Tilastotieteen jatkokurssi

5+5 OP. 21.1.–6.5.2016. Luennoi yliopistonlehtori Pekka Pere.

Versio 10.5.2016.

Sisältö

Kuvat

1 Taustaa

1.1 Opiskelija ja tilastotiede

Tilastotiede on maailman jännittävin oppiaine. Sitä voi soveltaa likipitäen kaikille elämän- ja mielenkiinnonalueille. Vahvan teorian ja modernin tietoteknologian avulla se röntgensäteiden tapaan paljastaa silmälle näkymättömän todellisuuden.[1]

Paitsi että tilastotiede on hauskaa, se on hyödyllinen sivuaine ja monissa opinnäytteissä välttämätön työkalu. Yksi opiskelijan elämän säännönmukaisuuksia on katumus pro gradu –vaiheessa, että tilastotiedettä ei tullut opiskeltua enempää tai perusteellisemmin.

Opiskeluasenteeksi ei sovi kiire. Luetun sivumäärän päivää kohden ei ole suotavaa olla kovin suuri. On oleellista, että teoriaa opiskellaan kynä kädessä merkintöjä ja laskuja apupapereille ja sivujen syrjään tekemällä. Tilastotieteen ymmärtämiseen tarvitaan aikaa; tilastotieteen soveltamiseen tilasto-ohjelmisto (jollei aineisto ole pieni ja tehtävä hyvin yksinkertainen). Tilastotiedettä ei kannata yrittää opetella vain ”lukemalla”.

Tilastotieteen peruskurssit ovat ehkä pisimmälle kantavat kurssit elämässä. Oman alansa tehtäviin sijoittuva maisteri hyödyntää uransa alkuvaiheessa erityistietämystään — ja mahdollisesti kvantitatiivista tietoa. Työelämässä edetessään hän tyypillisesti erkaantuu asiantuntijaroolista ja siirtyy johtajuutta vaativiin tehtäviin. Edelleen ellei enenevässä määrin hän tekee päätöksiä kvantitatiiviseen tietoon perustuen.

Tilastotieteen perusopintojen tavoite on, että maisteri pystyy ymmärtämään ja lukemaan kriittisesti empiirisiä kvantitatiivisia tutkimuksia sekä tekemään itse pienimuotoisia sellaisia.

Tilastotiede ei ole helppoa. Onko se liian vaikeaa vaikkapa sosiaali- tai käyttäytymistieteiden opiskelijalle, joka ei ole tarvinnut matematiikkaa muissa opinnoissaan? Ei ole. Kaikki on suhteellista. Karl Pearson piti toivottomana Francis Galtonin yritystä 1889 saada antropologit ymmärtämään keksimäänsä korrelaatiota (Pearson 1930, 56–57). Harva ajatellee tänä päivänä, että antropologi ei voisi ymmärtää korrelaatiota, joka kuuluu nykyään yleissivistykseen. Entinen mahdottomuus on nykyinen itsestäänselvyys. Kaikki alla seuraava on opittavissa lukion lyhyen matematiikan tietopohjalta, jos on halukas ja valmis käyttämään aikaa opiskeluun. Ilo uudesta kielestä ja tavasta hahmottaa maailmaa sekä kyky ymmärtää ja tutkia sitä ovat palkintoja aherruksesta.

1.2 Tilastotieteen historiaa, nykyisyyttä ja tulevaisuutta

Nykyisenlaista tilastotiedettä pohjusti matematiikan ja todennäköisyyslaskennan kehittyminen. Todennäköisyyslaskennan syntyä motivoivat voiton mahdollisuuden arviointi uhkapelissä tai oikeudessa. Girolamo Cardano oli aktiivinen

[1] Nämä ajatukset sekä moni tieto tässä ja seuraavassa jaksossa löytyvät David Handin kirjasta ja artikkelista (2008, 2009). Hand oli Iso-Britannian kuninkaallisen tilastotieteellisen seuran presidentti 2008–2009 sekä *pro tem* 2010.

uhkapelaaja ja määritteli 1564 ehkä ensimmäisenä klassisen todennäköisyyden (jakso 2.2.1). Niklaus I Bernoulli väitteli 1709 todennäköisyyslaskennan soveltamisesta oikeustieteisiin. Thomas Bayes kuvasi sanallisesti Bayesin kaavan (jakso 2.6). Richard Price formalisoi ja julkaisi sen Bayesin kuoleman 1761 jälkeen 1763. Price käytti kaavaa kritisoidakseen David Hume'in kuuluisaa kritiikkiä ihmeiden mahdollisuutta kohtaan ja pohjimmiltaan pyrki kaavalla todistamaan, että Jumala on olemassa (Hooper 2013 ja Stigler 2016, luku 3). Pierre Simon Laplace sovelsi todennäköisyyslaskentaa oikeuden päätösten oikeellisuuden todennäköisyyden arviointiin 1700-luvun lopulla. Oleellisia läpimurtoja olivat Abraham de Moivren Normaalijakauman (jakso 4.3.1) johto 1733, Adrien Marie Legendren 1805 julkaisema pienimmän neliösumman menetelmä (luku 12) ja Laplacen 1810 johtama keskeinen raja-arvolause (jakso 4.4).

John Arbuthnot teki ensimmäisenä tilastotieteelliseksi testiksi tulkittavan dataan ja todennäköisyyteen pohjautuvan laskun 1710. Hänkin pyrki todistamaan Jumalan olemassaolon laskullaan.

Numeerista aineistoa eli dataa on summeerattu pitkään. Sumerilainen frekvenssitaulukko (jakso 11.2.2) tunnetaan ajalta noin 3000 vuotta eKr. (Stigler 2016, 27). Vaikka keskiarvo oli tunnettu matemaattisena käsitteenä niinikään kauan, datan kuvaamisessa se otettiin käyttöön vasta 1600-luvun lopulla. Keskiarvo oli yleisessä käytössä tähtitieteessä ja maanmittauksessa 1800-luvulle tultaessa ja yleistyi yhteiskunnassa 1830-luvulta lähtien (mts:t 26, 29 ja 33). Ensimmäinen yhteiskuntatilastotieteilijä belgialainen Adolphe Quetelet loi tuolloin käsitteen *keskivertoihminen* (*l'homme moyen*). Quetelet mittasi ihmisryhmistä erilaisia fyysisiä ominaisuuksia tai moraalisia tai muita taipumuksia ja vertasi ryhmiä niitä edustavien keskivertoihmisten avulla. Quetelet kutsui tiedettään sosiaalifysiikaksi (*sociale physique*). (Stigler 1986, luku 5 ja 1999, luku 2.)

Tilastotiede-sanan (*Statswissenschaft*) julkaisi ensimmäisenä Gottfried Achenwall 1749. Sanalla tarkoitettiin valtion suorittamaa tiedonkeruuta ja -käyttöä. (Rao 2007 sekä Larsen ja Marx 2001, 8.) Zachris Topelius kuvasi tilastotieteen tässä hengessä (1905, s. 448):

> Kuinka maamme edistyy varallisuudessa uuden ajan elähyttävien voimien kautta, sen näemme silmäimme edessä, ja siitä kertoo nykyinen tilastotiede eli valtiotiede. Samoin kuin historia meille kuvaa menneen, jo päättyneen ajan, samoin tilastotiede merkitsee nykyajan olot, jotka vielä ovat tekeillä, ja vertaa niitä toisiinsa tai entisiin aikoihin. Tämä on opettavaista. Se osoittaa meille selvillä numeroilla, edistyykö vai taantuuko maa varallisuudessa, väkiluvussa, hyvissä tavoissa, tiedossa ja monessa muussa. Vuodesta 1865 on Helsingissä tilastollinen virkakunta, joka vuosittain ikäänkuin vaa'alla punnitsee minkä arvoinen maa on.

Suomenkielinen tilasto-sana painettiin ilmeisesti ensimmäisen kerran vuonna 1848 Paavo Tikkasen kirjassa "Suomen suuriruhtinanmaan nykynen tilasto: yritös alkeis- ja rahwaan-kouluin tarpeeksi".

Vähitellen tilastotiede vapautui kytköksestä valtioon. Tutkittavat aineistot saattoivat olla elämän kaikilta alueilta. Pitkään tilastotiede miellettiin opiksi aineiston kuvailusta ja yleisten empiiristen lakien hauksi. Kun Statistical Section of the British Association perustettiin 1832, todettiin sen intressinä olevan

> – – tosiasiat, jotka liittyvät ihmisiin ja ovat ilmaistavissa numeroilla ja joista

vaikuttaa voitavan päätellä yleisiä lakeja.[2]

Tieteellisen murroksen lähtölaskenta alkoi 1877, jolloin Francis Galton havaitsi regression (luku 12). Uuden aikakauden portit todella avautuivat 1885 ja 1888, jolloin Galton oivalsi regression syvällisemmin ja keksi korrelaation ja 1896, jolloin Karl Pearson muotoili nykymuotoisen korrelaatiokertoimen. Tilastotieteen fokus muuttui aineistojen kuvailusta hypoteesien testaukseen ja tilastolliseen päättelyyn Pearsonin 1900 julkaiseman χ^2-testin ja William Gossetin (Student-nimimerkillä) 1908 julkaiseman t-testin jälkeen (luku 11). Ehkä kautta-aikojen merkittävin tilastotieteilijä Ronald Fisher kehitti tilastotiedettä voimallisesti seuranneina vuosikymmeninä ja julkaisi 1935 empiiristä tutkimusta mullistaneen kirjan *The Design of Experiments*. Vuosia 1885–1935 on sanottu tilastotieteelliseksi valistuksen ajaksi (*statistical enlightenment*), koska tuona aikana tilastotieteellinen ymmärrys kehittyi räjähdysmäisesti ja tapa hahmottaa maailmaa myllertyi (Stigler 2010).

Fisher korosti todennäköisyyden käsitteen merkitystä ja tilastotieteen matemaattisuutta tilastotieteen nykyisiä perusteita pohjustaneissa artikkelissaan ja kirjassaan (1922, 311–312 ja 1925, i):

> Tilastotieteellisten menetelmien tavoite on aineiston tiivistäminen. Aineisto – – korvataan muutamalla suureella, jotka sisältävät mahdollisimman kattavasti aineistossa olevan relevantin informaation. – – todennäköisyys on perustavanlaatuisin tilastotieteellinen käsite.
>
> Tilastotiede on oleellisesti matematiikkaa sovellettuna havaittuun aineistoon.

Vaikka Fisher piti matematiikkaa tärkeänä, hän vastusti todellisista aineistoista ja ongelmista irrotettua matemaattista tilastotiedettä (Box 1978, 435–437). Nykyisenlaisen tilastotieteen isäksi kutsuttu Pearson samoin painotti tilastotieteen käytännön merkitystä (Porter 1986, 305).

Leo Harmaja (1939, 12) liitti tilastotieteen oppikirjassaan vielä väestötieteeseen:

> Tilastollista tutkimusta, varsinkin yhteiskuntaoloihin ja tapahtumiin kohdistuvaa, sanotaan tavallisesti tilastotieteeksi. Mutta useiden arvovaltaisten tutkijain taholla on esitetty epäilyksiä, voidaanko tilastollista tutkimusta katsoa itsenäiseksi tieteenhaaraksi vai onko sitä pidettävä vain monien eri tieteenhaarojen käyttämänä aputieteenä tai erikoisena tutkimusmenetelmänä eli metodina tai ehkä kumpanakin. Niiden tutkijain taholla, joiden mielestä tilastollinen tutkimus on kehittynyt itsenäiseksi tieteeksi, sen varsinaiseen alaan on luettu ihmisten muodostaman yhteiskuntaelämän ilmiöt ja nimenomaan väestöolot. Täten tilastotiede on yhteiskuntatieteiden osa, ja ahtaammassa mielessä se on pääasiallisesti sama tieteenhaara kuin väestötiede – – .

Harmaja kertoi tilastotieteellisistä käsitteistä kuten korrelaatiosta ja regressiosuorasta (luku 12) mutta nipisti matematiikan minimiin (esitti kaavan yhdellä sivulla). Harmaja ilmeisesti katsoi kirjansa olevan ensimmäinen suomenkielinen tilastotieteen ”yleisesitys” mutta kertoi (mts. 5) ”todennäköisyyslaskusta” olevan ennestään suomenkielisiä ”esityksiä”. Sellaisia olivat J.W. Lindebergin (1927) ”Todennäköisyyslasku ja sen käytäntö tilastotieteessä. Alkeellinen esitys” ja A.M. Ritalan (1934) ”Todennäköisyyslaskennan taulukkoja”. Lindeberg kuvasi

[2] Tämä ja myöhemmät suomennokset ovat luennoitsijan.

kirjaansa matemaattisen tilastotieteen oppikirjaksi, mutta nykypäivän näkökulmasta se on ensimmäinen suomenkielinen tilastotieteen oppikirja. Jälkimmäinen oli tehty ”lääketieteellisissä y.m. tilastollisissa tutkimuksissa” käytettäväksi Pearsonin aloittaman menetelmällisen vallankumouksen jalanjäljissä.

Tilastotieteen historian kirjan alaotsikko kuvaa napakasti seuranneen: ”Kuinka tilastotiede mullisti tieteen 1900-luvulla” (Salsburg 2001). On vaikea keksiä empiiristä tutkimusalaa, joka olisi välttynyt tilastotieteen invaasiolta.

Kielitoimiston sanakirja pitää todennäköisyyslaskentaa tilastotieteen perustana[3]:

> Todennäköisyyslaskentaan perustuva tiede, joka tutkii tilastollisten tietojen keräämistä, käsittelyä ja tältä pohjalta tehtävää päättelyä – – .

Tämä on myös suomenkielisen Wikipedian määritelmä.

Monissa moderneissa määritelmissä todennäköisyyslaskenta ei ole enää esillä. Englanninkielinen Wikipedia ehdottaa[4]

> Tilastotiede on tiede datan keruusta, analyysistä, tulkinnasta, esittämisestä ja organisoinnista.

Samanhenkinen on Härdlen, Klinken ja Rönzin (2015, 1) määritelmä:

> Tilastotiede on tiede aineiston keruusta, kuvaamisesta ja tulkitsemista, eli se on empiirisen tutkimuksen työkalulaatikko.

Tilastotieteen ensimmäinen professuuri Suomessa perustettiin 1945 Helsingin yliopistoon (Liski ja Puntanen 1987, 19). Viran ensimmäinen haltija Leo Törnqvist määritteli tilastotieteen myös ilman viittausta satunnaisvaihteluun (Vasama ja Vartia 1971, 9):

> Tilastotiede on tietotuotannon tekniikkaa, jonka avulla voidaan suorittaa kvantitatiivisten tietojen joukkotuotantoa ja havaintoihin perustuvia tieteellisiä ja käytännöllisiä päätöksiä. Tilastotiede on siis yksikköjen muodostamaan joukkoon liittyvän numeerisen tietoaineiston keräämistä, analysointia ja tulkintaa käsittelevä tiede.

Törnqvistin määritelmä kattaa esimerkiksi Larsenin ja Marxin (2001), Digglen ja Chetwyndin (2011) sekä Dudewiczin ja Mishran (1988) lyhyet määritelmät:

> Tilastotiede on tiede otannasta.
>
> Tilastotiede on tiede aineiston keräämisestä ja tulkinnasta.
>
> Tilastotiede on tiede päätöksenteosta.

Rossin kuvaus (2010, 3):[5]

> Tilastotiede on aineistosta oppimisen taito. Se kattaa aineiston keruun, kuvauksen ja analyysin, joka usein johtaa johtopäätöksiin.

Ugarte ym. (2016, 97):

> Tilastotiede käsittelee datan keruuta, organisointia, summeerausta, analysointia ja esittämistä.

---

[3] Kielitoimiston sanakirjan verkkoversio. Kotimaisten kielten keskuksen verkkojulkaisuja 35. URN:NBN:fi:kotus-201433, ISSN 2323-3370. Julkaistu verkossa 11.11.2014.

[4] `Https://en.wikipedia.org/wiki/Statistics` (viitattu 20.1.2016).

[5] Ross listaa tilastotieteen määritelmiä vuodesta 1849 alkaen (mts. 11).

Näissäkään määritelmissä ei viitata matematiikkaan tai todennäköisyyslaskentaan.

Uudet sovellusalueet ovat perinteisesti vieneet tilastotiedettä eteenpäin:

- Maanviljelyskokeet inspiroivat kokeensuunnittelun teorian.
- Lääketieteen kysymykset veivät elinaika-analyysiin.
- Käyttäytymistieteiden ongelmat johtivat faktorianalyysiin.
- Yhteiskuntatieteelliset teemat tuottivat survey-tutkimusten teorian.

Näin on edelleen. Esimerkiksi geeniteknologia ja Internet ovat johtaneet uusiin menetelmiin valtavien tietomäärien tutkimiseksi.

Tietokoneen keksiminen ja kehitys ovat mullistaneet tilastotieteen soveltamistavan. Tilastotieteilijät — sekä sen soveltajat — voivat keskittyä entistä enemmän ymmärtämiseen laskennan sijaan. Samanlaista mullistusta ei ole tapahtunut tilastotieteen määritelmissä. Käsitteet ja periaatteet ovat pysyneet pitkälti samoina.

Tietotekniikan edistyminen, Internetin mukanaan tuoman tiedon hankkimisen helpottumisen ja kvantitatiivisten aineistojen analysoinnin kätevöittyminen ovat johtaneet monilla aloilla tilastotieteellisten menetelmien suosion kasvuun entisestään. Taloustieteessä empiiristen aineistojen analyyseihin perustuvien artikkelien osuus on kasvanut 48 %:sta 72 %:iin vuosina 1963–2011. Eritoten itse kerättyjen aineistojen analysointi on yleistynyt. (Hamermesh 2013.)

Nykyisen tilastotieteen valtavirtaa vahvasti muokanneen John Nelderin (1999) mielestä tilastotiede tarvitsisi uuden nimen, joka painottaisi tilastotieteen yhteyttä käytäntöön ja joka tekisi käsitteen ”soveltava tilastotiede” (*applied statistics*) tarpeettomaksi.

On esitetty, että tilastotieteen tärkein käsite olisi aineisto tai havainto. Aineisto voi olla empiirinen (soveltava tilastotiede) tai oletettu (teoreettinen tilastotiede). Todennäköisyys tai satunnaisuus ja aineisto (havainnot) ovat ehkä tilastotieteen ”syvimmät” käsitteet. Jälkimmäinen on vielä merkityksellisempi, sillä tilastotiedettä voidaan tehdä ilman todennäköisyyslaskentaa muttei ilman aineistoa (aineiston on oltava vähintään oletettu).

Olisiko aineisto- tai datatiede tilastotiedettä osuvampi nimitys?! Aineistosanaan eteen tulisi lisätä kvantitatiivinen tai sen tapainen määre, sillä kvalitatiivisten aineistojen analyysi on oma tieteenalansa. Data-sana viittaa onnistuneemmin numeeriseen aineistoon, mutta datatiede (*data science*) -nimitys on jo käytössä. Sillä tarkoitetaan tutkimusalaa, joka on tilastotieteen ja tietojenkäsittelytieteen välimaastoa. Datatieteessä aineistot ovat usein valtavia ja analyysit laskentaintensiivisiä. Tilastotieteen nimenvaihtoa datatieteeksi on pohdittu (esim. Speed 2014).

Penttisen (2014) mukaan

> tilastotieteellisen läpimurron seuraus on, että käsite tilasto tai tilastoaineisto on häviämässä korvautuen uusilla käsitteillä kuten data tai tietovaranto. Myös tilastoaineisto on monimuotoisempi sisältäen digitaalisia kuvia (joita myös analysoidaan), karttoja ja jopa tekstejä, ei pelkästään numeroita. Myös ammattinimi-

> ke tilastotieteilijä on häviämässä ja tilalle on tullut data-analyytikko (tai data scientist).

Työpaikkailmoituksissa haetaan nykyään sekä tilastotieteilijöitä että data-analyytikkoja.

Iso-Britannian kuninkaallisen tilastotieteellisen seuran presidentti Peter Diggle (2015) totesi virkaanastujaispuheessaan, että (englanninkielisen) Wikipedian määritelmät datatieteelle, informaatiotieteelle ja tilastotieteelle ovat päällekkäisiä:

> Datatiede on – – tiedon eristämistä datasta. – – Datatiede käyttää tekniikoita, jotka ovat peräisin monilta aloilta kuten matematiikasta, tilastotieteestä ja informaatioteknologiasta.
>
> Informaatiotiede on tieteidenvälinen ala, joka käsittelee lähinnä analyysia, keräystä, luokittelua, manipulaatiota, tallentamista, jäljitystä (*retrieval*), siirtoa, levittämistä ja informaation suojelemista.
>
> Tilastotiede on tiede aineiston keräämisestä, analyysista, tulkinnasta, esittämisestä ja organisoinnista.

Diggle kertoo oman näkemyksensä tilastotieteestä olevan lähempänä Wikipedian datatieteen määritelmää kuin sen tilastotieteen määritelmää. Informaatiotiede vaikuttaa hänestä liittyvän paljolti teknologiaan ja ohjelmistoihin. Datatieteen hän ajtteli aiemmin olevan vain uusi nimi tilastotieteelle muttei ajattele enää. Diggle painottaa, että tilastotieteessä pyritään minimoimaan aineiston tulkintaan liittyvää epävarmuutta.

Datatieteellisissä analyyseissa sivuutetaan monesti epävarmuus — osin ehkä koska tutkittavat aineistot ovat niin suuria, että esimerkiksi Pearsonin perinteen mukainen testaus ei olisi aina järkevää. Testien voima (luku 10) muodostuu niin suureksi, että käytännön kannalta mitättömät poikkeamat testattavasta nollahypoteesista muodostuvat tilastollisesti merkitseviksi. Tilastotiede on taas murroksessa.

Peter Phillips visioi (2003), että tulevaisuudessa tilastollisen mallin valinnan ja sovittamisen voi ”koneistaa” ja ”ulkoistaa”. Hänen ajatuksensa on, että aineistot ympäri maailmaa lähetetään Internetin välityksellä tietokoneelle, jonka automaattinen mallin valinta ja sovitus -ohjelmisto putkauttaa vastauksena valitsemansa ja estimoimansa mallin. Käyttäjän kannalta kyseessä olisi pitkälti ”musta laatikko”, joka toimii ”nappia painamalla”. Se mahdollistaisi nykyistä laajemman tilastotieteen käyttäjäkunnan ja soveltamisen.

Ei ole selvää, että näin äärimmäinen ”käyttäjäystävällisyys” olisi hyvä asia. Kyky tulkita tuloksia on keskeistä tilastotieteen hedelmällisessä soveltamisessa.

1.3 Kurssi ja tilastotiede

Kurssilla opiskeltavassa tilastotieteessä epävarmuuden arviointi on keskiössä, joten otetaan se mukaan tilastotieteen määritelmään. Määritelmä voisi olla vaikka tällainen:

> *Tilastotiede on tiede aineiston keräämisestä, kuvauksesta ja analyysista. Tilastotieteellä voidaan arvioida aineistoon ja siitä tehtäviin päätelmiin liittyvää epävarmuutta.*

Kurssilla pitkälti sivuutetaan konkreettiset aineiston keräämiseen, tallentamiseen, analyysiin ja kuvaukseen sekä oletusten tarkistamiseen liittyvät kysymykset. Niihin tavataan perehtyä survey- ja data-analyysin -erikoiskursseilla. Otanta on tilastotieteellisen analyysin peruskallio, joten sitä pohditaan silti luvussa 5.

Yhteiskunta- ja käyttäytymistieteelliset aineistot eivät yleensä ole suuria. Kurssilla painotus on peruskäsitteiden ja -menetelmien sekä pienehköjen aineistojen tutkimisessa. Niidenkin analyysi perustuu usein oletukselle suurehkosta havaintomäärästä. Pienten otosten analyysiin kehitettyihin tekniikoihin viitataan maininnoilla, jotka on todettu kurssivaatimuksiin kuulumattomiksi.

Kurssin tavoite on oppia ymmärtämään tilastotieteen teoriaa intuitiivisesti ja soveltamaan sitä. Matemaattiset perusteet esitetään pääpiirteissään. Matemaattisia yksityiskohtia sivuutetaan paikoin.

Esimerkeissä käytetään R-ohjelmistoa apuna (R Core Team 2015).

2 Todennäköisyyslaskentaa

2.1 Otosavaruus, tapahtuma ja satunnaismuuttuja

Koe on menettely, josta seuraa tulos, jonka tulosvaihtoehdot ovat määritellyt. Mielenkiinnon kohteena ovat kokeet, joiden lopputulokseen voi liittyä ja tyypillisesti liittyy satunnaisuutta. Kokeen tulos ei silloin ole välttämättä sama, jos koe uusitaan, vaikka olosuhteet olisivat muuttumattomat. Koe on tässä lavea käsite eikä viittaa esimerkiksi kokeeseen, jonka tarkoitus on vahvistaa tai saada uutta tietoa.

Kokeen otosavaruus S on sen kaikkien mahdollisten tulosvaihtoehtojen joukko. Monesti otosavaruuden tulosvaihtoehtoja on äärellinen määrä ja ne voidaan luetella: $S = \{a_1,a_2,\ldots,a_n\}$.

Tapahtuma A on otosavaruuden osajoukko. Alkeistapahtuma on yksinkertaisin mahdollinen tapahtuma. Luettelon tilanteessa alkeistapahtumia ovat a_i:t, $i = 1,\ldots,n$.

Esimerkki. Lantin heitto. Otosavaruus $S = \{kruuna, klaava\}$. Alkeistapahtumia on kaksi. □

Esimerkki. Nopan heitto. Otosavaruus $S = \{1,2,3,4,5,6\}$. Alkeistapahtumia on kuusi. Tapahtuma A voisi olla, että nopan silmäluku on parillinen: $A = \{2,4,6\}$. □

Esimerkki. Kahden nopan heitto. Otosavaruus $S = \{(1,1),(1,2),\ldots,(1,5),(1,6),(2,1),(2,2),\ldots,(6,5),(6,6)\}$. Sen alkiot ovat siis:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Alkeistapahtumia on kolmekymmentäkuusi. Tapahtuma A voisi olla esimerkiksi, että molempien noppien silmäluku on sama ja parillinen: $A = \{(2,2),(4,4),(6,6)\}$. □

Satunnaismuuttuja on kuvaus otosavaruudesta reaalilukujen joukkoon.[6] Satunnaismuuttujaa merkitään usein isolla ja sen reaalisaatiota pienellä kirjaimella (esim. Y ja y).

Esimerkki. Lantin heitto (jatkoa). Kuvataan "kruuna" ykköseksi ja "klaava" nollaksi. Lantin heittoon liittyy nyt satunnaismuuttuja Y, joka voi saada arvon 1 tai 0. Kun on saatu kruuna, $y = 1$. □

2.2 Todennäköisyyden määritelmiä

Todennäköisyys voidaan määritellä monin tavoin. Seuraavassa selitetään kolme keskeisintä määritelmää. Bruno de Finettin (1974, x) kuuluisa näkemys on, että todennäköisyyttä ei ole olemassa. Hän tarkoitti, ettei ole yhtä objektiivista todennäköisyyttä vaan on monia subjektiivisia näkemyksiä siitä. Savagen (1972, 2) mukaan sitten Baabelin tornin on harvoin ollut täydellisempää erimielisyyttä ja kommunikaatiokatkosta kuin kysymyksistä, mitä todennäköisyys on ja kuinka se liittyy tilastotieteeseen. Nykytilanne ei ole yhtä jyrkkä. Määrittelytavasta alla riippumatta todennäköisyys noudattaa samoja laskusääntöjä. Todennäköisyyden muita tulkintoja, filosofiaa ja historiaa selvittävät Gorroochurn (2012, luku 14), Gillies (2000), Hacking (2006), Niiniluoto (1975) ja von Plato (1994).

2.2.1 Klassinen todennäköisyys

Olkoot otosavaruuden $S = \{a_1,a_2,\ldots,a_n\}$ kaikki tulosvaihtoehdot symmetrisiä eli yhtätodennäköisiä:

$$\mathsf{P}(a_i) = \frac{1}{n}.$$

Yllä on merkitty todennäköisyyttä $\mathsf{P}(\cdot)$:llä.

Määritellään tapahtuma A symmetristen tulosvaihtoehtojen avulla. Tapahtuman A klassinen todennäköisyys on tällöin

$$\mathsf{P}(A) = \frac{A\text{:lle suotuisten tulosvaihtoehtojen lukumäärä}}{S\text{:n tulosvaihtoehtojen lukumäärä}}. \tag{1}$$

Esimerkki. Pallojen poimiminen. Olkoon pussissa sekaisin 10 vihreää, 20 oranssia ja 30 keltaista palloa. Poimitaan pussista pallo. Todennäköisyys, että saatu pallo on vihreä, on $1/6$:

$$\mathsf{P}(\text{vihreä}) = \frac{10}{10+20+30} = \frac{1}{6}. \ \square$$

[6] Satunnaismuuttuja tavataan määritellä tähän tapaan. Vaihtoehtoinen määritelmiä (Everitt 2003): Satunnaismuuttuja on muuttuja, joka saa arvoja jonkin todennäköisyysjakauman määräämällä tavalla.

Esimerkki. Kahden nopan heitto (jatkoa). Kaikki 36 silmälukuparia ovat yhtä todennäköisiä (tämä perustellaan myöhemmin). Jos $A = \{(2,2),(4,4),(6,6)\}$, niin tapahtuman A todennäköisyys on 1/12:

$$\mathsf{P}(A) = \frac{3}{36} = \frac{1}{12}. \ \square$$

Monet tilanteet eivät sovi klassisen todennäköisyyden määritelmään. Näin käy, jos vaikkapa tulosvaihtoehdot eivät ole symmetrisiä tai tulosvaihtoehtoja on ääretön määrä, jolloin ne eivät ole lueteltavissa.

2.2.2 Frekventistinen todennäköisyys

Frekventistinen todennäköisyys tapahtumalle A määritellään kaavalla

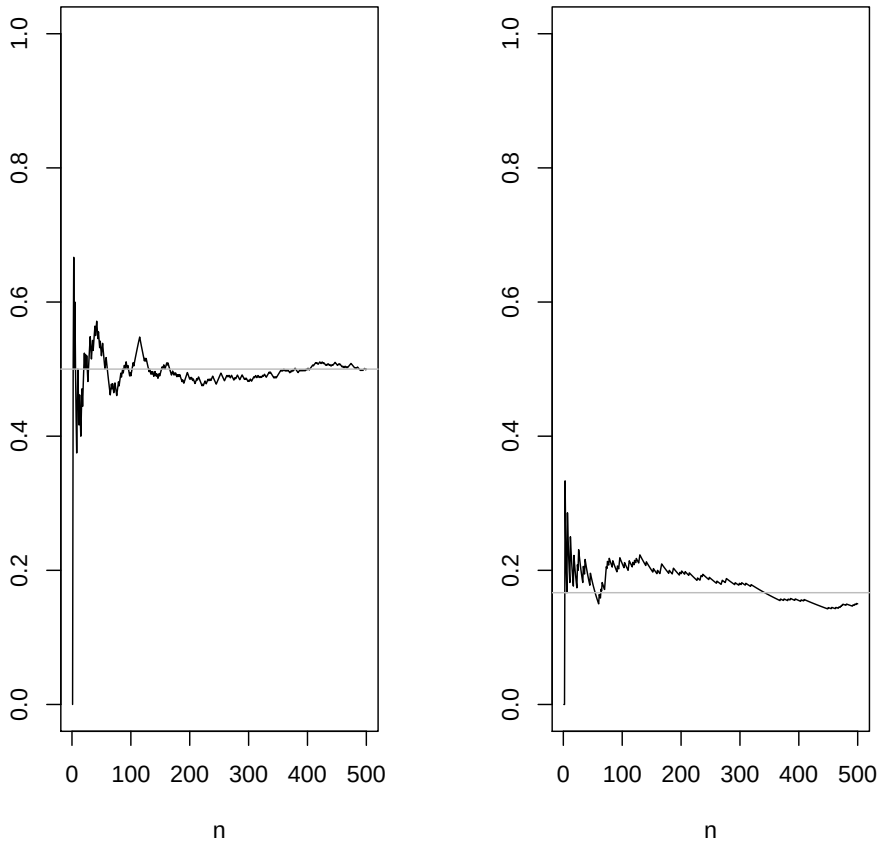
$$\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}_n(A) = \lim_{n\to\infty} \frac{f_n}{n}. \tag{2}$$

Kaava voidaan tulkita siten, että tapahtuman A todennäköisyys saadaan toistamalla riippumattomasti satunnaiskoetta n kertaa ja laskemalla tapahtuman A esiintymiskertojen (frekvenssin) f_n ja toistojen määrän n osamäärä, kun n kasvaa suureksi. Raja-arvon laskun oletuksena on, että satunnaiskoetta voidaan toistaa samanlaisissa olosuhteissa rajatta ja kokeet ovat toisistaan riippumattomia.

Esimerkki. Pallojen poimiminen (jatkoa). Todennäköisyys vihreän pallon nostamiselle on frekventistisen näkemyksen mukaan vihreiden pallojen nostojen määrä f_n jaettuna kaikkien poimintojen määrällä n, kun se kasvaa kohti ääretöntä (olettaen, että nostettu pallo palautetaan pussiin ja pallot sekoitetaan kunkin poiminnan jälkeen). Sadan poiminnan jälkeen $f_{100} = 20$ ja $\mathsf{P}_{100}(A) = 20/100 = 1/5$. Kahdensadan poiminnan kohdalla $f_{200} = 35$ ja $\mathsf{P}_{200}(A) = 35/200 = 0.175$. Kun poimintoja on 1000, $f_{1000} = 160$ ja $\mathsf{P}_{1000}(A) = 161/1000 = 0.16 \approx 1/6$. Vihreiden pallojen suhteellinen osuus menee suurilla havaintomäärillä kohti 1/6:tta. (Teoreettinen perustelu tälle esitetään myöhemmin.) $\square$

Esimerkki. Lantin heitto ja pallojen poimiminen (jatkoa). R-ohjelmistolla simuloitiin lantin heittämistä ja pallon nostoa. Kuvassa 1 vasemmalla havainnollistetaan kruunujen suhteellisen frekvenssin (f_n/n yhtälössä (2)) kehittymistä, kun heittojen lukumäärä $n = 1,\ldots,500$. Oikealla kuvassa ovat vastaavat vihreän pallon poimimisen suhteelliset frekvenssit. Osuudet näyttävät suppenevan kohti 1/2:hta ja 1/6:tta. $\square$

Todennäköisyyttä ei voida määritellä frekventistisesti, jos kyseessä on ainutkertainen tapahtuma. Tällaisiksi tapahtumiksi voisi mieltää esimerkiksi alkuräjähdyksen tai että tietty henkilö päättää ostaa maitoa kaupasta tiettynä päivämääränä tiettynä kellonaikana. Kaava yllä ei anna myöskään vastausta sellaisiin kysymyksiin kuin, mikä on todennäköisyys, että Jumala on olemassa.



Kuva 1: Kruunujen ja vihreiden pallojen suhteellisen osuuden kehitykset lantin heitossa ja pallon poiminnassa.

2.2.3 Subjektiivinen todennäköisyys

Ihmisillä on eri mielipiteitä ja niin myös todennäköisyyksistä. Näitä yksilöllisiä todennäköisyyksiä — tai uskomuksen asteita — kutsutaan subjektiivisiksi todennäköisyyksiksi. Ne ovat subjektiivisia, koska ihmiset saattavat arvioida todennäköisyyksiä erilailla jopa samojen tietojen perusteella.

Esimerkki. Talouden asiantuntijat esittivät Ylen uutisissa 15.5.2012[7] suuresti toisistaan poikkeavia arvioita (0.50:n ja 0.80:n välillä) todennäköisyydestä, jol-

[7] `Hhttp://yle.fi/uutiset/asiantuntijat_kreikka_voi_aiheuttaa_suomelle_jopa_miljarditappiot/6096707` (viitattu 24.1.2016).

la Kreikka eroaa eurosta (uutisessa todennäköisyydet on esitetty prosentteina 0:n ja 100:n välillä):

> Johtavat talousasiantuntijat pitävät valtiovarainministeriön arviota Kreikasta mahdollisesti aiheutuvista tappioista liian optimistisena. Ministeriö arvioi viime viikolla, että tappioita tulisi Suomelle korkeintaan 400 miljoonaa euroa. – – Kaikki asiantuntijat pitivät todennäköisenä, että Kreikka eroaa eurosta. Arviot eron todennäköisyydestä vaihtelevat 50–80 prosentin välillä.
>
> Pasi Holm, PTT:n toimitusjohtaja
>
> Saattaa olla, että kahdenvälisen lainan tappiot ovat 500 miljoonaa euroa ja väliaikaisen kriisirahaston kautta 140 miljoonaa euroa. Lisäksi EKP:n ja eurojärjestelmän kautta joitakin satoja miljoonia euroa. – Kreikan eurosta eroamisen todennäköisyys: 70 prosenttia.
>
> Seija Ilmakunnas, Palkansaajien tutkimuslaitoksen johtaja
>
> Pidän mahdollisena, että koko kahdenvälinen laina (miljardi euroa) menetetään. ERVV-osuuden riski on pienempi, mutta sitä ei voi arvioida tuntematta vakuusjärjestelyn yksityiskohtia. – Kreikan eurosta eroamisen todennäköisyys: 50 prosenttia.
>
> Vesa Kanniainen, Kansantaloustieteen professori
>
> Kokonaistappio 2–3 miljardia euroa, sisältäen lainat ja takuut sekä eurojärjestelmän kautta syntyvät pääomatappiot. – Kreikan eurosta eroamisen todennäköisyys: 75 prosenttia.
>
> Vesa Vihriälä, Etlan toimitusjohtaja
>
> Maksimitappio voisi olla 1.84 miljardia euroa. Tappio jäänee tätä pienemmäksi, mutta ei välttämättä VM:n arvioimaan 400 miljoonaan. Lisäksi Suomen Pankille voi tulla tappioita siitä, jos Kreikan keskuspankki ei pysty vastaamaan veloistaan EKP:lle. – Kreikan eurosta eroamisen todennäköisyys: 80 prosenttia.

Todennäköisyyden frekventistinen tulkinta ei ole luonteva Kreikan eurosta eroamisen mahdollisuutta arvioitaessa. Yksikään maa ei ole aiemmin eronnut eurosta, eikä eroon johtavista seikoista ole kaikkea tarvittavaa tietoa. Jokaikinen Ylen haastattelema asiantuntija kertoo omista lähtökohdistaan eri suuruisen arvion eron todennäköisyydestä. Ne ovat heidän subjektiivisia arvioitaan Kreikan eurosta eroamisen todennäköisyydestä. □

Subjektiivinen todennäköisyys voi olla mahdollista selvittää, vaikka ihminen ei sitä kertoisi tai osaisi kertoa yhtä avoimesti kuin esimerkissä edellä. Yksilön subjektiivinen todennäköisyys määritellään usein vedonlyöntisuhteen (*odds*)

$$\frac{S}{V} = \frac{\mathsf{P}(A)}{1 - \mathsf{P}(A)}, \tag{3}$$

avulla. Yllä $S > 0$ ja $V > 0$ (euroa) ovat sijoitus (mahdollisesti tappio) ja voitto vedonlyönnissä tapahtumasta A ja $\mathsf{P}(A) \in (0,1)$ on yksilön subjektiivinen todennäköisyys tapahtumalle A. ("$\in$" luetaan "kuuluu joukkoon" tai tässä yhteydessä "kuuluu välille".) Yhtäsuuruus perustellaan alla.

Idea on, että yksilö ilmaisee subjektiivisen todennäköisyytensä reilussa (*fair*) vedonlyönnissä. Oletetaan, että pelaaja sijoittaa vedonlyöntiin S euroa, voittaa V euroa, jos A tapahtuu mutta häviää sijoituksensa S euroa, jos A:ta ei tapahdu. Sille todennäköisyys on pelaajan mielestä $1-\mathsf{P}(A)$. Tällainen vedonlyönti on reilu, jos pelaajan mielestä hän ei odotetun tuoton (jakso 4.1) mielessä hyödy vedonlyönnistä eli odotettu tuotto on 0 euroa:

$$\mathsf{P}(A) \times V + [1 - \mathsf{P}(A)] \times (-S) = 0.$$

Jaetaan yhtälö puolittain V:llä ja todetaan, että yhtäsuuruuden (3) täytyy olla voimassa, jotta yhtälö pätisi:

$$\mathsf{P}(A) \times 1 - [1 - \mathsf{P}(A)] \times \frac{S}{V} = 0.$$

Pelaajan subjektiivinen todennäköisyys tapahtumalle A saadaan ratkaisemalla yhtäsuuruus (3) $\mathrm{P}(A)$:n suhteen:

$$\mathsf{P}(A) = \frac{S}{S+V}.$$

Subjektiivinen todennäköisyys voidaan siten selvittää, kun tiedetään vedonlyöntisuhde pelaajan mielestä reilussa vedonlyönnissä.

