(x, y) f(x, y)$$

on satunnaismuuttujan  $h(X, Y)$  odotusarvo, mikäli summa on olemassa.

Huomaa, että odotusarvon  $E[h(X, Y)]$  olemassaolo tarkoittaa sitä, että summa

$$\sum_{(x,y) \in S} h(x, y) f(x, y)$$

suppenee itseisesti eli summa

$$\sum_{(x,y) \in S} |h(x, y)| f(x, y)$$

suppenee ja on äärellinen. Tästä seuraa, että  $E[h(X, Y)]$  on olemassa. Funktio  $V = h(X, Y)$  on satunnaismuuttuja, jonka todennäköisyysfunktio  $g(v)$  on määritelty arvoavaruudessa  $S_v = \{v \mid v = h(x, y), (x, y) \in S\}$ . Silloin

$$E[h(X, Y)] = E(V) = \sum_{v \in S_v} v g(v).$$

### 4.7.1 Momentit

Monilla odotusarvoilla on omat nimensä, koska niillä on tärkeä rooli jakaumateoriassa. Olkoot  $X_1$  ja  $X_2$  diskreetit satunnaismuuttujat, joiden yhteisjakauman todennäköisyysfunktio  $f(x_1, x_2)$  on on määritelty arvoavaruudessa  $S$ . Olkoon  $h(X_1, X_2)$  satunnaismuuttujien  $X_1$  ja  $X_2$  reaaliarvoinen funktio. Määritellään esimerkiksi seuraavat odotusarvot:

1. Jos  $h(X_1, X_2) = X_i$ , niin

$$E[h(X_1, X_2)] = E(X_i) = \mu_i$$

on  $X_i$ :n odotusarvo,  $i = 1, 2$ .

2. Jos  $h(X_1, X_2) = (X_i - \mu_i)^2$ , niin

$$E[h(X_1, X_2)] = E[(X_i - \mu_i)^2] = \sigma_i^2$$

on  $X_i$ :n varianssi,  $i = 1, 2$ .

3. Jos  $h(X_1, X_2) = (X_1 - \mu_1)(X_2 - \mu_2)$ , niin

$$E[h(X_1, X_2)] = E[(X_1 - \mu_1)(X_2 - \mu_2)] = \sigma_{12}$$

on  $X_1$ :n ja  $X_2$ :n *kovarianssi*.

Odotusarvo  $\mu_i$  ja varianssi  $\sigma_i^2$  voidaan laskea joko yhteisjakauman todennäköisyysfunktion  $f(x_1, x_2)$  tai reunajakauman todennäköisyysfunktion  $f_i(x_i)$  avulla.

Vastaavalla tavalla voidaan määritellä kaikkien kertalukujen momentit: Olkoon  $r$  positiivinen kokonaisluku.

1. Jos  $h(X_1, X_2) = X_i^r$ , niin

$$E[h(X_1, X_2)] = E(X_i^r)$$

on  $X_i$ :n  $r$ . *momentti*,  $i = 1, 2$ .

2. Jos  $h(X_1, X_2) = (X_i - \mu_i)^r$ , niin

$$E[h(X_1, X_2)] = E[(X_i - \mu_i)^r]$$

on  $X_i$ :n  $r$ . *keskusmomentti*,  $i = 1, 2$ .

3. Jos  $h(X_1, X_2) = X_1^r X_2^s$ , niin

$$E[h(X_1, X_2)] = E(X_1^r X_2^s)$$

on  $X_1$ :n ja  $X_2$ :n kertalukua  $r + s$  oleva *yhteismomentti*.

Esimerkiksi kovarianssin laskemisessa tarvitaan  $X_1$ :n ja  $X_2$ :n yhteismomentti  $E(X_1 X_2)$ .

## 4.7.2 Satunnaisvektorin momenttifunktio

Satunnaisvektorin  $(X, Y)$  yhteisjakauman momenttifunktio on

$$\begin{aligned} M(t, s) &= E[\exp(tX + sY)] \\ &= \sum_{x_i \in S_X} \sum_{y_j \in S_Y} \exp(tx_i + sy_j) f(x_i, y_j), \end{aligned}$$

mikäli odotusarvo on olemassa nollan ympäristössä. Silloin on siis olemassa sellainen positiiviluku  $a > 0$ , että odotusarvo  $E[\exp(tX + sY)]$  on olemassa kaikilla  $(t, s) \in \{(t, s) \mid t^2 + s^2 < a\}$  jollain  $a > 0$ . Edellä on käytetty merkintää  $\exp(tX + sY) = e^{tX + sY}$ .

## 4.8 Riippumattomat satunnaismuuttujat

Riippumattomuuden määritelmän mukaan tapahtumat  $\{X = x\}$  ja  $\{Y = y\}$  ovat riippumattomat jos ja vain jos  $P(X = x, Y = y) = P(X = x)P(Y = y)$  eli

$$(4.8.1) \quad f(x, y) = f_X(x)f_Y(y),$$

missä  $f(x, y)$  on  $X$ :n ja  $Y$ :n yhteisjakauman todennäköisyysfunktio,  $f_X(x)$  on  $X$ :n ja  $f_Y(y)$   $Y$ :n todennäköisyysfunktio. Satunnaismuuttujat  $X$  ja  $Y$  ovat riippumattomat jos ja vain jos yhtäsuuruus (4.8.1) pitää paikkansa kaikilla  $x \in S_X$  ja  $y \in S_Y$ , missä  $S_X$  on  $X$ :n ja  $S_Y$  on  $Y$ :n arvojoukko. Voidaan helposti osoittaa, että  $X$  ja  $Y$  ovat riippumattomat, jos ja vain jos

$$(4.8.2) \quad F(x, y) = F_X(x)F_Y(y)$$

kaikilla  $x \in S_X$  ja  $y \in S_Y$ , missä  $F_X(x)$  on  $X$ :n ja  $F_Y(y)$   $Y$ :n kertymäfunktio (reunakertymäfunktio).

Satunnaismuuttujien  $X$  ja  $Y$  ovat riippumattomuus voidaan luonnehtia myös ehdollisten jakaumien avulla. Jos Määritelmässä 4.7

$$f(y | x) = f_Y(y)$$

kaikilla  $x \in S_X$  ja  $y \in S_Y$ , kun  $f_X(x) \neq 0$ , niin  $X$  ja  $Y$  ovat riippumattomat. Tämä tarkoittaa sitä, että tieto  $X$ :n arvosta ei vaikuta  $Y$ :n todennäköisyyteen. Vastaavasti pitää paikkansa, että  $X$  ja  $Y$  ovat riippumattomat jos ja vain jos  $X$ :n ehdollinen todennäköisyysfunktio ehdolla  $Y = y$  ei riipu  $y$ :stä.

Olkoot  $Y_1, Y_2, \dots, Y_n$  jossain otosavaruudessa määritellyt diskreetit satunnaismuuttujat. Muuttujien  $Y_1, Y_2, \dots, Y_n$  yhteisjakauman todennäköisyysfunktio on

$$f(y_1, y_2, \dots, y_n) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n).$$

Muuttujat  $Y_1, Y_2, \dots, Y_n$  ovat riippumattomat, jos

$$(4.8.3) \quad f(y_1, y_2, \dots, y_n) = f_1(y_1)f_2(y_2) \cdots f_n(y_n)$$

kaikilla  $y_i \in S_i$ ,  $i = 1, 2, \dots, n$ , missä  $f_i(y_i)$  on  $Y_i$ :n todennäköisyysfunktio ja  $S_i$  on  $Y_i$ :n arvoavaruus.

### 4.8.1 Riippumattomat kokeet

Olkoot  $\mathcal{E}_1$  ja  $\mathcal{E}_2$  riippumattomat satunnaiskokeet. Oletetaan, että satunnaismuuttujan  $X$  arvo määräytyy vain satunnaiskokeen  $\mathcal{E}_1$  tuloksen ja  $Y$ :n arvo vain satunnaiskokeen  $\mathcal{E}_2$  tuloksen perusteella. Silloin tapahtumat  $\{X = x\}$  ja  $\{Y = y\}$  määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat (katso alaluku 3.6 ja määritelmä (3.6.1)). Siksi tapahtumat  $\{X = x\}$  ja  $\{Y = y\}$

ovat riippumattomat. Koska riippumattomuus pätee kaikilla mahdollisilla  $x$ :n ja  $y$ :n arvoilla, niin  $X$  ja  $Y$  ovat riippumattomat. *Jos satunnaismuuttujien arvot määräytyvät eri satunnaiskokeista, jotka ovat riippumattomat, niin satunnaismuuttujat ovat riippumattomat.*

Tehdään esimerkiksi Bernoullin toistokoe, jossa on  $r + s$  toistoa ja onnistumistodennäköisyys  $p$ . Olkoon  $X$  onnistumisten lukumäärä  $r$ :ssä ensimmäisessä kokeessa ja  $Y$  onnistumisten lukumäärä  $s$ :ssä viimeisessä kokeessa. Koska  $X$  ja  $Y$  riippuvat eri kokeista, jotka ovat riippumattomat, niin  $X$  ja  $Y$  ovat riippumattomat. Silloin (4.8.1):n mukaan

$$f(x, y) = f_X(x)f_Y(y) = \binom{r}{x} \binom{s}{y} p^{x+y} (1-p)^{r+s-x-y},$$

missä  $x = 0, 1, \dots, r$  ja  $y = 0, 1, \dots, s$ .

### 4.8.2 Samoin jakautuneet riippumattomat (SJR) satunnaismuuttujat

Riippumattomia satunnaismuuttujia  $Y_1, Y_2, \dots, Y_n$ , joista jokainen noudattaa samaa jakaumaa, sanotaan samoin jakautuneiksi riippumattomiksi (sjr) satunnaismuuttujiksi. Silloin puhutaan usein lyhyesti sjr satunnaismuuttujista. Vastaava englanninkielinen termi on iid (independent, identically distributed).

Jos esimerkiksi  $Y_i \sim \text{Poi}(\lambda)$ ,  $i = 1, 2, \dots, n$ , niin

$$f_1(y) = f_2(y) = \dots = f_n(y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

Silloin (4.8.3):n nojalla

$$\begin{aligned} f(y_1, y_2, \dots, y_n) &= \prod_{i=1}^n f_i(y_i) = \prod_{i=1}^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \\ &= \frac{1}{y_1! y_2! \dots y_n!} \lambda^y e^{-n\lambda}, \end{aligned}$$

missä  $y = \sum_{i=1}^n y_i$ .

### 4.8.3 Riippumattomien satunnaismuuttujien funktio

Olkoot  $X$  ja  $Y$  riippumattomat satunnaismuuttujat. Silloin

$$f(x, y) = f_X(x)f_Y(y) \quad \text{kaikilla } x, y \in S,$$

missä  $S$  on satunnaisvektorin  $(X, Y)$  arvoavaruus. Määritellään nyt satunnaismuuttujat  $U = g(X)$  ja  $V = h(Y)$ , missä  $g(\cdot)$  riippuu vain  $X$ :stä ja  $h(\cdot)$  vain  $Y$ :stä. Silloin Lauseen 3.6 mukaan  $U$  ja  $V$  ovat riippumattomat.

Väite todistettiin tarkastelemalla  $U$ :n ja  $V$ :n mielivaltaisten arvojen  $u$  ja  $v$  todennäköisyyttä. Olkoon  $A_u = \{x \in S_X \mid g(x) = u\}$  ja  $B_v = \{y \in S_Y \mid g(y) = v\}$ . Koska kaikilla  $U$ :n ja  $V$ :n arvoilla  $u$  ja  $v$

$$\begin{aligned}
 (4.8.4) \quad & P(U = u, V = v) \\
 &= P(X \in A_u, Y \in B_v) \\
 &= P(X \in A_u) P(Y \in B_v) \quad (X \text{ ja } Y \text{ riippumattomat}) \\
 &= P(U = u) P(V = v),
 \end{aligned}$$

niin  $U$  ja  $V$  ovat riippumattomat. Kun sijoitetaan identiteettiin (4.8.4)

$$\begin{aligned}
 P(U = u, V = v) &= \sum_{x \in A_u} \sum_{y \in B_v} f(x, y), \\
 P(U = u) &= \sum_{x \in A_u} f_X(x) \quad \text{ja} \quad P(V = v) = \sum_{y \in B_v} f_Y(y),
 \end{aligned}$$

niin saadaan

$$\sum_{x \in A_u} \sum_{y \in B_v} f(x, y) = \left( \sum_{x \in A_u} f_X(x) \right) \left( \sum_{y \in B_v} f_Y(y) \right).$$

## 4.9 Multinomijakauma ja moniulotteinen hypergeometrinen jakauma

Binomijakauma ja multinomijakauma ovat keskeisen tärkeitä tilastollisissa sovelluksissa, koska niitä tarvitaan esimerkiksi riippumattomien koetoistojen tulosten frekvenssijakaumien käsittelyssä. Alaluvussa 3.6 esitettiin, miten binomijakauma saadaan Bernoullin kokeiden avulla. Kun toistetaan kokeita, joissa on useampia kuin kaksi tulostulovaihtoehtoa, tulosten frekvenssijakauma voidaan kuvata multinomijakauman avulla. Laajennetaan ensin binomijakauma *trinomijakaumaksi*.

Tarkastellaan koetta, jossa on kolme toisensa poissulkevaa tulostulovaihtoehtoa. Esimerkiksi tuotantoprosessissa syntyvä tuote luokitellaan yhteen ja vain yhteen seuraavista kategorioista: ensiluokkainen (1), sekunda (2) tai viallinen (3). Olkoot ensiluokkaisen, sekundan ja viallisen todennäköisyydet vastaavasti  $p_1$ ,  $p_2$  ja  $p_3 = 1 - p_1 - p_2$ . Valmistetaan  $n$  tuotetta. Olkoon  $X_1 =$  ensiluokkaisten lukumäärä,  $X_2 =$  sekundatuotteiden lukumäärä ja  $X_3 = n - X_1 - X_2 =$  viallisten lukumäärä tuote-erässä. Jos  $x_1$  ja  $x_2$  ovat sellaiset epänegatiiviset kokonaisluvut, että  $x_1 + x_2 \leq n$ , niin todennäköisyys saada  $x_1$  ensiluokkaista,  $x_2$  sekunda ja  $n - x_1 - x_2$  viallista jossain annetussa järjestyksessä on

$$p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2}.$$

Sellaisia  $n$ :n tuotteen järjestyksiä, joissa on  $x_1$  ensiluokkaista,  $x_2$  sekunda ja  $n - x_1 - x_2$  viallista, on

$$\binom{n}{x_1} \binom{n - x_1}{x_2} = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!}$$

kappaletta. Siksi *trinomijakauman todennäköisyysfunktio* on

$$(4.9.1) \quad f(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2},$$

missä  $f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$ . Kun  $(X_1, X_2)$  noudattaa trinomijakaumaa parametrein  $n$ ,  $p_1$  ja  $p_2$ , merkitään

$$(X_1, X_2) \sim \text{Tri}(n, p_1, p_2).$$

On helppo todeta, että  $X_1 \sim \text{Bin}(n, p_1)$  ( $X_1$ :n reunajakauma) ja  $X_2 \sim \text{Bin}(n, p_2)$ .

Multinomijakauma voidaan johtaa samalla periaatteella kuin trinomijakauma. Toistetaan  $n$  kertaa koe, jossa on  $k$  toisensa poissulkevaa tulosvaihtoehtoa. Merkitään tulosvaihtoehtoja  $1, 2, \dots, k$  ja olkoon  $p_i =$  tuloksen  $i$  todennäköisyys ja  $X_i$  on tuloksen  $i$  lukumäärä  $n$ :n kokeen sarjassa. Silloin  $k$ -ulotteisen satunnaisvektorin  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  arvoalue on

$$S = \{ (x_1, x_2, \dots, x_k) \mid 0 \leq x_i \leq n, x_1 + x_2 + \dots + x_k = n \}.$$

$X_i$ :t ovat siis epänegatiivisia kokonaislukuarvoisia satunnaisuuttujia, joiden summa on  $n$ . Satunnaisvektori  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  noudattaa *k-ulotteista multinomijakaumaa* parametrein  $n$  ja  $\mathbf{p} = (p_1, p_2, \dots, p_k)$ , jota merkitään  $\text{Mult}(n, \mathbf{p})$ . Multinomijakauman todennäköisyysfunktio on

$$(4.9.2) \quad f(x_1, x_2, \dots, x_k) = \binom{n}{x_1 \ x_2 \ \dots \ x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k},$$

missä  $p_1 + p_2 + \dots + p_k = 1$  ja  $\binom{n}{x_1 \ x_2 \ \dots \ x_k} = \frac{n!}{x_1! x_2! \dots x_k!}$ . Multinomialauseen 2.9 nojalla voidaan helposti osoittaa, että todennäköisyyksien (4.9.2) summa yli arvoalueen  $S$  on 1, joten kyseinen funktio on todellakin todennäköisyysfunktio. Multinomijakaumassa jokaisen  $X_i$ :n reunajakauma on binomijakauma, eli  $X_i \sim \text{Bin}(n, p_i)$ ,  $i = 1, 2, \dots, k$ . Esitetään nyt multinomijakaumaa koskevat perustulokset lauseen muodossa.

**Lause 4.14** 1. *Funktio (4.9.2) on multinomijakauman todennäköisyysfunktio kaikilla positiivisilla kokonaisluvulla  $n$  ja kaikilla sellaisilla  $p_1, \dots, p_k$ , että  $0 \leq p_i \leq 1$  ja  $p_1 + \dots + p_k = 1$ .*

2. *Jos  $\mathbf{X} \sim \text{Mult}(n, \mathbf{p})$ , niin*

$$X_i \sim \text{Bin}(n, p_i) \quad \text{ja} \quad (X_i, X_j) \sim \text{Tri}(n, p_i, p_j),$$

3.

$$E(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i), \quad \text{Cov}(X_i, X_j) = -np_i p_j, \\ M(t) = E[\exp(t_1 X_1 + \dots + t_k X_k)] = (p_1 e^{t_1} + \dots + p_k e^{t_k})^n.$$

Multinomijakauma liittyy otantaan palauttaen. Olkoon uurnassa erivärisiä palloja yhteensä  $N$ , värien lukumäärä on  $k$  ja väriä  $i$  olevia palloja on  $N_i$  kappaletta ( $i = 1, 2, \dots, k$ ) ja  $N_1 + N_2 + \dots + N_k = N$ . Otannassa palauttaen todennäköisyys  $p_i$  saada väri  $i$  on  $N_i/N$  jokaisessa nostossa. Valitaan uurnasta  $n$  palloa palauttaen ja olkoon  $X_i$  väriä  $i$  olevien pallojen lukumäärä otoksessa. Silloin satunnaismuuttujien  $X_1, X_2, \dots, X_k$  yhteisjakauma on multinomijakauma.

Otannassa palauttamatta uurnan sisältö muuttuu ja siten myös valintatodennäköisyydet muuttuvat valintaprosessin aikana. Satunnaismuuttujien  $X_1, X_2, \dots, X_k$  yhteisjakauman johtamiseksi meidän on yleistettävä aluvuossa 2.6.1 esitetty hypergeometrinen jakauma.

Olkoon  $x_i$  väriä  $i$  ( $1 \leq i \leq k$ ) olevien pallojen lukumäärä otannassa palauttamatta. Millä todennäköisyydellä saadaan otos, jossa eri väriä olevien lukumäärät ovat  $(x_1, x_2, \dots, x_k)$ ? Koska uurnassa on  $N_i$  kappaletta väriä  $i$ , niin  $0 \leq x_i \leq N_i$ . Otoskoko on  $n$  ja  $n = x_1 + x_2 + \dots + x_k$ . Nyt väriä 1 olevat  $x_1$  palloa voidaan valita  $\binom{N_1}{x_1}$  tavalla, väriä 2 olevat  $\binom{N_2}{x_2}$  tavalla ja lopulta väriä  $k$  olevat pallot  $\binom{N_k}{x_k}$  tavalla. Suotuisten otosten lukumäärä on tuloperiaatteen nojalla

$$\binom{N_1}{x_1} \binom{N_2}{x_2} \dots \binom{N_k}{x_k}.$$

Koska kaikkien mahdollisten  $n$ :n kokoisten otosten lukumäärä on  $\binom{N}{n}$ , niin todennäköisyys saada lukumäärät  $(x_1, x_2, \dots, x_k)$  erivärisiä palloja on

$$(4.9.3) \quad f(x_1, x_2, \dots, x_k) = \frac{\binom{N_1}{x_1} \binom{N_2}{x_2} \dots \binom{N_k}{x_k}}{\binom{N}{n}},$$

missä  $x_1 + x_2 + \dots + x_k = n$ . Tämä on *moniulotteisen hypergeometrisen jakauman* todennäköisyysfunktio.

## Diskreetit jakaumat: Yhteenvedo

<b>Bernoulli</b>	$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$
Ber( $p$ )	$E(X) = p, \quad \text{Var}(X) = p(1-p)$
	$M(t) = 1 - p + pe^t$

<b>Binomi</b>	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$
Bin( $n, p$ )	$E(X) = np, \quad \text{Var}(X) = np(1-p)$
	$M(t) = (1 - p + pe^t)^n$



**Negatiivinen binomi**  
NBin( $r, p$ )

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots$$

$$E(X) = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

$$M(t) = \left[ \frac{pe^t}{1 - (1-p)e^t} \right]^r, \quad t < -\log(1-p)$$

**Geometrisen**  
Geo( $p$ )

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \dots$$

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}$$

$$M(t) = \frac{pe^t}{1 - (1-p)e^t}, \quad t < -\log(1-p)$$

**Hypergeometrisen**  
HGeo( $n, N, p$ )

$$f(x) = \frac{\binom{Np}{x} \binom{N-Np}{n-x}}{\binom{N}{n}}, \quad \begin{array}{l} X \leq pN \text{ ja} \\ n - X \leq N - Np \end{array}$$

$$E(X) = np, \quad \text{Var}(X) = \frac{N-n}{N-1} np(1-p),$$

**Negatiivinen hypergeometrisen**  
NHGeo( $r, N, p$ )

$$f(x) = \binom{x-1}{r-1} \frac{\binom{N-x}{Np-r}}{\binom{N}{Np}}, \quad x = r, r+1, \dots, N$$

$$E(X) = r \cdot \frac{N+1}{Np+1}$$

$$\text{Var}(X) = \frac{r(1-p)N(N+1)(Np+1-r)}{(Np+1)^2(Np+2)}$$

**Poisson**  
Poi( $\lambda$ )

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \dots$$

$$E(X) = \lambda, \quad \text{Var}(X) = \lambda$$

$$M(t) = \exp(\lambda e^t - \lambda), \quad -\infty < t < \infty$$

**Poissonin prosessi.** Laskuri prosessi  $\{N(t), t \geq 0\}$ , jonka intensiteetti  $\lambda$ .

1.  $N(0) = 0$ .
2. Prosessin lisäykset ovat riippumattomat.
3. Tapahtumien lukumäärä jokaisella  $t$ :n pituisella välillä noudattaa Poissonin jakaumaa, jonka odotusarvo on  $\lambda t$ :

$$P[N(t+s) - N(s) = x] = e^{-\lambda t} \frac{(\lambda t)^x}{x!}, \quad x = 0, 1, \dots$$

kaikilla  $s, t \geq 0$ .

## Kaksiulotteiset jakaumat

### Satunnaisvektorin $(X, Y)$ yhteisjakauma.

Todennäköisyysfunktio  $f(x, y)$  toteuttaa ehdot

1.  $0 \leq f(x, y) \leq 1$ , kaikilla  $(x, y) \in S$  ja
2.  $\sum_{(x,y) \in S} f(x, y) = 1$ ,

missä  $S$  on satunnaisvektorin  $(X, Y)$  arvojoukko.

Reunajakaumien todennäköisyysfunktiot:

$$f_X(x) = \sum_{y \in S_Y} f(x, y), \quad x \in S_X; \quad f_Y(y) = \sum_{x \in S_X} f(x, y), \quad y \in S_Y.$$

Ehdolliset todennäköisyysfunktiot:

$$f_1(x | y) = \frac{f(x, y)}{f_Y(y)} \quad \text{ja} \quad f_2(y | x) = \frac{f(x, y)}{f_X(x)}, \quad (x, y) \in S.$$

### Bernoulli

Ber2( $p_1, p_2, p_{11}$ )

$$f(x, y) = p_{00}^{(1-x)(1-y)} p_{01}^{(1-x)y} p_{10}^{x(1-y)} p_{11}^{xy}, \quad \text{missä}$$

$$p_{00} + p_{01} + p_{10} + p_{11} = 1, \quad x \in \{0, 1\} \quad \text{ja} \quad y \in \{0, 1\}$$

$$E(X) = p_1 = P(X = 1), \quad E(Y) = p_2 = P(Y = 1)$$

$$E(XY) = p_{11} = P(X = 1, Y = 1)$$

### Multinomi

Mult( $n, \mathbf{p}$ )

$$\mathbf{p} = (p_1, p_2, \dots, p_k), \quad \text{missä}$$

$$0 \leq p_i \leq 1 \quad \text{ja} \quad p_1 + p_2 + \dots + p_k = 1$$

$$E(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i)$$

$$\text{Cov}(X_i, X_j) = -np_i p_j$$

$$X_i \sim \text{Bin}(n, p_i), \quad (X_i, X_j) \sim \text{Mult}(n; p_i, p_j, 1 - p_i - p_j)$$

$$M(t) = (p_1 e^{t_1} + \dots + p_k e^{t_k})^n$$

**Hypergeometrinen**  $f(x_1, x_2, \dots, x_k) = \frac{\binom{N_1}{x_1} \binom{N_2}{x_2} \dots \binom{N_k}{x_k}}{\binom{N}{n}}, \quad \text{missä}$

$$N_1 + N_2 + \dots + N_k = N \quad \text{ja} \quad x_1 + x_2 + \dots + x_k = n$$

## Harjoituksia

1. Olkoon  $X$  satunnaismuuttuja, jonka arvojoukko on  $S_X = \{x_1, x_2\}$  ja todennäköisyysfunktio  $P(X = x_1) = p$ ,  $P(X = x_2) = 1 - p$ .

(a) Laske  $E(X^r)$ ,  $r = 1, 2$  ja

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

- (c) Määritä  $X$ :n momenttifunktio.
2. Olkoon  $X \sim \text{Ber}(p)$  ja  $Y$  sellainen satunnaismuuttuja, jonka arvojoukko on  $S_Y = \{y_1, y_2\}$  ja todennäköisyysfunktio  $P(Y = y_1) = p$ ,  $P(Y = y_2) = 1 - p$ . Lausu  $Y$  satunnaismuuttujan  $X$  avulla.
3. Heitetään lanttia  $n$  kertaa ( $n$  riippumattonta 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  $n = 200$ ?
- (b) Laske  $\text{Var}(X)$ .
- (c) Mitä on toistosten RRRR lukumäärän odotusarvo?
- (Vihje: Katso Esimerkki 4.2.)
4. Jos  $X$  noudattaa binomijakaumaa, jonka odotusarvo on 6 ja varianssi 2.4, niin mitä on  $P(X = 5)$ ?
5. Hatussa on  $N$  yhdestä lähtien juoksevasti numeroitua arpalippua. Valitaan hatusta  $n$ :n arvan satunnaisotos palauttamatta (ks. Esimerkki 4.2). Olkoon  $X$  suurin valittujen arpalippujen järjestysnumeroista.
- (a) Piirrä  $X$ :n todennäköisyys- ja kertymäfunktion kuvaajat, kun  $N = 100$  ja  $n = 10$ .
- (b) Piirrä  $X$ :n odotusarvon kuvaaja  $n$ :n funktiona, kun  $N = 100$ .
6. Valitaan satunnaisesti ja toisistaan riippumatta 2000 pistettä yksikköneliöstä  $\{(x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq 1\}$ . Olkoon  $Z$  yksikköympyrään  $\{(x, y) \mid x^2 + y^2 \leq 1\}$  osuvien pisteiden lukumäärä.
- (a) Mitä jakaumaa  $Z$  noudattaa?
- (b) Laske  $Z$ :n odotusarvo ja hajonta.
- (c) Satunnaismuuttujan  $\frac{Z}{500}$  odotusarvo?
- (d) Generoi 2000 satunnaislukuparia. Määritä  $Z$ :n arvo ja laske sen avulla  $\pi$ :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  $f_n = f_{n-1} + f_{n-2}$ ,  $n = 3, 4, \dots$

(a) Osoita, että  $X$ :n todennäköisyysfunktio on

$$f(x) = \frac{f_{x-1}}{2^x}, \quad x = 2, 3, 4, \dots$$

(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  $\text{Var}(X) = 22$ .

(e) Simuloi  $X$ :n arvoja ja tarkastele, vastaavatko simuloinnin tulokset teoreettisia tuloksia.

9. Eräessä 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. Tarkistus suunnitelman 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  $X \sim \text{Poi}(\lambda)$ . Osoita, että  $E(X) = \lambda$  ja  $\text{Var}(X) = \lambda$ .

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 millilitra. Hiiri säilyy hengissä jos ja vain jos kummassakaan ruiskeessa ei ole yhtään organismeja.

- (a) Millä todennäköisyydellä hiiri jää eloon?
- (b) Millä todennäköisyydellä kuolleista hiiristä löytyy molempia orgaanismeja?

(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 maalikuussa,
- (c) ainakin yksi onnettomuus vuoden jokaisena kuukautena.

(Vihje: Käytä Poissonin jakaumaa.)

- 15.** Olkoot  $X$  ja  $Y$  toisistaan riippumattomat Poissonin jakaumaa noudattavat satunnaismuuttujat. Olkoon  $E(X) = 1$  ja  $E(Y) = 2$ .

- (a) Laske todennäköisyys  $P(X + Y) = 5$ .
- (b) Millä kokonaislukuarvolla  $n$  todennäköisyys  $P(X + Y) = n$  saavuttaa maksiminsa?
- (c) Lausu todennäköisyys  $P(X + Y) = 5$  satunnaismuuttujien  $X$  ja  $Y$  todennäköisyysfunktioiden avulla.

- 16.** Kirjassa on 200 sivua. Painovirheiden lukumäärä jokaisella sivulla noudattaa Poissonin jakaumaa, jonka keskiarvo on 0.01. Painovirheiden lukumäärät eri sivuilla ovat toisistaan riippumattomat.

- (a) Mikä on virheettömien sivujen lukumäärän odotusarvo ja hajonta?
- (b) Kirjan oikolukija havaitsee minkä tahansa annetun virheen todennäköisyydellä 0.9. Mikä on oikolukijan havaitsemien virheellisten sivujen lukumäärän odotusarvo?

- 17.** Määritellään  $X$ :n ja  $Y$ : yhteisjakauman todennäköisyysfunktio seuraavasti:

$$f(x, y) = \frac{x + y}{32}, \quad \text{kun } x = 1, 2; \quad y = 1, 2, 3, 4.$$

Määritä

- (a)  $X$ :n reunajakauman todennäköisyysfunktio ja
- (b)  $Y$ :n reunajakauman todennäköisyysfunktio.
- (c) Laske  $P(X > Y)$ ,
- (d)  $P(Y = 2X)$  ja
- (e)  $P(X + Y = 3)$ .

