Todennäköisyyslaskenta I

Ville Hyvönen, Patrik Lauha, Topias Tolonen[1]

Kesä 2018

[1]Virheitä ja kehitysehdotuksia otetaan vastaan jatkuvasti osoitteeseen *topias.tolonen@helsinki.fi*. Kiitos palautteestasi!

Sisältö

0 Intro **3**
0.1 Esipuhe . . . 3
0.2 Mitä on todennäköisyys? . . . 4

1 Joukko-oppia ja kombinatoriikkaa **8**
1.1 Joukko-opin termien kertaus . . . 8
1.2 Kombinatoriikkaa . . . 10
1.2.1 Tuloperiaate . . . 11
1.2.2 Summaperiaate . . . 13
1.2.3 Permutaatiot ja kombinaatiot . . . 13
1.3 Harjoitustehtäviä . . . 20

2 Matemaattinen todennäköisyys **22**
2.1 Klassinen todennäköisyys . . . 22
2.1.1 Otanta . . . 26
2.2 Todennäköisyyden aksioomat . . . 30
2.3 Todennäköisyyksien tulkinnasta . . . 41
2.4 Harjoitustehtäviä . . . 43

3 Ehdollinen todennäköisyys ja riippumattomuus **45**
3.1 Ehdollinen todennäköisyys . . . 45
3.2 Kokonaistodennäköisyys ja Bayesin kaava . . . 47
3.3 Tapahtumien riippumattomuus . . . 52
3.4 Toistokokeet . . . 55
3.5 Harjoitustehtäviä . . . 57

4 Satunnaismuuttujat **59**
4.1 Diskreettejä jakaumia . . . 69
4.1.1 Diskreetti tasajakauma . . . 69

4.1.2 Bernoulli-jakauma . . . 69
4.1.3 Binomijakauma . . . 70
4.1.4 Hypergeometrinen jakauma . . . 71
4.1.5 Geometrinen jakauma . . . 71
4.1.6 Poisson-jakauma . . . 73
4.2 Diskreetin satunnaismuuttujan odotusarvo . . . 74
4.3 Satunnaismuuttujien riippumattomuus . . . 77
4.4 Harjoitustehtäviä . . . 78

5 Jatkuvat satunnaismuuttujat **81**
5.1 Esimerkkejä jatkuvista jakaumista . . . 85
5.1.1 Jatkuva tasajakauma . . . 86
5.1.2 Eksponenttijakauma . . . 86
5.1.3 Normaalijakauma . . . 88
5.2 Jatkuvan satunnaismuuttujan odotusarvo . . . 90
5.3 Satunnaismuuttujan muunnos . . . 91
5.4 Harjoitustehtäviä . . . 92

6 Odotusarvon ominaisuuksia **93**
6.1 Odotusarvo . . . 94
6.2 Varianssi ja keskihajonta . . . 98
6.3 Kovarianssi ja korrelaatio . . . 102
6.4 Harjoitustehtäviä . . . 105

7 Raja-arvolauseet **107**
7.1 Epäyhtälöitä . . . 107
7.2 Raja-arvolauseita . . . 109
7.2.1 Suurten lukujen laki . . . 110
7.2.2 Keskeinen raja-arvolause . . . 112
7.2.3 Normaaliapproksimaatio . . . 113
7.3 Harjoitustehtäviä . . . 117

8 Lisämateriaalia **118**
8.1 Satunnaismuuttujan muunnos . . . 118
8.2 Yhteisjakaumat . . . 121
8.3 Todennäköisyyslaskuja R-ohjelmistolla . . . 126
8.4 Harjoitustehtäviä . . . 131

A Jakaumia **134**

Luku 0

Intro

0.1 Esipuhe

Tervetuloa todennäköisyyslaskennan pariin. Ajatus uudesta materiaalista todennäköisyyslaskennan peruskurssille on ollut ilmoilla jo usean vuoden. Paljon käytetyssä Tuomisen kirjassa [1] on havaittu ajan tuomia haasteita: merkinnät eivät vastanneet luennoitsijoiden käyttämiä ja ennen kaikkea teoksen laajuus on jo pitkään ollut eri tasolla kurssin sisällön kanssa. Hyviä todennäköisyyslaskennan materiaaleja on englanniksi saatavilla paljon, ja suomenkielisiä luentomateriaaleja on myös ollut saatavilla. Näin ollen taustatyötä oli huomattavasti tehty ja tilaisuus uudelle ja laajemmalle – ”oppikirjamaiselle” – luentomateriaalille oli otollinen. Materiaalin tekijät ovat pohtineet aikaisempien suomen kielisten materiaalien ongelmakohtia ja yhdistelleet eri lähestymistapoja eri materiaaleista yksiin kansiin. Materiaalia käytetään tulikokeessa Helsingin yliopiston avoimen yliopiston todennäköisyyslaskennan kesäkurssilla vuonan 2018. Tätä kirjoittaessa (heinäkuu 2018) materiaali on käyttövalmis, mutta materiaalia tullaan laajentamaan ja päivittämään syksyn 2018 aikana.

Materiaalin raakarunko on syntynyt Helsingin yliopiston avoimen yliopiston Todennäköisyyslaskenta 1-kurssien kesä 2016 (Ville Hyvönen) ja kesä 2017 (Ville Hyvönen, Topias Tolonen) luentomateriaaleista[1], jotka puolestaan pohjautuivat osin Petri Koistisen loistavan 2013 luentomateriaalin alkuosaan kurssille Todennäköisyyslaskenta (nyk. Todennäköisyyslaskenta 2). Raakarungosta eteenpäin materiaalia on suurimmaksi osaksi työstänyt Patrik Lauha Tolosen ehdotuksien pohjalta kesän 2018 aikana, poimien parhaita aiheen lähestymistapoja sekä Sheldon Rossin huikeasta kirjasta *A First Course In Probability*[2] että pitkään kurssimateriaalina käytetystä Pekka Tuomisen klassikkoteoksesta *To-*

[1] Vuonna 2017 avoimen kesäkurssin luennoinut Tolonen muokkasi alkuperäistä Hyvösen materiaalia runsaanpuoleisesti vastaamaan luentojen kulkua ja luennoitsijan huumorintajua.

[2] Pearson, 8. painos, 2010

dennäköisyyslaskenta 1[3]. Materiaalin tekoon on osallistunut myös pitkäaikainen todennäköisyyslaskennan kurssin luennoitsija Mika Koskenoja, jolta on varmistettu pedagogisia linjanvetoja ja jolta on saatu satoja esimerkkitehtäviä ja tehtäväideoita ratkaisuineen. Erityiskiitos vielä Sirkka-Liisa Varviolle projektin käynnistämisestä.

Materiaalissa on selitetty melko kattavasti esitietoja joukko-opista, ja materiaali pyrkii käymään järkevässä järjestyksessä todennäköisyyslaskennan peruskäsitteistöä: aloitamme kombinatoriikan perusteista ja etenemme formaalimman märittelyn kautta satunnaismuuttujien teoriaan. Olemme pitäneet parhaamme mukaan merkinnät selkeinä ja esitysmuodot ymmärrettävinä myös niille lukijoille, jotka eivät ole harrastaneet matematiikkaa pitkäaikaisesti. Tämän vuoksi materiaali soveltuu loistavasti myös niille, joille todennäköisyyslaskenta on ensimmäinen yliopistotason matematiikan kurssi. Ainoina merkittävinä esitietoina numeeristen vastausten saamiseen vaaditaan lukiotason integrointitaitoa, mutta keskeiset käsitteet aukeavat myös ilman tyhjentävää tietämystä integraalilaskennan saloista.

Toivottavasti tämä materiaali antaa lukijalle innostusta ja oppia todennäköisyyslaskennan peruskurssin suorittamiseen, sekä ehkä erityisesti edelleen motivaatiota kiinnostua yleisemmin satunnaisilmiöistä ja jatkamaan niiden opiskelua tämän kurssin ulkopuolelle.

Höyrylaiva Wennon kannelta heinäkuussa 2018,
Topias Tolonen

0.2 Mitä on todennäköisyys?

Mitä mielestäsi tarkoitetaan, jos sanotaan että Helsingissä sataa huomenna todennäköisyydellä 0.5*?*

Todennäköisyyslaskenta on matematiikan ala, jota käytetään apuna epävarmuuden hahmottamisessa. Sitä hyödynnetään lukuisilla eri tieteenaloilla, kuten tilastotieteessä, luonnontieteissä, taloustieteessä ja koneoppimisessa. Vaikka todennäköisyyslaskentaa on kehitetty jo 1600-luvulta lähtien, sen soveltamisen mahdollisuudet ovat viime aikoina laajentuneet huomattavasti tietokoneiden laskentatehon kasvamisen myötä.

Todennäköisyyslaskennan juuret ovat vahvasti kietoutuneet uhkapelaamiseen. Ranskalainen kirjailija ja innokas uhkapeluri Chevalier de Méré havaitsi pelatessaan seuraavat asiat:

- On kannattavaa lyödä vetoa siitä, että kun heitetään noppaa 4 kertaa, saadaan vähintään yksi kuutonen.

[3] Limes, 10. painos, 2010

- Ei ole kannattavaa lyödä vetoa siitä, että kun heitetään kahta noppaa 24 kertaa, saadaan vähintään yksi kuutospari.

De Méré ei osannut kuitenkaan perustella havaintoaan, joten hän pyysi apua ystävältään matemaatikko Blaise Pascalilta, joka ratkaisi ongelman yhdessä Pierre de Fermat'n kanssa luoden samalla perustan todennäköisyyslaskennalle. Myöhemmin todennäköisyyslaskentaa kehittivät mm. sveitsiläinen Jakob Bernoulli sekä ranskalaissyntyiset Abraham de Moivre ja Pierre Simon de Laplace. Todennäköisyys ymmärrettiin pitkään ainoastaan *kombinatoriikan* (luku 1.2) ja *klassisen todennäköisyyden* (2.1) kautta, kunnes 1900-luvun alussa neuvostoliittolainen matemaatikko Andrei Kolmogorov esitti *todennäköisyyslaskennan aksioomat* (2.15), jotka toimivat täsmällisenä matemaattisena määritelmänä todennäköisyyden käsitteelle.

Nykypäivänä todennäköisyyslaskennan sovellukset ulottuvat myös uhkapelien ulkopuolelle. Todennäköisyyslaskennan taitamista käytetään teollisuudessa ja sovelluksissa laajasti epävarmuuden mallintamisessa ja siten sen hallitsemisessa. Kun saamme lukuarvot satunnaisilmiöiden käyttäytymiselle ja siten lukuarvoja epävarmuudelle, mahdollistamme epävarmuuteen liittyvän rationaalisen keskustelun ja erilaisten vaihtoehtojen vertailun. Epävarmuuden hallitseminen ja sen tuoma riskienhallinta on elintärkeä taito nykyajassa, missä pienellä vaivalla voimme tietokoneiden avulla mallintaa jopa yllättävän komplekseja ilmiöitä. Esimerkiksi nykyaikana tekoäly ja koneoppiminen nojaavat vahvasti todennäköisyyslaskentaan. Lisäksi rahoitusala on pelottavan riippuvainen todennäköisyyslaskennan tuomista mahdollisuuksista niin riskienhallinnassa kuin arvopaperien hintojen ennustamisessa. Kuka mallintaa epävarmuutta, mallintaa huomista.

Mitä todennäköisyys sitten oikeastaan tarkoittaa? Jotkin ilmiöt ja tapahtumat ovat deterministisiä, eli niiden lopputuloksen voi suoraan päätellä riittävillä taustatiedoilla. Esimerkki deterministisestä tapahtumasta voisi olla vaikkapa lumipallon sulaminen. Kun lumipallo tuodaan sisätiloihin, se sulaa varmasti, koska lumi muuttuu yli 0 asteen lämpötilassa vedeksi. Tapahtumaan ei liity minkäänlaista satunnaisuutta, vaan kun tiedetään veden sulamispiste ja sisäilman lämpötila, voidaan varmuudella päätellä lumipallon sulavan.

Kaikki tapahtumat eivät kuitenkaan ole deterministisiä, vaan niihin liittyy satunnaisuutta. Perinteinen esimerkki tällaisesta tapahtumasta on painottamattoman nopan heitto. Saatua silmälukua ei voi ennen nopan heittämistä päätellä, voidaan ainoastaan todeta, mitä eri vaihtoehtoja on ja kuinka todennäköisiä ne ovat.

Ennen kuin siirrymme käsittelemään todennäköisyyden matemaattista määritelmää (luku 2), pohditaan hetki erilaisia 'maalaisjärkisiä' tulkintoja todennäköisyydelle. Yksinkertaisesti todennäköisyys tarkoittaa jonkin väitteen totuusarvon uskottavuuden astetta. Kun väitetään sateen todennäköisyyden olevan 0.5, sillä voidaan tarkoittaa olevan yhtä uskottavaa, että huomenna sataa ja että huomenna ei sataisi. Edellä mainittu nopanheittoesimerkki edustaa klassista todennäköisyyden määritelmää. Klassinen todennäköisyys määritellään yksinkertaisesti suotuisten tulosvaihtoehtojen osuutena kaikista mahdollista tuloksista. Esimerkiksi nopanheitossa mahdollisuus saada parillinen silmäluku on $\frac{3}{6}$. Mahdollisia silmälukuja on 6, joista 3 (2,4 ja 6) ovat parillisia.

Useissa tilanteissa kaikki mahdolliset tapahtumat eivät kuitenkaan ole yhtä todennäköisiä, eikä todennäköisyyttä voida tällöin tulkita suotuisten tapausten osuuden kautta. Esimerkkinä tällaisesta tapahtumasta voidaan pitää presidentinvaaleja, joissa vastakkain ovat henkilöt A ja B. Ennen vaaleja saatetaan esittää, että henkilö A voittaa vaalit 70% todennäköisyydellä. Tällaisessa tilanteessa todennäköisyyttä käytetään mittana tapahtuman mahdollisuudelle ja se on usein ainakin jossain määrin subjektiivinen tulkinta eri vaihtoehtojen mahdollisesta toteutumisesta. Ainutkertaisissa tapahtumissa, kuten vaaleissa edellä mainittujen todennäköisyyksien oikeellisuutta on vaikeaa jälkikäteen arvioida. Tieto siitä, että henkilö B vastoin odotuksia voittaakin vaalit, ei sinänsä kerro mitään siitä, kuinka täsmällinen ennuste oli, ja mitkä voiton mahdollisuudet 'todellisuudessa' olivat.

Sen sijaan tilanteissa, jossa samankaltaista ennustetta tuotetaan säännöllisesti tapahtuvasta tapahtumasta, voidaan empiirisesti tarkastella ennusteiden täsmällisyyttä ja tapahtumien todellisuudessa ilmenevää todennäköisyyttä. Esimerkiksi sääennusteiden toteutumisen tarkkuutta voidaan pitkällä aikavälillä arvioida vertaamalla toteutuneita säätiloja ennusteisiin. Jos valitaan usean vuoden ajalta ennusteet, joiden mukaan seuraavana päivänä sataa 90% todennäköisyydellä ja tarkastellaan näitä ennusteita seuranneiden päivien säätä, tulisi noin yhdeksänä päivänä kymmenestä olla sateista, jotta ennusteiden arvioita sateen todennäköisyydestä voitaisiin pitää onnistuneina. Ajattelutapa, jossa jonkin asian todennäköisyyttä tutkitaan tällä tavoin frekventistisesti tapahtuman toistumisen yleisyyden kautta, liittyy olennaisesti tilastolliseen päättelyyn. Todennäköisyyden tulkintaa toistuvien tapahtumien yhteydessä käsitellään tällä kurssilla muun muassa toistokokeiden (3.4) ja suurten lukujen lain (7.2.1) kautta.

Termin 'todennäköisyys' kanssa on syytä olla tarkkana, sillä todennäköisyyksien tulkinnassa sattuu usein virheitä. Hyvin yksinkertaistettuna esimerkkinä: lottovoiton todennäköisyys ei ole $\frac{1}{2}$, vaikka mahdollisia tapahtumia voisi ajatella olevan kaksi: voittaa tai ei voita[4]. Tai toisaalta: väite, jonka mukaan 'Etelä-Suomessa sataa huomenna 90% todennäköisyydellä' ei esimerkiksi tarkoita sitä, että Etelä-Suomessa sataisi seuraavana päivänä noin 90% ajasta, tai että noin 90% alueella Etelä-Suomesta saataisiin seuraavana päivänä sadetta, vaan että on hyvin todennäköistä, että jossain päin Etelä-Suomea sataa jossain vaiheessa seuraavan päivän aikana.

Tämän kurssin tarkoituksena on auttaa opiskelijaa ymmärtämään todennäköisyys-

[4]Jutun juoni on siinä, että *voiton mahdollistavia tapahtumia* eli voittavia rivejä on vain yksi. Vastaavasti kaikki muut rivit ovat ei-voittavia rivejä, joten selvästi tapahtumia on suomalaisessa Lotossa enemmän kuin kaksi

laskennan periaatteita ja laskemaan todennäköisyyksiä käytännössä. Kurssilla kerrataan kombinatoriikkaa, tutustutaan todennäköisyyslaskennan perusteisiin ja satunnaismuuttujiin, esitellään tavallisimpia todennäköisyysjakaumia, opetellaan jakaumien tulkintaa niiden tunnuslukujen avulla sekä esitellään muutamia todennäköisyyslaskennan teorian keskeisimpiä lauseita.

Luku 1

Joukko-oppia ja kombinatoriikkaa

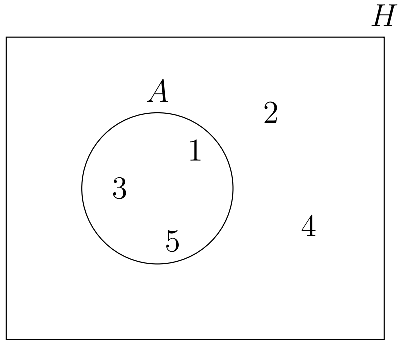
Kurssilla käsitellään matemaattisia käsitteitä, joita ovat esimerkiksi joukot ja niiden alkiot. Tässä kappaleessa lukijalle kerrataan matemaattisia termejä, joukko-oppia ja kombinatoriikkaa, joita tarvitaan todennäköisyyslaskennassa.

1.1 Joukko-opin termien kertaus

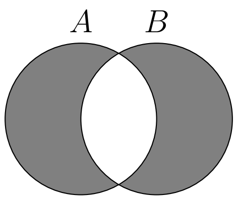
Todennäköisyyslaskenta perustuu joukko-oppiin, joten muistellaan aluksi muutamia joukko-opin merkintöjä ja sääntöjä. *Joukko* on järjestämätön kokoelma mitä tahansa asioita tai 'olioita'. Joukon jäseniä kutsutaan *alkioiksi*. Olkoot A ja B joukkoja ja x jokin alkio. Tällöin käytetään merkintöjä, joilla on seuraavat tulkinnat:

Merkintä	Tulkinta
$x \in A$	x *kuuluu* joukkoon A
$x \notin A$	x *ei kuulu* joukkoon A
A^C	joukon A *komplementti*, eli ne alkiot, jotka eivät kuulu joukkoon A
$A \subset B$	A on joukon B *osajoukko*, eli kaikki A:n alkiot ovat myös B:n alkioita
$A \subseteq B$	A on joukon B *osajoukko*, mutta joukot voivat olla myös sama joukko eli $A = B$.
$A = B$	A ja B ovat *samoja*, eli niillä on samat alkiot
$A \cup B$	joukkojen A ja B *yhdiste* $\{x \colon x \in A$ tai $x \in B\}$
$A \cap B$	joukkojen A ja B *leikkaus* $\{x \colon x \in A$ ja $x \in B\}$
$A \setminus B$	joukkojen A ja B *erotus* $\{x \colon x \in A$ ja $x \notin B\}$
$\emptyset$	*tyhjä joukko*

Taulukossa esiteltyjä joukko-operaatioita voi havainnollistaa esimerkiksi Vennin diagrammien avulla:



Kuva 1.1: Luonnolliset luvut 1, 2, 3, 4 ja 5 muodostavat joukon H, jonka osajoukon A muodostavat parittomat luonnolliset luvut 1, 3 ja 5. Tällöin voidaan merkitä esimerkiksi $1 \in A$ tai $5 \in H$.



Kuva 1.2: Joukoista A ja B on varjostettu joukko $(A \cup B) \setminus (A \cap B)$.

Joukkojen yhdisteillä ja leikkauksilla on muutamia tärkeitä ominaisuuksia, jotka liittyvät joukkojen 'laskujärjestykseen'. Seuraavat säännöt on hyvä pitää mielessä joukkoja käsiteltäessä:

- Vaihdantalait. Kaikilla A ja B pätee

$$A \cup B = B \cup A \qquad \text{ja} \qquad A \cap B = B \cap A.$$

- Liitäntälait. Kaikilla A, B ja C pätee

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

- Osittelulait. Kaikilla A, B ja C pätee

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

- De Morganin lait. Kaikilla A ja B pätee

$$(A \cup B)^C = A^C \cap B^C \qquad \text{ja} \qquad (A \cap B)^C = A^C \cup B^C.$$

  De Morganin lait yleistyvät myös useampien joukkojen yhdisteille ja leikkauksille.

Yhdisteen ja leikkauksen voi laskea myös äärettömän monelle joukolle. Jos $A_1, A_2, \ldots$ ovat joukkoja, niin niiden yhdistettä merkitään

$$\bigcup_{j=1}^{\infty} A_j = \{x : x \in A_j \text{ jollakin } j\}$$

ja leikkausta merkitään

$$\bigcap_{j=1}^{\infty} A_j = \{x : x \in A_j \text{ kaikilla } j\}.$$

Määritellään vielä lopuksi potenssijoukon käsite, jota tarvitsemme myöhemmin klassisen todennäköisyyden perusjoukon määrittelemiseksi.

Määritelmä 1.1. *Potenssijoukko.* Olkoon A mielivaltainen joukko. Joukon A *potenssijoukko* on tällöin joukon A kaikkien osajoukkojen joukko:

$$\mathcal{P}(A) = \{B : B \subseteq A\}.$$

Määritelmä 1.1 siis kertoo, että joukon A potenssijoukko muodostuu kaikista joukon A osajoukoista. Määritelmän mukaan joukon A osajoukkoja ovat myös tyhjä joukko ja joukko A itse. Näin ollen esimerkiksi joukon $A_1 = \{1, 2, 3\}$ potenssijoukko on

$$\mathcal{P}(A_1) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

1.2 Kombinatoriikkaa

Ennen todennäköisyyden matemaattisen käsitteen esittelyä, tarkastellaan vielä hetki kombinatoriikkaa[1]. Lähdetään liikkeelle yksinkertaisesta esimerkistä:

[1]Kombinatoriikassa tutkitaan tietyt ominaisuudet toteuttavien joukkojen lukumääriä ja näihin liittyviä ongelmia. Osa kombinatoriikan peruskäsitteistä on käyty läpi lukion todennäköisyyttä käsittelevillä kursseilla.

Esimerkki 1.2. Nopanheitto. Heitetään tavallista noppaa. Nopanheitossa todennäköisyys lasketaan jakamalla haluttujen silmälukujen lukumäärä kaikkien silmälukujen lukumäärällä eli kuudella. Mikä on todennäköisyys, että saatu silmäluku on pienempi kuin 5?

Lasketaan suotuisten silmälukujen määrä. Mahdollisia silmälukuja ovat 1, 2, 3, 4, 5 ja 6, joista pienempiä kuin 5 ovat 1, 2, 3 ja 4 (4 kpl). Vastaus on siis $\frac{4}{6}$.

Kun mahdollisia lopputuloksia on rajallinen määrä ja ne ovat kaikki yhtä todennäköisiä, on todennäköisyyksien laskeminen periaatteessa aina yhtä helppoa kuin ylläolevassa esimerkissä. Kuitenkin, kun mahdollisia lopputuloksia eli *alkeistapauksia* [2] on paljon, niitä ei voida enää erikseen luetella, vaan tarvitaan laskusääntöjä *perusjoukon*[2] ja sen osajoukkojen sisältämien alkeistapausten lukumäärän laskemiseksi. Näitä laskusääntöjä kutsutaan kombinatoriikaksi.

Lause 1.3. *Laskemisen perusperiaate.*[3] *Suoritetaan kaksi koetta, joista ensimmäisellä on m ja toisella n mahdollista lopputulosta $(m, n \in \mathbb{N})$. Tällöin kahdella kokeella on yhdessä yhteensä mn mahdollista lopputulosta.*

Todistus. Numeroidaan kokeiden mahdolliset lopputulokset $1\ \ldots, m$ ja $1 \ldots n$. Todistus voidaan nyt esittää luettelemalla kahden kokeen kaikki mahdolliset lopputulokset:

$$\begin{array}{cccc}
(1,1) & (1,2) & \ldots & (1,n) \\
(2,1) & (2,2) & \ldots & (2,n) \\
\vdots & \vdots & \ldots & \vdots \\
(m,1) & (m,2) & \ldots & (m,n).
\end{array}$$

Kullakin vaakarivillä on n tulosta ja kussakin pystysarakkeessa m lopputulosta, jolloin mahdollisia lopputuloksia on selvästi mn. □

Laskemisen perusperiaate laajentuu yleisempään muotoon *tuloperiaatteena*.

1.2.1 Tuloperiaate

Monissa tilanteissa haluamme tietää, kuinka monella tavalla jokin asia tai tehtävä voidaan valita tai suorittaa. Esimerkiksi haluamme valita astiayhdistelmän, johon on mahdollista valita 2 lautasvaihtoehtoa, 3 lasivaihtoehtoa ja yksi aterinvaihtoehto ja haluamme tietää, kuinka monta erilaista astiayhdistelmää voimme valita. Ensimmäiseen lautasvaihtoehdon

[2] Alkeistapauksen ja perusjoukon käsitteisiin tutustutaan tarkemmin seuraavassa luvussa.
[3]engl. *basic principle of counting*

kanssa voimme valita yhteensä 3 lasivaihtoehtoa ja lisäksi toisen lautasen kanssa on valittavissa samat 3 lasia, eli yhteensä lasi-lautas-yhdistelmiä on 6 kappaletta. Koska aterinvaihtoehtoja on vain yksi, niin voimme päätellä astiayhdistelmiä olevan 6 kappaletta. Samaan tulokseen päädytään myös *tuloperiaatteen* avulla: $6 = 2 \cdot 3 \cdot 1$.

Lause 1.4. *Tuloperiaate.*[4] *Oletetaan, että tehtävä voidaan suorittaa k:ssa peräkkäisessä vaiheessa siten, että ensimmäisessä vaiheessa on n_1 vaihtoehtoa, toisessa vaiheessa on n_2 vaihtoehtoa, ja niin edelleen viimeiseen vaiheeseen, jossa on n_k vaihtoehtoa, asti. Oletetaan myös, että i:nen vaiheen vaihtoehtojen määrä n_i ei riipu (huom. itse vaihtoehdot voivat kyllä riippua) edeltävissä vaiheissa tehdyistä valinnoista. Tällöin tehtävä voidaan suorittaa yhteensä*

$$\prod_{i=1}^{k} n_i = n_1 \cdot n_2 \cdot \ldots \cdot n_k$$

eri tavalla.

Todistus. Todistus tapahtuu lausetta 1.3 hyödyntäen. Jätetään harjoitustehtäväksi. □

Havainnollistetaan asiaa vielä todennäköisyyslaskennan historian hengen mukaisesti uhkapeliesimerkillä:

Esimerkki 1.5. Vakioveikkausrivi. Jokaisessa kohteessa valitaan merkki kolmesta eri vaihtoehdosta (1,X,2), ja kohteita on yhteensä 13. Siten erilaisia mahdollisia rivejä on

$$3^{13} = 1594323$$

kappaletta.

Tuloperiaatteella saadaan myös suoraan laskettua joukon erilaisten osajoukkojen määrä:

Lause 1.6. *Osajoukkojen lukumäärä. $n-$alkioisella joukolla $E = \{x_1, \ldots, x_n\}$ on 2^n erilaista osajoukkoa.*

Todistus. Joukon E mielivaltainen osajoukko voidaan muodostaa n:ssä vaiheessa siten, että ensimmäiseksi valitaan kuuluko x_1 osajoukkoon, seuraavaksi valitaan kuuluko x_2 osajoukkoon, ja niin edelleen aina viimeiseen alkioon x_n asti. Jokaisessa vaiheessa on 2 vaihtoehtoa: joko alkio kuuluu osajoukkoon, tai se ei kuulu siihen. Siten tuloperiaatteen nojalla yhteensä erilaisia vaihtoehtoja on 2^n kappaletta. □

Huomautus 1.7. Huomaa, että lause 1.6 kertoo siis joukon E potenssijoukon $\mathcal{P}(E)$ (ks. määritelmä 1.1) alkioiden lukumäärän. Tästä tiedosta on hyötyä myöhemmin.

[4]engl. *generalized basic principle of counting*

1.2.2 Summaperiaate

Tuloperiaatteen lisäksi toinen hyödyllinen laskemismenetelmä on summaperiaate, jota voidaan käyttää tilanteissa, joissa asia tai tehtävä voidaan valita tai suorittaa useilla **toisensa poissulkevilla** menetelmillä.

Lause 1.8. *Summaperiaate.*[5] *Oletetaan, että koe voidaan suorittaa k toisensa poissulkevalla (eli vaihtoehtoisella) tavalla. Jos tavan $i \in \{1, \dots, k\}$ mahdollisuuksien määrää merkitään n_i:llä, kokeella on yhteensä*

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \dots n_k$$

mahdollista lopputulosta.

Esimerkki 1.9. Opiskelija matkustaa kotoaan kampukselle. Hän voi käyttää joko linjojen 10 ja 11 suoria bussiyhteyksiä, matkustaa ensin bussilla 20 ja vaihtaa sen jälkeen linjaan 21, 22 tai 23, tai matkustaa ensin metrolla ja vaihtaa sen jälkeen bussilinjaan 30 tai 31. Opiskelija valitsee aina sen liikennevälineen, joka tulee paikalle ensimmäisenä, joten opiskelijaa käyttää satunnaisesti kaikkia eri tapoja. Kuinka monella eri tavalla opiskelija voi päästä kampukselleen?

Ratkaisu voidaan esittää summaperiaatteen nojalla toistensa poissulkevien vaihtoehtojen summana. Vaihtoehdot ovat suora bussiyhteys, kahdella bussilla matkustaminen tai metrolla ja bussilla matkustaminen. Tällöin $n_1 = 2$, $n_2 = 3$ ja $n_3 = 2$, jolloin vastaus on $2 + 3 + 2 = 7$.

1.2.3 Permutaatiot ja kombinaatiot

Joukon alkioiden järjestyksillä eli *permutaatioilla* ja joukon osajoukoilla eli *kombinaatioilla* on keskeinen rooli todennäköisyyslaskennassa. Havainnollistetaan seuraavaksi permutaation ja kombinaation käsitteitä esimerkkien kautta.

Esimerkki 1.10. Tarkastellaan 4-alkioista joukkoa $A = \{a, b, c, d\}$.

- Kuinka monella tavalla joukko voidaan järjestää? Järjestäminen voidaa jakaa neljään vaiheeseen. Jonon ensimmäiseksi alkioksi voidaan valita mikä tahansa joukon alkio, eli erilaisia vaihtoehtoja on 4. Jonon toiseksi alkioksi voidaan valita mikä tahansa paitsi ensimmäiseksi valittu alkio, eli erilaisia vaihtoehtoja on 3. Kolmanneksi

[5] engl. *rule of sum*

alkioksi valitaan toinen jäljelläolevista alkioista, ja viimeiseksi alkioksi vaihtoehtoja on jäljellä enää yksi. Siten tuloperiaatteen nojalla joukko A voidaan järjestää

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

erilaisella tavalla.

- Kuinka monta erilaista kahden alkion järjestettyä jonoa (eli järjestettyä paria) joukosta A voidaan valita? Kaikki mahdollisuudet voidaan luetella ja laskea:

$$\begin{aligned}&(a,b),(a,c),(a,d)\\&(b,a),(b,c),(b,d)\\&(c,a),(c,b),(c,d)\\&(d,a),(d,b),(d,c),\end{aligned}$$

eli yhteensä 12 kappaletta. Tämä voidaan laskea myös tuloperiaatteella, sillä kahden alkion järjestetyn jonon valinta joukosta voidaan suorittaa kahdessa vaiheessa: ensin valitaan jonon ensimmäinen alkio kaikista joukon neljästä alkiosta, ja sen jälkeen toinen alkio jäljellä olevista kolmesta alkiosta. Siten erilaisia jonoja on yhteensä

$$4 \cdot 3 = 12$$

kappaletta.

- Kuinka monta erilaista kahden alkion (järjestämätöntä) osajoukkoa joukosta A voidaan valita? Kaikki vaihtoehdot voidaan jälleen luetella ja laskea:

$$\begin{aligned}&\{a,b\},\{a,c\},\{a,d\}\\&\{b,c\},\{b,d\},\\&\{c,d\},\end{aligned}$$

eli yhteensä 6 kahden alkion osajoukkoa. Tämä voidaan myös laskea havaitsemalla, että jokainen kahden alkion joukko voidaan järjestää kahdella eri tavalla; esimerkiksi joukon $\{a,b\}$ mahdolliset järjestykset ovat (a,b) ja (b,a). Siten jokaista kahden alkion osajoukkoa kohti on 2 kahden pituista järjestettyä jonoa. Edellä todettiin, että joukon A kahden pituisia järjestettyjä jonoja on 12 kappaletta, joten kahden alkion osajoukkoja on

$$\frac{12}{2} = 6$$

kappaletta.

Nimetään seuraavaksi nämä käsitteet, ja todetaan edellisen esimerkin havainnot yleisessä n-alkioisen joukon tapauksessa.

Määritelmä 1.11. *Permutaatio ja kombinaatio.* Olkoon $E = \{x_1, \dots, x_n\}$ n-alkioinen joukko.

(i) Joukon E alkioiden järjestystä, eli sen alkioista muodostettua (jokainen alkio esiintyy vain kerran) järjestettyä n-alkioista jonoa

$$(x_{i_1}, x_{i_2} \dots, x_{i_n})^{6}$$

kutsutaan sen *permutaatioksi.*

(ii) Kaikille $k \in \{1, \dots, n\}$ joukon E alkioista muodostettua (jokainen alkio esiintyy korkeintaan kerran) järjestettyä k-alkioista jonoa

$$(x_{i_1}, x_{i_2}, \dots, x_{i_k})$$

kutsutaan sen *k-permutaatioksi.* [7]

(iii) Joukon E alkioista muodostettua k-alkioista osajoukkoa (järjestyksellä ei siis ole väliä)

$$\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$$

kutsutaan sen *k-kombinaatioksi.*

Lause 1.12. *Permutaatioiden ja kombinaatioiden määrä. Olkoon $E = \{x_1, \dots, x_n\}$ n-alkioinen joukko.*

(i) Joukolla E on $n!$ permutaatiota.

(ii) Kaikille $k \in \{1, \dots, n\}$ joukolla E on

$$\frac{n!}{(n-k)!}$$

k-permutaatiota.

(iii) Kaikille $k \in \{1, \dots, n\}$ joukolla E on

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

k-kombinaatiota.

---

[6] Muistutus lukijalle: *jonoissa* alkioiden järjestyksellä **on** merkitystä, *joukoissa* alkioilla **ei ole** järjestystä.

[7] joissain lähteissä myös *k-variaatio*

Todistus.

(i) Mikä tahansa joukon E permutaatio voidaan valita n:ssä vaiheessa siten, että ensimmäiseksi valitaan jonon ensimmäinen alkio, seuraavaksi toinen alkio, ja niin edelleen aina viimeiseen alkioon saakka. Ensimmäiseksi alkioksi voidaan valita mikä tahansa joukon E alkio, joten ensimmäisessä vaiheessa on n vaihtoehtoa. Toiseksi alkioksi voidaan valita mikä tahansa paitsi ensimmäiseksi valittu alkio, joten toisessa vaiheessa on $n-1$ vaihtoehtoa. Näin jatketaan aina viimeiseen vaiheeseen, jossa jäljellä on vain yksi vaihtoehto, saakka. Siten tuloperiaatteen nojalla permutaatio voidaan valita

$$n \cdot (n-1) \cdot ... \cdot 1 = n!$$

erilaisella tavalla.

(ii) Vastaavasti kuin (i)-kohta, mutta vaiheita on k kappaletta, jolloin viimeisessä vaiheessa jäljellä on $(n-k+1)$ vaihtoehtoa jonon viimeiseksi alkioksi. Siten tuloperiaatteen nojalla k-permutaatio voidaan valita

$$n \cdot (n-1) \cdots (n-k+1) = \frac{(n-1) \cdots (n-k+1) \cdot (n-k) \cdot (n-k-1) \cdots 1}{(n-k) \cdot (n-k-1) \cdots 1} = \frac{n!}{(n-k)!}$$

erilaisella tavalla.

(iii) Olkoon x joukon E k-kombinaatioiden määrä. Jokainen k-kombinaatio on k-alkioinen joukko, joten kohdan (i) nojalla se voidaan järjestää $k!$:lla eri tavalla. Kaikki joukon k-permutaatiot saadaan järjestämällä sen k-kombinaatiot, joten joukolla E on $x \cdot k!$ k-permutaatiota. Siten kohdan (ii) nojalla saadaan yhtälö, josta voidaan ratkaista k-kombinaatioiden määrä:

$$\begin{aligned} x \cdot k! &= \frac{n!}{(n-k)!} \\ x &= \frac{n!}{k!(n-k)!} = \binom{n}{k}. \end{aligned}$$

$\square$

Kokonaisluvuille $n, k \in \mathbb{N}$ lukuja

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

kutsutaan *binomikertoimiksi*.[8] Kun $k < 0$ tai $k > n$, määritellään

$$\binom{n}{k} = 0.$$

Tarkastellaan seuraavaksi muutamaa esimerkkiä kombinaatioiden ja permutaatioiden laskemisesta.

Esimerkki 1.13. Pokerikäsi. Sekoitetusta korttipakasta nostetaan viisi korttia. Järjestyksellä, jossa kortit on nostettu pakasta, ei ole väliä pelin kannalta. Montako erilaista 5 kortin kättä on mahdollista nostaa?

Mahdollisia viiden kortin yhdistelmiä ovat esimerkiksi

$$\{\heartsuit 3, \diamondsuit 3, \heartsuit J, \spadesuit 5, \diamondsuit 9\}$$

ja

$$\{\diamondsuit 5, \spadesuit J, \diamondsuit K, \clubsuit 3, \heartsuit A\}.$$

Näitä korttipakan 5-kombinaatioita on lauseen 1.12 (iii)-kohdan nojalla yhteensä

$$\binom{52}{5} = \frac{52!}{5!47!} = 2598960$$

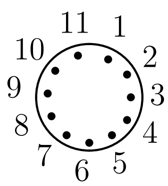
kappaletta.

Esimerkki 1.14. 11 maan YK-valtuuskuntia edustavat henkilöt istuvat turvallisuusneuvostossa pyöreän pöydän ääreen jossakin järjestyksessä. Kahta sijoittelua pidetään samana, jos ne saadaan toisistaan pöytää kiertämällä.

a) Kuinka monella tavalla henkilöt voivat asettua pöydän ääreen?

Numeroidaan valtuuskuntien edustajat numeroilla $1, \ldots, 11$ ja istumapaikat indekseillä $1, \ldots, 11$ (katso kuva).

Pöytä:



Edustajat:
1 2 3 4
5 6 7 8
9 10 11

Nyt lauseen 1.12 kohdan (i) nojalla 11 henkilöä voidaan asettaa 11 paikalle 11! eri tavalla. Näihin sijoitteluihin sisältyvät kuitenkin pöytää kiertämällä saatavat järjestykset, joten vastaus on vielä jaettava paikkojen (eli mahdollisten kiertojen) lukumäärällä, jolloin saadaan lopulliseksi vastaukseksi $\frac{11!}{11} = 10! = 3628000$.

[8] engl. *binomial coefficient*

b) Kuinka monella tavalla henkilöt voivat asettua pöydän ääreen siten, että Englannin ja Ranskan edustajat istuvat vierekkäin, mutta Yhdysvaltojen ja Venäjän eivät?

Tarkastellaan nyt suotuisia järjestyksiä ja tilanteita, joilla niihin päädytään. Ensin Englannin edustajalla on 11 mahdollista paikkaa, jolloin Ranskan edustaja voi istua kummalla puolella tahansa Englannin edustajaa (2 mahdollisuutta). Mikäli Yhdysvaltojen edustaja istuu Ranskan tai Englannin edustajan vieressä (2 mahdollisuutta), Venäjän edustajalle jää 7 mahdollista paikkaa, koska hän ei voi istua Yhdysvaltojen edustajan viereen. Mikäli taas Yhdysvaltojen edustaja ei istu Englannin tai Ranskan edustajan vieressä (7 mahdollisuutta), Venäjän edustajalle jää 6 mahdollista paikkaa. Loput 7 edustajaa saadaan paikoilleen 7! eri tavalla. Näin ollen ehdot täyttävien istumajärjestysten määrä on tulo- ja summaperiaatteiden nojalla

$$\begin{aligned}
&11 \cdot 2 \cdot 2 \cdot 7 \cdot 7! + 11 \cdot 2 \cdot 7 \cdot 6 \cdot 7! \\
=&11 \cdot 7 \cdot 2 \cdot (2+6) \cdot 7! = 11 \cdot 7 \cdot 2 \cdot 8 \cdot 7! \\
=&6209280
\end{aligned}$$

Ylläolevissa järjestyksissä on kuitenkin laskettu erikseen pöydän erilaiset kierrot (sillä samanlaisiin järjestyksiin voidaan mistä tahansa järjestyksestä päätyä asettamalla Englannin edustaja aluksi eri paikalle ja täyttämällä loput paikat tämän jälkeen identtisesti), joten jaetaan lopputulos vielä paikkojen lukumäärällä:

$$6209280/11 = 564480$$

Vaihtoehtoisesti oltaisiin suoraan voitu jättää Englannin edustajan paikka kiinnittämättä, jolloin oltaisiin päädytty samaan lopputulokseen. Tässäkin tehtävässä on apua kuvan piirtämisestä ja eri tilanteiden hahmottelemisesta.