Esimerkki. Tentin läpäisy. Opiskelija on varsin vakuuttunut, että hän läpäisee tentin. Hän on valmis lyömään siitä vetoa vedonlyöntisuhteella $S/V = 4/1$. Hänen subjektiivinen todennäköisyytensä tentin läpäisemiselle on tällöin $S/(S+V) = 4/(4+1) = 0.8$. □

Vedonlyöntiasetelma on keinotekoinen: Moni ei suostuisi uhkapeliin, jonka odotettu tuotto on nolla. Joku ei ryhtyisi uhkapeliin koskaan. Asetelma ei toimi monen mielestä tärkeimmän kysymyksen kohdalla, jatkuuko elämä kuoleman jälkeen. Uskomuksesta riippumatta kannattaisi aina lyödä tuonpuoleisen elämän puolesta vetoa, koska veto ratkeaisi ainoastaan elämän jatkuessa ja ainoastaan silloin olisi mahdollista kerätä vedon tuotto. Asiasta voisi huoletta lyödä vetoa mielivaltaisen suurella vedonlyöntisuhteella, mikä edellä olevan laskun mukaan merkitsisi uskoa tuonpuoleiseen elämään oleellisesti todennäköisyydellä 1. (Mellor 1973.) Johto edellä on silti tavanomainen subjektiivisen todennäköisyyden yhteydessä. Ajatus lienee, että ”pakottamalla” yksilö vedonlyöntiin, hänen subjektiivinen todennäköisyytensä määrittyy tarkasti. Vedonlyöntisuhde S/V on helpohko hahmottaa, ja johto yhdistää sen kätevästi subjektiivisen todennäköisyyden suuruuteen.

Muitakin menetelmiä selvittää subjektiivinen todennäköisyys on. Pähkinänkuoriesitys on Haighin (2012) kirjassa.

2.3 Joukko-oppia

Tapahtumat koostuvat otosavaruuden S osajoukoista. Olkoot A ja B kaksi tapahtumaa (osajoukkoa) siinä.

Tapahtumien A ja B yhdiste (unioni) koostuu niistä tulosvaihtoehdoista (alkioista), jotka kuuluvat joko A:han tai B:hen tai molempiin. A:n ja B:n yhdistettä merkitään $A \cup B$.

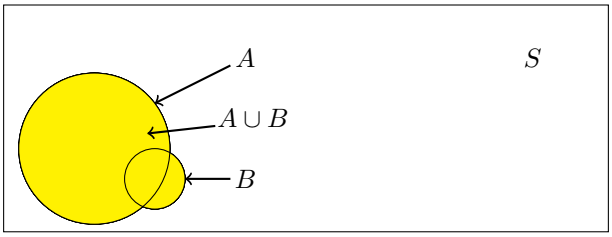
Tapahtumien A ja B leikkaus koostuu niistä tulosvaihtoehdoista, jotka kuuluvat sekä A:han tai B:hen. A:n ja B:n leikkausta merkitään $A \cap B$.

Jos A tapahtuu aina, kun B tapahtuu, B:n määrittävät tulosvaihtoehdot ovat A:n määrittävien tulosvaihtoehtojen osajoukko. Tällöin merkitään $B \subseteq A$ tai $B \subset A$, jos A ja B eivät ole samoja tapahtumia. Jälkimmäisessä tapauksessa B:n sanotaan olevan A:n aito osajoukko.

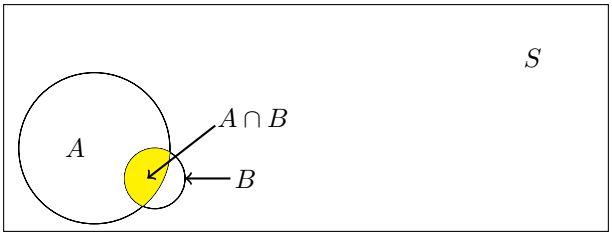
Jos $A \cap B = \emptyset$ (tyhjä joukko), niin A ja B ovat erillisiä eli toisensa poissulkevia tapahtumia.

Tapahtuman A komplementti muodostuu niistä tulosvaihtoehdoista, jotka eivät kuulu A:han. A:n komplementtia merkitään A^C:lla.

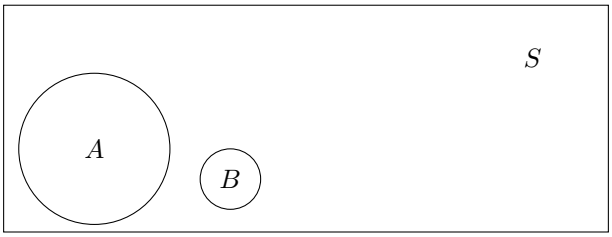
Joukko-opillisia operaatioita havainnollistetaan usein Venn-diagrammeilla (kuvat 2–8):



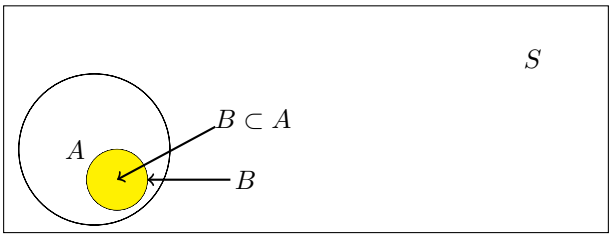
Kuva 2: A:n ja B:n yhdiste.



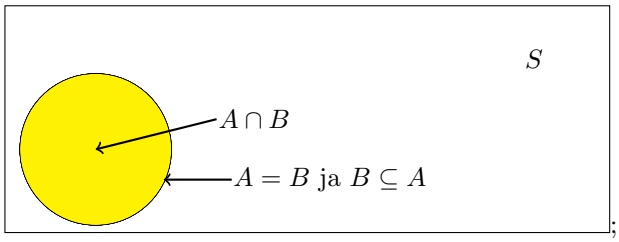
Kuva 3: A:n ja B:n leikkaus.



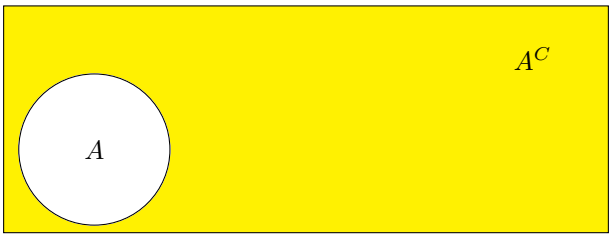
Kuva 4: A ja B ovat erillisiä.



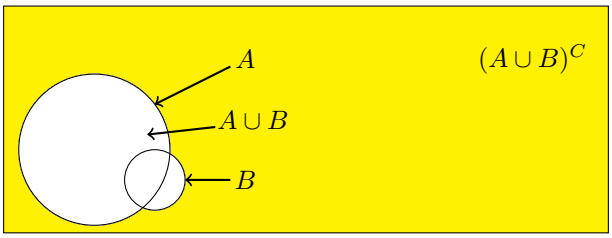
Kuva 5: Kaikki B:n tapahtumat ovat myös A:n tapahtumia.



Kuva 6: A ja B ovat sama tapahtuma.



Kuva 7: A:n komplementti



Kuva 8: A:n ja B:n yhdisteen komplementti.

2.4 Todennäköisyyslaskennan laskusääntöjä

Venkateshin mielestä (2013, xxvii–xviii ja 3) todennäköisyyslaskenta on matematiikan aloista geometrian ohella intuitiivisin ja käytännönläheisin, ja todennäköisyyslaskennan intuitiivisuus on myös hänen opiskelijoidensa kokemus. Hand (2014, 69) kuvaa todennäköisyyslaskennan päinvastoin tunnetusti matematiikan aloista intuitionvastaisimpana. Cobbista (2015) todennäköisyyden käsite on käsistälipsuva ja intuition ja teorian välinen kuilu on matematiikan aloista suurin todennäköisyyslaskennassa. Tijms (2012, 214) viittaa näkemykseen, että millään muulla matematiikan alalla ei asiantuntija erehdy yhtä helposti. Kirjoittajan näkemys on, että todennäköisyyslaskenta on käytännönläheistä, mutta intuitio ja todennäköisyys eivät aina kohtaa, ja se on osa aiheen viehätystä. Ensiksi perehdytään yksinkertaisimpiin tilanteisiin.

Kuulukoon tapahtuma A otosavaruuteen S. Tapahtuman A todennäköisyys on vähintään 0:

$$\mathsf{P}(A) \geq 0.$$

Todennäköisyys, että tapahtuu jokin otosavaruuden tapahtumista, on 1:

$$\mathsf{P}(S) = 1.$$

Olkoot A ja B otosavaruuteen kuuluvia erillisiä tapahtumia ($A \cap B = \emptyset$). Niiden yhdisteen todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B).$$

Edellisistä oletuksista voidaan johtaa todennäköisyyslaskennan laskusäännöt alla.[8]

Minkä tahansa tapahtuman A todennäköisyys on korkeintaan 1:

$$\mathsf{P}(A) \leq 1.$$

A:n komplementin todennäköisyys on

$$\mathsf{P}(A^C) = 1 - \mathsf{P}(A).$$

Otosavaruuteen kuulumattoman eli loogisesti mahdottoman tapahtuman todennäköisyys on 0:

$$\mathsf{P}(\emptyset) = 0.$$

Jos $B \subseteq A$, niin

$$\mathsf{P}(B) \leq \mathsf{P}(A).$$

[8] Mikäli otosavaruudessa on ääretön määrä tulosvaihtoehtoja, tarvitaan neljäs oletus: Olkoot tapahtumat $A_1, A_2, \ldots$ erillisiä ($A_i \cap A_j = \emptyset$ kaikille $i \neq j$) ja kuulukoot ne otosavaruuteen S. Tällöin yhdistetyn tapahtuman $\cup_{i=1}^{\infty} A_i$ todennäköisyys on $\mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$. Katso esimerkiksi tässä lähteenä käytettyä kirjaa Larsen ja Marx (mts:t 31–33). Samantapaisia todistuksia löytyy Blitzsteinin ja Hwangin (2015) sekä Venkateshin (2013) kirjoista.

Olkoot tapahtumat $A_1, \ldots, A_n$, erillisiä ($A_i \cap A_j = \emptyset$ kaikille $i \neq j$). Tällöin *erillisten tapahtumien yhteenlaskusäännön* mukaan yhdistetyn tapahtuman $\cup_{i=1}^n A_i$ todennäköisyys on

$$\mathsf{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathsf{P}(A_i). \tag{4}$$

Yllä $\cup_{i=1}^n A_i$ on kaikkien tapahtumien $A_1, \ldots, A_n$, yhdiste.

Yleistä yhteenlaskusääntöä käytetään, jos tapahtumat A ja B eivät ole erillisiä ($A \cap B \neq \emptyset$). Sen mukaan tapahtumien yhdisteen todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B). \tag{5}$$

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Yleinen yhteenlaskusääntö n:n tapahtuman $A_1, \ldots, A_n$ tilanteessa[9] on

$$\mathsf{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i=1}^{j} \sum_{j=1+1}^{n} \mathsf{P}(A_i \cap A_j) + \sum_{i=1}^{j} \sum_{j=1+1}^{k} \sum_{k=j+1}^{n} \mathsf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathsf{P}(A_1 \cap \cdots \cap A_n).$$

Kun tapahtumia on kaksi (A ja B), kaava typistyy säännöksi (5).

Säännöt ovat intuitiiviset. Ympyröiden kokoja kuvissa 2–8 voi ajatella tapahtumien todennäköisyyksinä, jos laatikon koko on yksi. Mahdottoman (todennäköisyyden mielessä) tapahtuman todennäköisyys on nolla ja varman yksi. Muunlaisen tapahtuman todennäköisyys on nollan ja yhden välillä. Erillisten tapahtumien A ja B yhdiste vastaa tapahtumaa, että jompikumpi niistä tapahtuu. Tällaisen yhdistetyn tapahtuman todennäköisyys on erillisten tapahtumien todennäköisyyksien summa (kuva 4). Tapahtuma A ja sen komplementti A^C ovat erillisiä ja kattavat koko otosavaruuden, joten niiden todennäköisyyksien summa on yksi (kuva 7). Kuva 3 selventää, että tapahtumista, jotka eivät ole erillisiä, muodostetun yhdistetyn tapahtuman todennäköisyyttä ei voi laskea suoraviivaisesti summaamalla tapahtumien todennäköisyyksiä. Summasta pitää vähentää todennäköisyys tulosvaihtoehdolle "molemmat tapahtuvat". Sen todennäköisyys tulisi muuten ynnättyä kahdesti (kuvat 2 ja 3). Kuvat 5 ja 6 havainnollistavat, että $\mathsf{P}(B) \leq \mathsf{P}(A)$, jos B tapahtuu aina, kun A tapahtuu.

Esimerkki. Nopan heitto (jatkoa). Silmäluku 3.5 ei kuulu otosavaruuteen, ja sen todennäköisyys on 0. Varmasti eli todennäköisyydellä 1 saadaan jokin otosavaruuden {1,2,3,4,5,6} muodostavista silmäluvuista eli alkeistapahtumista $\{1\}, \ldots, \{6\}$. Kunkin otosavaruuteen kuuluvan silmäluvun todennäköisyys on $0 < 1/6 < 1$. Kukin silmäluku on erillinen tapahtuma, sillä kahta silmälukua ei voi tulla samanaikaisesti.

[9] Blitzstein ja Hwang (2015, 26).

Olkoon tapahtuma $A = \{2,4,6\}$ ja tapahtuma $B = \{1\}$. Tapahtumat ovat erillisiä. Yhdistetyn tapahtuman $A \cup B = \{1,2,4,6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) = \frac{3}{6} + \frac{1}{6} = \frac{2}{3}.$$

Komplementtitapahtuman $(A \cup B)^C = \{3,5\}$ todennäköisyys voidaan laskea summana erillisten alkeistapahtumien todennäköisyyksistä

$$\mathsf{P}(A \cup B)^C = \frac{2}{6} = \frac{1}{3}$$

tai säännön

$$\mathsf{P}((A \cup B)^C) = 1 - \mathsf{P}(A \cup B) = 1 - \frac{2}{3} = \frac{1}{3}$$

avulla (vrt. kuva 8).

Olkoon $B = \{2\}$. Nyt A ja B eivät ole erillisiä: $A \cap B = \{2\}$. $\mathsf{P}(B) = \mathsf{P}(A \cap B) = 1/6$. Yhdistetyn tapahtuman $A \cup B = \{2,4,6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B) = \frac{3}{6} + \frac{1}{6} - \frac{1}{6} = \frac{1}{2}.$$

Se voitaisiin laskea myös vertaamalla (yhtätodennäköisten) suotuisten tulosvaihtoehtojen lukumäärää kaikkien tulosvaihtoehtojen lukumäärään:

$$\mathsf{P}(A \cup B) = \frac{3}{6} = \frac{1}{2}.$$

Olkoon $B = \{1,2\}$. A ja B eivät ole erillisiä: $A \cap B = \{2\}$. $\mathsf{P}(B) = 1/3$ ja $\mathsf{P}(A \cap B) = 1/6$. Yhdistetyn tapahtuman $A \cup B = \{1,2,4,6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \frac{3}{6} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}. \ \square$$

Esimerkki. Nopan heitto (jatkoa). Todennäköisyyslaskut nopan heitosta kuvaavat yhtälailla mitä tahansa koetta, jossa otosavaruus koostuu kuudesta yhtä todennäköisestä erillisestä tulosvaihtoehdosta. Sellaisia voisivat olla

- iskuryhmän valinta kuuden tasavahvan vapaaehtoisen sotilaan joukosta.
- lääkärin pohdinta potilaan oireiden syystä kuuden erillisen tasaveroisen selityksen joukosta.
- yhden tai useamman työntekijän valinta kuuden tasaveroisen hakijan joukosta. Hakijoiden ominaisuuksia kuten sukupuoli tai etninen tausta voidaan käsitellä silmälukujen tapaan ja laskea todennäköisyys, että valituksi tulee vaikkapa ei-kantasuomalainen nainen. (Parillinen silmäluku voisi vastata naissukupuolta, silmäluku $\{6\}$ ei-kantasuomalaista jne., jos hakijoista kolme on naisia ja heistä yksi on ei-kantasuomalainen.)

- äänestettävän kansanedustaehdokkaan valinta kuuden tasavahvan ehdokkaan joukosta. Ehdokkaisiin voitaisiin taas liittää ominaisuuksia silmälukujen tilalle kuvaamaan puoluetta, kantaa tiettyyn poliittiseen kysymykseen jne.
- sivuaineiden valinta yliopistossa kuuden yhtäkiinnostavan vaihtoehdon joukosta. (Niiden sisältö ja todellinen luonne voi olla opiskelijan näkökulmasta tuntematon ja satunnainen.)
- ylipäänsä yhden tai useamman asian, esineen tai ihmisen valinta kuuden tasavahvan vaihtoehdon joukosta.

Nopan heitto -laskut yleistyvät myös tilanteisiin, joissa tasavahvoja erillisiä vaihtoehtoja on kuudesta poikkeava määrä. Laskujen numeeriset tulokset eivät ole samat mutta periaatteet ovat. □

2.5 Ehdollinen todennäköisyys ja riippumattomuus

Olkoot A ja B tapahtumia otosavaruudessa ja $\mathsf{P}(B) > 0$ (mahdottomalle tapahtumalle ei voi ehdollistaa). A:n todennäköisyys ehdolla B on

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}. \tag{6}$$

Merkinnän "$\mid$" vasemmalla puolella on tapahtuma, jonka todennäköisyyttä määritellään ja oikealla puolella tapahtuma, jolle ehdollistetaan. Kaava on tärkeä. Siitä johdetaan myöhemmin tulosääntö (7) ja riippumattomuusehdot (8) ja (9).

Venn-diagrammit auttavat taas ymmärtämistä. Kuvassa 3 osajoukko $A \cap B$ on pieni osuus S:stä eli $\mathsf{P}(A \cap B)$ on pieni. Suhteutettuna osajoukon B kokoon $A \cap B$ on silti suuri ja siten $\mathsf{P}(A \mid B)$:kin on. A tapahtuu peräti aina kuvassa 5, jos B tapahtuu. Tällöin $\mathsf{P}(A \mid B) = 1$, vaikka $\mathsf{P}(A)$ olisi pieni. A:n ehdollinen todennäköisyys on 1 myös kuvan 6 tilanteessa, jossa A ja B ovat samat tapahtumat. Kun ehdollistetaan tapahtumalle B, otosavaruus S korvautuu B:llä.

Esimerkki. Nopan heitto (jatkoa). Olkoon $A = \{2,4,6\}$ ja $B = \{1\}$. Tällöin $A \cap B = \emptyset$, jonka todennäköisyys on nolla. Ehdollistaminen romauttaa todennäköisyyden:

$$\mathsf{P}(A) = \frac{1}{2}$$

mutta

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{0}{1/6} = 0$$

(kuva 4). Jos $B = \{2,4\}$, niin $A \cap B = \{2,4\}$. Nyt ehdollistaminen räjäyttää todennäköisyyden:

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{1/3}{1/3} = 1$$

(kuva 5). □

Esimerkki.[10] Olkoon todennäköisyys

- 0.2, että A tapahtuu mutta B ei tapahdu ($A \cap B^C$).
- 0.1, että B tapahtuu mutta A ei tapahdu ($B \cap A^C$).
- 0.6, että kumpikaan ei tapahdu ($(A \cup B)^C$). Mikä on todennäköisyys, että A tapahtuu, jos B tapahtuu eli ehdollinen todennäköisyys $\mathsf{P}(A \mid B)$?

Merkitään

$$\mathsf{P}(\text{kumpikaan ei tapahdu}) = \mathsf{P}((A \cup B)^C) = 0.6.$$

Toisaalta

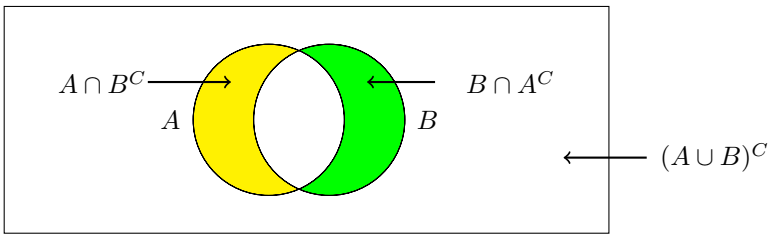
$$\mathsf{P}(A \cup B) = 1 - 0.6 = 0.4 = \mathsf{P}(A \cap B^C) + \mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C).$$

Kuva 9 havainnollistaa. Näin ollen

$$\mathsf{P}(A \cap B) = 0.4 - 0.2 - 0.1 = 0.1.$$

Kysytty ehdollinen todennäköisyys on

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C)} = \frac{0.1}{0.1 + 0.1} = 0.5. \ \square$$



Kuva 9: $\mathsf{P}(A \cup B) = \mathsf{P}(A \cap B^C) + \mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C)$.

Joskus sanotaan, että B myötävaikuttaa (*facilitate*) A:han, jos

$$\mathsf{P}(A \mid B) > \mathsf{P}(A).$$

Vastaavasti saatetaan sanoa, että B estää (*inhibit*) A:ta, jos

$$\mathsf{P}(A \mid B) < \mathsf{P}(A).$$

[10] Larsen ja Marx (2001, 53).

Verbejä ”myötävaikuttaa” ja ”estää” ei tule tulkita kirjaimellisesti. Tapahtumien välillä ei tarvitse olla syy-seuraus -suhdetta, vaikka tapahtuman A todennäköisyys muuttuu B:n tapahtumisen myötä. Ylipäänsäkään todennäköisyyksien laskuun ei tule liittää ajatusta kausaliteetista ellei ole muuta tietoa siitä.

Esimerkki. Harrastukset. Olkoot tapahtumat A=”harrastaa innokkaasti videopelejä” ja C=”kuntoilee aktiivisesti”.[11] Innokkaat videopelaajat ovat ehkä keskimääräistä harvemmin aktiivisia kuntoilijoita ja päinvastoin. Tällöin $\mathsf{P}(A \mid C) < \mathsf{P}(A)$ ja $\mathsf{P}(C \mid A) < \mathsf{P}(C)$. Kuntoilu voi vaikuttaa videopelaamisen todennäköisyyteen muttei välttämättä vaikuta esimerkiksi haluun videopelata saati estä sitä. Biologiset ominaisuudet tai vanhempien ohjaus jommankumman harrastuksen pariin voisivat olla syitä harrastukseen, eivätkä harrastukset sinällään vaikuta toisiinsa. □

Kaavasta (6) seuraa *tulosääntö*

$$\mathsf{P}(A \cap B) = \mathsf{P}(A \mid B) \times \mathsf{P}(B). \tag{7}$$

Sille on paljon käyttöä.

Esimerkki. Pallojen poimiminen (jatkoa). Pussissa on 10 vihreää, 20 oranssia ja 30 keltaista palloa. Poimitaan pussista pallo, jota ei palauteta pussiin, ja ongitaan sen jälkeen pussista toinen pallo. Mikä on todennäköisyys, että 1. pallo on vihreä ja 2. pallo keltainen? Entä todennäköisyys, että molemmat pallot ovat vihreitä?

Todennäköisyys saada ensin vihreä pallo on $\mathsf{P}(\text{1. pallo vihreä}) = 1/6$. Tämän tapahtuman jälkeen pussissa on 9 vihreää, 20 oranssia ja 30 keltaista palloa. Todennäköisyys saada seuraavaksi keltainen pallo on

$$\mathsf{P}(\text{2. pallo keltainen} \mid \text{1. pallo vihreä}) = \frac{30}{9+20+30} = \frac{30}{59}.$$

Todennäköisyys saada ensin vihreä ja sitten keltainen pallo on

$$\begin{aligned}&\mathsf{P}(\text{1. pallo vihreä ja 2. pallo keltainen})\\ &\quad = \mathsf{P}(\text{2. pallo keltainen} \mid \text{1. pallo vihreä}) \times \mathsf{P}(\text{1. pallo vihreä})\\ &\quad = \frac{30}{59} \times \frac{1}{6} = \frac{5}{59} \approx 0.085.\end{aligned}$$

Todennäköisyys saada kaksi vihreää palloa on

$$\begin{aligned}&\mathsf{P}(\text{1. ja 2. pallo vihreä})\\ &\quad = \mathsf{P}(\text{2. pallo vihreä} \mid \text{1. pallo vihreä}) \times \mathsf{P}(\text{1. pallo vihreä})\\ &\quad = \frac{9}{59} \times \frac{1}{6} = \frac{3}{118} \approx 0.025. \ \square\end{aligned}$$

Tapahtumat A ja B ovat *riippumattomia*, jos

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \times \mathsf{P}(B). \tag{8}$$

[11] C symbolisoi rivillä tapahtumaa ja yläindeksissä komplementtia. Tässä C on tapahtuma.

Kaava seuraa tulosäännöstä (7), jos $\mathsf{P}(B) > 0$ (ehdollista todennäköisyyttä $\mathsf{P}(A \mid B)$ ei ole muuten määritelty). Tällöin

$$\mathsf{P}(A \mid B) = \mathsf{P}(A). \tag{9}$$

Riippumattomien tapahtumien todennäköisyyteen ei vaikuta, onko toista tapahtunut vai ei. Jos A on mahdoton tapahtuma, riippumattomuusehto (8) toteutuu. Mahdoton tapahtuma on siten riippumaton mahdollisesta tapahtumasta.

Kaava (8) määrittelee riippumattomuuden, vaikka pätisi $\mathsf{P}(B) = 0$. Ehdon (8) mukaan riippumattomuus on symmetrinen ominaisuus: Jos A on riippumaton B:stä, myös B on riippumaton A:sta. Sama symmetrisyysominaisuus seuraa tulosäännöstä (7) ja ehdosta (9), jos sekä $\mathsf{P}(B) > 0$ että $\mathsf{P}(A) > 0$ (jolloin molemmille tapahtumille voidaan ehdollistaa):

$$\mathsf{P}(A \mid B) \times \mathsf{P}(B) = \mathsf{P}(A) \times \mathsf{P}(B) = \mathsf{P}(B) \times \mathsf{P}(A) = \mathsf{P}(B \mid A) \times \mathsf{P}(A).$$

Reunimmaiset yhtäsuuruudet seuraavat riippumattomuusoletuksesta. Kummassakaan tapahtumassa ei ole informaatiota toisesta.

Esimerkki. Kortin peluu. Korttipakka koostuu 52 kortista (pakassa ei ole jokereita). Kortit jakaantuvat neljään maahan: pataan, ristiin, herttaan ja ruutuun. Kunkin maan kortit on numeroitu 1–13. Numeroa 1 kutsutaan ässäksi ja numeroita 11–13 sotilaaksi, kuningattareksi ja kuninkaaksi. Vedetään korttipakasta sattumanvaraisesti kortti. Kunkin yksittäisen kortin todennäköisyys tulla valituksi on $1/52$. Ovatko tapahtumat ”hertta” ja ”kuningas” riippumattomia?

Tapahtumat ovat riippumattomia, mikäli

$$\mathsf{P}(\text{hertta}) \times \mathsf{P}(\text{kuningas}) = \frac{1}{52},$$

joka on todennäköisyys, että saadaan sekä ”hertta” että ”kuningas” (herttakuningas).

Todennäköisyys saada hertta tai kuningas ovat

$$\mathsf{P}(\text{hertta}) = \frac{13}{52} = \frac{1}{4}$$

ja

$$\mathsf{P}(\text{kuningas}) = \frac{4}{52} = \frac{1}{13}.$$

Niiden tulo on

$$\mathsf{P}(\text{hertta}) \times \mathsf{P}(\text{kuningas}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52}.$$

Saatiin herttakuninkaan todennäköisyys. Tapahtumat ”hertta” ja ”kuningas” ovat riippumattomia.

Todetaan lisäksi, että ehdolliset ja ehdollistamattomat todennäköisyydet ovat samat:

$$\mathsf{P}(\text{hertta} \mid \text{kuningas}) = \frac{\mathsf{P}(\text{hertta ja kuningas})}{\mathsf{P}(\text{kuningas})} = \frac{1/52}{4/52} = \frac{1}{4} = \mathsf{P}(\text{hertta}) \text{ ja}$$

$$\mathsf{P}(\text{kuningas} \mid \text{hertta}) = \frac{\mathsf{P}(\text{kuningas ja hertta})}{\mathsf{P}(\text{hertta})} = \frac{1/52}{13/52} = \frac{1}{13} = \mathsf{P}(\text{kuningas}). \ \square$$

Riippumattomuus on eri asia kuin erillisyys. Jos A ja B ovat erillisiä, ne eivät voi tapahtua yhtäaikaa (kuva 4). Tällöin toinen on ikään kuin este toisen tapahtumiselle, informaatiota on, eivätkä tapahtumat voi olla riippumattomia, jos niillä on positiivinen todennäköisyys. Tällöin riippumattomuusehto ei toteudu:

$$\mathsf{P}(A \cap B) = 0 \neq \mathsf{P}(A) \times \mathsf{P}(B).$$

Yhtäsuuruus seuraa erillisyysoletuksesta. Erilliset tapahtumat ovat riippumattomia vain, jos toisen tapahtuman (tai molempien) todennäköisyys on 0.

Esimerkki.[12] Jos tapahtumat A ja B ovat riippumattomia ja sekä $\mathsf{P}(A) > 0$ että $\mathsf{P}(B) > 0$, ovatko A ja B^C riippumattomia? Kyllä:

$$\mathsf{P}(B^C \mid A) = 1 - \mathsf{P}(B \mid A) = 1 - \mathsf{P}(B) = \mathsf{P}(B^C).$$

Toinen yhtäsuuruus seuraa A:n ja B:n riippumattomuudesta. Koska ehdollinen on ehdollistamaton todennäköisyys, A ja B^C ovat riippumattomia (riippumattomuusehto (9)). Lisähuomio: Vaihtamalla A:n ja B:n rooleja seuraa, että myös A^C ja B ovat riippumattomia, jos A ja B ovat. $\square$

Esimerkki.[13] Olkoot A ja B riippumattomia tapahtumia. Ovatko A^C ja B^C riippumattomia tapahtumia?

Yleisestä yhteenlaskusäännöstä (5) saadaan

$$\mathsf{P}(A^C \cup B^C) = \mathsf{P}(A^C) + \mathsf{P}(B^C) - \mathsf{P}(A^C \cap B^C).$$

Toisaalta

$$\mathsf{P}(A^C \cup B^C) = \mathsf{P}((A \cap B)^C) = 1 - \mathsf{P}(A \cap B)$$

(DeMorganin sääntö). Venn-diagrammit kuvassa 10 havainnollistavat.
Yhtälöistä seuraa, että

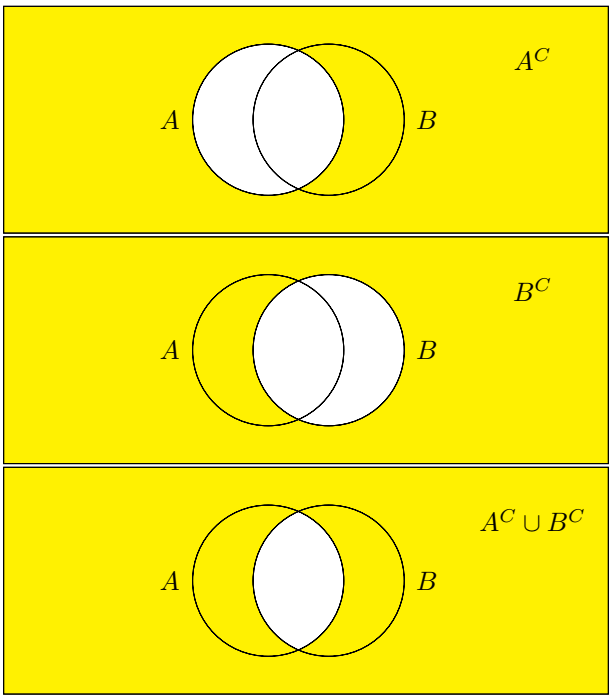
$$\begin{aligned} 1 - \mathsf{P}(A \cap B) &= \mathsf{P}(A^C) + \mathsf{P}(B^C) - \mathsf{P}(A^C \cap B^C) \\ &= 1 - \mathsf{P}(A) + 1 - \mathsf{P}(B) - \mathsf{P}(A^C \cap B^C). \end{aligned}$$

Riippumattomuusoletuksen mukaan $\mathsf{P}(A \cap B) = \mathsf{P}(A) \times \mathsf{P}(B)$. Näin ollen

$$\begin{aligned} \mathsf{P}(A^C \cap B^C) &= 1 - \mathsf{P}(A) + 1 - \mathsf{P}(B) - [1 - \mathsf{P}(A) \times \mathsf{P}(B)] \\ &= [1 - \mathsf{P}(A)] \times [1 - \mathsf{P}(B)] \\ &= \mathsf{P}(A^C) \times \mathsf{P}(B^C). \end{aligned}$$

Koska $\mathsf{P}(A^C \cap B^C) = \mathsf{P}(A^C) \times \mathsf{P}(B^C)$, niin A^C ja B^C ovat riippumattomia tapahtumia. $\square$

[12] Blitzstein ja Hwang (2015, 57).
[13] Larsen ja Marx (2001, 70).



Kuva 10: DeMorganin sääntö.

Esimerkki. Kahden nopan heitto (jatkoa). Aiemmin todettiin, että kaikki 36 silmälukuparia (1,1),(1,2),...,(6,5),(6,6) ovat yhtä todennäköisiä. Perustellaan se. Noppien heitot ovat riippumattomia. Kunkin silmäluvun todennäköisyys on $1/6$. Kunkin silmälukuparin (i,j) $(i,j = 1,\ldots,6)$ todennäköisyys on riippumattomuuden perusteella $1/36$:

$$\mathsf{P}(i,j) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

Vaikkapa silmälukuparien (1,2), (2,1) ja (2,2) todennäköisyys on sama $1/36$. □

Esimerkki. Sisarukset. Oletetaan, että tytöt (T) ja pojat (P) syntyvät toisistaan riippumatta samalla todennäköisyydellä $1/2$.[14] Perheessä on kaksi lasta, joista vanhempi on tyttö. Mikä on todennäköisyys, että nuorempi lapsista on tyttö?

Otosavaruus on

$$S = \{\text{TT,TP,PT,PP}\}.$$

Sisarukset ovat pareissa ikäjärjestyksessä. Kunkin parin todennäköisyys on $(1/2)\times(1/2) = 1/4$.

[14] Todellisuudessa poikavauvan todennäköisyys on noin 0.52 (esim. Pawitan 2013, 75).

Nuorempi sisarus on tyttö pareissa TT ja PT ja vanhempi pareissa TT ja TP. Sovelletaan ehdollisen todennäköisyyden kaavaa (6):

$$\begin{aligned}\mathsf{P}(\{\mathrm{TT}\cup\mathrm{PT}\}\mid\{\mathrm{TT}\cup\mathrm{TP}\}) &= \frac{\mathsf{P}(\{\mathrm{TT}\cup\mathrm{PT}\}\cap\{\mathrm{TT}\cup\mathrm{TP}\})}{\mathsf{P}(\mathrm{TT}\cup\mathrm{TP})}\\ &= \frac{\mathsf{P}(\mathrm{TT})}{\mathsf{P}(\mathrm{TT}\cup\mathrm{TP})}\\ &= \frac{1/4}{1/4+1/4} = \frac{1}{2}.\end{aligned}$$

Todennäköisyys, että nuorempi sisaruksista on tyttö, on $1/2$. □

Esimerkki. Sisarukset (jatkoa). Sisarusparadoksi I. Perheessä on kaksi lasta, joista ainakin toinen on tyttö. Mikä on todennäköisyys, että toinenkin on tyttö?

Ehdon ”ainakin toinen on tyttö” rajaama otosavaruus on {TT, TP, PT}. Ehdollinen todennäköisyys on

$$\begin{aligned}\mathsf{P}(\mathrm{TT}\mid\mathrm{TT},\mathrm{TP},\mathrm{PT}) &= \frac{\mathsf{P}(\mathrm{TT}\cap\{\mathrm{TT},\mathrm{TP},\mathrm{PT}\})}{\mathsf{P}(\mathrm{TT},\mathrm{TP},\mathrm{PT})} = \frac{\mathsf{P}(\mathrm{TT})}{\mathsf{P}(\mathrm{TT},\mathrm{TP},\mathrm{PT})}\\ &= \frac{1/4}{1/4+1/4+1/4} = \frac{1}{3}.\end{aligned}$$

Kun satunnaisesti poimitaan kaksilapsinen perhe, jossa toinen lapsista on tyttö, todennäköisyydellä $1/3$ molemmat lapset ovat tyttöjä. □

Riippumattomuus ei ole transitiivinen ominaisuus: Jos sekä A ja B että B ja C ovat riippumattomia tapahtumia, niin A ja C eivät välttämättä ole.

Esimerkki.[15] Otosavaruus koostuu kymmenestä yhtä todennäköisestä alkeistapahtumasta $\{1,\ldots,10\}$ ($\mathsf{P}(\{i\}) = 1/10, i = 1,\ldots,10$). Määritellään tapahtumat $A = \{2,3,4,5,6\}$, $B = \{4,5,6,7,8,9\}$ ja $C = \{2,4,6,8,10\}$. Niiden todennäköisyydet ovat

$$\mathsf{P}(A) = \frac{5}{10} = \frac{1}{2},\ \mathsf{P}(B) = \frac{6}{10} = \frac{3}{5} \quad \text{ja} \quad \mathsf{P}(C) = \frac{5}{10} = \frac{1}{2}.$$

Ehdollisten todennäköisyyksien laskua varten kirjataan $A\cap B = \{4,5,6\}$, $B\cap C = \{4,6,8\}$ ja $A\cap C = \{2,4,6\}$ sekä niiden todennäköisyydet:

$$\mathsf{P}(A\cap B) = \mathsf{P}(B\cap C) = \mathsf{P}(A\cap C) = \frac{3}{10}.$$

Huomataan, että tapahtumat A ja B sekä B ja C ovat riippumattomia, sillä niiden ehdollistetut ja ehdollistamattomat todennäköisyydet ovat yhtäsuuret:

$$\mathsf{P}(A\mid B) = \frac{3/10}{6/10} = \frac{1}{2} = \mathsf{P}(A),$$

$$\mathsf{P}(B\mid A) = \frac{3/10}{5/10} = \frac{3}{5} = \mathsf{P}(B),$$

[15] Venkatesh (2013, 37).

$$\mathsf{P}(B \mid C) = \frac{3/10}{5/10} = \frac{3}{5} = \mathsf{P}(B) \quad \text{ja}$$
$$\mathsf{P}(C \mid B) = \frac{3/10}{6/10} = \frac{1}{2} = \mathsf{P}(C).$$

Tapahtuman A todennäköisyys muuttuu, jos se ehdollistetaan tapahtumalle C tai toisinpäin:

$$\mathsf{P}(A \mid C) = \frac{3/10}{5/10} = \frac{3}{5} \neq \mathsf{P}(A) \quad \text{ja}$$
$$\mathsf{P}(C \mid A) = \frac{3/10}{5/10} = \frac{3}{5} \neq \mathsf{P}(C).$$

A ja B ovat riippumattomia, B ja C ovat riippumattomia, mutta A ja C eivät ole. □

Esimerkki. Harrastukset (jatkoa). Määritellään tapahtumiksi A = ”harrastaa innokkaasti videopelejä”, B = ”pelaa shakkia” ja C = ”kuntoilee aktiivisesti″. Voisi kuvitella, että harrastukset A ja B olisivat riippumattomia eli että videopelaamisen todennäköisyys ei riippuisi shakinpelaamisesta ja päinvastoin: $\mathsf{P}(A/B) = \mathsf{P}(A)$ ja $\mathsf{P}(B/A) = \mathsf{P}(B)$. Kuntourheilu ja shakki saattaisivat nekin olla riippumattomia harrastuksia: $\mathsf{P}(B/C) = \mathsf{P}(B)$ ja $\mathsf{P}(C/B) = \mathsf{P}(C)$. Silti saattaisi päteä aiemmin ounasteltu videopelaamisen ja kuntoilun käänteinen yhteys eli riippuvuus: $\mathsf{P}(A/C) < \mathsf{P}(A)$ ja $\mathsf{P}(C/A) < \mathsf{P}(C)$. □

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Monen tapahtuman riippumattomuus. Luennolla selitettiin, kuinka kaavasta $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$ seuraa erillisten tapahtumien yhteenlaskusääntö (4) tapahtumille $A_1, \ldots, A_n$. Niiden riippumattomuus ei yleisty aivan yhtä helpoksi kaavaksi. Ne ovat riippumattomia, jos kaikki yhtäsuuruudet alla pätevät:

$$\begin{aligned}
\mathsf{P}(A_i \cap A_j) &= \mathsf{P}(A_i) \times \mathsf{P}(A_j), \\
\mathsf{P}(A_i \cap A_j \cap A_k) &= \mathsf{P}(A_i) \times \mathsf{P}(A_j) \times \mathsf{P}(A_k), \\
&\vdots \\
\mathsf{P}(A_1 \cap \cdots \cap A_n) &= \mathsf{P}(A_1) \times \cdots \mathsf{P}(A_n).
\end{aligned}$$

Yllä alaindeksit i,j,k,… saavat kaikki mahdolliset arvot $1, \ldots n$ ja ovat kaikki eri suuria kullakin rivillä.[16] Pareittainen riippumattomuus ($\mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i) \times \mathsf{P}(A_j)$, $i \neq j$) ei takaa tapahtumien riippumattomuutta, jos tapahtumia on kolme tai enemmän.