18.  $X$ :n ja  $Y$ :n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.
- Laske odotusarvot  $\mu_X$  ja  $\mu_Y$ ,
  - variانسsit  $\sigma_X^2$  ja  $\sigma_Y^2$  sekä
  - korrelaatiokerroin  $\rho$ .
  - Ovatko  $X$  ja  $Y$  riippumattomat?
19.  $X$ :n ja  $Y$ :n yhteisjakauman todennäköisyysfunktio on määritelty 17. tehtävässä.
- Määritä  $X$ :n ehdolliset todennäköisyysfunktiot  $f_1(x | y)$  ehdolla  $y = 1, 2, 3$  ja  $y = 4$ .
  - Määritä  $Y$ :n ehdolliset todennäköisyysfunktiot  $f_2(y | x)$  ehdolla  $x = 1$  ja  $x = 2$ .
  - Laske  $P(1 \leq Y \leq 3 | X = 1)$ ,  $P(Y \leq 2 | X = 2)$ , ja  $P(X = 2 | Y = 3)$ .
  - Laske  $E(Y | X = 1)$  ja  $\text{Var}(Y | X = 1)$ .
20. Testataan kymmenen suojakypärän iskukestävyys. Kypärät jaetaan kahden viiden ryhmään. Ensimmäisen ryhmän kypärille annetaan isku, joka särkee kypärän todennäköisyydellä 0.1. Toisen ryhmän kypäriä isketään voimalla, joka särkee kypärän todennäköisyydellä 0.3. Millä todennäköisyydellä ensimmäisen ryhmän kypäriä rikkoontuu enemmän kuin toisen ryhmän kypäriä?
21. Oletetaan, että satunnaismuuttujat  $X_1, X_2, X_3$  noudattavat multinomijakaumaa  $\text{Mult}(5; 0.1, 0.3, 0.6)$  [Toisin sanoen  $(X_1, X_2) \sim \text{Tri}(5; 0.1, 0.3)$ ].
- Määritä  $X_1$ :n reuna jakauma ja  $X_2$ :n
  - ehdollinen todennäköisyysfunktio  $f_2(x_2 | x_1 = 1)$ .
22. Nostetaan tavallisesta korttipakasta (52 korttia) satunnaisesti palauttamatta 13 korttia. Olkoon  $X_1$  patojen lukumäärä,  $X_2$  herttojen lukumäärä ja  $13 - X_1 - X_2$  ruutujen ja ristien lukumäärä otoksessa.
- Määritä  $X_1$ :n ja  $X_2$ :n yhteisjakuman todennäköisyysfunktio. Mikä on satunnaisvektorin  $(X_1, X_2)$  arvoalue?
  - $X_1$ :n todennäköisyysfunktio ja arvoalue?
23. Oletetaan, että  $(X, Y)$  noudattaa trinomijakaumaa  $\text{Tri}(3, \frac{1}{6}, \frac{1}{2})$ .
- Laske odotusarvot  $\mu_X$  ja  $\mu_Y$ ,
  - variانسsit  $\sigma_X^2$  ja  $\sigma_Y^2$  sekä
  - kovarianssi  $\text{Cov}(X, Y)$  ja

(d) korrelaatiokerroin  $\rho$ .

24. Satunnaismuuttujat  $X \geq 0$  ja  $Y \geq 0$  ovat riippumattomat ja saavat vain kokonaislukuarvoja. Osoita, että

(a) 
$$P(X + Y = n) = \sum_{k=0}^n P(X = k) P(Y = n - k).$$

(b) Heitetään 4:ää harhatonta noppaa ja lasketaan silmälukujen summa. Mikä on todennäköisyys, että summa on 8? (Vihje: Olkoon  $X$  kahden nopan silmälukujen summa ja  $Y$  kahden muun nopan silmälukujen summa.)





# Luku 5

## Jatkuvat jakaumat

Sellaiset suureet kuten esimerkiksi aika, lämpötila, pituus ja paino ajatellaan tavallisesti jatkuviksi muuttujiksi, ts. muuttujiksi, jotka voivat saada mitä tahansa reaaliarvoja annetulla välillä. Esimerkiksi henkilön ikä on *jatkuva satunnaismuuttuja*, joka voi saada positiivisia reaalitylukuarvoja. Diskreetin satunnaismuuttujan arvoavaruus on äärellinen tai numeroituva, mutta jatkuvan satunnaismuuttujan arvoavaruus on ylinumeroituva.

### 5.1 Jatkuvat satunnaismuuttujat

Jokaiseen satunnaismuuttujaan liittyy kertymäfunktio. Satunnaismuuttujan  $X$  kertymäfunktio määriteltiin alaluvussa 2.5.2 (Määritelmä 2.4) funktiona

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

Diskreetin satunnaismuuttujan kertymäfunktio on porraskfunktio, joka voidaan lausua hyppöfunktioiden summana (4.1.1) [ks. alaluku 4.1].

Lauseen 2.10 mukaan funktio  $F(x)$  on kertymäfunktio jos ja vain jos seuraavat kolme ehtoa toteutuvat:

1.  $\lim_{x \rightarrow -\infty} F(x) = 0$  ja  $\lim_{x \rightarrow \infty} F(x) = 1$ .
2.  $F(x)$  on kasvava (ei-vähenevä) funktio.
3.  $F(x)$  on oikealta jatkuva eli  $\lim_{x \rightarrow x_0+} F(x) = F(x_0)$  kaikilla  $x_0 \in \mathbb{R}$ .

Jos meillä on jokin satunnaismuuttuja  $X$ , niin ominaisuudet 1.–3. voidaan todeta todennäköisyysfunktion  $P(X \leq x)$  ominaisuuksien avulla. Jos jokin funktio  $F(x)$  toteuttaa ehdot 1.–3., ei ole aivan helppoa todistaa, että  $F(x)$  on todella jonkin satunnaismuuttujan kertymäfunktio. Todistus löytyy vaaivista todennäköisyyslaskennan oppikirjoista.

**Esimerkki 5.1** Funktio

$$(5.1.1) \quad F(x) = \frac{1}{1 + e^{-x}}$$

on esimerkki jatkuvasta kertymäfunktioista, joka siis toteuttaa Lauseen 2.10 ehdot 1.–3. Koska

$$\lim_{x \rightarrow -\infty} e^{-x} = \infty, \quad \text{niin} \quad \lim_{x \rightarrow -\infty} F(x) = 0$$

ja

$$\lim_{x \rightarrow \infty} F(x) = 1, \quad \text{koska} \quad \lim_{x \rightarrow \infty} e^{-x} = 0.$$

Funktio  $F(x)$  on kasvava, koska sen 1. derivaatta

$$F'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0.$$

On myös helppo todeta, että  $F(x)$  ei ole ainoastaan oikealta jatkuva vaan *jatkuva*.  $\square$

Satunnaismuuttujan jatkuvuus voidaankin määritellä siihen liittyvän kertymäfunktion jatkuvuuden avulla.

**Määritelmä 5.1** Satunnaismuuttuja  $X$  on jatkuva, jos sen kertymäfunktio  $F_X(x)$  on  $x$ :n jatkuva funktio. Satunnaismuuttuja  $X$  on diskreetti, jos sen kertymäfunktio on  $x$ :n porraskäyrä.

Vastaavalla tavalla kuin diskreetin satunnaismuuttujan kertymäfunktio voidaan lausua summana, voidaan jatkuvan satunnaismuuttujan kertymäfunktio lausua integraalina:

$$(5.1.2) \quad P(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(t) dt.$$

Jos  $f_X(t)$  on jatkuva, niin integraalilaskennan peruslauseen mukaan

$$(5.1.3) \quad F'_X(x) = f_X(x),$$

missä  $F'_X(x)$  on kertymäfunktion  $F_X(x)$  derivaatta.

**Määritelmä 5.2** Jatkuvan satunnaismuuttujan  $X$  tiheysfunktio  $f_X(x)$  on funktio, joka toteuttaa yhtälön

$$(5.1.4) \quad F_X(x) = \int_{-\infty}^x f_X(t) dt \quad \text{kaikilla} \quad x \in \mathbb{R}.$$

**Esimerkki 5.2** Olkoon  $X$  tiettyyn palvelunumeroon tulevien puheluiden pituus. Oletetaan, että  $X$ :n tiheysfunktio on

$$f(x) = \frac{1}{20} e^{-x/20}, \quad 0 \leq x < \infty.$$

Silloin  $X$  noudattaa ns. *eksponenttijakaumaa* keskiarvolla 20. Nyt

$$S = \{x \mid 0 \leq x < \infty\} \quad \text{ja} \quad f(x) > 0 \quad \text{kun} \quad x \in S.$$

Kertymäfunktio on

$$\begin{aligned} F(x) &= \int_{-\infty}^x \frac{1}{20} e^{-t/20} dx = \int_0^x \frac{1}{20} e^{-t/20} dx \\ &= \int_0^x -e^{-t/20} = 1 - e^{-x/20}. \end{aligned}$$

Silloin

$$F'(x) = \frac{d}{dx}(1 - e^{-x/20}) = \frac{1}{20} e^{-x/20} = f(x), \quad x \geq 0$$

ja  $f(x) = 0$ , kun  $x < 0$ . □

Huomaa, että yksittäisen pisteen  $a \in \mathbb{R}$  todennäköisyys  $P(X = a)$  on aina nolla, jos  $X$  on jatkuva satunnaismuuttuja. Silloin erityisesti kaikilla reaalityyppisillä  $b > a$

$$\begin{aligned} F(b) - F(a) &= P(a \leq X \leq b) = P(a < X \leq b) \\ &= P(a \leq X < b) = P(a < X < b). \end{aligned}$$

**Esimerkki 5.3** Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio on  $f(x) = 2x$ , kun  $0 < x < 1$ . Silloin  $X$ :n kertymäfunktio on

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

Huomaa, että

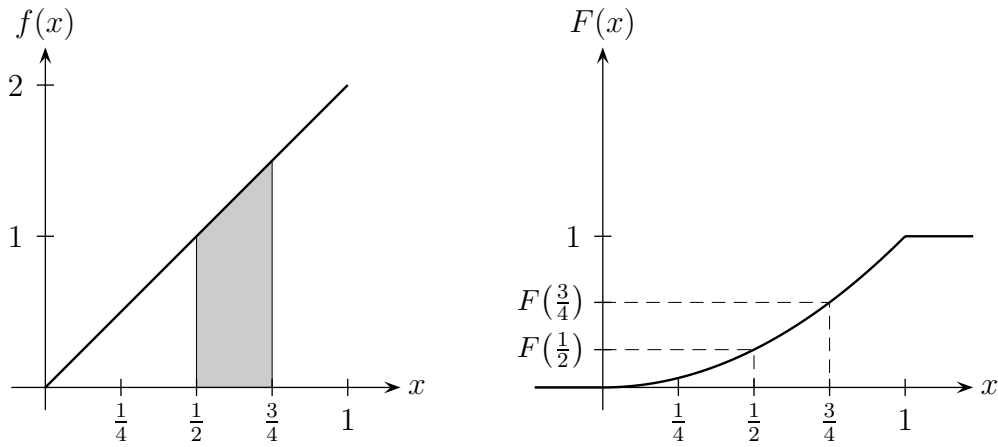
$$F(x) = \int_0^x 2t dt = x^2, \quad \text{kun } 0 \leq x < 1.$$

Jos kertymäfunktio on annettu, niin tiheysfunktio saadaan derivoimalla kertymäfunktio:

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

Kertymäfunktion avulla voidaan laskea todennäköisyyksiä. Esimerkiksi todennäköisyys

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



**Kuvio 5.1.** Jatkuvan satunnaismuuttujan  $X$  tiheysfunktio  $f(x) = 2x$  ja kertymäfunktio  $F(x) = x^2$ .

ja

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

Toisaalta tietysti  $P\left(\frac{1}{2} \leq X \leq \frac{3}{4}\right)$  voidaan laskea suoran  $y = 2x$  ja  $x$ -akselin väliin jäävänä pinta-alana:

$$P\left(\frac{1}{2} \leq X \leq \frac{3}{4}\right) = \int_{1/2}^{3/4} 2x \, dx = \frac{5}{15},$$

joka tietysti voidaan esittää kertymäfunktion avulla. □

Jatkuvan satunnaismuuttujan momentit määritellään vastaavasti kuin diskreetin satunnaismuuttujan tapauksessa, mutta määritelmässä summa korvataan integraalilla. Jatkuvan satunnaismuuttujan  $r$ . *momentti* on

$$\alpha_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) \, dx,$$

missä  $f(x)$  on  $X$ :n tiheysfunktio. Satunnaismuuttujan  $X$   $r$ . *keskusmomentti* on

$$\mu_r = E[(X - \mu)^r],$$

missä  $\mu = E(X) = \alpha_1$  on  $X$ :n *odotusarvo*. Satunnaismuuttujan  $X$  odotusarvo on siis integraali

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) \, dx$$

ja  $X$ :n varianssi  $\sigma^2$  on 2. keskusmomentti

$$\begin{aligned}\sigma^2 = \mu_2 &= E[(X - \mu)^2] \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.\end{aligned}$$

Merkitsemme myös  $E[(X - \mu)^2] = \text{Var}(X)$ , jolloin  $X$ :n hajonta on

$$\sigma = \sqrt{\text{Var}(X)}.$$

Momenttifunktio on

$$(5.1.5) \quad M(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx,$$

jos integraali 5.1.5 on olemassa jollakin avoimella välillä  $(-a, a)$ , missä  $a > 0$ . Tietysti esimerkiksi tulokset

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2, \\ \mu &= M'(0), \\ \alpha_2 &= E(X^2) = M''(0)\end{aligned}$$

pitävät edelleen paikkansa samalla tavalla kuin diskreettien satunnaismuuttujien tapauksessa.

**Esimerkki 5.4** Lasketaan nyt Esimerkissä 5.3 määritellyn satunnaismuuttujan  $X$  odotusarvo ja varianssi:

$$\mu = E(X) = \int_0^1 x(2x) dx = \frac{2}{3} \int_0^1 x^2 dx = \frac{2}{3}$$

ja

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 \\ &= \int_0^1 x^2(2x) dx - \left(\frac{2}{3}\right)^2 = \frac{1}{2} \int_0^1 x^3 dx - \frac{4}{9} = \frac{1}{18}.\end{aligned}$$

Kolmas momentti on

$$\alpha_3 = E(X^3) = \int_0^1 x^3(2x) dx = \frac{2}{5} \int_0^1 x^4 dx = \frac{2}{5}$$

ja 3. keskusmomentti on

$$\begin{aligned}
 \mu_3 &= E[(X - \mu)^3] = \int_0^1 (x - \mu)^3 (2x) dx \\
 &= \int_0^1 (x^3 - 3\mu x^2 + 3\mu^2 x - \mu^3)(2x) dx \\
 &= \int_0^1 x^3 (2x) dx - 3\mu \int_0^1 x^2 (2x) dx + 3\mu^2 \int_0^1 x (2x) dx - \mu^3 \int_0^1 2x dx \\
 &= \alpha_3 - 3\mu\alpha_2 + 3\mu^3 - \mu^3 = \alpha_3 - 3\mu\alpha_2 + 2\mu^3 \\
 &= \frac{2}{5} - 3 \cdot \frac{2}{3} \cdot \frac{1}{2} + 2 \cdot \left(\frac{2}{3}\right)^3 = -\frac{3}{5} + \frac{16}{27} = \frac{1}{15}.
 \end{aligned}$$

□

Myös prosenttipisteet ovat tärkeitä jakauman tunnuslukuja. Jakauman  $100p$ -prosenttipiste  $\pi_p$  määritellään seuraavasti:

$$p = \int_{-\infty}^{\pi_p} f(x) dx = F(\pi_p), \quad 0 \leq p \leq 1.$$

Prosenttipistettä  $\pi_{0.50}$  kutsutaan *mediaaniksi* ja pistettä  $\pi_{0.25}$  ja  $\pi_{0.75}$  *alakvartiiliksi* ja *yläkvartiiliksi*. Esimerkissä 5.3 käsitellyn jakauman 36 %:n piste on 0.6, koska

$$F(\pi_{0.36}) = \pi_{0.36}^2 = 0.6^2 = 0.36.$$

**Esimerkki 5.5** Olkoon satunnaismuuttujan  $X$  kertymäfunktio määritelty seuraavasti

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

Tarkistamme ensin, että  $F$  on todella kertymäfunktio. Toteamme helposti, että

- 1)  $\lim_{x \rightarrow -\infty} F(x) = 0$  ja  $\lim_{x \rightarrow \infty} F(x) = 1$ ,
- 2)  $F(x)$  on  $x$ :n kasvava (ei-vähenevä) funktio ja
- 3)  $F(x)$  on oikealta jatkuva, koska se on jatkuva.

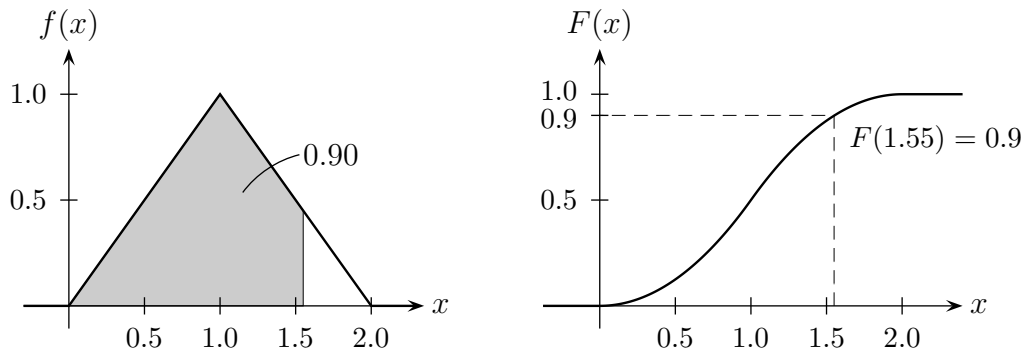
Tiheysfunktio saadaan derivoimalla  $F(x)$ . Nyt siis  $F'(x) = x$  välillä  $0 < x \leq 1$  ja  $F'(x) = 2 - x$  välillä  $1 \leq x \leq 2$ . Näin siis tiheysfunktio on

$$f(x) = \begin{cases} x, & 0 < x \leq 1; \\ 2 - x, & 1 \leq x \leq 2; \\ 0 & \text{muualla.} \end{cases}$$

Tiheysfunktio voidaan kirjoittaa lyhyesti muodossa

$$f(x) = 1 - |x - 1|, \quad 0 \leq x \leq 2.$$

Koska  $X$ :n tiheysfunktion kuvaaja on kolmion muotoinen,  $X$ :n jakaumaa kutsutaan kolmiojakaumaksi.



**Kuvio 5.2.** Kolmiojakauman tiheysfunktion ja kertymäfunktion kuvaajat.

Kolmiojakauman odotusarvo on

$$\begin{aligned} \mu &= \int_0^2 x f(x) dx = \int_0^1 x \cdot x dx + \int_1^2 x(2-x) dx \\ &= \int_0^1 \frac{x^3}{3} + \int_1^2 \left(x^2 - \frac{x^3}{3}\right) = \frac{1}{3} + \left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right) \\ &= 1. \end{aligned}$$

Koska jakauma on symmetrinen odotusarvon 1 suhteen, on 1 myös jakauman mediaani  $\pi_{0.50}$ . Se voidaan todeta helposti myös määritelmän perusteella, sillä  $F(1) = \frac{1^2}{2} = 0.5$ . Jakauman 90 %:n piste  $\pi_{0.90}$  saadaan ratkaisemalla yhtälö

$$1 - \frac{(2 - \pi_{0.90})^2}{2} = 0.90.$$

Ratkaisu on  $\pi_{0.90} = 2 - \sqrt{0.2} = 1.55$ . □

Itse asiassa relaatio (5.1.4) ei välttämättä ole voimassa kaikilla  $x$ :n arvoilla, sillä  $F(x)$  voi olla jatkuva, mutta ei derivoituva. Jos  $f(x)$  on jatkuva, niin silloin tietysti yhtälö (5.1.4) pitää paikkansa. Huomattakoon, että jatkuvan satunnaismuuttujan tiheysfunktio ei välttämättä ole jatkuva, mutta kertymäfunktio on.

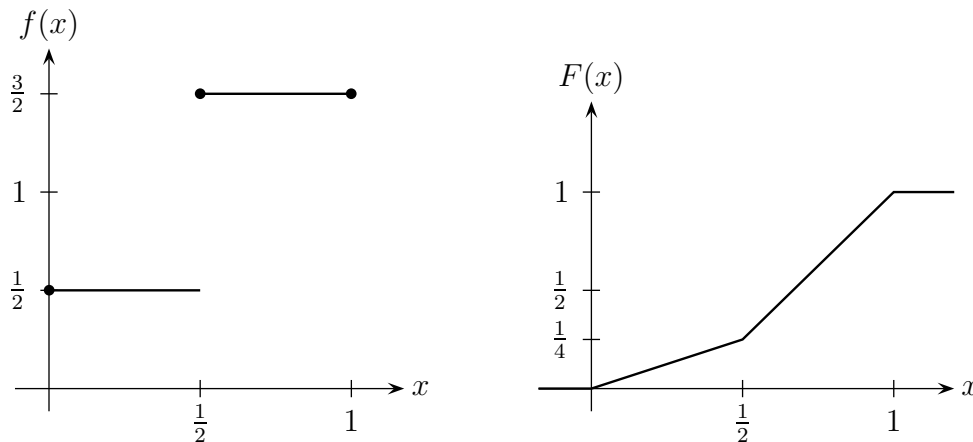
**Esimerkki 5.6** Tarkastellaan nyt satunnaismuuttujaa  $X$ , jonka tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{2}, & 0 \leq x < \frac{1}{2}; \\ \frac{3}{2}, & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Vastaavasti  $X$ :n kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{2}x, & 0 \leq x < \frac{1}{2}; \\ \frac{1}{4} + \frac{3}{2}(x - \frac{1}{2}), & \frac{1}{2} \leq x \leq 1; \\ 1, & 1 \leq x. \end{cases}$$

Havaitsemme nyt, että  $X$ :n tiheysfunktio ei ole jatkuva. Nyt myöskään  $F$  ei



**Kuvio 5.3.** Satunnaismuuttujan  $X$  tiheysfunktion ja kertymäfunktion kuvaajat.

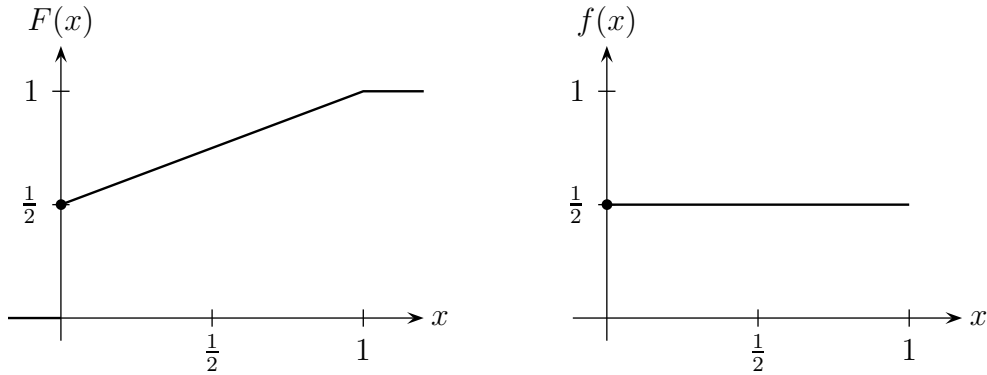
ole derivoituva pisteessä  $\frac{1}{2}$ . Pisteessä  $x = \frac{1}{2}$  ei ole voimassa, että  $F'(x) = f(x)$ . Tässä on esimerkki jatkuvasta satunnaismuuttujasta, jonka tiheysfunktio ei ole jatkuva ja jonka kertymäfunktio ei ole koko määrittelyalueella  $S$  derivoituva.  $\square$

Jatkuvan satunnaismuuttujan tiheysfunktioilla voi olla äärellinen määrä epäjatkuvuuspisteitä, mutta kertymäfunktio on jatkuva. Esimerkin 5.6 satunnaismuuttujan tiheysfunktioilla on määrittelyalueellaan yksi epäjatkuvuuspiste ja kertymäfunktio on jatkuva. Relaatio (5.1.3) pitää paikkansa vain tiheysfunktion jatkuvuuspisteissä, mutta ei epäjatkuvuuspisteissä.



**Esimerkki 5.7** Määritellään satunnaismuuttuja  $X$  siten, että sen kertymäfunktio on

$$(5.1.6) \quad F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{2}, & x = 0; \\ \frac{1}{2} + \frac{x}{2}, & 0 < x < 1; \\ 1, & 1 \leq x. \end{cases}$$



**Kuvio 5.4.** Satunnaismuuttujan  $X$  kertymäfunktion ja 'tiheysfunktion' kuvaajat.

Kertymäfunktio ei ole nyt jatkuva, koska funktio hyppää pisteessä  $x = 0$ . Kertymäfunktio ei ole myöskään porraskontinuuksinen. Nyt myös yksittäisellä pisteellä  $X = 0$  on positiivinen todennäköisyys  $P(X = 0) = \frac{1}{2}$ , joten  $f(x)$  ei ole tiheysfunktio. Itse asiassa kertymäfunktio (5.1.6) voidaan kirjoittaa porraskontinuuksisen kertymäfunktion (kertymäfunktio) ja jatkuvan kertymäfunktion summana. Alaluvussa 4.1 määriteltiin hyppyfunktio  $\varepsilon(x)$  siten, että  $\varepsilon(x) = 1$  epänegatiivisilla  $x$ :n arvoilla ja  $\varepsilon(x) = 0$ , kun  $x < 0$ . Funktio  $\varepsilon(x)$  on porraskontinuuksinen ja siis diskreetin satunnaismuuttujan kertymäfunktio. Puoliavoimella välillä  $(0, 1]$  tasajakaumaa noudattavan satunnaismuuttujan kertymäfunktio on

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

Nyt kertymäfunktio (5.1.6) voidaan kirjoittaa muodossa

$$F(x) = \frac{1}{2} \varepsilon(x) + \frac{1}{2} F_c(x).$$

Esimerkiksi todennäköisyys

$$\begin{aligned} P\left(X \leq \frac{1}{2}\right) &= \frac{1}{2} \varepsilon\left(\frac{1}{2}\right) + \frac{1}{2} F_c\left(\frac{1}{2}\right) \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Satunnaismuuttuja  $X$  ei ole diskreetti eikä jatkuva. □

Yleisesti jatkuva satunnaismuuttuja voidaan määritellä identiteetin (5.1.4) avulla olettamatta tiheysfunktion  $f(x)$  jatkuvuutta. Jos on olemassa sellainen epänegatiivinen funktio  $f(x)$  [ts.  $f(x) \geq 0$  kaikilla  $x \in \mathbb{R}$ ], että (5.1.4) pitää paikkansa kaikilla  $x \in \mathbb{R}$ , niin kertymäfunktion  $F(x)$  sanotaan olevan *absoluuttisesti jatkuva*. Absoluuttisesti jatkuva funktio on jatkuva. Kaikkien tässä luvussa käsiteltäviät jatkuvien satunnaismuuttujien kertymäfunktiot ovat absoluuttisesti jatkuvia.

## 5.2 Tasajakauma ja eksponenttijakauma

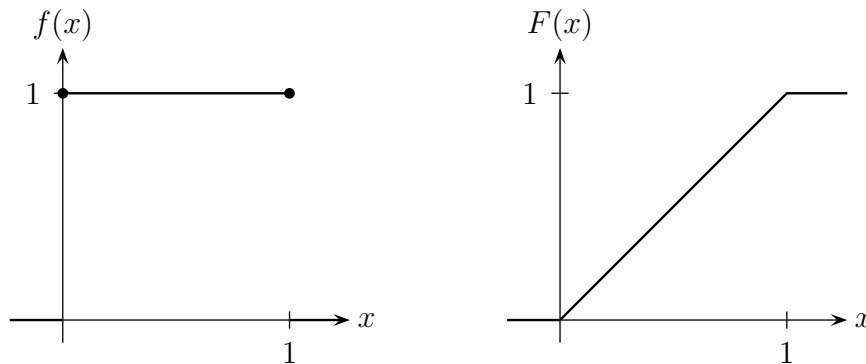
### 5.2.1 Tasajakauma

Jatkuva satunnaismuuttuja  $X$  noudattaa *tasajakaumaa* välillä  $[0, 1]$ , jos sen tiheysfunktio on 1 tällä välillä ja 0 muualla:

$$(5.2.1) \quad f(x) = \begin{cases} 1, & \text{kun } x \in [0, 1], \\ 0 & \text{muualla.} \end{cases}$$

Silloin merkitään  $X \sim \text{Tas}(0, 1)$ . On helppo todeta, että  $f(x)$  on tiheysfunktio, koska  $f(x) \geq 0$  ja

$$\int_0^1 f(x) dx = \int_0^1 dx = 1.$$



**Kuvio 5.5.** Tasajakauman  $\text{Tas}(0, 1)$  tiheysfunktio ja kertymäfunktio.

Tasajakauman keskiarvo ja varianssi ovat:

$$E(X) = \int_0^1 x dx = \frac{1}{2}$$

ja

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \int_0^1 x^2 dx - \frac{1}{4} = \frac{1}{12}.$$

Satunnaismuuttujan  $X$  momenttifunktio on

$$M_X(t) = \int_0^1 e^{tx} dx = \int_0^1 \frac{1}{t} e^{tx} = \frac{e^t - 1}{t}.$$

Huomaa, että  $M_X(0) = 1$ .

Olkoon  $[a, b]$  annettu suljettu väli,  $a < b$ . Silloin satunnaismuuttuja  $U = (b - a)X + a$  noudattaa tasajakaumaa välillä  $[a, b]$ . Silloin merkitään  $U \sim \text{Tas}(a, b)$ . Koska  $E(U) = (b - a)E(X) + a$  ja  $\text{Var}(U) = (b - a)^2 \text{Var}(X)$ , niin

$$E(U) = \frac{a + b}{2} \quad \text{ja} \quad \text{Var}(U) = \frac{(b - a)^2}{12}.$$

Satunnaismuuttujan  $U$  tiheysfunktio on

$$(5.2.2) \quad f(u) = \begin{cases} \frac{1}{b-a}, & \text{kun } u \in [a, b]; \\ 0 & \text{muualla} \end{cases}$$

ja  $U$ :n momenttifunktio on

$$M_U(t) = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)}, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

## 5.2.2 Eksponenttijakauma

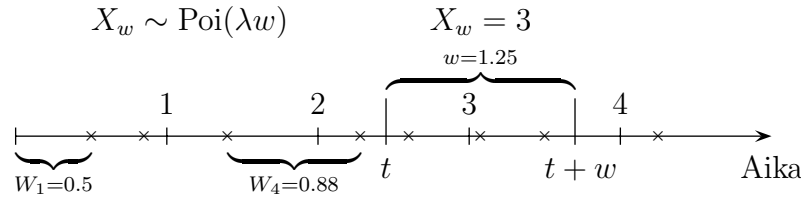
Poissonin prosessissa tarkastellaan, montako tapahtumaa (lisäystä) sattuu jollain aikavälillä. Merkitään  $w$ :n pituisella välillä sattuvien tapahtumien lukumäärää satunnaismuuttujalla  $X_w$ . Jos Poissonin prosessin intensiteetti on  $\lambda$ , niin Määritelmän 4.3 mukaan todennäköisyys, että  $w$ :n pituisella välillä sattuu  $x$  tapahtumaa, on

$$(5.2.3) \quad P(X_w = x) = e^{-\lambda w} \frac{(\lambda w)^x}{x!}.$$

Poissonin prosessilla voidaan mallintaa esimerkiksi asiakkaiden saapumista palvelupisteeseen, puheluiden tuloa vaihteeseen, onnettomuuksien sattumista tarkasteltavalla tieosuudella tai autojen kulkua liikenteen tarkkailupisteen ohi. Tällöin ajatellaan, että yksittäiset tapahtumat sattuvat toisistaan riippumatta täysin satunnaisesti.