Muistetaan, että binomikerroin kertoo, kuinka monella tavalla n alkion kokoinen joukko voidaan jakaa kahteen osaan, joista toisessa on k ja toisessa $n-k$ alkiota (kuten esimerkissä 1.13 käteen nostettuihin ja nostamattomiin kortteihin). Binomikerroin voidaan yleistää käsittämään useampaan kuin kahteen osaan jaettava joukko, jolloin puhutaan *multinomikertoimesta*.[9]

Lause 1.15. *(Multinomikerroin): Olkoon $E = \{x_1, \dots, x_n\}$ n-alkioinen joukko, ja $n_1, \dots, n_k \in \mathbb{N}$ siten, että $\sum_{i=1}^{k} n_i = n$. Joukko E voidaan jakaa k:n osajoukkoon siten, että kaikille $i \in \{1, \dots, k\}$ i:nnessä osajoukossa on n_i alkiota*

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

erilaisella tavalla.

[9]engl. *multinomial coefficient*

Todistus. Osajoukkojen valinta jakautuu luontevasti k:n vaiheeseen:

- Ensimmäiseen osajoukkoon voidaan valita n:stä alkiosta mitkä tahansa n_1, joten lauseen 1.12 kohdan (iii) perusteella erilaisia vaihtoehtoja on $\binom{n}{n_1}$ kappaletta
- Toiseen osajoukkoon valitaan n_2 alkiota niistä alkiosta joita ei valittu ensimmäiseen osajoukkoon, eli vaihtoehtoja on $\binom{n-n_1}{n_2}$ kappaletta.
- Ja niin edelleen aina k:een osajoukkoon asti, jossa vaihtoehtoja on $\binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k}$ kappaletta.

Siten tuloperiaatteen nojalla erilaisia vaihtoehtoja on

$$\begin{aligned}
&\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k} \\
=&\frac{n!}{n_1!(n-n_1)!}\cdot\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdot\cdots\cdot\frac{(n-n_1-n_2-\cdots-n_{k-1})!}{n_k!(n-\sum_{i=1}^{k} n_i)!} \\
=&\frac{n!}{n_1!n_2!\ldots n_k!}=\binom{n}{n_1 n_2 \ldots\ n_k}
\end{aligned}$$

kappaletta. Tässä käytettiin hyväksi sitä, että peräkkäisten osoittajien ja nimittäjien termit kumoavat toisensa tulossa. Viimeisen termin nimittäjään jää termi $(n-\sum_{i=1}^{k} n_i)!$, mutta oletuksen nojalla $\sum_{i=1}^{k} n_i = n$, joten tämä termi on edelleen $0! = 1$. □

Havainnollistetaan multinomikertoimen käyttöä esimerkillä:

Esimerkki 1.16. Luennoitsijalla on kaksi siskoa ja yksi veli. Lisäksi hänellä on mansikka, banaani, omena ja appelsiini. Kuinka monella tavalla hän voi jakaa hedelmät sisaruksilleen siten, että veli saa kaksi hedelmää ja molemmat siskot yhden hedelmän?

Lauseen 1.15 mukaan hän voi antaa hedelmät yhteensä

$$\binom{4}{112} = \frac{4!}{1!1!2!} = \frac{24}{2} = 12 \tag{1.17}$$

eri tavalla. Jaoimme siis neljän hedelmän joukon kolmeen osajoukkoon siten, että ensimmäiseen osajoukkoon tulee kaksi hedelmää ja loppuihin osajoukkoihin yksi hedelmä.

Käy myös ilmi, että multinomikerroin on hyödyllinen työkalu silloin, kun haluamme katsoa että kuinka monta permutaatiota jostain jonosta voimme luoda, missä jokainen permutaatio sisältää eri alkiot. Seuraava esimerkki havainnollistaa tätä tilannetta kirjainjonoilla eli sanoilla, mistä haluamme järjestää permutaatiot siten, että jokainen permutaatio on oma sanansa.

Esimerkki 1.18. Kuinka monta sanaa (= kirjainjonoa) voidaan muodostaa sanasta KOKKOLA?

Huomataan, että sanassa KOKKOLA on 7 kirjainta, jotka voidaan järjestää 7! eri tavalla. Kuitenkin alkuperäinen sana sisältää samoja kirjaimia, ja esimerkiksi järjestykset $K_1O_1K_2K_3O_2LA$ ja $K_3O_2K_1K_2O_1LA$ ovat eri permutaatioita, mutta muodostavat saman sanan. Näin ollen vastaus pitää vielä jakaa kunkin kirjaimen mahdollisten järjestysten lukumäärillä. Näin ollen vastaus on

$$\frac{n!}{n_K!n_O!n_L!n_A!} = \frac{7!}{3!2!1!1!} = \frac{7!}{3!2!} = 420,$$

mikä on täsmälleen sama luku kuin multinomikerroin

$$\binom{7}{3211} = 420. \tag{1.19}$$

1.3 Harjoitustehtäviä

1. Luettele joukon $A = \{a, b, c, d\}$ potenssijoukko $\mathcal{P}(A)$.
2. Olkoot A ja B joukkoja. Lausu joukko-operaatioiden avulla seuraavat asiat:
   a) x kuuluu molempiin joukkoihin A ja B
   b) x ei kuulu kumpaankaan joukkoon
   c) x kuuluu ainakin toiseen joukkoon
   d) x kuuluu täsmälleen toiseen joukkoon
3. Kansainvälisessä aakkostossa on 26 kirjainta. Rekisteritunnus muodostetaan valitsemalla ensin kolme kirjainta ja sitten kolme numeroa.
   a) Montako erilaista rekisteritunnusta on mahdollista muodostaa?
   b) Oletetaan, ettei sama kirjain tai numero saa esiintyä rekisteritunnuksessa useammin kuin kerran. Montako rekisterikilpeä nyt on mahdollista muodostaa?
4. Osoita ja selitä havainnollistaen, miksi $\binom{n}{k} = \binom{n}{n-k}$.
5. 8 erilaisen lukon ja näiden 8 avaimen joukosta halutaan valita 6 esinettä siten, ettei valitussa joukossa ole yhtään avainta joka sopisi valitun joukon lukkoihin.
   a) Montako erilaista 6 esineen joukkoa on mahdollista valita?
   b) Entä, jos halutaan, että joukossa on 3 lukkoa ja 3 avainta?

6. Opiskelija osallistuu tenttiin, jossa hänen on vastattava 7 kysymykseen kymmenestä. Jokainen tehtävä on monivalintatehtävä, jossa on 3 vaihtoehtoa. Kuinka monta erilaista mahdollista vastausta tentissä on?

7. Osoita, että $\binom{n+m}{k} = \binom{n}{0}\binom{m}{k} + \binom{n}{1}\binom{m}{k-1} + \ldots + \binom{n}{k}\binom{m}{0}$. (VINKKI: Mieti n aikuisen ja m lapsen joukkoa. Montako k ihmisen ryhmää on mahdollista muodostaa?)

8. Eräässä aakkostossa on 7 kirjainta. Kuinka monta

   a) tasan 10-kirjaimista

   b) enintään 10-kirjaimista

   sanaa (kirjainjonoa) voidaan muodostaa?

9. Juniorijalkapallojoukkueella on 15 pelaajaa. Avauskokoonpanoon pitää valita 2 hyökkääjää, 3 keskikenttäpelaajaa ja 3 puolustajaa, sekä maalivahti. Monellako tavalla avauskokoonpano voidaan valita? (VINKKI: Multinomikaava. Muista, että avauskokoonpanon lisäksi jäljelle jää ns. 'vaihtopelaajien osajoukko' eli niiden pelaajien joukko, joita ei valita avauskokoonpanoon)

10. Kuinka monta erilaista 11-kirjaimista jonoa voidaan muodostaa kirjaimista M, I, S, S, I, S, S, I, P, P, I?

11. Kuinka moni joukon $\{1,2, \ldots, 20\}$ 4-alkioisista osajoukoista sisältää jonkin alkioista 1,2,3,4,5?

12. Eräässä kerrostalossa asuu 2 yhden vanhemman ja yhden lapsen perhettä, 1 yhden vanhemman ja kahden lapsen perhe, 5 kahden vanhemman ja yhden lapsen perhettä, 6 kahden vanhemman ja kahden lapsen perhettä ja 3 kahden vanhemman ja kolmen lapsen perhettä. Monellako tapaa talosta voidaan valita samasta perheestä peräisin oleva vanhempi ja lapsi.

Luku 2

Matemaattinen todennäköisyys

2.1 Klassinen todennäköisyys

Tällä kurssilla käsitellään todennäköisyyttä matemaattisena käsitteenä, jonka tarkoituksena on mallintaa satunnaisuutta reaalimaailman ilmiöissä. Usein puhutaan *satunnaiskokeesta*, jonka mallintaminen on tavoitteena.

Yksinkertaisimmassa tapauksessa satunnaiskokeen mahdollisten lopputulosten määrä on äärellinen ja ne kaikki ovat yhtä todennäköisiä. Tällöin kyseessä on *äärellinen* ja *symmetrinen* todennäköisyysavaruus (kuten esimerkissä 1.2).

Esimerkkejä tällaisesta satunnaiskokeesta ovat kolikon tai arpakuution heitto, viiden kortin pokerikäden nostaminen sekoitetusta korttipakasta, lottoarvonta tai ruletin pyöräytys. Kuten luvussa 0 mainittiin, tämä historiallisesti varhaisin tapa mallintaa todennäköisyyttä saikin innoituksensa juuri uhkapeleistä.

Tapauksessa, jossa kaikki lopputulokset ovat symmetrisiä, tapahtuman todennäköisyys saadaan jakamalla sille suotuisten lopputulosten määrä kaikkien mahdollisten lopputulosten määrällä. Tätä kutsutaan todennäköisyyden klassiseksi määritelmäksi.

Olemme jo aiemmin laskeneet yksinkertaisia todennäköisyyslaskuja määrittelemättä sen tarkemmin perusjoukon, osajoukon tai klassisen todennäköisyyden käsitteitä. Määritellään nyt nämä käsitteet, ja esitetään sen jälkeen lisää esimerkkilaskuja.

Määritelmä 2.1. *Perusjoukko.*[1] Olkoon Ω satunnaiskokeen mahdollisten loputulosten joukko. Kutsumme sitä *perusjoukoksi*, sen alkioita $\omega \in \Omega$ *alkeistapauksiksi.*

Määritelmä 2.2. Äärellisen perusjoukon $\Omega = \{\omega_1, \dots, \omega_n\}$ tapauksessa kutsumme kaikkia perusjoukon osajoukkoja $A \subseteq \Omega$ *tapahtumiksi.* Merkitsemme kaikkien tapahtumien

[1] engl. *sample space*

joukkoa $\mathcal{F}$:llä. Äärellisen perusjoukon tapauksessa tämä on perusjoukon kaikkien osajoukkojen joukko, eli $\mathcal{F} = \mathcal{P}(\Omega)$.[2]

Määritelmä 2.3. *Klassinen todennäköisyys.*[3] Olkoon $A \subseteq \Omega$ äärellisen ja symmetrisen perusjoukon Ω tapahtuma. A:n todennäköisyys on

$$P(A) = \frac{|A|}{n},$$

missä $|A|$ on A:n alkioiden lukumäärä ja $n := |\Omega|$ on koko perusjoukon alkioiden lukumäärä.

Huomaa, että klassisessa todennäköisyyden mallissa kaikki alkeistapaukset ovat yhtä todennäköisiä, sillä

$$P(\{\omega\}) = \frac{|\{\omega\}|}{n} = \frac{1}{n} \quad \text{kaikille } \omega \in \Omega.$$

Tarkastellaan seuraavaksi esimerkkien kautta, mitä perusjoukon, alkeistapauksen ja klassisen todennäköisyyden käsitteet käytännössä tarkoittavat.

Esimerkki 2.4. Heitetään yhtä kolikkoa. Alkeistapaukset ovat ω_1 = kruuna ja ω_2 = klaava. Siten perusjoukko

$$\Omega = \{\text{kruuna}, \text{klaava}\},$$

sisältää $n = 2$ alkeistapausta, ja kaikkien tapahtumien joukko on

$$\mathcal{F} = \{\emptyset, \{\text{kruuna}\}, \{\text{klaava}\}, \{\text{kruuna, klaava}\}\}.$$

Näiden tapahtumien todennäköisyydet ovat:

$$\begin{aligned}
&P(\emptyset) = 0 \quad \text{(ei kruunaa eikä klaavaa)}\\
&P(\{\text{kruuna}\}) = \frac{1}{n} = \frac{1}{2}\\
&P(\{\text{klaava}\}) = \frac{1}{n} = \frac{1}{2}\\
&P(\{\text{kruuna, klaava}\}) = P(\Omega) = 1 \quad \text{(kruuna tai klaava).}
\end{aligned}$$

Huomautus 2.5. Yleensä tapahtuman todennäköisyyttä merkittäessä jätetään kaksista sulkeista toiset pois, esimerkiksi tapahtuman $\{\omega_1, \omega_2\}$ todennäköisyys voidaan kirjoittaa muodossa

$$P(\{\omega_1, \omega_2\}) = P\{\omega_1, \omega_2\}$$

[2] Huomaathan, että kaikkien tapahtumien joukkona $\mathcal{F}$ voidaan käyttää perusjoukon potenssijoukkoa $\mathcal{P}(\Omega)$ vain siinä erikoistapauksessa, missä perusjoukko on äärellinen. Palaamme tähän myöhemmin ei-äärellisten perusjoukkojen yhteydessä.

[3] engl. *classical definition of probability*

tai muodossa
$$P(\{\omega_1, \omega_2\}) = P(\omega_1, \omega_2).$$
Jatkossa käytämme näistä jälkimmäistä muotoa. Siten edellisen esimerkin todennäköisyys tapahtumalle {kruuna, klaava} (saadaan kruuna tai klaava), voidaan kirjoittaa kompaktimmin muodossa
$$P(\text{kruuna, klaava}).$$

Esimerkki 2.6. Yhden kortin nostaminen korttipakasta. Oletetaan, että meillä on (hyvin) sekoitettu 52:n kortin korttipakka. Alkeistapauksia ovat esimerkiksi ♠2, ♣7 ja $\heartsuit J$. Perusjoukko on siis

$$\Omega = \{(x, y) \,|\, x \in \{\clubsuit, \spadesuit, \diamondsuit, \heartsuit\}, y \in \{2, 3, \ldots, 10, J, Q, K, A\}\}.$$

Perusjoukossa on $n = 52$ alkiota, jotka voitaisiin edelleen periaatteessa kaikki luetella. Kuitenkin erilaisia tapahtumia on 2^{52} kappaletta[4], joten niiden luetteleminen ei (kätevästi) onnistu. Tarkastellaan sen sijaan muutamaa esimerkkitapahtumaa. Tapahtumia ovat esimerkiksi A : ”Kortti on musta”:

$$A = \{(x, y) \in \Omega \,|\, x \in \{\spadesuit, \clubsuit\}\}$$

ja B : ”Kortti on ruutu ja arvoltaan pienempi kuin 5”:

$$B = \{(x, y) \in \Omega \,|\, x \in \{\diamondsuit\} \text{ ja } y \in \{2, 3, 4\}\}.$$

Näiden tapahtumien todennäköisyydet saadaan jälleen suoraan klassisesta todennäköisyyden määritelmästä:
$$P(A) = \frac{|A|}{n} = \frac{26}{52} = \frac{1}{2}$$
ja
$$P(B) = \frac{|B|}{n} = \frac{3}{52}.$$

Esimerkki 2.7. Kahden kuusisivuisen nopan heitto. Alkeistapauksia ovat järjestetyt parit
$$(1, 1), (1, 2), \ldots, (6, 5), (6, 6),$$
eli $\Omega = E \times E$, $E = \{1, 2, \ldots, 6\}$.

[4]Kuten määritelmässä 2.2 todetaan, tapahtumien joukko on alkeistapausten joukon potenssijoukko. Jokaisen alkeistapauksen voidaan ajatella joko kuuluvan tai olevan kuulumatta kuhunkin tapahtumaan, jolloin erilaisten tapahtumien lukumäärä saadaan laskulla 2^{52} (Lause 1.6)

Nyt erilaisia tapahtumia on 2^{36} kappaletta. Merkitään A:lla tapahtumaa "Noppien silmälukujen summa on yli yhdeksän", jolloin A voidaan kirjoittaa muodossa

$$\begin{aligned} A &= \{(x,y) \in \Omega \,|\, x+y > 9\} \\ &= \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}. \end{aligned}$$

Tällöin klassisen todennäköisyyden määritelmän nojalla

$$P(A) = \frac{|A|}{n} = \frac{6}{36} = \frac{1}{6}.$$

Merkitään B:llä tapahtumaa "molemmat silmäluvut ovat pienempiä kuin kolme". Tällöin B voidaan esittää muodossa

$$\begin{aligned} B &= \{(x,y) \in \Omega \,|\, x < 3 \text{ ja } y < 3\} \\ &= \{(1,1),(1,2),(2,1),(2,2)\}. \end{aligned}$$

Jälleen klassisen todennäköisyyden määritelmän nojalla

$$P(B) = \frac{|B|}{n} = \frac{4}{36} = \frac{1}{9}.$$

Esimerkki 2.8. Muistellaan vielä hetki kombinaatioita ja esimerkkiä 1.13 (pokerikäsi). Perusjoukoksi Ω kannattaa tässä yhteydessä valita kaikki mahdolliset 52:n kortin 5:n alkion osajoukot. Esimerkistä 1.13 muistamme, että näitä korttipakan 5-kombinaatioita, eli erilaisia 5 kortin käsiä on yhteensä $\binom{52}{5} = 2598960$ kappaletta.

Näin ollen jokaisen viiden kortin yhdistelmän, esimerkiksi herttareedin (herttavärisuora kympistä ässään), todennäköisyys on

$$P\{\heartsuit 10, \heartsuit J, \heartsuit Q, \heartsuit K, \heartsuit A\} = \frac{1}{\binom{52}{5}} = \frac{1}{2598960} \approx 3.85 \cdot 10^{-7}.$$

Lasketaan seuraavaksi todennäköisyys tapahtumalle A : "nostetaan käsi, jossa on neljä ässää". Tälläisen 5-kombinaation tulee sisältää ainakin kortit $\heartsuit A$, $\diamondsuit A$, $\clubsuit A$ ja $\spadesuit A$. Viimeinen kortti voi olla mikä tahansa jäljelläolevasta 48 kortista, joten tapahtuma A sisältää 48 perusjoukon alkiota (pokerikättä, eli korttipakan 5-kombinaatiota). Siten tapahtuman A todennäköisyys on

$$P(A) = \frac{48}{\binom{52}{5}} \approx 1.85 \cdot 10^{-5}.$$

Esimerkki 2.9. Lasketaan vielä todennäköisyydet tapahtumille B : "nostetaan käsi, jossa on täsmälleen kolme ässää", ja C : "nostetaan käsi, jossa on vähintään kolme ässää".

Tapahtumien B ja C alkioiden määrää laskettaessa tulee olla tarkkana, ettei samaa kättä laske mukaan useampaan kertaan.

Lasketaan ensin käsien, joissa on *täsmälleen* kolme ässää, määrä. Näiden käsien valitseminen voidaan jakaa kahteen osaan: valitaan ensin kolme ässää, ja sen jälkeen kaksi muuta korttia. Kolme ässää voidaan valita neljällä erilaisella tavalla:

$$\{\heartsuit A, \diamondsuit A, \clubsuit A\},$$
$$\{\heartsuit A, \diamondsuit A, \spadesuit A\},$$
$$\{\heartsuit A, \spadesuit A, \clubsuit A\},$$

tai

$$\{\spadesuit A, \diamondsuit A, \clubsuit A\}.$$

Toisessa vaiheessa valitaan 2 korttia muuta korttia. Vaihtoehtoja on 48, koska kolme ässistä on jo käytetty, ja neljättä ei voida valita, koska muuten kädessä olisi *yli* kolme ässää. Lauseen 1.12 kohdan (iii) nojalla 48:sta kortista voidaan valita $\binom{48}{2} = 1128$ kahden kortin joukkoa eli 2-kombinaatiota. Siten tuloperiaatteen nojalla erilaisia käsiä, joissa on täsmälleen 3 ässää, on $4 \cdot 1128 = 4512$ kappaletta, joten todennäköisyys nostaa täsmälleen kolme ässää on

$$P(B) = \frac{4512}{\binom{52}{5}} \approx 0.001736$$

Käsien, joissa on *vähintään* kolme ässää, lukumäärä saadaan laskemalla yhteen käsien, joissa on täsmälleen kolme ässää, ja käsien, joissa on täsmälleen neljä ässää, määrä: erilaisia vaihtoehtoja on yhteensä $4512 + 48 = 4560$ kappaletta. Siten todennäköisyys nostaa vähintään kolme ässää on

$$P(C) = \frac{4560}{\binom{52}{5}} \approx 0.001755.$$

2.1.1 Otanta

Monesti tilastollisessa tutkimuksessa halutaan tutkia jonkin populaation ominaisuuksia. Selvitettäviä kysymyksiä voisivat olla esimerkiksi, kuinka suuri osa suomalaisista kannattaa sotilasliitto NATO:on liittymistä, tai kuinka suuri osa jonkin järven kaloista on alamittaisia. Kaikkien suomalaisten NATO-kannan kysyminen, tai kaikkien järven kalojen kalastaminen on käytännössä lähes mahdotonta, joten yleensä tilastollisessa tutkimuksessa suoritetaan otanta, eli poimitaan tutkittavasta populaatiosta satunnaisotos, jonka koko on yleensä paljon koko populaation kokoa pienempi. Otanta voidaan suorittaa takaisinpanolla eli palauttaen, jolloin sama yksilö voi päätyä otokseen useamman kerran, tai ilman takaisinpanoa, jolloin sama yksilö voi päätyä otokseen vain kerran.

Esimerkki 2.10. Laatikossa on 8 palloa, joista 5 on valkoista ja $8 - 5 = 3$ mustia. Laatikko voidaan kuvata joukkona

$$E = \{v_1, v_2, v_3, v_4, v_5, m_1, m_2, m_3\},$$

missä v:llä merkitään valkoista, ja m:llä mustaa palloa.

(i) **Otanta ilman takaisinpanoa.** Nostetaan laatikosta satunnaisesti 4 palloa palauttamatta nostettuja palloja nostojen välillä laatikkoon. Mikä on tapahtuman $A_2 =$ ”Nostetaan tasan 2 valkoista palloa” todennäköisyys?

Perusjoukoksi Ω kannattaa valita kaikkien 8 pallon joukon 4 pallon osajoukot, eli 4-kombinaatiot; järjestyksellä, jossa pallot on nostettu, ei ole tutkittavan tapahtuman kannalta merkitystä. Perusjoukon alkioita $\omega \in \Omega$ ovat esimerkiksi

$$\{v_3, v_4, m_1, m_3\}$$

ja

$$\{v_1, v_2, v_4, m_1\}.$$

Näitä on lauseen 1.12 kohdan (iii) nojalla $\binom{8}{4}$ kappaletta. Näistä tapahtumalle A_3 suotuisia alkeistapauksia ovat ne, joissa on tasan 2 valkoista ja $4 - 2 = 2$ mustaa palloa, eli esimerkiksi

$$\{v_3, v_4, m_1, m_3\}$$

ja

$$\{v_1, v_4, m_1, m_2\}.$$

Lauseen 1.12 kohdan (iii) nojalla valkoiset pallot voidaan valita $\binom{5}{2}$ erilaisella, ja mustat pallot $\binom{3}{2}$ erilaisella tavalla. Siten erilaisia mahdollisia yhdistelmiä, joissa on tasan 2 valkoista palloa, on tuloperiaatteen nojalla $\binom{5}{2}\binom{3}{2}$ kappaletta, joten todennäköisyys nostaa tasan 2 valkoista palloa on

$$P(A_2) = \frac{\binom{5}{2}\binom{3}{2}}{\binom{8}{4}} = \frac{10 \cdot 3}{70} = \frac{3}{7} \approx 0.4286.$$

(ii) **Otanta takaisinpanolla.** Otanta takaisinpanolla. Nostetaan laatikosta satunnaisesti 4 palloa palauttaen nostojen välillä nostettu pallo takaisin laatikkoon. Mikä on tapahtuman $A_2 =$ ”Nostetaan tasan 2 valkoista palloa” todennäköisyys?

Nyt perusjoukoksi Ω kannattaa valita kaikki 4:n pallon jonot, jotka on mahdollista muodostaa 8 pallosta siten, että sama pallo voi esiintyä useampaan kertaan. Perusjoukon alkioita $\omega \in \Omega$ ovat esimerkiksi:

$$(m_3, m_3, m_1, m_3)$$

ja

$$(v_3, v_2, v_3, m_2).$$

Tuloperiaatteen (luku 1.4) nojalla näitä jonoja on 8^4 kappaletta.

Lasketaan, kuinka monessa näistä jonoista on täsmälleen 2 valkoista palloa. Selvitetään ensimmäiseksi, kuinka monta erilaista tapaa on järjestää 4 palloa, joista 2 on valkoisia ja 2 mustia, jonoiksi. Tällaisiä jonoja ovat esimerkiksi

$$(m, m, v, v)$$

ja

$$(v, m, v, m).$$

Lauseen 1.12 (iii)-kohdan nojalla nämä ehdot täyttäviä permutaatioita on $\binom{4}{2}$ kappaletta. Näiden erilaisten järjestysten valkoiset pallot ja mustat pallot voidaan edelleen valita useammalla eri tavalla. Esimerkiksi järjestyksen (m, m, v, v) toteuttavat

$$(m_2, m_2, v_5, v_3)$$

ja

$$(m_3, m_1, v_1, v_3).$$

Tuloperiaatteen nojalla valkoiset pallot kuhunkin järjestykseen voidaan valita 5^2 erilaisella, ja mustat pallot 3^2 erilaisella tavalla.

Edelleen tuloperiaatteen nojalla erilaisia 4:n pallon jonoja, joissa 2 palloista on valkoisia, on $\binom{4}{2} \cdot 5^2 \cdot 3^2$ kappaletta, joten todennäköisyys nostaa tasan 2 valkoista palloa on

$$P(A_2) = \frac{\binom{4}{2} \cdot 5^2 \cdot 3^2}{8^4} = \frac{1350}{4096} \approx 0.3296.$$

Tarkastellaan kahdentyyppisistä alkioista koostuvaa N-alkioista joukkoa $E = A \cup B$, missä $A = \{a_1, \ldots, a_K\}$ ja $B = \{b_1, \ldots, b_{N-K}\}$. E voi olla esimerkiksi laatikossa olevien pallojen, joista K on valkoisia ja $N - K$ mustia, joukko. Edellinen esimerkki yleistyy seuraavasti.

Lause 2.11. *Otanta ilman takaisinpanoa ja takaisinpanolla.*[5] *Olkoon* $n \in \{1, \ldots, N\}$ *ja* $k \in \{0, \ldots, n\}$.

(i) Valitaan yllämainitusta joukosta E *n:n kappaleen otos satunnaisesti ilman takaisinpanoa. Todennäköisyys tapahtumalle* A_k*: "otoksessa on täsmälleen* k *joukon* A *alkiota" on*

$$P(A_k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

(ii) Valitaan yllämainitusta joukosta E *n:n kappaleen otos satunnaisesti takaisinpanolla. Todennäköisyys tapahtumalle* A_k*: "otoksessa on täsmälleen* k *joukon* A *alkiota" on*

$$P(A_k) = \binom{n}{k}\frac{K^k(N-K)^{(n-k)}}{N^n}.$$

Todistus. Kuten esimerkissä 2.10. □

Pokerikäden nostamisessa korttipakasta kyse on otannasta ilman takaisinpanoa. Siten esimerkin 2.9 tulos täsmälleen kolmen ässän todennäköisyydelle seuraa suoraan edellisestä lauseesta, kun valitaan $N = 52$, $K = 4$, $n = 5$ ja $k = 3$:

$$P(B) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{K}{n}} = \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \approx 0.001736.$$

Jos taas nostetaan korttipakasta 5 korttia siten, että sekoitetaan nostettu kortti pakkaan aina nostojen välillä, kyse on otannasta takaisinpanolla. Siten todennäköisyys nostaa täsmälleen 3 ässää (merkitään tätä tapahtumaa D:llä) saadaan myös suoraan edellisesta lauseesta

$$P(D) = \binom{n}{k}\frac{K^k(N-K)^{(n-k)}}{N^n} = \binom{5}{3}\frac{4^3 \cdot 48^2}{52^5} \approx 0.003878.$$

Huomautus 2.12. Perusjoukon valinta on aina viime kädessä mielivaltainen, ja samaa satunnaiskoetta voidaan mallintaa usealla eri perusjoukolla. Usein, kuten tämän luvun esimerkeissä, perusjoukon valinta on ilmeinen. Muitakin mahdollisuuksia kuitenkin on: esimerkiksi jos meitä kiinnostaa vain kahden nopan yhteenlaskettu silmäluku, voimme käyttää alkeistapauksina kahden nopan mahdollisten yhteenlaskettujen silmälukujen joukkoa

$$\Omega_1 = \{2, \ldots, 12\}.$$

[5] engl. *sampling without/with replacement*

Tällä perusjoukolla voimme esittää esimerkin 2.7 tapahtuman A (silmälukujen summa yli 9) muodossa

$$A = \{\omega \in \Omega_1 \,|\, \omega > 9\} = \{10, 11, 12\}.$$

Sen sijaan saman esimerkin tapahtumaa B (molemmat silmäluvut pienempää kuin 3) emme voi esittää perusjoukon Ω_1 avulla. Perusjoukon tuleekin olla riittävän rikas, jotta kaikki meitä kiinnostavat tapahtumat voidaan esittää sen osajoukkoina.

Myöskään tapahtumaan A emme voi perusjoukolla Ω_1 soveltaa klassisen todennäköisyyden määritelmää, koska tilanne ei ole enää symmetrinen (alkeistapaukset eivät ole yhtä todennäköisiä).

Klassisen todennäköisyyden käsite ei siis aina ole paras mahdollinen. Kuvitellaan tilanne, jossa heitämme kolikkoa niin kauan, kunnes tulee klaava. Mikä on tämän tilanteen perusjoukko? Huomaat, että perusjoukolle pätee $|\Omega| = n = \infty$. Miksi? Klassisen todennäköisyyskäsitteen rajoitukset tulevat nopeasti vastaan, ja siksi tarvitsemme lisää todennäköisyysmalleja. Tällaisia ovat esimerkiksi frekventistiset ja subjektiiviset todennäköisyyksien tulkinnat[6].

2.2 Todennäköisyyden aksioomat

Jos perusjoukko Ω ei ole *symmetrinen*, eli ei ole perusteita olettaa että alkeistapaukset olisivat yhtä todennäköisiä, klassisen todennäköisyyden määritelmää (määritelmä 2.3), ei voida käyttää satunnaiskokeeseen liittyvän epävarmuuden mallintamiseen. Esimerkkejä tällaisesta satunnaiskokeesta ovat (viiden) tikan heitto, jossa perusjoukoksi voidaan asettaa $\Omega = \{0, 1, 2, \ldots, 10\}^5$, jalkapallo-ottelun lopputulos, jossa luonteva perusjoukko on $\mathbb{N}^2$, tai keihäänheiton tulos, jossa $\Omega = (0, \infty)$.

Kaksi viimeksimainittua esimerkkiä poikkeavat aiemmista tällä kurssilla käsitellyistä esimerkeistä myös siinä suhteessa, että perusjoukko Ω on ääretön; lisäksi viimeiseksi mainitussa esimerkissä perusjoukko on myös (ainakin teoriassa) ylinumeroituva[7].

Tarvitaankin määritelmää 2.3 yleisempää todennäköisyyden määritelmää toisaalta epäsymmetristen satunnaiskokeiden, ja toisaalta laajempien perusjoukkojen käsittelemiseen.

[6] Katso kappale 2.3

[7] Numeroituvuuden käsitettä käytetään mittateoriassa. Käytännössä numeroituva joukko on sellainen, jonka alkiot voidaan luetella. Tällaisia joukkoja ovat mm. luonnolliset luvut $\{1,2,3, \ldots\}$ tai parilliset kokonaisluvut $\{2,4,6 \ldots\}$. Joukko, joka ei ole numeroituva, on ylinumeroituva. Tällaisia joukkoja ovat esimerkiksi reaaliakselin luvut, tai välin $[0,1]$ reaaliluvut, joiden kaikkia alkioita ei voida edes teoriassa luetella.

Jos keihäänheiton tulosta voitaisiin mitata äärettömän tarkasti olisi keihäänheiton tulosten perusjoukko ylinumeroituva. Käytännössä tulokset ilmoitetaan kuitenkin yhden senttimetrin välein, jolloin perusjoukko on numeroituva.

Määrittelemme seuraavaksi todennäköisyyden tietyt ominaisuudet, joita kutsumme *todennäköisyyden aksioomiksi*, täyttävänä joukkofunktiona tapahtumien $\mathcal{F}$ joukolta välille $[0, 1]$.

Kun perusjoukko Ω on äärellinen tai korkeintaan numeroituvasti ääretön, niin tapahtumien joukoksi $\mathcal{F}$ voidaan aina valita (voidaan periaatteessa valita myös pienempi joukko) kaikkien perusjoukon osajoukkojen joukko $\mathcal{P}(\Omega)$. Ylinumeroituvan perusjoukon tapauksessa vaaditaan, että kaikkien tapahtumien joukko on ns. *σ-algebra*.

Määritelmä 2.13. *σ-algebra.* Kokoelma $\mathcal{F}$ perusjoukon Ω osajoukkoja on σ-algebra, jos

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Käytännössä σ-algebran ehdot todennäköisyyslaskennan viitekehyksessä tarkoittavat sitä, että tapahtumina voidaan tarkastella vain sellaisia joukkoja, jotka ovat riittävän 'siistejä'. Seuraava lause helpottaa kokoelman $\mathcal{F}$ käsittelyä:

Lause 2.14. *σ-algebran ominaisuuksia. Olkoon $\mathcal{F}$ σ-algebra. Tällöin*

(i) $\emptyset \in \mathcal{F}$,

(ii) Jos $A, B \in \mathcal{F}$, *niin* $A \cup B, A \cap B, A \setminus B \in \mathcal{F}$,

(iii) $A_i \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} \in \mathcal{F}$.

Todistus. Jätetään harjoitustehtäväksi. Perusidea: kohta (i) seuraa siitä, että $\Omega^c = \emptyset$. Kohta (ii) saadaan soveltamalla σ-algebran määritelmän kohtaa 3, ja edelleen kohta (iii) saadaan ns. De Morganin lauseita soveltamalla. □

Ennen mittateorian tuomaa tulikoneistoa todennäköisyyslaskennalle σ-algebran lähempi tarkastelu ei ole kovin mielekästä, joten jätämme σ-algebraan liittyvät lisätarkastelut myöhemmille kursseille. Määrittelemme nyt todennäköisyyden[8] seuraavasti todennäköisyyden aksioomien (*Kolmogorov: Foundations of the Theory of Probability (1933)*) mukaisesti:

[8]Joissain lähteissä mainitaan myös todennäköisyyskuvauksena tai todennäköisyysmittana.

Määritelmä 2.15. *Todennäköisyyden aksioomat.*[9] Olkoon Ω perusjoukko ja σ-algebra $\mathcal{F}$ sen tapahtumien joukko . Kuvaus $P : \mathcal{F} \to \mathbb{R}$ on todennäköisyys (lyh. tn), jos

(TN1) $P(A) \geq 0$ kaikille tapahtumille $A \in \mathcal{F}$,

(TN2) $P(\Omega) = 1$,

(TN3) Kaikille tapahtumille $A_1, A_2, \ldots \in \mathcal{F}$ pätee

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

kun ne ovat erillisiä, eli $A_i \cap A_j = \emptyset$ kaikille $i \neq j$.

Ensimmäinen aksiooma (TN1) vaatii, että kaikkien tapahtumien todennäköisyyden on oltava epänegatiivinen, eli vähintään nolla. Toinen aksiooma (TN2) taas vaatii, että koko perusjoukon Ω todennäköisyyden on oltava 1, eli että jokaisella satunnaiskokeen suorituskerralla joku lopputulos tapahtuu. Kolmas aksiooma (TN3) vaatii, että kaikkien erillisten tapahtumien (numeroituvan) yhdisteen todennäköisyyden täytyy olla näiden tapahtumien todennäköisyyksien summa. Toisin sanoen todennäköisyys sille, että ainakin yksi **erillisistä** tapahtumista A_i tapahtuu saadaan laskemalla yhteen näiden yksittäisten tapahtumien tapahtumistodennäköisyydet $P(A_i)$. Tätä todennäköisyyden ominaisuutta kutsutaan yleensä *täysadditiivisuudeksi.*

Todennäköisyyden käsitteen määrittelemisen jälkeen voimme muodostaa *todennäköisyysavaruuden*, jota tarvitaan perustaksi todennäköisyyksien mallintamiselle.

Määritelmä 2.16. *Todennäköisyysavaruus.*[10] Kolmikkoa $(\Omega, \mathcal{F}, P)$, missä Ω on perusjoukko, σ-algebra $\mathcal{F}$ sen kaikkien tapahtumien joukko, ja $P : \mathcal{F} \to \mathbb{R}$ todennäköisyys, kutsutaan *todennäköisyysavaruudeksi* (lyh. tn-avaruudeksi).

Esitellään seuraavaksi todennäköisyysavaruuden erikoistapaukset *diskreetti todennäköisyysavaruus* ja *äärellinen todennäköisyysavaruus*, sekä *symmetrinen todennäköisyysavaruus*, jota käsittelimme jo klassisen todennäköisyyden ytheydessä.

Määritelmä 2.17. *Diskreetti todennäköisyysavaruus.*[11] Jos perusjoukko Ω on äärellinen, eli $\Omega = \{\omega_1, \ldots, \omega_n\}$, tai numeroituvasti ääretön, eli $\Omega = \{\omega_1, \omega_2, \ldots\}$, tarkasteltavien

[9] engl. *probability axioms*
[10] engl. *probability space*
[11] engl. *discrete probability space*

tapahtumien joukoksi $\mathcal{F}$ voidaan aina valita perusjoukon kaikkien osajoukkojen joukko $\mathcal{P}(\Omega)$, ja todennäköisyysavaruutta $(\Omega, \mathcal{P}(\Omega), P)$ kutsutaan diskreetiksi tn-avaruudeksi.

Koska tapahtumien joukoksi voidaan diskreetissä tapauksessa aina valita kaikkien perusjoukon osajoukkojen joukko, diskreettiä tn-avaruutta merkitään usein lyhyemmin parilla (Ω, P).

Määritelmä 2.18. *Äärellinen todennäköisyysavaruus.*[12] Diskreettiä todennäköisyysavaruutta (Ω, P) kutsutaan äärelliseksi tn-avaruudeksi, jos sen perusjoukko Ω on äärellinen, eli $\Omega = \{\omega_1, \dots, \omega_n\}$ jollekin $n \in \{1, 2, \dots\}$.

Osoitetaan seuraavaksi, että diskreetin perusjoukon tapauksessa kaikkien tapahtumien todennäköisyydet määräytyvät yksikäsitteisesti alkeistapausten todennäköisyyksien summana.

Lause 2.19. *Olkoon Ω joko äärellinen, eli $\Omega = \{\omega_1, \dots, \omega_n\}$, tai ääretön, eli $\Omega = \{\omega_1, \omega_2, \dots\}$, diskreetti perusjoukko, ja vastaavasti luvut $p_1, \dots, p_n$ tai $p_1, p_2, \dots \geq 0$ s.e. $\sum_{i=1}^{\infty} p_i = 1$ [13]. Nyt on olemassa yksi ja vain yksi todennäköisyys, eli aksioomat (TN1)-(TN3) toteuttava joukkofunktio $P : \mathcal{P}(\Omega) \to \mathbb{R}$, jolle pätee*

$$P(\omega_i) = p_i \quad \text{kaikille} \quad i \in \{1, 2, \dots\}. \tag{2.20}$$

Todistus. "$\Rightarrow$": Osoitetaan, että on olemassa ainakin yksi todennäköisyys, jolle 2.20 pätee.

Määritellään funktion $P : \mathcal{P}(\Omega) \to \mathbb{R}$ arvoksi jokaiselle joukolle $A \in \mathcal{P}(\Omega)$:

$$P(A) := \sum_{\omega_i \in A} p_i. \tag{2.21}$$

Tästä P:n määritelmästä seuraa suoraan, että P toteuttaa ehdon 2.20. Osoitetaan vielä, että P toteuttaa todennäköisyyden aksioomat:

(TN1) Määritelmän P:lle (2.21) nojalla jokaiselle tapahtumalle $A \in \mathcal{P}(\Omega)$ pätee

$$P(A) = \sum_{\omega_i \in A} p_i \geq 0.$$

Epänegatiivisuus seuraa siitä, että oletuksen nojalla $p_i \geq 0$ kaikille $i \in \{1, 2, \dots\}$.

(TN2) Jälleen määritelmän 2.21 ja oletuksen nojalla perusjoukolle Ω pätee

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

[12] engl. *finite probability space*

[13] Äärelliselle perusjoukolle käytetään tässä lauseessa merkintöjen selkeyttämiseksi samaa merkintää $\sum_{i=1}^{\infty} p_i$ kuin äärettömälle perusjoukolle. Tällöin asetetaan $p_i = 0$ kaikille $i > n$.

(TN3) Olkoon $A_1, A_2, \dots \in \mathcal{P}(\Omega)$ jono erillisiä tapahtumia. Myös tapahtumien yhdiste $A := \bigcup_{j=1}^{\infty} A_j \in \mathcal{P}(\Omega)$ voidaan esittää numeroituvana yhdisteenä perusjoukon alkioista: $A = \bigcup_{\omega_i \in A}\{\omega_i\}$. Siten määritelmästämme seuraa täysadditiivisuus:

$$\begin{aligned} P\left(\bigcup_{j=1}^{\infty} A_j\right) &= P\left(\bigcup_{j=1}^{\infty} \bigcup_{\omega_i \in A_j} \{\omega_i\}\right) = P\left(\bigcup_{\omega_i \in A} \{\omega_i\}\right) = P(A) \\ &\overset{(2.21)}{=} \sum_{\omega_i \in A} p_i = \sum_{j=1}^{\infty} \sum_{\omega_i \in A_j} p_i \overset{(2.21)}{=} \sum_{j=1}^{\infty} P(A_j). \end{aligned}$$

Toiseksi viimeisessä yhtäsuuruudessa hyödynnettiin oletusta tapahtumien $A_1, A_2, \dots$ erillisyydestä.