Esimerkki. Tapahtumat A, B ja C ovat riippumattomia, jos kaikki yhtälöt alla pätevät:

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

[16] Asiaa pohditaan tarkemmin kirjoissa Blitzstein ja Hwang (2015, 57), Larsen ja Marx (2001, jakso 2.7) sekä Venkatesh (2013, luku 3).

Monesti oletetaan, että tapahtumat ovat riippumattomia. Oletuksen taustalla on enemmän ehtoja kuin asiaan perehtymättä ehkä arvaisi.

Riippumattomuusehto (8) yleistyy usean tapahtuman tilanteeseen: Tapahtumien $A_1, \ldots, A_n$, ollessa riippumattomia

$$\mathsf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathsf{P}(A_1) \times \mathsf{P}(A_2) \times \cdots \times \mathsf{P}(A_n). \tag{10}$$

Todennäköisyyslaskut ovat usein yksinkertaisia tai yksinkertaistuvat tavattomasti, jos voidaan olettaa riippumattomuus ja kaava on käytettävissä.

Esimerkki. Kolmen lantin heitto. Heitetään lanttia kolme kertaa peräjälkeen riippumattomasti. Merkitään kruuna = H (*heads*) ja klaava = T (*tails*). Heittosarjat ovat alkeistapahtumia ja muodostavat otosavaruuden $S = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HTT}, \mathrm{THH}, \mathrm{THT}, \mathrm{TTH}, \mathrm{TTT}\}$. Riippumattomuuden perusteella kunkin alkeistapahtuman todennäköisyys on

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}. \ \square$$

Esimerkki. Kolmen lantin heitto (jatkoa). Asiaan perehtymätön saattaisi ajatella, että heittosarjan HHH todennäköisyys olisi pienempi kuin vaihtelevammin kruunuja ja klaavoja sisältävän heittosarjan. Ajatukselle on nimikin: Uhkapelurin virhepäätelmä (*gambler's fallacy*). Siihen sortunee helpommin, jos on heitetty vieläkin useampia kertoja peräjälkeen kruuna. Maallikko voi ajatella, että seuraavaksi täytyy tulla klaava.

Lasketaan todennäköisyys kolmannelle kruunalle, kun kaksi aiempaa heittoa ovat tuottaneet kruunan. Kahden ensimmäisen heiton tuloksen rajaama otosavaruus on $B = \{\mathrm{HHH,HHT}\}$. Merkitään $A = \{\mathrm{HHH}\}$. $\mathsf{P}(B) = 1/8 + 1/8 = 1/4$ (riippumattomuuden ja erillisyyden perusteella), $\mathsf{P}(A) = 1/8$ ja

$$\begin{aligned}\mathsf{P}(A \mid B) &= \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(\mathrm{HHH} \cap \{\mathrm{HHH},\ \mathrm{HHT}\})}{\mathsf{P}(\mathrm{HHH},\ \mathrm{HHT})} = \frac{\mathsf{P}(\mathrm{HHH})}{\mathsf{P}(\mathrm{HHH},\ \mathrm{HHT})} \\ &= \frac{1/8}{1/4} = \frac{1}{2}.\end{aligned}$$

Kruunan todennäköisyys kahden kruunan heiton jälkeen on edelleen $1/2$. Vastaavasti voitaisiin osoittaa, että vaikkapa 11. kruunan todennäköisyys on $1/2$, vaikka sitä ennen olisi heitetty 10 kruunaa. $\square$

Esimerkki. Tentin läpäisy (jatkoa). Uhkapelurin virhepäätelmä ei rajoitu tapahtumasarjoihin, joissa yksittäisen tapahtuman todennäköisyys on $1/2$. Tekeekö opiskelija uhkapelurin virhepäätelmän, jos hän on reputtanut monta kertaa tentin ja ajattelee, että seuraavalla kerralla tentin täytyy mennä läpi? Ei välttämättä. Jos opiskelija on ahkeroinut tenttien välissä, on todennäköisyys läpäistä tentti kasvanut. Mikäli opiskelija yrittää kerta toisensa jälkeen tenttiä samoilla tiedoilla ja uskoo, että seuraavalla kerralla hänen täytyy jo läpäistä tentti, hän

tekee uhkapelurin virhepäätelmän — riippumatta siitä, mikä on tentin läpäisemisensä todennäköisyys. □

Esimerkki. Ylen uutinen 9.5.2011[17]:

Japanilainen ydinvoimayhtiö Chubu Electric Power Co. on päättänyt sulkea toistaiseksi Hamaokan ydinvoimalan. Japanin hallitus pyysi yhtiötä sulkemaan laitoksen, koska se on erityisen altis luonnononnettomuuksille. – – Hamaokan ydinvoimala sijaitsee 200 kilometriä Tokiosta länteen. Voimalaa on pidetty maailman vaarallisimpana, koska se sijaitsee mannerlaattojen saumakohdassa. Voimalassa on viisi 1970-luvulla rakennettua reaktoria, mutta kaksi niistä suljettiin jo vuonna 2009.

Japanin maanjäristysasiantuntijoiden mukaan on 87 prosentin todennäköisyys sille, että alueella tapahtuu seuraavien 30 vuoden aikana kahdeksan richterin suuruinen maanjäristys. Järistyksen odotetaan synnyttävän vastaavanlaisen tsunamin kuin Fukushiman ydinvoimalan hukuttanut jättiaalto.

Oletetaan, että maanjäristykset ovat toisistaan riippumattomia ja että todennäköisyys ainakin yhdelle maanjäristykselle 30 vuoden aikana on 0.87.[18] Mikä on todennäköisyys π maanjäristykselle vuoden aikana?

Ilmaistaan todennäköisyys maanjäristykselle vuoden aikana komplementtitapahtuman avulla:

$$\begin{aligned}\pi &= \mathsf{P}(\text{kahdeksan richterin suuruinen maanjäristys vuoden aikana})\\ &= 1 - \mathsf{P}(\text{ei kahdeksan richterin suuruista maanjäristystä vuoden aikana})\\ &= 1 - (1-\pi).\end{aligned}$$

Riippumattomuusoletuksesta ja oletuksesta 0.87:n todennäköisyydestä seuraa, että

$$\begin{aligned}&\mathsf{P}(\text{kahdeksan richterin suuruinen maanjäristys ainakin kerran 30 vuoden aikana})\\ &= 1 - \mathsf{P}(\text{ei kahdeksan richterin suuruista maanjäristystä 30 vuoden aikana})\\ &= 1 - (1-\pi)^{30}\\ &= 0.87.\end{aligned}$$

Ratkaistaan π:

$$\begin{aligned}(1-\pi)^{30} &= 1 - 0.87 &&= 0.13 &&\Leftrightarrow\\ 1-\pi &= (0.13)^{1/30} &&&&\Leftrightarrow\\ \pi &= 1 - (0.13)^{1/30} &&\approx 0.066.\end{aligned}$$

Todennäköisyys maanjäristykselle vuoden aikana on noin 0.066. □

Tulosääntö (7) yleistyy useamman tapahtumaan tilanteeseen. Tekniikka on ajatella useampaa ehdollistavaa tapahtumaa yhtenä. Olkoon tapahtumia kolme: A,

[17] `Http://yle.fi/uutiset/ulkomaat/2011/05/japanissa\_suljetaan\_quotmaailman\_vaarallisin\_ydinvoimalaquot\_2576095.html` (viitattu 9.5.2011).

[18] Riippumattomuusoletus kuvannee todellisuutta hyvin tässä yhteydessä. Maanjäristyksiä pidetään vaikeina ellei mahdottomina ennustaa. (Esim. `http://www.earthquakes.bgs.ac.uk/education/faqs/faq19.html` ja `https://en.wikipedia.org/wiki/Earthquake_prediction` (viitattu 4.2.2016).

B ja C. Merkitään $D = B \cap A$. Tulosäännön mukaan

$$\begin{aligned}\mathsf{P}(C \cap B \cap A) &= \mathsf{P}(C \cap D) \\ &= \mathsf{P}(C \mid D) \times \mathsf{P}(D) \\ &= \mathsf{P}(C \mid B \cap A) \times \mathsf{P}(B \cap A) \\ &= \mathsf{P}(C \mid B \cap A) \times \mathsf{P}(B \mid A) \times \mathsf{P}(A).\end{aligned}$$

Jos tapahtumia on n $(A_1, \ldots, A_n)$, tulosääntö yleistyy samalla tekniikalla näin:

$$\begin{aligned}\mathsf{P}(A_n \cap A_{n-1} \cap \cdots \cap A_1) \quad = \quad & \mathsf{P}(A_n \mid \{A_{n-1} \cap \cdots \cap A_1\}) \\ & \times \mathsf{P}(A_{n-1} \mid \{A_{n-2} \cap \cdots \cap A_1\}) \\ & \times \mathsf{P}(A_{n-2} \mid \{A_{n-3} \cap \cdots \cap A_1\}) \\ & \vdots \\ & \times \mathsf{P}(A_2 \mid A_1) \times \mathsf{P}(A_1).\end{aligned}$$

Esimerkki.[19] Sisarukset (jatkoa). Perheessä on kaksi lasta ($S = \{$TT, TP, PT, PP$\}$ ja kunkin parin todennäköisyys on $1/4$). Tapaat sattumalta toisen lapsista kutsuilla ja huomaat, että hän on tyttö. Mikä on todennäköisyys, että lapsista toinenkin on tyttö? Oletetaan, että tapaat lapsen samalla todennäköisyydellä $1/2$ riippumatta hänen sukupuolestaan.

Merkitään $A =$ "vanhempi lapsi on tyttö", $B =$ "nuorempi lapsi on tyttö" ja $C =$ "lapsi, jonka tapaat on tyttö". Tällöin $\mathsf{P}(A) = \mathsf{P}(B) = 1/2$, koska ehdot täyttäviä sisaruspareja on kaksi otosavaruudessa. Myös $\mathsf{P}(C) = 1/2$, sillä oletettiin, että sattumanvaraisesti tavattava lapsi on yhtätodennäköisesti tyttö kuin poika. Tapahtuman, että molemmat lapset ovat tyttöjä, todennäköisyys on $\mathsf{P}(A \cap B) = 1/4$, koska sellaisia sisaruspareja on yksi (TT) otosavaruudessa. Huomataan, että $A \cap B \cap C = A \cap B$, sillä jos molemmat lapset ovat tyttöjä, myös satunnaisesti tavattavan lapsen täytyy olla tyttö! Havaitaan, että perheen lapsista toinenkin on tyttö todennäköisyydellä $1/2$, jos sattumalta tavattu lapsi on tyttö:

$$\mathsf{P}(A \cap B \mid C) = \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(C)} = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(C)} = \frac{1/4}{1/2} = \frac{1}{2}. \ \square$$

Esimerkki. Sisarukset (jatkoa). Sisarusparadoksi II. Perheessä on kaksi lasta. Mikä on todennäköisyys, että lapsista toinenkin on tyttö, jos tiedetään, että ainakin toinen lapsista on talvella syntynyt tyttö? Oletetaan, että lapset syntyvät toisistaan riippumattomasti todennäköisyydellä $1/4$ kunakin vuodenaikana ja että sukupuoli ja vuodenaika ovat riippumattomia tapahtumia.

Laskettava todennäköisyys on

$$\begin{aligned}&\mathsf{P}(\text{molemmat tyttöjä} \mid \text{ainakin yksi talvityttö}) \\ &= \frac{\mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvityttö})}{\mathsf{P}(\text{ainakin yksi talvityttö})}\end{aligned}$$

[19] Tämä ja seuraavaa esimerkki ovat Blitzsteinin ja Hwangin (2015, 46–47) kirjasta.

Riippumattomuuden perusteella todennäköisyys syntyä tyttönä talvella on $(1/2) \times (1/4) = 1/8$. Todennäköisyys nimittäjässä saadaan näin:

$$\begin{aligned}\mathsf{P}(\text{ainakin yksi talvityttö}) &= 1 - \mathsf{P}(\text{ei yhtään talvityttöä}) \\ &= 1 - \mathsf{P}(\text{1. lapsi ei ole talvityttö ja 2. lapsi ei ole talvityttö}) \\ &= 1 - \left(1 - \frac{1}{8}\right)^2 = 1 - \left(\frac{7}{8}\right)^2 = \frac{15}{64}.\end{aligned}$$

Todennäköisyyden osoittajassa päättely:

$$\begin{aligned}&\mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvityttö}) \\ &= \mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvilapsi}) \\ &= \frac{1}{4} \times \mathsf{P}(\text{ainakin yksi talvilapsi}) \\ &= \frac{1}{4} \times [1 - \mathsf{P}(\text{kumpikaan ei talvilapsi})] \\ &= \frac{1}{4} \times [1 - \mathsf{P}(\text{1. ei talvilapsi ja 2. ei talvilapsi})] \\ &= \frac{1}{4} \times \left[1 - \left(\frac{3}{4}\right)^2\right] \\ &= \frac{7}{64}.\end{aligned}$$

Toinen yhtäsuuruus pätee, koska viitatut osajoukot (joukkojen leikkaukset) ovat samat! Seuraava yhtäsuuruus tulee sukupuoli- ja vuodenaikatapahtumien riippumattomuudesta, minkä johdosta kahden tytön todennäköisyys $(1/4)$ voidaan kertoa talvilapsen syntymiseen liittyvällä todennäköisyydellä.

Todennäköisyys, että perheessä on kaksi tyttöä, kun perheessä on ainakin yksi talvella syntynyt tyttö, on

$$\mathsf{P}(\text{molemmat tyttöjä} \mid \text{ainakin yksi talvityttö}) = \frac{15/64}{7/64} = \frac{7}{15}\ !$$

Todennäköisyys $7/15$ sijoittuu aiemmissa sisarusesimerkeissä laskettujen todennäköisyyksien $1/3$ ja $1/2$ väliin. □

Tapahtumat A ja B ovat *ehdollisesti riippumattomia*, jos

$$\mathsf{P}(A \cap B \mid C) = \mathsf{P}(A \mid C) \times \mathsf{P}(B \mid C).$$

Yllä C on tapahtuma, jolle ehdollistetaan. Tapahtumat voivat olla ehdollisesti riippumattomia, vaikka ne eivät olisi riippumattomia, tai ne voivat olla riippumattomia, vaikka ne eivät olisi ehdollisesti riippumattomia. Jos tapahtumat ovat riippumattomia ehdolla C, ne eivät välttämättä ole riippumattomia ehdolla C^C.

Esimerkki.[20] Kaksi lanttia. Ehdollisesta riippumattomuudesta ei seuraa riippumattomuutta. Lanteista toinen (”kultainen lantti” tai ”lantti1”) on harhaton: ”kruuna”:n todennäköisyys on $1/2$. Toinen lantti (”hopeinen lantti” tai ”lantti2”) on harhainen: ”kruuna”:n todennäköisyys on $3/4$. Arvotaan jommankumman värinen lantti ja heitellään sitä. Heittojen tulokset (”kruuna” vai ”klaava”) ovat riippumattomia ehdolla arvotunvärinen lantti. ”Kruunu”:jen todennäköisyys kylläkin riippuu arvotusta lantista.

Jos lantteja ei voi erottaa väristä, heittojen tulokset eivät ole riippumattomia. Arvottua lanttia heitetään ja saadaan ”kruuna”. Se viittaa siihen, että on valittu ”lantti2”. Seuraavankin heiton tulos olisi tällöin todennäköisemmin ”kruuna”. Heittojen tulokset eivät ole riippumattomia.

Olkoon C lantin arvonnan tulos (”lantti1” tai ”lantti2”) ja A ja B ensimmäisen ja toisen lantin heiton tulos (”kruuna” tai ”klaava”). A ja B ovat nyt riippumattomia ehdolla C: Ensimmäinen heitto ei anna informaatiota toisen heiton tuloksesta. A ja B eivät ole riippumattomia, sillä A:ssa on informaatiota B:stä: Jos A oli ”kruuna”, se viittaa siihen, että on valittu ”lantti1”, jolloin on todennäköisempää, että B:kin on ”kruuna”. □

Esimerkki. Kahdenlaiset kurssit. Riippumattomuudesta ehdolla C^C ei seuraa riippumattomuus ehdolla C. Opiskelija on erikoisessa yliopistossa. Toisilla kursseilla ahertaminen palkitaan hyvällä arvosanalla; toisilla arvosana on sama opiskelun määrästä riippumatta. Merkitään ensimmäisenlaista kurssia C:llä (jälkimmäisenlaista C^C:lla), hyvää arvosanaa A:lla (huonoa A^C:lla) ja ahertamista B:llä (ahertamattomuutta B^C:lla). Tällöin A ja B ovat riippumattomia ehdolla C^C mutteivät ole riippumattomia ehdolla C. □

Esimerkki. Puhelimen soiminen. Riippumattomuudesta ei seuraa ehdollista riippumattomuutta. Vanhuksella on kaksi ystävää A ja B. Kunakin päivänä A voi soittaa riippumatta siitä, soittaako B ja päinvastoin. Kun puhelin soi (ehdollistava tapahtuma), A ja B eivät ole riippumattomia: Jos soittaja on A, soittaja ei voi olla B ja päinvastoin. □

Vanhuksella on kaksi ystävää A ja B. Kunakin päivänä A voi soittaa riippumatta siitä, soittaako B ja päinvastoin. Kun puhelin soi (ehdollistava tapahtuma), A ja B eivät ole riippumattomia: Jos soittaja on A, soittaja ei voi olla B ja päinvastoin. □

2.6 Kokonaistodennäköisyys ja Bayesin kaava

Olkoon otosavaruus ositettu erillisiin tapahtumiin A_i, joilla on kaikilla positiivinen todennäköisyys ($S = \cup_{i=1}^{n} A_i, A_i \cap A_j = \emptyset$ ja $\mathsf{P}(A_i) > 0, i = 1, \ldots, n$). Tällöin tapahtuman B todennäköisyys on

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i). \tag{11}$$

[20] Tämä ja kaksi seuraavaa esimerkkiä ovat kirjasta Blitzstein ja Hwang (2015, 58).

Sitä kutsutaan *kokonaistodennäköisyyden laiksi*. Se perustellaan näin: $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$. Leikkaukset ovat erillisiä. Erillisten tapahtumien yhteenlaskusäännön (4) ja tulosäännön (7) mukaan $\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \cap A_i) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i)$.

Esimerkki.[21] Helsingin Sanomat 5.2.2016:

> Aarnio-tutkinnan puolueettomuus nousi keskeiseksi teemaksi torstaina Helsingin käräjäoikeudessa jatkuneessa istunnossa. Jari Aarnion asianajaja Riitta Leppiniemi ja jutun tutkinnanjohtaja, kihlakunnansyyttäjä Jukka Haavisto ottivat tiukasti yhteen, kun Haavistoa kuultiin todistajana. Istunnossa kiisteltiin pitkään lähes huutamalla muun muassa siitä, mitä he olivat jutelleet keskenään vangitsemisoikeudenkäyntien käytäväkeskusteluissa. Leppiniemi kyseenalaisti koko Aarnio-tutkinnan ja Haaviston puolueettomuuden. Hän huomautti, että Haavisto oli huumepoliisien ensimmäisessä virkarikosjutussa toisena syyttäjänä. Siinä syytteet Aarniota vastaan hylättiin. – – Leppiniemi kummeksui sitä, miksi jutun toiseksi tutkinnanjohtajaksi tuli nimenomaan Haavisto, vaikka vaihtoehtoja olisi ollut tarjolla. – – Aarnio on esittänyt myös toiveita tutkintaryhmän kokoonpanosta. Hän on esimerkiksi pyytänyt, ettei keskusrikospoliisin rikosylikomisario Rabbe von Hertzen osallistuisi jutun tutkintaan. Aarnio on nimennyt von Hertzenin julkiseksi vihamiehekseen. – – Von Hertzeniltä kysyttiin myös salanauhoituksesta, jonka kuopiolainen naispoliisi oli tehnyt huumerikostutkijoiden risteilyllä huhtikuussa 2015. Keskustelun lomassa von Hertzen luonnehtii Aarniota rosvoksi. – – Von Hertzeniltä kysyttiin, onko hän sanonut Aarniota täydeksi narsistiksi. ”En muista sanoneeni, mutta pidän häntä jossain määrin narsistina.” Istunnossa myös selviteltiin useiden todistajien avulla, miten Aarnion tontilta Porvoosta löydettiin 65 000 euron rahakätkö toukokuussa 2014. – – Aarnion mukaan kätkö on lavastus.

Oikeudessa puolustus väittää poliisin väärentäneen todistusaineistoa tai muuten toimineen väärin. Todennäköisyys, että puolustus pystyy vakuuttamaan oikeuden, että niin on tapahtunut, on 0.70. Tällöin oikeus tuomitsee syytetyn todennäköisyydellä 0.15. Muussa tilanteessa tuomion todennäköisyys on 0.80. Mikä on todennäköisyys, että oikeus tuomitsee syytetyn?

Merkitään $B =$ ”oikeus tuomitsee syytetyn” ja $A_1 =$ ”oikeus katsoo poliisin toimineen väärin” ja $A_2 =$ ”oikeus ei katso poliisin toimineen väärin”. Annettujen tietojen mukaan $\mathsf{P}(B \mid A_1) = 0.15$, $\mathsf{P}(B \mid A_2) = 0.80$, $\mathsf{P}(A_1) = 0.70$ ja $\mathsf{P}(A_2) = 1 - 0.70 = 0.30$. Tuomion todennäköisyys on 0.345:

$$\begin{aligned}\mathsf{P}(B) &= \mathsf{P}(B \mid A_1)\mathsf{P}(A_1) + \mathsf{P}(B \mid A_2)\mathsf{P}(A_2) \\ &= 0.15 \times 0.70 + 0.80 \times 0.30 \\ &= 0.345. \ \square\end{aligned}$$

Bayesin kaava on

$$\mathsf{P}(A_j \mid B) = \frac{\mathsf{P}(B \mid A_j)\mathsf{P}(A_j)}{\sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i)}, \tag{12}$$

[21] Larsen ja Marx (2001, 62–63) ja `http://www.hs.fi/kotimaa/a1454569374503`) (viitattu 5.2.2016).

jossa $1 \leq j \leq n$. Kaavaa kutsutaan myös käänteistodennäköisyyden kaavaksi, koska tapahtuman (B) ja sen ehdon (A_j) roolit on käännetty kaavan vasemmalla puolella. Kaavan perustelu on lyhyt: $\mathsf{P}(A_j \mid B) = \mathsf{P}(A_j \cap B)/\mathsf{P}(B) = \mathsf{P}(B \mid A_j)\mathsf{P}(A_j)/\mathsf{P}(B)$. Sijoittamalla nimittäjään kokonaistodennäköisyyden kaava (11) saadaan Bayesin kaava.

Esimerkki. HIV. HIV-infektoituneita oli 14.2.2010 mennessä Suomessa tilastoitu 2618, joista 497 oli kuollut. Infektiot esiintyvät valtaosin ikäryhmässä 15–64-vuotta. Tämänikäisiä suomalaisia oli 3 543 084 vuoden 2008 lopussa. HIV-infektion todennäköisyys tässä ikäryhmässä on siten noin $(2618 - 497)/3543084$ eli 0.0006. Tietty lääketieteellinen testi indikoi infektiota todennäköisyyksillä 0.997 ja 0.015, kun henkilö on infektoitunut tai ei ole (testi ei anna aina oikeata vastausta kummassakaan tilanteessa).[22]

Mikä oli vuonna 2010 todennäköisyys, että satunnaisesti valittu ja testattu 15–64-vuotias suomalainen on HIV-infektoitunut, kun hänelle tehty testi indikoi infektiota?

Merkitään E:llä ei-infektoitunutta, I:llä infektoitunutta ja $+$:lla positiivista testitulosta (testi indikoi infektiota). Annetuista tiedoista saadaan todennäköisyydet $\mathsf{P}\ (I) = 0.0006$, $\mathsf{P}(E) = 0.9994$, $\mathsf{P}(+ \mid I) = 0.997$ ja $\mathsf{P}(+ \mid E) = 0.015$. Sijoitetaan ne Bayesin kaavaan:

$$\begin{aligned} \mathsf{P}(I \mid +) &= \frac{\mathsf{P}(+ \mid I)\mathsf{P}(I)}{\mathsf{P}(+ \mid I)\mathsf{P}(I) + \mathsf{P}(+ \mid E)\mathsf{P}(E)} \\ &= \frac{0.997 \times 0.0006}{0.997 \times 0.0006 + 0.015 \times 0.9994} \\ &\approx 0.038. \end{aligned}$$

Todennäköisyys, että satunnaisesti valittu 15–64-vuotias suomalainen on HIV-infektoitunut, kun HIV-testitulos on positiivinen, on noin 0.04. Tauti on niin harvinainen, että positiivinen testitulos johtuu tavattomasti useammin virheellisestä testituloksesta kuin infektiosta. □

Esimerkki. HIV (jatkoa). Kansanterveyslaitos (nykyinen Terveyden ja hyvinvoinnin laitos) ja Aids-tukikeskus lähettivät kyselylomakkeen sekä HIV:tä testaavan sylkitestin seksuaali- ja sukupuolivähemmistöjen lehden tilaajarekisterin miesten osoitteisiin 2006. Homo- tai biseksuaalisista miehistä HIV-infektoituneeksi ilmoittautui tai tutkimuksessa havaittiin infektoituneiksi 4.6 %.[23] Mikä on to-

[22] Lähteet: `Http://www.ktl.fi/ttr/gen/rpt/hivsuo.html`, `http://www.ktl.fi/ttr/gen/rpt/hivaidskuo.html`, `http://www.tilastokeskus.fi/til/vaerak/2008/vaerak_2008_2009-03-27_tie_001_fi.html` ja `http://en.wikipedia.org/wiki/HIV_test#Accuracy_of_HIV_testing` (viittaukset 2010). Luvut ovat suuntaa antavia, muun muassa koska kaikki HIV-infektiot eivät ole tiedossa. Arviolta joka kolmas tartunta on diagnosoimaton (Suomen hiv-strategia 2013–2016. Terveyden ja hyvinvoinnin laitos 7/2012. S. 7).

[23] Kansanterveyslaitoksen HIV-yksikön johtajan Mika Salmisen artikkeli (2006) HIV-riskit kasvaneet (urlhttp://demo.seco.tkk.fi/tervesuomi/item/ktl:11672 (viitattu 14.7.2014)) sekä Terveyden ja hyvinvoinnin laitoksen erikoistutkija Henrikki Brummer-Korvenkontio (henkilökohtainen tiedonanto 16.2.2010). Tutkimuksen tuloksia ei voida yleistää koskemaan suomalaisia homo- tai biseksuaalisia miehiä ylipäänsä mm., koska lehden lukijakunta voi olla valikoitunutta. Esimerkiksi lehden lukijakunta painottuu yli 30-vuotiaisiin miehiin. Myös voi olla,

dennäköisyys, että tilaajarekisteristä satunnaisesti valittu homo- tai biseksuaalinen mies on HIV-infektoitunut, kun hänelle tehty testi indikoi infektiota?

Infektion yleisyyttä kuvaavat todennäköisyydet ovat nyt $\mathsf{P}(I) = 0.046$ ja $\mathsf{P}(E) = 0.954$. Sijoitetaan ne yhdessä testin ominaisuuksia kuvaavien todennäköisyyksien (samat kuin edellä) kanssa Bayesin kaavaan:

$$\begin{aligned} \mathsf{P}(I \mid +) &= \frac{\mathsf{P}(+ \mid I)\mathsf{P}(I)}{\mathsf{P}(+ \mid I)\mathsf{P}(I) + \mathsf{P}(+ \mid E)\mathsf{P}(E)} \\ &= \frac{0.997 \times 0.046}{0.997 \times 0.046 + 0.015 \times 0.954} \\ &\approx 0.762. \end{aligned}$$

Todennäköisyys, että satunnaisesti valittu lehden tilaajarekisterin homo- tai biseksuaalinen mies on HIV-infektoitunut, kun HIV-testitulos on positiivinen, on noin 0.76. Ero todennäköisyyteen edellisessä esimerkissä on suuri ja johtuu HIV-infektion huomattavasti suuremmasta yleisyydestä tässä ryhmässä verrattuna 15–64-vuotiaisiin suomalaisiin. Edelleenkin kuitenkin testin virhemahdollisuudesta johtuen ($\mathsf{P}(+ \mid E) = 0.015$) HIV-infektoituneeksi testattu homo- tai biseksuaalinen tilaajamies on todennäköisyydellä noin 0.24 infektoitumaton. □

Diagnostisen tai muun luokittelevan testin todennäköisyyttä tunnistaa sairaus tai muu ominaisuus kutsutaan testin *herkkyydeksi* (*sensitivity*). Todennäköisyys, jolla testi tunnistaa sairauden tai ominaisuuden puuttumisen, on testin *tarkkuus* (*specifity*).

Esimerkki. HIV (jatkoa). HIV-testin herkkyys ja tarkkuus ovat 0.997 ja $1 - 0.015 = 0.985$. □

Olkoon tapahtumat A ja B. Lasketaan Bayesin kaavalla (12) todennäköisyydet $\mathsf{P}(A \mid B)$ ja $\mathsf{P}(A^C \mid B)$, ja jaetaan saadut yhtälöt puolittain. $\mathsf{P}(B)$:t supistuvat pois. Tulos on

$$\frac{\mathsf{P}(A \mid B)}{\mathsf{P}(A^C \mid B)} = \frac{\mathsf{P}(B \mid A)}{\mathsf{P}(B \mid A^C)} \times \frac{\mathsf{P}(A)}{\mathsf{P}(A^C)}. \tag{13}$$

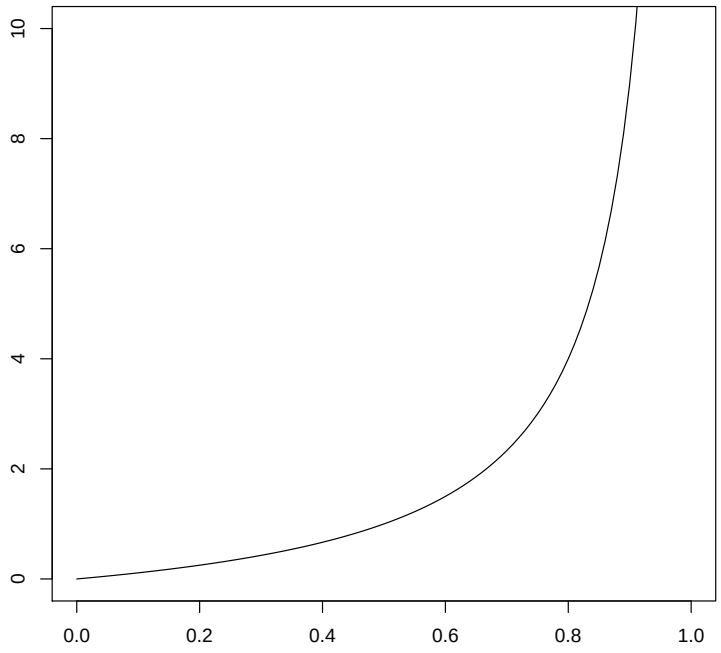
Oikeanpuoleisin osamäärä on A:n ja sen komplementin todennäköisyyksien suhde *prioritapahtumakerroin* (*prior odds*). Yhtälön vasemmalla puolella on A:n ja sen komplementin ehdollisten todennäköisyyksien suhde *posterioritapahtumakerroin* (*posterior odds*).

Tapahtumakerroin (*odds*) kertoo, kuinka monta kertaa todennäköisempi on tapahtuman todennäköisyys verrattuna sen komplementin todennäköisyyteen.[24] Prioritapahtumakerroin on tämä suhde ennen ehdollistamista; posterioritapahtumakerroin sen jälkeen. Mitä enemmän priori- ja posterioritapahtumakertoimet eroavat, sitä enemmän informaatiota on ehdollistaminen B:lle tuottanut.

että vastaajiksi valikoitui erityisen riskialttiisti käyttäytyviä homoseksuaaleja, joille tarjoutui kyselyn kautta tilaisuus teettää HIV-testi kätevästi. Kyselyyn vastasi 410 (vastausprosentti 29) ja sylkinäytteen antoi 368 miestä. (Em. artikkeli.)

[24] Veto (esim. Rita 2004) ja vedonlyöntisuhde ovat paljon käytetympiä suomennoksia *odds*ille. Jälkimmäinen voi sekoittua kahden tapahtumakertoimen suhteeseen *odds ratio*'hun (Rita ym. 2008). Vedonlyöntisuhde-suomennosta käytettiin jaksossa 2.2.3, koska siellä pohdittiin vedonlyöntitilannetta.

Tapahtumakerroin saa arvoja nollasta ylöspäin (kuva 11, jossa todennäköisyyttä merkitään π:llä). Mahdottomalle tapahtumalle kerroin on 0. Jos todennäköisyys on pieni, se ja kerroin ovat suurinpiirtein yhtäsuuria. Jos tapahtuman todennäköisyys on sama kuin sen komplementin (0.5), kerroin on 1. Jos kerroin on yhtä suurempi, todennäköisyys tapahtumalle on suurempi kuin sen komplementille. Kerroin suurenee nopeasti todennäköisyyden lähestyessä yhtä.



Kuva 11: Tapahtumakerroin $\pi/(1-\pi)$.

Esimerkki. Tentin läpäisy (jatkoa). Opiskelija arvioi, että hän läpäisee tentin todennäköisyydellä 0.8. Tapahtumakerroin — tässä yhteydessä onnistumiskerroin — on varsin suuri:

$$\frac{0.8}{0.2} = 4.$$

Opiskelijasta on 4 kertaa todennäköisempää, että hänen suorituksensa hyväksytään kuin ettei hyväksytä. □

Esimerkki. Lastensuojelun asiakkuus. 16–17-vuotiaista 9.9 % oli lastensuojelun

avohuollon asiakkaita vuonna 2014.[25] Poimitaan satunnaisesti 16–17-vuotias. Asiakkuuden todennäköisyys ei ole suuri, joten se ja tapahtumakerroin poikkevat vähän:

$$\frac{0.099}{1-0.099} = \frac{0.099}{0.901} \approx 0.110.$$

Jokaista lastensuojelun nuorta asiakasta kohden on keskimäärin 9 nuorta, jotka eivät ole asiakkaita. On noin 9 kertaa todennäköisempää, ettei nuori ole lastensuojelun asiakas kuin että on. □

Esimerkki. HIV (jatkoa).[26] HIV-infektion ($I; E = I^C$) prioritapahtumakerroin suomalaisten ikäryhmässä 15–64-vuotta on lähes sama kuin infektion todennäköisyys 0.0006:

$$\frac{\mathsf{P}(I)}{\mathsf{P}(E)} = \frac{0.0006}{0.9994} \approx 0.0006.$$

Selitys on, että tapahtumakertoimen jakaja on lähes 1. Positiivisen HIV-testituloksen jälkeinen todennäköisyys infektiolle on noin 0.038. Sekin on niin pieni, että posterioritapahtumakerroin on lähes sama:

$$\frac{\mathsf{P}(I \mid +)}{\mathsf{P}(E \mid +)} \approx \frac{0.038}{1-0.038} \approx 0.040.$$

Seksuaali- ja sukupuolivähemmistöjen lehden tilaajarekisterin miesten keskuudessa priori- ja posterioritapahtumakertoimet eroavat selvästi:

$$\frac{\mathsf{P}(I)}{\mathsf{P}(E)} = \frac{0.046}{0.954} \approx 0.048$$

ja

$$\frac{\mathsf{P}(I \mid +)}{\mathsf{P}(E \mid +)} \approx \frac{0.762}{0.238} \approx 3.205.$$

On noin kolme kertaa todennäköisempää, että tilaajarekisterin mies on HIV-infektoitunut kuin että ei ole, kun testin tulos on positiivinen. □

Olkoon tapahtumakerroin k. Sen määrittävästä yhtälöstä voidaan ratkaista todennäköisyys:

$$\begin{aligned} \frac{\mathsf{P}(A \cdot)}{\mathsf{P}(A^C \cdot)} &= \frac{\mathsf{P}(A \cdot)}{1-\mathsf{P}(A \cdot)} = k \quad \Leftrightarrow \\ \mathsf{P}(A \cdot) &= \frac{k}{1+k}. \end{aligned} \tag{14}$$

Yllä $\mathsf{P}(A \cdot)$ on ehdollistamaton tai ehdollinen todennäköisyys riippuen siitä, onko k priori- vai posterioritapahtumakerroin. Jos tapahtumakerroin on muotoa

[25] Lastensuojelu 2014. Terveyden ja hyvinvoinnin laitos. `Http://urn.fi/URN:NBN:fi-fe2015120422151` (viitattu 10.2.2016). Kuvio 11 ja liitetaulukko 3.

[26] Tässä ja seuraavissa esimerkeissä laskut on tehty suuremmalla tarkkuudella kuin tekstistä ilmenee. Raportointitarkkuudella tehtyjen laskujen lopputulos poikkeaisi paikoin esitetystä.

$k = a/b$, jossa a ja b ovat kokonaislukuja, niin todennäköisyyden voi päätellä erityisen kätevästi näin:

$$\begin{aligned} \frac{\mathsf{P}(A \cdot)}{\mathsf{P}(A^C \cdot)} &= \frac{a}{b} \qquad \Leftrightarrow \\ \mathsf{P}(A \cdot) &= \frac{a}{a+b}. \end{aligned} \tag{15}$$

Esimerkki. HIV (jatkoa). Seksuaali- ja sukupuolivähemmistöjen lehden tilaajamiehille $k = 3.205$. Kaavasta (14) voidaan laskea ehdollinen todennäköisyys HIV-kantajuudelle:

$$\mathsf{P}(I \mid +) \approx \frac{3.205}{1 + 3.205} \approx 0.762.$$

Saatiin aiemmin laskettu todennäköisyys. □

Esimerkki. Tentin läpäisy (jatkoa). Onnistumiskerroin tentin läpimenolle on opiskelijan mielestä $4 = 4/1$. Läpimenon subjektiivinen todennäköisyys kaavan (15) mukaisesti on tällöin $4/(1 + 4) = 4/5 = 0.8$. □

Bayesin kaava on *johdonmukainen* (*coherent*): On samantekevää, tuleeko uusi ehdollistava tieto kerralla vai ehdollisesti riippumattomissa erissä. Saman tiedon huomioiva ehdollinen todennäköisyys on molemmilla tavoilla laskettuna sama. Kaava (13) päivittää kätevästi ehdollisen todennäköisyyden, kun saadaan uutta informaatiota.

Esimerkki. HIV (jatkoa). Tehdään ikäryhmään 15–64-vuotta kuuluvalle suomalaiselle kaksi HIV-testiä. Testit ovat riippumattomia ehdolla tutkittavan infektoituneisuusstatus (on tai ei ole). Testien herkkyys ja tarkkuus ovat 0.997 ja 0.985. Molempien testien mukaan tutkittu kantaa HIV:tä. Mikä on todennäköisyys, että hänessä on HIV?