Tarkkaillaan nyt Poissonin prosessia, jonka intensiteetti on  $\lambda$ . Olkoon  $W$  odotusaika siihen hetkeen, kunnes seuraava tapahtuma sattuu. Odotusaika on jatkuva satunnaismuuttuja. Jos tarkkailemme prosessia hetkestä  $t$  hetkeen  $t + w$  eli  $w$ :n pituisen aikavälin  $[t, t + w]$ , niin tapahtuma  $\{W > w\}$  sattuu jos ja vain jos Poissonin prosessissa ei satu yhtään tapahtumaa välillä  $[t, t + w]$ . Siksi identiteetin (5.2.3) mukaan

$$P(W > w) = P(X_w = 0) = e^{-\lambda w}.$$



**Kuvio 5.6.** Kaaviokuva esittää Poissonin saapumisprosessia, esimerkiksi autojen kulkemista liikenteen tarkkailupisteen ohi. Esimerkiksi  $W_1$  on 1. auton odotusaika ja  $W_4$  on 3. ja 4. auton välinen aika. Kiinnitetyllä  $w$ :n pituisella välillä on kulkenut ohi  $X_w = 3$  autoa. Peräkkäiset odotusajat  $W_1, W_2, W_3, \dots$  ovat toisistaan riippumattomat ja noudattavat samaa jakaumaa.

Odotusajan  $W$  kertymäfunktio on siis

$$\begin{aligned} F(w) &= P(W \leq w) \\ &= 1 - P(W > w) = 1 - P(X_w = 0) \\ &= 1 - e^{-\lambda w}. \end{aligned}$$

Koska odotusaika  $W$  on epänegatiivinen, niin  $F(w) = 0$ , kun  $w < 0$ .

Odotusajan  $W$  tiheysfunktio on

$$F'(w) = f(w) = \lambda e^{-\lambda w}$$

derivointisäännön (5.1.3) nojalla. Usein merkitään  $\lambda = \frac{1}{\theta}$ , missä  $\theta > 0$ . Sanomme, että  $W$  noudattaa *eksponenttijakaumaa* parametrilla  $\theta$  ja merkitsemme  $W \sim \text{Exp}(\theta)$ . Parametri  $\theta$  on jakauman keskiarvo. Eksponenttijakauman tiheysfunktio on silloin muotoa

$$(5.2.4) \quad f(w) = \frac{1}{\theta} e^{-w/\theta}.$$

Eksponenttijakauman  $\text{Exp}(\theta)$  momenttifunktio on

$$\begin{aligned} M(t) &= \int_0^{\infty} e^{tw} \frac{1}{\theta} e^{-w/\theta} dw = \int_0^{\infty} -\frac{e^{-(1-\theta t)w/\theta}}{1-\theta t} \\ &= \frac{1}{1-\theta t}, \quad t < \frac{1}{\theta}. \end{aligned}$$

Eksponenttijakaumalla on vastaava ”unohtamisominaisuus” kuin geometrisella jakaumalla. Jos  $T \sim \text{Exp}(\theta)$ , niin

$$(5.2.5) \quad P(T > a + b \mid T > a) = P(T > b)$$

kaikilla epänegatiivisilla  $a$  ja  $b$ . Tulos voidaan todistaa laskemalla ehdollinen todennäköisyys

$$\begin{aligned} P(T > a + b \mid T > a) &= \frac{P(T > a, T > a + b)}{P(T > a)} = \frac{P(T > a + b)}{P(T > a)} \\ &= \frac{e^{-(a+b)/\theta}}{e^{-a/\theta}} = e^{-b/\theta} = P(T > b). \end{aligned}$$

Huomattakoon, että edellä on käytetty tulosta

$$P(T > t) = 1 - P(T \leq t) = 1 - F(t) = e^{-t/\theta}, \quad t \geq 0.$$

**Esimerkki 5.8** Oletetaan, että asiakkaiden saapuminen liikkeeseen noudattaa Poissonin prosessia intensiteetillä 20 asiakasta tunnissa. Mikä on todennäköisyys, että myyjä joutuu odottamaan seuraavaa asiakasta yli 5 minuuttia? Olkoon  $X$  odotusaika, kunnes seuraava asiakas saapuu. Silloin prosessissa (5.2.3)  $\lambda = 1/3$  asiakasta minuutissa ja  $X \sim \text{Exp}(3)$ , koska eksponenttijakauman keskiarvo  $\theta = 1/\lambda$ . Jakauman  $\text{Exp}(3)$  tiheysfunktio on

$$f(x) = \frac{1}{3}e^{-x/3}, \quad 0 \leq x < \infty$$

ja

$$P(X > 5) = \int_5^{\infty} \frac{1}{3}e^{-x/3} dx = \int_5^{\infty} -e^{-x/3} = e^{-5/3} \approx 0.1889.$$

Jatkuvan jakauman *mediaani*  $m$  on sellainen piste, että  $F(m) = 1/2$ . Nyt jakauman  $\text{Exp}(3)$  mediaanin  $m$  tulee toteuttaa ehto  $F(m) = 1 - e^{-m/3} = \frac{1}{2}$ , joten

$$m = 3 \log(2) \approx 2.0794.$$

□

### 5.2.3 Elinajakajakauma

Ominaisuuden (5.2.5) perusteella eksponenttijakauma on sopiva elinajan jakauma silloin, kun jäljellä oleva elinaika ei riipu tämänhetkisestä iästä. Olkoon  $T$  esimerkiksi jonkin elektronisen komponentin ikä tunteina. Silloin  $P(T > b)$  on todennäköisyys, että uusi komponentti kestää ainakin  $b$  tuntia, kun taas  $P(T > a + b \mid T > a)$  on todennäköisyys, että  $a$  tuntia käytössä ollut komponentti kestää vielä  $b$  tuntia. Jos elinaika noudattaa eksponenttijakaumaa, niin ominaisuuden (5.2.5) nojalla todennäköisyydet  $P(T > b)$  ja  $P(T > a + b \mid T > a)$  ovat samat kaikilla  $a$  ja  $b$ . Todennäköisyys, että komponentti rikkoontuu  $b$ :n seuraavan tunnin aikana, ei riipu lainkaan siitä, kuinka kauan komponentti on jo ollut käytössä.

Funktiota  $G(t) = P(T > t)$  kutsutaan *eloonjäämisfunktioiksi*. Eksponenttijakauma määrittelee eloonjäämisfunktion  $G(t) = e^{-t/\theta}$ , jolla on unohtamisominaisuus

$$(5.2.6) \quad G(t + s) = G(t)G(s), \quad t > 0, \quad s > 0.$$

Määritelmänsä nojalla  $G(0) = 1$  ja  $G(t) \rightarrow 0$ , kun  $t$  kasvaa. Onko eksponenttifunktion lisäksi muita eloonjäämisfunktioita, joilla on unohtamisominaisuus (5.2.6)? Voidaan osoittaa, että ehdon (5.2.6) toteuttavat eloonjäämisfunktiot ovat aina muotoa  $e^{-\lambda t}$ ,  $\lambda > 0$ .

Jos elinaika  $T$  noudattaa eksponenttijakaumaa  $\text{Exp}(\theta)$ , niin vakio  $\lambda = \frac{1}{\theta}$  on hetkellinen *kuolleisuusaste* tai *vaaran aste*. Parametri  $\lambda$  säätelee todennäköisyyttä kuolla hetken  $T = t$  jälkeisellä yksikön pituisella aikavälillä.

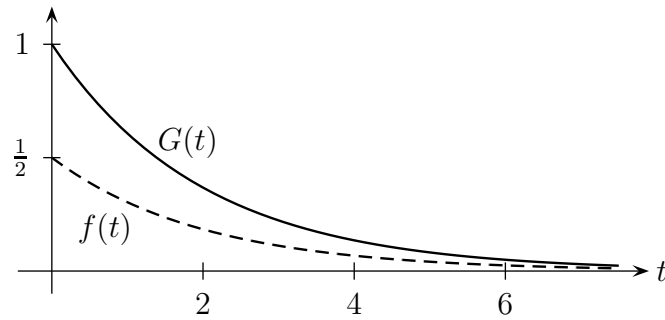
Olkoon  $\Delta$  tarkasteltavan aikavälin pituus. Määritellään todennäköisyys

$$\begin{aligned} P(T \leq t + \Delta \mid T > t) &= 1 - P(T > t + \Delta \mid T > t) \\ &= 1 - P(T > \Delta) = 1 - e^{-\lambda\Delta}, \end{aligned}$$

missä viimeistä edellinen yhtäsuuruus saadaan unohtamisominaisuuden (5.2.6) nojalla. Kun funktiota  $e^{-\lambda\Delta}$  arvioidaan Taylorin polynomin avulla, saadaan

$$\begin{aligned} 1 - e^{-\lambda\Delta} &= 1 - (1 - \lambda\Delta + \frac{1}{2}\lambda^2\Delta^2 - \dots) \\ &= \lambda\Delta - \frac{1}{2}\lambda^2\Delta^2 + \dots \\ &\approx \lambda\Delta, \quad \text{kun } \Delta \text{ on pieni.} \end{aligned}$$

Arviointivirhe pienenee merkityksettömäksi verrattuna  $\Delta$ :aan, kun  $\Delta \rightarrow 0$ . Silloin siis  $P(T \leq t + \Delta \mid T > t) \approx \lambda\Delta$ .



**Kuvio 5.7.** Eksponenttijakauman  $\text{Exp}(2)$  tiheysfunktio  $f(t) = \frac{1}{2}e^{-t/2}$  ja vastaava eloonjäämisfunktio  $G(t) = e^{-t/2}$ .

Nyt nähdään, että

$$\lim_{\Delta \rightarrow 0} \frac{P(T \leq t + \Delta \mid T > t)}{\Delta} = \lambda$$

on riippumaton ajasta  $t$ . Eksponentiaalisesti jakautuneen elinajan tapauksessa kuolleisuusaste  $\lambda$  on iästä riippumaton vakio. Yleisesti kuolleisuusaste  $\lambda(t)$  on tietysti iän funktio.

### 5.3 Gammajakauma ja $\chi^2$ -jakauma

Gammajakaumajakauma on välillä  $[0, \infty)$  määritelty jakauma tai jakaumaperhe, koska parametrien vaihdellessa saadaan hyvinkin erinäköisiä jakaumia,

vaikka ne ovat matemaattisesti samaa muotoa. Gammafunktio

$$(5.3.1) \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

määriteltiin jo Pykälässä 2.4.7. Jos  $\alpha > 0$ , niin  $\Gamma(\alpha)$  on äärellinen. Jos  $\alpha$  on positiivinen kokonaisluku, niin  $\Gamma(\alpha)$  voidaan lausua suljetussa muodossa, muutoin ei.

Gammafunktio toteuttaa rekursiivisen relaation

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha),$$

joka voidaan osoittaa osittaisintegroinnilla. Jos  $\alpha = n$  on positiivinen kokonaisluku, niin

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

Koska  $\Gamma(1) = 1$ , niin

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

kaikilla positiivisilla kokonaisluvuilla. Myös  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$  on tärkeä erikoistapaus.

Funktio

$$(5.3.2) \quad f(t) = \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)}, \quad 0 < t < \infty$$

määrittelee tiheysfunktion, sillä gammafunktiossa integroitava on positiivinen välillä  $(0, \infty)$ . Sanokaamme, että (5.3.2) on satunnaismuuttujan  $T$  tiheysfunktio. Kaikkien gammajakaumien perhe saadaan määrittelemällä satunnaismuuttuja  $X = \beta T$ , missä  $\beta$  on positiivinen vakio.  $X$ :n tiheysfunktio voidaan johtaa soveltamalla Lauseen 5.5 muunnostekniikkaa. Merkitsemme  $X \sim \text{Gamma}(\alpha, \beta)$  ja sanomme, että  $X$  noudattaa gammajakaumaa parametrein  $\alpha$  ja  $\beta$ . Jakauman  $\text{Gamma}(\alpha, \beta)$  tiheysfunktioiksi saadaan

$$(5.3.3) \quad f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0.$$

Esitämme nyt gammajakauman perusominaisuudet seuraavassa lauseessa.

**Lause 5.1** *Oletetaan, että  $X \sim \text{Gamma}(\alpha, \beta)$ .*

1. *Funktio (5.3.3) määrittelee tiheysfunktion kaikilla  $\alpha > 0$ ,  $\beta > 0$ .*

2.

$$E(X) = \alpha\beta, \quad \text{Var}(X) = \alpha\beta^2$$

ja

$$M(t) = E(e^{tX}) = \left( \frac{1}{1 - \beta t} \right)^\alpha, \quad t < \frac{1}{\beta}.$$

3.

$$E(X^c) = \frac{\Gamma(\alpha + c)\beta^c}{\Gamma(\alpha)}$$

kaikilla  $c > -\alpha$ .

4. Olkoon  $U = bX$ ,  $b > 0$ . Silloin  $U \sim \text{Gamma}(\alpha, b\beta)$ .

Eksponettijakauma on gammajakauman erikoistapaus. Kun sijoitetaan tiheysfunktioon (5.3.3)  $\alpha = 1$ , saadaan

$$f(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0.$$

Havaitaan siis, että  $\text{Gamma}(1, \beta) = \text{Exp}(\beta)$ .

### $\chi^2$ -jakauma

Toinen tärkeä gammajakauman erikoistapaus on  $\chi^2$ -jakauma. Jos valitaan  $\alpha = \frac{r}{2}$ , missä  $r$  on positiivinen kokonaisluku, ja  $\beta = 2$ , tulee tiheysfunktio (5.3.3) muotoon

$$(5.3.4) \quad f(x) = \frac{1}{\Gamma(\frac{r}{2})2^{r/2}} x^{(r/2)-1} e^{-x/2}, \quad 0 < x < \infty,$$

mikä on  $\chi^2$ -jakauman tiheysfunktio vapausastein  $r$ . Jos  $X$  noudattaa  $\chi^2$ -jakaumaa vapausastein  $r$ , merkitään  $X \sim \text{Khi2}(r)$ .  $\chi^2$ -jakauman keskiarvo, varianssi ja momenttifunktio saadaan nyt suoraan gammajakauman avulla. Jos  $X \sim \text{Khi2}(r)$ , niin

$$E(X) = r, \quad \text{Var}(X) = 2r$$

ja

$$M(t) = (1 - 2t)^{-r/2}, \quad t < \frac{1}{2}.$$

### Odotusaika Poissonin prosessissa

Seuraavan tapahtuman odotusaika Poissonin prosessissa noudattaa eksponettijakaumaa. Olkoon  $W$  nyt odotusaika, kunnes sattuu  $\alpha$  tapahtumaa, missä  $\alpha$  on siis positiivinen kokonaisluku. Jos Poissonin prosessin intensiteetti on  $\lambda$ , niin todennäköisyys, että  $w$ :n pituisella aikavälillä sattuu  $x$  tapahtumaa, saadaan kaavalla (5.2.3):

$$P(X_w = x) = e^{-\lambda w} \frac{(\lambda w)^x}{x!}.$$



Odotusajan  $W$  kertymäfunktio, kun  $W \geq 0$ , on

$$\begin{aligned} F(w) &= P(W \leq w) = 1 - P(W > w) \\ &= 1 - P(\text{vähemmän kuin } \alpha \text{ tapahtumaa välillä } [t, t + w]) \\ &= 1 - \sum_{x=0}^{\alpha-1} e^{-\lambda w} \frac{(\lambda w)^x}{x!}, \end{aligned}$$

koska tapahtumien lukumäärä aikavälillä  $[t, t + w]$  noudattaa Poissonin jakaumaa keskiarvolla  $\lambda w$  [ks. (5.2.3)]. Laskemalla derivaatta  $F'(w) = f(w)$  saadaan tiheysfunktio

$$f(w) = \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} e^{-\lambda w}.$$

Jos  $w < 0$ , niin  $F(w) = 0$  ja  $f(w) = 0$ . Nyt huomaamme, että

$$W \sim \text{Gamma}\left(\alpha, \frac{1}{\lambda}\right).$$

## 5.4 Normaalijakauma

### 5.4.1 Standardimuotoinen normaalijakauma

Tarkastelemme nyt todennäköisyysteorian ja tilastotieteen tärkeintä jakaumaa, normaalijakaumaa. Olkoon  $Z$  jatkuva satunnaismuuttuja, jonka tiheysfunktio on

$$(5.4.1) \quad f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty.$$

Silloin  $Z$  noudattaa standardimuotoista normaalijakaumaa. Käytetään myös sanontaa ” $Z$  noudattaa standardoitua normaalijakaumaa”.

Tarkistamme nyt, että (5.4.1) on todellakin tiheysfunktio. Koska  $f(z) > 0$ , pitää vain osoittaa, että

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 1.$$

Osoitamme siis, että

$$(5.4.2) \quad \int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}.$$

Emme pysty suoraan integroimaan funktiota  $e^{-z^2/2}$ , koska sen integraalifunktio ei ole lausuttavissa suljetussa muodossa. Osoittautuu kuitenkin, että integraalin (5.4.2) neliö on helppo laskea.

Integraalin arvo ei muutu, jos integrointimuuttuja nimetään uudelleen, joten

$$I = \int_{-\infty}^{\infty} e^{-z^2/2} dz = \int_{-\infty}^{\infty} e^{-x^2/2} dx = \int_{-\infty}^{\infty} e^{-y^2/2} dy.$$

Riittää osoittaa, että  $I^2 = 2\pi$ . Nyt

$$\begin{aligned} I^2 &= \left( \int_{-\infty}^{\infty} e^{-x^2/2} dx \right) \left( \int_{-\infty}^{\infty} e^{-y^2/2} dy \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy = \left( \int_0^{\infty} r e^{-r^2/2} dr \right) \left( \int_0^{2\pi} d\theta \right) \\ &= 2\pi \int_0^{\infty} e^{-u} du = 2\pi. \end{aligned}$$

Näin siis tulos (5.4.2) pitää paikkansa. Edellä kolmas yhtäsuuruus saadaan siirtymällä napakoordinaatteihin:

$$x = r \cos \theta \quad \text{ja} \quad y = r \sin \theta.$$

Silloin  $x^2 + y^2 = r^2$ ,  $dx dy = r d\theta dr$  ja integrointirajat ovat  $0 < r < \infty$ ,  $0 < \theta < 2\pi$ .

Integraalilla (5.4.2) on myös läheinen yhteys gammafunktioon. Koska integraalissa (5.4.2) integroitava on symmetrinen nollan suhteen, niin integraalit yli välien  $(-\infty, 0)$  ja  $(0, \infty)$  ovat yhtä suuret. Siksi

$$(5.4.3) \quad \int_0^{\infty} e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}}.$$

Tekemällä sijoitus  $x = \frac{1}{2}z^2$  integraaliin (5.4.3) saadaan integraali, joka on  $\Gamma\left(\frac{1}{2}\right)$ . Silloin

$$(5.4.4) \quad \Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} x^{-1/2} e^{-x} dx = \sqrt{\pi}.$$

**Lause 5.2** *Oletetaan, että  $Z$  noudattaa standardoitua normaalijakaumaa. Silloin*

1.  $Z$ :n momenttifunktio on

$$M(t) = e^{t^2/2}, \quad -\infty < t < \infty.$$

2.  $E(Z) = 0$  ja  $\text{Var}(Z) = 1$ .

**Todistus.** 1. Määritelmän mukaan

$$M(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Tehdään sijoitus  $x = z - t$ . Silloin  $dz = dx$  ja  $e^{tz} e^{-z^2/2} = e^{(t^2-x^2)/2}$ , joten

$$M(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(t^2-x^2)/2} dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = e^{t^2/2}.$$

Viimeinen yhtäsuuruus seuraa siitä, että integraali yli normaalijakauman tiheysfunktion  $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$  on 1.

2. Koska  $M(t) = e^{t^2/2}$ , niin  $M'(t) = te^{t^2/2}$  ja  $M''(t) = e^{t^2/2} + t^2 e^{t^2/2}$ . Silloin  $M'(0) = 0$ ,  $M''(0) = 1$  ja  $\text{Var}(Z) = M''(0) - [M'(0)]^2 = 1$ .  $\square$

Merkitään  $Z \sim N(0, 1)$ , missä siis  $E(Z) = 0$  ja  $\text{Var}(Z) = 1$ . Seuraavassa pykälässä määritellään normaalijakauma, jonka keskiarvo on  $\mu$  ja varianssi  $\sigma^2$ .

## 5.4.2 Yleinen normaalijakauma

Satunnaismuuttuja  $X$  noudattaa normaalijakaumaa keskiarvolla  $\mu$  ja varianssilla  $\sigma^2 > 0$ , jos se voidaan esittää muodossa

$$X = \mu + \sigma Z,$$

missä  $Z \sim N(0, 1)$ . Silloin merkitään  $X \sim N(\mu, \sigma^2)$ . Jos  $X \sim N(\mu, \sigma^2)$ , niin vastaavasti

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Seuraavassa lauseessa esitetään jakaumaa koskevat perustulokset.

**Lause 5.3** Jos  $X \sim N(\mu, \sigma^2)$ , niin

1.  $E(X) = \mu$ ,  $\text{Var}(X) = \sigma^2$  ja

2.

$$M_X(t) = E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}, \quad -\infty < t < \infty.$$

3.  $X$ :n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

**Todistus.** 1. Koska  $X \sim N(\mu, \sigma^2)$ , niin  $X = \mu + \sigma Z$ , missä  $Z \sim N(0, 1)$ . Silloin

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

ja

$$\text{Var}(X) = \text{Var}(\mu + \sigma Z) = \sigma^2 \text{Var}(Z) = \sigma^2.$$

2. Määritelmän mukaan (ks. myös Lause 3.14)

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E[e^{t(\mu + \sigma Z)}] = e^{t\mu} E(e^{t\sigma Z}) \\ &= e^{t\mu} M_Z(t\sigma) = e^{t\mu} e^{t^2\sigma^2/2} = e^{t\mu + t^2\sigma^2/2}. \end{aligned}$$

3. Tehdään muunnos  $x = h(z) = \mu + \sigma z$ . Silloin  $h$ :lla on käänteisfunktio  $g$  ja  $z = g(x) = \frac{x-\mu}{\sigma}$  sekä  $g'(x) = \frac{1}{\sigma}$ . Alaluvussa 5.5 esitettävän muunnostekniikan avulla saadaan  $X$ :n tiheysfunktioiksi

$$\begin{aligned} f_X(x) &= f_Z\left(\frac{x-\mu}{|\sigma|}\right) \frac{1}{|\sigma|} \\ (5.4.5) \quad &= \frac{1}{\sqrt{2\pi}|\sigma|} e^{-(x-\mu)^2/2\sigma^2}. \end{aligned}$$

Tavallisesti tiheysfunktio kirjoitetaan muodossa

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

missä

$$\sigma = +\sqrt{\text{Var}(X)} = +\sqrt{\sigma^2}$$

on  $X$ :n hajonta. Todistuksessa ei oletettu, että  $\sigma > 0$ . □

**Esimerkki 5.9** Jos  $X$ :n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{32\pi}} e^{-(x+7)^2/32}, \quad -\infty < x < \infty,$$

niin  $X \sim N(-7, 16)$  ja

$$M_X(t) = e^{-7t+8t^2}.$$

□

**Esimerkki 5.10** Jos  $X$ :n momenttifunktio on

$$M_X(t) = e^{5t+12t^2},$$

niin  $X \sim N(5, 24)$  ja  $X$ :n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{48\pi}} e^{-(x-5)^2/48}, \quad -\infty < x < \infty.$$

□

Jos  $X \sim N(\mu, \sigma^2)$ , niin  $X$ :n tiheysfunktio saavuttaa maksimin pisteessä  $x = \mu$  ja käänneispisteet ovat  $x = \mu \pm \sigma$ . Todennäköisyysmassa on jakautunut siten, että

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = 0.6826, \\ P(|X - \mu| \leq 2\sigma) &= P(|Z| \leq 2) = 0.9544, \\ P(|X - \mu| \leq 3\sigma) &= P(|Z| \leq 3) = 0.9974, \end{aligned}$$

missä  $Z \sim N(0, 1)$ . Esimerkiksi

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = P(-1 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-1) = 0.8413447 - 0.1586553 = 0.6826895, \end{aligned}$$

missä

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv$$

on standardimuotoisen normaalijakauman kertymäfunktio. Sen arvot on taulukoitu ja se saadaan laskettua useilla ohjelmistoilla. Edellä esitettyjen todennäköisyyksien kahden numeron likiarvoina käytetään tavallisesti lukuja 0.68, 0.95 ja 0.99, jotka eivät ole pyöristettyjä vaan katkaistuja arvoja. Myös yllä esitetyt neljän numeron likiarvot ovat katkaistuja arvoja.

#### Lause 5.4

1. Olkoon  $X \sim N(\mu, \sigma^2)$  ja  $U = aX + b$ , missä  $a \neq 0$  ja  $b$  ovat annettuja vakioita. Silloin

$$U \sim N(a\mu + b, a^2\sigma^2).$$

2. Olkoot  $X_1, X_2, \dots, X_n$  riippumattomat,  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$  ja  $a_1, a_2, \dots, a_n, b$  ovat annetut vakiot, joista ainakin yksi  $a_i$  poikkeaa nolasta. Silloin  $Y = \sum_{i=1}^n a_i X_i + b$  noudattaa normaalijakaumaa

$$Y \sim N\left(\sum_{i=1}^n a_i \mu_i + b, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

**Esimerkki 5.11** Riippumattomat satunnaismuuttujat  $X_1, X_2, X_3$  noudattavat normaalijakaumaa siten, että  $X_i \sim N(2^i, i^i)$ ,  $i = 1, 2, 3$ . Silloin  $Y = X_1 + X_2 + X_3 \sim N(14, 32)$ , sillä

$$E(Y) = 2 + 2^2 + 2^3 = 14 \quad \text{ja} \quad \text{Var}(Y) = 1 + 2^2 + 3^3 = 32$$

ja Lauseen 5.4 mukaan  $Y$  noudattaa normaalijakaumaa. Satunnaismuuttuja  $Y = X_1 + 2X_2 + 3X_3 \sim N(34, 260)$ , koska

$$E(Y) = 2 + 2 \cdot 2^2 + 3 \cdot 2^3 = 34$$

ja

$$\text{Var}(Y) = 1 + 2^2 \cdot 2^2 + 3^2 \cdot 3^3 = 260.$$

□

## 5.5 Muuttujien vaihto

Oletetaan, että  $X$  on jatkuva satunnaismuuttuja, jonka kertymäfunktio on  $F(x)$ . Lukuisissa sovelluksissa tarvitaan satunnaismuuttujan  $X$  jonkin funktion  $Y = h(X)$  jakaumaa, kun  $X$ :n jakauma tunnetaan. Tehtävänä on nyt siis määrittää satunnaismuuttujan  $Y = h(X)$  jakauma, missä  $h(x)$  on  $x$ :n reaaliarvoinen funktio.

### 5.5.1 Muunnos kertymäfunktio avulla

Voimme pyrkiä johtamaan  $Y$ :n kertymäfunktion

$$G(y) = P(Y \leq y)$$

suoraan  $X$ :n kertymäfunktion  $F(x)$  avulla.  $Y$ :n tiheysfunktio  $g(y)$  voidaan määrittää sitten identiteetin (5.1.3) avulla, kun  $G(y)$  on derivoituva.

**Esimerkki 5.12** Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio on

$$f(x) = \frac{3x^2}{2}, \quad -1 \leq x \leq 1.$$

Tarkastellaan satunnaismuuttujan  $Y = X^2$  jakaumaa. Silloin  $Y$ :n arvoavaruus on  $S_Y = [0, 1]$  ja  $Y$ :n kertymäfunktio on

$$\begin{aligned} G(y) &= P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \geq X \geq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{3x^2}{2} dx = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{x^3}{2} = y^{3/2}, \quad 0 \leq y \leq 1. \end{aligned}$$

Derivoimalla saadaan  $Y$ :n tiheysfunktioiksi

$$g(y) = G'(y) = \frac{3y^{1/2}}{2}, \quad 0 \leq y \leq 1.$$

□

**Esimerkki 5.13** Olkoon  $X$  jatkuva satunnaismuuttuja, jonka kertymäfunktio on

$$F(x) = 1 - (1 + x)e^{-x}, \quad x > 0.$$

Johdetaan satunnaismuuttujan  $Y = e^{-X}$  jakauma. Merkitään  $Y$ :n kertymäfunktioita  $G$ :llä. Silloin

$$\begin{aligned} G(y) &= P(Y \leq y) = P(e^{-X} \leq y) = P[-X \leq \log(y)] \\ &= P[X \geq -\log(y)] = 1 - P[X < -\log(y)] \\ &= 1 - F[-\log(y)], \end{aligned}$$

missä  $F(x)$  on  $X$ :n kertymäfunktio. Sijoittamalla  $x = -\log(y)$   $X$ :n kertymäfunktioon saadaan

$$G(y) = [1 - \log(y)]e^{\log(y)} = [1 - \log(y)]y.$$

Koska  $S_X = (0, \infty)$ , niin  $S_Y = (0, 1)$ .  $Y$  on jatkuva satunnaismuuttuja, koska  $G(y)$  on jatkuva ja sillä on jatkuva derivaatta muualla paitsi pisteessä  $y = 0$ .  $Y$ :n tiheysfunktio on

$$g(y) = G'(y) = \begin{cases} -\log(y), & \text{kun } 0 < y < 1; \\ 0 & \text{muualla.} \end{cases}$$

Huomaa, että  $-\log(y) > 0$ , kun  $0 < y < 1$ . Nyt siis  $g(y) \geq 0$  kaikilla  $y \in S_Y = (0, 1)$ .  $\square$

## 5.5.2 Muunnos tiheysfunktion avulla

Seuraavaksi esitetään yleinen menetelmä, jonka avulla voidaan johtaa satunnaismuuttujan  $X$  funktion  $Y = h(X)$  tiheysfunktio suoraan  $X$ :n tiheysfunktion  $f_X(x)$  avulla. Menetelmän edellyttää kuitenkin, että funktiolla  $h(x)$  on tarkasteltavalla välillä *käänteisfunktio*. Esimerkiksi funktion  $y = e^x$  käänteisfunktio on  $x = \log(y)$ . Myös funktio  $y = x^2$  on *kääntävä*, kun  $x > 0$ , sillä silloin  $x = \sqrt{y}$ . Funktio  $y = x^2$  ei ole kääntävä koko reaaliakselilla, koska silloin  $x = \pm\sqrt{y}$ , joka ei ole funktio. Huomattakoon, että jatkuva funktio  $h(x)$  on *kääntävä*, jos ja vain jos se on joko aidosti kasvava tai aidosti vähenevä.