Tällä tavoin voimme siis aina löytää todennäköisyyden P, joka toteuttaa ehdon 2.20. Osoitetaan vielä, että tämä P on yksikäsitteinen, eli jos todennäköisyys P^* toteuttaa ehdon 2.20, on välttämättä $P^* = P$, eli $P^*(A) = P(A)$ kaikille $A \in \mathcal{P}(\Omega)$.

"$\Leftarrow$": Olkoon P^* todennäköisyys, eli aksioomat (TN1)-(TN3) toteuttava joukkofunktio $P^* : \mathcal{P}(\Omega) \to \mathbb{R}$, jolle 2.20 pätee. Osoitetaan, että jokaiselle joukolle $A \in \mathcal{P}(\Omega)$ pätee 2.21, eli että

$$P^*(A) = \sum_{\omega_i \in A} p_i,$$

mistä seuraa $P^* = P$.

Tehdään vastaoletus: on olemassa joukko $A \in \mathcal{P}(\Omega)$ s.e.

$$P^*(A) \neq \sum_{\omega_i \in A} p_i.$$

Vastaoletuksen nojalle erillisille joukoille $\{\omega_i\} \in A$ pätee (Huomaa, että sekä yhdiste $\bigcup_{\omega_i \in A}\{\omega_i\}$, että summa $\sum_{\omega_i \in A} P^*(\omega_i)$ ovat numeroituvia. Jos A on äärellinen, valitaan yhdisteen lopuiksi joukoiksi tyhjiä joukkoja.):

$$P^*\left(\bigcup_{\omega_i \in A} \{\omega_i\}\right) = P^*(A) \neq \sum_{\omega_i \in A} p_i \overset{2.20}{=} \sum_{\omega_i \in A} P^*(\omega_i).$$

Mutta tämä on ristiriidassa P^*:n oletetun täysadditiivisuuden kanssa. Siten vastaoletus on väärä, joten on oltava $P^* = P$.

$\square$

Esimerkki 2.22. Symmetrinen tn-avaruus. Olkoon $\Omega = \{\omega_1, \dots, \omega_n\}$ perusjoukko, ja asetetaan

$$P(\omega_i) = p_i = \frac{1}{n} \quad \text{kaikille} \quad i \in \{1, \dots, n\}.$$

Nyt lauseen 2.19 nojalla (huomaa, että $\sum_{i=1}^n p_i = 1$) P määrää todennäköisyyden yksikäsitteisesti kaikille tapahtumille $A \in \mathcal{P}(\Omega)$ siten, että

$$P(A) = \sum_{\omega_i \in A} p_i = \frac{|A|}{n}. \tag{2.23}$$

Tällaista tn-avaruutta (Ω, P), jonka kaikkien alkeistapausten todennäköisyydet ovat samat, kutsutaan *symmetriseksi* todennäköisyysavaruudeksi.

Huomaa, että näin määritelty P on yhtenevä todennäköisyyden klassisen määritelmän (määritelmä 2.3) kanssa. Siten klassisen määritelmän mukainen todennäköisyys on erikoistapaus yleisemmän määritelmän (määritelmä 2.15) mukaisesta todennäköisyydestä.

Kaikki aiemmin tällä kurssilla käsitellyt esimerkit, kuten kolikonheitto, kahden nopan heitto, pokerikäsi ja 4 pallon nostaminen laatikosta, on mallinnettu symmetrisen tn-avaruuden avulla. Esitellään seuraavaksi esimerkit epäsymmetrisestä diskreetistä todennäköisyysavaruudesta äärellisellä ja äärettömällä[14] perusjoukolla.

Esimerkki 2.24. Heitä sikaa. Suositussa Heitä sikaa-pelissä heitetään sian muotoisia arpakuutioita. Pisteitä heitosta saa sen mukaan, mihin asentoon sika jää maahan tullessaan. Pisteet ja frekvenssit [15] yhdelle sialle ovat seuraavat:

	Nimi	Asento	Pisteet	Frekvenssi (%)
ω_1	Siankylki	Kyljellään	0 p.	65.1
ω_2	Sianselkä	Selällään	5 p.	22.5
ω_3	Tavallinen sika	Jaloillaan	5 p.	8.8
ω_4	Siansyltty	Etujalkojen ja kärsän varassa	10 p.	3.0
ω_5	Etupotka	Toisen korvan ja etujalan varassa	15 p.	0.6

[14]Itse asiassa ääretön diskreetti tn-avaruus ei voi olla symmetrinen. Tämän voi itse kokeilla osoittaa tn:n aksioomien perusteella.

[15]Simuloitu standardoidulla pinnalla ja automaattisella heittolaitteella (otoskoko $n = 11954$). Lähde: https://en.wikipedia.org/wiki/Pass_the_Pigs

Asetetaan P luvuille $p_1 = 0.651, p_2 = 0.225, p_3, = 0.088, p_4 = 0.030$ ja $p_5 = 0.006$ seuraavasti:

$$P(\omega_i) = p_i \quad \text{kaikille} \quad i \in \{1, 2, 3, 4, 5\},$$

Lauseen 2.19 nojalla P on todennäköisyys perusjoukolla $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ (huomaa, että $\sum_{i=1}^{5} p_i = 1$). Siten (Ω, P) on äärellinen (ja diskreetti) tn-avaruus.

Lisäksi kaikkien tapahtumien todennäköisyydet saadaan niiden sisältämien alkeistapauksien todennäköisyyksien summana. Esimerkiksi tapahtuman A : ”Heitolla saadaan tasan 5 pistettä” todennäköisyys on

$$P(A) = P(\omega_2, \omega_3) = p_2 + p_3 = 0.313.$$

Esimerkki 2.25. Kotijoukkueen tekemien maalien määrä jalkapallo-ottelussa. Periaatteessa, vaikkakaan ei käytännössä [16], joukkueen tekemien maalien määrä ei ole rajoitettu, vaan se voi olla mikä tahansa epänegatiivinen kokonaisluku. Perusjoukoksi on tällöin luontevaa asettaa

$$\Omega = \{0, 1, 2 \dots\} = \mathbb{N}.$$

Jos perusjoukko on ääretön, perusjoukon alkioiden todennäköisyyksien valinta siten, että ne summautuvat ykköseksi, on hieman hankalampaa kuin äärellisen perusjoukon tapauksessa, sillä nyt kyse on äärettömästä summasta. Yksi vaihtoehto valita todennäköisyydet on seuraava.

Olkoon $\lambda \in (0, \infty)$. Asetetaan

$$p_i = e^{-\lambda} \frac{\lambda^i}{i!} \quad \text{kaikille} \quad i \in \{0, 1, 2, \dots\}.$$

Nyt eksponenttifunktion sarjakehitelmän $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ nojalla

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1,$$

joten luvut p_i määräävät todennäköisyyden äärettömälle perusjoukolle $\Omega = \{0, 1, 2, \dots\} = \mathbb{N}$. Siten $(\mathbb{N}, P)$, missä

$$P(i) = e^{-\lambda} \frac{\lambda^i}{i!} \quad \text{kaikille} \quad i \in \mathbb{N}$$

[16] Jalkapallon peliaika on rajoitettu, joten mielivaltaisen suuren maalimäärän ei käytännössä olisi fyysisesti mahdollista toteutua. Suurin voittomarginaali ammattilaistason jalkapallo-ottelussa on 12. syyskuuta 1885, kun Arbroath voitti Bon Accordin 36-0 Skotlannin cupin ottelussa. Tosin lokakuussa 2002 AS Adema voitti SO l’Emyrne:n 149-0 Madagaskarin mestaruudesta pelatussa neljän joukkueen turnauksessa. Tätä ei kuitenkaan lasketa, sillä kaikki maalit olivat omia: SO l’Emyrne hävisi ottelun tahallaan protestiksi sarjan aiempien ottelujen tuomarityöskentelystä. Lähde: `https://en.wikipedia.org/wiki/Arbroath_36-0_Bon_Accord`, `https://en.wikipedia.org/wiki/AS_Adema_149-0_SO_l%27Emyrne`

on diskreetti tn-avaruus. Näin määriteltyä todennäköisyyttä kutsutaan *Poisson-jakaumaksi*[17].

Jos jonkin asian sattumisen todennäköisyys jollain aikavälillä on vakio, ja näiden sattumisten ajankohdat ovat toisistaan riippumattomia, näiden tapahtumien määrää tietyllä aikavälillä voidaan mallintaa Poisson-jakaumalla. Onnekkaasti osoittautuu, että esimerkiksi joukkueen jalkapallo-ottelussa tekemien maalien määrä noudattaa melko tarkkaan Poisson-jakaumaa. Seuraavaan taulukkoon on merkitty kotijoukkueen ottelussa tekemien maalien määrät ja frekvenssit La Ligassa (Espanjan pääsarja) kausilla 2010-2015 (yhteensä 2116 ottelua).

Maalit	Ottelut	Frekvenssi (%)	Poisson-jakaumasta ($\lambda = 1.631$) laskettu tn
0	466	22.02	0.1957
1	668	31.57	0.3192
2	498	23.53	0.2604
3	262	12.38	0.1416
4	147	6.95	0.0577
5	50	2.36	0.0188
6	15	0.71	0.0051
7	7	0.33	0.0012
8	2	0.09	0.0002
9	1	0.05	0.0000

Kotijoukkue on siis jäänyt 466 ottelussa nollaan maaliin, 668 ottelussa tehnyt 1 maalia, ja niin edelleen. Lisäksi taulukkoon on merkitty Poisson-jakaumasta parametrin $\lambda = 1.631$ arvolla [18] lasketut todennäköisyydet. Vaikka aineistossamme suurin maalimäärä on 9, Poisson-jakauman avulla saadaan positiiviset todennäköisyydet kaikille perusjoukon alkioille: esimerkiksi $P(10) = 7.20 \cdot 10^{-6}$.

Seuraavassa esimerkissä ratkaistaan todennäköisyyslaskennan aksioomien avulla epäsymmetristä todennäköisyysavaruutta käsittelevä ongelma.

Esimerkki 2.26. Olkoon erään todennäköisyysavaruuden perusjoukko $\Omega = \{1, 2, \ldots, 10\}$. Määritetään todennäköisyydet $p_k = \mathrm{P}(\mathrm{k})$ kaikille perusjoukon alkioille $k \in \Omega$ siten, että p_k on

[17] Poisson-jakaumaa ja muita todennäköisyysjakaumia käsitellään lisää myöhemmin satunnaismuuttujien yhteydessä.

[18] Kotijoukkueen maalimäärien keskiarvo aineistossa. Parametrien päättelyyn ja tilastollisen mallin sovittamiseen aineistoon perehdytään tarkemmin kurssilla Tilastollinen päättely I.

(a) suoraan verrannollinen lukuun k

(b) suoraan verrannollinen lukuun $\ln(k)$.

Lasketaan lisäksi kummassakin tapauksessa todennäköisyydet tapahtumille

A : ”tulos on suurempaa kuin 6”

ja

B: ”tulos on kokonaisluvun neliö”.

Todennäköisyyden aksioomien nojalla $p_k = P(k) \geq 0$ kaikilla $k = 1, 2, \ldots, 10$ ja $\sum_{k=1}^{10} p_k = 1$.

(a) Koska p_k on suoraan verrannollinen lukuun k, eli $\frac{p_k}{k} = c$ on vakio kaikilla $k = 1, 2, \ldots, 10$, niin $p_k = kc \geq 0$, jolloin myös $c \geq 0$ ja

$$1 = \sum_{k=1}^{10} p_k = \sum_{k=1}^{10} kc = c\Big(\sum_{k=1}^{10} k\Big) = c \cdot \frac{10 \cdot 11}{2} = 55c$$

(Yllä käytettiin summakaavaa $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$).
Nyt voidaan ratkaista $c = \frac{1}{55}$, joten $p_k = \frac{k}{55}$, kun $k = 1, 2, \ldots, 10$. Edelleen, koska A : ”tulos on suurempi kuin 6” = {7,8,9,10}, niin

$$P(A) = \sum_{k=7}^{10} p_k = \frac{7}{55} + \frac{8}{55} + \frac{9}{55} + \frac{10}{55} = \frac{34}{55} \approx 0.618.$$

Samoin, koska B : ”tulos on kokonaisluvun neliö” = {1, 4, 9}, niin

$$P(B) = \sum_{k \in \{1,4,9\}} p_k = \frac{1}{55} + \frac{4}{55} + \frac{9}{55} = \frac{14}{55} \approx 0.255.$$

(b) Nyt $\frac{p_k}{\ln(k)} = c$ on vakio kaikilla $k = 1, 2, \ldots, 10$, jolloin $p_k = c \cdot \underbrace{\ln(k)}_{\geq 0 \ \forall k \geq 1}$. Näin ollen $c \geq 0$ ja

$$1 = \sum_{k=1}^{10} p_k = \sum_{k=1}^{10} c \cdot \ln(k) = c\Big(\sum_{k=1}^{10} \ln(k)\Big) = c \cdot \ln(1 \cdot 2 \cdot \ldots \cdot 10) = c \cdot \ln 10!.$$

Nyt voidaan ratkaista $c = \frac{1}{\ln 10!}$, joten $p_k = \frac{\ln(k)}{\ln 10!}$ (koska p_k ja $\ln(k)$ olivat suoraan verrannollisia). Näin ollen

$$\begin{aligned} P(A) &= \sum_{k=7}^{10} p_k = \frac{\ln 7}{\ln 10!} + \frac{\ln 8}{\ln 10!} + \frac{\ln 9}{\ln 10!} + \frac{\ln 10}{\ln 10!} \\ &= \frac{\ln(7 \cdot 8 \cdot 9 \cdot 10)}{\ln 10!} = \frac{\ln 5040}{\ln 3628000} \approx 0.564 \end{aligned}$$

ja

$$P(B) = \sum_{k \in \{1,4,9\}} p_k = \frac{\ln 1}{\ln 10!} + \frac{\ln 4}{\ln 10!} + \frac{\ln 9}{\ln 10!} = \frac{\ln(1 \cdot 4 \cdot 9)}{\ln 10!} = \frac{\ln 36}{\ln 3628000} \approx 0.237.$$

Seuraavassa lauseessa osoitetaan joitain todennäköisyyslaskennan keskeisiä laskusääntöjä, jotka ovat johdettavissa suoraan todennäköisyyden aksioomista.

Lause 2.27. *Todennäköisyyslaskennan laskusääntöjä. Olkoon* $(\Omega, \mathcal{F}, P)$ *todennäköisyysavaruus ja* A *ja* B *sen tapahtumia. Tällöin*

(i) $P(\emptyset) = 0$.

(ii) (Äärellinen additiivisuus) Jos tapahtumat $A_1, \ldots, A_n$ *ovat erillisiä, eli* $A_i \cap A_j = \emptyset$ *kaikille* $i \neq j$*, niin* $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.

(iii) $P(A^C) = 1 - P(A)$. *kaikille* $A_1, \ldots, A_n \in \mathcal{F}$.

(iv) $0 \leq P(A) \leq 1$.

(v) (Monotonisuus) Jos $A \subseteq B$*, niin* $P(A) \leq P(B)$.

(vi) (Yhteenlaskukaava) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Todistus. Todistetaan kohdat (i)-(iii), ja jätetään loput harjoitustehtäväksi.

(i) Koska tyhjät joukot ovat erillisiä, todennäköisyyden additiivisuuden (TN3) nojalla tapahtumille $\emptyset, \emptyset, \ldots$ pätee

$$P(\emptyset) = P\Big(\bigcup_{i=1}^{\infty} \emptyset\Big) = P(\emptyset) + P(\emptyset) + \cdots = \sum_{i=1}^{\infty} P(\emptyset)$$

Mutta ainoa reaaliluku, joka toteuttaa tämän yhtälön on 0, joten on oltava $P(\emptyset) = 0$.

(ii) Kun $i > n$, valitaan $A_i = \emptyset$. Nyt väite seuraa täysadditiivisuudesta (TN3) erillisille tapahtumille $A_1, A_2, \ldots$ ja kohdasta (i).

(iii) Joukon komplementin määritelmän nojalla $\Omega = A \cup A^C$ ja $A \cap A^C = \emptyset$. Siten todennäköisyyden toisen aksiooman (TN2) ja äärellisen additiivisuuden (osoitettiin edellisessä kohdassa) nojalla

$$1 = P(\Omega) = P(A \cup A^C) = P(A) + P(A^C),$$

mistä väite seuraa.

$\square$

Seuraavassa taulukossa vielä muutamia todennäköisyyslaskennan kannalta keskeisiä joukko-opin merkintöjä tulkittuna todennäköisyyslaskennan termeinä. Todennäköisyyslaskennallisissa tulkinnoissa joukot A ja B voidaan tulkita tapahtumiksi, jotka joko realisoituvat eli tapahtuvat tai sitten eivät.

Joukko-opin merkintä	Tn-laskennan tulkinta
Ω	perusjoukko / varma tapahtuma
$\omega \in \Omega$	alkeistapaus / satunnaiskokeen tulosmahdollisuus
$\emptyset$	mahdoton tapahtuma
$\mathcal{F}$	kaikkien tapahtumien joukko
A	A tapahtuu
A^C	A ei tapahdu
$A \cup B$	A tai B tapahtuu
$A \cap B$	A ja B tapahtuvat
$\bigcup_{j=1}^{\infty} A_j$	ainakin yksi tapahtumista $A_1, A_2, \ldots$ tapahtuu
$\bigcap_{j=1}^{\infty} A_j$	kaikki tapahtumat $A_1, A_2, \ldots$ tapahtuvat
$A \setminus B$	A tapahtuu, mutta B ei tapahdu
$A \subseteq B$	jos A tapahtuu, niin myös B tapahtuu
$A \cap B = \emptyset$	A ja B eivät voi tapahtua yhtä aikaa

Tarkastellaan seuraavaksi muutamaa esimerkkiä, joissa sovelletaan lausetta 2.27.

Esimerkki 2.28. Opiskelija osallistuu kurssien Todennäköisyyslaskenta 1 ja Todennäköisyyslaskenta 2 kurssikokeisiin. Opiskelija läpäisee ensimmäisen kurssin todennäköisyydellä 0.8, toisen kurssin todennäköisyydellä 0.4 ja molemmat kurssit todennäköisyydellä 0.3. Millä todennäköisyydellä hän ei läpäise kumpaakaan kurssia?

Merkitään A_1 : "Opiskelija läpäisee ensimmäisen kurssin" ja A_2 : "Opiskelija läpäisee toisen kurssin". Nyt todennäköisyys, että hän läpäisee vähintään yhden kurssin, on lauseen 2.27 (vi)-kohdan nojalla

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.8 + 0.4 - 0.3 = 0.9.$$

Tällöin (De Morganin lauseiden ja) lauseen 2.27 (iii)-kohdan nojalla

$$P(A_1^C \cap A_2^C) = P((A_1 \cup A_2)^C) = 1 - P(A_1 \cup A_2) = 1 - 0.9 = 0.1.$$

Tässä luvussa olemme määritelleet todennäköisyysavaruuden käsitteitä ja luoneet puitteet, jossa voimme käsitellä aiempia esimerkkejä ja toisaalta myös myöhemmin esiteltäviä asioita. Vaikka luku on jokseenkin tekninen, käytännön laskut ja intuitio eivät ole klassisen todennäköisyyden yleistämisen myötä kadonneet mihinkään, kuten yllä olevat esimerkit osoittavat. Asiat eivät siis ole vaikeutuneet, vaan niiden käsitteistöä on täsmennetty. Täsmälliset määritelmät ja käsitteet ovat tarpeellisia, jotta todennäköisyyttä voidaan käsitellä matemaattisesti. Näitä ei kuitenkaan pidä säikähtää, eikä esimerkiksi σ-algebran käsitteen täydellinen hahmottaminen ole kurssin oppimistavoitteissa olennaista[19].

2.3 Todennäköisyyksien tulkinnasta

Olemme käsitelleet todennäköisyyttä ennen kaikkea sen matemaattisen määritelmän kautta. Tarkastellaan kuitenkin vielä muutamia käytännön tulkintoja todennäköisyydelle esimerkein havainnollistaen. Esimerkeissä 2.24 ja 2.25 alkeistapausten todennäköisyyksien määrittelyssä käytettiin apuna tapahtumien suhteellisia frekvenssejä. Tapahtuman todennäköisyys voidaan yrittää määritellä tapahtuman suhteellisen frekvenssin raja-arvona, kun satunnaiskoe toistetaan hypoteettisesti äärettömän monta kertaa. Toisin tapahtuman A todennäköisyys on

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n},$$

missä $n(A)$ on tapahtuman A toteutumiskertojen lukumäärä ja n satunnaiskokeen toistojen lukumäärä. Frekvenssitulkinta perustuu *suurten lukujen lakiin*[20], joka kuuluu todennäköisyyslaskennan merkittävimpiin tuloksiin. Suhteelliseen frekvenssiin perustuvalla tulkinnalla ei voi korvata todennäköisyyden aksioomia (ääretöntä raja-arvoa ei edes voisi tosielämässä saada kokeellisesti selville), mutta siitä on apua todennäköisyyden käsitteen hahmottamisessa.

[19] Toki huomautuksena, että σ-algebran käsitteen pyörittely ja hahmottaminen on hyvää aivojumppaa, vaikka onkin käsitteenä hyvin abstrakti.

[20] Suurten lukujen lakia käsitellään kappaleessa 7.2.1

Ainutkertaisten tapahtumien yhteydessä frekvenssitulkinta ei ole enää käyttökelpoinen[21], vaan tällöin voidaan puhua muun muassa subjektiivisesta eli bayesilaisesta todennäköisyydestä. Subjektiivinen todennäköisyys kuvaa sitä, kuinka uskottavana henkilö tai jokin muu taho jonkun tapahtuman toteutumista pitää. Eri tahojen käsitykset saman tapahtuman todennäköisyydestä voivat vaihdella ja subjektiivinen todennäköisyys voi myös muuttua henkilön saadessa lisää tietoa tapahtumasta. Subjektiivisia todennäköisyyksiä käytetään usein vaikeasti ennakoitavissa olevien tapahtumien, kuten urheilu- tai laulukilpailuiden yhteydessä. Näissä tapahtumissa voittajasta voi lyödä vetoa joillakin kertoimilla, vaikka kyseessä on yleensä ainutkertainen tapahtuma, jonka lopputulosten todennäköisyyksiä ei olla voitu empiirisesti tutkia. Tarkastellaan kuitenkin rajatumpaa esimerkkiä tilanteesta, jossa subjektiivinen todennäköisyys voi muuttua.

Esimerkki 2.29. *Painotettu noppa*: Hannulla ja Kertulla on uusi noppa. Kerttu kysyy Hannulta, millä todennäköisyydellä seuraavan heiton silmäluku on 6. Hannulla ei ole ollut mahdollisuutta tutkia noppaa, mutta hän on lukenut klassista todennäköisyyttä ja vastaa näin ollen $\frac{1}{6}$. Siis Hannun subjektiivisen käsityksen mukaan $P(6) = \frac{1}{6}$.

Kerttu on kuitenkin Hannun tietämättä heittänyt noppaa 100 000 kertaa ja havainnut, että noppa onkin painotettu. Kertun havaintojen mukaan todennäköisyydet saada 5 tai 6 ovat kumpikin $\frac{1}{3}$, ja kaikkien muiden silmälukujen todennäköisyydet ovat kukin $\frac{1}{12}$.

Kerttu päättää jakaa tietonsa Hannun kanssa ja kysyy tämän jälkeen uudestaan, millä todennäköisyydellä seuraavan heiton silmäluku on 6. Nyt Hannu on saanut uutta tietoa ja vastaa luonnollisesti $\frac{1}{3}$. Näin ollen Hannun subjektiivinen todennäköisyys silmäluvulle 6 onkin nyt $P(6) = \frac{1}{3}$. Hannun arvio kutosen todennäköisyydestä muuttui huomattavasti, vaikka itse noppa ei muuttunut mitenkään.

On hyvä muistaa, että frekvenssitulkinta tai subjektiivinen todennäköisyys eivät ole kaikissa tilanteissa riittäviä matemaattisia määritelmiä todennäköisyydelle. Jotta voidaan täsmällisesti puhua todennäköisyyksistä, täytyy frekvenssitulkintaan perustuvan todennäköisyyden tai subjektiivisen todennäköisyyden toteuttaa todennäköisyyden aksioomat (kuten muun muassa aiemmin esitettyjen esimerkkien todennäköisyydet tekevät).

Esimerkki 2.30. Säätä ennustavalta etanalta kysytään seuraavien päivien säästä. Perinteisen "onko huomenna pouta" -kysymyksen sijaan etanalle esitetään kysymykset neljän tapahtuman todennäköisyyksistä, joihin se antaa seuraavat vastaukset:

- A : Huomenna on pouta —0.5
- B : Ylihuomenna on pouta —0.6
- C : Molempina päivinä on pouta —0.4

[21] Aidon frekvenssin arviointi yhden tapahtuman perusteella on mahdotonta.

- D : Ainakin toisena päivänä on pouta —0.8

Ovatko etanan antamat subjektiiviset todennäköisyydet todella todennäköisyyksiä, eli noudattavatko ne todennäköisyyden aksioomia?

Huomataan, että $A \cap B = C$ ja $A \cup B = D$, jolloin lauseen 2.27 (vi)-kohdan nojalla pitäisi päteä

$$\begin{aligned}
&P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \Leftrightarrow \\
&P(D) = P(A) + P(B) - P(C) \quad \Leftrightarrow \\
&0.8 = 0.5 + 0.6 - 0.4 = 0.7,
\end{aligned}$$

mikä ei selvästi pidä paikkaansa. Näin ollen etanan subjektiiviset todennäköisyydet eivät noudata todennäköisyyden aksioomia.

Huomaamme siis, että käsitteen todennäköisyys voi tulkita monella eri tavalla, ja kuten monissa muissa asioissa, yksi tulkintatapa ei ole yksiselitteisesti ylivertainen muihin tulkintatapoihin nähden. Nämä yrittävät selittää ja mallintaa reaalimaailman ilmiöitä ja ovat siten väistämättä jollain tavalla puutteellisia[22].

2.4 Harjoitustehtäviä

1. Puinen kuutio, jonka sivutahkot on maalattu, sahataan 1000 yhtä suureksi pikkukuutioksi. Kuutiot sekoitetaan ja niistä valitaan umpimähkään yksi. Mikä on todennäköisyys, että siinä on täsmälleen k maalattua tahkoa ($k = 0, 1, 2, 3$)?

2. Kolikkoa heitetään 3 kertaa. Mikä on todennäköisyys, että saadaan

   (a) tasan 1 kruuna,

   (b) tasan 3 kruunaa,

   (c) enintään 2 kruunaa,

   (d) pariton määrä kruunia?

3. Sekoitetusta korttipakasta nostetaan 5 korttia. Millä todennäköisyydellä nostetuista korteista 2 on kuvakortteja?

4. Korttipakka jaetaan kortti kerrallaan pöydälle. Millä todennäköisyydellä

   (a) 14. jaettu kortti on ässä,

[22] Tilastotieteilijä George Boxin kuuluisa lainaus ”all models are wrong but some are useful” kuvaa tilannetta hyvin.

(b) ensimmäinen ässä tulee 14. kortin kohdalla?

5. (Ross Exercise 2.5) Korttipakasta vedetään neljä korttia. Millä todennäköisyydellä kortit ovat

   (a) eri maita,

   (b) eri arvoja,

   (c) sekä eri maita, että eri arvoja?

6. Laatikossa on 3 valkoista ja 5 mustaa palloa. Laatikosta nostetaan kolme palloa takaisinpanolla. Millä todennäköisyydellä nostettujen pallojen joukossa on sekä mustia että valkoisia palloja?

7. Edellisen tehtävän laatikosta poimitaan palloja ilman takaisinpanoa. Millä todennäköisyydellä kaikki valkoiset pallot poimitaan ennen ensimmäistä mustaa palloa?

8. Heitetään yhtä tikkaa tavalliseen tauluun, jossa on 10 sisäkkäistä rengasta, joista saa 1-10 pistettä. 1 pisteen alueelle eli uloimpaan renkaaseen osuminen on 2 kertaa todennäköisempää kuin 2 pisteen alueelle (eli seuraavaksi sisempään renkaaseen) osuminen, 3 kertaa todennäköisempää kuin 3 pisteen alueelle osuminen, ..., ja 10 kertaa todennäköisempää kuin 10 pisteen alueelle eli 'häränsilmään' osuminen. Jos oletetaan, että olemme niin taitavia heittäjiä, että osumme varmasti tauluun, millä todennäköisyydellä saamme vähintään 9 pistettä?

9. Olkoot $P(A) = 0.6$, $P(B) = 0.4$ ja $P(A \cap B) = 0.2$. Määritä seuraavien tapahtumien todennäköisyydet:

   (a) $A \cup B$

   (b) A^C

   (c) $A \cap B^C$

   (d) $A \cup B^C$

   (e) $A^C \setminus B^C$

10. Oletetaan, että $P(A) = 0.45$ ja $P(B) = 0.75$, mitä voit saoa luvusta $P(A \cap B)$?

Luku 3

Ehdollinen todennäköisyys ja riippumattomuus

Edellisessä luvussa määrittelimme todennäköisyyden matemaattisena käsitteenä. Nyt etenemme tarkastelemaan useita tapahtumia ja tapahtumien vaikutusta toisiinsa. Tärkeitä tässä luvussa esiteltyjä käsitteitä ovat *ehdollinen todennäköisyys* ja *riippumattomuus*. Meitä saattaa esimerkiksi kiinnostaa, millä todennäköisyydellä suosikkijoukkueemme etenee kaksiosaisista pudotuspeleistä jatkoon, kun tiedämme, että ensimmäinen ottelu päättyi tasan. Tai saatamme haluta arvioida sateen todennäköisyyttä aamupäivän aikana. Tällöin arvioomme saattaa vaikuttaa se, että huomaamme ulkona olevan synkän pilvistä.

Molemmissa edellä mainituista tilanteista on kyse ehdollisesta todennäköisyydestä. Haluamme laskea todennäköisyyden suosikkijoukkueemme jatkoonpääsylle **ehdolla** että ensimmäinen ottelu päättyi tasan, ja todennäköisyyden sateelle **ehdolla** että sää on pilvinen. Riippumattomuuden käsite ja ehdollinen todennäköisyys kietoutuvat voimakkaasti yhteen. Sääesimerkissä pilvinen sää epäilemättä vaikuttaa sateen todennäköisyyteen, joten pilvisyys ja sade eivät ole toisistaan riippumattomia tapahtumia. Sen sijaan edellisillan ottelun lopputuloksen ei voida ajatella vaikuttavan tämän päivän sateen mahdollisuuteen, joten näitä tapahtumia voidaan sanoa riippumattomiksi.

3.1 Ehdollinen todennäköisyys

Määritellään aluksi todennäköisyyslaskennassa monesti hyödyllinen ehdollisen todennäköisyyden käsite. Ehdollista todennäköisyyttä tarvitaan, kun haluamme laskea todennäköisyyttä tilanteessa, jossa meillä jo on jotain osittaista tietoa tapahtuman lopputuloksesta. Toisaalta ehdollisen todennäköisyyden avulla on usein helpompaa laskea haluttuja todennäköisyyksiä myös, vaikka mitään osittaista tietoa ei edes olisi tarjolla.

Määritelmä 3.1. *Ehdollinen todennäköisyys.*[1] Olkoot A ja B tapahtumia siten että $P(B) \geq 0$. Tapahtuman A *todennäköisyys ehdolla* B (eli tapahtuman A todennäköisyys, kun tiedetään, että B on tapahtunut) on

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

Voidaan myös osoittaa, että ehdollinen todennäköisyys $P(A|B)$ toteuttaa todennäköisyyden aksioomat (2.15). Tämä mahdollistaa kaikkien todennäköisyyden laskusääntöjen soveltamisen myös ehdolliselle todennäköisyydelle.

Tarkastellaan seuraavaksi ehdollista todennäköisyyttä esimerkkien kautta.

Esimerkki 3.2. Kahden nopan heitto. Valitaan perusjoukoksi jälleen $\Omega = \{1, \ldots, 6\}^2$, ja todennäköisyydeksi

$$P(\omega) = \frac{1}{|\Omega|} = \frac{1}{36} \quad \text{kaikille} \quad \omega \in \Omega,$$

jolloin (Ω, P) on symmetrinen tn-avaruus.

(i) Mikä on tn tapahtumalle A: ”Silmälukujen summa on vähintään 8”?

Laskemalla tapahtumalle A suotuisat alkeistapaukset saadaan

$$P(A) = \frac{15}{36}$$

(ii) Mikä on tapahtuman A todennäköisyys ehdolla B : ”Ensimmäisen nopan silmäluku on kuusi”?

Yhdistelmät, joilla kumpikin tapahtuma toteutuu, ovat $(6, 2), (6, 3), (6, 4), (6, 5)$ ja $(6, 6)$, joten

$$P(A \cap B) = \frac{5}{36}.$$

Siten ehdollisen todennäköisyyden kaavasta saadaan

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{5}{36}}{\frac{1}{6}} = \frac{5}{6}.$$

Tieto siitä, että ensimmäisen nopan silmäluku on kuusi, muutti siis huomattavasti arviotamme tapahtuman A todennäköisyydestä.

[1] engl. *conditional probability*

Kirjoittamalla ehdollisen todennäköisyyden määritelmä tulomuodossa saadaan *kertolaskusääntö*, jota kutsutaan myös todennäköisyyden ketjusäännöksi.

Lause 3.3. *Kertolaskusääntö.*[2] *Olkoot A ja B tapahtumia ja $P(B) > 0$. Nyt*

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

Todistus. Suoraan ehdollisen tn:n määritelmästä saadaan:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ P(A \cap B) &= P(A|B)P(B). \end{aligned}$$

□

Kertolaskusääntö on hyödyllinen tapahtumien leikkausten todennäköisyyksien laskennassa. Itse asiassa ehdollinen todennäköisyys esiintyy sovelluksissa yleensä juuri kertolaskusäännön kautta. Kertolaskulaskusääntö yleistyy suoraan myös useammalle tapahtumalle.

Lause 3.4. *Yleinen ketjusääntö. Olkoot $A_1, \ldots, A_n$ tapahtumia siten, että $P(A_1 \cap \cdots \cap A_{n-1}) \geq 0$. Nyt*

$$\begin{aligned} P(A_1 \cap \cdots \cap A_n) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}) \\ &= P(A_1) \prod_{i=2}^{n} P(A_i| \cap_{j=1}^{i-1} A_j). \end{aligned}$$

Todistus. Induktiolla kertolaskusäännöstä. □

3.2 Kokonaistodennäköisyys ja Bayesin kaava

Joissakin tilanteissa satunnaiskoe voi saavuttaa saman lopputuloksen edeten useaa eri reittiä. Tällöin puhutaan *kokonaistodennäköisyydestä.* Tarkastellaan ensin havainnollistavaa esimerkkiä.

Esimerkki 3.5. Tarkastellaan satunnaiskoetta, jossa ensin valitaan satunnaisesti yksi kolmesta laatikosta, joissa on valkoisia ja mustia palloja seuraavan taulukon mukaisesti.

[2] engl. *multiplication rule*

Laatikko	Valkoiset	Mustat
1	1	1
2	2	1
3	3	1

Valitaan sen jälkeen ensimmäisessä vaiheessa valitusta laatikosta satunnaisesti yksi pallo. Mikä on tapahtuman A : ”valittu pallo on valkoinen” todennäköisyys?

Tätä satunnaiskoetta ei voi mallintaa symmetristen alkeistapauksien avulla, sillä eri pallojen todennäköisyydet poikkeavat toisistaan. Monissa tilanteissa tapahtumien ehdolliset todennäköisyydet on helpompi päätellä kuin ehdottomat todennäköisyydet. Niin tässäkin: kun merkitsemme B_i : ”valitaan i:s laatikko” kaikille $i \in \{1, 2, 3\}$, tapahtuman A ehdolliset todennäköisyydet ehdolla B_i ovat:

$$P(A|B_1) = \frac{1}{2}, \quad P(A|B_2) = \frac{2}{3}, \quad P(A|B_3) = \frac{3}{4}.$$

Koska laatikko valitaan satunnaisesti, $P(B_i) = \frac{1}{3}$ kaikille $i \in 1, 2, 3$.

Valkoinen pallo voidaan nostaa kolmesta eri laatikosta, joten tapahtuma A voidaan jakaa erillisiin toisensa poisulkeviin osiin $A \cap B_1$, $A \cap B_2$ ja $A \cap B_3$. Näiden todennäköisyydet saadaan kertolaskusäännöstä, ja koska kyseiset tapahtumat ovat erillisiä, A:n todennäköisyys saadaan todennäköisyyden additiviisuuden nojalla:

$$\begin{aligned} P(A) &= P\left(\bigcup_{i=1}^{3}(A \cap B_i)\right) = \sum_{i=1}^{3} P(A \cap B_i) \\ &= \sum_{i=1}^{3} P(B_i)P(A|B_i) = \left(\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{3}{4}\right) = \frac{23}{36}. \end{aligned}$$

Osoittautuu, että vastaavaa päättelyä voidaan soveltaa yleisemminkin: yleistä tapausta kutsutaan *kokonaistodennäköisyyden kaavaksi*. Määritellään ensin tarkemmin mitä tarkoittaa joukon, eli tapahtuman, *ositus*.

Määritelmä 3.6. *Ositus.*[3] Joukot $B_1, \dots, B_n$ osittavat joukon X, eli muodostavat joukon X *osituksen*, jos ne toteuttavat seuraavat ehdot:

(i) Ne ovat erillisiä:

$$B_i \cap B_j = \emptyset \quad \text{kaikille} \quad i \neq j.$$

(ii) Niiden yhdiste on koko tarkasteltava joukko

$$\bigcup_{i=1}^{n} B_i = X.$$

[3] engl. *partition*

Todennäköisyyslaskennassa tärkeä erityistapaus on $X = \Omega$, jolloin sanotaan, että joukot $B_1, \ldots, B_n$ ovat perusjoukon ositus.

Seuraavaksi voimme esittää kokonaistodennäköisyyden kaavan.

Lause 3.7. *Kokonaistodennäköisyyden kaava.*[4] *Olkoon* $(\Omega, \mathcal{F}, P)$ *tn-avaruus,* $A \in \mathcal{F}$ *tapahtuma, ja* $B_1, \ldots, B_n \in \mathcal{F}$ *perusjoukon* Ω *ositus. Nyt*

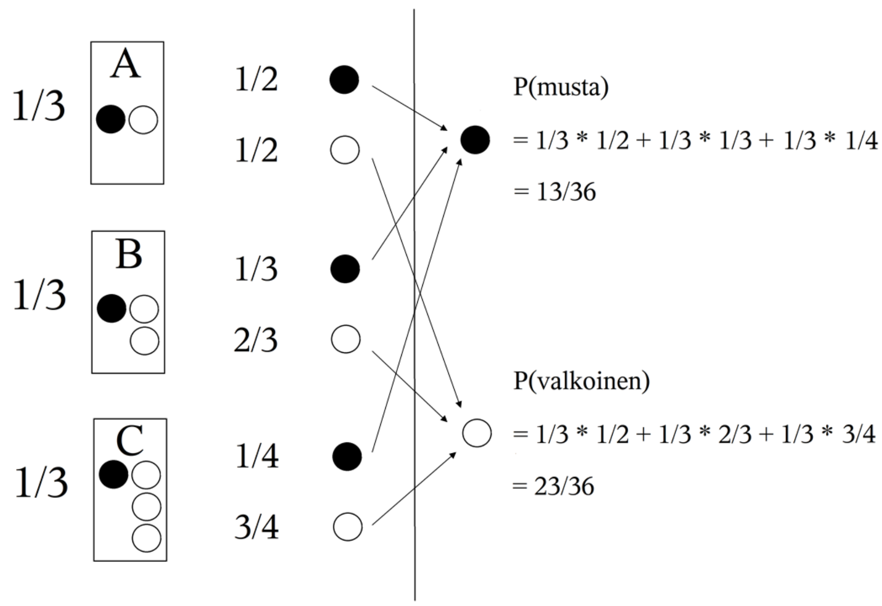
$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

Todistus. Joukot $(A \cap B_1), \ldots, (A \cap B_n)$ muodostavat joukon A osituksen, joten todennäköisyyden additiivisuuden ja kertolaskusäännön nojalla

$$P(A) = P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A|B_i).$$

$\square$

Kuva 3.1 havainnollistaa kokonaistodennäköisyyden kaavaa esimerkin 3.5 laatikoilla ja palloilla. Mustan ja valkoisen pallon kokonaistodennäköisyydet muodostuvat niihin johtavien ehdollisten osapolkujen todennäköisyyksien summana.



Kuva 3.1: Kokonaistodennäköisyyden kaava.

[4] engl. *total probability rule*

Katsotaan esimerkin 3.5 tilannetta seuraavaksi hieman eri näkökulmasta.

Esimerkki 3.8. Jatkoa esimerkkiin 3.5. Oletetaan, että koe suoritetaan verhon takana, eli emme näe mistä laatikosta pallo on nostettu, vaan pelkästään nostetun pallon värin. Pallo on valkoinen, eli havaitsemme tapahtuman A. Mikä on nyt tapahtuman B_1 : "Valittiin ensimmäinen laatikko" todennäköisyys ehdolla A, eli mikä on todennäköisyys sille, että pallo nostettiin laatikosta 1, kun tiedetään että nostettu pallo on valkoinen?