Merkitään virusta indikoivaa 1. ja 2. testitulosta $+_1$:llä ja $+_2$:lla ja infektoituneisuutta I:llä ($E = I^C$). Sovelletaan kaavaa (13):

$$\begin{aligned} \frac{\mathsf{P}(I \mid +_1 \cap +_2)}{\mathsf{P}(E \mid +_1 \cap +_2)} &= \frac{\mathsf{P}(+_1 \cap +_2 \mid I)}{\mathsf{P}(+_1 \cap +_2 \mid E)} \times \frac{\mathsf{P}(I)}{\mathsf{P}(E)} \\ &\approx \frac{0.997^2}{0.015^2} \times 0.0006 \\ &\approx 2.646. \end{aligned}$$

Prioritapahtumakerroin $\mathsf{P}(I)/\mathsf{P}(E) \approx 0.0006$ laskettiin edellä. Toinen (approksimatiivinen) yhtäsuuruus seuraa testien riippumattomuudesta ehdolla, että tutkittava on infektoitunut. Kaavasta (14) saadaan kysytyksi todennäköisyydeksi 0.726:

$$\mathsf{P}(I \mid +_1 \cap +_2) \approx \frac{2.646}{1 + 2.646} \approx 0.726.$$

Jos kaksi testiä indikoi HIV-infektiota, infektoituneisuus on melko todennäköistä.

Sama tulos voidaan laskea kahdessa vaiheessa. Testin 1 jälkeen posterioritapahtumakertoimeksi laskettiin edellä 0.040. Käytetään sitä uutena päivitettynä prioritapahtumakertoimena:

$$\begin{aligned}\frac{\mathsf{P}(I \mid +_1 \cap +_2)}{\mathsf{P}(E \mid +_1 \cap +_2)} &= \frac{\mathsf{P}(+_2 \mid I \cap +_1)}{\mathsf{P}(+_2 \mid E \cap +_1)} \times \frac{\mathsf{P}(I \mid +_1)}{\mathsf{P}(E \mid +_1)} \\ &\approx \frac{0.997}{0.015} \times 0.040 \\ &\approx 2.646.\end{aligned}$$

Posterioritapahtumakerroin on sama kuin edellä laskettu. Sijoittamalla $k = 2.646$ kaavaan (14) kysytyksi todennäköisyydeksi saadaan taas 0.726. □

2.7 Puudiagrammi

Puudiagrammi visualisoi satunnaisilmiöitä ja auttaa hahmottamaan, kuinka yhdistetty tapahtuma koostuu useammasta yksinkertaisemmasta tapahtumasta. Se helpottaa yhdistettyjen tapahtumien todennäköisyyksien ymmärtämistä ja laskemista.

Puudiagrammeilla voidaan laskea todennäköisyyksiä, jos tutkittavalla prosessilla on yksi alkutila, useita vaihtoehtoisia lopputiloja, joista yksi toteutuu ja välissä on tapahtumia, jotka johtavat toisensa poissulkeviin tapahtumaketjuihin. Puudiagrammin avulla voidaan tapahtuman todennäköisyys laskea kahden säännön avulla:

- Reitin todennäköisyys on siihen johtavien tapahtumien todennäköisyyksien tulo (tulosääntö).
- Tapahtuman todennäköisyys on siihen johtavien reittien todennäköisyyksien summa (yhteenlaskusääntö).

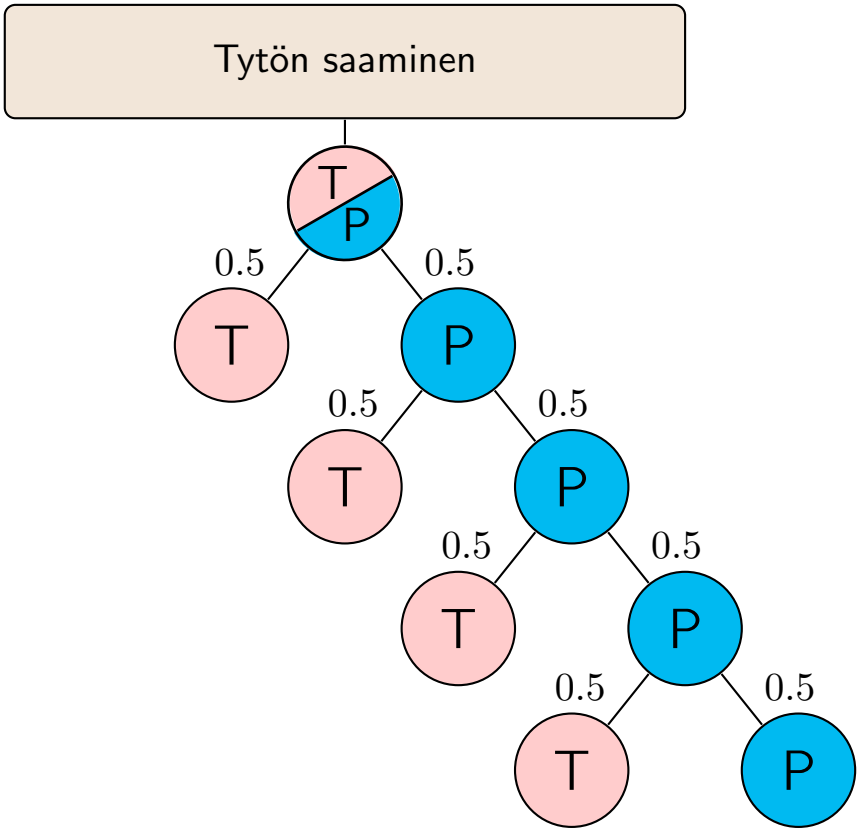
Esimerkki. Tytön saaminen.[27] Pariskunta päättää tehdä lapsia, kunnes on saatu tyttö. Neljää lasta enempää ei kuitenkaan ryhdytä tekemään.

Oletetaan, että tyttöjä ja poikia syntyy todennäköisyydellä 0.5 riippumatta aiemmin syntyneiden lasten sukupuolesta ja että kerrallaan syntyy yksi lapsi. Mikä on todennäköisyys, että pariskunta saa tytön?

Puudiagrammissa kuvassa 12 punertaviin tyttö-tapahtumiin (oksien kärkiin) päästään lähtöpisteestä (T/P; puun juuri) neljää eri reittiä (haarautunutta oksaa) pitkin. Reitin todennäköisyys saadaan lasten sukupuolien riippumattomuuden perusteella riippumattomien tapahtumien tulosäännöstä (10). Kukin tyttötapahtuma on erillinen, ja kuhunkin vie vain yksi reitti. Todennäköisyys saada 1:senä, 2:sena, 3:ntena tai 4:ntenä lapsena tyttö on riippumattomuuden johdosta $1/2$, $(1/2)^2 = 1/4$, $(1/2)^3 = 1/8$ tai $(1/2)^4 = 1/16$. Erillisyyden perusteella reittien todennäköisyydet voidaan laskea yhteen. Todennäköisyys saada tyttö on $15/16$:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}.$$

[27] Mellin (1996, 55–57).



Kuva 12: Tytön saamisen todennäköisyys korkeintaan neljällä yrityksellä.

□

2.8 Kokonaistodennäköisyyden ja Bayesin kaavan ehdollistaminen

Otosavaruus S on ositettu erillisiin tapahtumiin A_i, $i = 1, \ldots, n$. Tapahtuman B todennäköisyys ehdolla tapahtuma C on

$$\mathsf{P}(B \mid C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C). \tag{16}$$

Oletus on, että $\mathsf{P}(A_i \cap C) > 0$, $i = 1, \ldots, n$. (Muuten kaikkia ehdollisia todennäköisyyksiä $\mathsf{P}(B \mid A_i \cap C)$ ei ole määritelty.) Kaava yllä saadaan soveltamalla alkuperäistä kokonaistodennäköisyyden kaavaa (11) uuteen otosavaruuteen C, jolle on ehdollistettu.

Esimerkki. Terroristi-isku. Terroristi iskee. Vastuulliseksi ilmoittautuvat kan-

sainväliset terroristiryhmät $A_1,\ldots,A_n$, joista yksi on vastuullinen. Ryhmien aiemmasta aktiviteetista arvioituna ne ovat vastuullisia todennäköisyyksillä $\mathsf{P}(A_i)$. Iskun tekotapa ja siitä jäänyt muu todistusaineisto C viittaa ryhmään A_i todennäköisyydellä $\mathsf{P}(A_i \mid C)$ ja sen sisällä terroristin kansallisuuteen B todennäköisyydellä $\mathsf{P}(B \mid A_i \cap C)$. Se ei ole välttämättä sama kuin $\mathsf{P}(B \mid A_i)$, joka voisi olla vaikkapa B:n kansalaisten osuus A_i:ssä. Todennäköisyys, että terroristi on B:n kansalainen, lasketaan kaavalla (16). □

Esimerkki. Kybervakoilu. Helsingin Sanomat 17.1.2016[28]:

> Kybervakoilijoiden ryhmä iskenyt ministeriöihin, yrityksiin ja suurlähetystöihin – - . – – Suomen ulkoministeriön tietoverkosta vuonna 2013 paljastunutta vakoiluohjelmaa ja sen taustavoimia on jäljitetty jo kymmenkunta vuotta. – – Tietoturvayhtiöiden julkaisemissa raporteissa haittaohjelman sylttytehtaan on epäilty olevan Venäjällä. Päätelmiä on tehty esimerkiksi lähdekoodista löydetyistä venäjänkielisistä sanoista.

Ulkoministeriön sisäiseen verkkoon on murtauduttu. Hyökkäyksen takana voi olla mikä tahansa tietoteknisesti kehittynyt valtio: $S = \{A_1,\ldots,A_n\}$. Aiemman vakoiluaktiviteetin perusteella valtio A_i on syyllinen todennäköisyydellä $\mathsf{P}(A_i)$. Vakoiluohjelman piirteet C viittaavat valtioon A_i todennäköisyydellä $\mathsf{P}(A_i \mid C)$. Todennäköisyys, että kybervakoilija on valtio $B = A_j$ $(j = 1,\ldots,n)$, saadaan kaavasta (16). Koska $\mathsf{P}(B \mid A_j \cap C) = 1$ ja $\mathsf{P}(B \mid A_i \cap C) = 0$, jos $i \neq j$, todennäköisyys typistyy $\mathsf{P}(A_j \mid C)$:ksi. Ehdollistaminen ei välttämättä johda monimutkaiseen laskuun, vaikka se muuttaisi todennäköisyyttä. □

Olkoot kiinnostuksen kohteena tapahtumat A, B ja C. Bayesin kaavan ehdollinen versio on

$$\mathsf{P}(A \mid B \cap C) = \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\mathsf{P}(B \mid C)} = \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)}. \quad (17)$$

Yllä oletetaan, että $\mathsf{P}(A \cap C) > 0$ ja $\mathsf{P}(B \cap C) > 0$, jotta ehdolliset todennäköisyydet ovat määriteltyjä. Viimeinen muoto olettaa taustalle lisäksi otosavaruuden jaon erillisiin tapahtumiin A_i $(i = 1,\ldots,n)$ ja että niihin liittyvät kaavassa esiintyvät ehdolliset todennäköisyydet ovat määriteltyjä.

Perustelu:

$$\begin{aligned}\mathsf{P}(A \mid B \cap C) &\overset{1.}{=} \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(B \cap C)} = \frac{\mathsf{P}(B \cap (A \cap C))}{\mathsf{P}(B \cap C)} \overset{3.}{=} \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \cap C)}{\mathsf{P}(B \cap C)} \\ &= \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)\mathsf{P}(C)}{\mathsf{P}(B \mid C)\mathsf{P}(C)} \overset{5.}{=} \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\mathsf{P}(B \mid C)}.\end{aligned}$$

Kolmas ja neljäs yhtäsuuruus seuraavat tulosäännöstä (7). Sijoittamalla kaava (16) nimittäjään yllä seuraa kaavan (17) oikeanpuoleisin muoto. Ensimmäinen, kolmas ja viides yhtäsuuruus havainnollistavat, kuinka sama ehdollinen todennäköisyys voidaan laskea monella tavalla. Sopivin kannattaa valittaa tilanteen mukaan.

[28] `Http://www.hs.fi/kotimaa/a1452786303941` (viitattu 17.2.2016).

2.9 Simpsonin paradoksi

Esimerkki.[29] Kaksi maisteria. B ja B^C ovat juuri valmistuneet maistereiksi. He ovat molemmat opiskelleet vaikeana pidettyä ainetta C ja maineeltaan helpompaa ainetta C^C. B:n pääaine oli C; B^C:n C^C. Tutkintotodistukset dokumentoivat B:n ja B^C:n suorittamat kurssit ja menestyksen niissä:

		aine				aine		
		C	C^C	Σ		C	C^C	Σ
arvosana	hyvä	70	10	80	hyvä	2	81	83
	huono	20	0	20	huono	8	9	17
	Σ	90	10	100	Σ	10	90	100

B on opiskellut enimmäkseen C:tä. B^C:n tutkintotodistuksesta näkyy, kuinka hänen suorituksensa koostuvat enimmäkseen C^C:sta. Huomionarvoisinta todistuksissa on, että

- C-aineessa B:n hyvien arvosanojen osuus $70/90 = 7/9$ on suurempi kuin B^C:n $2/10 = 1/5$.
- C^C-aineessa B:n hyvien arvosanojen osuus $10/10 = 1$ on suurempi kuin B^C:n $81/90 = 9/10$.
- kokonaisuutena katsoen hyvien arvosanojen osuus $80/100$ B:llä on silti pienempi kuin $83/100$ B^C:lla!

Jos maisterit hakevat samaa työpaikkaa, kumman työnantaja valitsee? Jos molemmat aineet antavat pätevyyden haettuun tehtävään, ehkä B:n, joka on pärjännyt molemmissa paremmin. B^C:n kokonaisuutena katsoen parempi opintomenestys saattaa johtua siitä, että hän on opiskellut helpompaa ainetta. Ylipäänsä ehdollistettu tieto on informatiivisempaa, ja siihen kannattaa kiinnittää huomiota enemmän kuin ehdollistamattomaan. □

Jos osajoukoissa tapahtumien todennäköisyyksien suuruusjärjestys on päinvastainen kuin koko joukossa, on kohdattu *Simpsonin paradoksi*. Kolmen tapahtuman A, B ja C tilanteessa pätevät tällöin suuruusjärjestykset

$$\begin{aligned} \mathsf{P}(A \mid C \cap B) &> \mathsf{P}(A \mid C \cap B^C) \ \text{ ja} \\ \mathsf{P}(A \mid C^C \cap B) &> \mathsf{P}(A \mid C^C \cap B^C) \ \text{ mutta} \\ \mathsf{P}(A \mid B) &< \mathsf{P}(A \mid B^C). \end{aligned}$$

Selitys on ehdollistettu kokonaistodennäköisyyden laki (16). Sen mukaan

$$\mathsf{P}(A \mid B) = \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B)$$

ja

$$\mathsf{P}(A \mid B^C) = \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C).$$

[29] Jakso perustuu sivuihin 67–69 kirjassa Blitzstein ja Hwang (2015).

Vaikka pätisi

$$\mathsf{P}(A \mid C \cap B) > \mathsf{P}(A \mid C \cap B^C) \text{ ja } \mathsf{P}(A \mid C^C \cap B) > \mathsf{P}(A \mid C^C \cap B^C),$$

niin $\mathsf{P}(C \mid B)$ tai $\mathsf{P}(C^C \mid B)$ voi olla niin pieni ja $\mathsf{P}(C \mid B^C)$ tai $\mathsf{P}(C^C \mid B^C)$ voi olla niin suuri, että

$$\mathsf{P}(A \mid B) < \mathsf{P}(A \mid B^C).$$

Esimerkki. Kaksi maisteria (jatkoa). Merkitään hyvää arvosanaa A:lla. Sijoitetaan tiedot todistuksista ehdollisen kokonaistodennäköisyyden kaavaan:

$$\begin{aligned}
\mathsf{P}(A \mid B) &= \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B) \\
&= \frac{70}{90} \times \frac{90}{100} + \frac{10}{10} \times \frac{10}{100} \\
&= \frac{80}{100} \\
&< \\
&= \frac{83}{100} \\
&= \frac{2}{10} \times \frac{10}{100} + \frac{81}{90} \times \frac{90}{100} \\
&= \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C) \\
&= \mathsf{P}(A \mid B^C).
\end{aligned}$$

B:n todistuksessa helpomman aineen pienehkö osuus $10/100 = \mathsf{P}(C^C \mid B)$ ja B^n:n todistuksessa vaikeamman aineen pienehkö osuus $10/100 = \mathsf{P}(C \mid B^c)$ myötävaikuttavat epäyhtälön suunnan kääntymiseen. B^C:n todistuksen hyvien arvosanojen osuutta laskettaessa termi $(2/10) \times (10/100) = 2/100$ on edellisestä johtuen pieni. B^C:n todistuksessa helpomman aineen osuus $\mathsf{P}(C^C \mid B^C) = 90/100$ on suuri ja jälkimmäinen summattava $(81/90) \times (90/100) = 81/100)$ niin suuri, että rivin ensimmäisen termin pienuudesta huolimatta muodostuu epäyhtälö yllä.

Empiirisiä esimerkkejä Simpsonin paradoksista on monia:

- Berkeley'in yliopisto haastettiin oikeuteen sukupuolisyrjinnästä vuoden 1973 opiskelijavalinnan takia. Mieshakijoista oli opiskelijoista hyväksytty selvästi suurempi osuus kuin naishakijoista. Asiaa tutkittaessa ilmeni, että monilla laitoksilla naisten hyväksymisprosentit olivat suurempia kuin miesten. Selitys miesten suurempaan hyväksymisprosenttiin ylipäänsä oli, että he pyrkivät laitoksille (esim. matemaattisiin tai teknisiin tieteisiin), joille oli helpompi päästä (hyväksyttyjen osuudella mitattuna) kuin naisten suosimiin laitoksiin (esim. psykologia).

- Baseballissa pelaaja voi pelata toista pelaajaa paremmin (esim. onnistumisosuudella mitattuna) sekä kauden ensimmäisellä että toisella kaudella mutta kokonaisuutena huonommin.

- USA:ssa Floridan osavaltiossa 1976–1986 valkoihoiset tuomittiin mustaihoisia useammin kuolemaan mutta aineiston kaikissa alaryhmissä mustaihoiset tuomittiin valkoihoisia useammin kuolemaan.

- Tupakoitsijoiden kuolleisuus on kaikissa ikäryhmissä suurempi kuin tupakoimattomien. Tupakoijien kuolleisuus ylipäänsä on kuitenkin pienempi kuin tupakoimattomien, koska tupakoijat ovat keskimäärin tupakoimattomia nuorempia. Simpsonin paradoksi ilmenee muunkinlaisissa tupakointiaineistoissa.

- Taloudellisia esimerkkejä on Man (2015) artikkelissa.

3 Kombinatoriikkaa

Mistä todennäköisyydet tulevat? Monissa yhteyksissä todennäköisyys voidaan päätellä kombinatorisilla laskuilla. Muun muassa myöhemmin esitettävät sovellusten kannalta tärkeät binomi- ja hypergeometriset jakaumat seuraavat tällaisista laskuista. Tilastotieteen ymmärtäminen edellyttää kombinatoriikan perusteisiin tutustumista.

Seuraavilla periaatteilla, varsinkin jälkimmäisellä, on paljon käyttöä.

Yhteenlaskuperiaate: Oletetaan, että

- operaatiot A ja B ovat erillisiä eli että niistä vain toinen voidaan suorittaa.

- A voidaan suorittaa n_1:llä ja B n_2:lla tavalla.

Tällöin yhdistetty operaatio ”A tai B” voidaan suorittaa $(n_1 + n_2)$:lla tavalla.

Kertolaskuperiaate: Oletetaan, että

- operaatiot A ja B voidaan suorittaa toisistaan riippumattomasti.

- A ja B voidaan suorittaa n_1:llä ja n_2:lla tavalla.

Tällöin yhdistetty operaatio ”A ja B” voidaan suorittaa $(n_1 \times n_2)$:lla tavalla.

Esimerkki. Maisteriopinnot. Opiskelija pohtii, miten erikoistua maisterivaiheessa. Linjoja on kaksi. Vaihtoehdon A voi suorittaa 56:lla ja vaihtoehdon B 126:lla erilaisella kurssikombinaatiolla. (Tarkoitus on, että suoritetaan vain toisen linjan opinnot.) Yhteensä opiskelijalla on $56 + 126 = 182$ erilaista mahdollisuutta suorittaa maisteriopinnot. (Syy kurssikombinaatioiden lukumäärille selviää myöhemmässä esimerkissä.) □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelija on niin innostunut oppiaineestaan, että pohtii, suorittaisiko (tarkoitetunvastaisesti) molempien linjojen kurssit. Hän voisi suorittaa ne $56 \times 126 = 7056$ erilaisella tavalla. □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelijalla on sivuaineopintoja suorittamatta. Ne voi suorittaa 6 tavalla. Opiskelija voisi suorittaa pääaineensa kahden linjan ja sivuaineensa opinnot $7056 \times 6 = 42336$ tavalla. □

Koostukoon joukko A erilaisista alkioista $a_1, \ldots, a_n$. Ne voidaan järjestää

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 \tag{18}$$

erilaiseen jonoon eli *permutaatioon*. Yllä $n!$ luetaan "n:n kertoma".

Perustelu: Jonon ensimmäiseksi alkioksi on mahdollista valita n alkiota, toiseksi alkioksi mahdollisuuksia on $n-1$, kolmanneksi $n-2$ jne. Toiseksi viimeinen alkio joudutaan valitsemaan kahden alkion väliltä. Viimeiseksi alkioksi jää aiemmin valitsematon alkio. Kertolaskuperiaatteen mukaan kaksi ensimmäistä operaatiota — tässä järjestämistä — voidaan tehdä $n \times (n-1)$ tavalla. Kolmas operaatio voidaan ajatella koostuvan ensimmäisestä yhdistetystä operaatiosta, joka voidaan tehdä $[n \times (n-1)]$ tavalla ja toisesta operaatiosta, joka voidaan tehdä $n-2$ tavalla. Kertolaskuperiaate tuottaa mahdollisten järjestettyjen jonojen lukumääräksi $[n \times (n-1)] \times (n-2)$. Neljäs operaatio ajatellaan jälleen koostuvaksi ensimmäisestä yhdistetystä operaatiosta $[n \times (n-1) \times (n-2)]$ vaihtoehdolla ja seuraavasta operaatiosta $(n-3)$ vaihtoehdolla. Näin jatkaen päädytään kaavaan yllä.

Esimerkki. Kertomia. 5:n kertoma on 120:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

Kertomat kasvavat nopeasti:

$$8! = 8 \times 7 \times \cdots \times 2 \times 1 = 40\ 320 \quad \text{ja}$$
$$14! = 14 \times 13 \times \cdots \times 2 \times 1 = 87\ 178\ 291\ 200.$$

Erikoistapauksena määritellään 0:n kertomaksi 1:

$$0! = 1.$$

Laskujen tulokset voi tarkistaa R:llä käskyillä `factorial(5)`, `factorial(8)` ja `factorial(14)`. □

Esimerkki. Maisteriopinnot (jatkoa). Linjalla A luennoidaan 8 kurssia. Extrainnokas opiskelija pohtii, kävisikö kaikki. Kuinka monessa järjestyksessä hän voisi suorittaa ne? Järjestyksiä on $8! = 40\ 320$. □

Monesti on tarpeen selvittää, kuinka monta k:n ($k \leq n$) pituista permutaatiota voidaan muodostaa n:stä erilaisesta alkiosta. Lukumäärä voidaan päätellä kuten edellä mutta katkaisemalla valintojen tekeminen k:n alkion järjestykseen laiton jälkeen:

$$\begin{aligned} & n \times (n-1) \times \cdots \times [n-(k-2)] \times [n-(k-1)] \\ &= n \times (n-1) \times \cdots \times (n-k+2) \times (n-k+1) \\ &= \frac{n \times \cdots \times (n-k+1) \times (n-k) \times \cdots \times 1}{(n-k) \times \cdots \times 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned} \tag{19}$$

Ensimmäisellä rivillä kerrottavia on k kappaletta.

Esimerkki. Maisteriopinnot (jatkoa). Linjan 8 kurssista 5 täytyy suorittaa maisterin tutkintoa varten. Kuinka monta vaihtoehtoista suoritusjärjestystä opiskelijalla on 5 kurssille?

Järjestyksiä on 6 720:

$$\frac{8!}{(8-5)!} = \frac{8!}{3!} = 8 \times 7 \times 6 \times 5 \times 4 = 6\ 720.$$

Samaan vastaukseen päätyy järkeilemällä, että ensimmäiseksi kurssiksi on 8 vaihtoehtoa, seuraavaksi 7 ja niin edespäin viidenteen kurssiin asti, johon on jäljellä $8 - 4 = 4$ vaihtoehtoa. Kertolaskuperiaatetta soveltamalla vastaus on $8 \times 7 \times 6 \times 5 \times 4$. □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelijaa kiinnostaa linjan 8 kurssista vain 5. Kuinka monessa järjestyksessä opiskelija voi suorittaa häntä kiinnostavat 5 kurssia?

Nyt vaihtoehtoisia kursseja on vain 5. Opiskelija voi tenttiä ne 120 järjestyksessä:

$$\frac{5!}{(5-5)!} = \frac{5!}{0!} = \frac{5!}{1} = 120.$$

Yltä ilmenee yksi syy, miksi oli hyödyllistä määritellä $0! = 1$: Sama laskusääntö $n!/(n-k)!$ pätee myös tilanteessa $n = k$. □

Olkoon joukossa A erilaisia alkioita n kappaletta: $A = \{a_1, \ldots, a_n\}$. Kuinka monta k:n kokoista ($0 \leq k \leq n$) erilaista osajoukkoa voidaan A:sta poimia? Vastaus on

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{20}$$

erilaista osajoukkoa. Suure $\binom{n}{k}$ on *binomikerroin*. Se luetaan "n yli k:n".

Tulos päätellään näin: Merkitään tuntematonta osajoukkojen lukumäärää N:llä. Niistä saadaan $N \times k!$ permutaatiota (kaava 18)). Toisaalta n alkiosta voidaan muodostaa k:n pituisia permutaatiota $n!/(n-k)!$ (kaava 19)). Täytyy siis päteä $N \times k! = n!/(n-k)!$. Ratkaisemalla yhtälöstä N saadaan tulos (20).

Esimerkki. Maisteriopinnot (jatkoa). Linjalla A luennoidaan 8 ja linjalla B 9 kurssia. Kursseja täytyy suorittaa 5. Kuinka monta erilaista kurssikokonaisuutta linjaa A opiskeleva voi muodostaa 5 kurssista? Entä linjalla B opiskeleva? Sivuainekursseja on tarjolla 4, joista 2 täytyy suorittaa. Kuinka monta 2 kurssin kombinaatiota sivuaineen kursseista voi muodostaa?

Linjalla A kursseista voi muodostaa 56 ja linjalla B 126 erilaista kokonaisuutta:

$$\binom{8}{5} = \frac{8!}{5!(8-5)!} = \frac{8!}{5!3!} = \frac{8 \times 7 \times 6}{3 \times 2} = 56 \quad \text{ja}$$
$$\binom{9}{5} = \frac{9!}{5!(9-5)!} = \frac{9!}{5!4!} = \frac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2} = 126.$$

Sivuaineen 2 kurssin kombinaatioita on 6:

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4 \times 3}{2} = 6.$$

Laskujen tulokset voi tarkistaa R:llä käskyillä

```
choose(8,5)
choose(9,5)
choose(4,2)
```

Näitä lukumääriä käytettiin ensimmäisessä ja kolmannessa Maisteriopinnot-esimerkissä. □

Esimerkki. Maisteriopinnot (jatkoa). Linjan B kurssit jakaantuvat 5 teoreettiseen ja 4 empiiriseen kurssiin. Teoreettisia kursseja pitää käydä 3 ja empiirisiä 2. Kuinka monta erilaista kurssikombinaatiota linjan B opiskelija voi suorittaa (5 kurssista)?

Opiskelija voi valita teoreettisia kursseja 10 tavalla:

$$\binom{5}{3} = \frac{5!}{3!2!} = \frac{5 \times 4}{2} = 10.$$

Empiirisissä kursseissa on 6 valintamahdollisuutta:

$$\binom{4}{2} = 6.$$

Kertolaskuperiaatteen mukaan opiskelija voi muodostaa 60 erilaista opintokokonaisuutta B-linjan maisteriopinnoista:

$$\binom{5}{3}\binom{4}{2} = 10 \times 6 = 60. \ \square$$

Perustellaan, että binomikerroin (20) on myös on erilaisten jonojen lukumäärä, kun joukko A koostuu kahdenlaisista (vaikkapa oransseista ja vihreistä) alkioista o ja v, joita on k ja $n-k$ kappaletta: Merkitään erilaisten jonojen lukumäärää N:llä. Mikäli voitaisiin erotella o-alkiot toisistaan, olisi erilaisia jonoja $N \times k!$ kappaletta, sillä o-alkiot voidaan järjestää $k!$ eri tavalla yhdessä jonossa (kaava (18)). Mikäli lisäksi voitaisiin erotella v-alkiotkin, olisi erilaisia jonoja $N \times k! \times (n-k)!$ kappaletta, sillä v-alkiot voidaan järjestää $(n-k)!$ eri tavalla yhdessä jonossa. Tällöin pystyttäisiin erottelemaan kaikki alkiot, jolloin erilaisia jonoja on $n!$. Näin ollen täytyy päteä

$$N \times k! \times (n-k)! = n!$$

eli

$$N = \frac{n!}{k! \times (n-k)!} = \binom{n}{k}. \tag{21}$$

Päätellään seuravaksi, kuinka monella tavalla A:n n erilaista alkiota voidaan jakaa k:hon erilaiseen osajoukkoon, joiden koko on $n_1, \ldots, n_k$ ($\sum_{i=1}^{k} n_i = n$). Kertolaskuperiaatteesta seuraa, että jakojen lukumäärä on

$$\begin{aligned}
&\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-n_1-n_2}{n_3} \times \cdots \times \\
&\binom{n-n_1-n_2-\cdots-n_{k-2}}{n_{k-1}} \times \binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k} \\
=&\frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \times \cdots \times \\
&\frac{(n-n_1-\cdots-n_{k-2})!}{n_{k-1}!(n-n_1-\cdots-n_{k-1})!} \times \frac{(n-n_1-\cdots-n_{k-1})!}{n_k!0!} \\
=&\frac{n!}{n_1!n_2!n_3!\ldots n_{k-1}!n_k!}.
\end{aligned}$$

Neljännellä rivillä on sijoitettu $(n-n_1-n_2-\cdots-n_{k-1})-n_k = n_k - n_k = 0$. Peräkkäisissä osamäärissä nimittäjien ja osoittajien vastaavat termit supistuvat. Laskettu lukumäärä on *multinomikerroin*:

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \ldots n_k!}. \tag{22}$$

Multinomikerroin voidaan binomikertoimen tapaan tulkita toisinkin: Multinomikerroin on n pituisten permutaatioiden lukumäärä, kun A:n alkioita on k:nlaisia ja kussakin osajoukossa on n_i alkiota ($i = 1, \ldots, k$ ja $\sum_{i=1}^{k} n_i = n$). Perustelu on samanlainen kuin binomikertoimelle esitettiin vastaavan tuloksen perustelemiseksi. Vastaavasti saadaan yhtälö

$$N \times n_1! \times n_2! \times \ldots \times n_k! = n!.$$

(N on permutaatioiden lukumäärä, $N \times n_1!$ on permutaatioiden lukumäärä, jos 1. osajoukon alkiot voitaisiin erotella, $N \times n_1! \times n_2!$, jos myös 2. osajoukon alkiot voitaisiin erotella jne.) Yhtälöstä ratkaistaan

$$N = \frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}. \tag{23}$$

Esimerkki. Maisteriopinnot (jatkoa). Linjan B kurssit jakaantuvat 5 teoreettiseen ja 4 empiiriseen kurssiin. Teoreettisia kursseja pitää käydä 3 ja empiirisiä 2. Linjan opiskelija on valinnut, mitkä 3 teoreettista ja 2 empiiristä kurssia ja mitkä 2 kurssia jäljellä olevista sivuaineopinnoistaan hän tenttii. Hän haluaa vaihtelua ja pohtii, kuinka monessa järjestyksessä hän voisi tenttiä nämä $3+2+2=7$ kurssia. Vastaus ”210:ssä järjestyksessä” saadaan multinomikertoimesta (22):

$$\binom{7}{3,2,2} = \frac{7!}{3!2!2!} = \frac{7 \times 6 \times 5 \times 4}{4} = 210. \quad \square$$

Binomikerroin on erikoistapaus multinomikertoimesta:

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

Binomikertoimet saadaan Pascalin kolmiosta:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

Ykkösestä poikkeavat luvut ovat kolmannesta rivistä alkaen luvun yllä olevien kahden luvun summa. Luvut ovat sijainniltaan vastaavia binomikertoimia kaaviossa alla:

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \\
 & & & & & \vdots & & & & &
\end{array}$$

Binomikerroin-nimitys tulee siitä, että binomikertoimet ovat kertoimia binomin $(x+y)^n$ aukikirjoitetussa muodossa

$$\begin{aligned}
(x+y)^n =& \binom{n}{0} x^n y^0 + \binom{n}{1} x^{n-1} y^1 + \binom{n}{2} x^{n-2} y^2 \\
&+ \cdots + \binom{n}{n-1} x^1 y^{n-1} + \binom{n}{n} x^0 y^n \\
=& \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i.
\end{aligned} \tag{24}$$

Binomilause (24) perustellaan näin: Termi $x^{n-i}y^i$ muodostuu kerrottaessa tulossa

$$(x+y)^n = (x+y) \times (x+y) \times \cdots \times (x+y) \times (x+y)$$

i kappaletta y:itä ja $n-i$ kappaletta x:iä keskenään. Tällainen tulo voidaan muodostaa lausekkeesta yllä $\binom{n}{i}$ kertaa. Tähän päädytään pohtimalla, kuinka monesta järjestyksestä i kappaletta y:itä ja $n-i$ kappaletta x:iä tulo $x^{n-i}y^i$ syntyy. Vastaus on binomikerroin $\binom{n}{i}$. Vaihtoehtoisesti voi ajatella, että poimitaan i kappaletta y:itä $(x+y)$-termeistä ja jäljellejäävistä $n-i$:stä $(x+y)$-termistä x:t. Binomikerroin $\binom{n}{i}$ kertoo, kuinka monta i:n kokoista osajoukkoa eriväristä y:tä voidaan poimia n:stä erivärisestä y:stä. Kukin osajoukko vastaa tiettyä valintaa y:itä ja x:iä tulossa yllä. Kaava (24) seuraa käymällä jompikumpi argumentti läpi i:n arvoille $0, \ldots, n$.

Esimerkki. Binomi astetta 5.

$$\begin{aligned}(x+y)^5 =& \binom{5}{0}x^5 + \binom{5}{1}x^4y^1 + \binom{5}{2}x^3y^2 + \binom{5}{3}x^2y^3 + \binom{5}{4}x^1y^4 + \binom{5}{5}y^5 \\ =& x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5. \quad \square\end{aligned}$$

4 Todennäköisyysjakaumia ja niiden ominaisuuksia

4.1 Ilkka Mellinin opetusmonisteen jakso 1.2

- Diskreetin ja jatkuvan satunnaismuuttujan todennäköisyysjakauma.
- Diskreetin satunnaismuuttujan pistetodennäköisyysfunktio ja jatkuvan satunnaismuuttujan tiheysfunktio.
- Diskreetin ja jatkuvan satunnaismuuttujan kertymäfunktio.
- Satunnaismuuttujan odotusarvo ja varianssi.
- Varianssin neliöjuuri standardipoikkeama (SD) (*standard deviation*) eli keskihajonta.
- Riippumattomien satunnaismuuttujien lineaarimuunnosten odotusarvo ja varianssi.

Mellinin opetusmonisteen jaksosta 1.2 sivuutetaan empiiriset jakaumat (s:t 68–69 ja 74–75), tiheysfunktion integrointi (s. 79) ja suurten lukujen laki (s:t 87–90). Viimeksi mainittu esitetään alla luvussa 6.

4.2 Diskreettien satunnaismuuttujien todennäköisyysjakaumia

4.2.1 Bernoulli-jakauma

Bernoulli-kokeessa on kaksi tulosvaihtoehtoa: A tapahtuu tai ei. *Bernoulli-satunnaismuuttuja* (X) saadaan, kun tulosvaihtoehtoihin liitetään luvut 1 (tapahtuu) ja 0 (ei tapahdu). Bernoulli-satunnaismuuttujan jakauman määrittelee parametri π, joka on todennäköisyys 1:lle eli tapahtumiselle. Jos X noudattaa Bernoulli-jakaumaa parametrilla π, merkitään $X \sim \mathsf{B}(\pi)$.[30]

Bernoulli-jakautuneen satunnaismuuttujan odotusarvo ja varianssi ovat π ja $\pi(1-\pi)$:

$$\mathsf{E}(X) = \pi \times 1 + (1-\pi) \times 0 = \pi \tag{25}$$

ja

$$\begin{aligned} \mathsf{V}(X) = \mathsf{E}(X-\mu)^2 = \mathsf{E}(X-\pi)^2 &= \pi \times (1-\pi)^2 + (1-\pi) \times (0-\pi)^2 \\ &= \pi(1-\pi)^2 + \pi^2(1-\pi) = \pi(1-\pi)(1-\pi+\pi) = \pi(1-\pi). \end{aligned} \tag{26}$$

Yllä on varianssin määritelmässä $\mathsf{E}(X-\mu)^2$ merkitty odotusarvoa μ:llä. Bernoulli-jakautunut satunnaismuuttuja on yksinkertaisin mahdollinen satunnaismuuttuja. Tapahtumaa tai tapahtumattomuutta kuvataan monesti ilmaisuilla ”onnistuminen” tai ”epäonnistuminen”, ”voitto” tai ”häviö”, ”sairastuminen” tai ”pysyminen terveenä” jne.

Esimerkki. Lantin heitto (jatkoa). Kuvataan ”kruuna” ykköseksi ja ”klaava” nollaksi. Lantin heitto -satunnaismuuttuja X on Bernoulli-jakautunut parametrilla $\pi = 0.5$ (harhaton lantti). Satunnaismuuttujan X odotusarvo ja varianssi ovat 0.5 ja $0.5^2 = 0.25$.

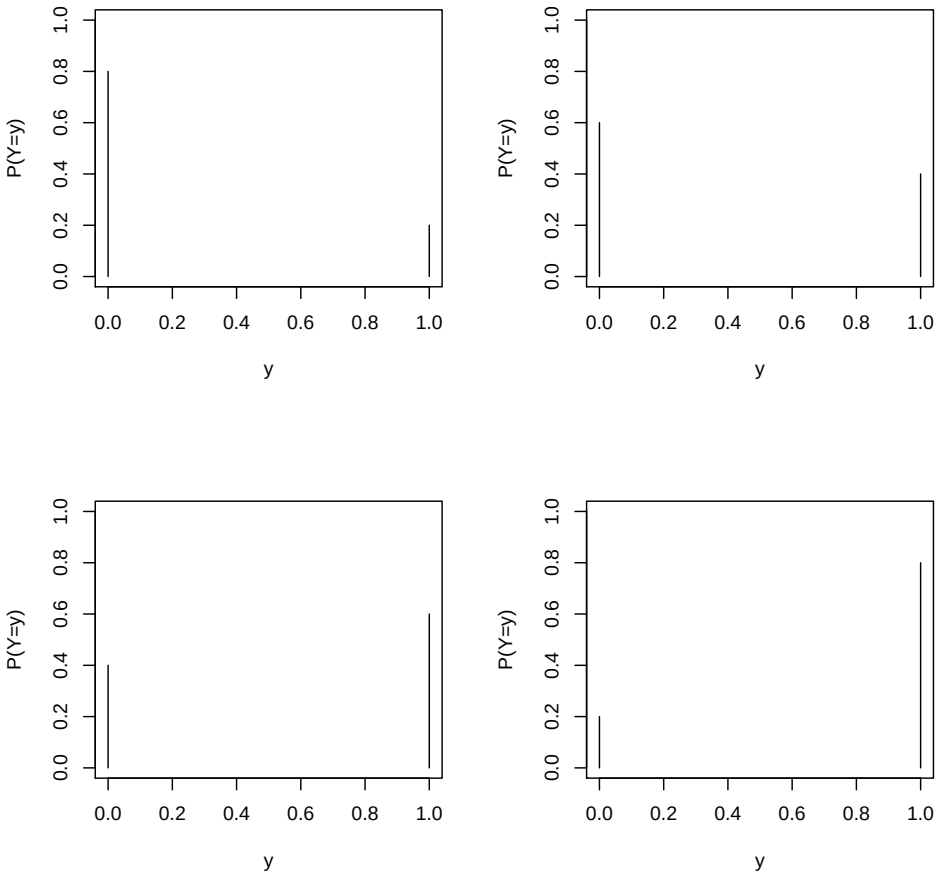
Tapahtumaa, joka vastaisi 0.5:ttä, ei ole. Odotusarvoa vastaava tapahtuma ei ole välttämättä ”yleinen” tai edes mahdollinen. □

Bernoulli-satunnaismuuttujan *pistetodennäköisyysfunktio* on

$$\mathsf{P}(Y=y) = \pi^y(1-\pi)^{1-y} = \begin{cases} \pi, & \text{jos } y=1, \\ 1-\pi, & \text{jos } y=0. \end{cases}$$

Kuva 13 havainnollistaa Bernoulli-jakaumaa π:n arvoilla 0.2, 0.4, 0.6 ja 0.8.