### Lineaarinen muunnos

Tarkastellaan ensin yksinkertaista lineaarista muunnosta  $Y = aX + b$ , missä  $a$  ja  $b$  ovat annettuja vakioita. Nyt siis  $h(X) = aX + b$ . Funktion  $y = h(x)$  derivaatta on

$$\frac{dy}{dx} = h'(x) = a.$$

Funktiolla  $h(x)$  on käänteisfunktio

$$g(y) = \frac{y - b}{a}, \quad a \neq 0$$

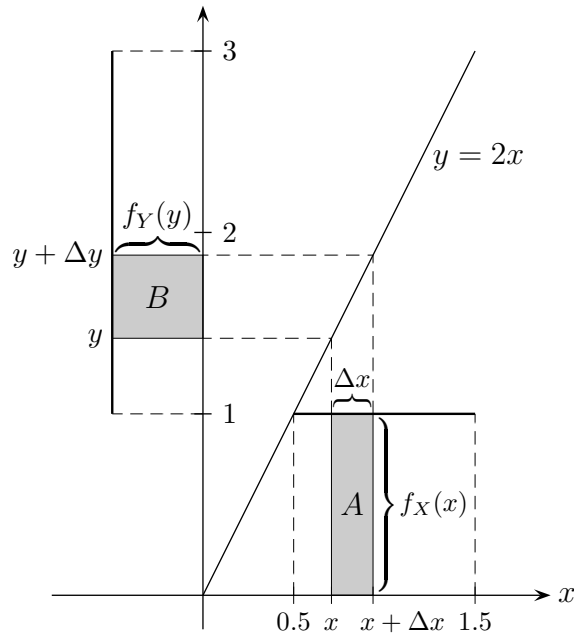
ja

$$\frac{dy}{dx} = g'(y) = \frac{1}{a}.$$

**Esimerkki 5.14** Oletetaan, että  $X \sim \text{Tas}(0.5, 1.5)$  ja  $Y = 2X$ . Mitä jakaumaa  $Y$  noudattaa?

Kuviossa 5.8 on alueen  $A$  pinta-ala

$$P[X \in (x, x + \Delta x)] = f_X(x) \cdot \Delta x = \Delta x$$



**Kuvio 5.8.** Tasajakaumaa  $\text{Tas}(0.5, 1.5)$  noudattavan satunnaismuuttujan  $X$  lineaarinen muunnos.

ja alueen  $B$  pinta-ala

$$P[Y \in (y, y + \Delta y)] = f_Y(y) \cdot \Delta y.$$

Tapahtumat  $X \in (x, x + \Delta x)$  ja  $Y \in (y, y + \Delta y)$  sattuvat täsmälleen samanaikaisesti, joten

$$(5.5.1) \quad P[X \in (x, x + \Delta x)] = P[Y \in (y, y + \Delta y)].$$

Koska  $y = 2x$  ja  $y + \Delta y = 2(x + \Delta x)$ , niin  $\Delta y = 2\Delta x$  ja identiteetistä (5.5.1) seuraa, että  $f_Y(y) = \frac{1}{2}$ . Koska  $0.5 < x < 1.5$ , niin  $1 < y < 3$ . Näin siis  $Y \sim \text{Tas}(1, 3)$ :

$$f_Y(y) = \begin{cases} \frac{1}{2}, & 1 < y < 3; \\ 0, & \text{muualla.} \end{cases}$$

□

Olkoon  $X$  satunnaismuuttuja, jonka arvoavaruus on  $S_X$ . Silloin satunnaismuuttujan  $Y = h(X)$  arvoavaruus  $S_Y$  määräytyy siten, että

$$X \in S_X \Leftrightarrow Y \in S_Y.$$

Seuraavassa lauseessa esitettävässä menetelmässä oletetaan, että funktio  $y = h(x)$  on tarkasteltavalla arvoalueella kääntyvä. Silloin on olemassa sellainen funktio  $x = g(y)$ , että

$$y = h(x) \Leftrightarrow x = g(y).$$



**Lause 5.5** Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio on  $f_X(x)$  ja arvoavaruus  $S_X$ . Olkoon  $Y = h(X)$  sellainen funktio, että sillä on käänteisfunktio  $x = g(y)$  ja käänteisfunktion derivaatta  $g'(y)$  on olemassa kaikilla  $y \in S_Y$ , missä  $S_Y$  on  $Y$ :n arvoavaruus. Silloin  $Y$ :n tiheysfunktio on

$$f_Y(y) = f_X(g(y))|g'(y)|, \quad y \in S_Y.$$

**Todistus.** Oletuksen mukaan  $g(y)$  on derivoituva, joten se on jatkuva. Koska  $h$  ja  $g$  ovat kääntyviä, niin  $h$  ja  $g$  ovat molemmat joko kasvavia tai väheneviä. Oletetaan  $h$  ja  $g$  ovat väheneviä. Silloin

$$F_Y(y) = P(Y \leq y) = P(h(X) \leq y) = P(X \geq g(y)) = 1 - F_X(g(y)).$$

Derivoidaan  $1 - F_X[g(y)]$  ketjusäännön avulla, jolloin saadaan

$$\begin{aligned} f_Y(y) &= F_Y'(y) = -F_X'(g(y))g'(y) \\ &= -f_X(g(y))g'(y) = f_X(g(y))|g'(y)|. \end{aligned}$$

Viimeinen yhtäsuuruus seuraa siitä, että  $g'(y)$  on negatiivinen, koska  $g$  on vähevä.

Jos  $h$  ja  $g$  ovat kasvavia, niin todistus on melkein samanlainen ja se jätetään harjoitustehtäväksi.  $\square$

**Esimerkki 5.15** Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio on  $f_X(x) = e^{-x}$  ja  $S_X = \{x \mid x > 0\}$ . Olkoon  $Y = X^{1/2}$ , joten  $X = Y^2 = g(Y)$  ja  $S_Y = S_X$ . Koska  $g'(y) = 2y$ , niin

$$f_Y(y) = f_X(y^2)|2y| = 2ye^{-y^2}, \quad y > 0.$$

Tarkastellaan vielä satunnaismuuttujaa  $V = e^{-X}$ . Silloin  $X = -\log(V)$ . Merkitään nyt  $-\log(V) = \tilde{g}(V)$ . Silloin  $S_V = [0, 1]$  ja  $\tilde{g}'(v) = -1/v$ . Siksi

$$f_V(v) = f_X[-\log(v)]\left|\frac{-1}{v}\right| = \frac{v}{v} = 1,$$

joten  $V$  noudattaa tasajakaumaa välillä  $[0, 1]$ .  $\square$

Mikäli muunnosfunktiolla  $h$  ei ole käänteisfunktiota  $X$ :n arvoavaruudessa  $S_X$ , niin Lauseen 5.5 muunnosmenetelmää ei voi suoraan soveltaa. Jos kuitenkin on olemassa sellainen  $S_X$ :n ositus yhteispisteettömiin osaväleihin  $A_1, A_2, \dots, A_m$ , että

$$(5.5.2) \quad S_X = A_1 \cup A_2 \cup \dots \cup A_m$$

ja  $h$  on kääntyvä jokaisella osavälillä, voidaan muunnos tehdä jokaisella osavälillä erikseen. Sitä varten määritellään funktiot

$$h(x) = \begin{cases} h_i(x), & \text{kun } x \in A_i; \\ 0 & \text{muualla.} \end{cases}$$

Silloin  $h(x)$  voidaan kirjoittaa muodossa  $h(x) = \sum_{i=1}^m h_i(x)$ , missä jokainen  $h_i(x)$  on kääntyvä välillä  $A_i$ . Olkoot funktioiden  $h_i$  käänteisfunktiot vastaavasti  $g_i$ ,  $i = 1, 2, \dots, m$ . Satunnaismuuttujan  $Y = h(X)$  tiheysfunktio voidaan nyt esittää Lauseen 5.5 avulla muodossa

$$(5.5.3) \quad f_Y(y) = \sum_{i=1}^m f_X(g_i(y)) |g_i'(y)|. \quad y \in S_Y.$$

Huomattakoon, että joskus tarvitaan äärellisen osituksen (5.5.2) sijasta ositus, jossa jakovälejä  $A_1, A_2, \dots$  on ääretön määrä ( $m = \infty$ ).

### 5.5.3 Normaalimuuttujan muunnokset

Jos  $X \sim N(0, 1)$ , niin  $X$ :n tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty,$$

joka on standardimuotoisen normaalijakauman tiheysfunktio. Johdetaan nyt satunnaismuuttujan  $U = X^2$  jakauma. Muunnosfunktio  $u = h(x) = x^2$  ei ole kääntyvä, koska  $x = \pm\sqrt{u}$  ei ole funktio. Siksi esitämme arvoavaruuden  $S_X = \{-\infty < x < \infty\}$  ositettuna muodossa

$$S_X = (-\infty, 0] \cup (0, \infty).$$

Silloin funktiolla  $h(x)$  on välillä  $(-\infty, 0]$  käänteisfunktio  $g_1(u) = -\sqrt{u}$  ja välillä  $(0, \infty)$  käänteisfunktio  $g_2(u) = \sqrt{u}$ . Nyt siis kaavan (5.5.3) mukaan  $U$ :n tiheysfunktio on

$$(5.5.4) \quad f_U(u) = f_X(-\sqrt{u}) \frac{1}{2\sqrt{u}} + f_X(\sqrt{u}) \frac{1}{2\sqrt{u}} = \frac{1}{\sqrt{2\pi u}} e^{-u/2},$$

kun  $u \in (-\infty, \infty)$ .  $U$  noudattaa  $\chi^2$ -jakaumaa vapausastein 1. Käsittelemme tilastotieteessä tärkeää  $\chi^2$ -jakaumaa vielä jatkossa tarkemmin.

**Lause 5.6** Jos  $X \sim N(\mu, \sigma^2)$ ,  $\sigma^2 > 0$ , niin silloin

$$\frac{(X - \mu)^2}{\sigma^2} \sim \text{Khi2}(1).$$

**Todistus.** Koska  $X \sim N(\mu, \sigma^2)$ , niin määritelmän mukaan  $\frac{X-\mu}{\sigma} = Z \sim N(0, 1)$ . Edellä näytettiin, että  $Z^2 \sim \text{Khi2}(1)$ . Näin on lause todistettu.  $\square$

**Lause 5.7** Jos  $Z_i$ :t ovat riippumattomat ja  $Z_i \sim N(0, 1)$ ,  $i = 1, 2, \dots, n$ , niin

$$Z_1^2 + Z_2^2 + \dots + Z_n^2 \sim \text{Khi2}(n).$$

Jos tehdään otos normaali jakaumasta  $N(0, 1)$ , niin Lauseen 5.7 mukaan havaintojen neliösumma noudattaa Khi2-jakaumaa vapausastein  $n$ , missä  $n$  on otoskoko.

**Seuraus 5.1** Jos  $X_i$ :t ovat riippumattomat ja  $X_i \sim N(\mu, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , niin

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \text{Khi2}(n).$$

Jos vastaavasti tehdään  $n$ :n suuruinen otos normaali jakaumasta  $N(\mu, \sigma^2)$ , niin Seurauslauseen 5.1 mukaan standardoitujen havaintojen neliösumma noudattaa Khi2-jakaumaa vapausastein  $n$ .

**Lause 5.8** Olkoot  $X_1$  ja  $X_2$  riippumattomat ja  $X_i \sim \text{Khi2}(n_i)$ ,  $i = 1, 2$ . Silloin

$$X_1 + X_2 \sim \text{Khi2}(n_1 + n_2).$$

## 5.6 Satunnaismuuttujan funktion odotusarvo

Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio  $f(x)$  on määritelty arvoavaruudessa  $S$ . Olkoon  $h(X)$  satunnaismuuttujan  $X$  reaaliarvoinen funktio, joka siis määrittelee uuden satunnaismuuttujan.

**Määritelmä 5.3** Jos  $X$  on jatkuva satunnaismuuttuja, niin satunnaismuuttujan  $h(X)$  odotusarvo on

$$(5.6.1) \quad E[h(X)] = \int_S h(x)f(x) dx,$$

mikäli  $E(|h(X)|) < \infty$ . Jos  $E(|h(X)|) = \infty$ , niin sanomme, että  $E[h(X)]$  ei ole olemassa.

**Huomautus 5.1** Odotusarvon  $E[h(X)]$  olemassaolo tarkoittaa siis sitä, että funktion  $|h(X)|$  odotusarvo on äärellinen. Jos  $X$  noudattaa esimerkiksi eksponenttijakaumaa keskiarvolla 1, niin  $f(x) = e^{-x}$  ja  $S = [0, \infty)$ . Silloin  $X$ :n odotusarvo on

$$\begin{aligned} E(X) &= \int_0^{\infty} xe^{-x} dx \\ &= \int_0^{\infty} (-xe^{-x}) + \int_0^{\infty} e^{-x} dx \quad (\text{osittaisintegrointi}) \\ &= \int_0^{\infty} e^{-x} dx = 1, \end{aligned}$$

joten odotusarvo on olemassa. Hyvin usein odotusarvot ovat epäoleellisia integraaleja, niin kuin tässäkin esimerkissä.

Jos  $h(X)$  integroituu itseisesti, eli

$$\int_S |h(x)|$$

on äärellisenä olemassa, niin  $E[h(X)]$  on olemassa. Funktio  $V = h(X)$  on satunnaismuuttuja, jonka tiheysfunktio  $g(v)$  on määritelty arvoavaruudessa  $S_V = \{v \mid v = h(x), x \in S\}$ . Silloin

$$E[h(X)] = E(V) = \int_{S_V} vg(v).$$

**Esimerkki 5.16** Tarkastellaan nyt *Cauchyn jakaumaa* noudattavaa satunnaismuuttujaa  $X$ , jonka tiheysfunktio on

$$(5.6.2) \quad f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

Kaava (5.6.2) todellakin määrittelee tiheysfunktion, koska

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = \frac{2}{\pi} \int_0^{\infty} \arctan(x) = \frac{2}{\pi} \cdot \frac{\pi}{2} = 1.$$

Osoitamme nyt, että  $E(|X|) = \infty$ , mistä seuraa, että *Cauchyn jakaumalla ei ole keskiarvoa*. Symmetrian nojalla voidaan kirjoittaa

$$E(|X|) = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx.$$

Jokaista reaalilukua  $M > 0$  kohti saadaan

$$\int_0^M \frac{x}{(1+x^2)} dx = \int_0^M \frac{\log(1+x^2)}{2} = \frac{\log(1+M^2)}{2}.$$

Tästä seuraa, että

$$E(|X|) = \lim_{M \rightarrow \infty} \frac{2}{\pi} \int_0^M \frac{x}{1+x^2} dx = \frac{1}{\pi} \lim_{M \rightarrow \infty} \log(1+M^2) = \infty,$$

joten  $E(X)$  ei ole olemassa. □

**Taulukko 5.1.** Tärkeitä odotusarvoja.

$h(x)$	$E[h(X)]$	Merkintä	Nimitys
$x$	$E(X)$	$\mu$	odotusarvo
$x^r$	$E(X^r)$	$\alpha_r$	$r$ . momentti
$x^{(r)}$	$E[X^{(r)}]$	$g_r$	$r$ . tekijämomentti
$(x - \mu)^2$	$E[(X - \mu)^2]$	$\sigma^2$	varianssi
$(x - \mu)^r$	$E[(X - \mu)^r]$	$\mu_r$	$r$ . keskusmomentti

### 5.6.1 Momentifunktio ja momentit

Kun  $h(X) = X^r$ , niin  $E[h(X)] = E(X^r)$  on  $X$ :n  $r$ . momentti. Jatkuvien satunnaismuuttujien momentit määritellään vastaavasti kuin diskreettien satunnaismuuttujien momentit. Summalausekkeet vain korvataan integraaleilla. Taulukossa 5.1 esitetään yhteenveto eri momenteista

Momenttifunktio määriteltiin 3. luvussa (Määritelmä 3.12) ja jatkuville satunnaismuuttujille alaluvussa 5.1 [ks. identiteetti (5.1.5)]. Jatkuvan satunnaismuuttujan  $X$  momentifunktio on

$$M(t) = E(e^{tX}) = \int_S e^{tx} f(x) dx, \quad t \in A,$$

missä  $f(x)$  on  $X$ :n tiheysfunktio ja  $A$  sellainen  $t$ :n arvojen joukko, että  $M(t)$  on äärellinen kaikilla  $t \in A$ . Koska  $M(0) = 1$ , niin  $0 \in A$ . Sanomme, että  $M(t)$  on olemassa, jos  $(-a, a) \subset A$  jollakin  $a > 0$ . Momenttifunktion perusominaisuudet esitettiin Pykälässä 3.5.2.

**Esimerkki 5.17** Huomautuksessa 5.1 laskettiin odotusarvo  $E(X)$ , kun  $X \sim \text{Exp}(1)$ . Silloin  $X$ :n tiheysfunktio on  $f(x) = e^{-x} \geq 0$  välillä  $S = [0, \infty)$  ja  $f(x) = 0$  muualla. Kaikki momentit  $E(X^r)$  voidaan määrittää osittaisintegroinnilla, mutta käytetään nyt momenttifunktiota, joka on

$$M(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} e^{-x} dx = \frac{1}{1-t}, \quad t < 1.$$

Derivoimalla  $M(t)$  toistuvasti  $r$  kertaa saadaan  $M^{(r)}(t) = \frac{r!}{(1-t)^{r+1}}$ . Siksi

$$E(X^r) = M^{(r)}(0) = r!,$$

joten

$$\mu = E(X) = 1, \quad E(X^2) = 2, \quad \sigma^2 = E(X^2) - \mu^2 = 1.$$

□

Erityisesti keskiarvo  $\mu$ , varianssi  $\sigma^2$  ja hajonta  $\sigma = \sqrt{\text{Var}(X)}$  ovat tavallimmat tunnusluvut, joilla jakaumaa luonnehditaan. Jakauman yksityiskohteisemmassa tarkastelussa voidaan käyttää myös korkeampia momenteja, mikäli ne ovat olemassa.

## Vinous ja huipukkuus

Satunnaismuuttujan 1. momentti  $\mu$  määrittää jakauman sijainnin. Keskitetyn muuttujan  $X - \mu$  toinen momentti (keskusmomentti) on varianssi  $\sigma^2$  ja se mittaa todennäköisyysmassan hajaantumista. Normeeratun muuttujan  $(X - \mu)/\sigma$  kolmas ja neljäs momentti luonnehtivat jakauman muotoa.

Jakauman *vinouskerroin*, josta käytetään merkintää  $\gamma_1$ , määritellään seuraavasti:

$$(5.6.3) \quad \gamma_1 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mu_3}{\sigma^3},$$

missä  $\mu_3$  on jakauman 3. keskusmomentti ja  $\sigma = \sqrt{\text{Var}(X)}$  on hajonta. Olkoon  $X$ :n tiheysfunktio  $f(x)$ . Silloin  $X$ :n jakauma on *symmetrinen pisteen*  $a$  suhteen, jos

$$f(a - x) = f[-(a - x)]$$

kaikilla  $x$ :n arvoilla. Jos  $E(X)$  on olemassa, niin silloin  $E(X) = a$ . Symmetrisen jakauman vinouskerroin on nolla. Jos jakaumalla on pitkä häntä oikealle, kuten Poissonin jakaumalla ja geometrisella jakaumalla, niin jakauma on positiivisesti vino ja  $\gamma_1 > 0$ . Jos jakaumalla on pitkä häntä vasemmalle, niin  $\gamma_1 < 0$ . Jakaumalla on tietysti oltava 3. momentti, jotta vinouskerroin voidaan laskea. Huomaa, että Cauchyn jakauma, jonka tiheysfunktio on

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty,$$

on symmetrinen pisteen  $a = 0$  suhteen, mutta 0 ei ole jakauman keskiarvo, koska jakaumalla ei ole keskiarvoa (ks. Esimerkki 5.16). Cauchyn jakauman vinouskerrointa ei voida laskea, vaikka määritelmän nojalla voimme todeta jakauman olevan symmetrinen.

*Huipukkuuskerrointa* merkitään  $\gamma_2$  ja se määritellään normeeratun muuttujan 4. momentin avulla seuraavasti:

$$(5.6.4) \quad \gamma_2 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \frac{\mu_4}{\sigma^4} - 3,$$

missä  $\mu_4$  on  $X$ :n 4. keskusmomentti. Normaalijakauman huipukkuus on 0. Jos jakaumalla on paksummat hännät kuin normaalijakaumalla, niin silloin  $\gamma_2 > 0$ . Jos hännät ovat ohuempia kuin normaalijakaumalla, niin  $\gamma_2 < 0$ .

## 5.7 Kaksiulotteiset jakaumat

Tarkastellaan nyt kahden jatkuvan satunnaismuuttujan yhteisjakaumaa. Yleistys usean muuttujan tapaukseen on sen jälkeen suoraviivainen.

**Määritelmä 5.4** Olkoot  $X$  ja  $Y$  samassa otosavaruudessa määritellyt jatkuvat satunnaismuuttujat. Olkoon kaksiulotteisen jatkuvan satunnaismuuttujan  $(X, Y)$  arvoavaruus  $S$ . Funktio  $f(x, y)$  on  $(X, Y)$ :n tiheysfunktio ( $X$ :n ja  $Y$ :n yhteisjakauman tiheysfunktio), jos sillä on seuraavat ominaisuudet:

1.  $f(x, y) \geq 0$  kaikilla  $(x, y) \in \mathbb{R}^2$ ,
2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$  ja

3.

$$P[(X, Y) \in A] = \iint_{(x,y) \in A} f(x, y) dx dy,$$

missä  $(X, Y) \in A$  on tasossa määritelty tapahtuma.

**Esimerkki 5.18** Olkoon  $X$ :n ja  $Y$ :n yhteisjakauman tiheysfunktio

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \quad -1 < x < 1, \quad -1 < y < 1.$$

Määritellään  $A = \{(x, y) \mid 0 < x < 1, 0 < y < x\}$ . Todennäköisyys, että  $(X, Y) \in A$ , on

$$\begin{aligned} P[(X, Y) \in A] &= \int_0^1 \int_0^x \frac{3}{2}x^2(1 - y) dy dx = \int_0^1 \frac{3}{2}x^2 \Big/ \Big( y - \frac{y^2}{2} \Big) dx \\ &= \int_0^1 \frac{3}{2} \left( x^3 - \frac{x^4}{2} \right) dx = \frac{3}{2} \Big/ \left( \frac{x^4}{4} - \frac{x^5}{10} \right) = \frac{9}{40}. \end{aligned}$$

□

### 5.7.1 Reunajakauma ja ehdollinen jakauma

Kaksiulotteista satunnaismuuttujaa  $(X, Y)$  kutsutaan kaksiulotteiseksi satunnaisvektoriksi. Silloin  $X$  ja  $Y$  ovat tietysti (yksiulotteisia) satunnaismuuttujia.  $X$ :n *reunajakauman tiheysfunktio*, jota merkitään  $f_X(x)$ , on pelkästään  $X$ :n tiheysfunktio, jossa  $Y$ :tä ei oteta huomioon. Satunnaismuuttujan  $X$  *ehdollinen tiheysfunktio* ehdolla  $Y = y$  on  $X$ :n tiheysfunktio, kun  $Y$ :n arvo tunnetaan.  $X$ :n ehdollista tiheysfunktioita ehdolla  $Y = y$  merkitään  $f_X(x \mid Y = y)$  tai lyhyesti  $f_X(x \mid y)$ .

**Määritelmä 5.5** Olkoon  $f(x, y)$  jatkuvan satunnaisvektorin  $(X, Y)$  tiheysfunktio ja  $S$  sen arvoavaruus. Silloin satunnaismuuttujat  $X$  ja  $Y$  ovat jatkuvia ja niiden reunajakaumien tiheysfunktiot ovat

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad x \in S_X; \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx, \quad y \in S_Y,$$

missä  $S_X$  on  $X$ :n ja  $S_Y$  on  $Y$ :n arvoavaruus. Satunnaismuuttujat  $X$  ja  $Y$  ovat riippumattomat jos ja vain jos

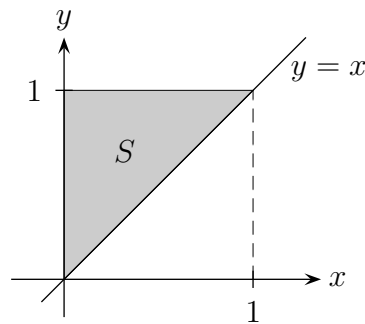
$$(5.7.1) \quad f(x, y) = f_X(x)f_Y(y) \quad \text{kaikilla } x \in S_X \text{ ja } y \in S_Y;$$

muutoin  $X$  ja  $Y$  riippuvat toisistaan.

**Esimerkki 5.19** Olkoon  $X$ :n ja  $Y$ :n yhteisjakauman tiheysfunktio

$$f(x, y) = 2, \quad 0 \leq x \leq y \leq 1,$$

muualla  $f(x, y) = 0$ . Satunnaisvektorin  $(X, Y)$  arvoavaruus on  $S = \{(x, y) \mid 0 \leq x \leq y \leq 1\}$ .



$$S = \{(x, y) \mid 0 \leq x \leq y \leq 1\}$$

**Kuvio 5.9.** Tasajakauman  $f(x, y) = 2$  määrittelyalue  $S$ .

Silloin esimerkiksi todennäköisyys

$$\begin{aligned} P\left(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2}\right) &= P\left(0 \leq X \leq Y, 0 \leq Y \leq \frac{1}{2}\right) \\ &= \int_0^{1/2} \int_0^y 2 \, dy \, dx = \int_0^{1/2} 2y \, dy = \frac{1}{4}. \end{aligned}$$

Reunajakaumien tiheysfunktiot ovat

$$f_X(x) = \int_x^1 2 \, dy = 2(1 - x), \quad 0 \leq x \leq 1,$$

ja

$$f_Y(y) = \int_0^y 2 \, dx = 2y, \quad 0 \leq y \leq 1.$$



Lasketaan vielä  $X$ :n ja  $Y$ :n odotusarvot sekä  $Y$ :n 2. momentti.

$$E(X) = \int_0^1 \int_x^1 2x \, dy \, dx = \int_0^1 2x(1-x) \, dx = \frac{1}{3},$$

$$E(Y) = \int_0^1 \int_0^y 2y \, dx \, dy = \int_0^1 2y^2 \, dy = \frac{2}{3},$$

$$E(Y^2) = \int_0^1 \int_0^y 2y^2 \, dx \, dy = \int_0^1 2y^3 \, dy = \frac{1}{2}.$$

Odotusarvot  $E(X)$ ,  $E(Y)$  ja  $E(Y)$  voidaan laskea joko suoraan reunajakaumasta tai sitten yhteisjakaumasta.  $\square$

Nähdään helposti, että Esimerkissä 5.19 satunnaismuuttujat  $X$  ja  $Y$  eivät ole riippumattomat, koska

$$f_X(x)f_Y(y) = 2(1-x)2y \neq f(x,y) = 2, \quad (x,y) \in S.$$

Sen sijaan voidaan osoittaa, että Esimerkissä 5.18 satunnaismuuttujat  $X$  ja  $Y$  ovat riippumattomat.

Jatkuvan satunnaismuuttujan ehdollinen tiheysfunktio määritellään seuraavasti:

**Määritelmä 5.6** Jos jatkuvan satunnaisvektorin  $(X, Y)$  tiheysfunktio on  $f(x, y)$  ja arvoavaruus  $S$ , niin  $X$ :n ehdollinen tiheysfunktio ehdolla  $Y = y$  on

$$f_X(x | y) = \frac{f(x, y)}{f_Y(y)}, \quad (x, y) \in S$$

ja  $Y$ :n ehdollinen tiheysfunktio ehdolla  $X = x$  on

$$f_Y(y | x) = \frac{f(x, y)}{f_X(x)}, \quad (x, y) \in S.$$

Huomattakoon, että Määritelmässä 5.6 oletetaan, että  $f_Y(y) > 0$  ja  $f_X(x) > 0$ .

**Esimerkki 5.20** Olkoot satunnaismuuttujat  $X$  ja  $Y$  samat kuin Esimerkissä 5.19 Silloin

$$\begin{aligned} f(x, y) &= 2, & 0 \leq x \leq y \leq 1, \\ f_X(x) &= 2(1-x), & 0 \leq x \leq 1, \\ f_Y(y) &= 2y, & 0 \leq y \leq 1. \end{aligned}$$

Määritetään nyt  $Y$ :n ehdollisen jakauman tiheysfunktio, kun  $X = x$  on annettu. Määritelmän 5.6 mukaan

$$f(y | x) = \frac{f(x, y)}{f_X(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad x \leq y \leq 1, \quad 0 \leq x \leq 1.$$

$Y$ :n ehdollinen odotusarvo ehdolla  $X = x$  on

$$E(Y | x) = \int_x^1 y \frac{1}{1-x} dy = \int_x^1 \frac{y^2}{2(1-x)} = \frac{1+x}{2}, \quad 0 \leq x \leq 1.$$

Samalla tavalla voidaan osoittaa, että

$$E(X | y) = \frac{y}{2}, \quad 0 \leq y \leq 1.$$

Suoraan määritelmän perusteella  $Y$ :n ehdollinen varianssi ehdolla  $X = x$  on

$$\begin{aligned} E([Y - E(Y | x)]^2 | x) &= \int_x^1 \left(y - \frac{1+x}{2}\right)^2 \frac{1}{1-x} dy \\ &= \int_x^1 \frac{1}{3(1-x)} \left(y - \frac{1+x}{2}\right)^3 \\ &= \frac{(1-x)^2}{12}. \end{aligned}$$

Jos  $U \sim \text{Tas}(a, b)$ , niin  $E(U) = \frac{a+b}{2}$  ja  $\text{Var}(U) = \frac{(b-a)^2}{12}$ . Koska  $Y$ :n ehdollinen jakauma ehdolla  $X = x$  on  $\text{Tas}(x, 1)$ , niin olisimme voineet tasajakau-  
man ominaisuuksien perusteella suoraan todeta, että

$$E(Y | x) = \frac{x+1}{2} \quad \text{ja} \quad \text{Var}(Y | x) = \frac{(1-x)^2}{12}.$$

Lasketaan vielä ehdollinen todennäköisyys

$$P(3/4 \leq Y \leq 7/8 | X = 1/4) = \int_{3/4}^{7/8} f(y | 1/4) dy = \int_{3/4}^{7/8} \frac{1}{3/4} dy = \frac{1}{6}.$$

□

Havaitsimme edellisessä esimerkissä, että  $Y$ :n ehdollinen odotusarvo on  $x$ :n lineaarinen funktio:

$$E(Y | x) = \frac{1}{2} + \frac{1}{2}x, \quad 0 \leq x \leq 1.$$

Jos  $E(Y | x)$  on lineaarinen, niin pitää yleisesti paikkansa, että

$$E(Y | x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X),$$

missä  $\rho = \text{Cor}(X, Y)$  on  $X$ :n ja  $Y$ :n välinen korrelaatio,  $\sigma_X$  on  $X$ :n hajonta ja  $\sigma_Y$  on  $Y$ :n hajonta. Jos  $E(X | y)$  on lineaarinen, niin

$$E(X | y) = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y).$$

Ehdollisten odotusarvojen  $E(Y | x)$  ja  $E(X | y)$  yhtälöissä kertoimien  $\rho \frac{\sigma_Y}{\sigma_X}$  ja  $\rho \frac{\sigma_X}{\sigma_Y}$  tulo on  $\rho^2$ . Esimerkissä 5.20 näiden kertoimien tulo on  $\rho^2 = \frac{1}{4}$ . Siksi  $\rho = \frac{1}{2}$ , koska molemmat kertoimet ovat positiiviset. Näiden kertoimien suhde on  $\sigma_Y^2 / \sigma_X^2$  ja esimerkissä tämä suhde on 1. Tästä voimme päätellä, että Esimerkissä 5.20  $\sigma_X^2 = \sigma_Y^2$ .

Satunnaismuuttujien  $X$  ja  $Y$  riippumattomuuden tarkistaminen suoraan relaation (5.7.1) perusteella edellyttää reuna-jakaumien tiheysfunktioiden  $f_X(x)$  ja  $f_Y(y)$  tuntemista. Seuraava apulause tekee riippumattomuuden tarkistamisen jonkin verran helpommaksi, koska siinä ei edellytetä reuna-jakaumien tuntemista.