Ehdollisen todennäköisyyden määritelmää ja kertolaskusääntöä soveltamalla saadaan:

$$P(B_1|A) = \frac{P(B_1 \cap A)}{P(A)} = \frac{P(B_1)P(A|B_1)}{P(A)} = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{23}{36}} = \frac{6}{23}.$$

Vastaavasti voidaan laskea todennäköisyydet toiselle ja kolmannelle laatikolle:

$$P(B_2|A) = \frac{\frac{1}{3} \cdot \frac{2}{3}}{\frac{23}{36}} = \frac{8}{23}$$
$$P(B_3|A) = \frac{\frac{1}{3} \cdot \frac{3}{4}}{\frac{23}{36}} = \frac{9}{23}.$$

Tämä päättely on esimerkki todennäköisyyden keskeisen laskusäännön, *Bayesin kaavan* käytöstä. Bayesin kaavaa voidaan soveltaa tilanteessa, jossa olemme havainneet tapahtuman A, ja tiedämme tapahtuman A ehdolliset todennäköisyydet jonkin perusjoukon osituksen suhteen:

$$P(A|B_i) \quad \text{kaikille} \quad i \in \{1, \dots, n\}.$$

Tällöin Bayesin kaavan avulla saadaan laskettua tapahtumien $B_1, \dots, B_n$ ehdolliset todennäköisyydet ehdolla A:

$$P(B_i|A) \quad \text{kaikille} \quad i \in \{1, \dots, n\}.$$

Bayesin kaava on nimetty brittiläisen matemaatikon Thomas Bayesin mukaan ja sen saa johdettua suoraan kokonaistodennäköisyyden kaavasta 3.7. Bayesin kaavaan perustuu toinen tilastotieteen merkittävistä suuntauksista, *bayesiläinen tilastotiede.*

Lause 3.9. *Bayesin kaava.*[5] *Olkoon A tapahtuma, jolle $P(A) \geq 0$, ja $B_1, \dots, B_n$ perusjoukon ositus. Jokaisen tapahtuman B_i ehdollinen todennäköisyys ehdolla A saadaan Bayesin kaavasta:*

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

[5]engl. *Bayes' rule*

Todistus. Käyttämällä ehdollisen todennäköisyyden määritelmää ja kertolaskusääntöä saadaan

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

Jälkimmäinen muotoilu saadaan soveltamalla kokonaistodennäköisyyden kaavaa nimittäjään. □

Tarkastellaan vielä klassista Bayesin kaavan sovellusta.

Esimerkki 3.10. Oletetaan, että harvinainen sairaus esiintyy 0.1 prosentilla väestöstä, ja että meillä on erittäin tarkka testi sen toteamiseksi. Testin sensitiivisyys, eli positiivisen testituloksen todennäköisyys henkilölle, jolla on tutkittava sairaus, on 0.95. Testin spesifisyys, eli positiivisen testituloksen tn terveelle henkilölle on 0.01. Mikä on todennäköisyys, että testattavalla (satunnaisesti väestöstä valitulla) henkilöllä on tutkittava sairaus, jos testin tulos positiivinen?

Merkitään tapahtumia S : "henkilöllä on tutkittava sairaus", T : "henkilön testitulos on positiivinen". Tehtävänannosta saadaan seuraavat todennäköisyydet:

$$\begin{aligned} P(S) &= 0.001 \\ P(S^C) &= 0.999 \\ P(T|S) &= 0.95 \\ P(T|S^C) &= 0.01. \end{aligned}$$

Tapahtumat S ja S^C osittavat perusjoukon, joten todennäköisyys, että henkilö on sairas ehdolla että testitulos on positiivinen, saadaan Bayesin kaavasta:

$$\begin{aligned} P(S|T) &= \frac{P(S)P(T|S)}{P(S)P(T|S) + P(S^C)P(T|S^C)} \\ &= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.01} \approx 0.087. \end{aligned}$$

Todennäköisyys, että positiivisen testituloksen saanut henkilö on oikeasti sairas, saatiin siis tutkimalla, mikä on odotettu oikeiden positiivisten (positiivinen testitulos sairaalle) osuus kaikista positiivisista testituloksista, eli yhteenlasketuista oikeista positiivista ja vääristä positiivista (positiivinen testitulos terveelle) testituloksista.

Sairaiden pieni odotettu osuus positiivisen testituloksen saaneista selittyy tietenkin sairaiden pienellä osuudella väestöstä: testin tarkkuudesta huolimatta on huomattavasti todennäköisempää, että satunnaisesti väestöstä valittu testihenkilö on terve **ja** saa positiivisen testituloksen, kuin että henkilö on sairas **ja** saa positiivisen testituloksen.

3.3 Tapahtumien riippumattomuus

Intuitiivisesti lienee selvää, mitä *riippumattomilla* tapahtumilla tarkoitetaan. Muistellaan esimerkiksi luvun alussa mainittua esimerkkiä säästä ja urheiluottelun lopputuloksesta. Kahden tapahtuman sanotaan olevan riippumattomia, jos yhden tapahtuminen ei vaikuta toisen tapahtumiseen[6]. Esimerkkejä tällaisista tapahtumista voi keksiä lukemattomia: peräkkäisten arvontojen lottorivit, vuoden keskilämpötila ja jääkiekon pääsarjatason lopputulokset, yksittäisen viljelijän perunasato ja Suomessa syntyneiden lasten määrä, jne. Riippumattomuuden oletusta käytetään hyvin yleisesti todennäköisyyslaskennan sovelluksissa esimerkiksi tilastotieteessä, joten sen käsite on tärkeä ymmärtää. Oletuksissa on myös syytä olla tarkkana, sillä kaikki asiat eivät välttämättä ole riippumattomia, vaikka voivat siltä vaikuttaa. Lukija voi esimerkiksi yrittää perustella itselleen, ovatko luvun alussa mainittujen kaksiosaisten pudotuspelien tulokset riippumattomia toisistaan.

Määritelmä 3.11. *(Kahden tapahtuman) riippumattomuus.*[7] Tapahtumia A ja B kutsutaan riippumattomiksi, jos
$$P(A \cap B) = P(A)P(B).$$
Tällöin merkitään $A \perp\!\!\!\perp B$.

Tapahtumien riippumattomuus voidaan yhtäpitävästi esittää myös ehdollisen todennäköisyyden avulla, kuten seuraavasta lauseesta nähdään. Tapahtuma A on siis riippumaton tapahtumasta B, jos ehdollistaminen B:llä ei muuta sen todennäköisyyttä; toisin sanoen tieto siitä, että B on tapahtunut, ei muuta arviotamme A:n todennäköisyydestä.

Lause 3.12. *Olkoot tapahtumat A ja B siten että $P(B) \geq 0$. Nyt $A \perp\!\!\!\perp B$ jos ja vain jos $P(A|B) = P(A)$.*

Todistus. "$\Rightarrow$": Ehdollisen tn:n määritelmästä saadaan tapahtumien A ja B riippumattomuuden nojalla:
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

"$\Leftarrow$": Kertolaskusäännön ja oletuksen nojalla:
$$P(A \cap B) = P(A|B)P(B) = P(A)P(B).$$
$\square$

[6] tapahtumisen todennäköisyyteen
[7] engl. *independence*

Huomautus 3.13. Tässä vaiheessa lukijan olisi syytä tarkistaa, että ymmärtää, mitä eroa on tapahtumien *riippumattomuudella* ja *erillisyydellä*[8].

Havainnollistetaan seuraavaksi riippumattomuuden käsitettä esimerkeillä.

Esimerkki 3.14. Esimerkki riippumattomista tapahtumista ja tapahtumista, jotka eivät ole riippumattomia.

(i) Heitetään kahta noppaa, jolloin perusjoukko on jälleen $\Omega = \{1, 2, \dots, 6\}^2$. Tapahtumien A : "Ensimmäisen nopan silmäluku on 6" ja tapahtuman B : "toisen nopan silmäluku on kuusi" todennäköisyydet ovat $P(A) = \frac{1}{6}$ ja $P(B) = \frac{1}{6}$. Toisaalta vain alkeistapaus $(6, 6)$ toteuttaa kummankin näistä tapahtumista, joten

$$P(A \cap B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(A)P(B).$$

Tapahtumat A ja B ovat siis riippumattomia eli ensimmäisen nopan silmäluku ei vaikuta toisen nopan silmälukuun.

(ii) Nostetaan kaksi korttia pakasta palauttamatta ensimmäiseksi nostettua korttia välillä takaisin pakkaan. Perusjoukoksi voidaan valita korttipakan 2-permutaatiot, eli kahden alkion järjestetyt jonot. Sekä tapahtuman A : "Ensimmäiseksi nostettu kortti on ässä", että tapahtuman B : "Toiseksi nostettu kortti on ässä" tn on perusjoukon symmetrian nojalla

$$P(A) = \frac{4 \cdot 51}{52 \cdot 51} = \frac{1}{13} = \frac{51 \cdot 4}{52 \cdot 51} = P(B).$$

Mahdollisia yhdistelmiä, joissa sekä ensimmäinen, että toinen nostettu kortti ovat ässiä, on $4 \cdot 3$, joten tapahtumien A ja B leikkauksen todennäköisyys on

$$P(A \cap B) = \frac{4 \cdot 3}{52 \cdot 51} = \frac{1}{221} \neq \frac{1}{169} = P(A)P(B),$$

joten tapahtumat A ja B eivät ole riippumattomia.

Nähdään myös, että todennäköisyys nostaa toisella kortilla ässä, jos ensimmäinen nostettu kortti oli ässä on

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{221}}{\frac{1}{13}} = \frac{1}{17} \neq \frac{1}{13} = P(B).$$

Siis tieto siitä, onko ensimmäisellä kortilla nostettu ässä, vaikuttaa todennäköisyyteen, että toisella kortilla nostetaan ässä.

[8] mm. luvun 2.2 taulukko joukko-opin merkintöjen tulkinnasta tn-laskennassa

Seuraava tulos on intuitiivinen seuraus riippumattomuuden määritelmästä.

Lause 3.15. *Komplementin riippumattomuus. Jos tapahtumat A ja B ovat riippumattomia, myös tapahtumat A ja B^C ovat riippumattomia.*

Todistus. Oletetaan, että $A \perp\!\!\!\perp B$. Koska $A = (A \cap B) \cup (A \cap B^C)$, ja joukot $A \cap B^C$ ja $A \cap B$ ovat selvästi erillisiä,

$$
\begin{aligned}
&P(A) = P(A \cap B) + P(A \cap B^C) = P(A)P(B) + P(A \cap B^C) \quad \Leftrightarrow \\
&P(A \cap B^C) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(B^C),
\end{aligned}
$$

eli $A \perp\!\!\!\perp B^C$. □

Riippumattomuuden määritelmä yleistyy myös useammalle kuin kahdelle tapahtumalle. Tällöin vaaditaan, että kaikkien mahdollisten tarkasteltavien tapahtumien keskinäisten leikkausten todennäköisyydet saadaan niiden todennäköisyyksien tulona.

Määritelmä 3.16. *Usean tapahtuman riippumattomuus.* Tapahtumia $A_1, \ldots, A_n$ kutsutaan riippumattomiksi, jos

$$
P\Big(\bigcap_{i \in I} A_i\Big) = \prod_{i \in I} P(A_i).
$$

kaikille indeksijoukoille $I \subseteq \{1, \ldots, n\}$. Tällöin merkitään $A_1, \ldots, A_n \perp\!\!\!\perp$.

Esimerkiksi kolmelle tapahtumalle A, B ja C riippumattomuus tarkoittaa, että tapahtumat ovat pareittain riippumattomia, eli että

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

ja lisäksi, että

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

Huomautus 3.17. Useamman kuin kahden tapahtuman riippumattomuuteen vaaditaan siis, että **kaikki** yllä mainitut riippumattomuudet toteutuvat. Siitä, että tapahtumat A, B ja C ovat pareittan riippumattomia toisistaan, ei välttämättä seuraa, että $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Esimerkki 3.18. Heitetään kahta noppaa. Merkitään

$$\begin{aligned}&A\text{: ”noppien silmälukujen summa on 7”},\\ &B\text{: ”ensimmäisen nopan silmäluku on 3”},\\ &C\text{: ”toisen nopan silmäluku on 4”}.\end{aligned}$$

Selvästikin esimerkin 3.14 perusteella tapahtumat B ja C ovat riippumattomia. Tarkastellaan nyt tapahtumaa A. Jos ensimmäisen nopan silmäluku on $k \in \{1, \dots, 6\}$, niin toisella nopanheitolla tulee saada $7 - k$, jotta noppien silmälukujen summa olisi 7. Huomataan, että $7 - k \in \{1, \dots, 6\}$, jolloin riippumatta ensimmäisen nopan silmäluvusta, toisella nopalla on mahdollista saada sellainen silmäluku, että silmälukujen summa on 7 ja tämän todennäköisyys on $\frac{1}{6}$. Tämän voi tarkistaa mm. esimerkin 4.10 (ii)-kohdan taulukosta. Siis $P(A) = \frac{1}{6}$. Nyt

$$P(A \cap B) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(A)P(B),$$

eli tapahtumat A ja B ovat riippumattomia. Samoin voidaan osoittaa, että tapahtumat A ja C ovat riippumattomia. Nyt kuitenkin huomataan että tapahtuma $A \cap B$ voitaisiin muotoilla myös: ”ensimmäisen nopan silmäluku on 3 ja toisen 4”, jolloin siis $A \cap B = B \cap C$, mistä seuraa, että

$$P(A \cap B \cap C) = P((A \cap B) \cap C) = P((B \cap C) \cap C) = P(B \cap C) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

mutta toisaalta

$$P(A)P(B)P(C) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

eli

$$P(A \cap B \cap C) \neq P(A)P(B)P(C)$$

jolloin tapahtumat A, B ja C eivät ole riippumattomia.

3.4 Toistokokeet

Joskus on mielekästä tarkastella tilannetta, jossa haluamme toistaa tiettyä satunnaiskoetta n kertaa. Rajoitamme tilanteen siten, että tarkastelemme vain sellaisten satunnaiskokeiden toistoa, jotka ovat riippumattomia toisistaan. Lisäksi kiinnitämme yksinkertaisuuden vuoksi huomiota yksittäisessä kokeessa vain siihen, sattuuko joku tietty tapahtuma A vai ei. Jos merkitsemme $P(A) = p$, $P(A^c) = 1 - p = q$, ja edelleen jos

$$B_k = \text{”}A \text{ tapahtuu vain ja ainoastaan } k \text{ kertaa”},$$

voimme laskea todennäköisyyden $P(B_k)$.

Ennen toistokokeen määrittelyä, tarkastellaan yksinkertaista esimerkkiä.

Esimerkki 3.19. Tarkastellaan tilannetta, jossa haluamme toistaa yhden nopan heittoa 5 kertaa ja olemme kiinnostuneita silmäluvun 6 ilmentymistä. Nyt $P(A) =$"kutonen", ja $B_3 =$"viidestä heitosta saamme 3 kutosta". Lisäksi $p = 1/6$ ja $q = 5/6$. 5-kertaista toistokoetta kuvailevan todennäköisyysavaruuden alkeistapauksiksi valitsemme kaikki jonot $(\omega_1, \omega_2, \dots, \omega_5)$, missä $\omega_i \in \{A, A^c\}$ kaikilla $i \in \{1, 2, 3, 4, 5\}$.

Koska erilliset toistot ovat toisistaan riippumattomia, kunkin sellaisen jonon todennäköisyys, missä saamme kutosia 3 kappaletta ja ei-kutosia 2 kappaletta, on

$$p^3 q^2 = \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2.$$

Tämän lisäksi tällaisia jonoja (eli mahdollisia tapahtumajärjestyksiä, missä saamme 3 kutosta) on yhteensä $\binom{5}{3}$ kappaletta. Edelleen tuloperiaatteen nojalla

$$P(B_3) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 \approx 0.032.$$

Yleisesti toistokoe ja siihen liittyvä todennäköisyys $P(B_k)$ käyttäytyy seuraavasti:

Lause 3.20. *Toistokoe.* [9] *Merkitään $P(A) = p$. Tällöin n-kertaisessa toistokokeessa tapahtuman $B_k =$ "A tapahtuu tasan k kertaa" todennäköisyys on*

$$P(B_k) = \binom{n}{k} p^k q^{n-k},$$

missä $k = 0, 1, \dots, n$ ja $q = 1 - p$.

Todistus. Todistus kuten esimerkissä 3.19. □

Muistellaan tässä välissä otantaa 2.1.1. Huomataan, että otannassa takaisinpanolla suoritetaan peräkkäisiä, toisistaan riippumattomia satunnaiskokeita. Näin ollen toistokoe soveltuu todennäköisyyksien laskemiseen otanta takaisinpanolla -tyylisissä tilanteissa.

Esimerkki 3.21. Jatkoa esimerkin 2.10 (*ii*)-kohdalle. Otanta takaisinpanolla -tyylinen tilanne voidaan tulkita n-kertaiseksi toistokokeeksi, missä $A =$"satunnaisesti valittu pallo on valkoinen", ja edelleen $p = K/N$ lauseen 2.11 merkinnöin.

Laatikossa on 5 valkoista ja 3 mustaa palloa. Nostetaan laatikosta 4 palloa palauttaen nostojen välillä pallot takaisin laatikkoon. Millä todennäköisyydellä nostetaan tasan 2 valkoista palloa?

[9] engl. *Bernoulli trial / binomial trial*

Merkitään B_k : ”nostetaan yhteensä k valkoista palloa.” Nyt $p = \frac{5}{8}$ ja lauseen 3.20 nojalla

$$P(B_2) = \binom{4}{2}\left(\frac{5}{8}\right)^2\left(\frac{3}{8}\right)^{4-2} \approx 0.3296,$$

mikä on sama vastaus, kuin minkä esimerkissä 2.10 ratkaisimme.

Toistokokeita voidaan mallintaa *binomijakaumalla*, johon tutustutaan tarkemmin kappaleessa 4.1.3.

3.5 Harjoitustehtäviä

1. Heitetään kahta noppaa. Mikä on todennäköisyys, että vähintään toinen saatu silmäluku on 6 ehdolla, että saadaan eri silmäluvut?

2. Oletetaan, että $A \subset B$. Ilmaise seuraavat tapahtumat mahdollisimman yksinkertaisesti:

   a) $P(A|B)$
   b) $P(A|B^C)$
   c) $P(B|A)$
   d) $P(B|A^C)$

3. Laatikossa on 6 punaista ja 9 valkoista palloa. Kokeessa laatikosta nostetaan 3 palloa ilman takaisinpanoa, laske seuraavat todennäköisyydet:

   a) kaikki ovat punaisia ehdolla, että ainakin yksi on punainen,
   b) kaikki ovat punaisia ehdolla, että kaikki ovat samanvärisiä,
   c) saadaan 1 punainen ja 2 valkoista ehdolla, että kaikki pallot eivät ole samanvärisiä.

4. Heitetään kahta noppaa. Olkoon A: ”silmälukujen summa on 6” ja B: ”1. nopan silmäluku on 4”. Onko $A \perp\!\!\!\perp B$? Perustele käyttäen riippumattomuuden määritelmää.

5. Heitetään reilua kolikkoa kaksi kertaa. Määritellään tapahtumat A, B ja C seuraavasti:

   - A: Ensimmäisellä heitolla saadaan kruuna
   - B: Toisella heitolla saadaan kruuna

- C: Molemmilla heitoilla saadaan sama puoli

Osoita, että A, B ja C ovat pareittain riippumattomia, mutta eivät riippumattomia. (Eli, että $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp C$ ja $B \perp\!\!\!\perp C$, mutta A, B ja C eivät ole riippumattomia)

6. Rouva K on töissä kansallisen pankkilaitoksen rahapainossa. Työpäivän päätyttyä kaikki työntekijät joutuvat satunnaistarkastukseen, siten että todennäköisyys joutua tarkastetuksi on 0.03. Laske tn sille, että Rouva K joutuu työviikon (= 5 päivää) aikana tarkastukseen

   a) tasan kerran,

   b) ainakin kerran,

   c) useammin kuin kerran.

7. Laatikossa A on 2 valkoista ja 3 mustaa palloa, laatikossa B 1 valkoinen ja 4 mustaa palloa ja laatikossa C 5 valkoista ja 2 mustaa palloa. Nostetaan yksi pallo siten, että valitaan ensin todennäköisyydellä $\frac{3}{6}$ laatikko A, todennäköisyydellä $\frac{2}{6}$ laatikko B ja todennäköisyydellä $\frac{1}{6}$ laatikko B, ja nostetaan tämän jälkeen valitusta laatikosta umpimähkään yksi pallo. Millä todennäköisyydellä nostettu pallo on valkoinen?

8. Vanhoilla päivillään näkökykynsä menettänyt Robin Hood ampuu jousella ja hänen lapsenlapsensa ilmoittaa, onko laukaus osunut tauluun vai ei. Tn, että Robin osuu on 0.3 ja tn, että lapsenlapsen ilmoitus on virheetön on 0.8. Mikä on tn, että Robin on osunut lapsenlapsen näin ilmoittaessa?

Luku 4

Satunnaismuuttujat

Olemme jo aiemmin kurssilla tarkastelleet satunnaiskokeita ja niiden mahdollisia lopputuloksia. Monissa tilanteissa olemme kiinnostuneita jostakin satunnaiskokeen tuloksen funktiosta ennemmin kuin itse tuloksesta. Tämä tarkoittaa sitä, että esimerkiksi kahta noppaa heitettäessä meitä saattaa kiinnostaa noppien silmälukujen summa enemmän kuin se, mistä luvuista summa muodostuu[1]. Vastaavasti kolikonheitossa meitä saattaa kiinnostaa kruunien osuus sadasta heitosta enemmän kuin se nimenomainen kruuna-klaava-jono, joka satutaan saamaan. Tällaisissa tilanteissa meidän kannattaa mallintaa satunnaiskoetta *satunnaismuuttujan* avulla.

Esimerkiksi monissa esimerkeissämme käsitelty yksittäisen nopanheiton silmäluku on satunnaismuuttuja, vaikka sitä ei toistaiseksi kurssilla ole sellaisena esitetty. Satunnaismuuttujan määritteleminen on kuitenkin tärkeää, sillä sen kautta päästään hyödyntämään tärkeitä käsitteitä, kuten *jakaumia* ja niiden *odotusarvoa* sekä *hajontaa*, ja mallintamaan lukuisia erilaisia tilanteita matemaattisesti.

Satunnaismuuttujaksi kutsutaan satunnaiskokeeseen liittyvää funktiota, joka liittää jokaiseen satunnaiskokeen mahdolliseen tulokseen, eli alkeistapaukseen, jonkin reaaliluvun. Satunnaismuuttuja on siis numeerinen muuttuja, jonka arvo määräytyy satunnaiskokeen lopputuloksen mukaan. Tässä luvussa käsitellään satunnaismuuttujien erikoistapauksia, *diskreettejä satunnaismuuttujia.* Määritellään aluksi satunnaismuuttujan ja diskreetin satunnaismuuttujan käsitteet, ja havainnollistetaan tämän jälkeen diskreetin satunnaismuuttujan käsitettä esimerkeillä.

Määritelmä 4.1. *Satunnaismuuttuja (lyh. sm.).*[2] Olkoon $(\Omega, \mathcal{F}, P)$ tn-avaruus. Funktiota[3] $X : \Omega \to \mathbb{R}$ kutsutaan *satunnaismuuttujaksi.*

[1]Meille olennaista on siis tieto siitä, että silmälukujen summa on 6, eikä se, muodostuuko tämä summa silmäluvuista 2 ja 4, 1 ja 5 vai 5 ja 1, jne...

[2]engl. *random variable, r.v.*

[3]Tämä määritelmä riittää hyvin tämän kurssin tarpeisiin, mutta mainittakoon että tarkemmin ot-

Joskus käytämme satunnaismuuttujasta lyhennettyä merkintää $X(\omega) = X$.

Huomautus 4.2. Tässä vaiheessa lukijan on syytä huomata, että nimestään huolimatta satunnaismuuttuja ei ole tarkalleen ottaen satunnainen eikä muuttuja.

Määritelmä 4.3. *Diskreetti satunnaismuuttuja.* [4] Tn-avaruuden $(\Omega, \mathcal{F}, P)$ satunnaismuuttujaa $X : \Omega \to \mathbb{R}$ kutsutaan diskreetiksi satunnaismuuttujaksi, jos sen arvojoukko $X(\Omega)$ on numeroituva, eli äärellinen tai numeroituvasti ääretön.

Tämä tarkoittaa siis sitä, että satunnaismuuttujan mahdolliset arvot voidaan edes teoriassa luetella. Diskreettejä satunnaismuuttujia ovat esimerkiksi nopanheiton lopputulos (jonka mahdolliset tulokset ovat 1, ..., 6) tai maapallon väkiluku tiettynä ajanhetkenä (mahdollisia arvoja ovat kaikki luonnolliset luvut 0,1,2,...).

Esimerkki 4.4. Tarkastellaan kolmea yksinkertaista esimerkkiä diskreetistä satunnaismuuttujasta.

(i) Yhden nopan heitto. Määritellään perusjoukolle $\Omega = \{1, 2, 3, 4, 5, 6\}$ satunnaismuuttuja $X : \Omega \to \mathbb{R}$, $X(\omega) = \omega$. Satunnaismuuttuja X siis kuvaa heitetyn nopan silmälukua. Esimerkiksi joukko $\{X(\omega) = X = 3\}$ kuvaa tapahtumaa, jossa heitetty noppa saa silmäluvun kolme.

Tämän satunnaismuuttujan arvojoukko on $X(\Omega) = \{1, 2, 3, 4, 5, 6\}$, joten kyseessä on diskreetti satunnaismuuttuja.

(ii) Kahden nopan heitto. Määritellään perusjoukolle $\Omega = \{(a, b) : a, b \in \{1, 2, 3, 4, 5, 6\}\}$ satunnaismuuttuja $Y(\omega) = Y(a, b) = a + b$. Satunnaismuuttuja Y kuvaa nyt heitettyjen noppien silmälukujen summaa.

Satunnaismuuttujan Y arvojoukko on $Y(\Omega) = \{2, 3, \dots, 12\}$, joten kyseessä on myös diskreetti satunnaismuuttuja.

(iii) Heitä sikaa (vrt. esimerkki 2.24). Olkoon perusjoukko $\Omega = \{\omega_1, \dots, \omega_5\}$ seuraavilla todennäköisyyksillä.

---

taen yleisissä tapauksissa funktiolta X vaaditaan, että se on Borel-mitallinen, eli että jokaisen riittävän Borel-joukon, eli karkeasti sanottuna riittävän säännöllisen joukon, kuten välien ja välien numeroituvien yhdisteiden ja leikkausten, alkukuva on perusjoukon tapahtuma, eli että $X^{-1}(B) \in \mathcal{F}$ kaikille riittävän 'siisteille' $B \subseteq \mathbb{R}$.

[4] engl. *discrete random variable*

	Nimi	Asento	Pisteet	$P(\omega_i)$
ω_1	Siankylki	Kyljellään	0 p.	0.651
ω_2	Sianselkä	Selällään	5 p.	0.225
ω_3	Tavallinen sika	Jaloillaan	5 p.	0.088
ω_4	Siansyltty	Etujalkojen ja kärsän varassa	10 p.	0.030
ω_5	Etupotka	Toisen korvan ja etujalan varassa	15 p.	0.006

Määritellään satunnaismuuttuja $Z : \Omega \to \mathbb{R}$ kunkin perusjoukon alkion, eli sian laskeutumisasennon, tuottamaksi pistemääräksi; esimerkiksi $Z(\omega_1) = 0$. Satunnaismuuttujan Z arvojoukko on $Z(\Omega) = \{0, 5, 10, 15\}$, eli kyseessä on myös diskreetti satunnaismuuttuja.

Tapahtumaa, että satunnaismuuttuja X saa arvoja joukosta $B \in \mathbb{R}$ merkitään seuraavasti:

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}.$$

Tällöin todennäköisyyttä sille, että satunnaismuuttuja saa arvoja joukosta B voidaan merkitä jollain seuraavista tavoista:

$$P(X \in B) = P\{X \in B\} = P(\{X \in B\}) = P(\{\omega \in \Omega : X(\omega) \in B\}).$$

Jälleen[5] toiset sisäkkäisistä sulkeista yleensä tiputetaan pois merkinnöistä. Tällä kurssilla sovelletaan jatkossa merkintää $P(X \in B)$.

Esitellään seuraavaksi käsitteet *jakauma* ja *pistetodennäköisyysfunktio*. Määritellään ensin nämä käsitteet ja havainnollistetaan sen jälkeen esimerkkien avulla, mitä ne käytännössä tarkoittavat.

Satunnaismuuttujan mahdollisten arvojen todennäköisyyksiä kuvataan satunnaismuuttujan jaukaman avulla. Satunnaismuuttujan jakauma ja sen määräävä pistetodennäköisyysfunktio kertovat, millä todennäköisyydellä satunnaismuuttuja mitäkin arvoja saavuttaa, eli kuinka todennäköinen satunnaiskokeen kukin lopputulos on. Samalla tavoin kuin todennäköisyys on joukkofunktio, joka on määritelty perusjoukon osajoukoille eli tapahtumille, satunnaismuuttujan jakauma on joukkofunktio, joka on määritelty satunnaismuuttujan arvojoukon osajoukoille.

[5] vrt huom. 2.5

Määritelmä 4.5. *Satunnaismuuttujan jakauma.*[6] Olkoon $X : \Omega \to \mathbb{R}$ satunnaismuuttuja. Satunnaismuuttujan X jakauma on joukkofunktio

$$P_X := P(X \in B),$$

joka on määritelty kaikille satunnaismuuttujan X arvojoukon riittävän siisteille[7] osajoukoille $B \subseteq X(\Omega)$.

Huomautus 4.6. Voidaan osoittaa, että satunnaismuuttujalle X sen jakauma P_X toteuttaa todennäköisyyden aksioomat (TN1-TN3), ja siten määrittää todennäköisyyden satunnaismuuttujan arvojoukossa. Tämän takia monissa käytännön sovelluksissa alkuperäistä todennäköisyysavaruutta ja perusjoukkoa ei tarvitse erikseen määritellä, vaan puhutaan yleensä pelkästään satunnaismuuttujasta ja sen jakaumasta. Voidaankin ajatella, että perusjoukko on satunnaismuuttujan arvojoukko $X(\Omega)$, tapahtumat ovat sen osajoukkoja ja että todennäköisyys on satunnaismuuttujan jakauma P_X.

Määritelmä 4.7. *Pistetodennäköisyysfunktio (lyh. ptnf).*[8] Satunnaismuuttujan $X : \Omega \to \mathbb{R}$ pistetodennäköisyysfunktio on funktio $f : \mathbb{R} \to \mathbb{R}$, jolle

$$f(x) = P(X = x) \quad \text{kaikille} \quad x \in \mathbb{R}.$$

Pistetodennäköisyysfunktio on periaattessa määritelty kaikille reaaliluvuille, mutta sen arvo on nolla satunnaismuuttujan X arvojoukon ulkopuolella[9]:

$$f(x) = 0 \quad \text{kaikille} \quad x \notin X(\Omega).$$

Useampia satunnaismuuttujia käsiteltäessä satunnaismuuttujaa, jonka ptnf on kyseessä, merkitään usein sekaannusten välttämiseksi alaindeksillä:

$$f_X(x) = P(X = x).$$

Huomautus 4.8. Pistetodennäköisyysfunktio siis kertoo sen tapahtuman todennäköisyyden, että satunnaismuuttujan X arvo on x[10].

$$f(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\}).$$

[6] engl. *probability distribution*

[7] Jälleen tarkemmin reaalilukujen Borel-joukoille.

[8] engl. *probability mass function, pmf*

[9] Tämä tarkoittaa sitä, että tapahtuma, jossa satunnaismuuttuja saavuttaa arvojoukkonsa ulkopuolisen arvon (esim. yhden nopan heitossa saataisiin silmäluvuksi 7) on mahdoton.

[10] Tämän asian hahmottamiseen auttaa intuitiivisesti seuraavanlainen tilanne: teemme satunnaiskokeen, ja oletamme että satunnaiskokeen tulos noudattaa satunnaismuuttujaa X. Tällöin *ennen* satunnaiskokeen tekemistä luku X on vielä *tuntematon* eli satunnainen. Vastaavasti voimme merkitä, että luku x on tämän satunnaiskokeen tulos eli satunnaisilmiön *realisoitunut* arvo eli se arvo, jonka onnetar on päättänyt meille suoda. Tällöin pistetodennäköisyysfunktio $f(x)$ kuvaa todennäköisyyttä sille, että vielä arpomaton luku X realisoituu satunnaiskokeessa luvuksi x, toisin sanoin $P(X = x)$.

Kun $(\Omega, \mathcal{F}, P)$ on alkuperäinen tn-avaruus, niin $\{X = x\} \in \mathcal{F}$, ja yllä oleva P on alkuperäisen tn-avaruuden todennäköisyys. Ptnf siis antaa satunnaismuuttujan arvojen todennäköisyydet alkuperäisen perusjoukon tapahtumien todennäköisyyksinä.

Diskreetin satunnaismuuttujan, kuten aiemmin diskreetin tn-avaruuden (vrt. lause 2.19), tapauksessa tilanne yksinkertaistuu siten, että arvojoukon pisteiden todennäköisyydet määrittävät satunnaismuuttujan jakauman yksikäsitteisesti, ja tapahtumien todennäköisyydet saadaan yksittäisten arvojen todennäköisyyksien summana:

Lause 4.9. *Olkoon X diskreetti satunnaismuuttuja ja f sen ptnf. Nyt f määrää satunnaismuuttujan X jakauman kaavalla*

$$P(X \in B) = \sum_{x \in B} f(x).$$

Todistus. Harjoitustehtävä. Kokonaistodennäköisyyden kaavasta käyttämällä perusjoukolle $\Omega = \{x_1, x_2, \dots\}$ ositusta

$$\{X \notin X(\Omega)\}, \{X = x_1\}, \{X = x_2\}, \dots$$

□

Havainnollistetaan seuraavaksi äsken määriteltyjä käsitteitä esimerkkien avulla.

Esimerkki 4.10. Jatkoa esimerkille 4.4. Määritellään edellisen esimerkin satunnaismuuttujien pistetodennäköisyysfunktiot.

(i) Yhden nopan tapaus. Oletetaan symmetrinen perusjoukko, eli että kaikkien perusjoukon alkoiden tn on

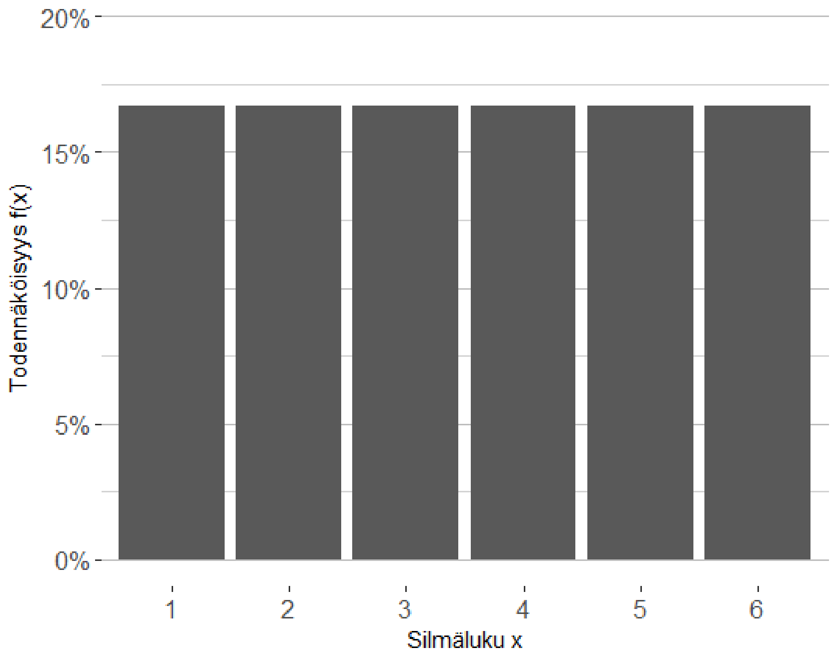
$$P(\omega) = \frac{1}{6}.$$

Nyt voimme merkitä lyhennetysti $\{\omega \in \Omega : \omega = x\} := \{X = x\} := \{x\}$, joten myös

$$f_X(x) = P(X = x) = P(x) = \frac{1}{6} \quad \text{kaikille} \quad x \in X(\Omega).$$

Lisäksi $f(x) = 0$ kaikille $x \notin X(\Omega) = \{1, 2, \dots, 6\}$; monesti tästä ei erikseen mainita.

Pistetodennäköisyysfunktion määrittelemiä jakaumia voidaan graafisesti havainnollistaa kuvaajalla, jossa x-akselilla on satunnaismuuttujan X arvojoukon alkiot x_i ja y-akselilla vastaavat todennäköisyydet $P(X = x_i)$. Tällöin satunnaismuuttujan pistetodennäköisyysfunktion kuvaaja näyttää yhden nopan tapauksessa seuraavalta:



Kuva 4.1: Yhden nopan heittotuloksen jakauma.

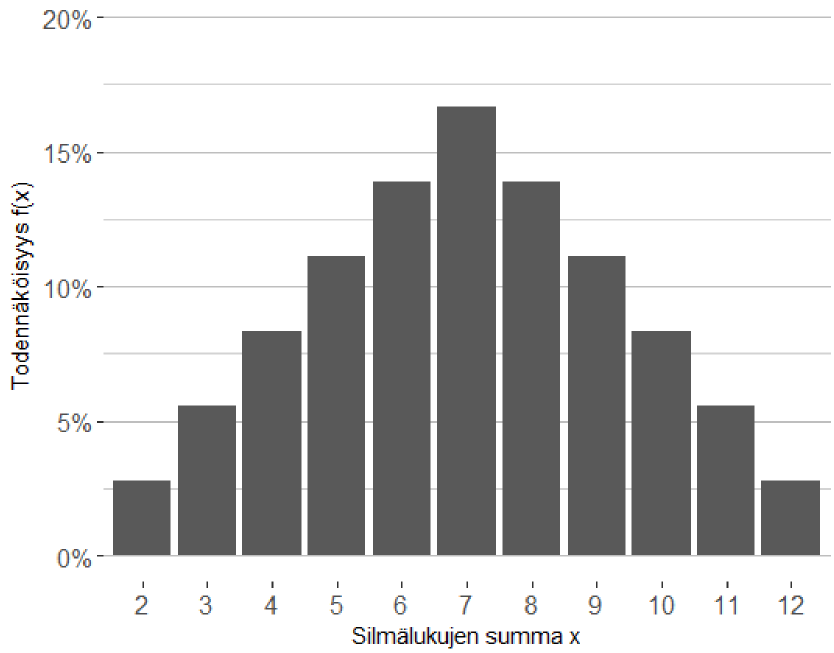
(ii) Kahden nopan tapaus. Oletetaan jälleen symmetrinen perusjoukko, eli että kaikkien perusjoukon alkoiden tn on

$$P(\omega) = \frac{1}{36}.$$

Nyt pistetodennäköisyydet saadaan laskemalla satunnaismuuttujan arvoille suotuisat alkeistapaukset:

y	Suotuisat alkeistapaukset	$f_Y(y)$
2	$(1,1)$	$1/36$
3	$(1,2),(2,1)$	$2/36$
4	$(1,3),(2,2),(3,1)$	$3/36$
5	$(1,4),(2,3),(3,2),(4,1)$	$4/36$
6	$(1,5),(2,4),(3,3),(4,2),(5,1)$	$5/36$
7	$(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$	$6/36$
8	$(2,6),(3,5),(4,4),(5,3),(6,2)$	$5/36$
9	$(3,6),(4,5),(5,4),(6,3)$	$4/36$
10	$(4,6),(5,5),(6,4)$	$3/36$
11	$(5,6),(6,5)$	$2/36$
12	$(6,6)$	$1/36$

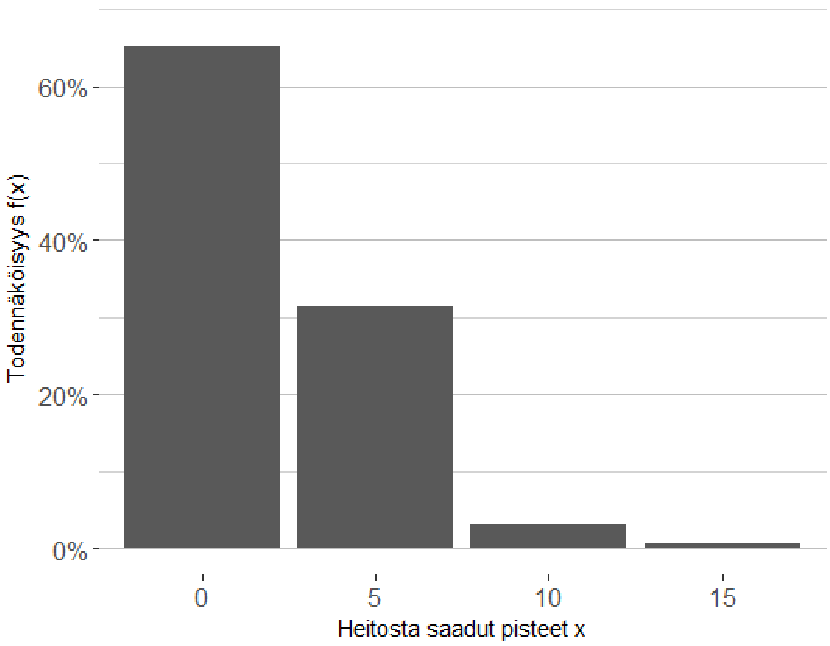
Kahden nopan tapauksessa pistetodennäköisyysfunktion kuvaaja on esitetty alla. Kuvaajasta nähdään esimerkiksi, että luku 7 on kaikista todennäköisin summa kahden nopan silmäluvuille.



Kuva 4.2: Kahden nopan silmälukujen summan jakauma.

(iii) Heitä sikaa -tapaus. Satunnaismuuttujan pistetodennäköisyydet saadaan jälleen laskemalla yhteen kunkin satunnaismuuttujan arvon tuottavien perusjoukon alkioiden todennäköisyydet:

z	Suotuisat alkeistapaukset	Asento	$f_Z(z)$
0	ω_1	Siankylki	0.651
5	ω_2, ω_3	Sianselkä, Tavallinen sika	0.313
10	ω_4	Siansyltty	0.030
15	ω_5	Etupotka	0.006



Kuva 4.3: Yhden sian heiton tuomien pisteiden jakauma.