[30] Jakaumien merkinnöissä on vaihtelua kirjallisuudessa. Tämä tai muut merkinnät alla eivät ole vakiintuneita.



Kuva 13: Bernoulli-jakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita π:n arvoilla 0.2, 0.4, 0.6 ja 0.8.

4.2.2 Diskreetti tasainen jakauma

Satunnaismuuttuja X noudattaa *diskreettiä tasaista jakaumaa*, jos sen pistetodennäköisyysfunktio on

$$\mathsf{P}(X = x) = \begin{cases} \dfrac{1}{n}, & \text{jos } x = x_1, \ldots, x_n, \\ 0, & \text{muutoin.} \end{cases}$$

Mikäli luvut $x_1, \ldots, x_n$ ovat peräkkäisiä kokonaislukuja, kätevät esitysmuodot odotusarvolle ja varianssille ovat

$$\mathsf{E}(X) = (x_1 + x_n)/2$$

ja

$$\mathsf{V}(X) = [(x_n - x_1 + 1)^2 - 1]/12$$

(esim. Tijms 2012, 313).

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Lukujen $x_1, \ldots, x_n$ ollessa peräkkäisiä kokonaislukuja odotusarvo on

$$\mathsf{E}(X) = \sum_{i=0}^{n-1} \frac{x_1 + i}{n} = x_1 + \frac{1}{n} \sum_{i=0}^{n-1} i = x_1 + \frac{1}{n} \frac{n(n-1)}{2} = x_1 + \frac{n-1}{2} = \frac{x_1 + x_1 + n - 1}{2}$$
$$= \frac{x_1 + x_n}{2}.$$

Yllä on hyödynnetty kaavaa $\sum_{i=1}^{n} i = n(n+1)/2$. Sen todistus: Merkitään $S = \sum_{i=1}^{n} i$. Selvästikin pätee

$$S = 1 + 2 + \cdots + (n-1) + n \quad \text{ja} \quad S = n + (n-1) + \cdots + 2 + 1.$$

Lasketaan yhtälöt puolittain yhteen. Saadaan

$$2S = (n+1) + (n+1) + \cdots + (n+1) + (n+1) = n(n+1).$$

Ratkaistaan S:

$$S = \frac{n(n+1)}{2}.$$

Esimerkki. Nopan heitto (jatkoa). Nopan silmäluku (X) on diskreetisti tasaisesti jakautunut satunnaismuuttuja. Kunkin silmäluvun todennäköisyys on 1/6. Silmäluvun odotusarvo ja varianssi ovat

$$\mathsf{E}(X) = \sum_{i=1}^{6} i \times \frac{1}{6} = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5 \quad \text{ja}$$

$$\mathsf{V}(X) = \sum_{i=1}^{6} (i - 3.5)^2 \times \frac{1}{6} = (1 - 3.5)^2 \times \frac{1}{6} + \cdots + (6 - 3.5)^2 \times \frac{1}{6} = \frac{35}{12} \approx 2.92.$$

Ne voi laskea kätevämmin näin: $(1+6)/2 = 3.5$ ja $[(6-1+1)^2 - 1]/12 = 35/12$. □

4.2.3 Binomijakauma

On tehty tai havaittu n toisistaan riippumatonta samanlaista Bernoulli-koetta (kussakin tapahtumatodennäköisyys on π). Merkitään tapahtumien lukumäärää niissä y:llä. Sitä vastaavan satunnaismuuttujan pistetodennäköisyys on

$$\mathsf{P}(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}. \tag{27}$$

Jakaumaa kutsutaan *binomijakaumaksi* (otoskoolla n ja) parametrilla π ja sitä merkitään $\mathsf{Bin}(n,\pi)$.

Selitys pistetodennäköisyydelle (27): Kaikkien n:n pituisten tapahtumajonojen, joissa on y tapahtumaa, todennäköisyys on $\pi^y(1-\pi)^{n-y}$. Esimerkiksi jos kolme ensimmäistä koetta onnistuvat, seuraava epäonnistuu ja kaksi viimeistä koetta onnistuu ja epäonnistuu ja onnistumisia on yhteensä y, havaitun tapahtumajonon todennäköisyys on

$$\pi\pi\pi(1-\pi) \times \cdots \times \pi(1-\pi) = \pi^y(1-\pi)^{n-y}.$$

Muoto oikealla saadaan kokoamalla tulon termit. Järjestyksestä riippumatta muotoilu oikealla pätee, jos onnistumisia on y kappaletta. (Vrt. uhkapelurin virhepäätelmä.) Vaihtoehtoisia järjestyksiä y:lle onnistumiselle ja $n-y$:lle epäonnistumiselle on binomikertoimen $\binom{n}{y}$ mukainen määrä. Kukin järjestys on erillinen. Todennäköisyys $\mathsf{P}(Y = y)$ saadaan erillisyyden perusteella laskemalla kaikkien mahdollisten jonojen, joissa on y onnistumista, todennäköisyydet yhteen:

$$\mathsf{P}(Y = y) = \pi^y(1-\pi)^{n-y} + \cdots + \pi^y(1-\pi)^{n-y} = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

Merkitään binomijakauman taustalla olevia yksittäisiä Bernoulli-satunnaismuuttujia X_i:llä (saa arvon 1 tai 0). Tällöin $Y = \sum_{i=1}^{n} X_i$. Nyt voidaan päätellä binomijakautuneen satunnaismuuttujan odotusarvo ja varianssi $n\pi$ ja $n\pi(1-\pi)$:

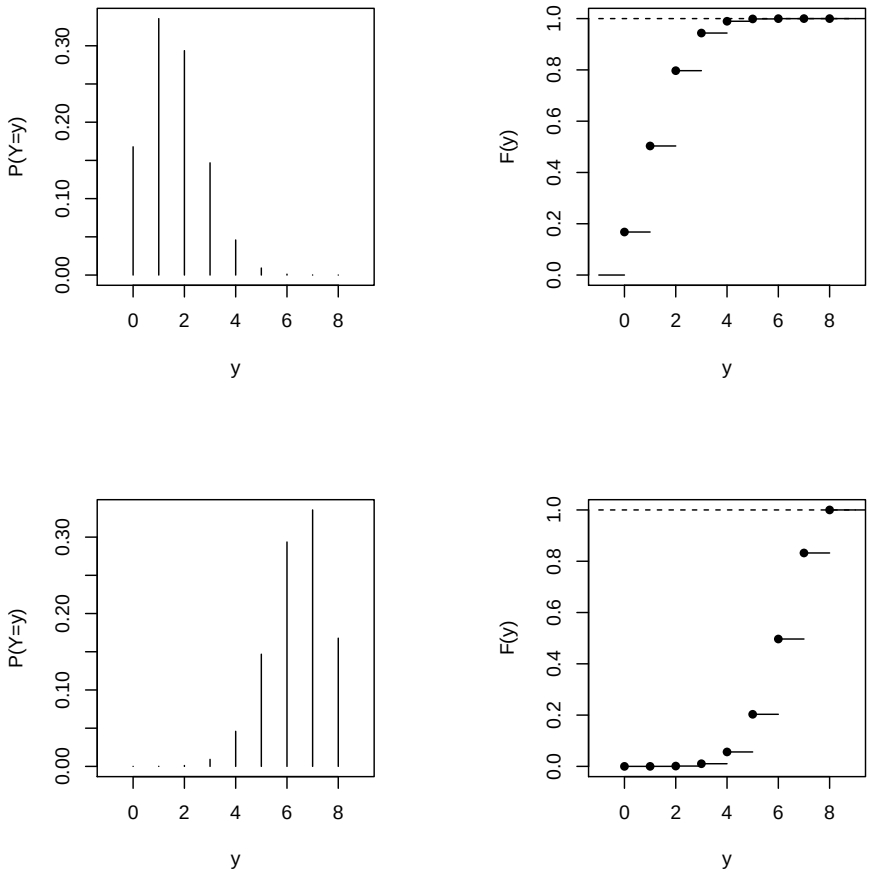
$$\mathsf{E}(Y) = \mathsf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} \pi = n\pi \tag{28}$$

ja

$$\mathsf{V}(Y) = \mathsf{V}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{V}(X_i) = \sum_{i=1}^{n} \pi(1-\pi) = n\pi(1-\pi). \tag{29}$$

Ylle on sijoitettu Bernoulli-satunnaismuuttujan odotusarvo π ja varianssi $\pi(1-\pi)$ kaavoista (25) ja (26). Varianssin laskussa on hyödynnetty oletusta satunnaismuuttujien X_i riippumattomuudesta.

Binomijakauma on epäsymmetrinen, paitsi jos $\pi = 0.5$. Kuva 14 havainnollistaa binomijakauman pistetodennäköisyys- ($\mathsf{P}(Y = y)$) ja kertymäfunktiota ($F(y)$) π:n arvoilla 0.2 ja 0.8.[31] Pistetodennäköisyysfunktiot ovat toistensa peilikuvia, koska ne määrittävät parametriarvot sijaitsevat yhtäkaukana 0.5:stä.



Kuva 14: Binomijakautuneen satunnaismuuttujan pistetodennäköisyys- ja kertymäfunktiot π:n arvoilla 0.2 ja 0.8, kun $n = 8$.

Havaintojen poimiminen ei saa vaikuttaa tapahtuman todennäköisyyteen myöhemmin, jotta satunnaismuuttuja voisi olla binomijakautunut. Tällaisia ovat tilanteet, joissa kukin poimittu alkio palautetaan takaisin perusjoukkoon tai perusjoukko on äärettömän (tai hyvin) suuri, jolloin alkion poiminta ei vaikuta mitään (tai vaikuttaa mitättömästi) seuraavan Bernoulli-kokeen tuloksen todennäköisyyteen.

[31] Kuva pohjautuu Alan Arnholtin R-koodiin (`https://github.com/alanarnholt/PASWR2E-Rscripts/blob/master/ChapterScripts/chapter04.R`; viitattu 26.2.2016). Koodi on luvusta 4 kirjasta Ugarte ym. (2016).

Binomijakauma on sovellusten kannalta tärkeimpiä jakaumia. Yksi syy on, että binomijakauma kuvaa luontevasti monia ilmiöitä. Toinen syy on, että monesti kiinnostuksen kohde on suhteellinen osuus Y/n. Sen jakauma on sama kuin Y:n, kun n on kiinteä. Y on vain jaettu vakiolla suhteellisen osuuden laskemiseksi. Myös jakauman yksinkertaisuus on ilmeinen syy sen suosiolle. Siksi sillä approksimoidaan monia tilanteita, jotka ovat lähes mutteivät täsmälleen kuvattavissa binomijakaumalla.

Jos satunnaismuuttuja $Y \sim \mathsf{Bin}(n,\pi)$, niin R laskee binomitodennäköisyyden tapahtumalle $Y = y$ käskyllä

`dbinom(y,n,π)`

kun siihen sijoittaa sopivat arvot.

Esimerkki. Kortin peluu (jatkoa). Vedetään hyvinsekoitetusta korttipakasta sattumanvaraisesti kortti, katsotaan se, palautetaan kortti pakkaan ja sekoitetaan pakka. Toistetaan tämä 5 kertaa. Lasketaan todennäköisyydet, että kuninkaita tulee 0, 1, 2, 3, 4 tai 5 kappaletta sekä todennäköisyys, että ainakin 1 nostetuista korteista on ollut kuningas.

Kukin kortin nosto on Bernoulli-koe, jossa tapahtuman ”kuningas” todennäköisyys on $4/52 = 1/13$. Lasketaan kysytyt binomijakauman pistetodennäköisyydet (Y on ”kuninkaiden lukumäärä”):

$$\mathsf{P}(Y=0) = \binom{5}{0}\left(\frac{1}{13}\right)^0\left(1-\frac{1}{13}\right)^{5-0} = \left(\frac{12}{13}\right)^5 = 0.6701769,$$

$$\mathsf{P}(Y=1) = \binom{5}{1}\left(\frac{1}{13}\right)^1\left(1-\frac{1}{13}\right)^{5-1} = 5\times\frac{1}{13}\times\left(\frac{12}{13}\right)^4 = 0.2792404,$$

$$\mathsf{P}(Y=2) = \binom{5}{2}\left(\frac{1}{13}\right)^2\left(1-\frac{1}{13}\right)^{5-2} = 10\times\left(\frac{1}{13}\right)^2\times\left(\frac{12}{13}\right)^3 = 0.04654006,$$

$$\mathsf{P}(Y=3) = \binom{5}{3}\left(\frac{1}{13}\right)^3\left(1-\frac{1}{13}\right)^{5-3} = 10\times\left(\frac{1}{13}\right)^3\times\left(\frac{12}{13}\right)^2 = 0.003878339,$$

$$\mathsf{P}(Y=4) = \binom{5}{4}\left(\frac{1}{13}\right)^4\left(1-\frac{1}{13}\right)^{5-4} = 5\times\left(\frac{1}{13}\right)^4\times\frac{12}{13} = 0.0001615974 \quad \text{ja}$$

$$\mathsf{P}(Y=5) = \binom{5}{5}\left(\frac{1}{13}\right)^5\left(1-\frac{1}{13}\right)^{5-5} = \left(\frac{1}{13}\right)^5 = 0.000002693291.$$

Todennäköisyydet on saatu sijoittamalla $\pi = 1/13$ ja $n = 5$ kaavaan (27) ja laskemalla ne R-komennoilla `dbinom(0,5,1/13)`, `dbinom(1,5,1/13)` jne.

Todennäköisyys saada ainakin 1 kuningas on todennäköisyys saada 1 tai enemmän kuninkaita. Erillisten tapahtumien yhteenlaskusäännön (4) perusteella todennäköisyydet näille tapahtumille voidaan laskea yhteen:

$$0.279 + 0.047 + 0.004 + 0.000 + 0.000 = 0.330.$$

Tulos 0.330 saataisiin helpommin laskemalla se komplementtitapahtuman todennäköisyyden ($\mathsf{P}(Y = 0)$) avulla:

$$1 - \mathsf{P}(Y = 0) = 1 - 0.6701769 \approx 0.330.$$

Todetaan lisäksi, että odotusarvo ja varianssi kuninkaiden lukumäärälle ovat noin 0.385 ja 0.355 (kaavat (28) ja (29)):

$$\mathsf{E}(Y) = 5 \times \frac{1}{13} = \frac{5}{13} \approx 0.385 \quad \text{ja}$$
$$\mathsf{V}(Y) = 5 \times \frac{1}{13} \times \frac{12}{13} = \frac{60}{169} \approx 0.355. \ \square$$

Esimerkki. Huoltoriidat käräjäoikeuksissa Suomessa. Oikeuspoliittinen tutkimuslaitos (nykyinen Kriminologian ja oikeuspolitiikan instituutti) tutki huoltoriitoja lapsista suomalaisissa käräjäoikeuksissa 14.11.2005–13.2.2006 (529 havaintoa). Tutkitaan päätöksiä, joissa lapset määrättiin asumaan vain jommankumman vanhemman luona. Niissä lapsi osoitettiin asumaan 35:ssä isän ja 83:ssa äidin luona (118 havaintoa).[32] Vastaavat prosenttiosuudet ovat noin 29.7 ja 70.3. Oletetaan, että käräjäoikeudet ylipäänsä määräisivät lapset asumaan eri sukupuolta olevien vanhempien luona yhtä todennäköisesti eli prosenttiosuuksilla 50 ja 50 ja että päätökset olisivat riippumattomia toisistaan. Mikä olisi odotusarvo lapsille, jotka oikeus osoittaa asumaan isän luona? Entä äidin luona? Mikä olisi todennäköisyys, että 118:n suuruisessa satunnaisessa aineistossa 35:ssä tai vähäisemmässä lukumäärässä lapset määrätään asumaan isän luona?

Kukin päätös voidaan mieltää Bernoulli-kokeena, ja päätösten lukumäärä ajatella binomijakautuneeksi. Päätökset ajatellaan poimituiksi äärettömästä määrästä potentiaalisia päätöksiä, joita käräjäoikeudet olisivat voineet tehdä.

Binomijakauman mukaan odotusarvo sekä isän että äidin luokse osoitettavien lasten lukumäärälle on $118 \times 0.5 = 59$ (kaava (28)). Kysytty todennäköisyys saadaan laskemalla $\mathsf{Bin}(118,0.5)$-jakauman kertymäfunktion arvo pisteessä 35. R-käsky

```
pbinom(35,118,0.5)
```

antaa täksi todennäköisyydeksi noin 0.00006. Jos $\pi = 0.5$, niin havaittu lukumäärä 35 on poikkeuksellisen pieni. $\square$

4.2.4 Multinomijakauma

On tehty tai havaittu n toisistaan riippumatonta koetta. Kussakin kokeessa on c tulosvaihtoehtoa tai solua, jonka todennäköisyys on π_c ($\sum_{i=1}^{c} \pi_i = 1$). Satunnaismuuttuja koostuu frekvensseistä eli lukumääristä N_i kussakin solussa eli on moniulotteinen ($\sum_{i=1}^{c} N_i = n$). Havaitun tapahtumajonon todennäköisyys on

$$\pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c},$$

[32] Lapset määrättiin asumaan isän luona 27.3 %:ssa, äidin luona 65.2 %:ssa ja 7.5 %:ssa päätöksistä molemmilla (lapset "jaettiin" tai määrättiin "vuoroasumisesta"). Havaintoja oli 127. Periaatteessa on mahdollista, että samoista lapsesta olisi riidelty samanvuoden aikana useamman kerran aineistossa. Se mahdollisuus sivuutetaa, ja esimerkissä oletetaan, että havainnot ovat riippumattomia. E. Valkama ja M. Litmala (2006): Lasten huoltoriidat käräjäoikeuksissa. OPTL:n julkaisuja 224. `Https://helda.helsinki.fi/handle/10138/152456` (viitattu 25.2.2016).

jossa n_i:t ovat lukumääriä, joilla i. tulosvaihtoehto on toteutunut. Todennäköisyys on sama riippumatta järjestyksestä, jossa tulosvaihtoehdot ovat toteutuneet, jos n_i:t ovat samat. Tapahtumajonoja on yhteensä

$$\frac{n!}{n_1!n_2!\ldots n_c!}$$

(kaava (23)). Ne ovat erillisiä, joten todennäköisyys n_1:lle havainnolle luokassa 1, n_2:lle havainnolle luokassa 2 jne. on

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1!n_2!\ldots n_c!}\pi_1^{n_1}\pi_2^{n_2}\ldots\pi_c^{n_c} \qquad (30)$$

($\sum_{i=1}^{c} N_i = n$). Jakaumaa kutsutaan *multinomijakaumaksi* (otoskoolla n ja) parametreilla $\pi_1, \ldots, \pi_c$, ja sitä merkitään $\mathsf{Mul}(n, \pi_1, \ldots, \pi_c)$. Jakauma määräytyy jo c:stä π_i-parametrista, sillä viimeinen voidaan ratkaista rajoituksesta $\sum_{i=1}^{c} \pi_i = 1$.

Solun i lukumäärän odotusarvo ja varianssi ovat

$$\mathsf{E}(N_i) = n\pi_i$$

ja

$$\mathsf{V}(N_i) = n\pi_i(1 - \pi_i).$$

Ne seuraavat Binomijakaumasta: Solun i lukumäärä on binomijakautunut karkeistamalla Multinomijakauman solujako kahdeksi: i. soluksi ja muiksi soluiksi.

Pearsonin χ^2-testi on käytetyimpiä tilastollisia testejä. Sitä käytettäessä aineisto voidaan monesti ajatella syntyneeksi Multinomijakaumasta. Se on relevantti lukuisissa muissakin yhteyksissä kuten gallupeissa.

Esimerkki. Gallup. Tehdään otantatutkimus äänestysikäisten suomalaisten puoluekannoista. On päätetty selvittää 1000:lta riippumattomasti ja sattumanvaraisesti poimitulta äänestysikäiseltä (kullakin yhtäsuuri todennäköisyys tulla poimituksi otokseen) kannattavatko he vasemmistoa, keskustaa vai oikeistoa (V, K ja O). Kukin haastattelu on koe, jossa on kolme tulosvaihtoehtoa. Kunkin haastattelun jälkeen haastateltu palautetaan perusjoukkoon ja hänet saatetaan haastatella uudestaan, jos hän tulee poimituksi myöhemmin.[33]

Otantatutkimus tuottaa kolmesta solusta koostuvan taulukon alla. Haastattelujen järjestyksestä riippumatta kaikki tulosvaihtoehtojonot, joissa olisi yhtämonta havaintoa kutakin puoluekantaa, olisivat yhtä todennäköisiä.

V	K	O
320	350	330

Oletetaan, että kunkin solun todennäköisyys on 1/3 eli että vasemmistolla, keskustalla ja oikeistolla on yhtäsuuri kannatus. (Poliittinen kanta noudattaa Diskreettiä tasaista jakaumaa.) Havaittujen kannatuslukumäärien todennäköisyys on tällöin kaavan (30) mukaan noin 0.0004. Se on laskettu R:n komennoilla

[33] Käytännössä kerran haastateltua ei haastateltaisi uudestaan. Tällöinkin Multinomijakaumalla voitaisiin approksimoida gallupin vastausten jakaumaa, sillä haastateltujen kansalaisten lukumäärä gallupissa olisi pieni suhteessa äänestysikäisten suomalaisten lukumäärään.

```
x <- c(320, 350, 330)
prob <- c(1/3, 1/3, 1/3)
dmultinom(x, size = 1000, prob)
```

Tällaisen yksittäisen solufrekvenssikombinaation todennäköisyys on tyypillisesti hyvin pieni. Mielenkiintoisempi on esimerkiksi kysymys, mikä on todennäköisyys havaita vasemmistolle 32 %:n tai sitä pienempi kannatus, jos todellisuudessa vasemmiston kannatus on 33.33 %. Tähän kysymykseen vastaus voidaan laskea binomijakauman kertymäfunktiosta ajatellen kyselyä binomikokeena, jossa on kaksi vaihtoehtoa "vasemmisto" ja "muut" vasemmiston todennäköisyyden ollessa $1/3$. Kysytty todennäköisyys on noin 0.19 (laskettu R:n käskyllä pbinom(320,1000,1/3)). Ei ole poikkeuksellista havaita 1000 henkilöä kattavassa otantatutkimuksessa 32 %:n kannatus puolueelle, jonka kannatus on todellisuudessa 33.33 %. □

4.2.5 Hypergeometrinen jakauma

Hypergeometrinen jakauma on relevantti tilanteissa, joissa perusjoukon koko on pienehkö tai ei ainakaan suuri ja havainnot poimitaan perusjoukosta palauttamatta niitä siihen poiminnan jälkeen. Kukin poiminta on Bernoulli-koe muttei riippumaton sellainen. Jakauma tavataan perustella pallojen poimimis -analogialla.

Pussissa on l liilaa ja m mustaa palloa. Poimitaan pussista yksitellen sattumanvaraisesti n palloa palauttamatta niitä pussiin. Hypergeometrisen jakauman pistetodennäköisyydet

$$\mathsf{P}(Y = y) = \frac{\binom{l}{y}\binom{m}{n-y}}{\binom{l+m}{n}} \tag{31}$$

ovat todennäköisyyksiä saada y liilaa palloa. Yllä $0 \leq y \leq l$ ja $0 \leq n - y \leq m$; muulloin $\mathsf{P}(Y = y) = 0$.

Jakauman (31) johto: Kaikki poimitut pallokombinaatiot ovat yhtä todennäköisiä. Erilaisia n pallon kombinaatioita on $\binom{l+m}{n}$. Liiloista palloista voidaan poimia y palloa $\binom{l}{y}$ tavalla. Mustista palloista saa $n - y$ pallon erilaisia kombinaatiota $\binom{m}{n-y}$. Kombinaatiot liiloista ja mustista palloista voidaan muodostaa toisistaan riippumattomasti, joten kertolaskuperiaatteesta seuraa, että erilaisia tapoja muodostaa y liilan ja $n - y$ mustan pallon kombinaatiota on $\binom{l}{y}\binom{m}{n-y}$. Suhteuttamalla se klassisen todennäköisyyden määritelmän mukaisesti kaikkien mahdollisten kombinaatioiden lukumäärään saadaan y liilan pallon poimimistodennäköisyydeksi kaava (31).

Kun satunnaismuuttuja Y noudattaa hypergeometrista jakaumaa parametreilla l ja m (otoskoolla n), merkitään $Y \sim \mathsf{HG}(l, m, n)$. Sen odotusarvo ja varianssi ovat

$$\mathsf{E}(Y) = n \times \frac{l}{l+m} = n\pi \tag{32}$$

ja

$$\mathsf{V}(Y) = n \times \frac{l}{l+m} \times \frac{m}{l+m} \times \frac{l+m-n}{l+m-1} = n\pi(1-\pi) \times \frac{l+m-n}{l+m-1} \qquad (33)$$

(esim. Lindgren 1976, 169–170). Yllä π on todennäköisyys saada liila pallo ensimmäisellä nostolla.

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Osoitetaan, että todennäköisyys saada liila pallo myös i:llä nostolla on π $(i = 1, \ldots, n)$ ja että $\mathsf{E}(Y) = n\pi$. Määritellään satunnaismuuttuja X_i:

$$X_i = \begin{cases} 1, & \text{jos pallo on liila } i. \text{ nostolla,} \\ 0, & \text{jos pallo on musta } i. \text{ nostolla.} \end{cases}$$

Pallot (n kappaletta) voidaan poimia

$$(l+m)(l+m-1)\cdots[l+m-(n-1)] = (l+m)(l+m-1)\cdots(l+m-n+1)$$

järjestyksessä (ajatellen palloja yksilöinä; kaava (19)). Kaikki permutaatiot ovat yhtä todennäköisiä. Permutaatioita, joissa i. pallo on liila, on

$$l(l+m-1)\cdots[l+m-(n-1)] = l(l+m-1)\cdots(l+m-n+1).$$

Yllä i. liilaksi palloksi on l vaihtoehtoa. Loput $n-1$ palloa voidaan järjestää $(l+m-1)\cdots(l+m-n+1)$ tavalla. Todennäköisyys, että i. pallo on liila, on permutaatioiden lukumäärä, joissa ehto toteutuu suhteessa kaikkien permutaatioiden lukumäärään:

$$\mathsf{P}(X_i = 1) = \frac{l(l+m-1)\cdots(l+m-n+1)}{(l+m)(l+m-1)\cdots(l+m-n+1)} = \frac{l}{l+m} = \pi.$$

Liilojen pallojen lukumäärä n:ssä nostossa on $Y = \sum_{i=1}^{n} X_i$. Sen odotusarvo on

$$\mathsf{E}(Y) = \mathsf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} \pi = n\pi.$$

Satunnaismuuttujat X_i eivät ole riippumattomia, mutta summan odotusarvo lasketaan silti yllä esitetyllä tavalla.

Vastaava summan varianssin laskusääntö edellyttää summattavien satunnaismuuttujien riippumattomuuden. Liilojen pallojen lukumäärän varianssin lasku on monimutkaisempi ja sivuutetaan.

Esimerkki.[34] Liilojen ja mustien pallojen poimiminen. Pussissa on sekaisin 3 liilaa ja 7 mustaa palloa. Poimitaan 3 palloa pussista palauttamatta niitä pussiin nostojen jälkeen. Määritellään satunnaismuuttuja X_i, joka saa arvon 1 tai 0, jos i. pussista nostettu pallo on liila tai musta.

Todennäköisyys, että 1. pussista nostettu pallo on liila, on 3/10.

$$\mathsf{P}(X_1) = \frac{3}{3+7} = \frac{3}{10}.$$

Osoitetaan, että todennäköisyys 3. pussista nostetun pallon liilaudelle on myös 3/10. Liila pallo voi tulla kolmantena $2 \times 2 \times 1 = 4$:llä eri tavalla. Kokonaistodennäköisyyden lakia (11) soveltamalla saadaan todettu tulos:

$$\mathsf{P}(X_3) = \mathsf{P}(''001'') + \mathsf{P}(''011'') + \mathsf{P}(''101'') + \mathsf{P}(''111'')$$

[34] Lindgren (1976, 168).

$$= \frac{7}{10} \times \frac{6}{9} \times \frac{3}{8} + \frac{7}{10} \times \frac{3}{9} \times \frac{2}{8} + \frac{3}{10} \times \frac{7}{9} \times \frac{2}{8} + \frac{3}{10} \times \frac{3}{9} \times \frac{1}{8} = \frac{3}{10}.$$

Esimerkiksi $''001''$ tarkoittaa, että on nostettu ensin kaksi mustaa ja kolmanneksi liila pallo.

Huom! Todennäköisyys, ehdolla aiempien nostojen tulokset, ylipäänsä poikkeaa edellä lasketusta ehdollistamattomasta todennäköisyydestä. Jos on nostettu 3. ensimmäisellä nostolla liila pallo, todennäköisyys, ehdolla aiemmat nostot, nostaa seuraavaksi liila pallo on 0.

Tulos yleistyy: Jos palloja on n, liilan pallon nostamisen todennäköisyys on sama kaikilla i:n arvoilla, $i = 1, \ldots, n$. □

$\mathsf{HG}(l, m, n)$-jakauman pistetodennäköisyyden pisteessä y voi laskea R:llä komennolla

```
dhyper(y,l,m,n).
```

Komentoon sijoitetaan l:n, m:n ja n:n arvot.

Esimerkki. Kortin peluu (jatkoa). Vedetään hyvinsekoitetusta korttipakasta sattumanvaraisesti kortti ja katsotaan se, muttei palauteta sitä pakkaan. Toistetaan tämä 5 kertaa (jaetaan pelaajalle "käsi"). Lasketaan todennäköisyydet, että kuninkaita tulee 0, 1, 2, 3, tai 4 kappaletta (5 kappaletta ei voida saada, koska kortteja ei palauteta pakkaan) sekä todennäköisyys, että ainakin 1 nostetuista korteista on kuningas.

Kortin nostojen tulos on hypergeometrisesti jakautunut, koska kortteja ei palauteta pakkaan nostojen jälkeen. Pistetodennäköisyydet ovat:

$$\mathsf{P}(Y=0) = \frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} \approx 0.658842, \quad \mathsf{P}(Y=1) = \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} \approx 0.2994736,$$

$$\mathsf{P}(Y=2) = \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} \approx 0.0399298, \quad \mathsf{P}(Y=3) = \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \approx 0.0017361$$

ja

$$\mathsf{P}(Y=4) = \frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}} \approx 0.0000185.$$

Todennäköisyydet voidaan laskea R:n `choose(,)`-komennon avulla. Esimerkiksi todennäköisyys 0.658842 on laskettu käskyllä

```
choose(4,0)*choose(48,5)/choose(52,5)
```

Saman tuloksen saa R:n hypergeometrisia pistetodennäköisyyksiä tuottavalla käskyllä

```
dhyper(0,4,48,5)
```

Siinä 0 on tapahtumien lukumäärä, jolle pistetodennäköisyys lasketaan, 4 on kuninkaiden lukumäärä pakassa (edellä liilat pallot pussissa), 48 on muiden korttien lukumäärä pakassa (edellä mustat pallot pussissa) ja 5 on pakasta vedettävien korttien lukumäärä (poimittujen pallojen lukumäärä edellä). Käskyssä ei tarvitse määritellä muista korteista nostettavien korttien lukumäärää, koska se määräytyy nostettavien kuninkaiden ja yhteensä nostettavien korttien lukumääristä.

Todennäköisyys, että ainakin 1 nostetuista korteista on kuningas, on noin 0.341:

$$\begin{aligned}&0.2994736 + 0.03992982 + 0.001736079 + 0.00001846893\\&= 0.341158\\&= 1 - 0.658842.\end{aligned}$$

Todennäköisyys on laskettu ensin tapahtumien erillisyyteen (kaava (4)) perustuen ja lopuksi lyhyemmin komplementtitapahtuman todennäköisyyden kautta.

Todennäköisyydet poikkeavat vähän aiemmasta kortin peluu -esimerkistä, jossa kukin vedetty kortti palautettiin pakkaan ja kuninkaiden lukumäärä noudatti binomijakaumaa. Todennäköisyys jäädä ilman yhtään kuningasta on tässä pienempi ($0.659 < 0.670$), koska kortteja pakasta vedettäessä on todennäköisempää, että saadaan muu kortti kuin kuningas. Se kasvattaa todennäköisyyttä saada myöhemmin kuningas. Odotusarvo kuninkaiden lukumäärälle on kuitenkin sama kuin binomijakautuneessa versiossa (kaavat (28) ja (32)).

Myöskään todennäköisyys saada 1 tai enemmän kuninkaita ei poikkea juurikaan esimerkeissä ($0.341 > 0.330$), vaikka binomijakautuneessa tilanteessa on mahdollista saada 5 kuningasta. Sen todennäköisyys on merkityksettömän pieni. □

Esimerkki.[35] Toisilleen tuntemattomat Antti ja Anna tutustuvat miljoonakaupungissa — vaikkapa pääkaupunkiseudulla. Molemmilla on 500 tuttua kaupungissa. Oletetaan, että molempien tutut voidaan ajatella satunnaisotoksiksi kaupunkilaisista. Mikä on todennäköisyys, että Antilla ja Annalla on ainakin yksi yhteinen tuttu?

Hahmotetaan tehtävä pallojen poimimis -analogian avulla. Pussissa on 500 liilaa (Antin tuttua) ja 999 500 mustaa palloa (muita kaupunkilaisia). Poimitaan pussista 500 palloa (Annan tutut). Todennäköisyys saada ainakin 1 liila pallo (yhteinen tuttu) on 1 miinus todennäköisyys, ettei saada yhtään liilaa palloa:

$$1 - \frac{\binom{500}{0}\binom{999\,500}{500}}{\binom{1\,000\,000}{500}} \approx 0.221.$$

Todennäköisyys Antin ja Annan yhteisille tutuille 0.221 on melko suuri.

[35] Tijms (2012, 136).

Kombinaatioiden lukumäärä osamäärässä on niin valtaisa, että R ei pysty suorittamaan komentoa `choose(500,0)*choose(999500,500)/choose(1000000,500)`. Todennäköisyys (0.2212965) on laskettu komennolla `1-dhyper(0,500,999500,500)`. □

4.2.6 Poisson-jakauma

Poisson-jakautuneen satunnaismuuttujan $Y \sim \mathsf{Poi}(\mu)$ pistetodennäköisyysfunktio on

$$\mathsf{P}(Y = i) = \begin{cases} e^{-\mu}\dfrac{\mu^i}{i!}, & i = 0,1,2,\ldots \\ 0, & \text{muulloin.} \end{cases} \tag{34}$$

Yllä $\mu > 0$ ja e on Neperin luku $2.71828\ldots$ Tällaisen satunnaismuuttujan odotusarvo ja varianssi ovat molemmat μ:

$$\mathsf{E}(Y) = \mathsf{V}(Y) = \mu \tag{35}$$

(esim. Blitzstein ja Hwang 2015, 161–162). Poisson-jakautunut satunnaismuuttuja saa kokonaislukuarvoja. $\mathsf{Poi}(\mu)$-jakauman moodi on $\mathsf{int}[\mu]$, jos μ ei ole kokonaisluku. Jos on, moodeja on kaksi: $\mu - 1$ ja μ. Edellä $\mathsf{int}[x]$ on argumentin x kokonaislukuosa (esim. $\mathsf{int}[0.5] = 0$).

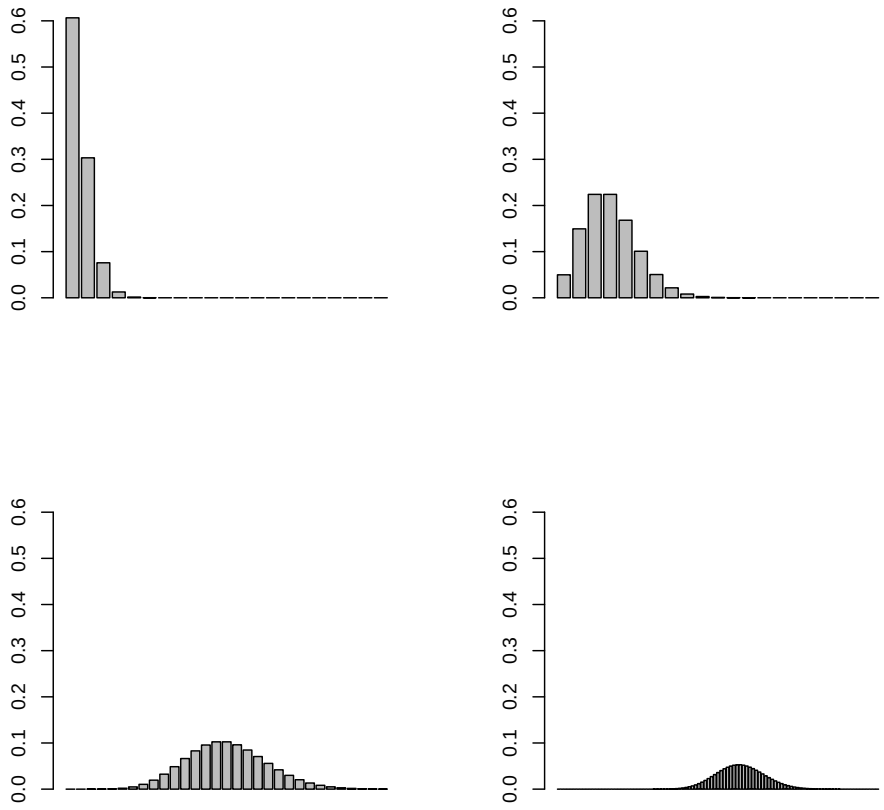
Jakauman varianssi kasvaa odotusarvon kanssa. Se on intuitiivista: Poisson-jakautunut satunnaismuuttuja ei voi saada nollaa pienempiä arvoja. Alarajalla on merkitystä, jos odotusarvo on pieni. Jos odotusarvo on suuri, satunnaismuuttujalla on enemmän tilaa vaihdella.

Jakauma on paljon käytetty, koska sillä voidaan approksimoida monia tilanteita. Taustalla on usein ajatus lukuisista kokeista pienellä todennäköisyydellä, niin että lopputuloksena havaitaan kokonaislukuarvoinen satunnaismuuttuja, jonka odotusarvo on μ. Pienen todennäköisyyden voi tulkita intensiteettinä, jolla tapahtumia tulee esimerkiksi aikayksikössä, ja tarkasteltavana on tapahtumien lukumäärä aikavälillä. Lukumäärä voi yksinkertaisimmillaan noudattaa binomijakaumaa, jolloin $\mu = n\pi$ (tästä lisää kurssivaatimuksiin kuulumattomassa jaksossa alla). Tapahtumilla voi kuitenkin olla eri todennäköisyydet ja ne voivat jossain määrin riippua toisistaan, ja silti Poisson-jakauma pätee. Tuloksen johto sivuutetaan matemaattisesti vaativana.

Jakauman soveltamista avittaa tulos riippumattomien Poisson-jakautuneiden satunnaismuuttujien summalle: Jos $Y_1 \sim \mathsf{Poi}(\mu_1)$ ja $Y_2 \sim \mathsf{Poi}(\mu_2)$ ja Y_1 ja Y_2 ovat riippumattomia, niin

$$Y_1 + Y_2 \sim \mathsf{Poi}(\mu_1 + \mu_2). \tag{36}$$

Kuva 15 havainnollistaa Poisson-jakaumaa μ:n arvoilla 0.5, 3, 15 ja 57. Jakauma on hyvin vino, kun odotusarvo on pieni, mutta symmetrisoituu odotusarvon suuretessa. Approksimaatio on varsin hyvä, jos $\mu > 20$. Moodi on 0, kun $\mu = 0.5$. Kokonaislukuarvoilla $\mu = 3$ sekä $\mu = 15$ jakaumissa näkyy molemmissa kaksi moodia (i:n arvoilla 2 ja 3 sekä 14 ja 15).