**Apulause 5.1** *Olkoon  $(X, Y)$  kaksiulotteinen satunnaisvektori, jonka yhteisjakauman tiheysfunktio on  $f(x, y)$ . Silloin satunnaismuuttujat  $X$  ja  $Y$  ovat riippumattomat, jos ja vain jos on olemassa sellaiset funktiot  $g(x)$  ja  $h(y)$ , että*

$$f(x, y) = g(x)h(y) \quad \text{kaikilla } x \in \mathbb{R} \text{ ja kaikilla } y \in \mathbb{R},$$

missä  $g$  riippuu vain  $x$ :stä ja  $h$  vain  $y$ :stä.

### Kertymäfunktio

Kaksiulotteinen jakauma voidaan täydellisesti luonnehtia kertymäfunktion avulla. Satunnaisvektorin  $(X, Y)$  yhteisjakauman kertymäfunktio  $F(x, y)$  määritellään relaatiolla

$$F(x, y) = P(X \leq x, Y \leq y)$$

missä  $(x, y) \in \mathbb{R}^2$ . Tiheysfunktion avulla lausuttuna kertymäfunktio on

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) ds dt.$$

Integraalilaskennan peruslause kahden muuttujan tapauksessa sanoo, että

$$(5.7.2) \quad \frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y)$$

kaikissa  $f(x, y)$ :n jatkuvuuspeisteissä. Relaatio (5.7.2) on hyödyllinen silloin, kun kertymäfunktio tunnetaan ja halutaan johtaa tiheysfunktio. Silloin tiheysfunktio  $f(x, y)$  saadaan derivoimalla  $F(x, y)$  sekä  $x$ :n että  $y$ :n suhteen eli laskemalla osittaisderivaatta  $\frac{\partial^2 F(x, y)}{\partial x \partial y}$ .

**Esimerkki 5.21** Olkoon  $X$ :n ja  $Y$ :n yhteisjakauman kertymäfunktio

$$F(x, y) = \begin{cases} xy, & 0 \leq x \leq 1 \text{ ja } 0 \leq y \leq 1; \\ y, & x > 1, 0 \leq y \leq 1; \\ x, & y > 1, 0 \leq x \leq 1; \\ 1, & x > 1 \text{ ja } y > 1; \\ 0, & x < 0 \text{ tai } y < 0. \end{cases}$$

Laskemalla osittaisderivaatta  $\frac{\partial^2 F(x,y)}{\partial x \partial y}$  saadaan

$$f(x,y) = \begin{cases} 1, & 0 \leq x \leq 1, 0 \leq y \leq 1; \\ 0 & \text{muualla.} \end{cases}$$

Satunnaisvektori  $(X, Y)$  noudattaa siis kaksiulotteista tasajakaumaa  $\text{Tas}[(0, 1) \times (0, 1)]$ . Todennäköisyys voidaan lausua kertymäfunktion avulla seuraavasti:

$$\begin{aligned} P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) &= \int_{x_1}^{x_2} \int_{y_1}^{y_2} dy dx = \int_{x_1}^{x_2} \int_{y_1}^{y_2} xy = (x_2 - x_1)(y_2 - y_1) \\ &= x_2 y_2 - x_2 y_1 - x_1 y_2 + x_1 y_1 \\ &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1). \end{aligned}$$

Yleisesti pitää paikkansa, että

$$\begin{aligned} P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) \\ = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1). \end{aligned}$$

Kahden muuttujan tasajakauman  $\text{Tas}[(0, 1) \times (0, 1)]$  tapauksessa todennäköisyys  $P\left(\frac{1}{4} \leq X \leq \frac{1}{2}, \frac{1}{2} \leq Y \leq \frac{3}{4}\right)$  on

$$\begin{aligned} F\left(\frac{1}{2}, \frac{3}{4}\right) - F\left(\frac{1}{2}, \frac{1}{2}\right) - F\left(\frac{1}{4}, \frac{3}{4}\right) + F\left(\frac{1}{4}, \frac{1}{2}\right) \\ = \frac{1}{2} \cdot \frac{3}{4} - \frac{1}{2} \cdot \frac{1}{2} - \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{16}. \end{aligned}$$

□

### 5.7.2 Yhteisjakauman momenttifunktio

Kaksiulotteisen diskreetin satunnaisvektorin momenttifunktio määriteltiin alaluvussa 4.7.2. Jatkuvien satunnaismuuttujien  $X_1$  ja  $X_2$  yhteisjakauman eli jatkuvan satunnaisvektorin  $(X_1, X_2)$  jakauman momenttifunktio määritellään samalla tavalla kuin diskreetissä tapauksessa. Olkoon  $(X_1, X_2)$  jatkuva satunnaisvektori ja  $t_1 X_1 + t_2 X_2$  satunnaismuuttujien  $X_1$  ja  $X_2$  lineaarinen yhdiste, missä  $t_1, t_2 \in \mathbb{R}$ . Satunnaisvektorin  $(X_1, X_2)$  jakauman momenttifunktio on

$$M(t_1, t_2) = E(e^{t_1 X_1 + t_2 X_2}).$$

Jatkuvien satunnaismuuttujien tapauksessa odotusarvon lauseke on muotoa

$$E(e^{t_1 X + t_2 X_2}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2} f(x_1, x_2) dx_1 dx_2.$$

Merkitään

$$M_i(t_1, t_2) = \frac{\partial M(t_1, t_2)}{\partial t_i},$$

$$M_{ii}(t_1, t_2) = \frac{\partial^2 M(t_1, t_2)}{\partial t_i^2},$$

$$M_{ij}(t_1, t_2) = \frac{\partial^2 M(t_1, t_2)}{\partial t_i \partial t_j},$$

missä  $M_i(t_1, t_2)$  on  $M$ :n osittaisderivaatta  $t_i$ :n suhteen,  $M_{ii}(t_1, t_2)$  on  $M$ :n 2. osittaisderivaatta  $t_i$ :n suhteen ja  $M_{ij}(t_1, t_2)$  on osittaisderivaatta  $t_i$ :n ja  $t_j$ :n suhteen ( $i = 1, 2; j = 1, 2$ ). Esitämme nyt seuraavassa lauseessa, miten momenttifunktio generoi satunnaisvektorin momentit.

**Lause 5.9** *Oletetaan, että satunnaisvektorilla  $(X_1, X_2)$  on momenttifunktio. Silloin  $E(X_i)$ ,  $E(X_i^2)$  ja  $E(X_i X_j)$  ovat äärelliset ja*

$$E(X_i) = M_i(0, 0), \quad E(X_i^2) = M_{ii}(0, 0), \quad E(X_i X_j) = M_{ij}(0, 0)$$

kaikilla  $i = 1, 2$  ja  $j = 1, 2$ .

Esimerkiksi  $X_1$ :n odotusarvo saadaan derivoimalla ensin momenttifunktio  $t_1$ :n suhteen ja sijoittamalla sitten derivaatan lausekkeeseen  $t_1 = 0$  ja  $t_2 = 0$ . Sekamomentti  $E(X_1 X_2)$  saadaan määrittämällä toisen kertaluvun osittaisderivaatta  $M_{12}(t_1, t_2)$  (derivoidaan momenttifunktio  $t_2$ :n ja  $t_1$ :n suhteen) ja laskemalla osittaisderivaatan arvo  $M_{ij}(0, 0)$  pisteessä  $(t_1, t_2) = (0, 0)$ .

**Esimerkki 5.22** Jos  $Z_1$  ja  $Z_2$  ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa, niin  $(Z_1, Z_2)$  noudattaa kaksiulotteista standardimuotoista normaalijakaumaa.  $(Z_1, Z_2)$ :n momenttifunktio on

$$M(t_1, t_2) = E(e^{t_1 Z_1 + t_2 Z_2}) = E(e^{t_1 Z_1}) E(e^{t_2 Z_2})$$

$$= e^{t_1^2/2} e^{t_2^2/2} = e^{(t_1^2 + t_2^2)/2}.$$

Tässä tapauksessa  $M_1(t_1, t_2) = t_1 e^{(t_1^2 + t_2^2)/2}$ , joten  $E(X_1) = M_1(0, 0) = 0$ . Vastaavasti  $M_{11} = e^{(t_1^2 + t_2^2)/2} + t_1^2 e^{(t_1^2 + t_2^2)/2}$  ja  $E(X_1^2) = M_{11}(0, 0) = 1$ .  $\square$

Huomattakoon, että myös satunnaisvektoreiden tapauksessa pätee momenttifunktioiden yksikäsitteisyttä koskeva lause (vrt. Lause 3.12). Jos siis satunnaisvektoreilla  $(X_1, X_2)$  ja  $(Y_1, Y_2)$  on sama momenttifunktio, niin niillä on sama jakauma. Reunajakaumien momenttifunktiot saadaan kätevästi yhteisjakuman momenttifunktiosta.

**Lause 5.10** *Oletetaan, että satunnaisvektorin  $(X, Y)$  momenttifunktio on  $M(s, t)$  sekä  $X$ :n ja  $Y$ :n momenttifunktiot vastaavasti  $M_X(s)$  ja  $M_Y(t)$ .*

1. Silloin

$$M_X(s) = M(s, 0) \quad \text{ja} \quad M_Y(t) = M(0, t).$$

2.  $X$  ja  $Y$  ovat riippumattomat jos ja vain jos

$$M(s, t) = M_X(s) M_Y(t).$$

## 5.8 Kahden muuttujan normaalijakauma

### 5.8.1 Standardimuoto

Oletetaan, että satunnaismuuttujat  $Z$  ja  $V$  ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa. Silloin  $Z$ :n ja  $V$ :n riippumattomuuden nojalla satunnaisvektorin  $(Z, V)$  yhteisjakauman tiheysfunktio on

$$(5.8.1) \quad f_{Z,V}(z, v) = f(z)f(v) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} \frac{1}{\sqrt{2\pi}}e^{-v^2/2} = \frac{1}{2\pi}e^{-(z^2+v^2)/2}.$$

Sanomme, että satunnaisvektori  $(Z, V)$  noudattaa kaksiulotteista standardimuotoista normaalijakaumaa ja funktio (5.8.1) on tämän jakauman tiheysfunktio. Merkitään  $(Z, V) \sim N_2(\mathbf{0}, \mathbf{I})$ , missä  $\mathbf{0}$  on  $2 \times 1$ -nollavektori eli  $\mathbf{0} = (0, 0)^T$ . Merkintä  $(0, 0)^T$  tarkoittaa vektorin  $(0, 0)$  transponointia, joka muuntaa vaakavektorin  $(0, 0)$  pystytoriksi. Matriisi

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

on  $2 \times 2$ -identiteettimatriisi. Yhteisjakauman reunajakaumien keskiarvot ovat  $E(Z) = E(V) = 0$  ja varianssit  $\text{Var}(Z) = \text{Var}(V) = 1$  sekä  $\text{Cov}(Z, V) = 0$ . Satunnaisvektorin  $(Z, V)$  odotusarvovektori on  $[E(Z), E(V)]^T = \mathbf{0}$  ja kovarianssimatriisi

$$\begin{pmatrix} \text{Var}(Z) & \text{Cov}(Z, V) \\ \text{Cov}(V, Z) & \text{Var}(V) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Huomaa, että aina  $\text{Cov}(Z, V) = \text{Cov}(V, Z)$ , joten kovarianssimatriisi on symmetrinen. Voidaan merkitä myös  $(Z, V) \sim N_2(0, 0; 1, 1, 0)$ , missä odotusarvot, varianssit ja korrelaatio on annettu sulkeissa.

### 5.8.2 Korreloivat muuttujat

Oletetaan, että  $X \sim N(0, 1)$  ja  $Z \sim N(0, 1)$  ovat riippumattomat. Niiden avulla voidaan konstruoida normaalijakaumaa noudattava satunnaismuuttuja  $Y$  siten, että  $X$  ja  $Y$  korreloivat. Kiertämällä  $x$ -akselia kulman  $\theta$  verran vastapäivään saadaan  $y$ -akseli (Kuvio 5.10). Projisoidaan satunnaispiste  $(X, Z)$   $y$ -akselille ja merkitään tätä projektiota  $Y$ :llä. On helppo todeta geometrisen päättelyn avulla (Kuvio 5.10), että

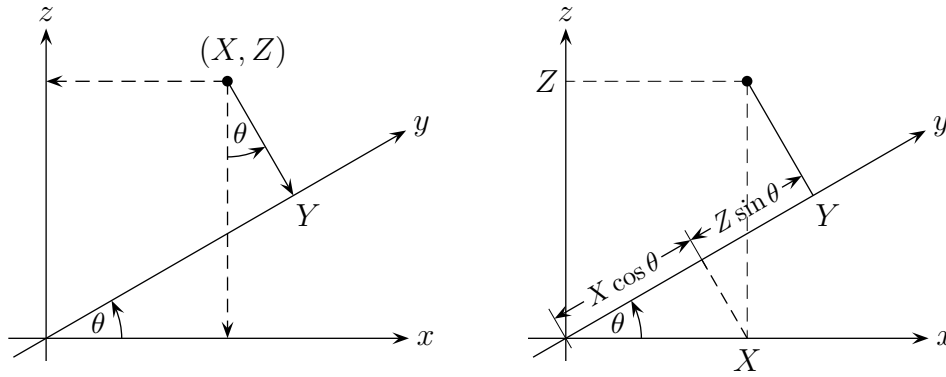
$$Y = X \cos \theta + Z \sin \theta.$$

Satunnaismuuttuja  $Y$  saadaan siis  $X$ :n ja  $Z$ :n lineaarisena muunoksena. Tästä seuraa, että

$$E(Y) = \cos \theta \cdot E(X) + \sin \theta \cdot E(Z) = 0$$

ja

$$\text{Var}(Y) = \cos^2 \theta \cdot \text{Var}(X) + \sin^2 \theta \cdot \text{Var}(Z) = 1,$$



Kuvio 5.10.

koska  $E(X) = E(Z) = 0$  ja  $\text{Var}(X) = \text{Var}(Z) = 1$ . Lauseen 5.4 mukaan  $Y \sim N(0, 1)$ . Satunnaisuuttujien  $X$  ja  $Y$  2. kertaluvun sekamomentti on

$$\begin{aligned} E(XY) &= E[X(X \cos \theta + Z \sin \theta)] \\ &= \cos \theta \cdot E(X^2) + \sin \theta \cdot E(XZ) \\ &= \cos \theta. \end{aligned}$$

Viimeinen yhtäsuuruus seuraa siitä, että  $E(X^2) = 1$  ja  $E(XZ) = E(X)E(Z) = 0$ . Satunnaisuuttujien  $X$  ja  $Y$  välinen korrelaatio  $\text{Cor}(X, Y) = E(X, Y)$ , koska  $E(X) = E(Y) = 0$  ja  $\text{Var}(X) = \text{Var}(Y) = 1$ .

## 5.9 Satunnaisvektoreiden muunnokset

Oletetaan, että jatkuvien satunnaismuuttujien  $X$  ja  $Y$  yhteisjakauman tiheysfunktio on  $f$ . Olkoon

$$(5.9.1) \quad U = h_1(X, Y); \quad V = h_2(X, Y)$$

sellainen satunnaisvektorin  $(X, Y)$  muunnos, että sillä on käänteismuunnos. Silloin mitä tahansa satunnaisvektorin  $(U, V)$  arvo  $(u, v) \in \mathbb{R}^2$  vastaa yksikäsitteinen satunnaisvektorin  $(X, Y)$  arvo  $(x, y) \in \mathbb{R}^2$ . Voimme silloin määritellä käänteiskuvauksen

$$x = g_1(u, v); \quad y = g_2(u, v).$$

Vektoreiden  $(u, v) \in \mathbb{R}^2$  ja  $(x, y) \in \mathbb{R}^2$  välillä on yksi-yksinen vastaavuus. Oletamme lisäksi, että funktioilla  $g_1$  ja  $g_2$  on jatkuvat osittaisderivaatat. Yksiulotteisen muunnoksen tapauksessa laskettavaa derivaattaa  $g'$  vastaa satunnaisvektorien muunnoksen *Jacobin determinantti*, joka on funktioiden  $g_1$  ja  $g_2$  osittaisderivaattojen matriisin determinantti. Jacobin determinanttia kutsutaan muunnoksen *Jakobiaaniksi*.

Muunnoksen (5.9.1) Jakobiaani on

$$(5.9.2) \quad \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v},$$

missä

$$\begin{aligned} \frac{\partial x}{\partial u} &= \frac{\partial g_1(u, v)}{\partial u}, & \frac{\partial x}{\partial v} &= \frac{\partial g_1(u, v)}{\partial v}, \\ \frac{\partial y}{\partial u} &= \frac{\partial g_2(u, v)}{\partial u}, & \frac{\partial y}{\partial v} &= \frac{\partial g_2(u, v)}{\partial v}. \end{aligned}$$

Jacobin determinanttia merkitään  $J = \frac{\partial(x, y)}{\partial(u, v)}$ . Oletetaan, että  $J \neq 0$ , kun  $f(x, y) > 0$ . Satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio on

$$(5.9.3) \quad f_{U, V}(u, v) = f(g_1(u, v), g_2(u, v)) |J|.$$

Satunnaisvektorin  $(U, V)$  arvoavaruus  $S_{U, V}$  saadaan tarkastelemalla kuvausta (5.9.1), joka kuvaa satunnaisvektorin  $(X, Y)$  arvojoukon  $S_{X, Y}$  kuvajoukoksi  $S_{U, V}$ .

**Esimerkki 5.23** Olkoot  $X$  ja  $Y$  jatkuvat satunnaismuuttujat, joiden yhteisjakauman tiheysfunktio on  $f$ . Määritellään satunnaismuuttujat  $U$  ja  $V$  siten, että

$$(5.9.4) \quad U = X + Y; \quad V = X - Y.$$

Johdetaan nyt  $(U, V)$ :n jakauman tiheysfunktio. Muunnoksen (5.9.4) käänteismuunnos on

$$x = g_1(u, v) = \frac{u + v}{2}, \quad y = g_2(u, v) = \frac{u - v}{2},$$

ja muunnoksen Jakobiaani on

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio on yhtälön (5.9.3) nojalla

$$(5.9.5) \quad f_{U, V}(u, v) = \frac{1}{2} f\left(\frac{u + v}{2}, \frac{u - v}{2}\right).$$

Jos esimerkiksi  $X$  ja  $Y$  ovat riippumattomat ja noudattavat tasajakaumaa  $\text{Tas}(0, 1)$ , niin  $(X, Y)$ :n yhteisjakauman tiheysfunktio  $f(x, y) = 1$ , kun  $x \in [0, 1]$  ja  $y \in [0, 1]$ . Silloin  $(U, V)$ :n tiheysfunktio on

$$f_{U, V}(u, v) = \begin{cases} \frac{1}{2}, & 0 \leq u + v \leq 2, \quad 0 \leq u - v \leq 2. \\ 0 & \text{muualla.} \end{cases}$$

□



### Yleinen muunnos

Tarkasteltavalla muunnoksella ei tietenkään aina ole käänteismuunnosta. Jos muunnos (5.9.1) ei ole yksi-yksinen eli bijektio, niin sillä ei ole käänteismuunnosta. Jos kuitenkin on olemassa sellainen arvoavaruuden  $S_{X,Y}$  ositus yhteispisteettömiin  $(x, y)$ -tason osaväleihin  $A_1, A_2, \dots, A_m$ , että

$$(5.9.6) \quad S_{X,Y} = A_1 \cup A_2 \cup \dots \cup A_m$$

ja muunnoksella

$$u = h_1(x, y), \quad v = h_2(x, y)$$

on käänteismuunnos

$$x = g_1(x, y), \quad y = g_2(x, y)$$

jokaisella osavälillä  $A_i$ ,  $i = 1, 2, \dots, m$ , niin kaavaa (5.9.3) voidaan soveltaa kullakin osavälillä erikseen vastaavalla tavalla kuin yhden muuttujan tapauksessa. Määritellään funktiot

$$h_{ki}(x, y) = \begin{cases} h_k(x, y), & \text{kun } (x, y) \in A_i; \\ 0 & \text{muualla,} \end{cases}$$

kun  $k = 1, 2$ . Silloin  $h_1(x, y) = \sum_{i=1}^m h_{1i}(x, y)$  ja  $h_2(x, y) = \sum_{i=1}^m h_{2i}(x, y)$ . Jokaisella muunnoksella

$$u = h_{1i}(x, y), \quad v = h_{2i}(x, y)$$

on käänteismuunnos välillä  $A_i$ ,  $i = 1, 2, \dots, m$ . Merkitään näitä käänteismuunnoksia

$$x = g_{1i}(u, v), \quad y = g_{2i}(u, v).$$

Satunnaisvektorin  $(U, V)$  tiheysfunktio voidaan silloin esittää kaavan (5.9.3) avulla seuraavasti

$$(5.9.7) \quad f_{U,V}(u, v) = \sum_{i=1}^m f_{X,Y}(g_{1i}(u, v), g_{2i}(u, v)) |J_i|,$$

missä  $J_i$  on muunnoksen  $x = g_{1i}(u, v)$ ,  $y = g_{2i}(u, v)$  Jakobiaani.

### Satunnaismuuttujien funktion jakauma

Usein tarkasteltavana on vain yksi satunnaismuuttujien  $X$  ja  $Y$  funktio  $U = h_1(X, Y)$ . Funktio  $h_1(X, Y)$  voi olla esimerkiksi muotoa  $X + Y$ ,  $XY$ ,  $X^2 + Y^2$  jne. Esitettyä muunnostekniikkaa voidaan edelleen käyttää, jos löydetään sellainen apumuuttuja  $V = h_2(X, Y)$ , että muunnoksella

$$U = h_1(X, Y), \quad V = h_2(X, Y)$$

on käänteismuunnos. Muunnoskaavan (5.9.3) avulla saadaan sitten satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio, josta voidaan määrittää  $U$ :n reunajakauman tiheysfunktio. Jos esimerkiksi  $U = h_1(X, Y) = X + Y$ , niin voidaan valita apumuuttuja  $V = h_2(X, Y) = X - Y$ . Silloin muunnoksella  $u = x + y$ ,  $v = x - y$  on käänteismuunnos

$$x = g_1(u, v) = \frac{u + v}{2}, \quad y = g_2(u, v) = \frac{u - v}{2}$$

ja  $(U, V)$ :n tiheysfunktio saadaan kaavalla (5.9.3), niinkuin Esimerkissä 5.23 osoitettiin. Kun tästä  $(U, V)$ :n tiheysfunktioista ”integroidaan  $v$  pois”, saadaan  $U$ :n tiheysfunktio. Jos ollaan kiinnostuneita esimerkiksi satunnaismuuttujan  $U = XY$  tiheysfunktioista, voidaan valita apumuuttuja  $V = X$ , sillä muunnoksella  $u = xy$ ,  $v = x$  on käänteismuunnos. Sitten sovelletaan jälleen edellä kuvattua tekniikkaa. Huomaa, että apumuuttujan valinta ei ole yksikäsitteinen, vaan useilla eri valinnoilla voidaan päästä haluttuun tulokseen.

Tässä yhteydessä on syytä palauttaa mieleen Lause 3.6. Siinä osoitettiin riippumattomille satunnaismuuttujille  $X$  ja  $Y$  seuraava tulos: Jos  $g(X)$  ei riipu  $Y$ :stä ja  $h(Y)$  ei riipu  $X$ :stä, niin silloin satunnaismuuttujat  $g(X)$  ja  $h(Y)$  ovat riippumattomat. Lause todistettiin diskreettien satunnaismuuttujien tapauksessa, mutta se pitää paikkansa myös jatkuville muuttujille.

**Esimerkki 5.24** Jos  $X$  ja  $Y$  ovat riippumattomat ja noudattavat standardimuotoista normaalijakumaa, niin mitä jakaumaa noudattaa  $X + Y$ ? Määritellään ensin apumuuttuja  $V = X - Y$ . Kuten esimerkissä 5.23 osoitettiin, muunnoksen  $u = x + y$ ,  $v = x - y$  käänteismuunnos on

$$x = g_1(u, v) = \frac{u + v}{2}, \quad y = g_2(u, v) = \frac{u - v}{2},$$

ja muunnoksen Jakobiaani  $J = -\frac{1}{2}$ . Yhtälön (5.9.5) perusteella  $(U, V)$ :n tiheysfunktio on

$$\begin{aligned} f_{U,V}(u, v) &= \frac{1}{4\pi} e^{-[(u+v)^2/8 + (u-v)^2/8]} \\ &= \frac{1}{4\pi} e^{-(u^2+v^2)/4} = \frac{1}{\sqrt{4\pi}} e^{-u^2/4} \frac{1}{\sqrt{4\pi}} e^{-v^2/4} \\ (5.9.8) \quad &= f_U(u) f_V(v). \end{aligned}$$

Näemme, että  $f_U(u)$  ja  $f_V(v)$  ovat kumpikin normaalijakauman  $N(0, 2)$  tiheysfunktioita. Siksi

$$\int_{-\infty}^{\infty} f_{U,V}(u, v) dv = f_U(u) \int_{-\infty}^{\infty} f_V(v) dv = f_U(u),$$

joten  $U = X + Y \sim N(0, 2)$ . Identiteetistä (5.9.8) seuraa, että  $X + Y$  ja  $X - Y$  ovat riippumattomat. Havaitsemme myös, että  $X - Y \sim N(0, 2)$ . Itse

asiassa voidaan todistaa seuraava tulos: Jos  $X$  ja  $Y$  ovat riippumattomat ja noudattavat samaa jakaumaa  $F$ , niin  $X + Y$  ja  $X - Y$  ovat riippumattomat jos ja vain jos  $F$  on normaalijakauma.  $\square$

**Esimerkki 5.25** Olkoot  $X$  ja  $Y$  jatkuvat satunnaismuuttujat, joiden yhteisjakauman tiheysfunktio on  $f$ . Määritellään lineaarinen muunnos

$$(5.9.9) \quad \begin{aligned} u &= ax + by, \\ v &= cx + dy. \end{aligned}$$

Ratkaisemalla yhtälöryhmästä (5.9.9)  $x$  ja  $y$  saadaan käänteismuunnos

$$\begin{aligned} x &= \frac{du - bv}{D}, \\ y &= \frac{av - cu}{D}, \end{aligned}$$

missä  $D = ad - bc$ . Käänteismuunnos on olemassa, jos  $D \neq 0$ . Muunnoksen Jakobiaani on

$$J = \begin{vmatrix} \frac{d}{D} & \frac{-b}{D} \\ \frac{-c}{D} & \frac{a}{D} \end{vmatrix} = \frac{ad - bc}{D^2} = \frac{1}{D}.$$

Satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio on yhtälön (5.9.3) nojalla

$$(5.9.10) \quad f_{U,V}(u, v) = \frac{1}{|D|} f[(du - bv)/D, (av - cu)/D].$$

Esimerkin 5.23 yhtälö (5.9.5) on yhtälön (5.9.10) erikoistapaus. Kun  $a = b = c = 1$  ja  $d = -1$  sijoitetaan yhtälöön (5.9.10), saadaan yhtälö (5.9.5). Lineaarinen muunnos

$$\begin{aligned} u &= ax + by + e, \\ v &= cx + dy + f. \end{aligned}$$

voidaan palauttaa muunnokseen (5.9.9) merkitsemällä  $u^* = u - e$  ja  $v^* = v - f$ , jolloin

$$\begin{aligned} u^* &= ax + by, \\ v^* &= cx + dy. \end{aligned}$$

Yhtälöstä (5.9.10) saadaan sitten  $(U^*, V^*)$ :n tiheysfunktio.  $\square$

### 5.9.1 Yleinen kahden muuttujan normaalijakauma

Standardimuotoinen kaksiulotteinen normaalijakauma määriteltiin alaluvussa 5.8.1. Yleinen kaksiulotteinen normaalijakauma voidaan määritellä standardimuotoisen normaalijakauman avulla vastaavasti kuin yhden muuttujan tapauksessa. Olkoon  $(X, Y)$  sellainen satunnaisvektori, että  $E(X) = \mu_1$ ,

$E(Y) = \mu_2$ ,  $\text{Var}(X) = \sigma_1^2$ ,  $\text{Var}(Y) = \sigma_2^2$  ja  $\text{Cov}(X, Y) = \sigma_{12}$ . Silloin satunnaisvektorin  $(X, Y)$  keskiarvovektori on  $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$  ja kovarianssimatriisi

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Satunnaismuuttujien  $X$  ja  $Y$  välinen korrelaatiokerroin on  $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ .

**Määritelmä 5.7** Satunnaisvektori  $(X, Y)$  noudattaa normaalijakaumaa, jonka keskiarvovektori on  $\boldsymbol{\mu}$  ja kovarianssimatriisi  $\boldsymbol{\Sigma}$ , jos se voidaan lausua muodossa

$$(5.9.11) \quad \begin{aligned} X - \mu_1 &= \sigma_1 \sqrt{1 - \rho^2} Z_1 + \rho \sigma_1 Z_2, \\ Y - \mu_2 &= \sigma_2 Z_2, \end{aligned}$$

missä  $Z_1$  ja  $Z_2$  ovat riippumattomat ja noudattavat standardimuotoista normaalijakaumaa  $N(0, 1)$  sekä  $\sigma_1 > 0$ ,  $\sigma_2 > 0$  ja  $|\rho| \leq 1$ .

Merkitsemme  $(X, Y) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , kun  $(X, Y)$  noudattaa normaalijakaumaa, jonka keskiarvovektori on  $\boldsymbol{\mu}$  ja kovarianssimatriisi  $\boldsymbol{\Sigma}$ . Normaalijakaumaa  $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  noudattava satunnaisvektori  $(X, Y)$  saadaan aina riippumattomista standardoiduista normaalimuuttujista lineaarisella muunnoksella (5.9.11)

Yleisen normaalijakauman tiheysfunktio saadaan suoraan Esimerkissä 5.25 esitetyllä tekniikalla. Merkitään  $x - \mu_1 = u$  ja  $y - \mu_2 = v$ . Koska nyt lineaarisessa muunnoksessa (5.9.9)  $c = 0$  ja  $D = ad$ , saa yhtälö (5.9.10) muodon

$$f_{U,V}(u, v) = \frac{1}{ad} f\left(\frac{du - bv}{ad}, \frac{v}{d}\right).$$

Koska  $f$  on kahden muuttujan standardimuotoisen normaalijakauman tiheysfunktio ja muunnoksessa (5.9.11)  $a = \sigma_1 \sqrt{1 - \rho^2}$ ,  $b = \rho \sigma_1$  ja  $d = \sigma_2$ , niin

$$\begin{aligned} f_{U,V}(u, v) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2u - \rho\sigma_1v)^2 + \frac{v^2}{\sigma_2^2}\right] \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{u^2}{\sigma_1^2} - 2\rho\frac{u}{\sigma_1}\frac{v}{\sigma_2} + \frac{v^2}{\sigma_2^2}\right)\right]. \end{aligned}$$

Kun edellä johdettuun tiheysfunktioon sijoitetaan  $u = x - \mu_1$  ja  $v = y - \mu_2$ , saadaan  $(X, Y)$ :n tiheysfunktio

$$(5.9.12) \quad f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)\right].$$

Seuraavassa lauseessa esitetään kaksiulotteisen normaalijakauman keskeiset ominaisuudet.