Edellisten esimerkkien tapauksessa huomattiin, että kaikkien tarkasteltujen satunnaismuuttujien ptnf:t summautuivat ykköseksi. Tämä pätee myös yleisesti:

Lause 4.11. *Pistetodennäköisyysfunktion ominaisuuksia. Olkoon $X : \Omega \to \mathbb{R}$ diskreetti satunnaismuuttuja. Sen ptnf:lle f pätee:*

(i) $f(x) \geq 0$ *kaikille* $x \in \mathbb{R}$

(ii) jos $f(x) > 0$, $x \in X(\Omega) = \{x_1, x_2 \dots\}$, *eli* x *kuuluu* X*:n arvojoukkoon.*

(iii) $\sum_{i=1}^{\infty} f(x_i) = 1$.

Voidaan osoittaa, että nämä ehdot pätevät myös toiseen suuntaan, eli että jokainen ei-negatiivinen funktio $f : \mathbb{R} \to \mathbb{R}$, joka saa positiivisia arvoja vain äärellisessä tai numeroituvasti äärettömässä joukossa ja summautuu kaikkien indeksiensä yli ykkökseksi, on jonkin diskreetin satunnaismuuttujan pistetodennäköisyysfunktio.

Pistetodennäköisyysfunktiosta voidaan johtaa satunnaismuuttujan *kertymäfunktio*. Siinä missä ptnf kertoo, kuinka todennäköistä on satunnaiskokeessa saavuttaa **tasan** jokin lopputulos, kertymäfunktio kuvaa, kuinka todennäköistä on, että saavutetaan **korkeintaan** tietyn suuruinen arvo. Esitetään seuraavaksi kertymäfunktion ja diskreetin kertymäfuntion määritelmät.

Määritelmä 4.12. *Kertymäfunktio (lyh. kf).* [11] Satunnaismuuttujan X kertymäfunktio on funktio $F : \mathbb{R} \to \mathbb{R}$, jolle

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

Määritelmä 4.13. *Diskreetti kertymäfunktio.*[12] Diskreetin satunnaismuuttujan X kertymäfunktio (lyh. kf) on funktio $F : \mathbb{R} \to \mathbb{R}$, jolle pätee

$$F(x) = P(X \leq x) = \sum_{x_k \leq x} P(X = x_k) = \sum_{x_k \leq x} f(x_k),$$

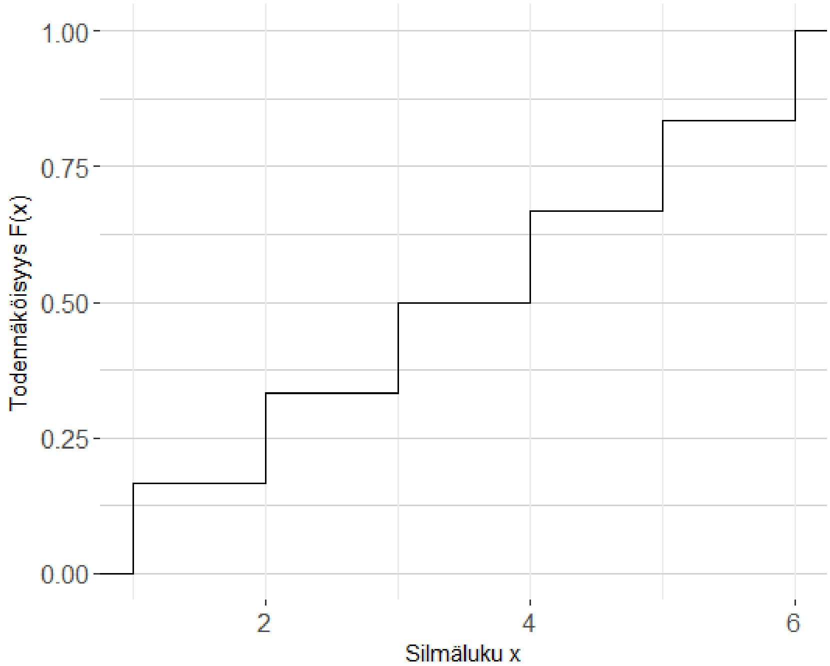
missä $k \in \{0, 1, 2, \dots, x\}$. Diskreetin satunnaismuuttujan kertymäfunktio saadaan siis ptnf:n f summana. Koska diskreetin sm:n arvojoukko on numeroituva, myös kyseinen summa on numeroituva.

Tarkastellaan seuraavaksi esimerkkiä diskreetistä kertymäfunktiosta.

Esimerkki 4.14. Yhden nopan silmäluku. Jatkoa esimerkkien 4.4 ja 4.10 i-kohdille. Satunnaismuuttujan X (yhden nopan silmäluku) kertymäfunktion kuvaajasta nähdään esimerkiksi, että todennäköisyys saada noppaa heittämällä silmäluvuksi 3 tai sitä pienempi luku on 0.5 ja että todennäköisyys saada silmäluvuksi 6 tai sitä pienempi luku on 1 (eli on varmaa, että nopanheiton tulos on korkeintaan 6).

[11] engl. *cumulative distribution function, CDF*

[12] engl. *discrete cumulative distribution function*



Kuva 4.4: Yhden nopan heittotuloksen kertymäfunktion jakauma

Seuraavassa lauseessa esitellyt kertymäfunktion ominaisuudet ovat suoraa seurausta kertymäfunktion määritelmästä.

Lause 4.15. *Kertymäfunktion ominaisuuksia. Satunnaismuuttujan X kertymäfunktiolla $F : \mathbb{R} \to \mathbb{R}$ on seuraavat ominaisuudet:*

(i) F on kasvava funktio

(ii) F on oikealta jatkuva funktio eli

$$F(x+) = F(x), \quad \textit{kaikilla } x \in \mathbb{R}.$$

(iii) $F(-\infty) = 0$ ja $F(\infty) = 1$.

Ja jälleen kääntäen: jos jollakin funktiolla F on nämä ominaisuudet, on olemassa satunnaismuuttuja X siten, että $F(X)$ on X:n kertymäfunktio. Lukija voi kuvasta 4.4 tarkistaa, että esimerkin 4.14 kertymäfunktio toteuttaa lauseen 4.15 kohdat.

Kertymäfunktiot ovat hyödyllisiä erityisesti *jatkuvien satunnaismuuttujien* tapauksissa, joita käsitellään seuraavassa luvussa. Diskreeteille jakaumille laskettu kertymäfunktio ei ole aivan yhtä hyödyllinen kuin jatkuville satunnaismuuttujille määritelty ystävänsä, sillä sen laskeminen symbolisesti on yleensä huomattavan työlästä summattavien termien lisääntyessä. Ennen tietokoneiden aikaa tämä saattoi olla iso ongelma, ja diskreettien satunnaismuuttujien kertymäfunktioiden arvoja approksimoitiin muilla jakaumilla. Tästä lisää myöhemmissä luvuissa.

4.1 Diskreettejä jakaumia

Äsken opimme, että jokainen ei-negatiivinen funktio $f : \mathbb{R} \to \mathbb{R}$, joka summautuu kaikkien indeksiensä yli ykkökseksi (ja saa positiivisia arvoja vain numeroituvassa joukossa), on jonkin diskreetin satunnaismuuttujan pistetodennäköisyysfunktio. Tämä tarkoittaa siis sitä, että erilaisia diskreettejä jakaumia on olemassa ääretön määrä. Jotkin ominaisuuksiltaan hyödylliset ja käytännön ongelmien ratkaisuun soveltuvat jakaumat on nimetty erikseen. Esitellään seuraavaksi tärkeimpiä näistä jakaumista.

4.1.1 Diskreetti tasajakauma

Määritelmä 4.16. Satunnaismuuttujalla X on *diskreetti tasajakauma*[13] joukossa $\{1, \dots, n\}$, jos sen ptnf on

$$f(k) = P(X = k) = \frac{1}{n} \quad \text{kaikille} \quad k \in \{1, \dots, n\}.$$

Esimerkkien 4.4 ja 4.10 kohdan (i) (nopanheitto) satunnaismuuttuja X noudattaa diskreettiä tasajakaumaa joukossa $\{1, \dots, 6\}$. Kaikki diskreettiä tasajakaumaa noudattavan satunnaismuuttujan arvojoukon arvot ovat yhtä todennäköisiä keskenään.

4.1.2 Bernoulli-jakauma

Määritelmä 4.17. Satunnaismuuttuja X noudattaa *Bernoulli-jakaumaa*[14] parametrilla[15] $p \in (0, 1)$, jos sen ptnf on

$$f(k) = P(X = k) = p^k(1-p)^{1-k}, \quad \text{missä } k \in \{0, 1\}.$$

Tällöin merkitään $X \sim \text{Bernoulli}(p)$.

Bernoulli-jakaumalla voidaan kuvata satunnaiskoetta, jossa on kaksi vaihtoehtoa, onnistuminen ($X = 1$) ja epäonnistuminen ($X = 0$), ja onnistumistodennäköisyys on p.

[13] engl. *(discrete) uniform distribution*

[14] engl. *Bernoulli distribution*

[15] Parametri on jokin numeerinen arvo, joka vaikuttaa jakauman pistetodennäköisyysfunktioon.

Esimerkiksi kolikonheitossa kruunaa voidaan merkitä ykkösellä ja klaavaa nollalla, jolloin tulos noudattaa Bernoullin jakaumaa parametrilla $\frac{1}{2}$, tai nopanheitossa kutonen voidaan määritellä onnistumiseksi ja muu tulos epäonnistumiseksi, jolloin tulos noudattaa Bernoulli-jakaumaa parametrilla $\frac{1}{6}$.

Esimerkki 4.18. Jos merkitään klaavan saamista nollalla, ja kruunan saamista ykkösellä, niin Bernoulli-jakauman tiheysfunktiosta nähdään, että kruunan todennäköisyys on

$$f(1) = 0.5^1 \cdot 0.5^0 = 0.5.$$

Huomataan, että (reilun) kolikonheiton tapauksessa Bernoulli-jakauman parametri p saa arvon 0.5, sillä 'onnistumisen' eli tässä yhteydessä kruunan saamisen todennäköisyys on 0.5. Painotetun kolikon tapauksessa kolikonheiton tulos voisi noudattaa Bernoulli-jakaumaa esimerkiksi parametrilla 0.7, jolloin pätisi $f(1) = 0.7$ ja $f(0) = 0.3$. Bernoulli-jakauman ptnf siis yksinkertaisesti palauttaa 'onnistumisen' ($= 1$) tai 'epäonnistumisen' ($= 0$) todennäköisyyden.

4.1.3 Binomijakauma

Määritelmä 4.19. Satunnaismuuttuja X noudattaa *binomijakaumaa*[16] parametreilla $n \in \{1, 2, \dots\}$ ja $p \in [0, 1]$, jos sen ptnf on

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{kaikille} \quad k \in \{0, 1, \dots, n\}.$$

Tällöin merkitään $X \sim \mathrm{Bin}(n, p)$.

Bernoulli-jakauman yleistä tapausta kutsutaan binomijakaumaksi. Binomijakaumaa voidaan käyttää tilanteissa, joissa samaa satunnaiskoetta (kuten kolikonheittoa) toistetaan monta kertaa eli binomijakautuneella satunnaismuuttujalla voidaan mallintaa onnistumisten lukumäärää toistokokeessa, jota käsittelimme kappaleessa 3.4.

Jos esimerkin 2.10 kohdassa (ii) nostettujen valkoisten pallojen määrää neljän pallon otoksessa kuvataan satunnaismuuttujalla X, niin X noudattaa binomijakaumaa parametreilla $n = 4$ ja $p = \frac{5}{8}$.

Samoin lauseen 2.11 kohdassa (ii) nostettujen valkoisten pallojen määrää n:stä nostetusta pallosta voidaan kuvata satunnaismuuttujalla X, joka noudattaa binomijakaumaa parametreilla n ja $p = \frac{K}{N}$, missä K on korissa olevien valkoisten pallojen, ja N kaikkien korissa olevien pallojen määrä. Helpolla laskulla nähdään, että lauseen kaava todellakin on binomijakautuneen satunnaismuuttujan pistetodennäköisyysfunktio.

[16]engl. *binomial distribution*

Huomautus 4.20. Totesimme äsken, että binomijakauma on Bernoulli-jakauman yleistys. Tämä pitää paikkansa, sillä voidaan huomata, että Bernoulli-jakaumaa paremetrillä p noudattava satunnaismuuttuja on oikeastaan binomijakaumaa parametreilla $(1, p)$ noudattava satunnaismuuttuja.

4.1.4 Hypergeometrinen jakauma

Määritelmä 4.21. Satunnaismuuttuja X noudattaa *hypergeometrista jakaumaa*[17] parametreilla $N \in \{1, 2, \dots\}$ ja $n, K \in \{1, \dots, N\}$, jos sen pistetodennäköisyysfunktio on

$$f(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \text{kaikille} \quad k \in \{0, 1, \dots, n\}.$$

Tällöin käytetään merkintää

$$X \sim \text{Hyperg}(N, K, n).$$

Hypergeometrisella jakaumalla voidaan mallintaa otantaa ilman takaisinpanoa. Jos esimerkin 2.10 kohdassa (i) nostettujen valkoisten pallojen määrää otoksessa kuvataan satunnaismuuttujalla X, se noudattaa hypergeometrista jakaumaa parametreilla $N = 8$, $K = 5$, ja $n = 4$. Samoin lauseen 2.11 kohdassa (i) nostettujen valkoisten pallojen määrää voidaan kuvata hypergeometrisella jakaumalla.

Huomautus 4.22. Huomaa, että koska binomikertoimen $\binom{x}{y}$ arvo on määritelmän mukaan 0, jos $y > x$, hypergeometrista jakaumaa noudattavan satunnaismuuttujan arvojoukkoon kuuluvat vain sellaiset $k \in \{0, 1, \dots, n\}$, joille $k \leq K$ ja $n - k \leq N - K$. Korista ei siis voida nostaa enempää valkoisia tai mustia palloja kuin niitä siellä on, jos palloja ei palauteta nostojen välillä!

4.1.5 Geometrinen jakauma

Määritelmä 4.23. Satunnaismuuttuja X noudattaa *geometrista jakaumaa*[18] parametrilla $p \in (0, 1)$, jos sen ptnf on

$$f(k) = P(X = k) = p(1 - p)^k \quad \text{kaikille} \quad k \in \mathbb{N}.$$

Tällöin merkitään

$$X \sim \text{Geom}(p).$$

[17] engl. *hypergeometric distribution*
[18] engl. *geometric distribution*

Toistetaan koetta, jonka onnistumistodennäköisyys on p, kunnes saadaan ensimmäinen onnistuminen. Tällöin sm X, joka kuvaa epäonnistuneiden koekertojen määrää ennen ensimmäistä onnistumista, noudattaa geometrista jakaumaa parametrilla p. Geometrisen jakauman tiheysfunktion takaa löytyvää intuitiota voi hahmotella ajattelemalla riippumattomia toistoja satunnaiskokeessa, jossa epäonnistumisia sattuu k kappaletta, ja tämän jälkeen viimein onnistuminen todennäköisyydellä p.

Ensimmäiseen onnistumiseen voi periaatteessa tarvita kuinka monta toistokertaa tahansa, joten geometrista jakaumaa noudattavan satunnaismuuttujan arvojoukkoon kuuluvat kaikki luonnolliset luvut.

Esimerkki 4.24. Heitetään noppaa, kunnes saadaan ensimmäinen kutonen. Merkitään satunnaismuuttujalla X heittojen määrää, joka tarvitaan ensimmäisen kutosen saamiseksi (heittoa, jolla saadaan ensimmäinen kutonen, ei lasketa näihin). Satunnaismuuttuja X noudattaa geometrista jakaumaa parametrilla $p = \frac{1}{6}$. Satunnaismuuttujan X ptnf on myös helppo päätellä suoraan:

- $X = 0$, jos heti ensimmäisellä heitolla saadaan kutonen, eli $P(X = 0) = \frac{1}{6} = p(1-p)^0$.
- $X = 1$, jos ensimmäisellä heitolla ei saada kutosta, mutta toisella saadaan, eli $P(X = 1) = \frac{5}{6} \cdot \frac{1}{6} = p(1-p)^1$
- $X = 2$, jos kahdella ensimmäisellä heitolla ei saada kutosta, mutta kolmannella saadaan, eli $P(X = 2) = \left(\frac{5}{6}\right)^2 \cdot \frac{1}{6} = p(1-p)^2$ ja niin edelleen.

Huomautus 4.25. Voidaan helposti tarkastaa, että myös geometrista jakaumaa noudattavan satunnaismuuttujan ptnf toteuttaa lauseen 4.11 ehdot. Erityisesti geometrisen sarjan (sarja suppenee, koska $1 - p < 1$) summan kaavaa soveltamalla nähdään, että pistetodennäköisyydet summautuvat ykköseksi:

$$\sum_{k=0}^{\infty} p(1-p)^k = p\frac{1}{1-(1-p)} = 1.$$

Huomautus 4.26. Geometrisen jakauman merkintätavat ja parametrisointi vaihtelevat lähteen mukaan. Joissakin lähteissä geometrisen jakauman ptnf on

$$f(k) = p(1-p)^{k-1},$$

jolloin geometrisesti jakautunut satunnaismuuttuja kuvaa epäonnistuneiden koekertojen lukumäärän sijaan ensimmäisen onnistumisen järjestyslukua.

4.1.6 Poisson-jakauma

Määritelmä 4.27. Satunnaismuuttuja X noudattaa *Poisson-jakaumaa*[19] parametrilla $\lambda > 0$, jos sen ptnf on

$$f(k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{kaikille} \quad k \in \mathbb{N}.$$

Tällöin merkitään

$$X \sim \text{Poisson}(\lambda).$$

Poisson-jakautuneella satunnaismuuttujalla voidaan kuvata joidenkin tapahtumien määrää aika- tai tilayksikköä kohti, jos tapahtumat oletetaan toisistaan riippumattomiksi ja niiden odotettu määrä aika- tai tilayksikköä kohti vakioksi. Poisson-jakautuneella satunnaismuuttujalla voidaan kuvata esimerkiksi:

- Joukkueen tekemien maalien määrää jalkapallo-ottelussa.
- Liikenneonnettomuuksien määrää tiekilometria kohti.
- Onkijan tunnissa saamien kalojen määrää.

Esimerkki 4.28. Kuvataan satunnaismuuttujalla X kotijoukkueen tekemien maalien määrä La Ligassa (Espanjan jalkapallon pääsarja) kausilla 2010-2015 (yhteensä 2116 ottelua). Oletetaan, että $X \sim \text{Poisson}(1.631)$[20]

Oheiseen taulukkoon on kirjattu kotijoukkueen maalimäärien havaitut määrät ja osuudet peleissä (esimerkiksi 466:ssä pelissä kotijoukkue on tehnyt 0 maalia, joten osuus on $466/2116 \approx 0.2202$) ja Poisson-jakauman pistetodennäköisyysfunktion (parametrilla $\lambda = 1.631$) arvot. Huomataan, että havaitut frekvenssit ovat hyvin lähellä Poisson-jakaumasta laskettuja, joten oletus siitä, että maalimäärät noudattavat Poisson-jakaumaa, vaikuttaa järkevältä.[21]

[19] engl. *Poisson distribution*

[20] Käytetty parametrin arvo $\lambda = 1.631$ on aineistosta laskettu maalien keskiarvo ottelua kohti. Jakauman parametrien arvojen päättelyyn aineiston perusteella, eli ns. tilastolliseen päättelyyn tutustutaan tarkemmin kurssilla Tilastollinen päättely I. Tällöin osoittautuu, että monille (tarkemmin eksponenttiperheen) jakaumille hyvä parametrin estimaatti tulee havaintojen otoskeskiarvosta.

[21] Tosin suuria maalimääriä on hieman enemmän kuin oletuksen Poisson-jakautuneisuudesta perusteella pitäisi olla. Tämä selittynee sillä, että sarjan joukkueet ovat eritasoisia, joten oletus siitä, että fmaalien syntymistodennäköisyys olisi vakio ottelua kohti, ei täyty. Maalimäärien tarkempaan mallintamiseen tulisikin ottaa huomioon ainakin pelaavien joukkueiden tasoero. Tällöin kysymyksessä olisi Poisson-regressio; tähän palataan kurssilla Yleistetyt lineaariset mallit.

Maalit	Ottelut	Osuus	$f_X(x)$
0	466	0.2202	0.1957
1	668	0.3157	0.3192
2	498	0.2353	0.2604
3	262	0.1238	0.1416
4	147	0.0695	0.0577
5	50	0.0236	0.0188
6	15	0.0071	0.0051
7	7	0.0033	0.0012
8	2	0.0009	0.0002
9	1	0.0005	0.0000
≥ 10	0	0.0000	0.0000

Tarkka todennäköisyys sille, että kotijoukkue tekee vähintään kymmenen maalia ottelussa, saadaan laskettua komplementtitapahtuman kautta:

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - \sum_{i=0}^{9} P(X = i) \approx 8.42 \cdot 10^{-6}.$$

Tämä todennäköisyys pyöristyy nollaan taulukossa, jossa käytetään neljän desimaalin tarkkuutta.

Vertaa esimerkkiin 2.25, jossa pistetodennäköisyydet asetettiin suoraan alkeistapauksille. Todennäköisyydet ovat samat, mutta satunnaismuuttujan käyttäminen yksinkertaistaa merkintöjä ja laskuja.

4.2 Diskreetin satunnaismuuttujan odotusarvo

Satunnaismuuttujan *odotusarvo* on yksi todennäköisyyslaskennan tärkeimmistä konsepteista. Odotusarvo on nimensä mukaisesti satunnaismuuttujan odotettu arvo, ja se määritellään diskreetille satunnaismuuttujalle todennäköisyyksillä painotettuna keskiarvona satunnaismuuttujan mahdollisista arvoista. Tässä kappaleessa käsitellään lähinnä odotusarvon laskemista diskreeteille satunnaismuuttujille. Odotusarvon ominaisuuksia käsitellään tarkemmin luvussa 6. Esitetään seuraavaksi diskreetin satunnaismuuttujan odotusarvon määritelmä, ja tarkastellaan tämän jälkeen muutamaa esimerkkiä.

Määritelmä 4.29. *(Diskreetin satunnaismuuttujan) odotusarvo.*[22] Olkoon X diskreetti satunnaismuuttuja, jonka ptnf on f, ja arvojoukko $X(\Omega) = \{x_1, x_2, \dots\}$. Satunnaismuut-

[22] engl. *expected value, EV*

tujan X odotusarvo on

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

jos kyseinen sarja suppenee itseisesti, eli jos

$$\sum_{i=1}^{\infty} |x_i| f(x_i) < \infty.$$

Mikäli sarja ei suppene itseisesti, sanotaan että X:llä ei ole odotusarvoa. Jos satunnaismuuttujan $X(\Omega)$ arvojoukko on äärellinen, niin myös kyseinen summa on aina äärellinen, ja siten odotusarvo on aina olemassa.

Satunnaismuuttujan odotusarvoa voi ajatella sen jakauman painopisteenä: Jos massattoman sauvan pisteisiin $\{x_1, x_2, \dots\}$ asetetaan pistetodennäköisyysfunktion arvoja näissä pisteissä vastaavat painot $\{f(x_1), f(x_2), \dots\}$, niin sauva pysyy tasapainossa, jos sitä tuetaan odotusarvon kohdalta.

Toinen tapa havainnollistaa odotusarvoa on uhkapelin keskimääräisenä voittosummana kierrosta kohti: jos peliä jatketaan monta kierrosta, keskimääräinen voitto kierrosta kohti lähestyy kierroksen voittosummaa kuvaavan satunnaismuuttujan odotusarvoa. Seuraavat esimerkit havainnollistavat näitä tulkintoja.

Esimerkki 4.30. Oletetaan, että satunnaismuuttuja X kuvaa yhden nopan heiton silmälukua, jolloin sen ptnf on

$$f(k) = P(X = k) = \frac{1}{6} \quad \text{kaikille} \quad k \in \{1, 2, \dots, 6\}.$$

Tällöin X:n odotusarvoksi saadaan

$$E(X) = \sum_{k=1}^{6} x_k f(x_k) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Esimerkki 4.31. Tarkastellaan peliä, jossa pelaaja heittää kahta noppaa, ja voittaa silmälukujen summan verran rahaa. Kannattaako tästä pelistä maksaa seitsemän euroa kierrokselta?

Merkitään silmälukujen summaa, joka kuvaa myös kierroksella voitettua rahamäärää, satunnaismuuttujalla X. Arvojoukko on $\{2, 3, \dots, 12\}$, ja ptnf laskettiin esimerkissä 4.10. Lasketaan X:n odotusarvo:

$$\begin{aligned} E(X) &= \sum_{i=2}^{12} x_i f(x_i) = \frac{2\cdot 1 + 3\cdot 2 + 4\cdot 3 + 5\cdot 4 + 6\cdot 5 + 7\cdot 6 + 8\cdot 5 + 9\cdot 4 + 10\cdot 3 + 11\cdot 2 + 12\cdot 1}{36} \\ &= \frac{252}{36} = 7. \end{aligned}$$

Voiton odotusarvo on tasan 7, joten seitsemän euron panoksella peli on *reilu*: keskimäärin kumpikaan pelaajasta ja pelin tarjoajasta ei jää voitolle.

Esitellään vielä esimerkin avulla, miten Poisson-jakaumaa noudattavan satunnaismuuttujan odotusarvo saadaan laskettua odotusarvon määritelmän kautta.

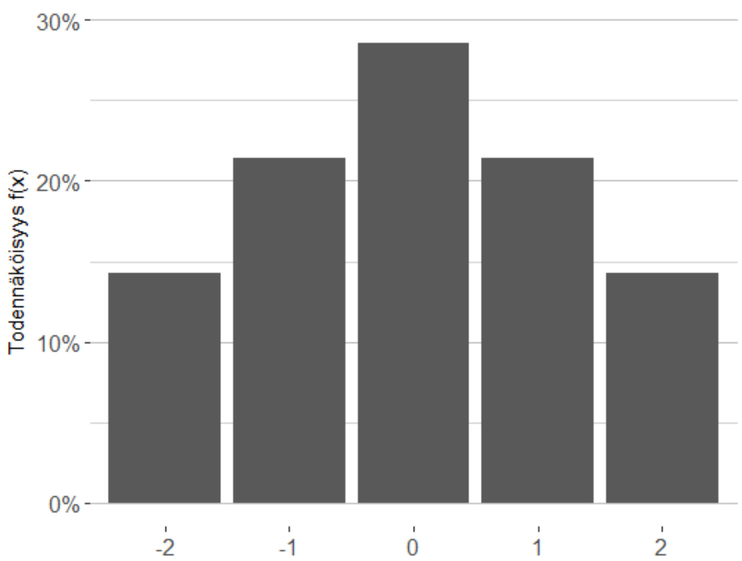
Esimerkki 4.32. Poisson-jakaumaa parametrilla λ noudattavan satunnaismuuttujan X odotusarvo saadaan laskettua käyttämällä eksponenttifunktion sarjakehitelmää

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

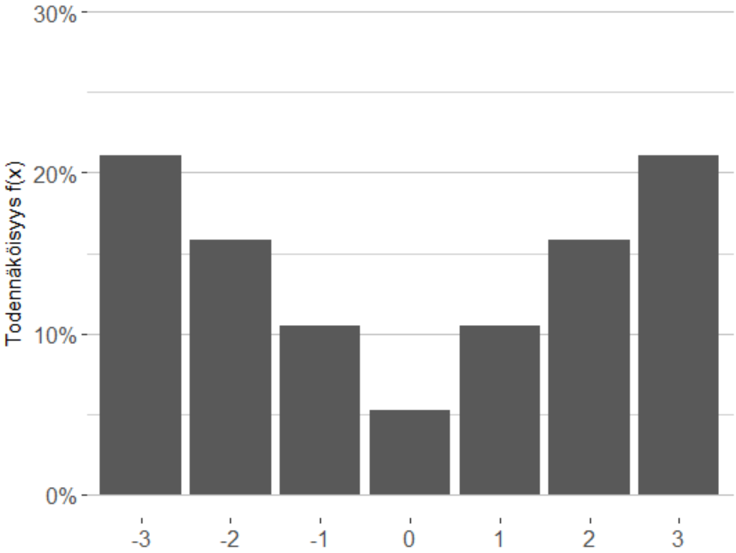
apuna:

$$E(X) = \sum_{k=1}^{\infty} kf(k) = 0 + \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Odotusarvon lisäksi satunnaismuuttujien jakaumien toinen tärkeä tunnusluku on niiden *varianssi*. Varianssi kuvaa jakaumien hajontaa, eli sitä kuinka paljon satunnaismuuttujan saavuttamat arvot poikkeavat jakauman odotusarvosta. Mitä suurempi jakauman varianssi on, sitä enemmän sen arvojoukon arvot poikkeavat jakauman odotusarvosta. Asiaa havainnollistavat seuraavat kuvat jakaumista, joilla on sama odotusarvo, mutta eri varianssit. Varianssia ja sen laskemista käsitellään tarkemmin luvussa 6, mutta lukijan kannattaa jo nyt pitää sen intuitiivinen käsite mielessä. Seuraavaksi kuvissa esiintyy havainnollistukset kahdesta eri todennäköisyysjakaumasta, joilla on sama odotusarvo mutta varianssit (ja jakauman muut ominaisuudet) eroavat.



Kuva 4.5: Odotusarvo 0, varianssi 1.69.



Kuva 4.6: Odotusarvo 0, varianssi 5.56.

4.3 Satunnaismuuttujien riippumattomuus

Muistellaan vielä edellistä lukua, jossa käsittelimme riippumattomuutta. Riippumattomuuden käsitettä voi soveltaa myös satunnaismuuttujiin aivan samoin kuin edellisluvussa mihin tahansa tapahtumiin, joihin liitetään todennäköisyyksiä. Intuitiivisesti satunnaismuuttujien riippumattomuus tarkoittaa, että niiden arvot eivät vaikuta toisiinsa: jos tiedämme toisen satunnaismuuttujan arvon, se ei vaikuta arvioomme toisen satunnaismuuttujan arvosta. Satunnaismuuttujien riippumattomuus määritellään tuttuun tyyliin tapahtumien riippumattomuuden kautta:

Määritelmä 4.33. *Satunnaismuuttujien riippumattomuus.* Satunnaismuuttujat X ja Y ovat riippumattomia, jos

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

eli jos

$$\{X \in A\} \perp\!\!\!\perp \{Y \in B\}$$

kaikille (riittävän säännöllisille) joukoille $A, B \subseteq \mathbb{R}$. Tällöin merkitään

$$X \perp\!\!\!\perp Y.$$

Huomautus 4.34. Jos $X \perp\!\!\!\perp Y$, ja $P(Y \in B) > 0$, niin joukkojen ehdolliset todennäköisyydet palautuvat ehdottomiin todennäköisyyksiin (vrt. lause 3.12):

$$\begin{aligned} P((X \in A)|(Y \in B)) &= \frac{P((X \in A) \cap P(Y \in B))}{P(Y \in B)} \\ &= \frac{P(X \in A)P(Y \in B)}{P(Y \in B)} \\ &= P(X \in A). \end{aligned}$$

Jos satunnaismuuttujat ovat keskenään riippumattomia, niin myös niiden *muunnokset*[23] ovat keskenään riippumattomia:

Lause 4.35. *Satunnaismuuttujien muunnosten riippumattomuus. Jos $X \perp\!\!\!\perp Y$, niin kaikille (riittävän säännöllisille) funktioille $g, h : \mathbb{R} \to \mathbb{R}$ pätee*

$$g(X) \perp\!\!\!\perp h(Y).$$

Riippumattomuus yleistyy useammalle satunnaismuuttujalle seuraavasti:

Määritelmä 4.36. *Usean satunnaismuuttujan riippumattomuus.* Joukko satunnaismuuttujia $X_1, \ldots, X_n$ on riippumaton, jos

$$P\left(\bigcap_{i=1}^{n} (X_i \in B_i)\right) = \prod_{i=1}^{n} P(X_i \in B_i)$$

kaikille (riittävän säännöllisille) joukoille $B_1, \ldots, B_n$. Tällöin merkitään

$$X_1, \ldots, X_n \perp\!\!\!\perp .$$

Useamman satunnaismuuttujan riippumattomuus on hyödyllistä esimerkiksi tilastollisessa analyysissä, missä ilmiöllä on useita selittäviä tekijöitä ja halutaan tarkastella näiden selittävien tekijöiden keskinäistä riippuvuutta. Tästä aiheesta lisää myöhemmillä kursseilla.

4.4 Harjoitustehtäviä

1. Satunnaismuuttuja X noudattaa tasajakaumaa joukossa 3, 4, ..., 11. Laske $E(X)$.

[23] Muunnoksella tarkoitetaan satunnaismuuttujasta jollain funktiolla saatua toista satunnaismuuttujaa. Esimerkiksi satunnaismuuttujan X yksinkertainen muunnos voi olla X^2. Satunnaismuuttujien muunnoksia käsitellään kappaleissa 5.3 ja 8.1

2. Ilmoita satunnaismuuttujan X jakauma (jakauman nimi ja parametrien arvot), jos

   (a) X on viallisten tuotteiden lukumäärä laatikossa, johon on pakattu 48 tuotetta, kun kukin tuote on toisista riippumatta viallinen todennäköisyydellä 0.5.

   (b) X on ässien lukumäärä vedettäessä 13 korttia ilman takaisinpanoa pakasta.

   (c) X on turhien kertojen lukumäärä toistuvassa kahden nopan heitossa ennen ensimmäistä kutosparia.

   (d) X on värisokeiden lukumäärä 10 hengen otoksessa takaisinpanolla 100 hengen populaatiossa, jossa on kolme värisokeata.

3. Satunnaismuuttuja X noudattaa geometrista jakaumaa parametrilla $p \in (0, 1)$. Näytä odotusarvon määritelmää ja geometrisen jakauman pistetodennäköisyysfunktiota käyttäen, että $E(X) = \frac{1-p}{p}$

4. Hajamielisellä herra H:lla on nipussaan n avainta, joista yksi sopii hänen oveensa. Herra H ei kuitenkaan muista, mikä. Olkoon X sen kerran järjestysluku, jolla ovi aukeaa. Laske X:n pistetodennäköisyysfunktio ja kertymäfunktio olettaen, että herra H valitsee avaimen umpimähkään ja

   (a) muistaa mitä avaimia on jo kokeillut,

   (b) ei muista mitä avaimia hän on kokeillut.

5. Lautta kulkee salmen yli rannalta A rannalle B säännöllisesti 10 minuutin välein ja sille mahtuu 8 autoa. Oletetaan, että rannalle A ei jää yhtään autoa lautan lähtiessä. Oletetaan lisäksi, että seuraavan 10 minuutin aikana rantaan A saapuvien autojen lukumäärä on Poisson-jakautunut satunnaismuuttuja odotusarvolla 5. Millä todennäköisyydellä lautta tulee täyteen käydessään seuraavan kerran rannalla A?

6. Isovanhemmat lähettävät lapsenlapselleen kaksi lahjaa, joiden arvot ovat 100e ja 150e. Postin toimituksissa on viime aikoina ilmennyt ongelmia ja paketin matkalla katoamisen todennäköisyys on 0,05. Isovanhemmat pohtivat, lähettäisivätkö he lahjat yhtenä vai kahtena eri pakettina. Vertaile menetelmiä seuraavilla kriteereillä:

   (a) katoamisesta johtuvan tappion odotusarvo,

   (b) tn, että lapsenlapsi saa molemmat lahjat,

   (c) tn, että lapsenlapsi saa ainakin yhden lahjan.

7. Herrat A, B ja C heittävät noppaa vuorotellen järjestyksessä A, B, C, A, B, C. . . , kunnes joku saa kutosen ja voittaa pelin. Laske kunkin pelaajan voiton todennäköisyys.

8. Tavallista noppaa heitetään 4 kertaa. Olkoon X suurin esiintyneistä silmäluvuista. Määritä $E(X)$.

Luku 5

Jatkuvat satunnaismuuttujat

Edellisessä luvussa esittelimme satunnaismuuttujan käsitteen ja tarkastelimme diskreettejä satunnaismuuttujia. Tässä luvussa siirrymme käsittelemään *jatkuvia satunnaismuuttujia*, jotka eroavat diskreeteistä satunnaismuuttujista siten, että niiden arvojoukon alkioita ei voi luetella (arvojoukko ei ole numeroituva). Tämän takia jatkuvien jakaumien laskutoimitukset täytyy määritellä hieman eri tavalla kuin diskreeteillä satunnaismuuttujilla. Jatkuvilla satunnaismuuttujilla on kuitenkin paljon samoja ominaisuuksia kuin diskreeteillä satunnaismuuttujilla (mm. kertymäfunktio, odotusarvo ja varianssi käyttäytyvät hyvin samalla tavalla), eikä satunnaismuuttujan intuitiivinen idea sinänsä muuutu mitenkään.

Jos satunnaismuuttujan arvojoukko on ylinumeroituva, esimerkiksi koko reaalilukujen joukko $\mathbb{R}$ tai väli $[0, 1]$, pistetodennäköisyyksiä ei voida enää käyttää satunnaismuuttujan jakauman määrittämiseen (itse asiassa osoittautuu, että jatkuvan jakauman tapauksessa kaikkien yksittäisten pisteiden todennäköisyydet ovat nollia).

Pistetodennäköisyysfunktiota vastaava käsite jatkuvasti jakautuneelle satunnaismuuttujalle on *tiheysfunktio*. Summaamisen sijaan satunnaismuuttujan arvojoukon osajoukkojen todennäköisyydet saadaan integroimalla tiheysfunktiota. Määritellään ensin täsmällisesti jatkuvan jakauman ja tiheysfunktion käsitteet.

Määritelmä 5.1. *Jatkuva satunnaismuuttuja*[1] *ja tiheysfunktio (lyh. tf).*[2] Satunnaismuuttujalla X on jatkuva jakauma tiheysfunktiolla f, jos

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

kaikille $a, b \in \mathbb{R}, a < b$.[3]

[1] engl. *continuous random variable*

[2] engl. *probability density function, PDF*

[3] Oikeastaan kyseessä ei edes tarvitse olla väli (a, b), vaan yleisesti kaikille mitallisille joukoille B pätee

Jatkuvan jakauman tapauksessa siis kaikkien välien todennäköisyydet $P(X \in [a,b])$ saadaan integroimalla tiheysfunktiota tämän välin yli. Kuten pistetodennäköisyysfunktio, myös tiheysfunktio on kaikkialla epänegatiivinen, ja sen integraali koko reaaliakselin yli on yksi. Samoin myös jokainen nämä ehdot täyttävä funktio on jonkin jatkuvan satunnaismuuttujan tiheysfunktio:

Lause 5.2. *Tiheysfunktion ominaisuudet. Funktio $f : \mathbb{R} \to \mathbb{R}$ on jonkin jatkuvan satunnaismuuttujan tiheysfunktio, jos ja vain jos*

(i) $f(x) \geq 0 \quad$ *kaikille* $\quad x \in \mathbb{R}$,

(ii) f *on integroituva koko reaaliakselilla ja*

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Todistus. Harjoitustehtävä. □

Tarkastellaan seuraavaksi esimerkkiä jatkuvista jakaumista.

Esimerkki 5.3. Jatkuva tasajakauma ja eksponenttijakauma.

(i) Metro kulkee tasaisesti 5 minuutin välein. Opiskelija saapuu satunnaiseen aikaan metroasemalle. Mikä on tn, että hän joutuu odottamaan yli 3.5 minuuttia?

Mallinnetaan odotusaikaa satunnaismuuttujalla X. Koska metro saattaa tulla yhtä todennäköisesti koska tahansa seuraavan viiden minuutin aikana, noudattaa X *jatkuvaa tasajakaumaa* (5.1.1) tiheysfunktiolla

$$f(x) = \frac{1}{5}.$$

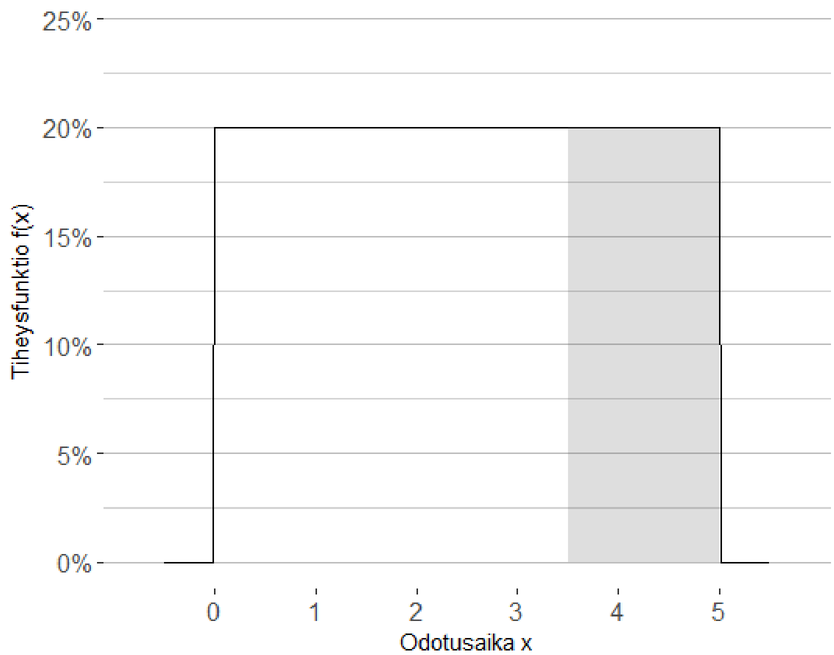
Nyt määritelmän 5.1 nojalla, todennäköisyys sille, että opiskelija joutuu odottamaan yli 3.5 minuuttia on

$$P(3.5 \leq X \leq 5) = \int_{3.5}^{5} \frac{1}{5}\,\mathrm{d}x = \Big/_{3.5}^{5} \frac{x}{5} = \frac{5}{5} - \frac{7}{10} = \frac{3}{10}.$$

Ero diskreettiin tasajakaumaan on se, että voimme mallintaa metron saapumisaikaa nyt mielivaltaisen tarkasti (emmekä ainoastaan esimerkiksi minuutin välein). Voimme ratkaista esimerkiksi, mikä on tn, että opiskelija joutuu odottamaan alle $0.3333\ldots$ minuuttia.

$P(X \in B) = \int_B f(x)\,\mathrm{d}x$

Myös jatkuvien jakaumien tiheysfunktioita havainnollistetaan kuvaajien avulla. Alla välillä [0,5] jatkuvan tasajakauman kuvaaja. Graafinen esitys metroesimerkille saadaan värittämällä tasajakauman kuvaajan alapuolinen alue väliltä [3.5, 5], jolloin väritetyn alueen pinta-ala on $\frac{3}{10}$ kuvaajan ja x-akselin väliin jäävän alueen pinta-alasta.