Kuva 15: Poisson-jakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita μ:n arvoilla 0.5, 3, 15 ja 57.

Esimerkki.[36] Helsingin Sanomat 12.4.2016: ”Vantaalaisperheen isä voitti jättipotin: perhe osti uuden asunnon ja auton – sitten päävoiton sai saman perheen äiti. ”En osaa edes hihkua ilosta, kun tämä on niin ihmeellistä — miten tämä on mahdollista”, nainen ihmettelee.” Iltalehti uutisoi 9.1.2015: ”Samalle miehelle jo toinen iso lottovoitto.” Uusi Suomi kertoi 8.11.2009 eteläafrikkalaisesta lottovoittajasta: ”Voitti jättipotin kahdesti — aikamoinen todennäköisyystemppu”. New York Times kirjoitti 14.2.1986: ”A Jersey lottery player defied odds of 1 in 17 trillion and won a second jackpot.”

Pohditaan, kuinka epätodennäköistä on, että joku voittaa kahdesti lotossa. Suomalaisessa lotossa arvotaan 7 numeroa väliltä 1–39. Päävoiton ”7 oikein” todennäköisyys on

$$\frac{1}{\binom{39}{7}} = \frac{1}{15\ 380\ 937}.$$

Jos lottoaa kahdesti, niin todennäköisyys voittaa molemmilla kerroilla päävoitto on

$$\frac{1}{15\ 380\ 937} \times \frac{1}{15\ 380\ 937} \approx 4.227021 \times 10^{-15} = 0.000000000000004227021.$$

Se ei kuitenkaan kuvaa todennäköisyyttä voittaa kahdesti lotossa ylipäänsä. New York Timesin toimittaja on ilmeisesti tehnyt yllä olevan tapaisen laskun.

Jos lottoaa viikottain viidellä eri lottorivillä, todennäköisyys voittaa kullakin viikolla on

$$\frac{5}{\binom{39}{7}} = \frac{5}{15\ 380\ 937}.$$

Jos tekee näin 20 vuoden ajan, tulee täyttäneeksi 5 lottoriviä $20 \times 52 = 1040$ kertaa. Koska voiton todennäköisyys on pieni, voidaan ajatella yksittäisen pelaajan voittojen lukumäärän noudattavan Poisson-jakaumaa parametrilla

$$\mu_0 = 1040 \times \frac{5}{15\ 380\ 937} = \frac{5\ 200}{15\ 380\ 937} \approx 0.0003380808.$$

Todennäköisyys, että näin pelaava henkilö voittaa ainakin kahdesti 20 vuoden aikana on edelleen hyvin pieni:

$$\begin{aligned} &1 - e^{-0.0003380808} \times \frac{0.0003380808^0}{0!} - e^{-0.0003380808} \times \frac{0.0003380808^1}{1!} \\ &\approx 0.00000005713645. \end{aligned}$$

[36] Tehtävä juontaa esimerkistä Tijmsin (2012, 114–115) kirjassa. Uutisten lähteet: `Http://www.hs.fi/kaupunki/a1460425765317`, `http://www.iltalehti.fi/uutiset/2015010918990956_uu.shtml`, `http://www.uusisuomi.fi/viihde/76355-voitti-jattipotin-kahdesti-\%E2\%80\%93-aikamoinen-todennakoisyys temppu` ja `http://www.nytimes.com/1986/02/14/nyregion/news-summary-friday-february-14-1986.html`. (Viittaukset 3.3. ja 12.4.2016).

Todennäköisyys on määritelty komplementtitapahtuman avulla vähentämällä 1:stä todennäköisyydet, että voittoja tulee 0 tai 1, ja on sovellettu pistetodennäköisyysfunktiota (34). Lasku on tehty R:n käskyllä

```
1-dpois(0,5200/15380937)-dpois(1,5200/15380937).
```

Määritellään nyt varsinaisen Poisson-jakautuneen tutkimuskohteen ainakin kahden lottovoiton saamisen 20 vuoden aikana parametri. Jos 2 000 000 suomalaista[37] toimii edellä kuvatulla tavalla 20 vuoden ajan, niin moninkertaisten lottovoittajien lukumäärä on Poisson-jakautunut parametrilla

$$\mu = 2\ 000\ 000 \times 0.00000005713645 = 0.1142729.$$

Moninkertaisten lottovoittajien todennäköisyys 20 vuoden aikana on

$$1 - e^{-0.1142729} \times \frac{-0.1142729^0}{0!} = 1 - e^{-0.1142729} \approx 0.1079855.$$

Ei ole tavatonta, että joku voittaa kaksi täyspottia lotossa Suomessa. □

4.3 Jatkuvia jakaumia

4.3.1 Normaalijakauma

Tärkein jatkuva-arvoisen satunnaismuuttujan (Y) jakauma on Normaalijakauma $\mathsf{N}(\mu,\sigma^2)$. Sen sijainnin ja muodon määräävät odotusarvo μ ja varianssi σ^2:

$$\mathsf{E}(Y) = \mu \quad \text{ja } \mathsf{V}(Y) = \sigma^2.$$

Normaalijakautuneen satunnaismuuttujan tiheysfunktio on

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

Jos $\mu = 0$ ja $\sigma^2 = 1$, jakaumaa kutsutaan Standardinormaalijakaumaksi $\mathsf{N}(0,1)$. Se on taulukoitu lukuisissa tilastotieteen oppikirjoissa. Lukemattomat tilastolliset tunnusluvut voidaan muokata noudattamaan Standardinormaalijakaumaa suurilla havaintomäärillä.

Normaalijakaumalla on miellyttävä ominaisuus, että lineaarikombinaatio riippumattomista normaalijakautuneista satunnaismuuttujista (Y_1 ja Y_2) on normaalijakautunut:

$$a_1 Y_1 + a_2 Y_2 \sim \mathsf{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2). \tag{37}$$

Tässä $a_1 \neq 0$ ja $a_2 \neq 0$ ovat vakioita.

Kohtaa täydennetään myöhemmin muilla jatkuvilla jakaumilla.

[TÄHÄN KAKSI-KOLME TYHJÄÄ SIVUA.]

[37] Suomalaisista täysikäisistä 40 % lottoaa viikottain (`http://www.aamulehti.fi/kotimaa/lottoa-pelattiin-arviolta-15-miljoonalla/`; Aamulehden uutiseen 13.2.2016 viitattu 6.3.2016).

4.4 Keskeinen raja-arvolause

Olkoot $X_1, X_2, \ldots, X_n$ toisistaan riippumattomia satunnaismuuttujia, joilla on odotusarvo $\mathsf{E}(X_i) = \mu$ ja varianssi $\mathsf{V}(X_i) = \sigma^2 > 0$. Muodostetaan keskiarvo $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Sen odotusarvo ja varianssi ovat μ ja σ^2/n:

$$\mathsf{E}(\bar{X}) = \mathsf{E}(n^{-1} \sum_{i=1}^{n} X_i) = n^{-1} \sum_{i=1}^{n} \mathsf{E}(X_i) = n^{-1} n\mu = \mu$$

ja

$$\mathsf{V}(\bar{X}) = \mathsf{V}[\sum_{i=1}^{n} (n^{-1} X_i)] = n^{-2} \mathsf{V}(\sum_{i=1}^{n} X_i) = n^{-2} \sum_{i=1}^{n} \mathsf{V}(X_i) = n^{-2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Standardoidun satunnaismuuttujan

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

odotusarvo on 0 ja varianssi 1:

$$\begin{aligned} \mathsf{E}(Z_n) &= \mathsf{E}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma/\sqrt{n}} \mathsf{E}(\bar{X} - \mu) = \frac{1}{\sigma/\sqrt{n}} [\mathsf{E}(\bar{X}) - \mu] = \frac{1}{\sigma/\sqrt{n}} (\mu - \mu) \\ &= 0 \end{aligned}$$

ja

$$\mathsf{V}(Z_n) = \mathsf{V}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma^2/n} \mathsf{V}(\bar{X} - \mu) = \frac{1}{\sigma^2/n} \mathsf{V}(\bar{X}) = \frac{\sigma^2/n}{\sigma^2/n} = 1.$$

Normaalijakauman suuri merkitys tilastotieteessä johtuu paljolti *Keskeisestä raja-arvolauseesta*. Sen mukaan X_i:den lukumäärän n kasvaessa kohti ääretöntä Z_n:n jakaumaksi muodostuu Standardinormaalijakauma $\mathsf{N}(0,1)$. Suurilla n:n arvoilla pätee siten approksimatiivisesti

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \quad \text{eli} \quad \bar{X} \sim \mathsf{N}(\mu, \sigma^2/n). \tag{38}$$

"Keskeinen" tarkoittaa tässä oleellista perustavaa laatua olevaa. Keskiarvon $\bar{X}$ normaalisuus pätee riippumatta satunnaismuuttujien X_i, joista keskiarvo lasketaan, jakaumasta.[38] Se on vaikuttava tulos ja selittää Normaalijakauman tärkeyttä tilastotieteessä.

Henri Poincaré kuvasi Keskeisen raja-arvolauseen merkitystä kieli poskella (Ross 2010, 311):

> Jokainen uskoo siihen: Empiirikot ajattelevat sen olevan matemaattinen välttämättömyys. Matemaatikot pitävät sitä empiirisenä tosiasiana.

[38] Lauseessa oletettiin, että satunnaismuuttujilla X_i on odotusarvo ja varianssi. Niin ei aina ole. Sellaisissa tilanteissa lause ei takaa keskiarvon normaalisuutta.

Keskeinen raja-arvolause kohoaa sellaisiin korkeuksiin Francis Galtonin mielessä, että käännös ei ehkä yltäisi samaan:[39]

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency Error" [Keskeinen raja-arvolause]. The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

4.5 Jakaumien kytkökset

4.5.1 Hypergeometrinen jakauma ja Binomijakauma

Hypergeometrisesti jakautuneen satunnaismuuttujan ja binomijakautuneen satunnaismuuttujan odotusarvo on sama. Muuttujien varianssit eroavat *äärellisen perusjoukon korjaustekijän*

$$\frac{l+m-n}{l+m-1} \leq 1$$

verran. *Otantasuhteen* $n/(l+m)$ supetessa kohti nollaa eli perusjoukon koon $l+m$ kasvaessa kohti ääretöntä korjaustekijä suppenee kohti yhtä. Varianssien ero häviää tällöin.

Karkea peukalosääntö on, että varianssien ja jakaumien erolla ei ole merkitystä, jos

$$\frac{n}{l+m} < 0.1.$$

Seber (2013,3) toteaa saman säännön mutta toteaa, että suhteen olisi suotavaa olla 0.05:ttä pienempi, jotta jakaumien ero olisi merkityksetön.

Esimerkki. Kortin peluu (jatkoa). Poimittaessa 5 korttia 52 kortin pakasta otantasuhde on $5/52 \approx 0.096 < 0.1$. Otantasuhde alittaa peukalosäännön ohjeen juuri ja juuri. Esimerkeissä edellä binomijakaumasta ja hypergeometrisesta jakaumasta lasketut pistetodennäköisyydet kuninkaiden saamisen todennäköisyyksille eivät eronneet paljoa toisistaan. Toisaalta suurilla panoksilla päivittäin pelaavalle pokeriammattilaiselle pienilläkin todennäköisyyseroilla on merkitystä. Hänen on syytä laskea pistetodennäköisyydet hypergeometrisesta jakaumasta. □

4.5.2 Binomijakauma ja Poisson-jakauma

Binomijakaumaa voidaan approksimoida Poisson-jakaumalla, kun toistojen lukumäärä n on suuri ja todennäköisyys π on pieni:

$$\mathsf{P}(Y=y) = \binom{n}{y}\pi^k(1-\pi)^{n-y} \approx e^{-n\pi}\frac{(n\pi)^y}{y!}.$$

Approksimaation avulla binomitodennäköisyys voidaan laskea, vaikka n tai π eivät olisi tiedossa, mutta Poisson-jakauman parametri $\mu = n\pi$ yllä tunnetaan tai voidaan arvioida.

[39] Galton (1899, 66).

Peukalosääntöjä[40] approksimaation käyttökelpoisuudelle:

- $\pi \leq 0.1$ ja $n \geq 40$
- $n\pi < 5$ ja $n > 50$
- $\pi \leq 0.05$ ja $n \geq 20$ tai
- $n\pi < 10$ ja $n \geq 100$ (approksimaation pitäisi olla ”erinomainen”).

Esimerkkejä approksimaation toimivuudesta löytyy monista oppikirjoista.[41] Taulukko alla on Tijmisin (2012, 112) kirjasta. Vasemmanpuolimmaisessa sarakkeessa on tapahtumien lukumäärä (y). Siitä oikealla on lukumäärään liittyvä pistetodennäköisyys binomijakaumasta laskettuna, kun π toteuttaa yhtälön $n\pi = 1$. Oikeanpuolimmaisissa sarakkeessa on $\mathsf{Poi}(1)$-jakaumasta laskettu pistetodenäköisyys lukumäärälle. Esimerkiksi jos $n = 100$ ja $\pi = 0.01$, niin 0:lle tapahtumalle Binomijakauman pistetodennäköisyys on 0.3660 ja Poisson-jakauman 0.3679. Jakaumien pistetodennäköisyydet ovat varsin samanlaisia jo, kun $n = 25$. Jos Poisson-jakauman μ-parametri olisi suurempi, yhteensopivuus ei välttämättä olisi näin hyvä (esim. Armitage ym. 2002, 76).

	Binomitodennäköisyys ($n\pi = 1$)				$\mathsf{Poi}(1)$
y	$n = 25$	$n = 100$	$n = 500$	$n = 1000$	
0	0.3604	0.3660	0.3675	0.3677	0.3679
1	0.3754	0.3697	0.3682	0.3681	0.3679
2	0.1877	0.1849	0.1841	0.1840	0.1839
3	0.0600	0.0610	0.0613	0.0613	0.0613
4	0.0137	0.0149	0.0153	0.0153	0.0153
5	0.0024	0.0029	0.0030	0.0030	0.0031

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Noudattakoon satunnaismuuttuja Y Binomijakaumaa $\mathsf{Bin}(n,\pi)$:

$$\mathsf{P}(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y},$$

jossa $0 < \pi < 1$.

Voidaan osoittaa, että

$$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^{-n} \approx 2.71828.$$

Yhtäsuuruuksista seuraa, että

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = \lim_{n\to\infty}\left(1 + \frac{1}{n/a}\right)^n$$

[40] Ramachandran ja Tsokos (2015, 119) sekä Engineering Statistics Handbook (`http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc331.htm` (viitattu 6.3.2015).

[41] Armitage, Berry ja Matthews (2002, 76), Larsen ja Marx (2001, 248), Lindgren (1976, 185) ja Ross (2010, 252). Lindgrenin ja Rossin esimerkeissä approksimaatio toimii hyvin jo, kun π on 0.1–0.125 ja n on 8–10.

$$= \lim_{n\to\infty} \left[\left(1 + \frac{1}{n/a}\right)^{n/a} \right]^a$$
$$= e^a.$$

Yllä a on vakio (jos $a = 0$, niin $\lim_{n\to\infty}(1 + 0/n)^n = 1 = e^0$).

Merkitään $\mu \equiv n\pi$. Ilmaistaan binomitodennäköisyys μ:n avulla ja annetaan havaintojen lukumäärän n kasvaa kohti ääretöntä siten, että tulo $\mu = n\pi$ on vakio (todennäköisyys π lähenee nollaa samalla vauhdilla kuin n kasvaa kohti ääretöntä):

$$\begin{aligned}
\lim_{\substack{n\to\infty,\pi\to 0,\\ n\pi=\mu}} \binom{n}{y}\pi^y(1-\pi)^{n-y} &= \lim_{\substack{n\to\infty,\pi\to 0,\\ n\pi=\mu}} \binom{n}{y}\left(\frac{n\pi}{n}\right)^y\left(1-\frac{n\pi}{n}\right)^{n-y} \\
&= \lim_{n\to\infty} \binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y} \\
&= \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n-y}\frac{\mu^y}{n^y}\frac{n!}{y!(n-y)!} \\
&= \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n-y}\frac{\mu^y}{y!}\frac{n(n-1)\times\cdots\times(n-y-1)}{n^y} \\
&= e^{-\mu}\frac{\mu^y}{y!}.
\end{aligned}$$

Tulos on Poisson-jakauma.

Esimerkki.[42] Suuren lottovoiton todennäköisyys määräytyy hypergeometrisesta jakaumasta. Lottoajia on paljon, kunkin mahdollisuus voittaa suuri lottovoitto on hyvin pieni ja riippumaton toisista lottoajista. Suurten lottovoittojen lukumäärä on siten binomijakautunut. Jollei täytettyjen lottorivien lukumäärä ole tiedossa, ei Binomijakauma ole käytettävissä. Kun Binomijakauman taustalla olevien Bernoulli-kokeiden lukumäärä on suuri ja kussakin kokeessa tapahtuman (tässä suuren lottovoiton) todennäköisyys on pieni, voidaan binomijakaumaa approksimoida Poisson-jakaumalla. Approksimaation käyttöön ei tarvita tietoa Bernoulli-kokeiden lukumäärästä.

Ontariolaisessa pikkukaupungissa asuva Bob Edmonds lottosi aina ainakin yhdellä samalla lottorivillä. Hän tapasi antaa lottokupongin myyjän tarkistaa, ovatko hänen lottorivinsä voittaneet. Edmonds osti kaksi lottoriviä 27.7.2001. Toinen lottoriveistä voitti 250 000 dollaria. Kupongin myyjä ei kertonut sitä Edmondsille vaan lunasti voiton itse. Edmonds ymmärsi myöhemmin mitä oli tapahtunut ja haastoi Ontarion veikkausyhtiön oikeuteen. Kolmen ja puolen vuoden oikeustaistelun jälkeen yhtiö suostui maksamaan Edmondsille 200 000 dollaria ehdolla, että hän ei paljasta julkisuuteen, mitä oli tapahtunut. Tällaisista väärinkäytöksistä nousi sittemmin kohu. Julkisuuden paineessa veikkausyhtiö maksoi Edmondsille koko voittosumman ja pyysi anteeksi tapahtunutta — vain päiviä ennen kuin Edmonds kuoli syöpään 2.4.2007.

Kanadan kansallinen TV-yhtiö CBC pyysi tilastotieteen professori Jeffrey Rosenthalia tutkimaan, onko lottovoittojen jakamisessa tapahtunut vastaavia huijauksia, joissa lottomyyjät olisivat lunastaneet itselleen asiakkaittensa voittoja. TV-yhtiön selvitysten mukaan Ontarion alueella oli vuosina 1999–2006 jaettu

[42] J.S. Rosenthal (2014): Statistics and the Ontario Lottery Retailer Scandal. *Chance*, 27, 4–9.

5713 suurta (yli 50 000 dollarin) voittoa. Niiden lunastajista (vähintään) 200 (3.5 %) oli lottokuponkien vähittäismyyjiä. Heitä oli Ontariossa noin 60 000 ja he lottosivat keskimäärin puolitoistakertaisella summalla keskiverto-ontariolaiseen verrattuna. Koska Ontariossa oli noin 8 900 000 täysi-ikäistä asukasta, Rosenthal arvio, että loton vähittäismyyjien suurten lottovoittojen lukumäärän voisi odottaa olevan noin 57:

$$5713 \times \frac{60000 \times 1.5}{8900000} \approx 57.$$

Se on luonteva estimaatti Poisson-jakauman $\mathsf{Poi}(\mu)$ odotusarvolle eli Poisson-jakauman määrittävälle parametrille μ. Rosenthal päätteli, että approksimatiivisesti $Y \sim \mathsf{Poi}(57)$, jossa satunnaismuuttuja Y on vähittäismyyjien saamien suurten lottovoittojen lukumäärä. Tällöin

$$P(Y \geq 200) = 1 - P(Y < 200) = 1 - \sum_{i=0}^{199} \frac{57^i e^{57}}{i!}.$$

R:n käsky

```
1-ppois(199,57)
```

laskee 0:ksi erotuksen 1:n ja Poisson-jakauman $\mathsf{Poi}(57)$ kertymäfunktion arvon pisteessä 199 välillä. Rosenthal kertoo artikkelissaan, että todennäköisyys on alle yhden suhde triljoonaan triljoonaan triljoonaan triljoonaan. Tehdyillä oletuksilla todennäköisyys, että loton vähittäismyyjistä vähintään 200 voittaa suuren lottovoiton, on oleellisesti nolla.

Rosenthal teki muitakin vastaavia analyysejä. CBC:n väärinkäytöksistä kertova dokumenttiohjelma lähetettiin 25.10.2006, ja ne olivat pääuutisia seuraavina päivinä televisiossa ja lehdistössä Kanadassa. Ontarion veikkausyhtiö yritti kiistää väärinkäytökset mutta joutui myöntämään väärinkäytökset ja muuttamaan käytäntöjään: Loton vähittäismyyjien kuponkien tarkastuskoneiden tulee nykyään olla asiakkaiden nähtävissä ja niiden pitää hälyttää voitoista äänimerkillä, asiakkaiden täytyy allekirjoittaa lottokuponkinsa eivätkä loton vähittäismyyjät saa enää ostaa lottokuponkeja omasta myymälästään. Yhtiön toimitusjohtaja erotettiin. Vastaavat tutkimukset käynnistettiin Brittiläisessä Kolumbiassa. Myös siellä havaittiin väärinkäytöksiä, ja Brittiläisen Kolumbian veikkausyhtiön toimitusjohtaja erotettiin. Tutkimukset levisivät muualle Kanadaan ja Yhdysvaltoihin. Suurin paljastunut huijaus oli 12 500 000 dollarin voiton väärä lunastus. Rosenthalin analyysit johtivat toimitusjohtajien erottamisen lisäksi useisiin vankilatuomioihin ja miljoonien dollarien maksatukseen ihmisille, joilta oli huijattu voitto.

Vastaavia huijauksia on paljastunut myöhemmin lisää. Iltalehti 23.12.2015[43]:

> 37:ssä osavaltiossa toimivan rahapeliyhdistyksen entisen turvallisuuspäällikön, Eddie Tiptonin, epäillään saaneen haittaohjelman avulla tietoonsa koneen arpomat numerot etukäteen ja lunastaneen petoksilla itselleen miljoonavoittoja.

[43] `Http://www.iltalehti.fi/digi/2015122320869989_du.shtml` (viitattu 26.12.2015).

> Tipton on jo tuomittu lottonumeroiden avulla tehdystä petoksesta kymmeneksi vuodeksi vankeuteen Iowan osavaltiossa. Nyt hänen epäillään tehtailleen vastaavia rikoksia veljensä ja ystävänsä kanssa myös kolmessa muussa osavaltiossa viimeisen kuuden vuoden aikana. Kolmikon epäillään huijanneen lottovoittoja Coloradon, Wisconsin ja Oklahoman osavaltioissa kahdeksan miljoonan dollarin edestä. – – On mahdollista, että huijauksia paljastuu vielä lisää.

Asiaa sivuava suomalainen uutinen Helsingin Sanomissa 3.4.2016:[44]

> Veikkaus-voitto maksettiin väärän henkilön tilille — vastaavia tapauksia kymmeniä vuodessa. Veikkaus vakuuttaa, että rahat päätyvät lopulta aina oikealle voittajalle. Veikkaus maksaa toistuvasti pelivoittoja väärän henkilön pankkitilille myyntipisteen työntekijän huolimattomuuden vuoksi. – – Veikkaukselle tilanne on tuttu. ”Tässä on todennäköisesti kyse tilanteesta, jossa myyjä ei ole vastoin ohjeita nollannut myyntipäätettä sen jälkeen, kun edellinen asiakas on käyttänyt Veikkaus-korttia. – – Oikea henkilö on aina saanut voiton itselleen sen jälkeen, kun asia on selvitetty.”

□

4.5.3 Binomijakauma ja Normaalijakauma

Binomijakautunut satunnaismuuttuja Y on summa Bernoulli-jakautuneista satunnaismuuttujista X_i: $Y = \sum_{i=1}^{n} X_i$. Muodostetaan keskiarvo $\bar{X} = \sum_{i=1}^{n} X_i/n$. Keskeisen raja-arvolauseen (38) ja Bernoulli-jakautuneen satunnaismuuttujan varianssin (26) perusteella suurilla n:n arvoilla

$$\bar{Y} \sim \mathsf{N}(\pi,\pi(1-\pi)/n).$$

Koska $\bar{Y}$ on vakiolla (n) skaalattu binomijakautunut Y, on Binomijakaumankin muoto suurilla havaintomäärillä Normaalijakauman

$$\mathsf{N}(n\pi,n\pi(1-\pi))$$

tapainen. Kuva 16 havainnollistaa, kuinka Binomijakauma ($\pi = 0.1$) muotoutuu Normaalijakaumaksi n:n kasvaessa.

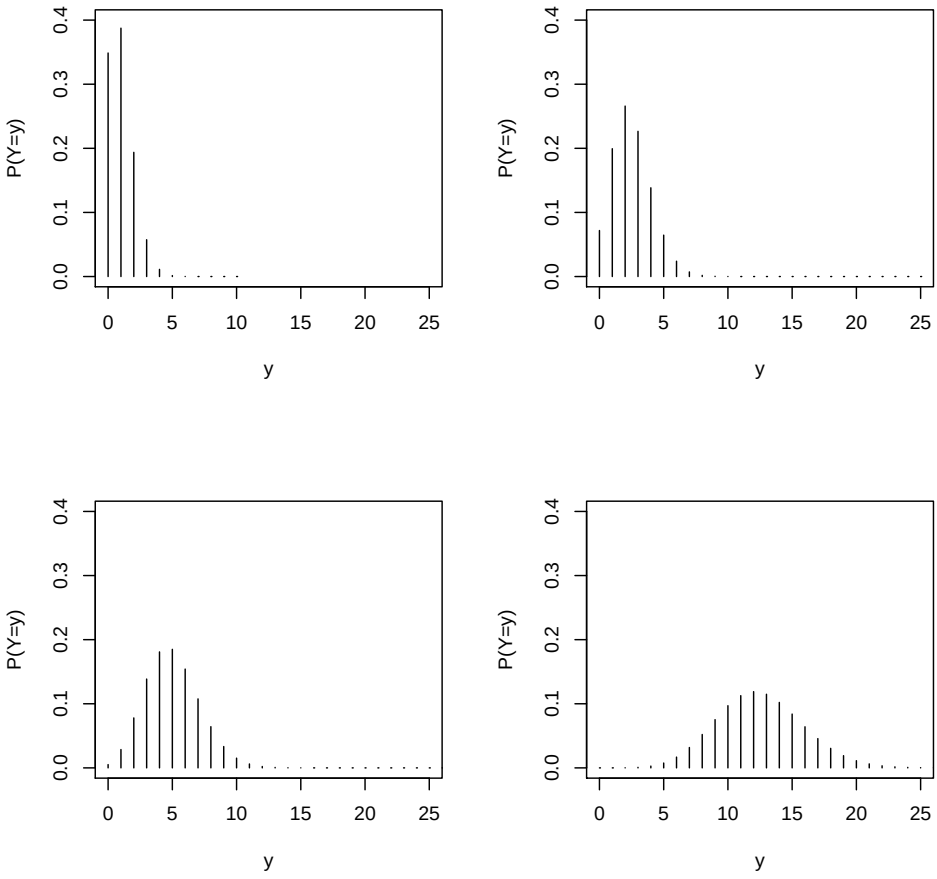
Approksimaation toimivuudelle on esitetty peukalosääntöjä kuten

- $\sqrt{n\pi(1-\pi)} > 3$ tai
- $n\pi > 5$ ja $n(1-\pi) > 5$.

Säännön toteutuminen ei takaa, että approksimaatio olisi riittävän hyvä. Riittävyys riippuu käyttötarkoituksesta.

Esimerkki. Huoltoriidat käräjäoikeuksissa Suomessa (jatkoa). Lapsista 35 osoitettiin asumaan isän ja 83 äidin luona (118 havaintoa). Edellä pohdittiin todennäköisyyttä tällaiselle tai pienemmälle isien voittojen lukumäärälle, jos isän ja äidin voittotodennäköisyys on sama. $\mathsf{Bin}(118,0.5)$-jakauman kertymäfunktion

[44] `Http://www.hs.fi/kotimaa/a1459564475529?ref=hs-prio1415-3` (viitattu 3.4.2016).



Kuva 16: Binomijakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita n:n arvoilla 10, 25, 50 ja 125, kun $\pi = 0.1$.

arvoksi pisteessä 35 laskettiin noin 0.00006. Lasketaan approksimaatio tälle todennäköisyydelle Normaalijakauma-approksimaatiolla.

Merkitään Y:llä isien luokse asumaan osoitettavien lasten lukumäärää. Tehtävän tilanteessa $\mathsf{E}(Y) = n\pi = 118 \times 0.5 = 59$ ja $\mathsf{V}(Y) = n\pi(1-\pi) = 118 \times 0.5 \times 0.5 = 29.5$. Normaalijakauma-approksimaation mukaan todennäköisyys, että satunnaisessa 118:n havainnon aineistossa lapset osoitetaan isälle asumaan päätöksistä 35:ssä tai pienemmässä lukumäärässä on

$$\begin{aligned}
\mathsf{P}(Y \leq 35) &= \mathsf{P}\left(\frac{Y-59}{\sqrt{29.5}} \leq \frac{35-59}{\sqrt{29.5}}\right) \approx \mathsf{P}(Z \leq \frac{35-59}{\sqrt{29.5}}) \\
&= \Phi\left(\frac{35-59}{\sqrt{29.5}}\right) = \Phi\left(-4.418758\right) \\
&\approx 0.000005.
\end{aligned}$$

Yllä Z on standardinormaalisti jakautunut satunnaismuuttuja ja $\Phi(z)$ on Standardinormaalijakauman kertymäfunktion arvo pisteessä z. Satunnaismuuttujan arvo -4.419 on niin pieni, että siihen liittyvää kertymäfunktion arvoa ei löytyne useimmista taulukoista. Kertymäfunktion arvo on laskettu R-ohjelmiston käskyllä

```
pnorm(-4.419)
```

Todennäköisyys on approksimaation mukaan noin 0.00005. Se on lähes sama kuin todellinen todennäköisyys 0.00006.

Todennäköisyys voitaisiin laskea myös suhteellisesta osuudesta lasketun standardoidun suureen

$$\frac{\bar{y}-\pi}{\sqrt{\pi(1-\pi)/n}} = \frac{0.2966102 - 0.5}{\sqrt{0.5 \times 0.5/118}} = -4.418758$$

avulla. Yllä $\bar{y} = 35/118 = 0.2966102 = \hat{\pi}$ on havaittu suhteellinen osuus. □

4.5.4 Galtonin kone

Francis Galton (1877) kuvasi sittemmin Galtonin koneena (*Galton's machine*, *Galton board*, *Galton box*, *Quincunx* tai *bean machine*) tunnetun mekaanisen laitteen. Sillä voidaan konkreettisella tavalla havainnollistaa keskeistä rajaarvolausetta sekä Binomi- ja Normaalijakauman yhteyttä. Alla on Galtonin luonnos koneesta (1889, 63), kuva Galtonille 1873 tehdystä koneesta ja kuva uudemmasta suuresta Galtonin koneesta.[45]

[45] Ensimmäinen kuva on Stephen M. Stiglerin ottama. Kaksi ensimmäistä kuvaa ovat artikkelista Burnett, D.G. (2009): Games of Chance. *Cabinet*. `Http://www.cabinetmagazine.org/issues/34/burnett.php` (viitattu 5.2.2014). Galtonin koneen R-kielisen emulaattorin ovat tehneet Y. Xie, L. Yu ja K. O'Rourke (2013): Demonstration of the Quincunx (Bean Machine/Galton Box). `Http://www.rforge.net/doc/packages/animation/quincunx.html` (viitattu 10.2.2015). Koneen "parannettu" versio on patentoitu Yhdysvalloissa 1990 (`http://www.google.nl/patents/US4900255`; viitattu 4.3.2016).

Suppilosta tippuu kuulia, jotka päätyvät koneen alaosassa oleviin numeroituihin ("0","1",...,"n") laareihin ($n+1$ kappaletta). Välissä on pyramidin muodossa n riviä pinnejä (*i.* rivillä i pinniä, $i = 1,\ldots,n$) niin, että 1. pinni on suppilon suun keskipisteen alapuolella.

Kuula, joka on päätynyt laariin "y" ($k = 0,\ldots,n$), on pompannut y kertaa oikealle todennäköisyydellä π ja $n-y$ kertaa vasemmalle todennäköisyydellä $1-\pi$. Kunkin laariin "y" johtavan polun todennäköisyys on $\pi^y(1-\pi)^{n-y}$.

Jos $\pi = 1-\pi = 0.5$ — kuten alla oletetaan — niin jokaisen polun mihin tahansa laariin todennäköisyys on $(0.5)^n$. Keskimmäisiin laareihin päätyy kuitenkin enemmän kuulia kuin laitimmaisiin, koska edellisiin johtaa useampia polkuja mutta laitimmaisiin vain yksi.

Kuinka monta polkua johtaa laariin "y"? Tarkastellaan oransseja ja vihreitä alkioita ("*o*" ja "*v*"), joita on y ja $n-y$ kappaletta (yhteensä n alkiota). Tuloksen (21) mukaan ne voidaan järjestää binomikertoimen $\binom{n}{y}$ mukaiseen määrään erilaisia n alkiota sisältävään jonoon. Korvataan *o*- ja *v*-alkiot edellä pompuilla oikealle ja vasemmalle. Polku laariin "y" koostuu y pompusta oikealle ja $n-y$ pompusta vasemmalle. Päättelyn edellä mukaan tällaisia järjestyksiä eli polkuja laariin "y" on $\binom{n}{y}$ kappaletta.

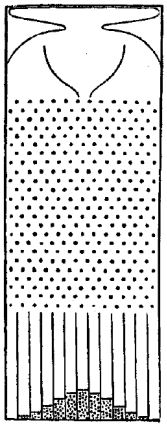
Polut ovat erillisiä, ja jokainen polku on yhtä todennäköinen. Todennäköisyys, että kuula päätyy laariin "y", saadaan summaamalla kaikkien laariin "y" johtavien polkujen todennäköisyys eli kertomalla yhden polun todennäköisyys polkujen lukumäärällä $\binom{n}{y}$ ($y = 0,\ldots,n$):

$$\begin{aligned}\mathsf{P}(\text{kuula päätyy laariin "}y\text{"}) &= \pi^y(1-\pi)^{n-y} + \ldots + \pi^y(1-\pi)^{n-y} \\ &= \binom{n}{y}\pi^y(1-\pi)^{n-y}.\end{aligned}$$

Kuulan päätyminen laariin "y" noudattaa binomijakaumaa. Lisäämällä luonteva oletus $\pi = 0.5$ todennäköisyydeksi saadaan

$$\begin{aligned}\mathsf{P}(\text{kuula päätyy laariin "}y\text{"}) &= \binom{n}{y}(0.5)^y(1-0.5)^{n-y} \\ &= \binom{n}{y}(0.5)^n.\end{aligned}$$

Galtonin kone osoittaa konkreettisesti, kuinka kuulien lukumäärän kasvaessa laareihin muodostuu normaalijakauman kuvio, vaikka jakauma on binomijakauma. Normaalijakauma approksimoi binomijakaumaa. Rakennetut Galtonin koneet havainnollistavat tilannetta $\pi = 0.5$ (kuula pomppaa samalla todennäköisyydellä vasemmalle tai oikealle), mutta approksimaatio toimii muillakin π:n arvoilla. Proschan ja Shaw (2016, 156–159) kuvaavat monimutkaisempia Galtonin koneita.



Kuva 17: Galtonin (1889, 63) luonnos koneestaan.

Kuva 18: Galtonille 1873 tehty ensimmäinen Galtonin kone.

Kuva 19: Suuri Galtonin kone.

4.5.5 Poisson-jakauma ja Normaalijakauma

Koska Binomijakaumaa voi approksimoida Poisson- tai Normaalijakaumalla, ei ole ehkä yllättävää, että Poisson-jakauman ja Normaalijakaumankin välillä on kytkös. Tuloksen (36) mukaan $\mathsf{Poi}(n)$-jakautunut satunnaismuuttuja Y voidaan ilmaista n:n $\mathsf{Poi}(1)$-satunnaismuuttujan Y_i summana: $Y = \sum_{i=1}^{n} Y_i$. Muodostetaan keskiarvo $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Summattavien odotusarvo ja varianssi ovat 1. Keskeisestä raja-arvolauseesta (38) seuraa, että approksimatiivisesti pätee

$$\bar{Y} \sim \mathsf{N}(1,1/n) \quad \text{eli} \quad Y \sim \mathsf{N}(n,n).$$

Tuloksen mukaan Poisson-jakauman tulisi normalisoitua Poisson-jakauman odotusarvon kasvaessa. Kuvan 15 mukaan niin näyttää käyvän.

5 Otantateoriaa

5.1 Käsitteitä

Aineisto on tilastotieteen keskeisin käsite (jakso 1.2), ja tilastotieteellisestä näkökulmasta aineistossa keskeisintä on, miten se on muodostettu. Menettelyä, jolla aineisto kootaan, kutsutaan *otannaksi*. Otannalla saatua aineistoa kutsutaan *otokseksi*. Otos on osajoukko perusjoukosta eli populaatiosta, josta se on kerätty.

Populaatio on varsinainen kiinnostuksen kohde. Sen tutkiminen kokonaisuudessaan voi olla vaikeaa tai mahdotonta vaikkapa sen suuren koon takia. Populaatio voi olla konkreettinen tai abstrakti. Jos populaatio on pieni ja tutkittavissa kokonaisuudessaan, ei otantaa tarvita.

Esimerkki. Suuri populaatio. Tieto suomalaisten hyvinvoinnista on terveyspoliittisesti tärkeää. Kaikkien verenpainetta, kolesteroliarvoja jne. ei voida mitata. Niiden tasot kansalaisten keskuudessa selviävät riittävällä tarkkuudella tutkimalla ne otoksesta suomalaisia. Otoksen kooksi riittää murto-osa suomalaisia. □

Esimerkki. Pieni populaatio. Kiinnostava populaatio voi olla hyvin tarkkaan rajattu ryhmä kuten tilastotieteen kokeeseen osallistuneet. Jos kiinnostuksen kohde on kokeeseen osallistuneiden sukupuolijakauma, se on helposti selvitettävissä. Otantaa ei tarvita. Jos kiinnostuksen kohde on heidän älykkyysosamääränsä tai keskipituutensa, otanta saattaa olla tarpeen. Kaikkia ei varmaankaan saataisi mittaukseen mutta otos kokeessa olijoista ehkä saataisiin. □

Otoksesta pyritään tekemään päätelmiä populaatiosta. Jotta päätelmät olisivat tilastotieteellisesti perusteltuja ja niiden luotettavuus mitattavissa, tulee otanta olla suoritettu tarkoituksenmukaisella tavalla ja olla kohdistettavissa tarkoitettuun kohdepopulaatioon (*target population*).

Esimerkki. Abstrakti populaatio I. Urheiluhullut, rakkausrunojen kirjoittajat tai musikaalisesti lahjakkaat voi olla kiinnostava populaatio. Tällaisia populaatioita voi olla vaikea määrittää, eikä otantaa voi kohdistaa niihin yksiselitteisesti. Tällöin otannalla ei välttämättä saada toivottua tietoa. □

Kiinnostus kohdistuu populaation numeerisesti mitattavissa olevaan piirteeseen tai piirteisiin, joita kuvaa yksi tai useampi *parametri*. Populaatiota kuvaavan jakauman odotusarvo (usein "μ") ja varianssi (usein "σ^2") ovat esimerkkejä parametreista.

Koostukoon otos satunnaismuuttujista $X_1, \ldots, X_n$. Niistä voidaan laskea *tunnusluku* (*statistic*)

$$t(X_1, \ldots, X_n).$$

Tunnusluku on funktio otoksesta. Tyypillisesti tunnusluvulla pyritään arvioimaan eli estimoimaan populaation parametria. Tunnusluvun sijaan on monesti luontevampaa puhua parametrin *estimaattorista* ja sen tietyssä otoksessa saamasta numeerista arvosta *estimaatista*

$$t(x_1, \ldots, x_n).$$

Yllä x_i:t ovat X_i:den reaalisaatioita.

Monesti puhutaan *piste-estimaattorista* tai *piste-estimaatista* (luku 6). Tällöin korostetaan, että yhdellä luvulla pyritään arvioimaan populaation parametria eikä yritetä arvioida, millä välillä parametri on. Tällaiseen *väliestimointiin* palataan luvussa 7).