**Lause 5.11** Oletetaan, että satunnaisvektori  $(X, Y)$  noudattaa kaksiulotteista normaalijakaumaa  $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , missä  $\boldsymbol{\mu} = (\mu_1, \mu_2)^T = [E(X), E(Y)]^T$  ja

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{pmatrix}$$

ja  $\rho = \text{Cor}(X, Y) = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ . Silloin pitävät paikkansa seuraavat ominaisuudet:

1.  $X \sim N(\mu_1, \sigma_1^2)$  ja  $Y \sim N(\mu_2, \sigma_2^2)$ ,
2.  $X$  ja  $Y$  ovat riippumattomat jos ja vain jos  $\rho = 0$ .
3.  $X$ :n ja  $Y$ :n ehdolliset jakaumat:

$$X | y \sim N\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right),$$

eli  $X$  noudattaa normaalijakaumaa ehdolla, että  $Y = y$  on annettu. Vastaavasti

$$Y | x \sim N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right).$$

4. Satunnaisvektorin  $(X, Y)$  momenttifunktio on

$$M(s, t) = \exp\left(\mu_1 s + \mu_2 t + \frac{\sigma_1^2 s^2 + \sigma_2^2 t^2 + \rho\sigma_1\sigma_2 st}{2}\right).$$

**Lause 5.12** Satunnaisvektori  $(X, Y)$  noudattaa kaksiulotteista normaalijakaumaa, jos ja vain jos satunnaismuuttujien  $X$  ja  $Y$  lineaarikombinaatiot  $aX + bY$  noudattavat yksiulotteista normaalijakaumaa kaikilla  $a \in \mathbb{R}$  ja  $b \in \mathbb{R}$ .

## 5.9.2 Studentin $t$ -jakauma, $F$ -jakauma ja beta-jakauma

Oletetaan, että satunnaismuuttujat  $Z$  ja  $U$  ovat riippumattomat,  $Z \sim N(0, 1)$  ja  $U$  noudattaa  $\chi^2$ -jakaumaa vapausastein  $r$  eli  $U \sim \text{Khi2}(r)$ . Tarkastellaan nyt satunnaismuuttujan

$$(5.9.13) \quad T = \frac{Z}{\sqrt{U/r}}$$

jakaumaa. Tätä muotoa olevalla satunnaismuuttujalla on erittäin keskeinen rooli tilastollisessa päättelyssä. Satunnaismuuttuja (5.9.13) noudattaa ns. *Studentin  $t$ -jakaumaa* (tai lyhyesti  $t$ -jakaumaa) vapausastein  $r$ . Jakauma on nimetty englantilaisen tilastotieteilijän W. S. Gossetin mukaan. Gosset esitti tämän jakauman *Biometrikassa* vuonna 1908 nimimerkillä ”Student”. Kun  $T$  noudattaa Studentin  $t$ -jakaumaa vapausastein  $r$ , merkitään  $T \sim t(r)$ .

Olkoon  $X_1, X_2, \dots, X_n$  on otos normaalijakaumasta  $N(\mu, \sigma^2)$ . Silloin siis satunnaismuuttujat  $X_1, X_2, \dots, X_n$  ovat toisistaan riippumattomat ja  $X_i \sim N(\mu, \sigma^2)$ ,  $1 \leq i \leq n$ . Otoksesta laskettu suure

$$(5.9.14) \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}$$

eli otossuure noudattaa  $t$ -jakaumaa vapausastein  $n - 1$ , missä  $\bar{X}$  on otoskeskiarvo ja

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

on otosvarianssi. Lausekkeen (5.9.14) osoittaja  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$  noudattaa normaalijakaumaa  $N(0, 1)$  ja satunnaismuuttuja  $S^2/\sigma^2$  noudattaa jakaumaa  $\text{Khi2}(n-1)$ . Koska  $\bar{X}$  ja  $S^2$  ovat toisistaan riippumattomat, niin lausekkeen (5.9.14) osoittaja ja nimittäjä ovat toisistaan riippumattomat.

Satunnaismuuttujan (5.9.13) eli Studentin  $t$ -jakauman tiheysfunktio vapausastein  $r$  on

$$(5.9.15) \quad f_T(t) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})} \frac{1}{\sqrt{r\pi}} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \quad -\infty < t < \infty.$$

Huomaa, että vapausastein  $r = 1$  tiheysfunktioista (5.9.15) tulee Cauchyn jakauman tiheysfunktio.

Tiheysfunktio (5.9.15) saadaan suoraviivaisesti edellä esitetyllä muunnostekniikalla. Lähdetään liikkeelle satunnaisvektorin  $(Z, U)$  yhteisjakaumasta. Koska  $Z$  ja  $U$  ovat riippumattomat, niin

$$f_{Z,U}(z, u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{\Gamma(r/2) 2^{r/2}} u^{r/2-1} e^{-u/2}, \quad -\infty < z < \infty, \quad 0 < u < \infty.$$

Tehdään muunnos

$$t = \frac{z}{\sqrt{u/r}}, \quad w = u,$$

jonka käänteismuunnos on

$$z = t\sqrt{w/r}, \quad u = w.$$

Muunnoksen jakobiaani on  $\sqrt{w/r}$ . Sen jälkeen voidaan soveltaa muunnoskaavaa (5.9.3).  $T$ :n tiheysfunktio saadaan  $(T, W)$ :n yhteisjakauman tiheysfunktioista integroimalla  $w$ :n yli:

$$f_T(t) = \int_0^\infty f_{Z,U}[t(w/r)^{1/2}, w](w/r)^{1/2} dw.$$

Laskennan lopputuloksena on  $t$ -jakauman tiheysfunktio (5.9.15). Yksityiskohdat jätetään lukijan tehtäväksi.

Studentin  $t$ -jakaumalla ei ole momenttifunktiota, koska sillä ei ole kaikkien kertalukujen momentteja. Jos  $T_r \sim t(r)$ , niin silloin  $T_r$ :llä on ainoastaan  $r - 1$  ensimmäistä momenttia. Esimerkiksi jakaumalla  $t(1)$  ei ole keskiarvoa ja jakaumalla  $t(2)$  ei ole varianssia. Voidaan laskemalla osoittaa, että

$$(5.9.16) \quad E(T_r) = 0, \quad \text{jos } r > 1, \quad \text{Var}(T_r) = \frac{r}{r-2}, \quad \text{jos } r > 2.$$

Toinen tilastollisessa päättelyssä keskeinen jakauma, *Snedecorin  $F$ -jakauma* tai vain lyhyesti  $F$ -jakauma, voidaan johtaa  $t$ -jakauman tapaan. Jos  $X_1, X_2, \dots, X_n$  on otos normaalijakaumasta  $N(\mu_1, \sigma_1^2)$  ja  $X_1, X_2, \dots, X_m$  otos normaalijakaumasta  $N(\mu_2, \sigma_2^2)$ , niin silloin satunnaismuuttuja

$$(5.9.17) \quad \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

noudattaa  $F$ -jakaumaa vapausastein  $n - 1$  ja  $m - 1$ . Lausekkeessa (5.9.17)  $S_1^2$  ja  $S_2^2$  ovat otosvariانسsit.  $S_1^2/\sigma_1^2$  ja  $S_2^2/\sigma_2^2$  ovat riippumattomat ja

$$(n-1)S_1^2/\sigma_1^2 \sim \text{Khi2}(n-1), \quad (m-1)S_2^2/\sigma_2^2 \sim \text{Khi2}(m-1).$$

$F$ -jakauma määritellään seuraavasti: Jos  $X \sim \text{Khi2}(r)$  ja  $Y \sim \text{Khi2}(s)$  ovat riippumattomat, niin satunnaismuuttuja

$$(5.9.18) \quad F = \frac{X/r}{Y/s}$$

noudattaa  $F$ -jakaumaa vapausastein  $r$  ja  $s$ . Silloin merkitään  $F \sim F(r, s)$ .  $F$ -jakauman tiheysfunktio voidaan johtaa vastaavalla tavalla kuin  $t$ -jakauman tiheysfunktio. Määritellään muunnos

$$U = X + Y, \quad V = \frac{X/r}{Y/s},$$

missä  $F$ -suuretta (5.9.18) on merkitty kirjaimella  $V$ . Muunnoksen käänteismuunnos on

$$X = \frac{r}{s} \left(1 + \frac{r}{s}V\right)^{-1} UV, \quad Y = \left(1 + \frac{r}{s}V\right)^{-1} U.$$

Koska  $X$  ja  $Y$  ovat riippumattomat, niin

$$f_{X,Y}(X, Y) = k_r k_s x^{(r/2)-1} y^{(s/2)-1} e^{-(x+y)/2}, \quad 0 < x < \infty, \quad 0 < y < \infty,$$

missä  $k_r = 1/\Gamma(\frac{r}{2})2^{r/2}$  ja  $k_s = 1/\Gamma(\frac{s}{2})2^{s/2}$ . Kun lasketaan Jakobiaani ja tehdään sijoitukset, saadaan satunnaisvektorin  $(U, V)$  tiheysfunktio  $f_{U,V}(u, v)$  kaavan (5.9.3) mukaisesti.  $V$ :n reunajakauman tiheysfunktio on sitten  $F$ -jakauman tiheysfunktio:

$$(5.9.19) \quad f_F(v) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \left(\frac{r}{s}\right)^{r/2} \frac{v^{(r/2)-1}}{[1 + \frac{r}{s}v]^{(r+s)/2}}, \quad 0 < v < \infty.$$

$F$ -jakauman keskiarvo ja varianssi ovat

$$(5.9.20) \quad E(F) = \frac{s}{s-2}, \quad s > 2$$

$$(5.9.21) \quad \text{Var}(F) = \frac{2s^2(r+s-2)}{r(s-2)^2(s-4)}, \quad s > 4.$$

### Beta-jakauma

Olkoot  $X \sim \text{Gamma}(\alpha, \theta)$  ja  $Y \sim \text{Gamma}(\beta, \theta)$  riippumattomat gamma-jakaumaa noudattavat satunnaismuuttujat. Silloin satunnaisvektorin  $(X, Y)$  tiheysfunktio on

$$f(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} x^{\alpha-1} y^{\beta-1} \exp\left(-\frac{x+y}{\theta}\right), \quad 0 < x, \quad y < \infty.$$

Tarkastellaan muunnosta

$$U = \frac{X}{X+Y}, \quad V = X+Y,$$

jonka käänteismuunnos on

$$X = UV, \quad Y = V - UV.$$

Muunnoksen Jakobiaani on

$$\begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1-u) + uv = v.$$

Satunnaisvektorin  $(U, V)$  tiheysfunktio on kaavan (5.9.3) mukaan

$$f_{U,V}(u, v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (uv)^{\alpha-1} (v-uv)^{\beta-1} e^{-v\theta} v,$$

missä  $0 < u < 1$  ja  $0 < v < \infty$ . Kun tästä  $(U, V)$ :n yhteisjakauman tiheysfunktioista määritetään  $U$ :n reunajakuman tiheysfunktio, saadaan

$$(5.9.22) \quad f_U(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1} (1-u)^{\beta-1}, \quad 0 < u < 1.$$

Sanomme, että  $U$  noudattaa betajakaumaa parametrein  $\alpha$  ja  $\beta$ . Silloin merkitsemme  $U \sim \text{Beta}(\alpha, \beta)$ . Koska (5.9.22) on tiheysfunktio, niin

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du.$$

Funktiota

$$(5.9.23) \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



kutsutaan *betafunktioksi*.

Betajakauma on yksi niitä harvoja nimettyjä jakaumia, joiden koko todennäköisyysmassa on äärellisellä välillä, eli välillä  $(0, 1)$ . Betajakauman momentit on helppo laskea tiheysfunktion erityisominaisuuksien avulla. Kun  $n > -\alpha$ , niin

$$(5.9.24) \quad E(X^n) = \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + n)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)}.$$

Sijoittamalla  $n = 1$  ja  $n = 2$  kaavaan (5.9.24) saadaan 1. ja 2. momentti ja niiden avulla varianssi:

$$(5.9.25) \quad E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{ja} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

**Esimerkki 5.26** Oletetaan, että on suoritettavana  $n + m$  työtä. Töiden suorittamiseen tarvittavat ajat noudattavat toisistaan riippumatta eksponenttijakaumaa keskiarvolla  $\theta > 0$  (ts. gammajakaumaa parametrein  $\alpha = 1$  ja  $\theta$ ). Oletetaan, että kaksi eri työntekijää tekee nämä työt siten, että työntekijä  $A$  tekee työt  $1, 2, \dots, n$  ja  $B$  tekee työt  $n + 1, n + 2, \dots, n + m$ . Jos merkitään  $X$ :llä työntekijän  $A$  käyttämää aikaa ja  $Y$ :llä työntekijän  $B$  käyttämää aikaa, niin toisistaan riippumatta  $X \sim \text{Gamma}(n, \theta)$  ja  $Y \sim \text{Gamma}(m, \theta)$ . Silloin  $(n + m)$ :n työn vaatima kokonaisaika  $X + Y$  noudattaa gammajakaumaa  $\text{Gamma}(n + m, \theta)$ . Työntekijän  $A$  käyttämä suhteellinen osuus  $X/(X + Y)$  kokonaisajasta noudattaa betajakaumaa  $\text{Beta}(n, m)$ .  $\square$

### 5.9.3 Hierarkkiset mallit ja yhdistetyt jakaumat

Tarkastellaan aluksi esimerkkinä tietyn laitteen, esimerkiksi kopiokoneen, rikkoontumista. Oletetaan, että rikkoontumisten lukumäärä  $X$  vuodessa noudattaa Poissonin jakaumaa parametrilla  $\lambda$ . Kun laite on rikkoontunut, sen korjaamiseen tarvittava aika noudattaa eksponenttijakaumaa keskiarvolla  $\theta > 0$ . Olkoon  $Y_i$  aika, joka tarvitaan  $i$ . rikkoontumisen korjaamiseen. Oletetaan lisäksi, että korjausajat eri kerroilla ovat riippumattomat. Jos vuodessa sattuu  $X = x$  rikkoontumista, niin kokonaiskorjausaika on

$$Y = Y_1 + Y_2 + \dots + Y_x.$$

Silloin

$$E(Y | x) = E(Y_1) + E(Y_2) + \dots + E(Y_x) = x\theta$$

ja

$$\text{Var}(Y | x) = \text{Var}(Y_1) + \text{Var}(Y_2) + \dots + \text{Var}(Y_x) = x\theta^2.$$

Edellä lasketut keskiarvo ja varianssi ovat siis ehdollisia ehdolla  $X = x$ . Ehdollinen keskiarvo ja varianssi

$$\mu(x) = E(Y | x) = x\theta \quad \text{ja} \quad \sigma^2(x) = \text{Var}(Y | x) = x\theta^2$$

ovat  $x$ :n funktioita. Koska  $X \sim \text{Poi}(\lambda)$  on satunnaismuuttuja, niin myös

$$\mu(X) = E(Y | X) = X\theta \quad \text{ja} \quad \sigma^2(X) = \text{Var}(Y | X) = X\theta^2$$

ovat satunnaismuuttujia. Silloin

$$\begin{aligned} E[\mu(X)] &= E[E(Y | X)] = \theta E(X) = \theta\lambda, \\ E[\sigma^2(X)] &= E[\text{Var}(Y | X)] = \theta^2 E(X) = \theta^2\lambda, \\ \text{Var}[\mu(X)] &= \text{Var}[E(Y | X)] = \theta \text{Var}(X) = \theta\lambda. \end{aligned}$$

Määritimme siis edellä kaksinkertaisen odotusarvon  $E[E(Y | X)]$ , missä sisimmäinen odotusarvo on otettu  $Y$ :n suhteen ja ulompi  $X$ :n suhteen.

Esitämme nyt kaksinkertaisia odotusarvoja koskevan lauseen, joka usein helpottaa huomattavasti odotusarvojen laskemista.

**Lause 5.13** *Olkoot  $X$  ja  $Y$  mitkä tahansa kaksi satunnaismuuttujaa. Silloin*

$$(5.9.26) \quad E(Y) = E[E(Y | X)],$$

*jos odotusarvot ovat olemassa.*

**Todistus.** Olkoon  $f(x, y)$  satunnaisvektorin  $(X, Y)$  tiheysfunktio. Määritelmän mukaan

$$(5.9.27) \quad E(Y) = \iint yf(x, y) \, dy \, dx = \int \left[ \int yf(y | x) \, dy \right] f_X(x) \, dx,$$

missä  $f(y | x)$  on  $Y$ :n ehdollisen jakauman tiheysfunktio ehdolla  $X = x$  ja  $f_X(x)$  on  $X$ :n reunajakauman tiheysfunktio. Integrointi tehdään yli  $(X, Y)$ :n arvoavaruuden. Koska sisempi integraali lausekkeessa (5.9.27) on ehdollinen odotusarvo  $E(Y | x)$ , niin odotusarvo (5.9.27) voidaan kirjoittaa muodossa

$$E(Y) = \int E(Y | x) f_X(x) \, dx = E[E(Y | X)],$$

niin kuin lauseessa väitetään. □

Voimme soveltaa nyt Lausetta 5.13 laitteen rikkoontumisten vaatimaa korjausaikaa koskevaan esimerkkiin. Kokonaiskorjausaika vuodessa on  $Y$  ja sen keskiarvo on Lauseen 5.13 mukaan

$$\begin{aligned} E(Y) &= E[E(Y | X)] \\ &= E(\theta X) = \theta E(X) = \theta\lambda. \end{aligned}$$

Tälläinen sovellus on selvintä ajatella kaksitasoisena hierarkisena mallina, missä 1. vaiheessa sattuvat rikkoontumiset Poissonin jakauman mukaan ja sitten 2. vaiheessa tarvittavat korjausajat jakaantuvat eksponenttijakauman

mukaan. Huomaa, että kokonaiskorjausajan  $Y$  keskiarvoparametri on nyt satunnaismuuttuja  $\theta X$ , missä  $X$  noudattaa Poissonin jakaumaa. Siksi  $Y$ :n jakaumaa on perusteltua kutsua yhdistetyksi jakaumaksi, koska siinä yhdistyvät parametrin Poissonin jakauma ja korjausajan eksponenttijakauama. Odotusarvoa  $E(Y)$  laskettaessa on yksinkertaisinta laskea ensin ehdollinen odotusarvo  $E(Y | X)$  ja sitten näiden ehdollisten odotusarvojen odotusarvo. Silloin laskennassa ei tarvita  $Y$ :n reunajakaummaa. Lopputulos on kuitenkin Lauseen 5.13 mukaan  $E(Y)$ .

Myös varianssi voidaan laskea ehdollisen jakauman varianssin avulla.

**Lause 5.14** *Olkoot  $X$  ja  $Y$  mitkä tahansa kaksi satunnaismuuttujaa. Silloin*

$$(5.9.28) \quad \text{Var}(Y) = E[\text{Var}(Y | X)] + \text{Var}[E(Y | X)],$$

*jos odotusarvot ovat olemassa.*

**Todistus.** Varianssin määritelmän mukaan

$$\text{Var}(Y) = E[Y - E(Y)]^2 = E([Y - E(Y | X) + E(Y | X) - E(Y)]^2),$$

missä viimeinen yhtäsuuruus on saatu lisäämällä ja vähentämällä  $E(Y | X)$ . Korottamalla lauseke neliöön ja ottamalla odotusarvot saadaan

$$(5.9.29) \quad \text{Var}(Y) = E([Y - E(Y | X)]^2) + E([(Y | X) - E(Y)]^2) \\ + 2 E([Y - E(Y | X)] E[(Y | X) - E(Y)]).$$

Lausekkeen (5.9.29) viimeinen termi on nolla, mikä voidaan osoittaa laskeamalla merkityt odotusarvot. Yhtälön (5.9.29) oikean puolen ensimmäinen termi voidaan kirjoittaa muodossa:

$$E([Y - E(Y | X)]^2) = E(E[Y - E(Y | X)]^2 | X) \\ = E[\text{Var}(Y | X)]$$

ja toinen termi muodossa

$$E([E(Y | X) - E(Y)]^2) = \text{Var}[E(Y | X)],$$

joten identiteetti (5.9.28) pitää paikkansa. □

## Jatkuvat jakaumat: Yhteenveto

- Satunnaismuuttuja  $X$  on (absoluuttisesti) jatkuva, jos  $X$ :llä on tiheysfunktio  $f(x) \geq 0$ , joka toteuttaa seuraavat ehdot:

1.  $f(x) > 0$ , kun  $x \in S$ ,
2.  $\int_S f(x) dx = 1$ ,
3.  $P(X \in A) = \int_A f(x) dx$  on tapahtuman  $\{X \in A\}$  todennäköisyys.

Satunnaismuuttujan  $X$  kertymäfunktio on

$$F(x) = \int_{-\infty}^x f(t) dt$$

ja momenttifunktio

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

- Tasajakauma  $X \sim \text{Tas}(a, b)$ . Tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{kun } x \in [a, b]; \\ 0 & \text{muualla} \end{cases} \quad \text{ja}$$

$$E(X) = \frac{a+b}{2} \quad \text{ja} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Momenttifunktio

$$M(t) = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)}, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

- Eksponenttijakauma  $X \sim \text{Exp}(\theta)$ ,  $\theta > 0$  ja  $x > 0$ ,

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \quad \text{ja} \quad F(x) = 1 - e^{-x/\theta}.$$

Silloin

$$E(X) = \theta \quad \text{ja} \quad \text{Var}(X) = \theta^2.$$

Momenttifunktio

$$M(t) = \frac{1}{1 - \theta t}, \quad t < \frac{1}{\theta}.$$

- Gammajakauma  $X \sim \text{Gamma}(\alpha, \beta)$ ,  $\alpha > 0$  ja  $\beta > 0$ . Silloin

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty,$$

$$E(X^c) = \frac{\Gamma(\alpha + c)\beta^c}{\Gamma(\alpha)} \quad \text{kaikilla } c > -\alpha.$$

Erityisesti

$$E(X) = \alpha\beta \quad \text{ja} \quad \text{Var}(X) = \alpha\beta^2.$$

Momenttifunktio

$$M(t) = \left( \frac{1}{1 - \beta t} \right)^\alpha, \quad t < \frac{1}{\beta}.$$

- $\chi^2$ -jakauma,  $X \sim \text{Khi2}(r)$ .  $\chi^2$ -jakauma saadaan, kun gammajakaumassa valitaan  $\alpha = \frac{r}{2}$  ja  $\beta = 2$ , missä  $r$  on positiivinen kokonaisluku. Silloin

$$E(X) = r, \quad \text{Var}(X) = 2r \quad \text{ja} \quad M(t) = (1 - 2t)^{-r/2}, \quad t < \frac{1}{2}.$$

Jos  $X_i \sim \text{Khi2}(r_i)$ ,  $i = 1, 2$ , ovat riippumattomat, niin  $X_1 + X_2 \sim \text{Khi2}(r_1 + r_2)$ .

- Normaalijakauma  $X \sim N(\mu, \sigma^2)$ . Silloin

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty,$$

$$E(X) = \mu, \quad \text{Var}(X) = \sigma^2 \quad \text{ja} \quad M(t) = e^{\mu t + (\sigma^2 t^2)/2}.$$

- Jos  $Z \sim N(0, 1)$ , niin  $Z^2 \sim \text{Khi2}(1)$ .

Jos  $Z_i \sim N(0, 1)$ ,  $i = 1, 2$ , ovat riippumattomat, niin  $Z_1^2 + Z_2^2 \sim \text{Khi2}(2)$ .

- Studentin  $t$ -jakauma,  $T \sim t(r)$ .

$$T = \frac{Z}{\sqrt{U/r}} \sim t(r),$$

jos  $Z \sim N(0, 1)$  ja  $U \sim \text{Khi2}(r)$  ovat riippumattomat. Silloin

$$E(T) = 0, \quad r > 1 \quad \text{ja} \quad \text{Var}(T) = \frac{r}{r-2}, \quad r > 2.$$

- $F$ -jakauma,  $F \sim F(r, s)$ .

$$F = \frac{X/r}{Y/s} \sim F(r, s),$$

jos  $X \sim \text{Khi2}(r)$  ja  $Y \sim \text{Khi2}(s)$  ovat riippumattomat. Silloin

$$E(F) = \frac{s}{s-2}, \quad s > 2; \quad \text{Var}(F) = \frac{2s^2(r+s-2)}{r(s-2)^2(s-4)}, \quad s > 4.$$

- Beta-jakauma,  $X \sim \text{Beta}(\alpha, \beta)$ ;  $\alpha > 0$ ,  $\beta > 0$ .

Tiheysfunktio

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{ja} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

## Muuttujien vaihto

Satunnaismuuttujan  $Y = h(X)$  tiheysfunktio, kun  $X$ :n tiheysfunktio on  $f_X(x)$ :

$$f_Y(y) = f_X[g(y)]|g'(y)|, \quad y \in S_Y,$$

missä  $g(y)$  on  $h(x)$ :n käänteisfunktio.

## Kaksiulotteiset jakaumat

- Jatkuvien satunnaismuuttujien  $X$  ja  $Y$  muodostaman satunnaisvektorin  $(X, Y)$  tiheysfunktio toteuttaa ehdot

1.  $f(x, y) \geq 0$  kaikilla  $(x, y)$ ,
2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ ,
3.  $P[(X, Y) \in A] = \iint_{(x,y) \in A} f(x, y) dx dy$ .

Reunajakaumien tiheysfunktiot

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad x \in S_X; \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx, \quad y \in S_Y$$

Ehdolliset tiheysfunktiot

$$f_X(x | y) = \frac{f(x, y)}{f_Y(y)} \quad \text{ja} \quad f_Y(y | x) = \frac{f(x, y)}{f_X(x)},$$

missä  $f_X(x) > 0$  ja  $f_Y(y) > 0$ .

Kertymäfunktio

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) ds dt.$$

- Normaalijakauma  $(X, Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ .

Tiheysfunktio

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}Q(x, y)\right],$$

missä

$$Q(x, y) = \left(\frac{x - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x - \mu_1}{\sigma_1}\right)\left(\frac{y - \mu_2}{\sigma_2}\right) + \left(\frac{y - \mu_2}{\sigma_2}\right)^2.$$

Reunajakaumat:

$$X \sim N(\mu_1, \sigma_1^2) \quad \text{ja} \quad Y \sim N(\mu_2, \sigma_2^2).$$

Ehdolliset jakaumat:

$X$ :n jakauma ehdolla  $Y = y$  on

$$X | y \sim N\left[\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right],$$

$Y$ :n jakauma ehdolla  $X = x$  on

$$Y | x \sim N\left[\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right].$$

## Muuttujien vaihto

Olkoon

$$U = h_1(X, Y); \quad V = h_2(X, Y)$$

muunnos, jolla on käänteismuunnos  $x = g_1(u, v)$ ,  $y = g_2(u, v)$ . Satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio on

$$f_{U,V}(u, v) = f(g_1(u, v), g_2(u, v))|J|,$$

missä  $f(x, y)$  on satunnaismuuttujien  $X$  ja  $Y$  yhteisjakauman tiheysfunktio ja

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v}.$$

## Ehdolliset odotusarvot ja varianssit

- Ehdollinen odotusarvo

$$E(Y) = E[E(Y | X)]$$

- Ehdollinen varianssi

$$\text{Var}(Y) = E[\text{Var}(Y | X)] + \text{Var}[E(Y | X)].$$

## Harjoituksia

1. Olkoon  $X$  jatkuva satunnaismuuttuja, jonka tiheysfunktio

$$f(x) = \begin{cases} \frac{1}{2}, & 0 \leq x < \frac{1}{3}; \\ 1, & \frac{1}{3} \leq x < \frac{2}{3}; \\ c, & \frac{2}{3} \leq x \leq 1; \\ 0 & \text{muualla.} \end{cases}$$

- (a) Laske  $c$ .
- (b) Määritä  $X$ :n kertymäfunktio.
- (c) Piirrä  $X$ :n tiheysfunktio ja kertymäfunktio.
2. Olkoon jatkuvan satunnaismuuttujan  $X$  tiheysfunktio  $f(x) = 2(1 - x)$ , kun  $0 \leq x \leq 1$  ja  $f(x) = 0$  muualla.
- (a) Piirrä  $X$ :n tiheysfunktio.
- (b) Määritä ja piirrä  $X$ :n kertymäfunktio.
- (c) Laske (i)  $P(0 \leq X \leq 1/2)$ , (ii)  $P(1/4 \leq X \leq 3/4)$ , (iii)  $P(X = 3/4)$  ja (iv)  $P(X \geq 3/4)$ .

3. Olkoon  $f(x)$  jatkuvan satunnaismuuttujan  $X$  tiheysfunktio. (i) Määritä jokaisesta alla määritellystä funktiosta  $f(x)$  vakio  $c$  siten, että  $f(x)$  on tiheysfunktio. (ii) Määritä kertymäfunktio  $F(x) = P(X \leq x)$  ja (iii) hahmottele tiheysfunktion  $f(x)$  ja kertymäfunktion  $F(x)$  kuvaajat.

- (a)  $f(x) = x^3/4$ ,  $0 < x < c$ ;
- (b)  $f(x) = (3/16)x^2$ ,  $-c < x < c$ ;
- (c)  $f(x) = c/\sqrt{x}$ ,  $0 < x < 1$ . Onko  $f(x)$  rajoitettu?

4. Olkoon  $X$ :n tiheysfunktio  $f(x) = c/x^2$ ,  $1 < x < \infty$ .

- (a) Määritä  $c$ :n arvo siten, että  $f(x)$  on tiheysfunktio.
- (b) Osoita, että  $E(X)$  ei ole äärellinen.

5. Olkoon  $X \sim \text{Khi2}(12)$ . Määritä vakiot  $a$  ja  $b$  siten, että

$$P(a < X < b) = 0.90 \quad \text{ja} \quad P(X < a) = 0.05.$$

6. Olkoon  $X \sim \text{Khi2}(23)$ .

- (a) Laske  $P(14.85 < X < 32.01)$ .
- (b) Määritä  $a$  ja  $b$  siten, että  $P(a < X < b) = 0.95$  ja  $P(X < a) = 0.025$ .



(c)  $X$ :n keskiarvo ja varianssi.

7. Olkoon  $X$ :n momenttifunktio  $M(t) = (1 - 2t)^{-12}$ ,  $t < 1/2$ . Laske

(a)  $E(X)$ ,

(b)  $\text{Var}(X)$  ja

(c)  $P(15.66 < X < 42.98)$ .

8. Olkoon  $X \sim N(7, 4)$ . Laske todennäköisyys  $P[15.364 \leq (X - 7)^2 \leq 20.096]$  (Vihje: Lause 5.6).

9. Oletetaan, että  $X$  ja  $Y$  ovat riippumattomat ja  $X \sim \text{Khi2}(8)$  ja  $Y \sim \text{Khi2}(12)$ .

(a) Laske

$$P(1.646 < X \leq 20.09), \quad P(Y > 6.304), \quad P(X + Y = 19.34).$$

(b) Määritä  $b$ ,  $c$  ja  $d$  siten, että

$$P(X \leq b) = 0.9, \quad P(Y > c) = 0.9, \quad P(X + Y > d) = 0.05.$$

(Vihje: Lause 5.8)

10. Olkoot  $Z_1$ ,  $Z_2$  ja  $Z_3$  riippumattomat ja ne noudattavat  $N(0, 1)$ -jakaumaa. Määritellään satunnaismuuttujat

$$\bar{Z} = \frac{1}{3}(Z_1 + Z_2 + Z_3) \quad \text{ja} \quad U = Z_1^2 + Z_2^2 + Z_3^2.$$

Määritä vakiot  $a$ ,  $b$  siten, että

$$P(|\bar{Z}| \leq a) = 0.95; \quad P(U > b) = 0.025.$$

(Vihje: Voit olettaa, että  $\bar{Z}$  noudattaa normaalijakaumaa. Ks. myös Lause 5.7).

11. Määritä Esimerkissä 5.13  $X$ :n ja  $Y$ :n jakaumien (reunajakaumien) tiheysfunktiot. Totea, että  $X$  ja  $Y$  ovat riippumattomat.