Kuva 5.1: Tasajakautuneen satunnaismuuttujan X tiheysfunktion kuvaaja.

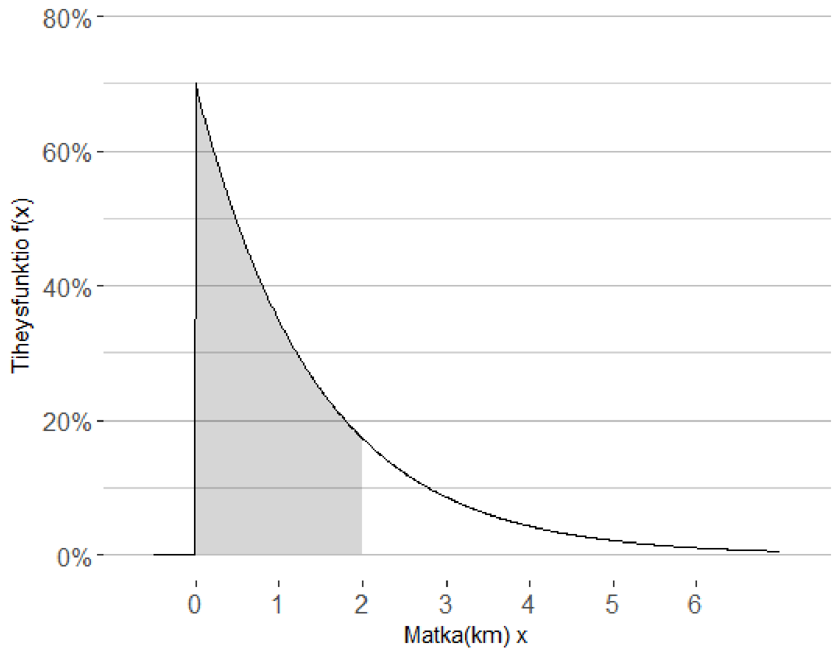
(ii) Auto ajaa tiellä, jossa tuulilasiin osuu hyönteinen keskimäärin 0.7 kilometrin välein. Mikä on tn, että seuraava hyönteinen osuu tuulilasiin seuraavien kahden kilometrin aikana?

Mallinnetaan matkaa hyönteisten osumien välillä satunnaismuuttujalla Z. Tällöin voidaan olettaa, että Z noudattaa *eksponenttijakaumaa* (joka esitellään myöhemmin tässä luvussa) tiheysfunktiolla

$$f(z) = 0.7e^{-0.7z}.$$

Nyt voimme ratkaista kysytyn todennäköisyyden integraalilla

$$P(Z \leq 2) = \int_0^2 0.7e^{-0.7z}\,\mathrm{d}z = \Big/_{0}^{2} -e^{-0.7z} = -e^{-1.4} - -e^{0} \approx 0.75$$



Kuva 5.2: Eksponenttijakautuneen satunnaismuuttujan Y tiheysfunktion kuvaaja.

Jatkuvia jakaumia käsitellään monesti kertymäfunktion avulla. Jatkuvan jakauman kertymäfunktio toimii täysin vastaavasti kuin diskreetin jakauman kertymäfunktio. Kertymäfunktio on määritelty jokaiselle reaaliluvulle $x \in \mathbb{R}$ ja kertoo, kuinka suuri osa satunnaismuuttujan 'todennäköisyysmassasta' on arvon x alapuolella, eli kuinka todennäköistä on, että satunnaismuuttuja saavuttaa korkeintaan arvon x.

Muistetaan. että määritelmän 4.12 mukaisesti satunnaismuuttujan X kertymäfunktio $F(x) = P(X \leq x)$ kaikilla $x \in \mathbb{R}$. Diskreeteille jakaumille kertymäfunktio saatiin ptnf:n summana. Vastaavasti jatkuvalle jakaumalle kertymäfunktio saadaan integroimalla tiheysfunktiota (ja kääntäen tiheysfunktio saadaan derivoimalla kertymäfunktiota):

Lause 5.4. *Jatkuvan satunnaismuuttujan kertymäfunktio. Jatkuvasti jakautuneelle satunnaismuuttujalle X (tiheysfunktiolla f) :*

(i) Kertymäfunktio F saadaan integroimalla tiheysfunktiota:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \quad \textit{kaikille} \quad x \in \mathbb{R}.$$

(ii) Toisaalta tiheysfunktiolle f pätee

$$f(x) = F'(x)$$

kaikissa f:n jatkuvuuspisteissä.

Seuraava lause osoittaa, että jatkuvalle jakaumalle kaikki pistetodennäköisyydet tosiaan ovat nollia. Tämän takia ei ole merkitystä, otetaanko välien päätepisteet mukaan tarkasteltaessa välien todennäköisyyksiä; ts. ei ole merkitystä, ovatko tarkasteltavat välit avoimia, puoliavoimia, vai suljettuja.

Lause 5.5. *Jatkuvan jakauman pistetodennäköisyydet. Jatkuvalle satunnaismuuttujalle X (tiheysfunktiolla f) pätee:*

(i) $P(X = x) = 0$ *kaikille* $x \in \mathbb{R}$,

(ii) $P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$ *kaikille* $a < b$.

Todistus. Ensimmäinen kohta seuraa jatkuvan satunnaismuuttujan määritelmästä, kun asetetaan $a = x$, ja annetaan b:n lähestyä x:ää oikealta. Toinen kohta seuraa todennäköisyyden additiivisuudesta ja ensimmäisestä kohdasta. $\square$

Huomautus 5.6. Kaikille $a < b$ tapahtuma, että jatkuva satunnaismuuttuja X on pienempi tai yhtä suuri kuin b, voidaan esittää yhdisteenä erillisistä tapahtumista:

$$(X \leq b) = (X < a) \cup (a \leq X \leq b).$$

Siten todennäköisyyden additiivisuuden ja lauseen 5.5 nojalla

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X < a) \\ &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a). \end{aligned}$$

Välien todennäköisyydet voidaan siis ilmaista erotuksena kertymäfunktion arvoista välin päätepisteissä. Jos muistellaan vielä esimerkkiä 5.3, huomataan, että itse asiassa integroidessamme satunnaismuuttujien tiheysfunktioita laskimmekin niiden kertymäfunktioiden arvoja.

5.1 Esimerkkejä jatkuvista jakaumista

Esitellään seuraavaksi muutamia tärkeimpiä ja yleisesti käytettyjä jatkuvia jakaumia.

5.1.1 Jatkuva tasajakauma

Määritelmä 5.7. Satunnaismuuttuja X noudattaa välin (a, b) *tasajakaumaa*[4] , jos sen tf on

$$f(x) = \frac{1}{b-a} \quad \text{kaikille} \quad x \in (a, b),$$

ja $f(x) = 0$, kun $x \leq a$ tai $x \geq b$. Tällöin merkitään $X \sim \mathrm{Tas}(a, b)$.

Kertymäfunktio saadaan integroimalla tiheysfunktiota. Kun $x \in (a, b)$:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{a}^{x} \frac{1}{b-a}\,\mathrm{d}t = \frac{x-a}{b-a}.$$

Siten tasajakautuneen satunnaismuuttujan kertymäfunktio on

$$F(x) = \begin{cases} 0, & \text{kun} \quad x \leq a \\ \frac{x-a}{b-a}, & \text{kun} \quad a < x < b \\ 1, & \text{kun} \quad x \geq b. \end{cases}$$

Esimerkki 5.8. Generoidaan satunnaisluku väliltä $(0, 1000)$. Millä todennäköisyydellä luku on väliltä $[500.00, 600.00]$?

Kyseessä on satunnaisluku, jolloin kaikki välin luvut ovat periaatteessa yhtä todennäköisiä. Tilannetta voidaan siis kuvata satunnaismuuttujalla $X \sim \mathrm{Tas}(0, 1000)$. Kysytty todennäköisyys saadaan erotuksena tasajakauman kertymäfunktion arvoista välin päätepisteissä:

$$P(500.00 \leq X \leq 600.00) = F(600.00) - F(500.00) = \frac{600}{1000} - \frac{500}{1000} = \frac{1}{10}.$$

5.1.2 Eksponenttijakauma

Määritelmä 5.9. Satunnaismuuttuja X noudattaa *eksponenttijakaumaa*[5] parametrilla $\lambda > 0$, jos sillä on jatkuva jakauma tiheysfunktiolla

$$f(x) = \lambda e^{-\lambda x} \quad \text{kaikille} \quad x > 0.$$

Tällöin merkitään $X \sim \mathrm{Exp}(\lambda)$.

Eksponenttijakauman kertymäfunktio saadaan integroimalla tiheysfunktiota:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{0}^{x} \lambda e^{-\lambda t}\,\mathrm{d}t = - \Big/_{0}^{x} e^{-\lambda t} = 1 - e^{-\lambda x},$$

kun $x > 0$ ja 0 muuten.

[4] engl. *(continuous) uniform distribution*
[5] engl. *exponential distribution*

Jos tapahtumien esiintymistodennäköisyys aikaväliä tai tilayksikköä kohti on vakio, ja ne esiintyvät toisistaan riippumatta, näiden tapahtumien esiintymisten välistä aikaa tai etäisyyttä kuvaava satunnaismuuttuja noudattaa eksponenttijakaumaa. Jos tapahtumien määrä aikayksikköä kohti noudattaa Poisson-jakaumaa, niin niiden välinen aika noudattaa eksponenttijakaumaa.[6]

Eksponenttijakautuneella satunnaismuuttujalla voidaan siis kuvata odotusaikaa tai välimatkaa toisistaan riippumattomien tapahtumien, joiden sattumistodennäköisyys on vakio, välillä. Tällaisia eksponenttijakautuneita satunnaismuuttujia ovat esimerkiksi:

- odotusaika seuraavaan maaliin jalkapallo-ottelussa,
- liikenneonnettomuuksien tapahtumispaikkojen väli tieosuudella,
- onkijan odotusaika seuraavaan kalaan.

Esimerkki 5.10. Oletetaan, että hehkulampun kestoaika tunneissa noudattaa eksponenttijakaumaa parametrilla $\lambda = 0.001$. Millä todennäköisyydellä hehkulamppu palaa yli 2000 tuntia?

Merkitään lampun palamisaikaa satunnaismuuttujalla X, jolloin $X \sim \mathrm{Exp}(0.001)$. Kysytyn tapahtuman tn saadaan laskettua kertymäfunktion avulla komplementtitapahtuman todennäköisyyden kautta:

$$P(X > 2000) = 1 - P(X \leq 2000) = 1 - F(2000) = e^{-2} \approx 0.135.$$

Toinen mahdollisuus laskea todennäköisyys on integroida suoraan tiheysfunktiota:

$$P(X > 2000) = \int_{2000}^{\infty} \lambda e^{-\lambda t}\,\mathrm{d}t = - \Big/_{2000}^{\infty} e^{-\lambda t} = e^{-2000\lambda} = e^{-2}.$$

Huomautus 5.11. Eksponenttijakaumaa noudattavalla satunnaismuuttujalla X on seuraava 'muistinmenetysominaisuus':

$$P(X > t + h \quad | \quad X > t) = P(X > h).$$

Ts. esimerkiksi laitteella, jonka toiminta-aika on eksponenttijakautunut todennäköisyys sille, että laite toimii ajan h kuluttua, ei riipu ajanhetkestä t, jolloin tilannetta tarkastelemme. Laitteen todennäköisyys pysyä ehjänä seuraavaan päivään saakka, on sama jokaisena päivänä.

[6] Tällaista prosessia, jonka tuottamien satunnaisten tapahtumien määrä aikayksikköä kohti noudattaa Poisson-jakaumaa, ja niiden väliaika eksponenttijakaumaa, kutsutaan Poisson-prosessiksi.

5.1.3 Normaalijakauma

Määritelmä 5.12. Jatkuva satunnaismuuttuja X noudattaa *normaalijakaumaa*[7] parametreilla $\mu \in \mathbb{R}$ ja $\sigma^2 > 0$, jos sen tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Tällöin merkitään $X \sim N(\mu, \sigma^2)$.

Normaalijakauman parametrit kertovat jakauman tiheysfunktion sijainnin ja leveyden. Näitä parametreja sanotaan usein *odotusarvoparametriksi* ja *varianssiparametriksi*[8].

Jos satunnaismuuttuja X noudattaa jakaumaa $N(0, 1)$, eli normaalijakaumaa parametreilla $\mu = 0$ ja $\sigma^2 = 1$, sanotaan että X noudattaa *standardinormaalijakaumaa.* Tällöin X:n tiheysfunktio yksinkertaistuu muotoon:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Huomautus 5.13. Normaalijakautunut satunnaismuuttuja $X \sim N(\mu, \sigma^2)$ voidaan aina muuntaa standardinormaalijakaumaa noudattavaksi satunnaismuuttujaksi vähentämällä siitä odotusarvoparametri μ ja jakamalla se keskihajonnalla σ. Satunnaismuuttujalle

$$Z = \frac{X - \mu}{\sigma}$$

siis pätee $Z \sim N(0, 1)$. Tätä muunnosta kutsutaan standardoinniksi.

Esitetään seuraavaksi esimerkki normaalijakautuneen satunnaismuuttujan todennäköisyyden laskemisesta standardoinnin avulla.

Esimerkki 5.14. Eräässä väestössä miespuolisten henkilöiden pituus noudattaa normaalijakaumaa parametreilla $\mu = 178$ ja $\sigma^2 = 25$ (cm). Mallinnetaan kyseisen väestön miesten pituutta satunnaismuuttujalla X. Mikä on todennäköisyys sille, että väestöön kuuluvan miehen pituus on 168 ja 188 cm välillä? Entä tn sille, että mies on yli 190 cm pitkä?

[7]engl. *normal distribution*

[8]Muistetaan, että varianssi kuvaa jakauman hajontaa. Varianssin ominaisuuksia käsitellään lisää seuraavassa luvussa.

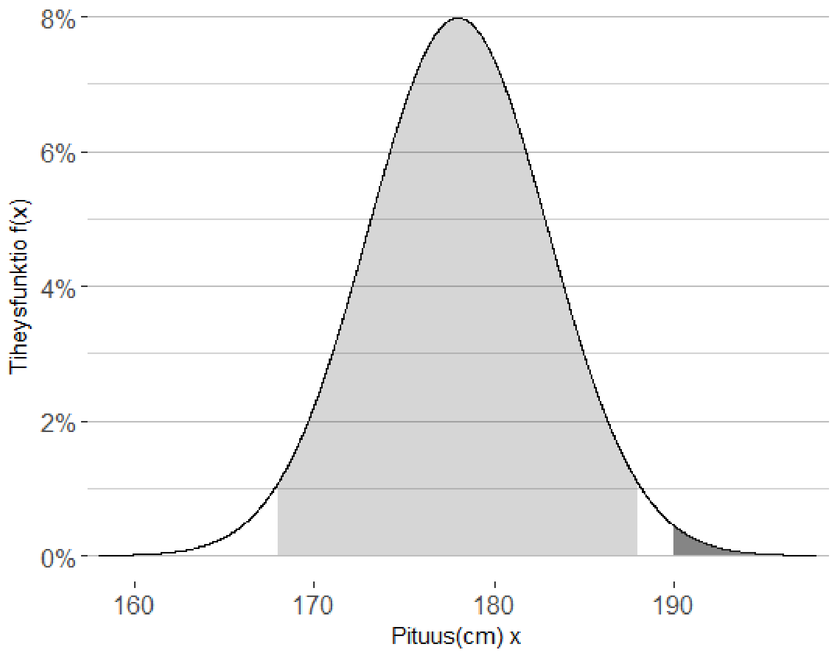
(a) Ratkaistaan todennäköisyys $P(168 \leq X \leq 188)$ sm:n X standardoinnilla Z vähentämällä X:stä jakauman odotusarvo ja jakamalla se jakauman keskihajonnalla.

$$\begin{aligned}
P(168 \leq X \leq 188) &= P\Big(\frac{168-178}{5} \leq \frac{X-178}{5} \leq \frac{188-178}{5}\Big) \\
&= P(-2 \leq Z \leq 2) \\
&= \Phi(2) - \Phi(-2) = \\
&= \Phi(2) - (1 - \Phi(2)) \\
&= 2\Phi(2) - 1 = 2 \cdot 0.9772499\ldots - 1 \\
&\approx 0.954.
\end{aligned}$$

Φ-symbolilla viitataan standardinormaalijakauman kertymäfunktioon, jonka arvoja voi laskea esimerkiksi R-ohjelman funktiolla `pnorm()` tai katsoa monista lähteistä tai Internetistä löytyvistä taulukoista.

(b) Edelliskohdan tapaan:

$$\begin{aligned}
P(190 < X) = 1 - P(X \leq 190) &= P\Big(\frac{X-178}{5} \leq \frac{190-178}{5}\Big) \\
&= 1 - P(Z \leq 2.4) = 1 - \Phi(2.4) = 1 - 0.9918025\ldots \\
&\approx 0.008.
\end{aligned}$$



Kuva 5.3: Normaalijakautuneen satunnaismuuttujan X tiheysfunktion kuvaaja. Kohdan (a) tilannetta kuvaava todennäköisyysmassa varjostettu vaalealla ja kohdan (b) tummalla.

Normaalijakauman keskeinen rooli todennäköisyyslaskennassa johtuu siitä, että riippumattomien satunnaismuuttujien summan (ja siten myös keskiarvon) jakauma lähestyy normaalijakaumaa otoskoon kasvaessa. Tähän *keskeisenä raja-arvolauseena* tunnettuun tulokseen palataan kurssin loppupuolella.

5.2 Jatkuvan satunnaismuuttujan odotusarvo

Muistetaan, että satunnaismuuttujan odotusarvoa voi ajatella eräänlaisena keskiarvona satunnaismuuttujan arvoista. Edellisluvussa esitelty odotusarvon määritelmä toimii vain diskreeteillä satunnaismuuttujilla, joten esitellään tässä kappaleessa, miten jatkuvan satunnaismuuttujan odotusarvo määritellään.

Määritelmä 5.15. *Jatkuvan satunnaismuuttujan odotusarvo.* Olkoon X jatkuva satunnaismuuttuja tiheysfunktiolla f. Sen odotusarvo on

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x,$$

mikäli kyseinen integraali suppenee itseisesti, eli

$$\int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x < \infty.$$

Jos integraali ei suppene itseisesti, sanotaan että X:llä ei ole odotusarvoa.

Esitetään seuraavaksi esimerkit odotusarvon laskemisesta jatkuville jakaumille.

Esimerkki 5.16. Olkoon X tasajakaumaa $\mathrm{Tas}(a, b)$ noudattava satunnaismuuttuja. Tällöin X:n odotusarvo on

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_{a}^{b} x \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{2(b-a)} \Big/_{\!a}^{\,b} x^2 = \frac{1}{2}(a+b).$$

Välin (a, b) tasajakaumaa noudattavan sm:n odotusarvo sijaitsee siis aina tämän välin puolessavälissä. Tämä sopii hyvin yhteen oletusarvon massakeskipistetulkinnan kanssa: tasajakaumassa todennäköisyysmassa on jakautunut tasaisesti koko välille, joten painopiste on välin puolessa välissä.

Esimerkki 5.17. Oletetaan, että hehkulampun kestoaikaa tunteina kuvaava satunnaismuuttuja X noudattaa eksponenttijakaumaa parametrilla $\lambda = 0.001$, eli $X \sim \mathrm{Exp}(0.001)$. Mikä on lampun kestoajan odotusarvo?

Lasketaan ensin eksponenttijakaumaa noudattavan sm:n odotusarvo yleisessä tapauksessa osittaisintegroinnilla:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_{0}^{\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x \\ &= \Big/_{0}^{\infty} x(-e^{-\lambda x}) - \int_{0}^{\infty} 1 \cdot (-e^{-\lambda x})\,\mathrm{d}x \\ &= 0 - \frac{1}{\lambda} \Big/_{0}^{\infty} e^{-\lambda x} = \frac{1}{\lambda}. \end{aligned}$$

Siten hehkulampun kestoiän odotusarvo on $E(X) = \frac{1}{\lambda} = 1000$ tuntia.

Olemme viimeisissä luvuissa käsitelleet diskreettejä ja jatkuvia satunnaismuuttujia ja esitelleet niiden eroja. Yleisimmät jakaumat, joita kurssilla käytämme, ovat joko diskreettejä tai jatkuvia. Jakauman ei kuitenkaan tarvitse olla yksinomaan diskreetti tai jatkuva, kuten seuraavassa huomautuksessa todetaan.

Huomautus 5.18. Satunnaismuuttujan jakauman ei välttämättä tarvitse olla diskreetti tai jatkuva, vaan satunnaismuuttuja voi noudattaa myös ns. *sekatyypin* jakaumaa. Esimerkki tällaisesta satunnaismuuttujasta on hehkulampun kestoikä, jos oletetaan että kymmenesosa lampuista on viallisia, eli niiden kestoikä on nolla, ja loppujen lamppujen kestoikä noudattaa eksponenttijakaumaa parametrilla 0.001. Tämän jakauman kertymäfunktio on

$$F(x) = \begin{cases} 0, & \text{kun} \quad x < 0 \\ \frac{1}{10}, & \text{kun} \quad x = 0 \\ \frac{1}{10} + \frac{9}{10}(1 - e^{-0.001x}), & \text{kun} \quad x > 0. \end{cases}$$

5.3 Satunnaismuuttujan muunnos

Satunnaismuuttujan muunnoksella tarkoitetaan satunnaismuuttujasta jonkin funktion avulla saatua toista satunnaismuuttujaa. Seuraavassa lauseessa esitellään muodollisemmin, mitä tällä tarkoitetaan.

Lause 5.19. *Satunnaismuuttujan muunnos. Jos $g : \mathbb{R} \to \mathbb{R}$ on funktio ja X on satunnaismuuttuja, myös $Y = g(X)$ on satunnaismuuttuja.*

Merkintä $Y = g(X)$ tarkoittaa tässä yhteydessä, että $Y(\omega) = g(X(\omega))$, missä $\omega \in \Omega$. Muunnos Y on siis yhdistetty funktio $(g \circ X)(\omega)$.

Oikeastaan olemme jo aiemmin tässä luvussa käsitelleet satunnaismuuttujan muunnosta normaalijakauman yhteydessä. Esimerkissä 5.14 johdimme normaalijakautuneesta satunnaismuuttujasta X funktion $\frac{x-178}{5}$ avulla standardoidun muunnoksen $Z \sim N(0,1)$. Vielä yksinkertaisempi muunnos sm:sta X voisi olla esimerkiksi $Z_2 = 2X$ tai $Z_3 = \sqrt{X}$.

Perehdymme myöhemmin kappaleessa 8.1 tarkemmin satunnaismuuttujien muunnoksiin ja niiden pistetodennäköisyys- tiheys- ja kertymäfunktioiden laskemiseen. Toistaiseksi riittää, että lukija on kuullut muunnoksen käsitteestä ja hahmottaa suurin piirtein, mitä se tarkottaa.

5.4 Harjoitustehtäviä

1. Paikallisjunan pitäisi saapua Helsinkiin klo 13:03. Junan myöhästymisaika on satunnaismuuttuja, joka jakautuu tasaisesti yli välin (-2, 8) (minuuttia). Laske todennäköisyys, että juna saapuu

   (a) ajoissa,

   (b) myöhemmin kuin 13:05:20,

   (c) korkeintaan 1,5 minuuttia myöhässä.

2. Määritä $E(X)$, kun X:llä on jatkuva jakauma tiheysfunktiolla f, missä

   (a) $f(x) = \frac{1}{2}e^{|-x|}$, missä $x \in \mathbb{R}$,

   (b) $f(x) = \frac{8}{x^3}$, missä $x > 2$,

   (c) $f(x) = xe^{-\frac{1}{2}x^2}$, missä $x > 0$.

3. Helsingin vuosittainen sademäärä noudattaa normaalijakaumaa parametreilla $\mu = 655$ ja $\sigma^2 = 900$ (mm). Laske todennäköisyys, että seuraavan vuoden sademäärä ylittää 800 mm.

4. Jatkuvalla satunnaismuuttujalla X on tiheysfunktio $f(x) = C\cos(x)$, kun $\frac{-\pi}{2} \leq x \leq \frac{\pi}{2}$, ja $f(x) = 0$ muulloin.

   (a) Selvitä vakion C arvo.

   (b) Laske $P(X < 1)$.

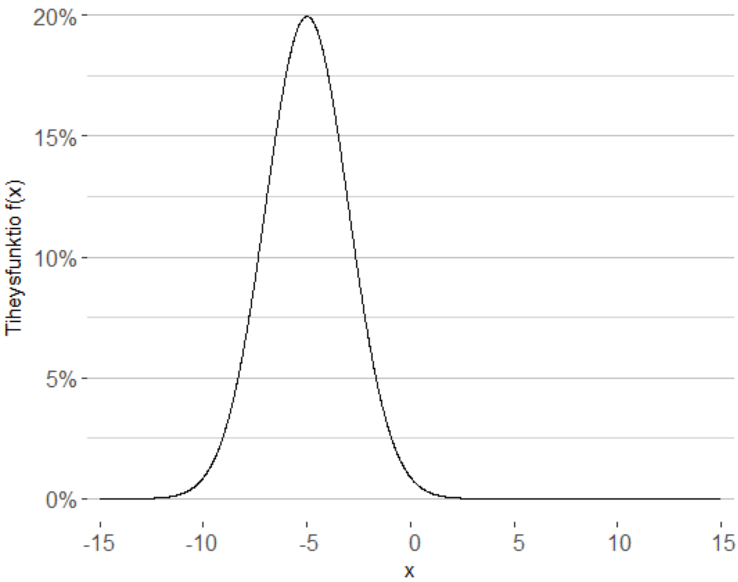
5. Tehdas valmistaa jääkaappeja, joiden kestoaika kulutuksessa on jakaumaltaan $Exp(\lambda)$. Tehtaassa voidaan säädellä parametria λ. Mikä λ:n tulisi olla, jotta todennäköisyys, että kestoaika olisi korkeintaan 5 vuotta olisi vähintään 0.5.

Luku 6

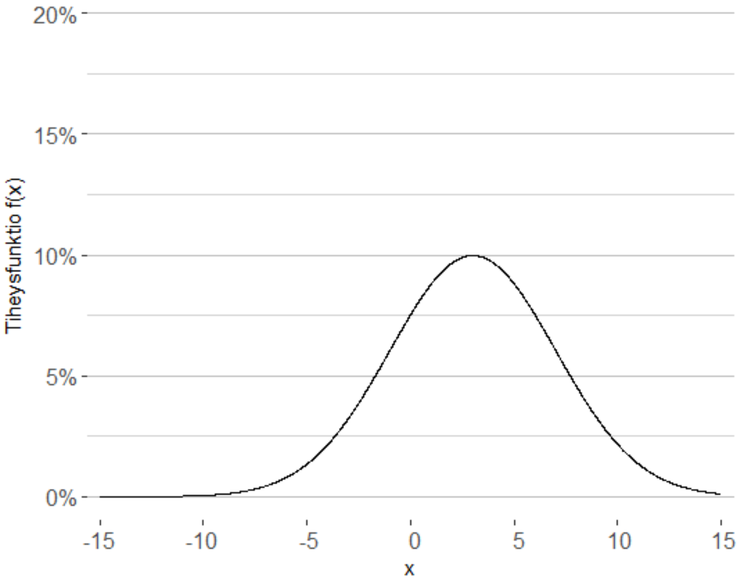
Odotusarvon ominaisuuksia

Olemme jo aiemmissa luvuissa käsitelleet odotusarvon laskemista diskreeteille ja jatkuville satunnaismuuttujille. Tässä kappaleessa tutustutaan tarkemmin odotusarvon ominaisuuksiin ja esitellään, miksi odotusarvon käsite on todennäköisyyslaskennassa käytännöllinen. Lisäksi tutustumme varianssin ja kovarianssin käsitteisiin. Varianssi on satunnaismuuttujan hajonnan mitta ja kovarianssi kahden satunnaismuuttujan välisen riippuvuuden mitta.

Havainnollistetaan vielä aluksi satunnaismuuttujan jakauman, odotusarvon ja varianssin käsitteitä jakaumien kuvaajien avulla. Kuten olemme aiemmissa luvuissa oppineet, satunnaismuuttujan jakauma määrää kuvaajan muodon, odotusarvo sijainnin (x-akselilla) ja varianssi leveyden.



Kuva 6.1: Normaalijakautuneen satunnaismuuttujan X tiheysfunktion kuvaaja. Odotusarvo $\mu = -5$, varianssi $\sigma^2 = 2^2 = 4$.



Kuva 6.2: Normaalijakautuneen satunnaismuuttujan X tiheysfunktion kuvaaja. Odotusarvo $\mu = 3$, varianssi $\sigma^2 = 4^2 = 16$.

6.1 Odotusarvo

Muistetaan vielä, että diskreetin satunnaismuuttujan odotusarvo määritellään summana

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i), \text{ missä } f \text{ on sm:n } X \text{ ptnf}$$

ja jatkuvan satunnaismuuttujan odotusarvo integraalina

$$E(X) = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x, \text{ missä } f \text{ on sm:n } X \text{ tf,}$$

mikäli kyseinen summa ja integraali suppenevat itseisesti. Odotusarvolla on seuraavat ominaisuudet, joita lukija voi yrittää itse osoittaa diskreetin tai jatkuvan satunnaismuuttujan tapauksessa.

Lause 6.1. *Odotusarvon ominaisuuksia. Satunnaismuuttujille X ja Y, joiden odotusarvot ovat olemassa, pätee:*

(i) *(Positiivisuus) Jos* $X \geq 0$*, niin* $E(X) \geq 0$.

(ii) *(Lineaarisuus)* $E(aX + bY) = aE(X) + bE(Y)$ *kaikille* $a, b \in \mathbb{R}$.

(iii) *(Monotonisuus) Jos* $X \leq Y$[1]*, niin* $E(X) \leq E(Y)$.

(iv) *Jos* $X \geq 0$ *ja* $E(X) = 0$*, niin sm* X *on vakio arvolla 0, eli* $P(X = 0) = 1$.

(v) *Jos* X*:llä ja* Y*:llä on sama jakauma, niin* $E(X) = E(Y)$.

(vi) *(Vakion odotusarvo) Jos vakio* $a \in \mathbb{R}$*, niin* $E(a) = a$.

(vii) *Jos* $X \perp\!\!\!\perp Y$*, niin satunnaismuuttujalla* XY *on odotusarvo* $E(XY) = E(X)E(Y)$.

Pureudutaan vielä hieman tarkemmin edellisen lauseen 6.1 kohtaan (*ii*). Suoraan kohdan (*ii*) ominaisuudesta seuraa todennäköisyyslaskennassa tärkeä tulos, jonka mukaan satunnaismuuttujien **odotusarvojen summa on yhtä kuin niiden summien odotusarvo**. Tätä tulosta kutsutaan odotusarvon lineaarisuudeksi. Esitetään ensin tulos ja sen todistus muodollisemmin, ja havainnollistetaan odotusarvon lineaarisuutta tämän jälkeen esimerkin avulla.

Lause 6.2. *Odotusarvon lineaarisuus.*[2] *Olkoot* $X_1, X_2, \ldots, X_n$ *satunnaismuuttujia, joilla on äärelliset odotusarvot, eli* $E(X_i) < \infty$ *kaikilla* $i = 1, \ldots, n$*. Tällöin satunnaismuuttujien summan odotusarvo saadaan niiden odotusarvojen summana:*

$$E\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} E(X_i).$$

Todistus. Todistetaan lause induktion avulla.

Kun $n = 2$, niin lauseen 6.1 kohdan (*ii*) nojalla

$$E(X_1 + X_2) = E(X_1) + E(X_2).$$

Oletetaan, että kun $n = k$, niin

$$E\Big(\sum_{i=1}^{k} X_i\Big) = \sum_{i=1}^{k} E(X_i).$$

Nyt, kun $n = k + 1$, niin

$$E\Big(\sum_{i=1}^{k+1} X_i\Big) = E\Big(\sum_{i=1}^{k} X_i + X_{k+1}\Big) = \sum_{i=1}^{k} E(X_i) + E(X_{k+1}) = \sum_{i=1}^{k+1} E(X_i).$$

Toinen yhtäsuuruus seuraa induktio-oletuksesta ja lauseen 6.1 kohdasta (*ii*). □

[1] Sanomme, että $X \leq Y$ mikäli $X(\omega) \leq Y(\omega)$ kaikilla $\omega \in \Omega$

[2] engl. *linearity of expectation*

Siirrytään seuraavaksi esimerkkiin. Tässä tapauksessa satunnaismuuttujien perusjoukko on äärellinen ja satunnaismuuttujia on vain kaksi, mutta kuten lause 6.2 osoittaa, tulos pätee yleisesti useammillekin satunnaismuuttujille, jotka voivat olla myös jatkuvia.

Esimerkki 6.3. Kasino tarjoaa seuraavanlaista peliä: 10 euron panoksella pelaaja saa heittää kolikkoa 3 kertaa. Jokaisesta kruunasta pelaaja voittaa 5 euroa. Lisäksi pelaaja saa heittää noppaa ja mikäli silmäluku on 6, pelaaja voittaa 10 euroa. Kannattaako peliä pelata?

Mallinnetaan tilannetta satunnaismuuttujien Xja Y avulla. Olkoon sm X yksittäisellä kolikonheitolla voitettu rahamäärä ja sm Y nopanheitolla voitettu rahamäärä. Selvästi $E(X) = 5 \cdot \frac{1}{2}$ ja $E(Y) = 10 \cdot \frac{1}{6}$. Nyt odotusarvo pelaajan voittamalle rahasummalle on odotusarvon lineaarisuuden nojalla

$$E(3X + Y) = 3E(X) + E(Y) = 3 \cdot \frac{5}{2} + \frac{10}{6} = \frac{45}{6} + \frac{10}{6} = \frac{55}{6}.$$

Kun vähennetään voiton odotusarvosta osallistumismaksu, huomataan että $\frac{55}{6} - 10 = -\frac{5}{6}$, joten pelin odotusarvo on negatiivinen, eli peliä ei kannata pitkällä tähtäimellä pelata.

Toistaiseksi olemme käsitelleet vain tilanteita, joissa satunnaismuuttujilla todella on odotusarvo. On kuitenkin hyvä muistaa, että näin ei aina välttämättä ole, vaikka asian hahmottaminen intuitiivisesti voi olla hankalaa. Tarkastellaan seuraavaksi esimerkkejä diskreetistä ja jatkuvasta jakaumasta, joilla ei ole odotusarvoa.

Esimerkki 6.4. Pietarin paradoksi ja *Cauchy-jakauma.*

(i) Pietarin paradoksi. Tarkastellaan seuraavaa peliä: heitetään kolikkoa, kunnes saadaan ensimmäinen kruuna. Pelaaja voittaa 2^k dukaattia, missä k on heittojen määrä. Kuinka paljon tästä pelistä kannattaa maksaa?

Merkitään pelistä saatavaa voittoa sm:lla X, jonka arvojoukko on $\{2, 4, 8, \dots\}$. Koska tn saada ensimmäinen kruuna k:nnella heitolla on $\frac{1}{2^k}$, X:n pistetodennäköisyysfunktio on

$$f(x) = \frac{1}{x} \quad \text{kaikille} \quad x \in \{2, 4, 8 \dots\}.$$

Voiton odotusarvoksi saadaan:

$$E(X) = \sum_{k=1}^{\infty} x_k f(x_k) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{i=1}^{\infty} 1 = \infty.$$

Siten pelistä kannattaa periaatteessa maksaa mielivaltaisen paljon!

Tämä perustuu siihen, että pelaajan panos on rajoitettu, mutta kasinon (= pelin tarjoajan) maksama voittosumma on periaatteessa rajoittamaton. 'Paradoksi' ratkeaa, kun oletetaan, että kasinon maksamalla voittosummalla on jokin yläraja.

(ii) Cauchy-jakauma. Satunnaismuuttuja X noudattattaa (standardi-)Cauchy-jakaumaa, jos sen tiheysfunktio on

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \quad \text{kaikille} \quad x \in \mathbb{R}.$$

Tällöin merkitään $X \sim \text{Cauchy}(0,1)$. Cauchy-jakaumaa käytetään yleisesti esimerkkinä jakaumasta, jolla ei ole äärellistä odotusarvoa eikä varianssia. Odotusarvolle tämä voidaan todeta integroimalla. Jakauman symmetrisyyden nojalla:

$$\int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x = \frac{1}{\pi}\int_{0}^{\infty} \frac{2x}{1+x^2}\,\mathrm{d}x = \Big/_{0}^{\infty} \ln(1+x^2) = \infty,$$

Tarkastellaan vielä hetki satunnaismuuttujien muunnosten odotusarvoja. Lineaarisen muunnoksen odotusarvoa käsittelimme jo esimerkissä 6.3. Yleisesti muunnoksen odotusarvo voidaan ajatella muunnoksen arvojen painotettuna keskiarvona, missä painoina toimivat kunkin arvon todennäköisyydet. Diskreetin satunnaismuuttujan muunnokselle odotusarvo saadaan summana, jossa muunnoksen arvoja $g(x)$ kerrotaan vastaavilla todennäköisyyksillä $f(x)$, ja jatkuvan satunnaismuuttujan muunnokselle vastaavana integraalina.

Lause 6.5. *Diskreetin satunnaismuuttujan muunnoksen odotusarvo. Jos X on diskreetti satunnaismuuttuja, jonka ptnf on f, ja arvojoukko $X(\Omega) = \{x_1, x_2, \dots\}$, ja $Y = g(x)$ sen muunnos, niin sm:n Y odotusarvo on*

$$E(Y) = E(g(X)) = \sum_i g(x_i) f(x_i),$$

mikäli kyseinen sarja suppenee itseisesti.

Ja vastaavasti jatkuvalle satunnaismuuttujalle:

Lause 6.6. *Jatkuvan satunnaismuuttujan muunnoksen odotusarvo. Jos X on jatkuva satunnaismuuttuja tf:lla f ja $Y = g(X)$ sen muunnos, niin sm:n Y odotusarvo on*

$$E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x,$$

mikäli kyseinen integraali suppenee itseisesti.

Havainnollistetaan muunnoksen odotusarvon laskemista seuraavan esimerkin avulla.

Esimerkki 6.7. Diskreetin ja jatkuvan sm:n muunnoksen odotusarvo.

(i) Eräässä nopanheittopelissä pelaajan voittosumma määräytyy nopan silmäluvun kautta siten, että voittosumma on nopan silmäluku korotettuna toiseen potenssiin. Kannattaako peliä pelata, jos osallistumismaksu on 15 euroa?

   Olkoon satunnaismuuttuja X nopan silmäluku, jolloin pistetodennäköisyysfunktio $f(x) = \frac{1}{6}$ kaikille $x \in \{1, \ldots, 6\}$. Olkoon satunnaismuuttuja Y sm:n X muunnos $Y = X^2$. Tällöin

$$E(Y) = E(X^2) = \sum_{i=1}^{6} i^2 \frac{1}{6} = \frac{91}{6} \approx 15,167.$$

   Toisin sanoen pelin pelaaminen on pitkällä aikavälillä kannattavaa, sillä voiton odotusarvo on osallistumismaksua suurempi.

(ii) Satunnaismuuttuja X noudattaa jatkuvaa tasajakaumaa välillä [0,5]. Laske $E(\sqrt{X})$.

   Satunnaismuuttujan X tiheysfunktio $f(x) = \frac{1}{5}$. Nyt

$$\begin{aligned} E(\sqrt{X}) &= \int_{-\infty}^{\infty} \sqrt{x} f(x) = \int_{-\infty}^{0} \sqrt{x} \cdot 0 + \int_{0}^{5} \sqrt{x}\frac{1}{5} + \int_{5}^{\infty} \sqrt{x} \cdot 0 \\ &= \int_{0}^{5} \sqrt{x}\frac{1}{5} = \Big/_{0}^{5} \frac{2x^{3/2}}{15} = \frac{2 \cdot \sqrt{5^3}}{15} - \frac{2 \cdot \sqrt{0^3}}{15} \\ &\approx 1.49 \end{aligned}$$

Huomaa, että koska esimerkin muunnokset eivät ole lineaarisia, niin $E(g(X))$ ja $g(E(X))$ eivät ole sama asia: Kohdassa (i):

$$E(X^2) = \frac{91}{6} \neq 3.5^2 = E(X)^2$$

ja kohdassa (ii):

$$E\big(\sqrt{X}\big) = \frac{2 \cdot \sqrt{5^3}}{15} \neq \sqrt{2.5} = \sqrt{E(X)}.$$

6.2 Varianssi ja keskihajonta

Olemme tähän asti käsitelleet jakaumien tunnusluvuista ainoastaan odotusarvoa. Siirrytään seuraavaksi toiseen tärkeään tunnuslukuun, varianssiin, ja sen johdannaiseen, keskihajontaan. Varianssi ja keskihajonta kuvaavat satunnaismuuttujan jakauman *hajontaa*, eli sitä kuinka paljon satunnaismuuttujan arvot keskimäärin poikkeavat sen odotusarvosta. Varianssi on odotusarvo satunnaismuuttujan arvojen neliöidystä poikkemasta sm:n odotusarvosta ja keskihajonta taas varianssin neliöjuuri. Esitetään seuraavaksi varianssin ja keskihajonnan määritelmät.

Määritelmä 6.8. *Varianssi.* [3] Olkoon X satunnaismuuttuja, jonka odotusarvo $E(X)$ on olemassa ja äärellinen. X:n varianssi on

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right],$$

mikäli kyseinen odotusarvo on olemassa. Jos $E(X)$ on olemassa, mutta $E\left[(X - E(X))^2\right] = \infty$, sanotaan, että X:n varianssi on ääretön. Varianssille voidaan käyttää myös merkintöjä $D^2(X)$ tai σ^2 (tai σ_X^2, jos käsitellään useampia satunnaismuuttujia ja halutaan täsmentää, että kyseessä on nimenomaan X:n varianssi).