Parametrin ja sen estimaattorin ero on keskeinen. Edellinen on kiintää luku, joka kuvaa populaatiota eli maailmaa, jota tutkitaan. Jälkimmäinen on satunnaismuuttuja, jonka realisaatio (estimaatti) havaitaan. Estimaatti on havaittu numeerinen arvio maailmasta.

Populaatiota kuvaavia parametreja tavataan merkitä kreikkalaisilla kirjaimilla ja niiden estimaattoreita sijoittamalla "$\wedge$" ("hattu") niiden päälle. Merkitään estimoitavaa parametria θ:lla ("theeta") ja sen estimaattoria $\hat{\theta}$:lla ("theeta hattu").[46] Estimaatti saa harvoin saman arvon kuin estimoitava parametri. Niiden erotus on *otantavirhe*

$$\hat{\theta} - \theta.$$

Sen suuruuden ja ominaisuuksien tutkiminen on tärkeää tilastotieteessä.

Estimaattoria ja estimaattia merkitään monesti samalla symbolilla ($\hat{\pi}$ esimerkissä alla). Silti on tärkeää hahmottaa niiden ero ($\bar{X}$ ja $\bar{x}$ esimerkissä alla).

Esimerkki. Gallup (jatkoa). Demokratia edellyttää äänestysikäisten suomalaisten poliittisten kantojen selvittämistä. Ne selvitetään vaaleilla. Vaaleja järjestetään harvoin, koska ne ovat suuritöisiä ja kalliita. Vaalien välissä poliittisia kantoja selvitetään vaivattomammin ja pienemmillä kustannuksilla gallupeilla, joissa kysytään äänestysikäisiltä suomalaisilta heidän puoluekantaansa. Kiinnostuksen kohteena oleva populaatio on äänestysikäiset suomalaiset.

Otos on haastatellut suomalaiset. Tutkittava parametri on $\pi \in [0,1]$ eli tiettyä puoluetta kannattavien suomalaisten osuus. Kukin otokseen tuleva havainto

[46] Kreikkalainen kirjaimisto selitetään Wikipediassa: `https://fi.wikipedia.org/wiki/Kreikkalainen_kirjaimisto` (viitattu 17.3.2016).

X_i on Bernoulli-jakautunut satunnaismuuttuja, joka voi saada arvon 1 (kannattaa puoluetta) tai 0 (ei kannata puoluetta). Otoksesta voidaan laskea kannatusosuuden estimaattori keskiarvo $\hat{\pi} = \bar{X} = \sum_{i=1}^{n} X_i/n$. Kun otos on kerätty ja saatu havainnot x_i, voidaan laskea osuuden estimaatti $\hat{\pi} = \bar{x} = \sum_{i=1}^{n} x_i/n$.

Jos populaatiossa K-puolueen kannatus on 33.33 % mutta otoksessa 32 %, otantavirhe on 1.33 %-yksikköä. □

Esimerkki. Abstrakti populaatio II. Populaatioksi voidaan ajatella ääretön määrä silmälukuja kuvitteellisista nopan heitoista. Abstraktius ei ole ongelma otannan kannalta. Äärellinen määrä riippumattomia heittoja kyseisellä nopalla ovat ikään kuin otos äärettömän suuresta silmälukujen populaatiosta. Otos kohdentuu tarkasti populaatioon, ja sillä voidaan selvittää populaation ominaisuuksia, kuten ovatko silmäluvut yhtä todennäköisiä kyseisellä nopalla. Parametreja ovat silmälukujen todennäköisyydet. Niitä voidaan estimoida heittämällä noppaa monta kertaa ja laskemalla kunkin silmäluvun osuus heitoista. □

5.2 Todennäköisyysotanta ja satunnaisotanta

Otannan pitää olla *todennäköisyysotantaa* (*probability sampling*), jotta otoksesta voidaan tehdä tilastotieteellisiä päätelmiä. Todennäköisyysotannassa populaation kunkin alkion todennäköisyys tulla otokseen eli *sisältymistodennäköisyys* (*inclusion probability*) tunnetaan ja se on positiivinen (nollaa suurempi) kaikille alkioille. Tällöin otantavirheen suuruutta voidaan arvioida tilastotieteellisesti.

Todennäköisyysotanta on yksinkertaisimmillaan *satunnaisotantaa* (*random sampling*) eli *yksinkertaista satunnaisotantaa* (*simple random sampling*). Siinä kaikilla populaation alkioilla on sama sisältymistodennäköisyys:

$$\mathsf{P}(\text{populaation tietty alkio tulee otokseen}) = \frac{n}{N}.$$

Yllä n on otoksen ja N on populaation koko.

Esimerkki. Ositettu otanta I. Ositetussa otannassa populaatio jaetaan homogeenisiin ryhmiin, joista kustakin poimitaan otos satunnaisotannalla. Näin voidaan taata, että otoksessa ovat kaikki ryhmät halutussa suhteessa edustettuina. Haluttu suhde voi olla ryhmien osuudet populaatiossa, jolloin otos edustaa mahdollisimman hyvin populaatiota ryhmäkoostumukseltaan. □

Esimerkki. Ositettu otanta II. Ositetulla otannalla voidaan hakea populaation osajoukoista tietoisesti epäsuhtainen määrä havaintoja. Pienen osajoukon alkioille asetetaan muita suurempi todennäköisyys tulla otokseen. Näin siihen saadaan pieneen osajoukkoon kuuluvia riittävästi osajoukkojen vertailua varten. Satunnaisotannalla koko populaatiosta ei välttämättä saataisi. □

Jakaumia opiskeltaessa tutustuttiin kahteen otantamenetelmään: Otanta takaisinpanolla (Binomijakauma) ja ilman (Hypergeometrinen jakauma). Molemmissa sovellettiin satunnaisotantaa. Otanta ilman takaisinpanoa on parempi otantamenetelmä kuin otanta takaisinpanolla. Jälkimmäisessä populaation alkio voi tulla useampaan kertaan eli yliedustetuksi otokseen.

Esimerkki. Kortin peluu (jatkoa). Yritetään päätellä kuninkaiden osuus korttipakassa poimimalla 13 korttia pakasta ilman takaisinpanoa ja takaisinpanolla. Jälkimmäisellä menetelmällä on pieni todennäköisyys saada otos, jossa on $5-13$ kuningasta. Sellainen otos johtaisi täysin väärään päätelmään kuninkaiden osuudesta pakassa. Otannassa ilman takaisinpanoa ei tätä virhemahdollisuutta ole. □

Monesti populaatio on siinä määrin suuri, ettei ole väliä, käytetäänkö otantaa takaisinpanolla vai ilman (jakso 4.5.1). Todennäköisyyslaskut voidaan tällöin tehdä matemaattisesti yksinkertaisemmalla tavalla olettaen otanta takaisinpanolla vaikkei sitä oltaisi käytetty.

Satunnaisotanta on yleispätevä otantamenetelmä. Jatkossa otos oletetaan kerätyksi satunnaisotannalla — paitsi seuraavassa jaksossa.

5.3 Ei-todennäköisyysotanta, valikoitumisharha ja muita ongelmia

Todennäköisyysotanta ei ole aina mahdollista tai on vaikeata tai kallista. Otos populaatiosta voi silti olla saatavissa. Otanta on tällöin ei-todennäköisyysotantaa (*nonprobability sampling*). Joidenkin populaation alkioiden todennäköisyys tulla otokseen on nolla tai alkioiden sisältymistodennäköisyyksiä ei tiedetä. Näin hankitusta otoksesta on vaikea tehdä tilastotieteellisiä johtopäätöksiä populaatiosta tai arvioida niiden luotettavuutta. Otos ei ole yleensä edustava, koska siinä on *valikoitumisharhaa*.

Valikoitumisharha voi syntyä monella tapaa:

- Kätevyysotos, jossa otokseen tulee mukaan vain helposti saatavilla olevia populaation alkioita (*convenience sample* tai *opportunity sample*).
- Vastaamattomuusharha (*nonresponse bias*) voi syntyä, jos kaikki otokseen valitut eivät vastaa kyselyyn. Tulokset vääristyvät, jos tietynlaiset ihmiset jättävät vastaamatta. Mitä suurempi vastaamattomien osuus, sitä suurempi vastaamattomuusharhan riski.
- Ihmiset hakeutuvat nettilinkin, lehti-ilmoituksen tai muun avoimen kutsun perusteella otokseen. Tällaisen itsevalikoituneen (*volunteer sampling* tai *self-selective sampling*) otoksen on syytä olettaa olevan poikkeuksellinen. Tutkimukseen tai kyselyyn hakeutuva ihmistyyppi löytää kutsun muita todennäköisemmin tai pitää tutkimusta itselleen erityisen tärkeänä.

Otoksiin voi liittyä muitakin ongelmia:

- *Vastausharhaa* (*response bias*) syntyy, jos kysymysten muotoilu tai järjestys tai haastatattelijan käyttäytyminen vaikuttavat vastaukseen.[47]

[47] Esim. V. Angelini, M. Bertoni ja L. Corazzini (2016): Unpacking the Determinants of Life Satisfaction: a Survey Experiment. *Journal of the Royal Statistical Society, A* (painossa).

- Jos kysymykset ovat arkaluonteisia, kysymysten esittämistapa (esim. puhelimitse tapahtuva haastattelu tai kutsu netissä tehtävään haastatteluun) saattaa vaikuttaa vastauksiin.[48]

- Haastateltava voi vastata tietoisesti väärin, jos hän haluaa vaikuttaa tutkimuksen tuloksiin.

- Tutkimuksen tekijä tai hänen apulaisensa saattavat vääristellä otosta tarkoitushakuisesti hyödyntämällä edellä lueteltuja seikkoja, poistamalla tai lisäämällä havaintoja tai raportoimalla tulokset virheellisesti.[49]

Esimerkki. Lumipallo-otanta. Lumipallo-otannassa (*snowball sampling*) otokseen haalitaan ihmisiä tuttujen kautta. Tutut tapaavat jakaa samankaltaisia piirteitä, jotka korostuvat otoksessa. □

Esimerkki.[50] Juutin laatu paalissa. Varhainen esimerkki kätevyysotannasta on Intiasta 1930-luvulta. Juuttipaaleja laivattiin Bombaystä Eurooppaan. Juutin laatua ja arvoa tarkkailtiin survaisemalla paalin kyljestä pyöreäreunainen terävä holkki. Sen kärjellä saatiin näyte paalissa olevasta juutista. Juutti pyrki tiivistymään paalin keskellä, ja holkki pysähtyi tyypillisesti paalin vahingoittumiselle aralle ulkokehälle. Tulos oli säännöllisesti todellista laatua huonompi arvio paalissa olevan juutin laadusta ja arvosta. □

Esimerkki.[51] Yhdysvaltojen presidentinvaali 1936. Klassinen esimerkki valikoitumisharhasta on Literary Digest -lehden kysely presidenttiehdokkaiden kannatuksesta 1936 Yhdysvalloissa. Lehti kokosi ilmeisesti suurimman otoksen koskaan 2.4 miljoonaa yhdysvaltalaista. He kertoivat, äänestävätkö demokraatti Franklin Rooseveltia vai republikaani Alfred Landonia. Kyselyn mukaan Landonista tulisi presidentti 57 %:n kannatuksella. Ennuste epäonistui täydellisesti. Roosevelt voitti ylivoimaisesti 67 % ääniosuudella. Ero lehden ennustamaan kannatukseen oli $67 - 43 = 24$ %-yksikköä, mikä on tiettävästi suurin ennustevirhe koskaan suureen otokseen perustuneessa vaaliennusteessa.

Selitys väärälle ennusteelle oli, että lehti oli lähettänyt 10 miljoonaa tiedustelua muun muassa puhelinluetteloista, lehden omasta tilaajarekisteristä, auton omistajien rekistereistä ja klubien jäsenlistoista kerättyihin osoitteisiin. Vasta 11 miljoonassa yksityistaloudessa oli tuolloin puhelin, ja myös auto oli nykyistä harvinaisempi omistus. Lehden tavoittamat äänestäjät lienevät olleet keskimääräistä vauraampia. Heidän poliittiset näkemyksensä erosivat keskivertoäänestä-

[48] Esim. S. Laaksonen ja M. Heiskanen (2014): Comparison of Three Modes for a Crime Victimization Survey. *Journal of Survey Statistics and Methodology*, 2, 459–483. J. Kuha ja J. Jackson (2014). The Item Count Method for Sensitive Survey Questions: Modelling Criminal Behavior. *Journal of the Royal Statistical Society, C*, 63, 321-341.

[49] Kuuluisa esimerkki on Gregor Mendelin 1865 julkaisema artikkeli perinnöllisyyskokeista. Ronald Fisher argumentoi 1936, että aineisto ei voi olla aito: Has Mendel's Work Been Rediscovered? *The Annals of Science*, 1, 115–137. Ks. myös Ross (2010, 606–608 ja 613–614).

[50] Salsburg (2001, 170).

[51] Freedman (1978, 302–304) sekä `https://en.wikipedia.org/wiki/The_Literary_Digest`, `https://en.wikipedia.org/wiki/George_Gallup` ja `https://en.wikipedia.org/wiki/Gallup_company` (viitattu 16.3.2016).

jän ajatuksista. Vastaamatta jättäneiden suuri osuus (n. 75 %) saattoi myös selittää suurta otantavirhettä.

Vaali 1936 teki kuuluisaksi George Gallupin ja hänen 1935 perustamansa yhtiön. Gallup ennusti Literary Digest -lehden ennusteen jo ennen sen julkaisua sekä vaalien voittajan Rooseveltin oikein. Lehden ennusteen hän selvitti kysymällä 3 000 Literary Digest -lehden käyttämään äänestäjäluetteloon kuuluvalta, kuinka he aikovat äänestää. Vaalien voittajan Gallup selvitti toisella 50 000 äänestäjän erilailla kerätyllä otoksella.[52] □

Esimerkki.[53] Kinsey ja Hite. Alfred Kinseyn ja Shere Hiten kyselytutkimukset miesten ja naisten seksuaalisuudesta ovat vaikuttaneet suuresti monien käsityksiin seksuaalisuudesta ja parisuhteesta. Hite kuvaa kotisivullaan tutkimustensa olleen uutispommi ja sensaatio.

Kinseyn 1940-luvulla suurta huomiota herättäneitä tuloksia olivat, että 70 % miehistä on käyttänyt prostituoitujen palveluita ja että maatiloilla asuvista miehistä 17 %:lla on ollut seksuaalinen suhde eläimeen. Hite raportoi 1978-tutkimuksessaan, että 70 % yli viisi vuotta naimisista olleista naisista harrastaa avioliiton ulkopuolisia suhteita. Osuus oli lähes sama kaikissa kuudessa Hiten tutkimassa etnisessä ryhmässä.

Kinseyn tulokset perustuivat 18 216 yksityiskohtaiseen haastatteluun, joista 11 246 muodosti ”perusotoksen”. Kinsey oli karsinut aineistosta kolmasosan, joka oli peräisin vankiloista ja homoseksuaalisesta yhteisöstä. Karsitussa otoksessa 84 % oli koulutettuja (*college-educated*), koska työväestön edustajat olivat alkuperäisessä otoksessa olleet järjestään vankilassa. Hite oli lähettänyt 100 000 kyselyä, joista 4 500 vastattiin (vastausprosentti 4.5).

Amerikan tilastotieteilijöiden yhdistys julkaisi raportin sekä artikkelin, joissa kritisoitiin Kinseyn tutkimusten uskottavuutta muun muassa siitä, että hänen aineistonsa ei ollut satunnaisotos. Kuuluisa tilastotieteilijä John Tukey opasti Kinseytä, että hän vaihtaisi hetimmiten Kinseyn keräämät 18000 tapaushistoriaa 400 havainnon todennäköisyysotokseen, jos se olisi mahdollista. Spiegelhalter (2015, 10) pitää mahdottomana, että avioliiton ulkopuolisia suhteita harrastavien vaimojen osuus olisi luonnostaan niin sama kuin Hite raportoi sen olevan etnisissä ryhmissä. Spiegelhalter arvioi Hiten aineiston luotettavuuden toiseksi alhaisimpaan luokkaan viisiportaisessa asteikossa. □

Esimerkki. Valikoitumisharha suomalaisessa tutkimuksessa.[54] Vuonna 2005 toteutettiin Suomessa rikosuhritutkimus. Kesällä soitettiin kiinteään puhelinverkkoon kuuluville kotitalouksille. Otokseen ei tullut henkilöitä, joilla oli vain matkapuhelin tai jotka olivat kesäilmoilla ulkona. Aromaan ja Heiskasen (2006) mukaan seuraus oli, että kerätty otos oli väestörakenteeltaan valikoitunut ja tulokset epäluotettavia. Otoksessa oli huomattavasti vähemmän uhreja kuin vuoden

[52] Robinsonin (1937) mukaan jälkimmäisen otoksen koko oli 125 000.

[53] Spiegelhalter (2015, 8, 10, 73 ja 317) ja `http://www.hiteresearchfoundation.org` (viitattu 16.3.2016).

[54] K. Aromaa ja M. Heiskanen (2006): Kansainvälinen rikosuhritutkimus vaikeuksissa. *Haaste*, 3/2006, 16–17.

2000 kyselyssä. Tulos saattaa johtua siitä, että nuoret ovat muita useammin väkivallan uhreja mutta heitä oli otoksessa tavanomaista vähemmän. □

Esimerkki. Vastausprosentit suomalaisessa tutkimuksessa. Yhteiskunta- ja käyttäytymistieteellisissä tutkimuksissa vastausprosentit ovat pienenneet trendimäisesti Suomessa ja kansainvälisesti. Suomalaisten seksuaalielämää kartoittavassa Finsex-tutkimushankkeessa vastausprosentit ovat laskeneet 91.4 %:sta 36.0 %:iin runsaassa kolmessakymmenessä vuodessa:[55]

vuosi	1971	1992	1999	2007	2015
vastaus-%	91.4	75.9	45.8	43.4	36.0

Suureen vastausprosenttiin voidaan edelleen yltää. Halon ym:iden (2011) tutkimuksessa vastausprosentti oli 97.[56] □

Esimerkki. Geeneistä mahdollisesti aiheutuva valikoitumisharha. Ylen uutiset 28.8.2015[57]:

> Professori: Monet tutkimukset liioittelevat liikunnan terveyshyötyjä. Liikunnan vaikutukset terveyteen eivät ehkä ole niin suuria kuin luullaan. Perimä selittää ihmisen fyysisestä kunnosta jopa puolet. – – Huomattava osa tieteellisistä tutkimuksista antaa liikunnan terveyshyödyistä liian valoisan kuvan, arvioi Jyväskylän yliopiston liikuntalääketieteen professori Urho Kujala. Terveysvaikutuksia tutkitaan useimmiten niin sanotuilla väestötutkimuksilla, joissa ensin kysytään isolta joukolta terveitä ihmisiä heidän liikuntatottumuksistaan ja sitten seurataan, kuka pysyy terveenä ja kuka sairastuu tai peräti kuolee pois. Yleensä tuloksia tulkitaan siten, että enemmän liikuntaa harrastavat elävät terveempinä ja pidempään kuin ne, jotka eivät liiku ja että kaikki tämä hyvä johtuu liikunnasta. Professori Kujalan – – mukaan tutkimuksiin liittyy kuitenkin useita ongelmia. – – liikunnan huikeista terveysvaikutuksista on syntynyt liian valoisa kuva. Hyväkuntoiset liikkuvat usein muita enemmän, koska se on heille helppoa. Pitkän päälle he saavat myös keskimäärin vähemmän sairauksia, mutta hyvän fyysisen kunnon taustalla voi olla myös muita asioita kuin liikunta, Kujala sanoo. – – Kaksosilla tehtyjen tutkimusten perusteella noin puolet ihmisen fyysisestä kunnosta selittyy perimällä ja puolet elämäntavoilla. – – väestötutkimuksia tarkempaa tietoa antavat niin sanotut interventiotutkimukset. Niissä otetaan joukko vaikkapa keski-ikäisiä terveitä miehiä ja arvotaan heidät kahteen ryhmään. Toinen ryhmistä pannaan liikkumaan ja toinen jatkaa elämäänsä entiseen malliin. Sitten esimerkiksi kymmenen vuoden kuluttua katsotaan, kuka on sairastunut ja kuka pysynyt terveenä. □

Esimerkki. Harkintaotanta. Harkintaotannassa (*purposive sampling* tai *judgement sampling*) tutkija valitsee otokseen mielestään edustavat havainnot. Se ei takaa edustavuutta. Tutkija ei välttämättä huomaa otantaa vääristävää tekijää. □

Esimerkki. Verrokkien valinta. Monesti vertaillaan kahta ryhmää kuten sairaat ja terveet. Sairaiden otos on olemassa. Vertailuryhmäksi haetaan ihmisiä, jot-

[55] `Http://www.vaestoliitto.fi/tieto_ja_tutkimus/vaestontutkimuslaitos/seksologinen_tutkimus/suomalaisten-seksuaalisuus-finse` (viitattu 16.3.2016).

[56] M.-L. Halko, M. Kaustia ja E. Alanko (2011): The Gender Effect in Risky Asset Holdings. *Journal of Economic Behaviour & Organisation*, 83, 66–81.

[57] `Http://yle.fi/uutiset/professori_monet_tutkimukset_liioittelevat_liikunnan_terveyshyotyja/8255216` (viitattu 16.3.2016)

ka eroavat tutkittavan ominaisuuden suhteen. Tällaisten verrokkien valinnassa voidaan erehtyä. Helsingin Sanomat 22.3.2016:[58]

> Alkoholista terveyshyötyjä löytäneistä tutkimuksista paljastui toistuva virhe. – – uutisissa on kerrottu maltillisen alkoholinkäytön pienentävän riskiä sairastua muun muassa syöpään, nivelreumaan ja sydäntautiin. Kohtuukäyttäjä elää pidempään kuin raitis, tutkimukset kielivät. Näyttö näiden väitteiden paikkansapitävyydestä on – – hutera, paljastui – – kokoomatutkimuksessa. Tim Stockwell Victorian yliopistosta kollegoineen perkasi 87 alkoholitutkimusta ja löysi useimmista virheitä. Yleisin niistä oli se, miten verrokit oli valittu. Kohtuukäyttäjiä – – verrattiin kyllä ihmisiin, jotka eivät juoneet alkoholia lainkaan, mutta tähän ryhmään saattoi kuulua myös raitistuneita juomareita, joiden terveys oli huono. ”Peruskysymys on siis se, keihin kohtuukäyttäjiä verrataan”, Stockwell painottaa – – . Vain 13 tutkimusta 87:stä oli valinnut verrokit oikein. Niiden tuloksissa ei alkoholin terveyshyötyjä ilmennyt. □

Esimerkki.[59] Vaaliennusteet voivat edelleen epäonnistua. Vuoden 1992 parlamenttivaaleissa Iso-Britanniassa 54 kyselyä ennusti Työväenpuolueen saavan Konservatiivipuoluetta enemmän ääniä vaaleja edeltävään päivään asti. Vaaleissa Konservatiivipuolue sai 7.6 %-yksikköä enemmän ääniä. Ero puolueiden kannatuseroon vaaliennusteissa oli 9 %-yksikköä. Ennustevirhe johtui suurelta osin otantaharhasta.

Vuoden 2015 vaaleissa Työväen- ja Konservatiivipuolue kilpailivat taas tasaväkisesti galluppien ennustaessa nyt keskimäärin %-yksikön voittoa Konservatiivipuolueelle. Vaaleissa kannatusero oli 6.6 %-yksikköä konservatiivien hyväksi. Ennustevirhe selittyy taas otantaharhalla. Työväenpuolueen kannattajat olivat osallistuneet konservatiiveja innokkaammin vaalikyselyihin. Gallupyhtiöt olivat kustannussyistä siirtyneet paljolti käyttämään nettipaneeleja, joihin oli ollut helpoin rekrytoida nuoria Työväenpuolueen kannattajia. Ennustevirheen syitä selvittänyt Patrick Sturgis arvioi, että pitäisi selkeästi erotella gallupfirmojen (*pollsters*) huonolaatuiset kyselyt (*polls*) satunnaisotannalla tehdyistä akateemisista tutkimuksista.

Perussuomalaisten kannatuksen arviointi epäonnistui Suomen eduskuntavaaleissa 2011. Perussuomalaiset saivat äänistä 19.1 %. Kolme viimeistä galluppia olivat ennustaneet osuudeksi 15.4–16.9 %. Ero ei ole niin suuri, että selitystä täytyisi hakea systemaattisesti vääristä otantakäytännöistä. □

Lukuisat tutkimustulokset mediassa perustuvat otantaan. Kyky lukea niitä kriittisesti on tilastotieteellistä perusosaamista.

---

[58] `Www.hs.fi/tiede/a1458625192348` (viitattu 22.3.2016).

[59] Jowell ym. (1993) ja B. Tarran (2015): The Failure of the Election Polls and the Future of Survey Research. `Https://www.statslife.org.uk/social-sciences/2573-cathie-marsh-lecture`. `Https://fi.wikipedia.org/wiki/Eduskuntavaalit_2011`. Viittaukset 9.4.2016.

6 Piste-estimointi

6.1 Hyvän estimaattorin ominaisuuksia

Olkoon parametri θ ja sen estimaattori $\hat{\theta}$. Estimaattorin *harha* (*bias*) on erotus

$$b(\theta) = \mathsf{E}(\hat{\theta}) - \theta.$$

Estimaattorin toivotaan monesti olevan *harhaton*, jolloin

$$\mathsf{E}(\hat{\theta}) = \theta.$$

Tällöin riippumattomista otoksista laskettu $\hat{\theta}$ saa keskimäärin oikean arvon θ. Harhattomuus on intuitiivinen ja yleensä toivottava hyvän estimaattorin ominaisuus — muttei olennainen! On olemassa hyviä estimaattoreita, jotka eivät ole harhattomia. On myös tilanteita, joissa harhattomuus ei ole hyvälle estimaattorille välttämätön (esimerkki alempana) tai ehkä edes toivottava ominaisuus (seuraava esimerkki).

Esimerkki. Parametrin θ tiedetään olevan välillä [0,1]. Olkoon $\theta = 1$. Jos θ:n estimaattori olisi harhaton, se saisi mahdollisesti usein 1:stä suurempia arvoja, jotka ovat mahdottomia. □

Oleellinen hyvän estimaattorin ominaisuus on, että se on *tarkentuva*: Havaintojen lukumäärän kasvaessa kohti ääretöntä estimaattorin arvo on varmasti parametrin todellinen arvo θ.

Esimerkki. Suurten lukujen laki. Olkoot satunnaismuuttujat X_i riippumattomia ja olkoon niillä sama odotusarvo $\mathsf{E}(X_i) = \mu$, $i = 1, \ldots, n$. Tällöin keskiarvo

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

saa arvon μ varmasti, kun havaintojen lukumäärä kasvaa kohti ääretöntä. Keskiarvo $\bar{X}$ on tällöin odotusarvon μ tarkentuva estimaattori.

Huom1! Suurten lukujen laki on todennäköisyyden frekvenssitulkinnan taustalla. Huom2! Suurten lukujen lain taustalla on hyvin vähän oletuksia. Esimerkiksi Suurten lukujen laki ei vaadi, että satunnaismuuttujilla X_i olisi sama varianssi. Laki takaa tarkentuvuuden muttei normaalisuutta, koska laki edellyttää vähemmän kuin Keskeinen raja-arvolause (jakso 4.4). □

Parametrille on olemassa usein monia estimaattoreita. Monesti niistä parhaana pidetään sitä, jonka *keskineliövirhe* (*mean-squared error*)

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{E}\,(\hat{\theta} - \theta)^2$$

on pienin. Määritelmä muistuttaa varianssin määritelmää. Mikäli estimaattori on harhaton, keskineliövirhe typistyy estimaattorin varianssiksi. Muulloin keskineliövirhe on estimaattorin varianssin ja harhan neliön summa:

$$\mathsf{E}\,(\hat{\theta} - \theta)^2 = \mathsf{E}\{\hat{\theta} - \mathsf{E}(\hat{\theta}) + [\mathsf{E}(\hat{\theta}) - \theta]\}^2$$

$$
\begin{aligned}
&= \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})]^2 + [\mathsf{E}(\hat{\theta}) - \theta]^2 + 2\mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})][\mathsf{E}(\hat{\theta}) - \theta] \\
&= \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})]^2 + [\mathsf{E}(\hat{\theta}) - \theta]^2 \\
&= \mathsf{V}(\hat{\theta}) + [b(\theta)]^2.
\end{aligned}
$$

Kolmas termi hävisi, koska $\mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})][\mathsf{E}(\hat{\theta}) - \theta] = [\mathsf{E}(\hat{\theta}) - \theta] \times \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})] = [\mathsf{E}(\hat{\theta}) - \theta] \times [\mathsf{E}(\hat{\theta}) - \mathsf{E}(\hat{\theta})] = 0$. (Siirrettiin vakio odotusarvon eteen, ja vakion odotusarvo on vakio. Jakso 4.1.)

Estimaattori $\hat{\theta}$ on keskineliövirheellä mitattuna tarkempi kuin estimaattori $\tilde{\theta}$, jos

$$\mathsf{E}\,(\hat{\theta} - \theta)^2 < \mathsf{E}\,(\tilde{\theta} - \theta)^2.$$

Jos estimaattorit ovat harhattomia, ehto pelkistyy $\hat{\theta}$:n varianssin pienemmyydeksi:

$$\mathsf{V}(\hat{\theta}) < \mathsf{V}(\tilde{\theta}).$$

Ehto voidaan yhtä hyvin ilmaista standardipoikkeamien

$$\mathsf{SD}(\hat{\theta}) < \mathsf{SD}(\tilde{\theta})$$

avulla. Niitä kutsutaan tässä yhteydessä *keskivirheiksi*.

Esimerkki. Odotusarvon estimointi I. Jakauman odotusarvoa voidaan estimoida keskiarvolla tai mediaanilla (otoksen keskimmäisellä havainnolla). Oletetaan, että otoksen havainnot ovat riippumattomia ja noudattavat Normaalijakaumaa $\mathsf{N}(\mu,\sigma^2)$. Jaksossa 4.4 osoitettiin, että keskiarvo on odotusarvon harhaton estimaattori, jonka varianssi on σ^2/n. Voidaan osoittaa, että esimerkin tilanteessa mediaani on myös harhaton estimaattori ja että sen varianssi on suurilla havaintomäärillä noin $1.57 \times \sigma^2/n$. Keskiarvo on keskineliövirheen mielessä tarkempi estimaattori kuin mediaani, kun havainnot ovat normaalijakautuneita. □

Esimerkki. Odotusarvon estimointi II. Havaintojen jakauma vaikuttaa estimaattoreiden ominaisuuksiin. Mediaani ei ole ylipäänsä harhaton estimaattori, jos jakauma on epäsymmetrinen. Noudattakoon satunnaismuuttuja X_i Diskreettiä tasaista jakaumaa

$$\mathsf{P}(X = x) = \begin{cases} \frac{1}{3}, & \text{jos } x = 1{,}2 \text{ tai } 300, \\ 0, & \text{muutoin.} \end{cases}$$

Odotusarvo on

$$\mathsf{E}(X) = \frac{1}{3} \times 1 + \frac{1}{3} \times 2 + \frac{1}{3} \times 300 = 101.$$

Jos otos on suuri, havainnot jakautuvat melko tasaisesti x:n arvoille 1, 2 ja 300. Mediaani tapaa tällöin olla odotusarvoa pienempi 2 ja on harhainen estimaattori. □

Esimerkki. Odotusarvon estimointi III. Olkoon tutkittava jakauma ”pitkähäntäinen”, jolloin suuresti odotusarvosta poikkeavat havainnot ovat todennäköisempiä kuin Normaalijakauman tilanteessa. Voidaan osoittaa, että mediaani voi olla tällöin keskiarvoa tarkempi estimaattori keskineliövirheellä mitattuna. □

Esimerkki. Normaalijakauman varianssin estimointi. Estimoidaan $\mathsf{N}(\mu,\sigma^2)$-jakautuneen satunnaismuuttujan X_i varianssia:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Voidaan osoittaa, että

$$\mathsf{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 < \sigma^2,$$
$$\mathsf{E}(s^2) = \sigma^2$$

ja että

$$\mathsf{MSE}(\hat{\sigma}^2) = \mathsf{E}\,(\hat{\sigma}^2 - \sigma^2)^2 < \mathsf{E}\,(s^2 - \sigma^2)^2 = \mathsf{MSE}(s^2)$$

(esim. Lindgren 1976, 216 ja 256). Estimaattori $\hat{\sigma}^2$ on harhainen mutta keskineliövirheen mielessä tarkempi kuin s^2. Jälkimmäinen on harhaton. Kumpaa tahansa voi perustellusti käyttää. □

6.2 Estimointimenetelmistä

Estimaattoreita voidaan johtaa eri tavoilla. *Suurimman uskottavuuden menetelmä* on niistä tärkeimpiä. Sen lähtökohta on oletus aineiston tuottaneesta jakaumasta. Aineiston perusteella haetaan estimaatit, jotka ovat tietyssä mielessä uskottavimpia arvioita jakauman parametreista. Tällaisten estimaattoreiden voidaan tietyin oletuksin osoittaa olevan suurilla havaintomäärillä keskineliövirheen mielessä tarkimpia mahdollisia ja normaalijakautuneita. Suurimman uskottavuuden estimaattorit eivät ole välttämättä harhattomia.

Toinen paljon käytetty estimointimenetelmä on pienimmän neliösumman menetelmä. Normaalijakauman tilanteessa se ja suurimman uskottavuuden menetelmä tuottavat samat estimaattorit. Menetelmä kuvataan myöhemmin regressioanalyysin yhteydessä (luku 12).

Alla ei yleensä perustella, millä menetelmällä estimaattori on johdettu. Estimaattorin intuitiivisuus ja järkeenkäypyys ovat riittäviä perusteita tutkia ja soveltaa sitä kurssilla. Havaintojen oletetaan alla olevan riippumattomia ja noudattavan otsikoissa nimettyjä jakaumia.[60] Jollei vaihtoehtoisia estimaattoreita parametrille eksplisiittisesti osoiteta, jatkossa viitatut estimaattorit ovat suurimman uskottavuuden estimaattoreita.

[60] Parametrit eivät saa sijaita niin sanotun parametriavaruuden reunapisteessä, jotta tekstissä osoitetut estimaattoreiden jakaumat suurilla havaintomäärillä pätisivät. Esimerkiksi Binomijakauman parametrin tulee sijaita välillä (0,1) (olla nollaa suurempi mutta yhtä pienempi). Jakson 6.1 esimerkinkaltaista tilannetta $\theta = 1$, kun parametriavaruus on [0,1], ei sallita.

6.3 Binomijakauman parametrin estimointi

Binomijakauman parametrille π luonteva estimaattori on tapahtumien osuus otoksessa eli tapahtumien lukumäärä (Y) jaettuna otoskoolla (n):

$$\hat{\pi} = \bar{Y} = \frac{Y}{n}.$$

Se on harhaton estimaattori:

$$\mathsf{E}(\hat{\pi}) = \mathsf{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathsf{E}(Y) = \frac{n\pi}{n} = \pi$$

(kaava (28)). Estimaattorin varianssi on

$$\mathsf{V}(\hat{\pi}) = \mathsf{V}\left(\frac{Y}{n}\right) = \frac{1}{n^2}\mathsf{V}(Y) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

(kaava (29)). Estimaattorin varianssi menee nollaan otoksen koon kasvaessa kohti ääretöntä. Estimaattori on tarkentuva. Se on normeerattuna binomijakautunut

$$n\hat{\pi} \sim \mathsf{Bin}(n,\pi)$$

ja suurilla havaintomäärillä approksimatiivisesti normaalijakautunut

$$\hat{\pi} \sim \mathsf{N}(\pi,\pi(1-\pi)/n)$$

(jakso 4.5.3).

6.4 Multinomijakauman parametrien estimointi

Multinomijakauman solutodennäköisyyksille luonteva estimaattori on solufrekvenssien (N_i) osuudet otoksessa:

$$\hat{\pi}_i = \frac{N_i}{n},$$

$i = 1,\ldots,c$. Normeeratut estimaattorit $n\hat{\pi}_i$ noudattavat Multinomijakaumaa $\mathsf{Mul}(n,\pi_1,\ldots,\pi_c)$. Yksittäinen solufrekvenssi N_i ja siten $n\hat{\pi}_i$ on binomijakautunut jaksoissa 4.2.4 ja 6.3 kuvatulla tavalla. Näin ollen $\hat{\pi}_i$ on harhaton ja tarkentuva estimaattori. Keskeisen raja-arvolauseen perusteella kukin $\hat{\pi}_i$ on normaalijakautunut suurilla havaintomäärillä:

$$\mathsf{E}(\hat{\pi}_i) = \pi_i,$$

$$\mathsf{V}(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n}$$

ja

$$\hat{\pi}_i \sim \mathsf{N}(\pi_i,\pi_i(1-\pi_i)/n).$$

6.5 Normaalijakauman parametrien estimointi

Normaalijakauman $\mathsf{N}(\mu,\sigma^2)$ odotusarvolle ilmeinen estimaattori on otoskeskiarvo

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Se on harhaton ja tarkentuva:

$$\mathsf{E}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{E}(X_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

ja

$$\mathsf{V}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{V}(X_i)}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Varianssin johdossa oletetaan, että havainnot ovat riippumattomia.

Voidaan osoittaa, että pienillä havaintomäärillä tunnusluku

$$\frac{\hat{\mu} - \mu}{s/\sqrt{n}}$$

noudattaa Studentin t-jakaumaa vapausasteilla $n-1$. Tilastollinen päättely odotusarvosta μ kannattaa perustaa ylipäänsä tunnuslukuun yllä. Keskeisen rajaarvolauseen (38) perusteella $\hat{\mu}$ noudattaa suurilla havaintomäärillä approksimatiivisesti

$$\mathsf{N}(\mu,\sigma^2/n)$$

-jakaumaa.

Varianssin estimointia pohdittiin jo esimerkissä jaksossa 6.1. Voidaan osoittaa, että sekä $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$ että $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ ovat σ^2:n tarkentuvia estimaattoreita ja että

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

Tunnusluku vasemmalla noudattaa χ^2-jakaumaa $n-1$ vapausasteella.

6.6 Poisson-jakauman parametrin estimointi

Kun $Y_i \sim \mathsf{Poi}(\mu)$, niin keskiarvo on luonteva estimaattori myös Poisson-jakauman parametrille μ:

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Estimaattori $\hat{\mu}$ on harhaton ja sen varianssi suppenee nollaan otoskoon kasvaessa:

$$\mathsf{E}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{E}(Y_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

ja

$$\mathsf{V}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{V}(Y_i)}{n^2} = \frac{\sum_{i=1}^{n} \mu}{n^2} = \frac{\mu}{n}.$$

Varianssin lasku perustuu havaintojen riippumattomuuteen. Estimaattori on tarkentuva.

Jakson 4.5.5 perusteella $\sum_{i=1}^{n} Y_i \sim \mathsf{Poi}(n\mu)$. Näin ollen normeerattu estimaattori $n\hat{\mu}$ on Poisson-jakautunut:

$$n\hat{\mu} \sim \mathsf{Poi}(n\mu).$$

Keskeisestä raja-arvolauseesta (38) seuraa, että suurilla havaintomäärillä pätee approksimatiivisesti

$$\hat{\mu} \sim \mathsf{N}(\mu,\mu/n).$$

Jos Poisson-jakauma kuvaa aineistoa, tulisi *hajontaindeksin* (*index of dispersion*) $s^2/\hat{\mu}$ olla karkeasti 1. Voidaan osoittaa, että

$$\frac{(n-1)s^2}{\hat{\mu}} \overset{n \text{ suuri}}{\sim} \chi^2(n-1).$$

6.7 Odotusarvon estimointi ilman jakaumaoletusta

Ilmiön taustalla olevaa jakaumaa ei voida aina määritellä. Tällöin voidaan turvautua Keskeiseen raja-arvolauseeseen (38) odotusarvoa estimoitaessa, jos lauseen oletukset ovat voimassa. Odotusarvon otosvastine ja ilmeinen estimaattori on keskiarvo. Keskeisen raja-arvolauseen mukaan suurilla otoskoilla keskiarvon jakauma on approksimatiivisesti

$$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \quad \text{eli} \quad \bar{Y} \sim \mathsf{N}(\mu,\sigma^2/n).$$

Jaksojen 4.4 ja 6.1 johdot keskiarvoestimaattorin harhattomuudelle ja varianssille pätevät tässäkin tilanteessa. Keskiarvo on harhaton ja tarkentuva estimaattori odotusarvolle vaikkei taustalla olevaa jakaumaa rajattaisi tarkasti.