12. Totea laskemalla, että Esimerkissä 5.20  $E(X | y) = \frac{y}{2}$ ,  $0 \leq y \leq 1$ .

13. Olkoon jatkuvien satunnaismuuttujien  $X$  ja  $Y$  yhteisjakauman kertymäfunktio

$$F(x, y) = \begin{cases} kxy(x + y), & \text{kun } 0 < x < 1 \text{ ja } 0 < y < 1; \\ 0 & \text{muualla.} \end{cases}$$

(a) Laske vakion  $k$  arvo ja määritä  $X$ :n ja  $Y$ :n yhteisjakauman tiheysfunktio.

- (b) Määritä  $X$ :n reunajakauman ja ehdollisen jakauman tiheysfunktio.  
 (c) Laske todennäköisyydet

$$P(X < 0.5, Y < 0.5); \quad P(X < 0.5); \quad P(X < 0.5 \mid Y < 0.5).$$

14. Määritä  $a$ ,  $b$ ,  $c$ ,  $d$  siten, että

(a)  $P(a \leq F_{8,12} \leq b) = 0.8$ ;  $P(c \leq F_{6,6} \leq d) = 0.98$ .

(b)  $P(|t_9| \geq a) = 0.05$ ;  $P(|t_{20}| > b) = 0.95$ .

(c)  $P(F_{1,12} \geq b) = 0.05$ ;  $P(F_{1,12} \leq c) = 0.2$ .

15. Olkoot  $Y$  ja  $Z$  sellaiset riippumattomat satunnaismuuttujat, että  $Z \sim N(0, 1)$  ja  $Y \sim \chi_4^2$ . Määritä  $a$ ,  $b$  siten, että

$$P(Z \leq a\sqrt{Y}) = 0.975; \quad P(Y + Z^2 \leq b) = 0.975.$$

16. Oletetaan, että  $X \sim t(\nu)$ . Osoita muuttujien vaihtotekniikalla, että  $X^2 \sim F_{1,\nu}$ . (Huomaa, että muunnos ei ole monotoninen).

17. Olkoon  $X \sim \text{Khi2}(12)$ .

- (a) Määritä vakiot  $a$  ja  $b$  siten, että

$$P(a < X < b) = 0.90 \quad \text{ja} \quad P(X < a) = 0.05.$$

- (b) Olkoon  $X \sim N(1, 4)$  ja  $P(a < X < b) = 0.50$ . Määritä  $a$  ja  $b$  siten, että  $b - a$  on mahdollisimman pieni.

18. Olkoot  $Z_1$ ,  $Z_2$  ja  $Z_3$  riippumattomat ja  $Z_i \sim N(0, 1)$ -jakaumaa,  $i = 1, 2, 3$ . Määritellään satunnaismuuttujat  $X_i = iZ_i + i$  ja

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3).$$

Määritä vakio  $a$  siten, että  $P(|\bar{X}| \leq a) = 0.95$ . (Vihje: Voit olettaa, että  $\bar{X}$  noudattaa normaalijakaumaa.)

19. Olkoon  $X \sim \text{Khi2}(23)$ .

- (a) Laske  $P(14.85 < X < 32.01)$ .

- (b) Määritä  $a$  ja  $b$  siten, että  $P(a < X < b) = 0.95$  ja  $P(X < a) = 0.025$ .

- (c) Laske  $X$ :n keskiarvo ja varianssi.

20. Oletetaan, että  $X \sim \text{Gamma}(\alpha, \beta)$ . Osoita gammajakauman momenttifunktion avulla, että  $E(X) = \alpha\beta$  ja  $\text{Var}(X) = \alpha\beta^2$ .

21. Sillalle saapuu autoja Poissonin prosessin mukaan keskimäärin 15 autoa 10:ssä minuutissa. Laske todennäköisyys, että siltamaksujen kerääjä joutuu odottamaan annetusta hetkestä alkaen kahdeksaa autoa (eli kahdeksatta autoa) ainakin puoli tuntia.

22. Oletetaan, että  $X$  noudattaa gammajakaumaa  $\text{Gamma}(3, 2)$ . Määritä satunnaismuuttujan  $Y = \sqrt{X}$  tiheysfunktio.

23. Logistista jakaumaa noudattavan satunnaismuuttujan  $X$  tiheysfunktio on

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}, \quad -\infty < x < \infty.$$

Osoita, että satunnaismuuttuja  $Y = \frac{1}{1+e^{-X}}$  noudattaa tasajakaumaa  $\text{Tas}(0, 1)$ .

24. Oletetaan, että  $X \sim \text{Tas}(-1, 3)$ . Määritä satunnaismuuttujan  $Y = X^2$  jakauma.

25. Olkoon momenttifunktio  $M(t) = (1 - 2t)^{-12}$ ,  $t < \frac{1}{2}$ . Määritä  $E(X)$ ,  $\text{Var}(X)$  ja  $P(15 < X \leq 42)$ .

26. Oletetaan, että matkustusaika kotoa töihin noudattaa normaalijakaumaa, jonka keskiarvo on 40 minuuttia ja hajonta 7 minuuttia. Jos haluat 95 %:n todelläköisyydellä olla työpaikalla klo 8:00, niin milloin viimeistään on lähdettävä kotoa?

27. Olkoon  $X$ :n momenttifunktio  $M(t) = (1 - 2t)^{-12}$ ,  $t < 1/2$ . Laske

- (a)  $E(X)$ ,
- (b)  $\text{Var}(X)$  ja
- (c)  $P(15.66 < X < 42.98)$ .

28. Oletetaan, että  $X \sim \text{Gamma}(3, 1.5)$ . Laske

- (a)  $P(X > 5)$ ,
- (b) jakauman moodi (tiheysfunktion maksimi) sekä
- (c)  $E(Y)$  ja  $\text{Var}(Y)$ , kun  $Y = \frac{1}{X}$ . (Ks. Lause 5.1.)

29. Tarkastellaan Poissonin prosessia, jonka intensiteetti on  $\lambda$ . Olkoon  $W$  odotusaika, kunnes  $\alpha$  tapahtumaa sattuu. Silloin  $W$ :n kertymäfunktio on

$$F(w) = 1 - \sum_{x=0}^{\alpha-1} e^{-\lambda w} \frac{(\lambda w)^x}{x!}.$$

Osoita, että tiheysfunktio  $f(w)$  on

$$f(w) = \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} e^{-\lambda w}.$$

30. Satunnaismuuttujan  $Z$  tiheysfunktio on  $f_Z(z)$ . Olkoon  $Y = aZ + b$ , missä  $a$  ja  $b$  ovat annettuja vakioita.

(a) Osoita, että  $Y$ :n tiheysfunktio on

$$f_Y(y) = \frac{1}{|a|} f_Z\left(\frac{y-b}{a}\right).$$

(b) Esitä  $Y$ :n tiheysfunktio, kun  $a = 2$ ,  $b = 1$  ja  $Z \sim N(0, 1)$ .

31. Olkoon  $X$ :n ja  $Y$ :n yhteisjakauman tiheysfunktio (Esimerkki 5.18)

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \quad -1 < x < 1, \quad -1 < y < 1.$$

Määritä  $X$ :n ja  $Y$ :n jakaumien (reunajakaumien) tiheysfunktiot. Totea, että  $X$  ja  $Y$  ovat riippumattomat.

32. Oletetaan, että riippumattomat satunnaismuuttujat  $X$  ja  $Y$  noudattavat tasajakaumaa  $\text{Tas}(0, 1)$ . Laske todennäköisyydet

(a)  $P(|X - Y| \leq \frac{1}{2})$  ja

(b)  $P(|\frac{X}{Y} - 1| \leq \frac{1}{2})$ .

33. Olkoot  $X$  ja  $Y$  riippumattomat standardoidut normaalimuuttujat. Määritellään muuttujat

$$U = \frac{X + Y}{\sqrt{2}}; \quad V = \frac{X - Y}{\sqrt{2}}.$$

Kirjoita  $U$ :n ja  $V$ :n yhteisjakauman tiheysfunktio. Näytä, että  $U$  ja  $V$  ovat riippumattomat.

34. Satunnaisvektori  $(X, Y)$  noudattaa kaksiulotteista normaalijakaumaa parametrein  $\mu_1 = \mu_2 = 0$ ,  $\sigma_1^2 = 4$ ,  $\sigma_2^2 = 9$  ja  $\rho = 0.8$ .

(a) Kirjoita  $(X, Y)$ :n yhteisjakauman tiheysfunktion lauseke,

(b)  $E(Y | x)$  ja

(c)  $\text{Var}(Y | x)$ .

(Ks. Lause 5.11.)

35. Erääseen naisten kunto-ohjelmaan osallistuneilta mitattiin kehon rasvaprosentti  $X$  ennen ohjelman alkua ja rasvaprosenttin muutos  $Y$  ohjelman lopussa. Oletetaan, että muuttujien yhteisjakauma on normaalin ja  $\mu_X = 24.5$ ,  $\sigma_X = 4.8$ ,  $\mu_Y = -0.2$ ,  $\sigma_Y = 3.0$  ja  $\rho_{XY} = -0.32$ . Laske

(a)  $P(1.3 \leq Y \leq 5.8)$ ,

(b)  $E(Y | X = x)$ ,

- (c)  $\text{Var}(Y \mid X = x)$ ,
- (d)  $P(1.3 \leq Y \leq 5.8 \mid X = 18)$ .

(Ks. Lause 5.11.)

- 36.** Oletetaan, että satunnaisvektori  $(X, Y)$  noudattaa standardimuotoista normaali-jakaumaa, missä  $\text{Cor}(X, Y) = 0.6$ . Laske  $P(X - Y < 1, X + Y > 2)$ .
- 37.** Olkoot  $X$  ja  $Y$  riippumattomat satunnaismuuttujat, joiden tiheysfunktio-  
tiot ovat

$$f(x) = e^{-x}, \quad f(y) = e^{-y}, \quad 0 < x < \infty, \quad 0 < y < \infty.$$

Esitä satunnaisvektorin  $(U, V)$  yhteisjakauman tiheysfunktio, kun  $U = X - Y$  ja  $V = X + Y$ .



# Luku 6

## Otantajakaumien teoria

### 6.1 Riippumattomat satunnaismuuttujat

Muistamme edellisistä luvuista, että satunnaismuuttujat  $X_1$  ja  $X_2$  ovat riippumattomat (määritelmät 4.6 ja 5.5), jos

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \quad \text{kaikilla } x_1 \in S_1, x_2 \in S_2,$$

missä  $f(x_1, x_2)$  on  $X_1$ :n ja  $X_2$ :n yhteisjakauman tiheysfunktio,  $f_1(x_1)$  on  $X_1$ :n ja  $f_2(x_2)$  on  $X_2$ :n tiheysfunktio. Määritelmä yleistyy suoraviivaisesti usean satunnaismuuttujan tapaukseen. Satunnaismuuttujat  $X_1, X_2, \dots, X_n$  ovat *riippumattomat*, jos

$$f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n).$$

Jos satunnaismuuttujat  $X_1, X_2, \dots, X_n$  ovat *riippumattomat*, niin myös niiden funktiot  $u_1(X_1), u_2(X_2), \dots, u_n(X_n)$  ovat riippumattomat, mikäli kukin funktio  $u_i, i = 1, 2, \dots, n$  riippuu vain satunnaismuuttujasta  $X_i$  eikä siis satunnaismuuttujista  $X_j, j \neq i$ . Silloin erityisesti Lauseen 3.10 mukaan

$$E[u_1(X_1)u_2(X_2) \cdots u_n(X_n)] = E[u_1(X_1)] E[u_2(X_2)] \cdots E[u_n(X_n)].$$

Jos riippumattomat satunnaismuuttujat  $X_1, X_2, \dots, X_n$  noudattavat samaa jakaumaa (RSJ), jonka kertymäfunktio on  $F(x)$ , niin sanomme, että  $X_1, X_2, \dots, X_n$  on  $n$ :n kokoinen otos jakaumasta  $F$ . Kertymäfunktio edustaa populaatiota, josta otos tehdään.

**Esimerkki 6.1** Olkoon  $X_1, X_2, \dots, X_n$  otos normaalijakaumasta  $N(\mu, \sigma^2)$ . Silloin  $X_i \sim N(\mu, \sigma^2), i = 1, 2, \dots, n$  ja  $X_1, X_2, \dots, X_n$  ovat riippumattomat.

Otoksen  $X_1, X_2, \dots, X_n$  yhteisjakauman tiheysfunktio on siis

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-[1/(2\sigma^2)](x_i-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-[1/(2\sigma^2)]\sum_{i=1}^n (x_i-\mu)^2}, \end{aligned}$$

missä

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-[1/(2\sigma^2)](x_i-\mu)^2}, \quad i = 1, 2, \dots, n.$$

□

**Lause 6.1 (Apulauseen 5.1 yleistys)** *Satunnaisvektorit  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  ja  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$  ovat riippumattomat jos ja vain jos on olemassa sellaiset funktiot  $g(\mathbf{X})$  ja  $h(\mathbf{Y})$ , että*

$$f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x})h(\mathbf{y})$$

kaikilla  $\mathbf{x}$ :n ja  $\mathbf{y}$ :n arvoilla, missä  $g$  ei riipu  $\mathbf{y}$ :stä ja  $h$  ei riipu  $\mathbf{x}$ :stä.

## 6.2 Riippumattomien satunnaismuuttujien summan jakauma

Tilastollisissa sovelluksissa tarkastellaan tavallisesti erilaisia satunnaismuuttujien funktioita. Otoksesta  $X_1, X_2, \dots, X_n$  laskettua reaali-, tai vektoriarvoista suuretta  $T(X_1, X_2, \dots, X_n)$  sanotaan otoksen *tunnusluvuksi* (*statistics*). Kaksi tärkeää otoksen tunnuslukua ovat otoskeskiarvo  $\bar{X}$  ja otosvarianssi  $S^2$ . Esimerkiksi  $T_1(X_1, X_2, \dots, X_n) = \bar{X}$  on reaaliarvoinen ja  $T_2(X_1, X_2, \dots, X_n) = (\bar{X}, S^2)$  on vektoriarvoinen.

**Lause 6.2** *Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka tiheysfunktio on  $f(x)$ . Silloin satunnaismuuttujien  $X_1, X_2, \dots, X_n$  yhteisjakuman tiheysfunktio on  $f_1(x_1)f_2(x_2)\cdots f_n(x_n)$ . Jos  $g(y)$  on satunnaismuuttujan  $Y = u(X_1, X_2, \dots, X_n)$  tiheysfunktio, niin*

$$\begin{aligned} E(Y) &= \int_{S_y} yg(y) dy \\ &= \int_S \int_S \cdots \int_S u(x_1, x_2, \dots, x_n) f_1(x_1) f_2(x_2) \cdots f_n(x_n) dx_1 dx_2 \cdots dx_n, \end{aligned}$$

mikäli odotusarvo on olemassa. Diskreettejä satunnaismuuttujia koskeva vastaava tulos saadaan korvaamalla integraalit summalausekkeilla. Satunnaismuuttujien  $X_1, X_2, \dots, X_n$  arvoalue on  $S$  ja  $Y$ :n arvoalue on  $S_y$ .



**Esimerkki 6.2** Heitetään kahta noppaa. Olkoon 1. nopan silmäluku  $X_1$  ja 2. nopan silmäluku  $X_2$ . Määritetään nyt silmälukujen summan  $Y = X_1 + X_2$  todennäköisyysfunktio  $g(y)$ . Tarkastellaan ensin yksittäisen arvon, esimerkiksi  $y = 4$ , todennäköisyyden  $g(4)$  laskemista. Tapahtuma  $\{Y = 4\}$  voi sattua kolmella toisensa poissulkevalla tavalla:  $\{X_1 = 1, X_2 = 3\}$ ,  $\{X_1 = 2, X_2 = 2\}$  ja  $\{X_1 = 3, X_2 = 1\}$ . Siksi

$$\begin{aligned} g(4) &= P(Y = 4) \\ &= P(X_1 = 1, X_2 = 3) + P(X_1 = 2, X_2 = 2) + P(X_1 = 3, X_2 = 1) \\ &= \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36}. \end{aligned}$$

Jatkamalla samalla periaatteella saadaan todennäköisyysfunktio  $g(y)$ :

$y$	2	3	4	5	6	7	8	9	10	11	12
$g(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

□

Yleisesti Esimerkin 6.2 todennäköisyydet voidaan laskea ns. *konvoluutio-kaavalla*

$$g(y) = P(Y = y) = \sum_{k=1}^{y-1} f(k)f(y-k),$$

missä

$$f(k) = \frac{1}{6}, \quad k = 1, 2, 3, 4, 5, 6$$

on nopan silmäluvun todennäköisyysfunktio.

Toinen tapa johtaa  $g(y)$ , on käyttää momenttifunktiota. Nopan silmäluvun momenttifunktio on

$$M_X(t) = E(e^{tX}) = \frac{1}{6}e^t + \frac{1}{6}e^{2t} + \frac{1}{6}e^{3t} + \frac{1}{6}e^{4t} + \frac{1}{6}e^{5t} + \frac{1}{6}e^{6t}.$$

Koska silmäluvut ovat riippumattomat, niin  $Y$ :n momenttifunktio on

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t) = [M_X(t)]^2.$$

Koska  $e^{kt}$ :n kerroin  $M_Y(t)$ :n lausekkeessa on todennäköisyys  $P(Y = k)$ ,  $k = 2, 3, \dots, 12$ , ne muodostavat  $Y$ :n todennäköisyysfunktion.

**Lause 6.3** *Olkoot riippumattomien satunnaismuuttujien  $X_1, X_2, \dots, X_n$  odotusarvot  $\mu_1, \mu_2, \dots, \mu_n$  ja varianssit  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ . Silloin satunnaismuuttujan  $Y = \sum_{i=1}^n a_i X_i$  odotusarvo ja varianssi ovat*

$$\mu_Y = \sum_{i=1}^n a_i \mu_i \quad \text{ja} \quad \sigma_Y^2 = \sum_{i=1}^n a_i^2 \sigma_i^2,$$

missä  $a_1, a_2, \dots, a_n$  ovat annettuja vakioita.

**Todistus.** Koska odotusarvo on lineaarinen operaattori, niin

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n E(a_i X_i) \\ &= \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i \mu_i. \end{aligned}$$

Vastaavasti

$$\begin{aligned} \sigma^2 &= E[(Y - \mu_Y)^2] = E\left[\left(\sum_{i=1}^n a_i X_i - \sum_{i=1}^n a_i \mu_i\right)^2\right] \\ &= E\left[\sum_{i=1}^n a_i (X_i - \mu_i)\right]^2 = E\left[\sum_{i=1}^n \sum_{j=1}^n a_i a_j (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j E[(X_i - \mu_i)(X_j - \mu_j)] = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \sigma_{ij}, \end{aligned}$$

missä

$$\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)].$$

Koska  $X_i$  ja  $X_j$  ovat riippumattomat, niin  $\sigma_{ij} = 0$ , kun  $i \neq j$ . Tästä seuraa, että

$$\sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

□

**Esimerkki 6.3** Olkoot  $X_1$  ja  $X_2$  riippumattomat satunnaismuuttujat, joiden odotusarvot ovat  $\mu_1 = -4$  ja  $\mu_2 = 3$  sekä varianssit vastaavasti  $\sigma_1^2 = 4$  ja  $\sigma_2^2 = 9$ . Silloin satunnaismuuttujan  $Y = 3X_1 - 2X_2$  odotusarvo ja varianssi ovat

$$\mu_Y = 3 \cdot (-4) + (-2) \cdot 3 = -18$$

ja

$$\sigma_Y^2 = 3^2 \cdot 4 + (-2)^2 \cdot 9 = 72.$$

□

**Esimerkki 6.4** Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka odotusarvo on  $\mu$  ja varianssi  $\sigma^2$ . Silloin otoskeskiarvon

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

odotusarvo ja varianssi ovat

$$\mu_{\bar{X}} = \sum_{i=1}^n \left(\frac{1}{n}\right) \mu = \mu \quad \text{ja} \quad \sigma_{\bar{X}}^2 = \sum_{i=1}^n \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

Otosvarianssi on muotoa

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right),$$

joten

$$E(S^2) = \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right].$$

Koska  $E(X_i^2) = \sigma^2 + \mu^2$  ja  $E(\bar{X}^2) = \sigma^2/n + \mu^2$ , niin laskemalla on helppo todeta, että  $E(S^2) = \sigma^2$ . Olemme siis osoittaneet, että  $\bar{X}$  on  $\mu$ :n ja  $S^2$  on  $\sigma^2$ :n harhaton estimaattori.  $\square$

**Lause 6.4** Jos  $X_1, X_2, \dots, X_n$  ovat riippumattomat satunnaismuuttujat, joiden momenttifunktiot ovat  $M_{X_i}(t)$ ,  $i = 1, 2, \dots, n$ , niin satunnaismuuttujan  $Y = \sum_{i=1}^n a_i X_i$  momenttifunktio on

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(a_i t).$$

**Todistus.** Satunnaismuuttujan  $Y$  momenttifunktio on

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{t(a_1 X_1 + a_2 X_2 + \dots + a_n X_n)}) \\ &= E(e^{a_1 t X_1} \cdot e^{a_2 t X_2} \dots e^{a_n t X_n}) \\ &= E(e^{a_1 t X_1}) E(e^{a_2 t X_2}) \dots E(e^{a_n t X_n}), \end{aligned}$$

koska satunnaismuuttujat  $e^{a_i t X_i}$  ovat keskenään riippumattomat. Momenttifunktion määritelmän mukaan

$$E(e^{tX_i}) = M_{X_i}(t),$$

joten

$$E(e^{a_i t X_i}) = M_{X_i}(a_i t).$$

Siksi

$$M_Y(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t) \dots M_{X_n}(a_n t) = \prod_{i=1}^n M_{X_i}(a_i t).$$

$\square$

**Esimerkki 6.5** Olkoon  $X_1, X_2, \dots, X_n$  otos Bernoullin jakaumasta  $\text{Ber}(\frac{1}{3})$ . Silloin

$$M(t) = \frac{2}{3} + \frac{1}{3}e^t.$$

Jos  $Y = X_1 + X_2 + \dots + X_n$ , niin

$$M_Y(t) = \prod_{i=1}^n \left( \frac{2}{3} + \frac{1}{3}e^t \right) = \left( \frac{2}{3} + \frac{1}{3}e^t \right)^n.$$

Tästä näemme, että  $Y \sim \text{Bin}(n, \frac{1}{3})$ .  $\square$

**Seuraus 6.1** Jos  $X_1, X_2, \dots, X_n$  on otos jakaumasta, jonka momenttifunktio on  $M(t)$ , niin

1. satunnaismuuttujan  $Y = \sum_{i=1}^n X_i$  momenttifunktio on

$$M_Y(t) = \prod_{i=1}^n M(t) = [M(t)]^n.$$

2. otoskeskiarvon  $\bar{X} = \sum_{i=1}^n (1/n)X_i$  momenttifunktio on

$$M_{\bar{X}}(t) = \prod_{i=1}^n M\left(\frac{t}{n}\right) = \left[M\left(\frac{t}{n}\right)\right]^n.$$

**Esimerkki 6.6** Olkoon  $X_1, X_2, X_3$  otos eksponenttijakaumasta, jonka odotusarvo on  $\theta$ . Eksponenttijakauman momenttifunktio on  $M(t) = 1/(1 - \theta t)$ ,  $t < 1/\theta$ . Silloin summan  $Y = X_1 + X_2 + X_3$  momenttifunktio on

$$M_Y(t) = [1/(1 - \theta t)]^3 = (1 - \theta t)^{-3}, \quad t < \frac{1}{\theta},$$

mikä on gammajakauman  $\text{Gamma}(3, \theta)$  momenttifunktio, joten  $Y \sim \text{Gamma}(3, \theta)$ . Toisaalta  $\bar{X}$ :n momenttifunktio on

$$M_{\bar{X}}(t) = \left[\left(1 - \frac{\theta t}{3}\right)^{-1}\right]^3 = \left(1 - \frac{\theta t}{3}\right)^{-3}, \quad \text{kun } t < \frac{3}{\theta}.$$

Otoskeskiarvo  $\bar{X}$  noudattaa siis gammajakaumaa  $\text{Gamma}(3, \theta/3)$ .  $\square$

## 6.3 Normaalijakaumaan liittyvät jakaumat

**Lause 6.5** Jos  $X_1, X_2, \dots, X_n$  on otos normaalijakaumasta  $N(\mu, \sigma^2)$ , niin otoskeskiarvon  $\bar{X} = \sum_{i=1}^n (1/n)X_i$  jakauma on  $N(\mu, \sigma^2/n)$ .

**Todistus.** Koska  $X_i \sim N(\mu, \sigma^2)$ , niin

$$M_{X_i}(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Seurauslauseen 6.1 mukaan

$$\begin{aligned} M_{\bar{X}}(t) &= \left[\exp\left(\mu \cdot \frac{t}{n} + \frac{\sigma^2 (t/n)^2}{2}\right)\right]^n \\ &= \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right), \end{aligned}$$

joka on normaalijakauman  $N(\mu, \sigma^2/n)$  momenttifunktio. Koska momenttifunktio määrittää yksikäsitteisesti satunnaismuuttujan jakauman, niin  $\bar{X} \sim N(\mu, \sigma^2/n)$ .  $\square$

**Lause 6.6** Olkoot  $X_i \sim \chi^2(r_i)$ ,  $i = 1, 2, \dots, n$ . Jos  $X_1, X_2, \dots, X_n$  ovat riippumattomat, niin satunnaismuuttujan  $Y = X_1 + X_2 + \dots + X_n$  jakauma on  $\chi^2(r_1 + r_2 + \dots + r_k)$ .

**Todistus.** Satunnaismuuttujan  $Y$  momenttifunktio voidaan kirjoittaa muodossa

$$\begin{aligned} M_Y(t) &= E[e^{t(X_1+X_2+\dots+X_n)}] \\ &= E(e^{tX_1}) \cdot E(e^{tX_2}) \dots E(e^{tX_n}) \\ &= \prod_{i=1}^n M_{X_i}(t). \end{aligned}$$

Koska

$$M_{X_i}(t) = (1 - 2t)^{-r_i/2}; \quad t < \frac{1}{2},$$

niin

$$M_Y(t) = \prod_{i=1}^n (1 - 2t)^{-r_i/2} = (1 - 2t)^{-(r_1+r_2+\dots+r_n)/2}, \quad t < \frac{1}{2}$$

on  $\chi^2$ -jakauman  $\chi^2(r_1 + r_2 + \dots + r_n)$  momenttifunktio. Tästä seuraa, että  $Y \sim \chi^2(r_1 + r_2 + \dots + r_n)$ .  $\square$

**Lause 6.7** Olkoon  $Z_1, Z_2, \dots, Z_n$  otos standardimuotoisesta normaalijakaumasta  $N(0, 1)$ . Silloin  $W = Z_1^2 + Z_2^2 + \dots + Z_n^2$  noudattaa jakaumaa  $\chi^2(n)$ .

**Todistus.** Koska  $Z_i \sim \chi^2(1)$ ,  $i = 1, 2, \dots, n$  ja  $Z_1^2, Z_2^2, \dots, Z_n^2$  ovat keskenään riippumattomat, niin tulos seuraa Lauseesta 6.6.  $\square$

**Seuraus 6.2** Olkoot  $X_1, X_2, \dots, X_n$  riippumattomat ja  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$ . Silloin satunnaismuuttuja

$$W = \sum_{i=1}^n \frac{(X_i - \mu_i)^2}{\sigma_i^2}$$

noudattaa jakaumaa  $\chi^2(n)$ .

**Lause 6.8** Olkoon  $X_1, X_2, \dots, X_n$  otos normaalijakaumasta  $N(\mu, \sigma^2)$ ,  $\bar{X} = \sum_{i=1}^n (1/n)X_i$  on otoskeskiarvo ja  $S^2 = [1/(n-1)] \sum_{i=1}^n (X_i - \bar{X})^2$  on otosvarianssi. Silloin

1.  $\bar{X}$  ja  $S^2$  ovat riippumattomat satunnaismuuttujat,
2.  $\bar{X}$  noudattaa normaalijakaumaa  $N(\mu, \sigma^2/n)$ ,
3.  $(n-1)S^2/\sigma^2$  noudattaa  $\chi^2$ -jakaumaa  $\chi^2(n-1)$ .

**Todistus.** Kohdan 1 todistus sivuutetaan tässä yhteydessä. Kohta 2 on lause 6.5. Todistetaan nyt väite, että

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Laskemalla voidaan todeta, että

$$\begin{aligned} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 &= \sum_{i=1}^n \left[ \frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right]^2 \\ &= \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \\ &= \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2. \end{aligned}$$

Koska  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ , niin  $Z^2 \sim \chi^2(1)$ . Vastaavasti Seurauslauseen 6.2 mukaan  $W = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$ .

Koska  $S^2$  ja  $Z^2$  ovat kohdan 1 mukaan riippumattomat, niin

$$\begin{aligned} E(e^{tW}) &= E(e^{t[(n-1)S^2/\sigma^2 + Z^2]}) = E(e^{t(n-1)S^2/\sigma^2} \cdot e^{tZ^2}) \\ &= E(e^{t(n-1)S^2/\sigma^2}) E(e^{tZ^2}). \end{aligned}$$

Koska  $W \sim \chi^2(n)$  ja  $Z^2 \sim N(0, 1)$ , niin

$$(1 - 2t)^{-n/2} = E[e^{t(n-1)S/\sigma^2}] \cdot (1 - 2t)^{-1/2}.$$

Tästä seuraa, että

$$E[e^{t(n-1)S/\sigma^2}] = (1 - 2t)^{-(n-1)/2}; \quad t < \frac{1}{2},$$

joka on jakauman  $\chi^2(n-1)$  momenttifunktio. Näin on lauseen väite 3 todistettu.  $\square$

**Lause 6.9** *Olkoot  $X_1, X_2, \dots, X_n$  keskenään riippumattomat normaalijakaumaa noudattavat satunnaismuuttujat, joiden odotusarvot ovat  $\mu_1, \mu_2, \dots, \mu_n$  ja varianssit  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ . Silloin lineaarikombinaatio*

$$Y = \sum_{i=1}^n a_i X_i$$

*noudattaa normaalijakaumaa*

$$N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

**Todistus.** Tulos saadaan soveltamalla Lausetta 6.4 normaalijakaumaan.  $\square$

## 6.4 Järjestyssuureet

Otoksen suurin ja pienin arvo sekä keskimääräinen arvo, mediaani, ovat tärkeitä otossuureiden arvojen järjestykseen perustuvia tunnuslukuja. Olkoon  $X_1, X_2, \dots, X_n$  otos. Merkitään otoksen pienintä arvoa  $X_{(1)}$  seuraavaksi pienintä  $X_{(2)}$  ja niin edelleen, joten

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

Tämä indeksointi tarkoittaa sitä, että otosarvot pannaan kasvavaan järjestykseen. Jos otos on esimerkiksi 5.0, 3.1, 2.7, 6.1, 5.3, niin järjestetty otos on 2.7, 3.1, 5.0, 5.3, 6.1. Nyt siis esimerkiksi  $X_1 = 5.0$ ,  $X_{(1)} = 2.7$  ja  $X_{(3)} = 5.0$  on mediaani ja  $X_3 = 2.7$ . Nyt siis

$$X_{(1)} = \min(X_1, \dots, X_n)$$

ja

$$X_{(n)} = \max(X_1, \dots, X_n).$$

Tunnusluku  $X_{(k)}$  on otoksen  $k$ . järjestystunnusluku.