Määritelmä 6.9. *Keskihajonta.* [4] Satunnaismuuttujan X varianssin neliöjuurta

$$\sqrt{\mathrm{Var}(X)}$$

kutsutaan X:n keskihajonnaksi. Keskihajonnalle voidaan käyttää myös merkintöjä $D(X)$, σ tai σ_X.

Varianssin laskeminen määritelmän 6.8 avulla voi olla monesti haastavaa. Osoittautuu kuitenkin, että varianssin kaava voidaan esittää myös satunnaismuuttujan *toisen momentin*[5] avulla.

Määritelmä 6.10. *Toinen momentti.* Satunnaismuuttujan neliön odotusarvoa $E(X^2)$ kutsutaan sen toiseksi momentiksi. Diskreetille satunnaismuuttujalle X, jonka ptnf on f se voidaan laskea seuraavana summana:

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 f(x_i)$$

ja jatkuvalle satunnaismuuttujalle Y, jonka tf on g vastaavana integraalina:

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 g(y)\,\mathrm{d}y.$$

Nyt satunnaismuuttujan varianssi voidaan laskea seuraavalla kaavalla:

[3] engl. *variance.* Lisäksi lukijalle huomautuksena, että varianssi määritellään tämän määritelmän mukaisesti lähinnä matemaattisen mukavuuden vuoksi. Satunnaismuuttujan "jakauman leveyttä" eli keskimääräistä poikkeamaa odotusarvosta voi mitata myös esimerkiksi keskihajonnalla tai keskipoikkeamalla $E(|X - E(X)|)$, mutta varinssin ominaisuudet tekevät siitä ylivertaisen teoreettisessa tarkastelussa. Mikäli varianssin sijaan käytettäisi jotain muuta tunnuslukua kuvaamaan hajonnan astetta, ei tässä kappaleessa esitetyt laskusäännöt pitäisi paikkaansa.

[4] engl. *standard deviation*

[5] Jakaumien tunnuslukuja eli momentteja voidaan laskea momentit generoivien funktioiden, eli *momenttiemäfunktioiden* avulla. Aihetta käsitellään lisää myöhemmillä kursseilla.

Lause 6.11. *Jos satunnaismuuttujan toinen momentti $E(X^2)$ on olemassa, myös sen varianssi on äärellinen, ja saadaan kaavasta*

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

Todistus. Odotusarvon olemassaolo seuraa aina toisen momentin olemassaolosta, sillä

$$|x| \leq 1 + x^2 \quad \text{kaikille} \quad x \in \mathbb{R}.$$

Nyt väite seuraa odotusarvon lineaarisuudesta (Lauseen 6.1 toinen kohta):

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right] = E\left[X^2 - 2XE(X) + E(X)^2\right] = E(X^2) - [E(X)]^2.$$

□

Huomautus 6.12. Olethan varovainen merkintöjen kanssa! Usein merkitään $[E(X)]^2 = E(X)^2$, joten pidäthän huolta siitä, että olet aina perillä neliöitävästä termistä. Käytämme myös tässä materiaalissa jatkossa merkintää $E(X)^2$.

Tarkastellaan seuraavaksi esimerkkiä varianssin laskemisesta tasajakautuneelle satunnaismuuttujalle.

Esimerkki 6.13. Olkoon X välin (a, b) jatkuvaa tasajakaumaa noudattava satunnaismuuttuja, eli $X \sim \mathrm{Tas}(a, b)$, jolloin X:n odotusarvo on $E(X) = \frac{1}{2}(a + b)$. Lasketaan ensin X:n toinen momentti:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{1}{3(b-a)} \Big/_{a}^{b} x^3 = \frac{b^3 - a^3}{3(b-a)}.$$

Varianssi saadaan nyt laskettua lauseen 6.11 nojalla:

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{b^3 - a^3}{3(b-a)} - \left[\frac{1}{2}(a+b)\right]^2 = \frac{1}{12}(b-a)^2.$$

Laskemalla varianssi suoraan määritelmästä 6.8 oltaisiin tässä päädytty hyvin samankaltaisiin laskuihin.

Muistetaan vielä satunnaismuuttujien muunnoksen käsite, jota olemme käsitelleet viimeisimmissä luvuissa. Satunnaismuuttujan muunnoksen yksi erikoistapaus on lineaarinen muunnos, jonka varianssi saadaan seuraavalla kaavalla.

Lause 6.14. *Lineaarimuunnoksen varianssi. Jos satunnaismuuttujan X varianssi on olemassa ja $a, b \in \mathbb{R}$, niin*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Todistus.
Odotusarvon lineaarisuuden nojalla $E(aX+b) = aE(X)+b$, joten

$$\mathrm{Var}(aX+b) = E([aX+b-aE(X)-b]^2) = a^2E([X-E(X)]^2) = a^2\mathrm{Var}(X).$$

□

Vakion lisääminen satunnaismuuttujaan ei siis muuta jakauman varianssia, mutta vakiolla kertomisen vaikutus on neliöllinen. Tämä on intuitiivista siinä mielessä, että varianssi mittaa nimenomaan keskimääräistä *neliöityä* poikkeamaa odotusarvosta.

Seuraavassa lauseessa esitetään vielä muutamia varianssin ominaisuuksia, joiden olemassaolo on hyvä muistaa.

Lause 6.15. *Varianssin ominaisuuksia. Satunnaismuuttujille X ja Y, joiden varianssit ovat olemassa, pätee:*

(i) $\mathrm{Var}(X) \geq 0$

(ii) Jos $\mathrm{Var}(X) = 0$*, niin X on vakio.*

(iii) Jos $X \perp\!\!\!\perp Y$, niin $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Todistus.

(i) Koska $(X-E(X))^2 \geq 0$, niin $\mathrm{Var}(X) \geq 0$, mikäli se on reaalilukuna olemassa.

(ii) Jos $\mathrm{Var}(X) = 0$, niin

$$P(X = E(X)) = P((X-E(X))^2 = 0) = 1,$$

joten X saa arvon E(X) todennäköisyydellä 1, eli X on vakio.

(iii) E(X+Y) = E(X)+E(Y), joten riippumattomuuden nojalla

$$\begin{aligned}\mathrm{Var}(X+Y) &= E[(X-E(X)+Y-(EY))^2] \\ &= E[(X-E(X))^2 + 2(X-E(X))(Y-E(Y)) + (Y-E(Y))^2] \\ &= E[(X-E(X))^2 + (Y-E(Y))^2] + E[2(X-E(X))(Y-E(Y))] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\cdot 0\cdot 0\end{aligned}$$

□

Edellisen lauseen kohta (*iii*) eli riippumattomien satunnaismuuttujien varianssin summan kaava yleistyy myös useammalle satunnaismuuttujalle:

Lause 6.16. *Jos satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomia, eli $X_1, \ldots, X_n \perp\!\!\!\perp$, ja niiden varianssit ovat äärellisiä, niin*

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

6.3 Kovarianssi ja korrelaatio

Satunnaismuuttujien välistä lineaarista riippuvuutta kuvataan niiden yhteisvaihtelulla, eli *kovarianssilla.* Esitellään heti seuraavaksi kovarianssin määritelmä:

Määritelmä 6.17. *Kovarianssi.* [6] Olkoot X ja Y satunnaismuuttujia, joilla on odotusarvot $E(X)$ ja $E(Y)$. Niiden välinen kovarianssi on

$$\operatorname{Cov}(X, Y) = E([X - E(X)][Y - E(Y)]),$$

mikäli kyseinen odotusarvo on olemassa. Osoittautuu myös, että kovarianssin olemassaololle riittävä ehto on, että X:n ja Y:n toiset momentit ovat äärellisiä.

Seuraava lause esittelee kovarianssin määritelmästä seuraavia ominaisuuksia. Erityisesti toinen kohta kertoo, että varianssi on erityistapaus kovarianssista: se on satunnaismuuttujan kovarianssi itsensä kanssa.

Lause 6.18. *Kovarianssin ominaisuuksia.*

(i) $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$

(ii) $\operatorname{Var}(X) = \operatorname{Cov}(X, X)$

(iii) $\operatorname{Cov}(aX, Y) = a\operatorname{Cov}(X, Y)$

Todistus. Ensimmäinen ja toinen kohta seuraavat suoraan kovarianssin määritelmästä (6.17), kolmas kohta jätetään harjoitustehtäväksi. □

Kuten varianssi, myös kovarianssi voidaan esittää muodossa, joka monesti yksinkertaistaa laskuja:

Lause 6.19. *Olkoot X ja Y satunnaismuuttujia, joiden toiset momentit ovat äärellisiä. Tällöin*

$$\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

[6] engl. *covariance*

Todistus. Harjoitustehtävä. Kannattaa rohkeasti aloittaa kertomalla odotusarvon sisällä oleva tulo auki. □

Yleensä kovarianssin sijaan satunnaismuuttujien välisen lineaarisen riippuvuuden kuvaamiseen käytetään normalisoitua kovarianssia, eli korrelaatiota, joka saadaan jakamalla tarkasteltavien muuttujien kovarianssi niiden keskihajonnoilla.

Määritelmä 6.20. *Korrelaatio*[7] Olkoot X ja Y satunnaismuuttujia, joiden odotusarvot ja toiset momentit ovat äärellisiä. Niiden välinen korrelaatio on

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Korrelaation arvot sijaitsevat aina välillä $[-1,1]$. Korrelaation arvo 1 kuvaa täydellistä lineaarista riippuvuutta, ja -1 täydellistä negatiivista lineaarista riippuvuutta. Korrelaation arvo 0 taas tarkoittaa, ettei tarkasteltavien muuttujien välillä ole *lineaarista* riippuvuutta.

Osoittautuukin, että riippumattomien satunnaismuuttujien välinen kovarianssi, ja siten myös korrelaatio, todellakin on 0. Seuraava lause seuraa suoraan lauseen 6.1 kohdasta (*vii*), jonka mukaan riippumattomien satunnaismuuttujien odotusarvo saadaan niiden odotusarvojen tulona.

Lause 6.21. *Jos X ja Y ovat riippumattomia, eli $X \perp\!\!\!\perp Y$, niin*

$$\mathrm{Cov}(X,Y) = 0,$$

ja siten myös

$$\mathrm{Corr}(X,Y) = 0.$$

Todistus.

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

□

Huomautus 6.22. Jos $\mathrm{Cov}(X,Y) = 0$, sanotaan, että X ja Y ovat *korreloimattomia.* Edellisen lauseen nojalla riippumattomuudesta seuraa aina korreloimattomuus. Mutta korreloimattomuudesta ei kuitenkaan välttämättä seuraa riippumattomus. Tämä johtuu siitä, että kovarianssi ja korrelaatio mittaavat pelkästään muuttujien välistä *lineaarista* riippuvuutta. Jos riippuvuus on epälineaarista, kovarianssi ei välttämättä tavoita sitä.

[7] engl. *correlation*

Esimerkiksi standardinormaalijakauman tiheysfunktio on symmetrinen odotusarvon 0 suhteen, joten $E(X) = E(X^3) = 0$. Tästä seuraa, että standardinormaalijakautuneelle muuttujalle $X \sim N(0,1)$ ja sen neliölle

$$\mathrm{Cov}(X, X^2) = E(X \cdot X^2) - E(X)E(X^2) = 0 - 0 = 0,$$

mutta tietenkään X ja X^2 eivät ole riippumattomia.

Satunnaismuuttujien summan varianssi saadaan laskettua niiden kovarianssin avulla. Satunnaismuuttujien erotukselle ja lineaarikombinaatioille voidaan laskea kaava vastaavalla tavalla. Erityisesti, jos satunnaismuuttujat ovat riippumattomia, niiden summan varianssi saadaan niiden varianssien summana, kuten lauseen 6.15 kohdasta (iii) muistamme.

Lause 6.23. *Jos satunnaismuuttujien X ja Y varianssit ovat äärellisiä, niin*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

Erityisesti, jos $X \perp\!\!\!\perp Y$, niin

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Todistus. Todistus saadaan laskemalla neliöt auki ja soveltamalla odotusarvon lineaarisuutta:

$$\begin{aligned}\mathrm{Var}(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).\end{aligned}$$

Väitteen toinen osa seuraa lauseesta 6.21, jonka mukaan riippumattomille satunnaismuuttujille $\mathrm{Cov}(X,Y) = 0$. □

Seuraava esimerkki niputtaa yhteen tässä luvussa oppimiamme asioita. Ennen esimerkkiä esitellään $Cauchyn - Schwarzin$ epäyhtälö, jota tarvitsemme esimerkissä.

Lause 6.24. *Cauchyn-Schwarzin epäyhtälö. Satunnaismuuttujille X ja Y pätee*

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

Nyt voimme siirtyä itse esimerkkiin. Lukija voi ensin yrittää ratkaista ongelmaa itse ja ottaa tarvittaessa mallia esimerkistä.

Esimerkki 6.25. Olkoot X ja Y satunnaismuuttujia, joilla on varianssit. Osoitetaan seuraavat ehdot yhtäpitäviksi:

(i) $\mathrm{Cov}(X, Y) = 0$,

(ii) $E(XY) = E(X)E(Y)$,

(iii) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Varianssien $\mathrm{Var}(X)$ ja $\mathrm{Var}(Y)$ olemassaolosta seuraa, että myös odotusarvot $E(X^2)$, $E(Y^2)$, $E(X)$ ja $E(Y)$ ovat olemassa, jolloin lauseen 6.24 nojalla myös odotusarvo $E(XY)$ on olemassa. Nyt satunnaismuuttujilla X ja Y on kovarianssin määritelmän (6.17) ja lauseen 6.19 perusteella kovarianssi

$$\mathrm{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

Satunnaismuuttujalla $X + Y$ on olemassa varianssi

$$\begin{aligned}
\mathrm{Var}(X + Y) &= E((X + Y)^2) - E(X + Y)^2 = E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\
&= (E(X^2) - E(X)^2) + 2(E(XY) - E(X)E(Y)) + (E(Y^2) - E(Y)^2) \\
&= \mathrm{Var}(X) + 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).
\end{aligned}$$

Nyt nähdään, että

$$\mathrm{Cov}(X, Y) = 0 \Leftrightarrow E(XY) - E(X)E(Y) = 0 \Leftrightarrow E(XY) = E(X)E(Y)$$

eli ehdot (i) ja (ii) ovat yhtäpitäviä. Lisäksi nähdään, että

$$\mathrm{Cov}(X, Y) = 0 \Leftrightarrow \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

eli ehdot (i) ja (iii) ovat yhtäpitäviä, jolloin kaikki ehdot ovat ylläpitäviä.

Huomaa, että yhdestäkään ehdosta ei seuraa X:n ja Y:n riippumattomuus.

6.4 Harjoitustehtäviä

1. Tasoon piirretyn kolmion kärkinä ovat origo, ja x- ja y-akseleilta satunnaisesti valitut pisteet X ja Y, jotka ovat riippumattomia ja noudattavat normaalijakaumaa parametreilla $\mu = 0$, $\sigma^2 = 1$. Laske kolmion pinta-alan odotusarvo.

2. Olkoot X, Y ja Z riippumattomia satunnaismuuttujia, joilla kaikilla on sama odotusarvo μ ja varianssi σ^2. Laske odotusarvo ja varianssi sm:ille

(a) $2X + 3$,

(b) $X - Y$,

(c) $X - \frac{1}{2}Y$,

(d) $X + 2Y + 3Z$,

(e) XY,

(f) XYZ.

3. Olkoon X satunnaismuuttuja, joka kuvaa tavallisen nopan silmälukua yhdellä heitolla. Etsi $\text{Var}(X)$.

4. Olkoon X satunnaismuuttuja, jonka tiheysfunktio $f(x) = 2x$, kun $0 \leq x \leq 1$. Etsi $\text{Var}(X)$.

5. (Ross Exercise 7.14) Korttipakasta jaetaan 13 kortin bridge-käsi. Olkoot satunnaismuuttujat X ja Y herttojen ja ässien lukumäärä kädessä. Näytä, että X ja Y ovat korreloimattomia. Entä ovatko ne riippumattomia?

6. (Ross Example 7.4a)Olkoot $X_1 \ldots X_n$ riippumattomia, samoin jakautuneita satunnaismuuttujia odotusarvolla μ ja varianssilla σ^2 ja $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ näiden otoskeskiarvo. Mikä on otoskeskiarvon varianssi $\text{Var}(\bar{X})$?

7. Kahta noppaa heitetään. Olkoon sm $X =$ ”1. nopan silmäluku” ja sm $Y =$ ”pistelukujen summa”. Laske $\text{Corr}(X, Y)$.

Luku 7

Raja-arvolauseet

Tässä luvussa siirrytään jakaumista ja niiden tunnusluvuista todennäköisyyden teorian ja todennäköisyyslaskennan merkittävimpien tulosten pariin. Luvussa esitellyt tulokset mahdollistavat teorian kehittämisen lisäksi myös tärkeitä käytännön menetelmiä, kuten luvun lopussa käsiteltävän *normaaliapproksimaation.*

7.1 Epäyhtälöitä

Seuraavaksi esittelemme kaksi todennäköisyyslaskennan keskeisintä epäyhtälöä: Markovin ja Tšebyševin epäyhtälöt[1]. Näiden avulla voidaan laskea ylärajoja satunnaismuuttujien "häntätodennäköisyyksille", eli todennäköisyyksille, että satunnaismuuttujien arvot poikkeavat satunnaismuuttujan odotusarvosta vähintään jonkin vakion verran, vaikka satunnaismuuttujan pistetodennäköisyys- tai tiheysfunktio ei olisi tiedossa.

Markovin epäyhtälössä oletetaan pelkästään, että epänegatiivisen satunnaismuuttujan odotusarvo on tunnettu: tämän perusteella todennäköisyys, että X:n arvo on vähintään a, on pienempää kuin X:n odotusarvo jaettuna a:lla. Markovin epäyhtälön todistuksessa tarvitaan satunnaismuuttujan karakteristisia funktioita, joita ei olla käsitelty tällä kurssilla, joten todistus ohitetaan.

Lause 7.1. *Markovin epäyhtälö.*[2] *Jos $X \geq 0$ ja X:n odotusarvo on olemassa, niin*

$$P(X \geq a) \leq \frac{E(X)}{a} \quad \textit{kaikille} \quad a > 0.$$

[1]Nimitykset vaihtelevat ulkomaisessa kirjallisuudessa: monesti myös Markovin epäyhtälöä kutsutaan Tšebyševin epäyhtälöksi.

[2]engl. *Markov's inequality*

Jos $\mu = E(X) > 0$, asettamalla $a = k\mu$ Markovin epäyhtälö voidaan myös kirjoittaa muodossa

$$P(X \geq k\mu) \leq \frac{1}{k} \quad \text{kaikille} \quad k > 0.$$

Markovin epäyhtälön mukaan siis todennäköisyys, että X:n arvo on vähintään k kertaa sen odotusarvo, on pienempää kuin $1/k$. Tämä muoto näyttää myös selkeämmin yhteyden Tšebyševin epäyhtälöön. Näitä kutsutaankin joskus ensimmäisen (Markov) ja toisen momentin (Tšebyšev) epäyhtälöiksi.

Tšebyševin epäyhtälöä voidaan soveltaa kaikille satunnaismuuttujille, joiden odotusarvo ja varianssi tunnetaan. Tšebyševin epäyhtälön mukaan todennäköisyys, että satunnaismuuttujan arvot poikkeavat vähintään k:n keskihajonnan verran satunnaismuuttujan keskiarvosta, on pienempää kuin $1/k^2$.

Lause 7.2. *Tšebyševin epäyhtälö.*[3] *Jos X on satunnaismuuttuja, jonka odotusarvo $\mu = E(X)$ ja varianssi $\sigma^2 = \mathrm{Var}(X) > 0$ ovat olemassa, niin*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \textit{kaikille} \quad k > 0.$$

Todistus. Sovelletaan Markovin epäyhtälöä satunnaismuuttujalle $|X - \mu|^2$ ja vakiolle $a = k^2\sigma^2$, sekä varianssin määritelmää:

$$P(|X - \mu| \geq k\sigma) = P(|X - \mu|^2 \geq k^2\sigma^2) \leq \frac{E(|X - \mu|^2)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

□

Sovelletaan seuraavaksi Markovin ja Tšebyševin epäyhtälöitä käytännön esimerkissä.

Esimerkki 7.3. Jatkoa esimerkeille 5.10 ja 5.17. Tarkastellaan jälleen hehkulamppuja, joiden kestoaika X noudattaa eksponenttijakaumaa, jonka odotusarvo on 1000 tuntia, eli $X \sim \mathrm{Exp}(0.001)$.

Koska $X \geq 0$, Markovin epäyhtälöstä saadaan ylärajat todennäköisyydelle, että lamppu palaa vähintään 3000 tuntia:

$$P(X \geq 3000) = P(X \geq 3\mu) \leq \frac{1}{3}.$$

Eksponenttijakauman varianssiksi saadaan osittaisintegroinnilla $1/\lambda^2$; siten X:n keskihajonta on $\sigma = 1/\lambda = 1000$. Kun tunnetaan varianssi, Tšebyševin epäyhtälöstä saadaan tiukempi yläraja:

$$P(X \geq 3000) \leq P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4}.$$

[3] engl. *Chebyshev's inequality*

Tässä tapauksessa tietenkin tiedämme X:n kertymäfunktion F, joten voimme laskea kyseiselle todennäköisyydelle tarkan arvon:

$$P(X \geq 3000) = 1 - F(3000) = e^{-0.001 \cdot 3000} = e^{-3} \approx 0.05.$$

Seuraavaan taulukkoon on vielä laskettu Markovin ja Tšebyševin epäyhtälöstä lasketut ylärajat, sekä tarkat arvot yllä tarkastellun X:n häntätodennäköisyyksille.

x	Markov	Tšebyšev	Tarkka kf:sta laskettu $P(X \geq x)$
3000	0.33	0.25	0.050
4000	0.25	0.11	0.018
5000	0.20	0.0625	0.007
10000	0.10	0.012	$4.54 \cdot 10^{-5}$

Huomataan, että epäyhtälöiden avulla saadut ylärajat ovat melko karkeita, ja toisaalta, että suurille poikkeamille Tšebyševin epäyhtälöstä saadaan huomattavasti tiukempi yläraja. Tämä selittyy sillä, että Tšebyševin epäyhtälössä käytössä on enemmän informaatiota: odotusarvon lisäksi myös varianssi.

Vaikka Markovin ja Tšebyševin epäyhtälöiden antamat rajat osoittautuivat melko epätarkoiksi, ovat nämä epäyhtälöt merkittäviä, sillä kuten jo aiemmin mainittiin, niiden avulla voidaan laskea rajoja satunnaismuuttujan arvojen todennäköisyyksille, vaikka sm:n jakauma ei olisi tiedossa. Lisäksi Markovin ja Tšebyševin epäyhtälöitä tarvitaan monien todennäköisyyslaskennan keskeisten tulosten todistuksissa. Tällä kurssilla näistä käsitellään suurten lukujen laki ja keskeinen raja-arvolause.

7.2 Raja-arvolauseita

Tässä kappaleessa esiteltävät suurten lukujen laki ja keskeinen raja-arvolause käsittelevät riippumattoman ja samoin jakautuneen[4] satunnaismuuttujajonon[5] $X_1, \ldots, X_n$ keskiarvon (tai vastaavasti summan) käyttäytymistä, kun satunnaiskoetta toistetaan rajatta. Osoittautuu, että keskiarvojen jono lähestyy satunnaismuuttujien yhteistä odotusarvoa, ja niiden jakauma lähestyy normaalijakaumaa, jonka odotusarvo on myös satunnaismuuttujien yhteinen odotusarvo.

[4] Suurten lukujen laista ja keskeisestä raja-arvolauseesta on olemassa useita eri versioita erilaisilla oletuksilla. Esittelemme tässä klassiset versiot näistä lauseista; vastaavat tulokset voi kuitenkin osoittaa myös näitä löyhemmillä oletuksilla.

[5] Tällaista satunnaismuuttujajonoa käytämme usein merkintää *i.i.d*, eli *independent and identically distributed*.

Nämä raja-arvotulokset ovat keskeisiä tilastollisen päättelyn teoriassa. Estimaattoreille[6] ei monesti pystytä laskemaan tarkkoja jakaumia, vaan sovelletaan keskeistä raja-arvolausetta, jonka mukaan suurella otoskoolla riippumattomien ja samoin jakautuneiden satunnaismuuttujien summat ja keskiarvot ovat likimain normaalisti jakautuneita: tämän perusteella estimaattorin jakaumaa voidaan approksimoida normaalijakaumalla.

7.2.1 Suurten lukujen laki

Aiemmin käytetty odotusarvon tulkinta uhkapelin keskimääräisenä voittona kierrosta kohti, kun peliä pelataan monta kierrosta, perustuu yhteen todennäköisyyslaskennan keskeisistä tuloksista, suurten lukujen lakiin. Sen mukaan riippumattomien satunnaismuuttujien $X_1, \dots X_n$, joilla on sama odotusarvo, keskiarvo lähenee niiden odotusarvoa, kun n kasvaa rajatta, eli kun satunnaiskoetta toistetaan n kertaa, ja $n \to \infty$. Matemaattisesti ilmaistuna keskiarvojen jonon raja-arvo (toistojen lukumäärän funktiona) on satunnaismuuttujien X_i yhteinen odotusarvo μ.

Suurten lukujen laki on myös monien tieteellisten koeasetelmien taustalla: jos otoskoko on riittävän suuri, niin sen nojalla voidaan olettaa, että keskimääräinen onnistumisten osuus kokeessa on lähellä todellista onnistumistodennäköisyyttä. Esimerkiksi jos testataan, parantaako uusi lääke sairauden, voidaan olettaa että parantuneiden potilaiden osuus otoksesta on lähellä todellista parantumistodennäköisyyttä, jos otoskoko on suuri.

Määritellään ensin, mitä tarkoitamme suppenemisella eli konvergenssilla satunnaismuuttujajonon (joka siis on jono funktioita perusjoukolta reaalilukujen joukolle) tapauksessa. Satunnaismuuttujien konvergenssi voidaan määritellä usealla eri tavalla; heikon suurten lukujen lain yhteydessä käytämme ns. *stokastista konvergenssia.*

Määritelmä 7.4. *Stokastinen konvergenssi.*[7] Satunnaismuuttujajono $X_1, X_2, \dots$ *suppenee stokastisesti* kohti satunnaismuuttujaa X, jos

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \text{kaikille} \quad \epsilon > 0.$$

Tällöin merkitään

$$X_n \xrightarrow{p} X.$$

[6] Estimaattorilla tarkoitamme arviota tunnusluvusta, jota mallinnetaan satunnaisena. Toisin sanoin, samalla tavalla kuin mallinnamme ei-realisoituneita eli ”tarkasteltaessa satunnaisia” satunnaismuuttujia, tarkastelemme näiden satunnaismuuttujien yhdistelmiä, yleisimmin keskiarvoa. Tällaista keskiarvoa kutsutaan odotusarvon *estimaattoriksi* ja jo-realisoitunutta satunnaismuuttujien keskiarvoa kutsumme *estimaatiksi.* Näistä termeistä lisää tilastollisen päättelyn kurssilla.

[7] engl. *stochastic convergence*

Huomaa, että vaikka stokastisen konvergenssin yleisessä muotoilussa X voi olla satunnaismuuttuja, seuraavassa tuloksessa keskiarvojen jono suppenee kohti vakiota $\mu = E(X_i)$.

Nyt voimme muotoilla heikon suurten lukujen lain ja todistaa sen Tšebyševin epäyhtälön avulla. Todistuksen helpottamiseksi teemme ylimääräisen lisäoletuksen, että satunnaismuuttujilla $X_1, X_2, \dots$ on yhteisen odotusarvon lisäksi sama varianssi.

Lause 7.5. *(Heikko) suurten lukujen laki, (SLL)*[8]*. Olkoon* $X_1, X_2, \dots$ *jono riippumattomia satunnaismuuttujia*[9]*, joilla on sama odotusarvo* $\mu = E(X_i)$ *ja varianssi* $\sigma^2 = \mathrm{Var}(X_i) \leq \infty$*. Nyt satunnaismuuttujien keskiarvojen*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

jono $\bar{X}_1, \bar{X}_2, \dots$ *suppenee stokastisesti kohti satunnaismuuttujien yhteistä odotusarvoa:*

$$\bar{X}_n \xrightarrow{p} \mu,$$

eli

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0 \quad \textit{kaikille} \quad \epsilon > 0.$$

Todistus. Satunnaismuuttujien riippumattomuuden nojalla niiden keskiarvon $\bar{X}_n$ varianssi on

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

ja siten keskihajonta on

$$\sqrt{\mathrm{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}$$

kaikille $n \in \{1, 2, \dots\}$.

Nyt kaikille $\epsilon > 0$ saadaan Tšebyševin epäyhtälöstä käyttäen arvoa $k = \sqrt{n}\epsilon/\sigma$:

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P\Big(|\bar{X}_n - \mu| \geq k\frac{\sigma}{\sqrt{n}}\Big) \leq \frac{1}{k^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \text{ kun } n \to \infty.$$

□

Huomautus 7.6. Jos satunnaismuuttujat $X_1, X_2, \dots$ ovat riippumattomia ja **samoin jakautuneita**, tulos pätee stokastista suppenemista voimakkaammalle konvergenssin muodolle, *melkein varmalle suppenemiselle*,[10] myös ilman oletusta varianssin olemassaolosta. Tätä *vahvaksi suurten lukujen laiksi* kutsuttua versiota ei käsitellä tällä kurssilla.

[8] engl. *(weak) law of large numbers, LLN*

[9] Huomaathan, että satunnaismuuttujien ei oikeastaan tarvitse edes noudattaa samaa jakaumaa vaan riittää, että jakaumien odotusarvo ja varianssi ovat samat.

[10] Melkein varma suppeneminen tarkoittaa, että jono $X_1, X_2, \dots$ konvergoi kohti satunnaismuuttujaa

7.2.2 Keskeinen raja-arvolause

Keskeisen raja-arvolauseen mukaan samoin jakautuneen ja riippumattoman satunnaismuuttujajonon keskiarvon (tai vastaavasti summan) jakauma suppenee kohti tiettyä normaalijakaumaa. Määrittelemme ensin täsmällisesti kertymäfunktioiden avulla, mitä tarkoittaa satunnaismuuttujajonon suppeneminen jakaumamielessä.

Määritelmä 7.7. *Jakaumasuppeneminen.* Olkoon $X_1, X_2, \ldots$ jono satunnaismuuttujia, joiden kertymäfunktiot ovat $F_1, F_2, \ldots,$ ja X satunnaismuuttuja, jonka kertymäfunktio on G. Jono (X_n) *suppenee jakaumaltaan* kohti X:ää, mikäli niiden kertymäfunktioiden jono suppenee pisteittäin kohti X:än kertymäfunktiota, eli

$$\lim_{n\to\infty} F_n(x) = G(x)$$

kaikissa G:n jatkuvuuspisteissä x. Tällöin merkitään

$$X_n \xrightarrow{d} X.$$

Seuraavaksi esitämme klassisen version keskeisestä raja-arvolauseesta. Todistuksessa tarvitaan jälleen karakteristisia funktioita, joten todistus ohitetaan.

Lause 7.8. *Keskeinen raja-arvolause.*[11] *Olkoon $X_1, X_2, \ldots$ jono riippumattomia samoin jakautuneita satunnaismuuttujia, joiden yhteinen odotusarvo on $\mu = E(X_i)$ ja varianssi $\sigma^2 = \mathrm{Var}(X_i) < \infty$.*

Tällöin standardoitujen keskiarvojen jono suppenee jakaumaltaan kohti standardinormaalijakaumaa:

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1),$$

tai yhtäpitävästi

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0,\sigma^2).$$

Keskeinen raja-arvolause tarkoittaa käytännössä sitä, että jos jostakin jakaumasta, jonka odotusarvo on μ, otetaan n kappaletta usean havainnon otoksia, niin näiden otosten keskiarvojen jakauma lähestyy normaalijakaumaa odotusarvolla μ, kun n kasvaa rajatta.

X todennäköisyydellä yksi, eli kaikkialla muualla paitsi mahdollisesti perusjoukon nollamittaisessa osajoukossa:

$$P\big(\lim_{n\to\infty} X_n = X\big) = 1.$$

Tällöin merkitään (a.s. = almost surely):

$$X_n \xrightarrow{a.s.} X.$$

[11] engl. *central limit theorem, CLT*

7.2.3 Normaaliapproksimaatio

Keskeistä raja-arvolausetta voidaan käyttää otoskeskiarvon jakauman approksimoimiseen, sillä jos n on riittävän suuri, riippumattomien ja samoin jakautuneiden satunnaismuuttujien $X_1, \ldots, X_n$ standardoitu ($\bar{X}_n$:stä vähennetään sen odotusarvo μ, ja tulos jaetaan $\bar{X}_n$:n keskihajonnalla $\frac{\sigma}{\sqrt{n}}$) keskiarvo

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma$$

noudattaa likimain standardinormaalijakaumaa, mitä voidaan merkitä

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \overset{d}{\approx} N(0,1).$$

Tällöin normaalijakauman ominaisuuksien nojalla keskiarvon $\bar{X}_n$ jakaumaa voidaan approksimoida normaalijakaumalla $N(\mu, \frac{\sigma^2}{n})$, eli

$$\bar{X}_n \overset{d}{\approx} N\Big(\mu, \frac{\sigma^2}{n}\Big),$$

missä $\mu = E(X_i)$ on satunnaismuuttujien yhteinen odotusarvo, ja $\sigma^2 = \mathrm{Var}(X_i)$ niiden yhteinen varianssi.

Keskeisen raja-arvolauseen mukaan myös riippumattomien ja samoin jakautuneiden satunnaismuuttujien $X_1, \ldots, X_n$ summan $\sum_{i=1}^n X_i$ jakaumaa voidaan approksimoida seuraavalla normaalijakaumalla:

$$\sum_{i=1}^{n} X_i \overset{d}{\approx} N(n\mu, n\sigma^2).$$

Approksimaation tarkkuus riippuu n:n lisäksi myös satunnaismuuttujien $X_1, \ldots, X_n$ jakaumasta. Jos se muistuttaa normaalijakaumaa, eli on yksihuippuinen ja suhteellisen symmetrinen, approksimaatio voi olla suhteellisen tarkka jo muutamalla kymmenellä havainnolla. Jos taas jakauma on hyvin vino, kuten esimerkiksi eksponenttijakauma, tarkkaan approksimaatioon vaaditaan huomattavasti suurempaa otoskokoa. Approksimaation tarkkuus riippuu myös siitä, mitä kohtaa jakaumasta tarkastellaan: normaaliapproksimaatio on usein tarkempi jakauman keskellä, ja epätarkempi sen hännissä. Esitetään seuraavassa esimerkissä, miten normaaliapproksimaatiota voidaan soveltaa käytännössä.

Esimerkki 7.9. Omenoita pakataan laatikkoon. Yhden omenan painon odotusarvo on 200 ja hajonta 20 grammaa. Pakkaaminen lopetetaan heti, kun omenien yhteispaino on vähintään 10 kg. Määritetään normaaliapproksimaatiota käyttäen $P(N \leq 49)$, missä N on laatikkoon sijoitettujen omenoiden lukumäärä.

Olkoon sm X_i i:nnen omenan paino, missä $i = 1, 2, \ldots$, ja sm N laatikkoon pakattujen omenoiden määrä. Nyt normaaliapproksimaatiota käyttäen

$$\begin{aligned}
P(N \leq 49) &= P\Big(\sum_{i=1}^{49} X_i \geq 10000\Big) = 1 - P\Big(\sum_{i=1}^{49} X_i < 10000\Big) \\
&= 1 - P\Big(\frac{\sum_{i=1}^{49} X_i - 49 \cdot 200}{20\sqrt{49}} < \frac{10000 - 49 \cdot 200}{20\sqrt{49}}\Big) \quad \text{standardointi} \\
&\approx 1 - \Phi\Big(\frac{10000 - 49 \cdot 200}{20\sqrt{49}}\Big) \quad \text{normaaliapproksimaatio} \\
&= 1 - \Phi\Big(\frac{10}{7}\Big) \approx 0.077.
\end{aligned}$$

Monien diskreettien jakaumien, kuten binomi- ja Poisson-jakaumien, kertymäfunktiot voidaan ilmaista pelkästään summina. Ennen tietokoneaikaa nämä olivat hankalia laskea, joten monesti turvauduttiin normaaliapproksimaatioon, jossa pystyttiin hyödyntämään standardinormaalijakauman kertymäfunktion[12] taulukoituja arvoja. Diskreettejä jakaumia approksimoitaessa ns. *jatkuvuuskorjaus* monesti parantaa approksimaation tarkkuutta.

Huomautus 7.10. Jatkuvuuskorjaus. Kun kokonaislukuarvoisen satunnaismuuttujan, eli satunnaismuuttujan jonka arvojoukko kuuluu kokonaislukujen joukkoon $\mathbb{Z}$, X jakaumaa approksimoidaan normaalijakaumalla (tai jollain muulla jatkuvalla jakaumalla), voidaan tarkastella jokaiselle satunnaismuuttujan arvolle $k \in \mathbb{Z}$ todennäköisyyttä

$$P\Big(k - \frac{1}{2} \leq X \leq k + \frac{1}{2}\Big)$$

k:n pistetodennäköisyyden

$$P(X = k)$$

sijaan, sillä kokonaislukuarvoiselle satunnaismuuttujalle

$$\{X = k\} = \Big\{k - \frac{1}{2} \leq X \leq k + \frac{1}{2}\Big\}.$$

Esimerkiksi jos X on yhden nopan heiton silmäluku, niin

$$\{X = 2\} = \{1.5 \leq X \leq 2.5\},$$

sillä nopan silmäluku voi saada arvoja vain joukosta $\{1, 2, \ldots, 6\}$.

[12] $\Phi(z) = F_Z(z) = P(Z < z)$ standardinormaalijakaumaa noudattavalle satunnaismuuttujalle $Z \sim N(0, 1)$

Tämä on myös erittäin järkevää, sillä normaalijakautuneelle (kuten kaikille muillekin jatkuvaa jakaumaa noudattaville) satunnaismuuttujalle Y kaikki pistetodennäköisyydet ovat nollia:

$$P(Y = k) = 0.$$

Koska kaikkien joukkojen todennäköisyydet saadaan diskreeteille satunnaismuuttujille joukon pisteiden pistetodennäköisyyksien summina, niin jatkuvuuskorjausta käytettäessä väleistä tulee $\frac{1}{2}$:n verran pidempiä kummastakin päästä:

$$P(k_1 \leq X \leq k_2) = P\Big(k_1 - \frac{1}{2} \leq X \leq k_2 + \frac{1}{2}\Big) \quad \text{kaikille} \quad k_1, k_2 \in \mathbb{Z}, k_1 \leq k_2.$$

Esimerkiksi jos X on yhden nopan heiton silmäluku, niin

$$P(2 \leq X \leq 4) = P(1.5 \leq X \leq 4.5).$$

Samoin vain toiselta puolelta rajatuille väleille

$$P(X \leq k) = P\Big(X \leq k + \frac{1}{2}\Big) \quad \text{kaikille} \quad k \in \mathbb{Z},$$

ja

$$P(X \geq k) = P\Big(X \geq k - \frac{1}{2}\Big) \quad \text{kaikille} \quad k \in \mathbb{Z}.$$

Pienellä otoskoolla jatkuvuuskorjaus voi parantaa normaaliapproksimaation tarkkuutta huomattavasti. Havainnollistetaan jatkuvuuskorjausta esimerkillä.

Esimerkki 7.11. Heitetään noppaa sata kertaa. Mikä on todennäköisyys, että saadaan vähintään 15, mutta enintään 20 kutosta?

Kyseessä on toistokoe, jossa satunnaismuuttujat $X_1, \ldots, X_{100}$, voidaan määritellä seuraavasti

$$X_i = \begin{cases} 1, & \text{jos } i\text{:nnellä heitolla saadaan kutonen} \\ 0, & \text{muuten,} \end{cases}$$

ovat riippumattomia ja noudattavat samaa Bernoullin jakaumaa: $X_i \sim \text{Bernoulli}(\frac{1}{6})$ kaikille $i \in \{1, 2, \ldots, n\}$. Tällöin niiden summalle $X = \sum_{i=1}^{100} X_i$ saataisiin tarkka jakauma: summa noudattaa binomijakaumaa otoskoolla $n = 100$, ja onnistumistodennäköisyydellä $p = \frac{1}{6}$.

Lasketaan kuitenkin likiarvo kysytylle todennäköisyydelle käyttäen normaaliapproksimaatiota. Odotusarvon lineaarisuuden nojalla summan odotusarvoksi saadaan

$$E(X) = E\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} E(X_i) = np$$

ja varianssiksi saadaan riippumattomuuden nojalla

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} \mathrm{Var}(X_i) = np(1-p).$$

Keskeisen raja-arvolauseen nojalla standardoitu keskiarvo

$$Z = \frac{X - E(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}}$$

noudattaa likimain standardinormaalijakaumaa $N(0,1)$ suurella otoskoolla n, joten (jatkuvuuskorjausta käyttäen) kysytyksi todennäköisyydeksi saadaan:

$$\begin{aligned} P(15 \le X \le 20) &= P(14.5 \le X \le 20.5) \\ &= P\Big(\frac{14.5 - np}{\sqrt{np(1-p)}} \le Z \le \frac{20.5 - np}{\sqrt{np(1-p)}}\Big) \\ &= \Phi\Big(\frac{20.5 - 100 \cdot \frac{1}{6}}{\sqrt{100 \cdot \frac{1}{6}(1 - \frac{1}{6})}}\Big) - \Phi\Big(\frac{14.5 - 100 \cdot \frac{1}{6}}{\sqrt{100 \cdot \frac{1}{6}(1 - \frac{1}{6})}}\Big) \\ &\approx \Phi(1.03) - \Phi(-0.58) \\ &\approx 0.848 - 0.281 \\ &= 0.567. \end{aligned}$$

Tarkaksi arvoksi saadaan binomijakauman kertymäfunktiosta F[13]

$$P(15 \le X \le 20) = P(X \le 20) - P(X \le 14) = F(20) - F(14) \approx 0.561,$$

eli normaaliapproksimaatio on tässä tapauksessa erittäin tarkka.