### 6.4.1 Maksimi ja minimi

Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka kertymäfunktio on  $F(x)$ . Maksimin kertymäfunktio on

$$\begin{aligned} F_{(n)}(x) &= P(X_{(n)} \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x) \cdots P(X_n \leq x), \end{aligned}$$

koska  $X_1, X_2, \dots, X_n$  ovat riippumattomat. Kertymäfunktion määritelmän mukaan  $P(X_i \leq x) = F(x)$ , joten

$$F_{(n)}(x) = [F(x)]^n.$$

Minimin kertymäfunktio on

$$\begin{aligned} F_{(1)}(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x, X_2 > x, \dots, X_n > x) \\ &= 1 - P(X_1 > x)P(X_2 > x) \cdots P(X_n > x) \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

**Esimerkki 6.7** Olkoon  $X_1, X_2, \dots, X_n$  otos eksponenttijakaumasta  $\text{Exp}(\lambda)$ . Määritetään minimin  $X_{(1)}$  jakauma. Eksponenttijakauman  $\text{Exp}(\lambda)$  kertymäfunktio on

$$F(x) = \begin{cases} 0, & \text{kun } x < 0; \\ 1 - e^{-\lambda x}, & \text{kun } x \geq 0. \end{cases}$$

Silloin minimin kertymäfunktio on

$$F_{(1)}(x) = \begin{cases} 0, & \text{kun } x < 0; \\ 1 - e^{-n\lambda x}, & \text{kun } x \geq 0. \end{cases}$$

Minimi noudattaa siis eksponenttijakaumaa  $\text{Exp}(n\lambda)$ .  $\square$

Jos otos on jatkuvasta jakaumasta, jonka tiheysfunktio on  $f(x)$ , saadaan  $X_{(1)}$ :n ja  $X_{(n)}$ :n jakaumien tiheysfunktiot derivoimalla kertymäfunktiot  $F_{(n)}(x)$  ja  $F_{(1)}(x)$ . Nyt siis maksimin tiheysfunktio on

$$f_{(n)}(x) = \frac{d}{dx}[F(x)]^n = n[F(x)]^{n-1}f(x)$$

ja minimin tiheysfunktio on

$$f_{(1)}(x) = \frac{d}{dx}(1 - [1 - F(x)]^n) = n[1 - F(x)]^{n-1}f(x).$$

### 6.4.2 Järjestyssuureen $X_{(k)}$ jakauma

Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka kertymäfunktio on  $F(x)$ . Johdetaan nyt järjestystunnusluvun  $X_{(k)}$ ,  $1 < k < n$ , jakauma. Jos  $\{X_{(k)} \leq x\}$ , niin silloin ainakin  $k$  otosarvoa on pienempiä tai korkeintaan yhtä suuria kuin  $x$ . Tarkastellaan yksinkertaisuuden vuoksi tapausta  $n = 3$ . Johdetaan mediaanin  $X_{(2)}$  jakauma. Tapahtuma  $\{X_{(2)} \leq x\}$  toteutuu täsmälleen silloin, kun  $\{X_1 \leq x, X_2 \leq x\}$  tai  $\{X_1 \leq x, X_3 \leq x\}$  tai  $\{X_2 \leq x, X_3 \leq x\}$  tai  $\{X_1 \leq x, X_2 \leq x, X_3 \leq x\}$ . Koska

$$P(X_i \leq x, X_j \leq x) = [F(x)]^2[1 - F(x)], \quad i \neq j$$

ja

$$P(X_1 \leq x, X_2 \leq x, X_3 \leq x) = [F(x)]^3,$$

niin

$$\begin{aligned} F_{(2)}(x) &= P(X_{(2)} \leq x) = 3[F(x)]^2[1 - F(x)] + [F(x)]^3 \\ (6.4.1) \quad &= \sum_{i=2}^3 \binom{3}{i} [F(x)]^i [1 - F(x)]^{3-i}. \end{aligned}$$

Yleisessä tapauksessa vastaava kaava voidaan johtaa samalla periaatteella. Emme kuitenkaan käsittele yleisen kaavan johtoa sen tarkemmin, toteamme vain, että  $X_{(k)}$ :n kertymäfunktio on

$$(6.4.2) \quad F_{(k)}(x) = P(X_{(k)} \leq x) = \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i}.$$



Jos otos on jatkuvasta jakaumasta, saadaan vastaava tiheysfunktio derivoimalla kertymäfunktio. Esitetään ensin  $X_{(2)}$ :n tiheysfunktio, kun  $n = 3$ . Kun kertymäfunktio (6.4.1) derivoidaan, saadaan

$$\begin{aligned} f_{(2)}(x) &= F'_{(2)}(x) = 3 \cdot 2F(x)f(x)[1 - F(x)] - 3[F(x)]^2f(x) + 3[F(x)]^2f(x) \\ &= 3!F(x)[1 - F(x)]f(x). \end{aligned}$$

Derivoimalla lauseke (6.4.2) saadaan satunnaismuuttujan  $X_{(k)}$  tiheysfunktio yleisessä tapauksessa ( $1 \leq k \leq n$ ):

$$(6.4.3) \quad f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x).$$

**Esimerkki 6.8** Olkoon  $X_1, X_2, X_3, X_4, X_5$  otos jakaumasta, jonka tiheysfunktio on  $f(x) = 2x$ ,  $0 < x < 1$ . Jakauman kertymäfunktio on

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

Silloin mediaanin tiheysfunktio on lausekkeen (6.4.3) nojalla

$$f_{(3)}(x) = \frac{5!}{2!2!} x^4 (1 - x^2)^2 \cdot 2x = 60x^5 (1 - x^2)^2, \quad 0 < x < 1.$$

Vastaavasti minimin tiheysfunktio on

$$f_{(1)}(x) = 10x(1 - x^2)^4, \quad 0 < x < 1$$

ja maksimin tiheysfunktio on

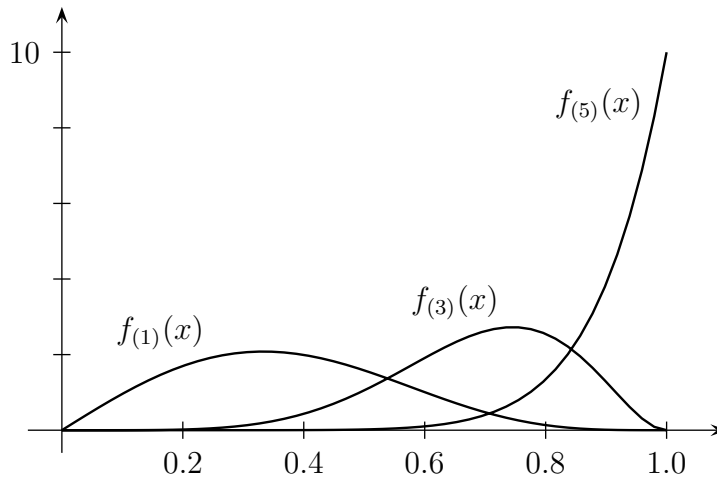
$$f_{(5)}(x) = 10x^9, \quad 0 < x < 1.$$

□

## 6.5 Keskeinen rajaväittäjä

Olemme havainneet, että otossuureen jakauma riippuu tavallisesti otoskoosta  $n$ . Jos  $X_1, X_2, \dots, X_n$  on otos Bernoullin jakaumasta  $\text{Ber}(p)$ , niin  $X = X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p)$ . Satunnaismuuttujan  $X$  jakauma riippuu siis otoskoosta  $n$ . Jos  $X_1, X_2, \dots, X_n$  on otos normaalijakaumasta  $N(\mu, \sigma^2)$ , niin otoskeskiarvon  $\bar{X}$  jakauma  $N(\mu, \sigma^2/n)$  riippuu  $n$ :stä.

Olkoon  $(X_i; i \geq 1) = X_1, X_2, X_3, \dots$  satunnaismuuttujien jono, missä satunnaismuuttujien  $X_n$ ,  $n = 1, 2, 3, \dots$  jakauma riippuu  $n$ :stä. Merkitään satunnaismuuttujan  $X_n$  kertymäfunktioita  $F_n(x)$ , joka siis riippuu  $n$ :stä. Seuraavassa määritellään satunnaismuuttujien jonon  $(X_i; i \geq 1)$  *suppeneminen jakaumamielellä*.



**Kuvio 6.1.** Minimim, maksimin ja mediaanin tiheysfunktio, kun otos on jakaumasta  $f(x) = 2x$ ,  $0 < x < 1$ .

**Määritelmä 6.1** Satunnaismuuttujien  $X_1, X_2, X_3, \dots$  jono *suppenee jakaumaltaan* kohti satunnaismuuttujaa  $X$ , jos  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$  kaikissa pisteissä  $x$ , joissa  $F(x)$  on jatkuva.

Kun jono  $\{X_n\}$  suppenee jakaumaltaan kohti satunnaismuuttujaa  $X$ , merkitään  $X_n \xrightarrow{d} X$ . Momenttifunktioiden yhteydessä esitettiin momenttifunktion ja jakauman (kertymäfunktion) yksikäsitteistä vastaavuutta koskeva Lause 3.12. Samassa yhteydessä esitettiin myös momenttifunktioiden suppenemista koskeva Lause 3.15, jota voidaan soveltaa raja-jakaumien määrittämiseen.

Merkitään satunnaismuuttujien  $X_1, X_2, \dots, X_n$  summaa ja keskiarvoa seuraavasti:

$$S_n = \sum_{i=1}^n X_i \quad \text{ja} \quad \bar{X}_n = \frac{S_n}{n}$$

**Lause 6.10 (Keskeinen rajaväittäjä)** Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka keskiarvo on  $\mu$  ja varianssi  $\sigma^2$ . Merkitään

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Silloin  $Z_n$ :n jakauma lähenee normaalijakaumaa  $N(0, 1)$ , kun  $n \rightarrow \infty$ .

Keskeisen rajaväittäjän mukaan riippumattomien satunnaismuuttujien summa noudattaa likimain normaalijakaumaa, kun  $n$  on suuri. Merkitsemme

$$Z_n \simeq N(0, 1),$$

kun  $n$  on suuri. Merkki  $\simeq$  tarkoittaa ”noudattaa likimain jakaumaa”. Käytännössä keskeisen rajaväittäjän avulla voidaan arvioida  $Z_n$ :n jakaumaa, kun

$n$  on riittävän suuri. Silloin

$$P(Z_n \leq z) \approx \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dx = \Phi(z),$$

missä  $\Phi(z)$  on normitetun normaalijakauman kertymäfunktio. Voimme merkitä saman asian myös seuraavasti:

$$P(Z_n \leq z) \rightarrow \Phi(z),$$

kun  $n \rightarrow \infty$ .

**Esimerkki 6.9** Olkoon  $X_1, X_2, \dots, X_{15}$  otos jakaumasta, jonka tiheysfunktio on  $f(x) = \left(\frac{3}{2}\right)x^2$ ,  $-1 < x < 1$ . Jakauman odotusarvo  $\mu = 0$  ja varianssi  $\sigma^2 = 3/5$ . Esimerkiksi todennäköisyys  $P(\bar{X} \leq 0.15)$  voidaan laskea joutamalla ensin  $\bar{X}$ :n jakauma ja määrittämällä siitä kysytty todennäköisyys. Keskeisen rajaväittämän avulla saadaan tämän todennäköisyyden tarkka arvio ilman tietoa  $\bar{X}$ :n tarkasta jakaumasta:

$$\begin{aligned} P(\bar{X} \leq 0.15) &= P\left(\frac{\bar{X} - 0}{\sqrt{3/5}/\sqrt{15}} \leq \frac{0.15 - 0}{\sqrt{3/5}/\sqrt{15}}\right) \\ &= P(Z_{15} \leq 0.75) \\ &\approx \Phi(0.75) = 0.7734. \end{aligned}$$

Arvion tarkkuudesta keskeinen rajaväittäjä ei kuitenkaan anna käsitystä.  $\square$

## 6.6 Jakaumien likiarvot normaalijakauman avulla

Olkoon  $X_1, X_2, \dots, X_n$  otos Bernoullin jakaumasta  $\text{Ber}(p)$ . Silloin  $S_n = X_1 + X_2 + \dots + X_n$  noudattaa binomijakaumaa  $\text{Bin}(n, p)$ . Keskeisen rajaväittämän mukaan

$$Z_n = \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

noudattaa likimain normaalijakaumaa  $N(0, 1)$ , kun  $n$  on suuri. Tuloksen mukaan binomijakauma lähenee normaalijakaumaa, kun  $n$  kasvaa. Peukalosääntönä voidaan pitää, että  $n$  on riittävän suuri, kun  $np \geq 5$  ja  $n(1-p) \geq 5$ . Mitä enemmän  $p$  poikkeaa 0.5:stä, sitä suurempi  $n$  tarvitaan.

**Esimerkki 6.10** Oletetaan, että  $X \sim \text{Bin}(10, 0.5)$ . Lasketaan todennäköisyys  $P(3 \leq X < 6)$ . Voidaan kirjoittaa

$$P(3 \leq X < 6) = P(2.5 \leq X \leq 5.5).$$

Arvioidaan nyt jälkimmäistä todennäköisyyttä keskeisen rajaväittämän nojalla normaalijakauman avulla. Silloin

$$\begin{aligned} P(2.5 \leq X \leq 5.5) &= P\left(\frac{2.5 - 5}{\sqrt{10/4}} \leq \frac{X - 5}{\sqrt{10/4}} \leq \frac{5.5 - 5}{\sqrt{10/4}}\right) \\ &\approx \Phi(0.316) - \Phi(-1.581) = 0.5670. \end{aligned}$$

Tarkka todennäköisyys binomijakauman avulla on  $P(3 \leq X < 6) = 0.5683$ .  $\square$

## 6.7 $t$ -jakauma ja $F$ -jakauma

Oletetaan, että  $X_1, X_2, \dots, X_n$  on otos jakaumasta  $N(\mu, \sigma^2)$ , jonka varianssi  $\sigma^2$  tunnetaan. Tarkastellaan lauseketta

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}},$$

joka tunnetaan  $t$ -testisuureena. Tiedämme, että

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

ja Lauseen 6.8 mukaan

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Lisäksi Lauseen 6.8 mukaan  $Z$  ja  $U$  ovat riippumattomat. Tällainen satunnaismuuttuja noudattaa  $t$ -jakaumaa vapausastein  $r = n - 1$ . Alaluvussa 5.9.2 esitettiin  $t$ -jakauman tiheysfunktio.

Usein halutaan verrata kahden normaalijakauman  $N(\mu_1, \sigma_1^2)$  ja  $N(\mu_2, \sigma_2^2)$  variansseja. Teemme  $n_1$ :n kokoisen otoksen jakaumasta  $N(\mu_1, \sigma_1^2)$  ja  $n_2$ :n kokoisen otoksen jakaumasta  $N(\mu_2, \sigma_2^2)$ . Oletetaan, että otokset ovat toisistaan riippumattomat. Olkoot  $S_1^2$  ja  $S_2^2$  näistä eri otoksista lasketut otosvarianssit. Lauseen 6.8 mukaan

$$U = (n_1 - 1) \frac{S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{ja} \quad V = (n_2 - 1) \frac{S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1).$$

Koska otokset ovat keskenään riippumattomat, niin satunnaismuuttujat  $U$  ja  $V$  ovat riippumattomat. Varianssien yhtäsuuruutta voidaan testata tarkastelemalla suhdetta

$$(6.7.1) \quad F = \frac{U/r_1}{V/r_2},$$

missä  $r_1 = n_1 - 1$  ja  $r_2 = n_2 - 1$ . Alaluvussa 5.9.2 osoitettiin, että suhde (6.7.1) noudattaa  $F$ -jakaumaa vapausastein  $r_1$  ja  $r_2$ .

## 6.8 Momenttifunktion rajafunktiot

Tarkastelemme nyt satunnaismuuttujan (usein otoksen tunnusluku) jakauman riippuvuutta otoskoosta  $n$ . Otoskeskiarvo ja otosvarianssi ovat tavallisimmat otoksesta lasketut tunnusluvut. Oletetaan esimerkiksi, että  $X_n \sim \text{Bin}(n, p)$ . Eri  $n$ :n arvoilla saamme eri binomijakauman. Miten jakauma muuttuu  $n$ :n kasvaessa? Olemme keskeisen rajaväittämän avulla jo osoittaneet, että  $\text{Bin}(n, p)$  lähenee normaalijakaumaa, kun  $n$  kasvaa.

Voimme tutkia  $\text{Bin}(n, p)$ :n rajajakaumaa myös ehdolla, että jakauman odotusarvo  $np$  pidetään vakiona  $\lambda$ . Jos  $np = \lambda$  on vakio ja  $n \rightarrow \infty$ , niin  $p \rightarrow 0$ . Satunnaismuuttujan  $X_n \sim \text{Bin}(n, p)$  momenttifunktio on

$$M_n(t) = (1 - p + pe^t)^n.$$

Koska  $p = \lambda/n$ , niin

$$M_n(t) = \left[1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right]^n = \left[1 + \frac{\lambda(e^t - 1)}{n}\right]^n.$$

Käyttäen hyväksi analyysin tulosta

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a,$$

saadaan

$$\lim_{n \rightarrow \infty} M_n(t) = e^{\lambda(e^t - 1)} = M(t),$$

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

$$M(t) = e^{\lambda(e^t - 1)}$$

on Poissonin jakauman  $\text{Poi}(\lambda)$  momenttifunktio, niin Lauseen 3.15 mukaan  $X_n$ :n jakauma lähestyy siis Poissonin jakaumaa  $\text{Poi}(\lambda)$ , kun  $n \rightarrow \infty$ .

## 6.9 Suppenemiskäsitteet

Olemme edellä jo useaan otteeseen tutustuneet suppenemiseen jakaumamielessä. Käsite määriteltiin alaluvussa 4.3.4 (Määritelmä 4.2). Satunnaismuuttujien jono  $(X_n, n \geq 1) = (X_1, X_2, \dots)$  suppenee jakaumaltaan kohti satunnaismuuttujaa  $X$  ( $X_n \xrightarrow{d} X$ , kun  $n \rightarrow \infty$ ), jos

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

kaikissa pisteissä  $x$ , joissa  $F_X(x)$  on jatkuva. Tässä yhteydessä on myös syytä muistaa, että momenttifunktioiden jonon suppenemisestä seuraa vastaavien jakaumien suppeneminen jakaumamielessä.

**Esimerkki 6.11** Olkoon  $\{X_n\}$  sellainen satunnaismuuttujien jono, että

$$p_n(x) = P(X_n = x) = \begin{cases} 1, & \text{kun } x = 2 + \frac{1}{n}; \\ 0, & \text{kun } x \neq 2 + \frac{1}{n}. \end{cases}$$

Huomaa, että  $p_n(2) = 0$  kaikilla  $n$ . Tästä seuraa, että  $p_n(x) \rightarrow p(x)$ , missä  $p(x) = 0$  kaikilla  $x$ . Satunnaismuuttujan  $X_n$  kertymäfunktio on muotoa

$$F_n(x) = \begin{cases} 0, & \text{kun } x < 2 + \frac{1}{n}; \\ 1, & \text{kun } x \geq 2 + \frac{1}{n}. \end{cases}$$

Kun  $n \rightarrow \infty$ , niin  $F_n(x) \rightarrow F(x)$ , missä

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

$F(x)$  on pisteeseen  $x = 2$  degeneroituneen jakauman kertymäfunktio eli  $P(X = 2) = 1$ . Todennäköisyysfunktioiden  $p_n(x)$  jono ei kuitenkaan supene kohti tämän jakauman todennäköisyysfunktioita.  $\square$

Olkoon  $\{X_n\}$  jono satunnaismuuttujia, joiden odotusarvo on  $\mu$  ja varianssi  $\sigma^2$ . Silloin keskeisen rajaväittämän mukaan

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

missä  $Z \sim N(0, 1)$ . Huomattakoon, että  $Z_n$ :n jakaumat ovat usein diskreettejä, mutta silti rajajakauma on normaalijakauma. Kun  $n$  on riittävän suuri, niin

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) \approx \Phi(b) - \Phi(a).$$

Jos esimerkiksi  $X_n \sim \text{Bin}(n, p)$ , niin silloin keskeisen rajaväittämän mukaan

$$\frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{d} Z,$$

missä  $Z \sim N(0, 1)$ . Tätä tulosta kutsutaan *De Moivren ja Laplacen* lauseeksi.

Osoitimme alavuvussa 3.4 Tšebyševin epäyhtälön avulla, että otoskeskiarvo  $\bar{X}$  on hyvä populaation keskiarvon tunnusluku. Tarkastelu ei perustunut suppenemiseen jakaumamielessä vaan ns. stokastiseen suppenemiseen.

**Määritelmä 6.2** Satunnaismuuttujien jono  $\{X_n\}$  suppenee stokastisesti kohti satunnaismuuttujaa  $X$ , jos kaikilla  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0$$

tai yhtäpitävästi

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1.$$

Stokastista suppenemista sanotaan myös suppenemiseksi todennäköisyyden mielessä ja merkitään  $X_n \xrightarrow{P} X$ . Usein tarkastellaan tilannetta, että satunnaismuuttuja, jota lähestytään, on vakio. Tällainen tilanne on heikossa suurten lukujen laissa (Lause 3.11, HSLI). Esitetty heikon suurten lukujen lain todistus oli sillä tavalla yleinen, että se on pätevä myös jatkuville satunnaismuuttujille. HSLI sanoo, että otoskeskiarvo suppenee stokastisesti kohti populaation keskiarvoa, kun otoskoko kasvaa.

Olkoon  $\{X_n\}$  sellaisten satunnaismuuttujien jono, että  $E(X_n) = \mu$  ja  $\text{Var}(X_n) = \sigma^2$ . Heikon suurten lukujen lain mukaan

$$\bar{X}_n \xrightarrow{P} \mu,$$

missä  $\bar{X}_n = (X_1 + X_2 + \dots + X_n)/n$ . Lause todistettiin Tšebyševin epäyhtälön avulla.

**Esimerkki 6.12** Olkoon  $\{X_n\}$  jono sellaisia diskreettejä satunnaismuuttujia, että

$$P(X_n = 1) = \frac{1}{n} \quad \text{ja} \quad P(X_n = 0) = 1 - \frac{1}{n}.$$

Silloin

$$P(|X_n| > \varepsilon) = \begin{cases} P(X_n = 1) = \frac{1}{n}, & \text{kun } 0 < \varepsilon < 1; \\ 0, & \text{kun } \varepsilon \geq 1. \end{cases}$$

Tästä nähdään, että  $P(|X_n| > \varepsilon) \rightarrow 0$ , kun  $n \rightarrow \infty$ . Voimme siis sanoa, että  $X_n \xrightarrow{P} 0$ .  $\square$

**Esimerkki 6.13 (Otosvarianssin tarkentuvuus)** Olkoon  $\{X_n\}$  sellainen satunnaismuuttujien jono, että  $E(X_n) = \mu$  ja  $\text{Var}(X_n) = \sigma^2 < \infty$ . Otosvarianssi on

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Tiedämme, että  $E(S_n^2) = \sigma^2$ . Tšebyševin epäyhtälön mukaan

$$P(|S_n^2 - \sigma^2| \geq \varepsilon) \leq \frac{E(S_n^2 - \sigma^2)^2}{\varepsilon^2} = \frac{\text{Var}(S_n^2)}{\varepsilon^2}.$$

Jos nyt  $\text{Var}(S_n^2) \rightarrow 0$ , kun  $n \rightarrow \infty$ , niin  $\lim_{n \rightarrow \infty} P(|S_n^2 - \sigma^2| \geq \varepsilon) = 0$  ja  $(S_n^2, n \geq 1)$  suppenee stokastisesti kohti populaation varianssia.  $\square$

Tässä yhteydessä on tietysti luonnollista kysyä, miten stokastinen suppeneminen ja suppeneminen jakaumamielessä suhteutuvat toisiinsa. Voidaan osoittaa, että stokastinen suppeneminen implikoi suppenemisen jakaumamielessä. Jos siis  $X_n \xrightarrow{P} X$ , niin  $X_n \xrightarrow{d} X$ . Jos jono  $\{X_n\}$  suppenee kohti vakiota  $\mu$ , niin silloin  $X_n \xrightarrow{P} \mu$  jos ja vain jos  $X_n \xrightarrow{d} \mu$ .

Rajoitumme tässä esityksessä kahteen edellä esitettyyn suppenemiskäsitteeseen: stokastiseen suppenemiseen ja suppenemiseen jakaumamielessä. Esitämme kuitenkin vielä ns. melkein varman (m.v.) suppenemisen.

**Määritelmä 6.3** Jono  $\{X_n\}$  suppenee melkein varmasti kohti satunnaisuuttujaa  $X$ , jos

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1.$$

Näennäisesti määritelmä muistuttaa stokastisen suppenemisen määritelmää, vaikka käsitteet ovat sisällöllisesti erilaisia.

## 6.10 Estimaattorit

### 6.10.1 Estimaattoreiden ominaisuuksia

Olkoon  $X_1, X_2, \dots, X_n$  otos jakaumasta, jonka tiheysfunktio on  $f(x, \theta)$ . Jos haluamme estimoida jakauman tunnuslukua  $\theta$  jollakin otoksen tunnusluvulla, merkitsemme usein tätä otoksen tunnuslukua  $\hat{\theta}$ . On siis muistettava, että  $\hat{\theta}$  on otoksen funktio ja täydellisempi merkintä olisi  $\hat{\theta}(X_1, X_2, \dots, X_n)$ . Havaitusta otoksesta  $x_1, x_2, \dots, x_n$  laskettua estimaattorin arvoa  $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$  sanotaan estimaatiksi. Olemme edellä jo tarkastelleet useita estimaattoreita. Tavanomaisia odotusarvon  $\mu$  ja varianssin  $\sigma^2$  estimaattoreita ovat otoskeskiarvo  $\bar{X}$  ja otosvariassi  $S^2$ , eli  $\hat{\mu} = \bar{X}$  ja  $\hat{\sigma}^2 = S^2$ .

**Määritelmä 6.4** Estimaattori  $\hat{\theta}$  on parametrin *harhaton estimaattori*, jos  $E(\hat{\theta}) = \theta$  kaikilla  $\theta$ :n arvoilla. Muutoin  $\hat{\theta}$  on harhainen ja  $\hat{\theta}$ :n harha on  $\text{harha}(\hat{\theta}) = E(\hat{\theta}) - \theta$ .

Olemme jo aikaisemmin osoittaneet, että  $\hat{\mu} = \bar{X}$  ja  $\hat{\sigma}^2 = S^2$  ovat harhatomia estimaattoreita.

Eräs intuitiivisesti hyväksyttävä estimaattorille asetettava vaatimus on, että se antaa 'tarkempia' estimaatteja kun otoskoko kasvaa. Tarkan estimaattorin arvot osuvat suurella todennäköisyydellä lähelle parametrin  $\theta$  oikeata arvoa. *Tarkentuvuus* sisältää tämän ajatuksen.

**Määritelmä 6.5** Tunnusluku  $\hat{\theta}$  on parametrin  $\theta$  *tarkentuva estimaattori*, jos  $\hat{\theta} \xrightarrow{P} \theta$ , kun otoskoko  $n$  kasvaa rajatta.

Selvempi olisi merkitä  $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \dots, X_n)$ , missä  $(\hat{\theta}_n; n \geq 1)$  on satunnaisuuttujien jono. Jos  $\hat{\theta}_n$  on  $\theta$ :n tarkentuva estimaattori, niin jono  $(\hat{\theta}_n; n \geq 1)$  suppenee stokastisesti kohti parametrin arvoa  $\theta$ .

**Määritelmä 6.6** Estimaattorin  $\hat{\theta}$  keskineliövirhe (MSE = Mean Square Error) on

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

Määritelmästä seuraa suoraviivaisesti, että

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{harha}(\hat{\theta})]^2.$$

Voidaan osoittaa, että  $\hat{\theta}$  on  $\theta$ :n tarkentuva estimaattori, jos  $\text{MSE}(\hat{\theta}) \rightarrow 0$  otoskoon  $n$  kasvaessa rajatta.



### 6.10.2 Delta-menetelmä

**Määritelmä 6.7** Funktion  $g(x)$   $r$ . asteen *Taylorin polynomi* pisteessä  $a$  on

$$(6.10.1) \quad T_r(x) = g(a) + g'(a)(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots + \frac{g^{(r)}(a)}{r!}(x-a)^r,$$

missä  $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$  on funktion  $g(x)$   $r$ . derivaatta.

*Taylorin lauseen* mukaan

$$(6.10.2) \quad \lim_{x \rightarrow a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0,$$

jos  $g^{(r)}(a)$  on olemassa. Funktio  $g(x)$  voidaan lausua pisteen  $x = a$  ympäristössä muodossa

$$g(x) = T_r(x) + R_{r+1}(x),$$

missä  $R_{r+1}(x) = g(x) - T_r(x)$  on jäännöstermi, joka siis toteuttaa ehdon 6.10.2.

Oletetaan, että  $X$  on satunnaismuuttuja, jonka odotusarvo on  $E(X) = \mu \neq 0$ . Jos estimoidaan funktiota  $g(\mu)$ , niin sen Taylorin polynomiin perustuva 1. kertaluvun likiarvo pisteessä  $\mu$  on

$$(6.10.3) \quad g(X) = g(\mu) + g'(\mu)(X - \mu).$$

Jos käytetään  $g(\mu)$ :n estimaattorina funktiota  $g(X)$ , niin

$$E[g(X)] \approx g(\mu)$$

ja

$$\text{Var}[g(X)] = [g'(\mu)]^2 \text{Var}(X).$$

**Esimerkki 6.14** Tarkastellaan odotusarvon  $E(X) = \mu \neq 0$  funktion  $g(\mu) = 1/\mu$  estimointia. Olkoon estimaattorina  $1/X$ . Silloin edellisen mukaan

$$E\left(\frac{1}{X}\right) \approx \frac{1}{\mu}$$

ja

$$\text{Var}\left(\frac{1}{X}\right) \approx \left(\frac{1}{\mu}\right)^4 \text{Var}(X).$$

□

**Lause 6.11 (Delta-menetelmä)** *Olkoon  $\{X_n\}$  sellainen satunnaismuuttujien jono, että  $\sqrt{n}(X_n - \theta)$  lähenee jakaumamielessä normaalijakaumaa  $N(0, \sigma^2)$ . Oletetaan, että annetulla funktiolla  $g$  on määrättyllä arvolla  $\theta$  derivaatta  $g'(\theta) \neq 0$ . Silloin*

$$\sqrt{n}[g(X_n) - g(\theta)] \rightarrow N(0, \sigma^2[g'(\theta)]^2)$$

*jakaumamielessä.*

**Esimerkki 6.15** Olkoon  $X_1, \dots, X_n$  otos jakaumasta  $\text{Ber}(p)$ . Onnistumisen todennäköisyyden  $p$  estimaattori on tavallisesti  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$ . Onnistumisen mahdollisuus (odds)  $p/(1-p)$  on vedonlyönnissä ja biostatistikassa tavantomainen parametri. Voimme käyttää  $p/(1-p)$ :n estimaattorina  $\hat{p}$ :n funktiota  $\hat{p}/(1-\hat{p})$ . Mitä voimme sanoa tämän estimaattorin ominaisuuksista? Nyt estimoidaan siis funktiota  $g(p) = p/(1-p)$ . Koska  $g'(p) = 1/(1-p)^2$ , niin lausekkeen 6.10.3 mukaan

$$\begin{aligned} \text{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx [g'(p)]^2 \text{Var}(\hat{p}) \\ &= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}. \end{aligned}$$

□