Jos heikon suurten lukujen lain ja keskeisen raja-arvolauseen käsitteet tuntuvat vielä epäselviltä, lukijan kannattaa tutustua muuhun Internetistä löytyvään materiaaliin näistä lauseista.[14] Vaikka tulosten muotoilu saattaa vaikuttaa hiukan tekniseltä ja niiden todistamiseen tarvitaan asioita, jotka eivät kuulu tämän kurssin sisältöön, on suurten lukujen lain ja keskeisen raja-arvolauseen sisältö intuiivisesti melko helposti hahmotettavissa, joten lisämateriaaliin kannattaa tarvittaessa perehtyä.

[13] `R`:ssä komennolla `pbinom(20,100,1/6) - pbinom(14,100,1/6)`.

[14] esimerkiksi Khan-academyn videot `https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library/expected-value-lib/v/law-of-large-numbers` ja `https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/central-limit-theorem` havainnollistavat näitä tuloksia hyvin

7.3 Harjoitustehtäviä

1. Eräs autokauppa myy keskimäärin 16 autoa viikossa. Anna yläraja todennäköisyydelle, että seuraavalla viikolla myytyjen autojen lukumäärä on 20.

2. Miljoonan luvun keskiarvo on 10. Lukujen neliöiden keskiarvo on 101. Anna mahdollisimman tarkka yläraja niiden lukujen määrälle, jotka ovat vähintään yhtä suuria kuin 14.

3. Noppaa heitetään n kertaa. Olkoon sm X kutosten suhteellinen frekvenssi (kutosten lkm/n) ja olkoon $\epsilon = 0.01$. Haluamme, että X olisi ϵ:n tarkkuudella likiarvo oikealle todennäköisyydelle P("silmäluku on 6"), eli haluamme, että epäyhtälö $|X - \frac{1}{6}| < \epsilon$ toteutuu. Tämän komplementtitapahtumaa $|X - \frac{1}{6}| \geq \epsilon$ kutsumme liian suureksi virheeksi ja merkitsemme kirjaimella L. Laske Tšebyševin epäyhtälön avulla yläraja todennäköisyydelle P(L). Kuinka suuri n:n on oltava, jotta P(L) < 0.05.

4. Eräässä ladontamenetelmässä sattuu 500-sivuiseen kirjaan keskimäärin 1000 painovirhettä.

   (a) Laske Poisson-jakauman avulla tn, että yhdellä sivulla olevien painovirheiden lkm on pienempi kuin 2.

   (b) Olkoon X niiden sivujen lkm, joilla painovirheitä on vähemmän kuin 2. Laske normaaliapproksimaatiota käyttäen likiarvo tn:lle $P(X > 215)$

5. Laskettaessa n reaaliluvun aritmeettinen keskiarvo luvut pyöristetään kokonaisluvuiksi. Oletamme, että pyöristysvirheet ovat riippumattomia ja niiden jakauma noudattaa Tas$(-\frac{1}{2}, \frac{1}{2})$-jakaumaa. Olkoon sm X aritmeettisen keskiarvon virhe. Arvioi valitsemallasi menetelmällä, mikä n:n on vähintään oltava, jotta

$$P(|X| \geq 0.01) < 0.05.$$

6. Juhlaillalliselle on saapunut 100 vierasta. Illalliselle on varattu 20 litraa boolia. Vieraat ottavat boolia itse toisistaan riippumattomasti siten, että odotusarvo yhden vieraan booliannokselle on 0.19 litraa ja keskihajonta 0.05 litraa. Mikä on tn, että viimeisenä saapunut vieras jää ilman boolia?

Luku 8

Lisämateriaalia

Tässä luvussa käsittelemme sellaisia todennäköisyyslaskennan teemoja, jotka eivät ole olleet kurssilla välttämättömiä, mutta joiden hahmottaminen syventää käsitystä todennäköisyyslaskennasta ja jotka kenties avaavat joitakin kurssilla käsiteltyjä asioita paremmin lukijalle. Luvun aiheisiin, eli satunnaismuuttujan muunnokseen ja moniuloitteisiin todennäköisyysjakaumiin tullaan tulevilla kursseilla paneutumaan syvällisemmin. Lisäksi luvun lopussa esitellään lyhyesti R-ohjelmiston käyttöä todennäköisyyslaskennassa.

8.1 Satunnaismuuttujan muunnos

Tässä kappaleessa käsitellään satunnaismuuttujan muunnoksen pistetodennäköisyysfunktion, tiheysfunktion ja kertymäfunktion laskemista. Muistetaan vielä satunnaismuuttujan muunnoksen käsite, jota olemme jo aiemmin sivunneet mm. normaalijakauman yhteydessä esimerkissä 5.14. Luvussa 6 opimme myös laskemaan satunnaismuuttujan muunnoksen odotusarvon ja **lineaarisen** muunnoksen varianssin. Palautetaan aluksi mieleen lause 5.19:

Jos $f : \mathbb{R} \to \mathbb{R}$ on funktio ja X on satunnaismuuttuja, myös $Y = g(X)$ on satunnaismuuttuja.

Esitellään seuraavaksi, miten muunnoksen Y jakauma lasketaan diskreetissä ja jatkuvassa tapauksessa.

Lause 8.1. *Satunnaismuuttujan muunnoksen pistetodennäköisyysfunktio. Olkoon X diskreetti sm ptnf:lla f_X ja olkoon g funktio. Tällöin g(X) on diskreetti satunnaismuuttuja ja sen ptnf f_Y on*

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x).$$

Merkinnällä $x \in g^{-1}(y)$ tarkoitetaan niitä X:n arvoja, joilla $g(x) = y$.

Todistus. Koska X:n arvojoukko on numeroituva, on selvää, että myös Y voi saavuttaa korkeintaan numeroituvasti äärettömän määrän arvoja ja on siis siten diskreetti. Kaikille $y \in \mathbb{R}$ pätee

$$\begin{aligned} P(Y = y) &= P(g(X) = y) = P(g(X) \in \{y\}) = P(X \in g^{-1}(y)) \\ &= \sum_{x \in g^{-1}(y)} f_X(x), \end{aligned}$$

missä viimeinen yhtäsuuruus seuraa täysadditiivisuudesta. □

Havainnollistetaan tilannetta esimerkillä.

Esimerkki 8.2. Jatkoa esimerkille 6.7. Tarkastellaan jälleen nopanheittopeliä, jossa pelaajan voittosumma määräytyy nopan silmäluvun kautta siten, että voittosumma on nopan silmäluku korotettuna toiseen potenssiin. Olkoon X sm, joka kuvaa nopan silmälukua ja sm $Y = X^2$ tämän muunnos. Mikä on sm:n Y ptnf?

Nyt

$$f_Y(y) = \sum_{x \in \sqrt{y}} f_X(x),$$

eli

$$f_Y(y) = \frac{1}{6}, \quad \text{kun } y = \{1, 4, 9, 16, 25, 36\}$$

ja $f_Y(y) = 0$ muulloin.

Jatkuvan jakauman tapauksessa tilanne on hieman monimutkaisempi. Jos satunnaismuuttujalla X on jatkuva jakauma, sm:n $Y = g(X)$ jakauma ei välttämättä ole jatkuva. Esimerkiksi, jos g on paloittain vakio funktio, niin Y:n jakauma on diskreetti. Jatkuvien jakaumien tapauksessa helpointa on johtaa muunnoksen jakauma kertymäfunktion kautta derivoimalla.

Esimerkki 8.3. Olkoon $a < b$, $X \sim \mathrm{Tas}(0, 1)$ ja $Y = (b - a)X + a$. Mikä on sm:n Y tiheysfunktio?

Koska X on jatkuva sm, lasketaan Y:n tiheysfunktio ratkaisemalla ensin Y:n kertymäfunktio.

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P((b - a)X + a \le y) = P\left(X \le \frac{y - a}{b - a}\right) \\ &= F_X\left(\frac{y - a}{b - a}\right) \end{aligned}$$

Selvitetään vielä, missä joukossa F_Y on määritelty. Koska $X \sim \mathrm{Tas}(0, 1)$, niin $F_X(x) = x$, kun $0 < x < 1$. Siten $F_X\left(\frac{y-a}{b-a}\right) = \frac{y-a}{b-a}$, kun $0 < \frac{y-a}{b-a} < 1$ eli, kun $a < y < b$. Y:n kertymäfunktio on näin ollen $F_Y(y) = F_X\left(\frac{y-a}{b-a}\right) = \frac{y-a}{b-a}$, kun $a < y < b$. Tiheysfunktio saadaan nyt kertymäfunktiota derivoimalla:

$$f_Y(y) = F'_Y(y) = \frac{1}{b-a}.$$

Huomataan, että kyseessä on tasajakauman tiheysfunktio, eli Y noudattaa tasajakaumaa välillä (a, b).

Tilanteissa, joissa g on aidosti monotoninen (eli aidosti kasvava tai aidosti vähenevä) ja derivoituva (ja siten jatkuva) funktio, voidaan suoraan todeta, että jatkuvan satunnaismuuttujan X muunnoksella $Y = g(X)$ on jatkuva jakauma, jonka tiheysfunktio voidaan laskea seuraavan lauseen avulla.

Lause 8.4. *Satunnaismuuttujan muunnoksen tiheysfunktio. Jos X on jatkuva satunnaismuuttuja ja g on aidosti monotoninen ja derivoituva funktio, niin sm $Y = g(X)$ on jatkuva ja sen tiheysfunktio on*

$$f_Y(y) = \begin{cases} f_X(h(y))|h'(y)| & \text{, kun } y = g(x) \text{ jollakin } x \\ 0 & \text{muuten,} \end{cases}$$

missä $h(x) = g^{-1}(x)$.

Todistus. Esitetään todistus tilanteessa, kun g on aidosti kasvava. Oletetaan, että $y = g(x)$ jollakin x ja $h = g^{-1}$. Nyt, kun $Y = g(X)$,

$$F_Y(y) = P(g(X) \leq y) = P(X \leq h(y)) = F_X(h(y)).$$

Derivoimalla saadaan

$$f_Y(y) = f_X(h(y)) \cdot h'(y),$$

mikä on yhtäpitävää lauseen 8.4 kanssa, koska $h(y)$ on kasvava, jolloin sen derivaatta on epänegatiivinen.

Kun $y \neq g(x)$ millään x, $F_Y(y)$ on joko 0 tai 1, ja $f_Y(y) = 0$. □

Havainnollistetaan lauseen 8.4 menetelmää esimerkillä.

Esimerkki 8.5. Olkoon X jatkuva, epänegatiivinen satunnaismuuttuja tiheysfunktiolla f ja $Y = X^n$. Etsi Y:n tiheysfunktio f_Y.

Jos $g(x) = x^n$, niin $h(y) = g^{-1}(y) = y^{\frac{1}{n}}$ ja $h'(y) = \frac{1}{n} y^{\frac{1}{n-1}}$. Nyt, koska g on aidosti derivoituva ja monotoninen funktio (kun x ≥ 0), niin

$$f_Y(y) = f(y^{\frac{1}{n}})\frac{1}{n} y^{\frac{1}{n-1}}.$$

8.2 Yhteisjakaumat

Tähän asti olemme tarkastelleet pääasiassa yksiulotteisia todennäköisyysjakaumia, eli jakaumia, jossa on vain yksi satunnaismuuttuja. Moniin satunnaisilmiöihin vaikuttavat kuitenkin samanaikaisesti useat satunnaiset tekijät, jolloin tarvitaan yhteisjakaumaa mallintamaan satunnaismuuttujien yhteisvaikutusta tutkittuun ilmiöön. Satunnaismuuttujien määrä ei periaatteessa ole mitenkään rajoitettu ja monissa tosielämän malleissa voikin olla kymmeniä tai satoja satunnaismuuttujia. Moniulotteisia jakaumia käsitellään lisää myöhemmillä tilastotieteen ja todennäköisyyslaskennan kursseilla. Tässä kappaleessa keskitymme käsittelemään yhteisjakauman erikoistapausta, kahden satunnaismuuttujan muodostamaa jakaumaa eli kaksiulotteista yhteisjakaumaa.

Yksiulotteisen satunnaismuuttujan määritelmä (4.1 ja luvun 4 alku) yleistyy helposti kaksiulotteisille (ja ylipäätään moniulotteisille) satunnaismuuttujille.

Määritelmä 8.6. *Yhteisjakauma.*[1] Olkoot X ja Y satunnaismuuttujia, $X : \Omega \to \mathbb{R}, Y : \Omega \to \mathbb{R}$. Satunnaismuuttujien X ja Y yhteisjakauma on kaksiulotteisen reaaliakselin $\mathbb{R}^2$ osajoukolla A määritelty funktio

$$P((X,Y) \in A) = P(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}).$$

Myös kaksiulotteisen jakauman kertymäfunktio, eli satunnaismuuttujien X ja Y *yhteiskertymäfunktio* määritellään hyvin intuitiivisesti samaan tapaan kuin yksiulotteisilla sm:illa.

Määritelmä 8.7. *Yhteiskertymäfunktio.*[2] Satunnaismuuttujien X ja Y yhteiskertymäfunktio määritellään kaavalla

$$F(x,y) = P(X \le x, Y \le y) \quad \text{kaikille} \quad x, y \in \mathbb{R}.$$

Satunnaismuuttujien X ja Y kertymäfunktiot saa johdettua yhteiskertymäfunktiosta raja-arvoina

$$\begin{aligned} F_X(x) &= P(X \le x) = P(X \le x, Y < \infty) = P\Big(\lim_{y\to\infty}(X \le x, Y \le y)\Big) \\ &= \lim_{y\to\infty} P(X \le x, Y \le y) = \lim_{y\to\infty} F(x,y) = F(x,\infty) \end{aligned}$$

Näitä X:n ja Y:n omia yksiulotteisia jakaumia kutsutaan *reunajakaumiksi.*[3]

[1]engl. *joint probability distribution*
[2]engl.*joint cumulative distribution function*
[3]engl. *marginal distribution*

Myös yhteisjakaumat voivat olla joko diskreettejä, jatkuvia tai eivät kumpaakaan. Esitetään ensin diskreetin yhteisjakauman määritelmä, sillä se on helppo hahmottaa. Satunnaismuuttujien X ja Y diskreettiydestä seuraa suoraan sm:ien yhteisjakauman diskreettiys.

Määritelmä 8.8. *Yhteispistetodennäköisyysfunktio (lyh. yptnf).*[4] Kun sekä X että Y ovat diskreettejeä satunnaismuuttujia, kutsutaan myös X:n ja Y:n yhteisjakaumaa diskreetiksi yhteisjakaumaksi. Tällöin satunnaismuuttjien X ja Y yhteispistetodennäköisyysfunktio määritellään todennäköisyytenä

$$f(x, y) = P(X = x, Y = y).$$

X:n ja Y:n pistetodennäköisyysfunktiot saadaan summaamalla toinen muuttuja pois yhteispistetodennäköisyysfunktiosta:

$$f_X(x) = P(X = x) = \sum_y f(x, y).$$

Seuraava esimerkki havainnollistaa diskreetin yhteisjakauman käsitettä.

Esimerkki 8.9. Trinomijakauma. Suomalainen, ruotsalainen ja norjalainen pelaavat neljä kierrosta mölkkyä. Koska suomalainen on harjoitellut kovasti, todennäköisyys, että hän voittaa yhden kierroksen on 0.8. Ruotsalaisen ja norjalaisen voittotodennäköisyydet ovat 0.1 kummallakin. Millä todennäköisyydellä suomalainen voittaa neljästä kierroksesta kaksi ollen siten paras ja ruotsalainen ja norjalainen voittavat kumpikin yhden, niin ettei kenellekään jää paha mieli?

Kyse on toistokokeesta, jossa on kolme mahdollista tulosvaihtoehtoa. Tällaista tilannetta voidaan mallintaa trinomijakaumalla. Jos $p_1, p_2 > 0$ siten, että $p_1 + p_2 < 1$ ja $n \geq 1$ positiivinen kokonaisluku, niin trinomijakauman $\mathrm{Trin}(n, (p1, p2))$ yptnf on

$$f(x, y) = \binom{n}{x, y, n-x-y} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y}, \quad \text{missä } x, y \geq 0, \quad x + y \leq n.$$

Nyt, kun sm X on suomalaisen voittojen lukumäärä, sm Y on ruotsalaisen voittojen lukumäärä, sm $4 - X - Y$ on norjalaisen voittojen lukumäärä ja merkitään A = ”suomalainen voittaa”, B = ”ruotsalainen voittaa” ja C = ”norjalainen voittaa”, niin voittojen jakautuminen noudattaa trinomijakaumaa, jossa $p_1 = P(A) = 0.8$, $p_2 = P(B) = 0.1$ ja $P(C) = 1 - p_1 - p_2$. Kysytty todennäköisyys saadaan nyt trinomijakauman ptnf:sta:

$$\begin{aligned} f(2, 1) &= \binom{4}{2, 1, 1} 0.8^2 0.1^1 (1 - 0.8 - 0.1)^{4-2-1} \\ &= \frac{4!}{2!1!1!} 0.64 \cdot 0.1 \cdot 0.1 \\ &= 0.0768 \end{aligned}$$

[4] engl. *joint probability mass function*

Taulukoidaan vielä kaikki mahdolliset lopputulokset ja niiden todennäköisyydet:

$P(X = i, Y = j)$

$i \quad \diagdown \quad j$	0	1	2	3	4	Rivisumma
0	0.0001	0.0004	0.0006	0.0004	0.0001	0.0016
1	0.0032	0.0096	0.0096	0.0032	0	0.0256
2	0.0384	0.0768	0.0384	0	0	0.1536
3	0.2048	0.2048	0	0	0	0.4096
4	0.4096	0	0	0	0	0.4096
Sarake-summa	0.6561	0.2916	0.0486	0.0036	0.0001	1

Taulukon rivi- ja sarakesummista voidaan myös tarkistaa, että reunajakaumat X ja Y ovat binomijakaumia. Tulos on intuitiivinen, sillä jos peliä tarkastellaan pelkästään suomalaisen näkökulmasta, on suomalaisen voittojen lukumäärä binomijakautunut $X \sim \text{Bin}(4, 0.8)$. Vastaavasti ruotsalaisen voittojen lukumäärä on binomijakautunut $Y \sim \text{Bin}(4, 0.1)$.

Huomautus 8.10. Reunajakaumat eivät yksiselitteisesti määrää yhteisjakaumaa. Esimerkiksi, jos reunajakaumat X ~Bernoulli($\frac{1}{2}$) ja Y ~Bernoulli($\frac{1}{2}$), niin sm:ien X ja Y yhteisjakaumalla voi olla mikä tahansa seuraavan taulukon määräämä yptnf, kun $0 \leq a \leq \frac{1}{2}$.

$P(X = i, Y = j)$

$i \quad \diagdown \quad j$	0	1	Rivisumma
0	a	$\frac{1}{2}$ - a	$\frac{1}{2}$
1	$\frac{1}{2} - a$	a	$\frac{1}{2}$
Sarake-summa	$\frac{1}{2}$	$\frac{1}{2}$	1

Siirrytään seuraavaksi jatkuviin yhteisjakaumiin. Nyt on syytä huomioida, että reunajakaumien jatkuvuudesta **ei välttämättä seuraa että yhteisjakauma olisi jatkuva**. Jatkuva yhteisjakauma määritellään seuraavasti:

Määritelmä 8.11. *Jatkuva yhteisjakauma.* Olkoot X ja Y jatkuvia satunnaismuuttujia. Nyt muuttujilla X ja Y on jatkuva yhteisjakauma jos on olemassa funktio $f(x, y)$, jolle pätee

$$P(X, Y) \in A = \iint_{(x,y)\in A} f(x, y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{kaikille} \quad A \subset \mathbb{R}^2,$$

Funktiota $f(x, y)$ kutsutaan satunnaismuuttujien X ja Y *yhteistiheysfunktioksi (lyh. ytf)*[5] . Jos voidaan määritellä

$$A = \{(x, y) : a < x < b, c < y < d\}$$

ja $\int_A |f| < \infty$ tai $f > 0$, yhteisjakauman todennäköisyyksiä voidaan laskea integroimalla ytf molempien muuttujien suhteen

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Jos X:n ja Y:n yhteisjakauma on jatkuva, ovat myös niiden reunajakaumat jatkuvia, ja niiden tiheysfunktiot saadaan integroimalla yhteistiheysfunktiota toisen muuttujan suhteen:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x.$$

Tarkastellaan esimerkkiä todennäköisyyksien laskemisesta jatkuvalle yhteisjakaumalle.

Esimerkki 8.12. Satunnaismuuttujien X ja Y yhteistiheysfunktio

$$f(x, y) = 3e^{-x}e^{-3y}, \quad \text{kun } x, y > 0$$

ja $f(x, y) = 0$, muuten. Lasketaan $P(X < 1, Y > 1)$.

Kahteen kertaan integroimalla saadaan

$$\begin{aligned}
P(X < 1, Y > 1) &= \int_1^{\infty} \int_0^1 3e^{-x}e^{-3y}\,\mathrm{d}x\,\mathrm{d}y = \int_1^{\infty} \Big/_{0}^{1} -3e^{-x-3y}\,\mathrm{d}y \\
&= \int_1^{\infty} -3e^{-1-3y} - -3e^{-0-3y}\,\mathrm{d}y = \int_1^{\infty} -3e^{-1-3y} + 3e^{-3y}\,\mathrm{d}y \\
&= \Big/_{1}^{\infty} e^{-1-3y} - e^{-3y} = e^{-\infty} - e^{-\infty} - e^{-1-3} + e^{-3} \\
&= -e^{-4} + e^{-3} \approx 0.031.
\end{aligned}$$

[5] engl. *joint probability density function*

Muistellaan vielä kappaleessa 3.3 esitetty riippumattomuuden käsite. Riippumattomuutta voi tulkita myös kaksiulotteisten jakaumien kautta. Satunnaismuuttujat X ja Y ovat riippumattomia jos ja vain jos niiden yptnf/ytf saadaan niiden reunapistetodennäköisyysfunktioiden/reunatiheysfunktioiden tulona.

Lause 8.13. *Jos diskreeteillä tai jatkuvilla satunnaismuuttujilla X ja Y on yhteisjakauma yhteispistetodennäköisyysfunktiolla tai yhteistiheysfunktiolla f, niin X ja Y ovat riippumattomia jos ja vain jos*

$$f(x,y) = f_X(x)f_Y(y).$$

Myös kaksiulotteisille satunnaismuuttujille voidaan määritellä muunnoksia, kuten teimme yksiulotteisille sm:ille kappaleessa 8.1. Emme tässä materiaalissa käsittele tiheysfunktion muuntokaavaa, mutta näytämme, miten kaksiulotteisen satunnaismuuttujan muunnoksen odotusarvo lasketaan. Käsitellään ensin diskreetti tapaus.

Lause 8.14. *Diskreetin yhteisjakauman odotusarvo. Jos diskreeteillä satunnaismuuttujilla X ja Y on yhteisjakauma pistetodennäköisyysfunktiolla f, ja on $Z = g(X,Y)$ jokin tämän jakauman diskreetti muunnos, niin*

$$E(Z) = \sum_y \sum_x g(x,y)f(x,y),$$

mikäli summa suppenee itseisesti.

Jatkuvalla yhteisjakaumalle odotusarvo lasketaan vastaavana integraalina.

Lause 8.15. *Jatkuvan yhteisjakauman odotusarvo. Olkoon f joidenkin jatkuvien satunnaismuuttujien X ja Y jatkuvan yhteisjakauman yhteistiheysfunktio, ja $Z = g(X,Y)$ jokin tämän jakauman muunnos. Tällöin*

$$E(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

mikäli summa suppenee itseisesti.

Kaksiulotteisten satunnaismuuttujien odotusarvot määritellään aina tällä tavoin muunnoksen kautta, eli muodostamalla jokin funktio, jolle voidaan laskea reaaliarvoinen odotusarvo. Mikäli ollaan kiinnostuneita kaksiulotteisen jakauman painopisteestä, ts. jakauman 'kaksiulotteisesta odotusarvosta', voidaan laskea reunajakaumien odotusarvot $E(X) = \mu_X$ ja $E(Y) = \mu_Y$ ja muodostaa näistä kaksiulotteinen odotusarvovektori (μ_X, μ_Y). Tarkastellaan vielä esimerkkiä kaksiulotteisen satunnaismuuttujan odotusarvon laskemisesta.

Esimerkki 8.16. Jatkoa esimerkille 8.12. Lasketaan E(XY).

Satunnaismuuttujilla X ja Y on yhteisjakauma yhteistiheysfunktiolla $f(x, y) = 3e^{-x}e^{-3y}$, kun $0 < x, y$. Lauseen 8.15 mukaan

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{0}^{\infty}\int_{0}^{\infty} xy \cdot 3e^{-x}e^{-3y}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{0}^{\infty} 3ye^{-3y}\,\mathrm{d}y \\ &= \frac{1}{3}. \end{aligned}$$

8.3 Todennäköisyyslaskuja R-ohjelmistolla

Tässä kappaleessa käsitellään lyhyesti muutaman esimerkin kautta todennäköisyyslaskujen ratkaisemista erityisesti tilastotieteilijöiden käyttämällä R-ohjelmistolla. R tarjoaa näppärän ympäristön todennäköisyyslaskennalle ja esimerkiksi eri jakaumien tiheys- ja kertymäfunktioiden arvot saa R:n avulla selville helpommin ja täsmällisemmin kuin perinteisistä taulukoista katsomalla. Materiaalin ei ole tarkoitus toimia kattavana R-oppaana, vaan pikemminkin esitellä R:ää auttavasti tuntevalle lukijalle muutamia tapoja ohjelman käyttöön todennäköisyyslaskennan parissa.

Aloitetaan esimerkillä kolikonheiton simuloimisesta.

Esimerkki 8.17. Seuraavalla koodinpätkällä voidaan simuloida reilun kolikon heittoa $n = 100$ kertaa. Koodia tulkitaan siten, että otoksen arvo 0 tarkoittaa klaavaa ja 1 kruunaa.

```
n = 100
tulos = sample(0:1, n, replace = TRUE)
lkm = sum(tulos == 1)
osuus = lkm / n
osuus
```

Viimeinen rivi tulostaa kruunien osuuden 100 heitolla. Kruunien teoreettinen osuus on tietenkin 0.5, mutta koska kyseessä on simulaatio, havaitut osuudet voivat poiketa teoreettisesta odotusarvosta joskus paljonkin. Toistetaan seuraavaksi koe ajamalla koodi uudestaan n:n arvoilla 10 000, 1 000 000, 100 000 000 ja arvioidaan silmämääräisesti, miten kruunien osuuden erotus odotetusta osuudesta pienenee, kun n kasvaa. Eräässä kokeessa saatiin seuraavat tulokset

n	kruunien lkm	kruunien osuus	virhe $\approx$
100	46	0.46	-0.04
10 000	4973	0.4973	-0.0027
1 000 000	499103	0.499103	-0.000897
100 000 000	50003060	0.50003060	+0.0000306

Karkeasti ottaen virhe vaikuttaa siis pienenevän noin kymmenesosaan otoskoon satakertaistuessa, eli kun n kasvatetaan satakertaiseksi, saadaan yksi merkitsevä desimaali lisää.

Seuraava esimerkki käsittelee jatkuvien jakaumien muunnoksen simuloimista.

Esimerkki 8.18. Opiskelija matkustaa bussilla Viikistä Kumpulaan. Bussin saapumisaika lähtöpysäkille on satunnaismuuttuja, joka on tasaisesti jakautunut välillä (0,5) (minuuttia). Bussimatkan kesto noudattaa normaalijakaumaa parametreilla $\mu = 14$, $\sigma^2 = 2^2$. Lisäksi opiskelija kävelee kotoaan bussipysäkille 2 minuuttia ja bussipysäkiltä kampukselle 3 minuuttia ja hän lähtee kotoaan 25 minuuttia ennen luennon alkua. Suoritetaan simulaatio opiskelijan ehtimisestä aamuluennoilleen, joita on syyslukukaudella 70. Seuraava koodi tulostaa, niiden luentojen lukumäärän, joilta opiskelija simulaatiossa myöhästyy.

```
n <- 70

x <- runif(n, 0, 5)
y <- rnorm(n, 14, 2)
s <- 2 + x + y + 3

ehti <- (s <= 25)
ehti_lkm <- sum(ehti)
myohastyi <- n - ehti_lkm
myohastyi
```

```
[1] 8
```

Eräässä simulaatiossa opiskelija myöhästyi siis aamuluennolta 8 kertaa. Selvitetään vielä kuinka moni myöhästymisistä olisi ollut vältettävissä, jos opiskelija olisikin juossut bussipysäkiltä kampukselle, jolloin 3 minuutin matka olisi taittunut kahdessa minuutissa. Tämä onnistuu muokkaamalla koodin loppuosaa.

```
s <- 2 + x + y + 2
```

```
ehti <- (s <= 25)
ehti_lkm <- sum(ehti)
myohastyi <- n - ehti_lkm
myohastyi
```

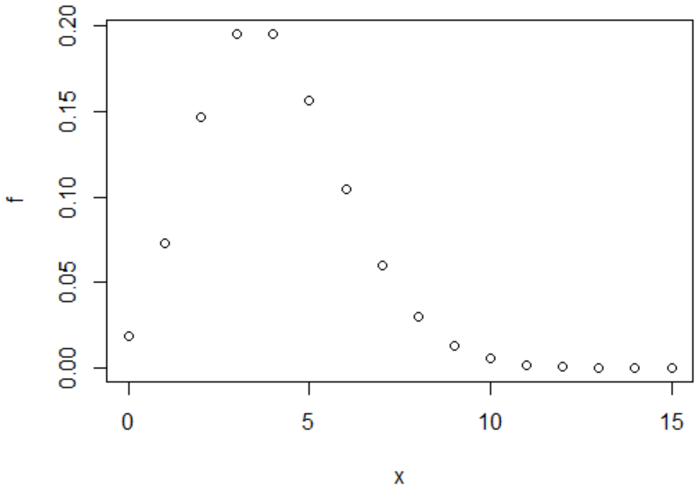
```
[1] 4
```

Opiskelija olisi siis myöhästynyt ainoastaan neljältä luennolta, eli neljä myöhästymistä olisi ollut vältettävissä viimeisen etapin juoksemisella.

Edellisessä esimerkissä käytettiin R:n komentoja `runif` ja `rnorm`, joilla generoitiin satunnaislukuja tasa- ja normaalijakaumasta. Komentojen nimet tulevat sanoista *random*, *uniform (distribution)* ja *normal (distribution)*. Tasa- ja normaalijakauman lisäksi R:stä löytyy valmiita funktioita monelle muullekin jakaumalle. Satunnaislukujen generoimisen lisäksi funktioilla voi laskea satunnaismuuttujien tiheys-, kertymä- tai kvantiilifunktioiden arvoja. Tästä lisää seuraavassa esimerkissä, jossa tutustumme funktioihin `dpois` ja `ppois` ja jakauman moodin ja mediaanin arvioimiseen jakauman kuvaajan avulla.

Esimerkki 8.19. Piirretään Poisson(4)-jakauman pistetodennäköisyysfunktion ja kertymäfunktion kuvaajat ja arvioidaan tämän perusteella jakauman moodia ja mediaania. Kuvaajat voidaan piirtää seuraavalla koodilla.

```
x <- 0:15
f = dpois(x, 4)
plot(x, f, 'p')
F = ppois(x, 4)
plot(x, F, 'p')
```



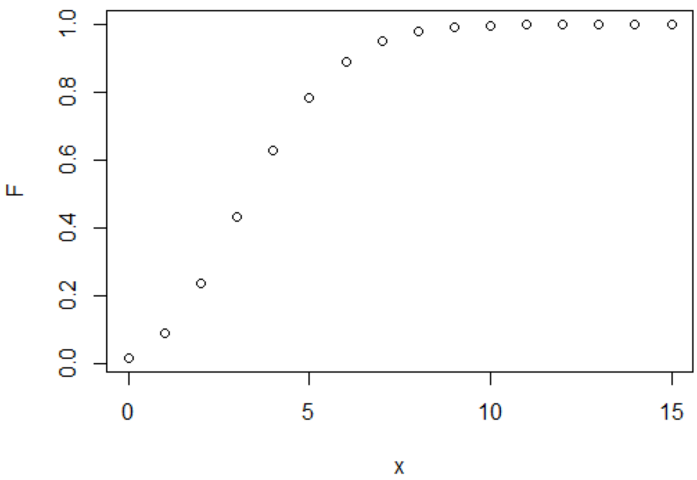
Kuva 8.1: Poisson(4)-jakauman pistetodennäköisyysfunktion kuvaaja.

Jakauman moodi eli tyyppiarvo on se muuttujan arvo, jonka esiintymisfrekvenssi on suurin. Yllä olevaa kuvaa tarkastelemalla Poisson(4)-jakauman moodi vaikuttaisi olevan pisteissä 3 ja 4.

Vastaavasti mediaani on järjestetyn joukon keskimmäinen piste. Todennäköisyysjakaumien yhteydessä tämä tarkoittaa siis pistettä x, jolle pätee

$$P(X \leq x) \geq \frac{1}{2} \quad \text{ja} \quad P(X \geq x) \geq \frac{1}{2}.$$

Alla olevaa kertymäfunktion kuvaajaa tarkastelemalla huomaamme, että kertymäfunktio vaikuttaisi arvolla 3 saavan arvon, joka on hieman suurempaa kuin 0.4, mutta kuitenkin alle 0.5. Arvolla 4 kertymäfunktion arvo on sen sijaan jo selvästi yli 0.6. Näin ollen jakauman mediaanin on oltava 4. Yksi tapa hahmottaa tilannetta on ajatella, että käymme läpi äärellistä, järjestettyä Poisson(4)-jakautunutta aineistoa. Kun olemme käyneet läpi kaikki luvut väliltä 0-3, olemme käsitelleet reilut 40% koko aineistosta. Kun olemme käyneet läpi vielä kaikki luvut 4, olemme käsitelleet yli 60 % koko aineistosta. Näin ollen sen luvun, jonka jälkeen olimme käsitelleet tasan 50% koko aineistosta, on oltava 4.



Kuva 8.2: Poisson(4)-jakauman kertymäfunktion kuvaaja.

Tarkistetaan havaintomme vielä laskemalla.

```
max(f)
dpois(3, 4)
dpois(4, 4)

[1] 0.1953668
[1] 0.1953668
[1] 0.1953668

ppois(4, 4)
1 - ppois(3, 4)

[1] 0.6288369
[1] 0.5665299
```

Jakauman tiheysfunktion maksimiarvo, oli sama kuin tiheysfunktion arvo kohdissa 3 ja 4 joten (ainakin) 3 ja 4 todella ovat jakauman moodeja. Lisäksi kertymäfunktion arvoja laskemalla saatiin selville, että

$$P(X \leq 4) \approx 0.629 \quad \text{ja} \quad P(X \geq 4) = 1 - P(X \leq 3) \approx 0.567,$$

joten 4 todella on jakauman mediaani. Oikeanpuoleisen yhtälön yhtäsuuruus seuraa siitä, että kyseessä on diskreetti jakauma, eli X ei voi saavuttaa muita kun kokonaislukuarvoja.

8.4 Harjoitustehtäviä

1. Olkoon $X \sim \mathrm{Tas}(0,1)$. Laske tn, että

   (a) $\sqrt{X}$:n ensimmäinen desimaali on 3,

   (b) X^2:n ensimmäinen desimaali on 3.

2. Olkoon $X \sim \mathrm{Exp}(1)$ ja $Y = 5X$. Johda Y:n kertymäfunktio ja tiheysfunktio ja tunnista Y:n jakauma.

3. Jatkoa esimerkeille 8.12 ja 8.16. Laske

   (a) $P(X > 2, Y < 2)$,

   (b) $P(X > 5)$,

   (c) $P(1 < X, Y < 3)$.

4. (Ross Exercise 6.2) Satunnaismuuttujilla X, Y ja Z on yhteisjakauma yptf:lla

   $$p(1,2,3) = p(3,2,1) = p(1,1,1) = p(2,3,2) = \frac{1}{4}.$$

   Etsi E(XYZ) ja E(XY + XZ + YZ).

5. (Ross Problem 6.9) Satunnaismuuttujilla X ja Y on yhteisjakauma yhteistiheysfunktiolla

   $$f(x,y) = \frac{6}{7}\Big(x^2 + \frac{xy}{2}\Big), \quad \text{missä } 0 < x < 1,\ 0 < y < 2.$$

   Laske

   (a) $P(X < \frac{1}{2}, Y > 1)$,

   (b) E(X),

   (c) E(Y).

6. Seuraava R-koodi arpoo n riippumatonta satunnaislukua väliltä (0,1), kertoo ne luvulla (b-a) ja lisää tuloksiin luvun a. Syntyneistä luvuista piirretään histogrammi.

   ```
   x = runif(n)
   y = (b-a)*x+a
   hist(y)
   ```

   (a) Kokeile koodia arvoilla $n = 10^6$, $a = 15$, $b = 20$. Mitä voit sanoa muunnettujen lukujen jakaumasta?

(b) Muokkaa koodia siten, että muunnosfunktio on `y = log(x)` (R:ssä `log` tarkoittaa luonnollista logaritmia). Mitä jakaumaa muunnetut luvut näyttävät nyt noudattavan?

Kiitos lukijoille!

Kirjallisuutta

[1] Pekka Tuominen: Todennäköisyyslaskenta I. 10. muuttumaton painos, Limes ry, 2010.

[2] Petri Koistinen: Todennäköisyyslaskenta. Luentomoniste, Helsingin yliopisto, 2013.

[3] Sheldon Ross: First Course in Probability. 8. painos, Upper Saddle River, N.J., Pearson Prentice Hall, 2010.

Liite A

Jakaumia

Diskreettejä jakaumia

Jakauma	Arvojoukko	Parametrit	Ptnf $f_X(k)$	$E(X)$	$\mathrm{Var}(X)$
Bernoulli(p)	$\{0,1\}$	$p \in (0,1)$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Bin(n,p)	$\{0,1,\ldots,n\}$	$n \in \mathbb{N}^+, p \in (0,1)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geom(p)	$\{0,1,2,\ldots\}$	$p \in (0,1)$	$p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\{0,1,2,\ldots\}$	$\lambda > 0$	$e^{-\lambda} \cdot \frac{\lambda^k}{k!}$	λ	λ
Hyperg(N,K,n)	$\{0,1,\ldots,n\}$	$N,K,n \in \mathbb{N}^+$, $n \leq N,\ K \leq N$	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	$n \cdot \frac{K}{N}$	$n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$

Jatkuvia jakaumia

Jakauma	Arvojoukko	Parametrit	Tf $f_X(x)$	$E(X)$	$\mathrm{Var}(X)$
Tas(a,b)	(a,b)	$a,b \in \mathbb{R},\ a < b$	$\frac{1}{b-a}$	$\frac{1}{2}(a+b)$	$\frac{(b-a)^2}{12}$
Exp(λ)	$(0,\infty)$	$\lambda > 0$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$N(\mu,\sigma^2)$	$\mathbb{R}$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

Standardinormaalijakauman kertymäfunktion arvoja

Standardinormaalijakauman kertymäunktion Φ arvoja, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Hakemisto

Alkeistapaus, 11, 22

Bayesin kaava, 50
Bernoulli-jakauma, 69
Binomijakauma, 70
Binomikerroin, 17

Cauchyn-Schwarxin epäyhtälö, 104

de Mérén ongelmat, 4
Diskreetti satunnaismuuttuja, 60

Ehdollinen todennäköisyys, 46
Eksponenttijakauma, 86

Geometrinen jakauma, 71

Hypergeometrinen jakauma, 71

Jakauma, 62
    Bernoulli-jakauma, 69
    binomijakauma, 70
    diskreetti tasajakauma, 69
    eksponenttijakauma, 86
    geometrinen jakauma, 71
    hypergeometrinen jakauma, 71
    jatkuva tasajakauma, 86
    kertymäfunktio, 67
    normaalijakauma, 88
    pistetodennäköisyysfunktio, 62
    Poisson-jakauma, 73
    tiheysfunktio, 81
Jakaumasuppeneminen, 112
Jatkuva satunnaismuuttuja, 81
Jatkuva tasajakauma, 86
Joukko, 8
    alkio, 8
    erotus, 8
    komplementti, 8
    kuuluminen, 8
    leikkaus, 8
    osajoukko, 8
    potenssijoukko, 10
    samat, 8
    tyhjä joukko, 8
    yhdiste, 8

Kertymäfunktio, 67
    diskreetti, 67
    jatkuva, 84
Keskeinen raja-arvolause, 112
Keskihajonta, 99
Klassinen todennäköisyys, 23
Kokonaistodennäköisyys, 49
Kombinaatio, 15
Kovarianssi, 102

Laskemisen perusperiaate, 11

Markovin epäyhtälö, 107
Multinomikerroin, 18
Muunnos, 91, 118
    odotusarvo, 97

Normaaliapproksimaatio, 113
    jatkuvuuskorjaus, 114
Normaalijakauma, 88

Odotusarvo
diskreetin satunnaismuuttujan, 74
jatkuvan satunnaismuuttujan, 90
lineaarisuus, 95
ominaisuuksia, 94
Ositus, 48
Otanta, 26
ilman takaisinpanoa, 29
takaisinpanolla, 29

Permutaatio, 15
Perusjoukko, 11, 22
Pistetodennäköisyysfunktio, 62
Poisson-jakauma, 73

Riippumattomuus, 52
komplementin, 54
satunnaismuuttujan, 77
satunnaismuuttujien muunnosten, 78
usean satunnaismuuttujan, 78
usean tapahtuman, 54

Satunnaiskoe, 22
Satunnaismuuttuja, 59
diskreetti, 60
jakauma, 62
jatkuva, 81
muunnos, 91
riippumattomuus, 77
varianssi, 99
Stokastinen konvergenssi, 110
Summaperiaate, 13
Suurten lukujen laki, 111

Tšebyševin epäyhtälö, 108
Tapahtuma, 22
Tasajakauma, 69
Tiheysfunktio, 81
Todennäköisyys
ehdollinen, 46
klassinen, 23
Todennäköisyysavaruus, 32
Toistokoe, 56
Tuloperiaate, 12

Variaatio, 15
Varianssi, 99
lineaarimuunnoksen, 100
ominaisuuksia, 101

Yhteisjakauma, 121
yhteiskertymäfunktio, 121
yhteispistetodennäköisyysfunktio, 122
yhteistiheysfunktio, 124