7 Väliestimointi

7.1 Idea

Tilastotieteen ytimessä on päätelmiin liittyvän epävarmuuden osoittaminen ja mittaaminen. Ajatus oli rivien välissä edellisessä luvussa, jossa estimaattorit noudattivat erilaisia jakaumia tilanteesta riippuen. Väliestimointi eli luottamusvälin lasku on paljon käytetty ja monilla aloilla yleistynyt tapa mitata estimaatin tarkkuutta.

Koostukoon otos satunnaismuuttujista $X_1,\ldots,X_n$, ja olkoon estimoitava parametri θ. Luottamusväli θ:lle toteuttaa yhtälön

$$\mathsf{P}[L(X_1,\ldots,X_n) \le \theta \le U(X_1,\ldots,X_n)] = \mathsf{P}(L \le \theta \le U) \ge 1-\alpha. \qquad (39)$$

Siinä $L = L(X_1,\ldots,X_n)$, $U = U(X_1,\ldots,X_n)$, $L \le U$ ja $\alpha \in (0,1)$. Luottamusvälin ala- ja ylärajat ovat L ja U. Luottamusväli $[L,U]$ on θ:n *väliestimaattori*,

ja havaitusta otoksesta laskettu väli $[l,u] = [L(x_1,\ldots,x_n),U(x_1,\ldots,x_n)]$ on θ:n *väliestimaatti*. Ala- ja ylärajat ovat otoksesta laskettavia tunnuslukuja, joiden määritelmä riippuu jakaumasta, johon θ liittyy.

Epäyhtälön (39) mukaan todennäköisyys, että väli $[L,U]$ peittää θ:n on vähintään $1-\alpha$. Tässä yhteydessä todennäköisyyttä $1-\alpha$ sanotaan luottamuskertoimeksi tai -tasoksi (*confidence coefficient* tai *confidence level*). Väliä kutsutaan $100 \times (1-\alpha)$ %:n luottamusväliksi. Tyypillisesti α on 0.05 tai 0.01, jolloin lasketaan 95 %:n tai 99 %:n luottamusväli.

Väliestimoinnissa puhutaan luottamuksesta todennäköisyyden sijaan, jotta parametriin θ ei sekoitettaisi todennäköisyyden käsitettä. Parametri on kiinteä populaatiota kuvaava suure eikä ole satunnaismuuttuja. Satunnaismuuttujia ovat luottamusvälin ala- ja ylärajat. Ne vaihtelevat otoksesta toiseen ja peittävät θ:n (vähintään) $100 \times (1-\alpha)$ %:ssa hypoteettisista riippumattomista toistokokeista.

Esimerkki. Sata luottamusväliä.[61] Kuvassa 20 on sata luottamusväliä odotusarvolle (100). Neljä väritettyä luottamusväliä ei peitä odotusarvoa. Tässä sadan otoksen toistokokeessa 95 %:n luottamusväleistä 96 % peittää estimoitavan parametrin. Jos toistoja olisi ääretön määrä, väleistä 95 % peittäisi odotusarvon. Kuvan luottamusvälien laskutapa selitetään jaksossa **??**. □

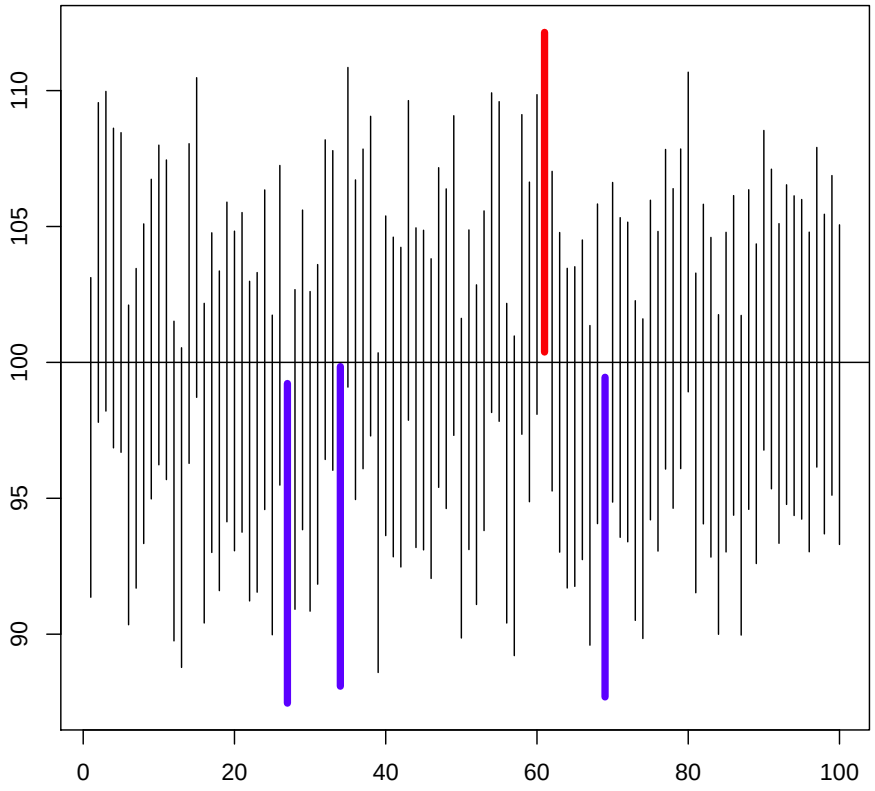
Väliestimaatin etu piste-estimaattiin verrattuna on, että väliestimaatista saa käsityksen estimaatin tarkkuudesta. Mikäli luottamusväli on suuri, θ ei sijaitse välttämättä lähellä piste-estimaattiansa $\hat{\theta}$. Jos väli on kapea, θ vaikuttaa tulleen estimoitua tarkasti.

$100 \times (1-\alpha)$ %:n luottamusväli kertoo myös todennäköisyyden, jolla väli peittää θ riippumattomissa toistokokeissa. Mikäli $\hat{\theta}$:n jakauma on jatkuva (voi saada äärettömän määrän arvoja), on piste-estimaattorin todennäköisyys peittää θ nolla. Todennäköisyydellä mitattuna väliestimoinnilla saavutetaan tavaton parannus piste-estimointiin verrattuna!

Epäyhtälön (39) määrittämä luottamusväli ei ole välttämättä symmetrinen. Useimmiten lasketaan symmetrisiä kaksisuuntaisia luottamusvälejä, jolloin $\theta - L = U - \theta$. Joskus on perusteltua laskea yksisuuntainen luottamusväli esimerkiksi muotoa $(-\infty, U]$, $[L, \infty)$, $[0, U]$ tai $[L, 1]$.

Luottamusvälejä voidaan laskea monella tavalla. Yleensä luottamusväli yritetään muodostaa niin, että se olisi mahdollisimman kapea mutta silti toteuttaisi epäyhtälön (39) ja että todennäköisyys siinä olisi tasan $1-\alpha$. Jos jakauma on diskreetti ja havaintoja on vähän, se ei ole aina mahdollista, minkä takia luottamusvälin peittävyyden todennäköisyys on määritelty epäyhtälönä kaavassa (39).

[61] Kuva pohjautuu Alan Arnholtin R-koodiin (`https://raw.githubusercontent.com/alanarnholt/PASWR2E-Rscripts/master/ChapterScripts/chapter08.R`; viitattu 23.3.2016). Koodi on Ugarten ym:iden (2016) kirjan luvusta 8.



Kuva 20: Tilastotieteellinen Sibelius-monumentti. Sadasta 36 havainnon otoksesta lasketut odotusarvon 95 %:n luottamusvälit. Havaintojen jakauma: $\mathsf{N}(100,18)$.

7.2 Suhteellisen osuuden luottamusväli

Binomijakauman todennäköisyyden π:n eli suhteellisen osuuden luottamusväli on lasketuimpia luottamusvälejä. Tilastotieteen perusoppikirjoissa järjestään opetetaan menettely alla sen laskemiseksi. Johto lähtee suhteellisen osuuden estimaattorin $\hat{\pi} = y/n$ normaalisuudesta suurilla havaintomäärillä:

$$\hat{\pi} \sim \mathsf{N}(\pi,\pi(1-\pi)/n)$$

(jakso 6.3). Tällöin pätee approksimatiivisesti

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

(vrt. kaava (38)). Tässä $z_{\alpha/2} = -z_{1-\alpha/2}$ on Standardinormaalijakauman $(100 \times \alpha/2)$. persentiili. Estimoidaan nimittäjässä oleva estimaattorin varianssi $\hat{\pi}(1-\hat{\pi})/n$:llä. Tämän approksimaation avulla saadaan yhtälöt

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} < z_{1-\alpha/2}\right) = 1 - \alpha \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} < \pi < \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1 - \alpha.$$

Suhteellisen osuuden $100(1-\alpha)$ %:n luottamusvälin ylä- ja alaraja ovat

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{40}$$

Jos havaintoja on paljon eikä π ole lähellä nollaa tai yhtä, luottamusvälin peittävyys on noin $100 \times (1-\alpha)$ %.

Luottamusvälin leveys riippuu luottamustasosta, π:n suuruudesta ja otoskoosta.

Luottamustason kasvattaminen $(1-\alpha)$:sta $(1-\alpha^*)$:iin leventää luottamusväliä. Luottamusvälin leveyden määräävä termi $z_{1-\alpha/2}[\hat{\pi}(1-\hat{\pi})/n]^{1/2}$ suurenee tällöin, koska $z_{1-\alpha^*/2} > z_{1-\alpha/2}$, jos $\alpha^* > \alpha$.

Esimerkki. Luottamusvälin leveys ja luottamustaso. Olkoon $\hat{\pi} = 0.5$ ja $n = 100$. Jos luottamustaso on 0.95, niin $\alpha = 0.05$ ja $z_{1-\alpha/2} = 1.960$. Jos luottamustaso on 0.99, niin $\alpha = 0.01$ ja $z_{1-\alpha/2} = 2.576$. Standardinormaalijakauman (97.5). ja (99.5). persentiilit 1.960 ja 2.576 on laskettu R-komennoilla:

```
qnorm(0.975)
qnorm(0.995)
```

Kaksisuuntaisen 95 %:n luottamusvälin rajat ovat

$$0.5 \pm 1.960\sqrt{\frac{0.5 \times 0.5}{100}} = 0.5 \pm 1.960 \times 0.05 = 0.5 \pm 0.098,$$

eli luottamusväli on

$$[0.402, 0.598].$$

Kaksisuuntaisen 99 %:n luottamusvälin rajat ovat

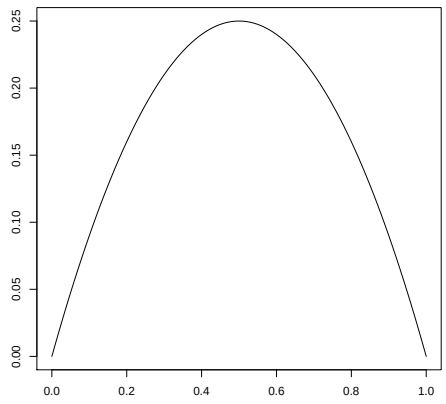
$$0.5 \pm 2.576\sqrt{\frac{0.5 \times 0.5}{100}} = 0.5 \pm 2.576 \times 0.05 = 0.5 \pm 0.1288,$$

jolloin luottamusväli on

$$[0.371,0.630].$$

Luottamustason nosto kasvattaa luottamusvälin leveyden $0.598-0.402 = 0.196$:sta $0.630 - 0.371 = 0.258$:aan. □

Luottamusväli tapaa olla sitä leveämpi, mitä lähempänä π ja siten $\hat{\pi}$ on 0.5:ttä. Ääriarvoja 0 tai 1 lähellä olevat osuudet estimoituvat tarkemmin kuin 0.5:n tienoilla olevat. Kuva 21 havainnollistaa tuloa $\pi(1-\pi)$, kun $\pi \in [0,1]$.



Kuva 21: Todennäköisyys π ja tulo $\pi(1-\pi)$.

Luottamusvälin leveys riippuu myös havaintomäärästä. Laskettaessa 95 %:n luottamusväli

$$z_{1-\alpha/2} = z_{1-0.05/2} = z_{0.975} \approx 1.960 \approx 2.$$

Tällöin estimaatin $\hat{\pi}$ pysyessä samana luottamusvälin (40) leveys noin puolittuu, jos n nelinkertaistuu:

$$1.960 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{4n}} \approx 2 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{4n}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Vasemmalla otoskoko on nelinkertaistettu. Seuraus on luottamusväli noin puoliintuminen kaavaan (40) verrattuna. Jaksossa 8.2 on numeerinen esimerkki.

On esitetty peukalosääntöjä, joiden pätiessä approksimaation pitäisi toimia (esim. "$n\pi > 10$ ja $n(1-\pi) > 10$", "$n\pi(1-\pi) \geq 10$" tai "$n\pi - 3[n\pi(1-\pi)]^{1/2} > 0$ ja $n\pi + 3[n\pi(1-\pi)]^{1/2} < n$"). Ne eivät takaa, että luottamusvälin peittävyys olisi tarkoitetunlainen. Luottamusvälin (40) peittävyys voi olla paljon pienempi kuin $100 \times (1-\alpha)$ %.

Esimerkki. Suhteellisen osuuden luottamusvälin (40) peittävyys.[62] Jos havaintoja on 25 ja π on noin 0.05, niin 95 %:n luottamusvälin (40) peittävyys on noin 70 %. □

Esimerkki. Luottamusväli, jos tapahtumia ei ole. Jos $\hat{\pi} = y/n = 0/n = 0$, luottamusväli (40) typistyy pisteeksi $[0, 0]$ kaikilla luottamustasoilla. Se on epätyydyttävää: Jos voitaisiin olla täysin varmoja, että tapahtuman todennäköisyys on 0, ei aineistoa olisi kerätty eikä luottamusväliä laskettu. □

Newcombista (1998a, 868) luottamusväliä (40) ei tulisi käyttää lainkaan tieteellisessä tutkimuksessa ja sen käyttö tulisi rajata otoskoon suunnitteluun (jakso 8.2) ja opetustarkoituksiin. Schilling ja Doi (2014) arvioivat samaan tapaan, että luottamusväli (40) ei ole käyttökelpoinen.

Paljon paremmin toimiva kaksisuuntainen 95 %:n luottamusväli on plus neljä -luottamusväli (Agresti ja Coull 1998). Havaintoihin lisätään neljä havaintoa taulukon alla mukaisesti (alunperin n havaintoa ja y tapahtumaa):

tapahtuma		
kyllä	ei	Σ
$y + 2$	$n - y + 2$	$n + 4$

Σ:lla on merkitty summaa rivin lukumääristä. Luottamusväli lasketaan muokatusta aineistosta kaavalla (40). Plus neljä -luottamusväli on hieman leveämpi ja peittää todellisen suhteellisen osuuden todennäköisyydellä, joka tapaa olla selvästi lähempänä 95 %:a kuin alkuperäisestä aineistosta kaavalla (40) laskettu luottamusväli (mt. ja Newcombe 2013, 106 ja 109). Jos π on hyvin lähellä nollaa tai yhtä, plus neljä -luottamusväli on liian leveä.

Joskus binomikokeessa tapahtumia ei tule lainkaan. Tällöin kätevä kaava π:n 95 %:n luottamusväliksi on *kolmen sääntö* (*rule of three*)

$$\left[0, \frac{3}{n}\right] \approx \left[0, \frac{3}{n+1}\right].$$

Jälkimmäinen approksimaatio on parempi (Jovanovic 2005). Approksimaatiot ovat toimivia, jos $n > 30$. Nämä ovat yksisuuntaisia luottamusvälejä.

Esimerkki.[63] Tehdas valmistaa 300 laskuvarjoa uudella menetelmällä. Jokaisen uuden laskuvarjon aukeamista käytössä kokeillaan, ja kaikki aukeavat. Laske 95 %:n luottamusväli uusille laskuvarjoille, jotka eivät aukea käytössä.

Kolmen säännön mukaan yksisuuntainen 95 %:n luottamusväli on

$$\left[0, \frac{3}{300}\right] = [0{,}0.01].$$

Laskuvarjon avautumattomuuden todennäköisyys vaikuttaa hyvin pieneltä. □

[62] Agresti (2013, 604).

[63] `Https://en.wikipedia.org/wiki/Rule_of_three_(statistics)` (viitattu 24.3.2016).

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Plus neljä -luottamusväli ja kolmen sääntö on tarkoitettu 95 %:n luottamusvälin laskemiseksi muttei muille luottamustasoille. Luottamusväli tällä ja muilla luottamustasoilla on suositeltavaa laskea esimerkiksi Wilsonin menetelmällä tai keski-p-korjattuna (*mid-p*) Clopper–Pearson-luottamusvälinä. Menetelmät selostetaan Agrestin (2007 ja 2013) sekä Newcomben (2013) kirjoissa. Newcombista (1998a) Wilsonin menetelmä on ainoa helppolaskuinen toimiva menetelmä. Keski-p-korjattu Clopper–Pearson-luottamusväli on laskennallisesti työläämpi, mutta silti helposti toteutettavissa Anna Gottardin R-koodin avulla.[64] Prattin estimaattori (Wilcox 2012, 165–167) toimi hyvin Blythin (1986) simulointikokeissa. Prattin estimaattori on helppo laskea Rand Wilcoxin WRS-paketilla R:lle.

Kolmen sääntö on yksisuuntaisen 95 %:n Clopper–Pearson-luottamusvälin approksimaatio. Hernández, Andrés ja Tejedor (2016) suosittelevat yksisuuntaista luottamusväliä laskettavaksi Wilsonin menetelmällä jatkuvuuskorjauksella tai plus neljä -luottamusväliä muistuttavalla Borkowfin menettelyllä. 95 %:n luottamusväliä laskettaessa se tuottaa lähes saman luottamusvälin kuin kolmen sääntö (Borkowf 2006).

7.3 Suhteellisten osuuksien erotuksen luottamusväli, jos osuudet ovat riippumattomia

Ollaan kiinnostuneita, kuinka paljon tapahtumien todennäköisyydet π_1 ja π_2 eroavat kahdessa binomijakautuneessa ilmiössä. Oletetaan, että käytettävissä on n_1:n ja n_2:n kokoiset riippumattomat otokset, joiden avulla voidaan estimoida tapahtumien havaitut osuudet $\hat{\pi}_1 = y_1/n_1$ ja $\hat{\pi}_2 = y_2/n_2$ ja niiden erotus $\hat{\pi}_1 - \hat{\pi}_2$. Näissä y_1 ja y_2 ovat tapahtumien lukumäärät otoksissa.

Molemmat suhteellisen osuuden estimaattorit ovat suurilla havaintomäärillä normaalijakautuneita:

$$\hat{\pi}_1 \sim \mathsf{N}(\pi_1, \pi_1(1-\pi_1)/n_1) \quad \text{ja} \quad \hat{\pi}_2 \sim \mathsf{N}(\pi_2, \pi_2(1-\pi_2)/n_2)$$

(jakso 6.3). Koska otokset ovat riippumattomia, erotuksen $\hat{\pi}_1 - \hat{\pi}_2$ varianssi on approksimatiivisesti

$$\mathsf{V}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

(jakso 4.1). Erotuksen keskihajonnan luonteva estimaatti on

$$\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

Erotuksen $100 \times (1-\alpha)$ %:n approksimatiivisen luottamusvälin ala- ja ylärajat ovat nyt

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}, \tag{41}$$

[64] Http://www.stat.ufl.edu/~aa/cda/R/one-sample/R1/index.html (viitattu 27.3.2016).

koska

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} < z_{1-\alpha/2} \right) \approx 1 - \alpha \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\pi}_1 - \hat{\pi}_2 - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} < \pi_1 - \pi_2 < \right.$$

$$\left. \hat{\pi}_1 - \hat{\pi}_2 + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \right) \approx 1 - \alpha.$$

Tämäkin luottamusväli perustuu jakson jakso 6.3 normaalisuusapproksimaatioon. Tässä se toimii paremmin kuin yhden osuuden suuruutta arvioitaessa mutta tapaa silti tuottaa liian lyhyitä luottamusvälejä (esim. Bilder ja Loughlin 2015, 33–34). Erilaisia peukalosääntöjä approksimaation kelvollisuudelle on annettu (esim. "$n_1 > 30$ ja $n_2 > 30$", "$n_i\hat{\pi}_i \geq 5$ ja "$n_i(1-\hat{\pi}_i) \geq 5$, $i = 1{,}2$" sekä "$y_i \geq 10$ ja $n_i - y_i \geq 10$, $i = 1{,}2$"). Ne eivät takaa, että luottamusväli olisi tarkoitetun kokoinen eli että se peittäisi parametrien erotuksen (noin) todennäköisyydellä $1 - \alpha$, jos π_1 tai π_2 on lähellä nollaa tai yhtä tai havaintoja on vähän.

Huomattavasti paremmin toimivia tapoja muodostaa luottamusväli suhteellisten osuuksien erotukselle on olemassa. Oletetaan, että on estimoitu suhteelliset osuudet $\hat{\pi}_1 = n_{11}/n_1$ ja $\hat{\pi}_2 = n_{21}/n_2$ kahdesta toisistaan riippumattomasta otoksesta:

	tapahtuma		
	kyllä	ei	Σ
ryhmä 1	n_{11}	n_{12}	n_1
ryhmä 2	n_{21}	n_{22}	n_2

Hyvin yksinkertainen parannuskeino on lisätä yksi havainto kuhunkin lukumäärään:

	tapahtuma		
	kyllä	ei	Σ
ryhmä 1	$n_{11} + 1$	$n_{12} + 1$	$n_1 + 2$
ryhmä 2	$n_{21} + 1$	$n_{22} + 1$	$n_2 + 2$

Näin muokatusta aineistosta kaavalla (41) laskettua luottamusväliä kutsutaan Agresti–Caffo-luottamusväliksi. Se on lähellä tarkoitetun levyistä pienilläkin havaintomäärillä (esim. $n_1 = n_2 = 20$). Jos havaintoja on tavattoman vähän (esim. $n_1 = n_2 = 10$), niin luottamusväli on selkeästi liian leveä, jos π_i:t ovat lähellä nollaa tai yhtä mutta tällöinkin voi muulloin toimia kohtuullisesti (Agresti ja Caffo 2000).

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Bilder ja Loughin (2015, 29) suosittelevat Agresti–Caffo-luottamusväliä. Sitä parempi on silti esimerkiksi neliöi ja summaa -Wilson-luottamusväli (*square–and–add* tai *hybrid score*; Newcombe 1998b ja 2013 sekä Agresti ja Caffo 2000). Se ei ole vaikea laskea mutta vaatii kvadraattisen yhtälön ratkaisemisen kuten suhteellisen osuuden luottamusvälikin Wilsonin menetelmällä laskettaessa.

7.4 Suhteellisten osuuksien erotuksen luottamusväli, jos osuudet eivät ole riippumattomia

Tutkitaan kaksiarvoisia satunnaismuuttujia X ja Y, jotka noudattavat Multinomijakaumaa $\mathsf{Mul}(1,\pi_{11},\pi_{12},\pi_{21},\pi_{22})$:

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

Todennäköisyydet on järjestetty 2×2 -taulukoksi. Siinä π_{ij} on *solutodennäköisyys* eli todennäköisyys, että X on saanut arvon x_i ja Y on saanut arvon y_j eli että molempien satunnaismuuttujien arvo osuus (i,j)-*soluun* $(i,j = 1,2)$. Rivien ja sarakkeitten todennäköisyydet on summattu *reunatodennäköisyyksiksi*. Ne kertovat todennäköisyyden, että X saa arvon x_1 tai x_2 (π_{1+} tai π_{2+}) tai Y saa arvon y_1 tai y_2 (π_{+1} tai π_{+2}). Jos alaindeksi "1" indikoi tapahtumaa, tapahtumisen todennäköisyydet ovat π_{1+} ja π_{+1} satunnaismuuttujille X ja Y.

Johdetaan luottamusväli erotukselle $\pi_{1+} - \pi_{+1}$, kun käytettävissä on n havaintoa:

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

Tässä n_{ij} on havaintojen frekvenssi eli lukumäärä (i,j)-solussa. Frekvenssit on summattu riveittäin ja sarakkeittain *reunafrekvensseiksi* (n_{1+}, n_{2+}, n_{+1} ja n_{+2}).

Ilmeiset estimaatit solu- ja reunatodennäköisyyksille ovat

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}, \ \hat{\pi}_{i+} = \frac{n_{i+}}{n} \quad \text{ja } \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

Erotuksen $\pi_{1+} - \pi_{+1}$ estimaatti on

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} = \frac{n_{1+}}{n} - \frac{n_{+1}}{n} = \frac{n_{11} + n_{12} - (n_{11} + n_{21})}{n} = \frac{n_{12} - n_{21}}{n}.$$

Erotus $\hat{\pi}_{1+} - \hat{\pi}_{+1}$ on sitä suurempi, mitä suurempi erotus $n_{12} - n_{21}$ on.

Suurilla havaintomäärillä $\hat{\pi}_{1+} - \hat{\pi}_{+1}$ on normaalijakautunut. Erotuksen varianssia ei voida laskea termien varianssien summana kuten jaksossa 7.3: Estimaattorit $\hat{\pi}_{1+}$ ja $\hat{\pi}_{+1}$ koostuvat osin frekvenssistä n_{11}, joten ne eivät ole riippumattomia eivätkä jakson 4.1 laskusäännöt ole käytettävissä. Voidaan osoittaa, että erotuksen estimoitu varianssi on suurilla havaintomäärillä

$$\frac{n_{12} + n_{21} - (n_{12} - n_{21})^2/n}{n^2}$$

(esim. Agresti 2007, 246). Erotuksen $\pi_{1+} - \pi_{+1}$ approksimatiivisen $100 \times (1-\alpha)$ %:n luottamusvälin rajat ovat

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} \pm z_{1-\alpha/2} \frac{\sqrt{n_{12} + n_{21} - (n_{12} - n_{21})^2/n}}{n}.$$

Esimerkki. Parisuhdeväkivalta. Taulukoissa on frekvenssit ja osuudet kuudes- ja yhdeksäsluokkalaisten lasten havainnoista vanhempiinsa kohdistuneesta parisuhdeväkivallasta.[65]

Nähnyt tai kuullut (lkm) parisuhdeväkivaltaa, joka kohdistuu		isään		
		kyllä	ei	Σ
äitiin	kyllä	516	674	1190
	ei	231	12038	12269
	Σ	747	12712	13459

Nähnyt tai kuullut (%) parisuhdeväkivaltaa, joka kohdistuu		isään		
		kyllä	ei	Σ
äitiin	kyllä	3.8	5.0	8.8
	ei	1.7	89.4	91.2
	Σ	5.6	94.4	100.0

Lapsista 8.8 % ($\hat{\pi}_{1+}$) on aistinut äitiin ja 5.6 % ($\hat{\pi}_{+1}$) isään kohdistunutta väkivaltaa. Osuuksien erotus on 3.3 %-yksikköä. Erotuksen $\pi_{1+} - \pi_{+1}$ 99 %:n luottamusvälin ylä- ja alarajat ovat

$$\begin{aligned} &\frac{674 - 231}{13459} \pm 2.575829 \frac{\sqrt{674 + 231 - (674 - 231)^2/13459}}{13459} \\ &= 0.03291478 \pm 0.005710859. \end{aligned}$$

Luottamusväli on noin [0.027,0.039]. Se on laskettu R-koodilla alla.

[65] M. Huttunen, M. Husso ja J. Hietamäki (2015): Sukupuoliero parisuhdeväkivallan yleisyydessä ja sen havaitsemisessa lasten ja nuorten näkökulmasta. *Janus*, 23, 109–126. Aineisto on vuonna 2008 kerätystä lapsiuhritutkimuksesta.

```
n12 <- 674
n21 <- 231
n <- 13459
z <- qnorm(0.995)
(n12-n21)/n+z*sqrt(n12+n21-(n12-n21)^2/n)/n
(n12-n21)/n-z*sqrt(n12+n21-(n12-n21)^2/n)/n
```

Luottamusväli ei peitä nollaa. Luottamustasolla 0.95 ero on $2.7 - 3.9$ %. Lapset ovat havainneet enemmän äitiinsä kuin isäänsä kohdistettua väkivaltaa. Luottamusvälin leveys on 1.2 %-yksikköä. Ero on saatu estimoitua melko tarkasti. □

7.5 Poisson-jakauman odotusarvon luottamusväli

$\mathsf{Poi}(\mu)$-jakautuneesta satunnaismuuttujasta on n riippumatonta havaintoa Y_i. Keskeisen raja-arvolauseen (kaava (38)) perusteella suurilla havaintomäärillä

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sqrt{\mu/n}} < z_{1-\alpha/2}\right) = 1 - \alpha,$$

jossa odotusarvon estimaattori on $\hat{\mu} = \bar{Y} = \sum_{i=1}^{n} Y_i/n$. Sijoitetaan varianssin paikalle sen estimaattori $\hat{\mu}/n$:

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sqrt{\hat{\mu}/n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha \Leftrightarrow$$
$$\mathsf{P}\left(\hat{\mu} - z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}} < \mu < \hat{\mu} + z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}\right) \approx 1 - \alpha.$$

Poisson-jakauman odotusarvon $100(1-\alpha)$ %:n luottamusvälin ala- ja yläraja ovat

$$\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}. \tag{42}$$

Tällainen luottamusväli kärsii samantapaisista ongelmista kuin suhteellisen osuuden luottamusväli (40). Välin peittävyys on pienillä μ:n ja n:n arvoilla yleensä tarkoitettua pienempi (Bilder ja Loughin 2015, 198–202 sekä Byrne ja Kabaila 2005). Jos tapahtumia ei ole ($\hat{\mu} = \sum_{i=1}^{n} Y_i/n = 0/n = 0$), luottamusväli surkastuu pisteeksi [0,0]. Approksimaation toimivuus paranee μ:n ja n:n kasvaessa. Molempien tulisi olla melko suuria tai n:n erityisen suuri. Peukalosääntö toimivuudelle on $n\mu > 100$ (Armitage ym. 2002, 154).

Myös muotoa

$$\hat{\mu}^* \pm z_{1-\alpha/2}\sqrt{\hat{\mu}^*} = y \pm z_{1-\alpha/2}\sqrt{y}$$

käytetään. Siinä $\mu^* = n\mu$ ja $\hat{\mu}^* = n\hat{\mu} = y$, joka on $Y \equiv \sum_{i=1}^{n} Y_i$:n arvo otoksessa. Perustelu: n:n riippumattoman $\mathsf{Poi}(\mu)$-jakautuneen satunnaismuuttujan summa on $\mathsf{Poi}(n\mu)$-jakautunut, joten $Y = \sum_{i=1}^{n} Y_i$ on $\mathsf{Poi}(n\mu)$-jakautunut ja suurilla n:n

arvoilla approksimatiivisesti $\mathsf{N}(n\mu,n\mu)$-jakautunut. (Kaava (36) ja jakso 4.5.5.) Tällöin pätee approksimatiivisesti

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu}^* - \mu^*}{\sqrt{\mu^*}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

Muoto yllä seuraa. Tässä aineisto hahmotetaan yhtenä yhdistettynä otoksena $\mathsf{Poi}(\mu^*)$-jakaumasta.

Esimerkki.[66] Rintasyöpä. Kohorttitutkimuksissa seurataan kahta ihmisryhmää — tyypillisesti otosta — joista toinen altistuu riskille ja toinen ei. Monesti altistumista mitataan henkilövuosilla eli otoksen koolla kerrottuna seuranta-ajalla.

Tuberkuloosia sairastaneiden naisten keuhkoja on tutkittu röntgenillä fluoresoivan varjostimen avulla (*fluoroscopy*). Seuraavien 28010 henkilövuoden aikana 41 naista sairastui rintasyöpään. Vertailuryhmässä, jonka keuhkoja ei oltu tutkittu, naisille kehittyi 15 rintasyöpää 19017 henkilövuoden aikana. Syöpään altistumisriskeiksi tuhatta henkilövuotta kohden estimoidaan $41/(28010/1000) = 41/28.01 = 1.463763$ ja $15/(19017/1000) = 15/19.017 = 0.7887679$. Mallitetaan sairastuneiden lukumääriä $\mathsf{Poi}(\mu_i)$-jakaumalla, $i = 1{,}2$. Estimoitu altistumisriski tuhatta henkilövuotta kohden vastaa $\hat{\mu}$:a ja tuhannet henkilövuodet otoskokoa n kaavassa (42). Otoskoko ei ole kokonaisluku aineiston muodostamistavasta johtuen.

Ensimmäisessä otoksessa 95 %:n luottamusvälin μ_1:lle rajat ovat

$$1.463763 \pm 1.959964 \times \sqrt{1.463763/28.01} = 1.463763 \pm 0.4480505.$$

Luottamusväli on noin [1.02,1.91]. Vastaavan luottamusvälin μ_2:lle rajat ovat

$$0.7887679 \pm 1.959964 \times \sqrt{0.7887679/19.017} = 0.7887679 \pm 0.3991643.$$

Luottamusväli on noin [0.39,1.19].

Peukalosääntö odotusarvojen luottamusvälien käyttökelpoisuudelle ei toteudu: $28.01 \times 1.463763 = 41 < 100$ ja $19.017 \times 0.7887679 = 15 < 100$. Vaikka havaintoja on paljon, on syöpäriski pieni. Niiden tulo ei ole tarpeeksi suuri takaamaan normaalisuusapproksimaation toimivuutta. □

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Poisson-jakauman odotusarvon luottamusväli voidaan laskea lukuisilla tavoilla, joista moni on edellä esitettyä parempi (Barker 2002, Byrne ja Kabaila 2005 sekä Patil ja Kulkarni 2012). Yhtä yksinkertaista ja toimivaa parannusta kuin plus neljä -luottamusväli (suhteellista osuutta väliestimoitaessa) ei ole kehitetty.

Monissa yhteyksissä käytetty jatkuvuuskorjaus on helppo ja tuottaa paremman peittävyyden (Byrne ja Kabaila 2005). Laskennallisesti helpoimpia on pistemääräperiaatteelle perustuva luottamusväli (esim. Agresti ja Coull 1998, Andersson 2015, Bilder ja Loughin 2015, 198–202, Byrne ja Kabaila 2005, Davison 2003, Newcombe 2013, luku 6, Swift 2009 tai Stuart ja Ord 1991, 755): Jos $Y \sim \mathsf{Poi}(\mu^*)$, niin suurilla havaintomäärillä

$$\frac{Y - \mu^*}{\sqrt{\mu^*}}$$

[66] Newcombe (2013, 118). Alkuperäistutkimus: Boice ja Monson (1977).

on standardinormaalijakautunut. Tällöin $1 \times (1-2\alpha)$ %:n luottamusvälin ala- ja yläraja saadaan ratkaisuna yhtälöstä

$$(Y-\mu^*)^2 = z_{1-\alpha/2}\mu^*.$$

Myös voidaan laskea eksakti luottamusväli, jonka peittävyys tiedetään liian suureksi (Agresti ja Coull 1998, Barker 2002, Casella ja Berger 2002, 434–435, Fleiss, Levin ja Paik 2013, 342 sekä Pawitan 2013, 134–135) tai keski-p-korjattu eksakti luottamusväli (Byrne ja Kabaila 2005, Cohen ja Young 1994 sekä Newcombe 2013, luku 6). Vaihtoehtoja on muitakin. Bilder ja Loughin (2015, 198–202) suosittelevat yllä esitettyä pistemääräluottamusväliä: Se on parempi kuin luottamusväli päätekstissä eikä muiden menetelmien edut siihen verrattuna ole suuria. Newcombe (2013, 120) on kriittisempi. Newcombe (mts. 122) laskee keski-p-korjatut eksaktit luottamusvälit rintasyöpäesimerkin aineistolla. Välit poikkeavat jonkin verran edellä lasketusta ([1.03,1.97] ja [0.46,1.27]).

7.6 Riippumattomien Poisson-jakautuneiden satunnaismuuttujien odotusarvojen erotuksen luottamusväli

Käytettävissä on kaksi riippumatonta otosta $\mathsf{Poi}(\mu_i)$-jakautuneista satunnaismuuttujista, $i = 1{,}2$. Estimaattorit $\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ ovat normaalijakautuneita

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\mu_i/n_i}} \sim \mathsf{N}(0{,}1)$$

suurilla havaintomäärillä (n_i) edellisen jakson tapaan. Edellä Y_{ij} on j. havainto satunnaismuuttujasta i. otoksessa. Erotuksen $\hat{\mu}_1 - \hat{\mu}_2$ varianssi on riippumattomuuden perusteella varianssien summa $\mu_1/n_1 + \mu_2/n_2$ (jakso 4.1). Erotuksen $100(1-\alpha)$ %:n luottamusväli voidaan jälleen perustaa normaalisuuteen:

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}}} < z_{1-\alpha/2} \right) \approx 1-\alpha \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\mu}_1 - \hat{\mu}_2 - z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}} < \mu_1 - \mu_2 < \right.$$

$$\left. \hat{\mu}_1 - \hat{\mu}_2 + z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}} \right) \approx 1-\alpha.$$

Luottamusvälin rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}}. \tag{43}$$

Vaihtoehtoinen muotoilu on jälleen mahdollinen. Erotuksen $\hat{\mu}_1 - \hat{\mu}_2$ komponentit ovat riippumattomia ja approksimatiivisesti pätee $\hat{\mu}_1 \sim \mathsf{N}(\mu_1,\mu_1/n_1)$, $\hat{\mu}_2 \sim \mathsf{N}(\mu_2,\mu/n_2)$, $Y_1 \equiv \sum_{j=1}^{n_1} Y_{1j} = n_1\hat{\mu}_1 \sim \mathsf{N}(n_1\mu_1,n_1\mu_1)$, $Y_2 \equiv \sum_{j=1}^{n_2} Y_{2j} = n_2\hat{\mu}_2 \sim \mathsf{N}(n_2\mu_2,n_2\mu_2)$ ja $\mathsf{V}(Y_1 - Y_2) = n_1\mu_1 + n_2\mu_2$. Merkitään $\mu_i^* = n_i\mu_i$. Erotuksen $\mu_1^* - \mu_2^*$ luottamusvälin luottamustasolla $(1-\alpha)$ rajat ovat

$$\hat{\mu}_1^* - \hat{\mu}_2^* \pm z_{1-\alpha/2}\sqrt{\hat{\mu}_1^* + \hat{\mu}_2^*} = y_1 - y_2 \pm z_{1-\alpha/2}\sqrt{y_1 + y_2}.$$

Yllä y_i on Y_i:n havaittu arvo.

Rajojen (43) määrittelemä luottamusväli toimii hyvin, jos $n_1 = n_2$, $\mu_1^* = n_1\mu_1 > 2$ ja $\mu_2^* = n_2\mu_2 > 2$. Muulloin välin peittävyys voi olla paljon pienempi kuin sen nimellinen peittävyys $1 - \alpha$. (Krishnamoorthy ja Lee 2012.)

Esimerkki. Rintasyöpä (jatkoa). 95 %:n luottamusvälin erotukselle $\mu_1 - \mu_2$ rajat ovat

$$1.463763 - 0.7887679 \pm 1.959964 \times \sqrt{1.463763/28.01 + 0.7887679/19.017}$$
$$= 0.6749951 \pm 0.6000678.$$

Luottamusväli on noin $[0.07, 1.28]$. Se ei peitä nollaa. Luottamusvälin mukaan röntgenillä fluoresoivan varjostimen avulla tutkitut naiset sairastuvat useammin rintasyöpään kuin näin tutkimattomat naiset. Luottamusväli on (tutkittavan asian kannalta) leveä, joten eron suuruutta ei ole saatu selvitettyä tarkasti.

Huom! Erotuksen luottamusväli ei kata nollaa, vaikka μ_1:n ja μ_2:n luottamusvälit [1.02,1.91] ja [0.39,1.19] lomittuvat toistensa päälle. Odotusarvojen luottamusvälien lomittumisesta ei pidä päätellä, että odotusarvot eivät eroaisi. □

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Jos havaintoja ei ole paljon, voidaan käyttää kehittyneempiä tekniikoita (Li ym. 2011 sekä Krishnamoorthy ja Lee 2012). Li ym. (2011) laskevat luottamusvälejä eri tekniikoilla Poisson-odotusarvojen erotukselle esimerkin rintasyöpäaineistolle. Kaikkien luottamusvälien mukaan odotusarvot eroavat. Ng, Gua ja Tang (2007) tutkivat vaihtoehtoisia piste-estimaattoreita ja arvioivat syöpäriskin eron suuruutta esimerkin aineiston avulla.

7.7 Luottamusvälejä havaintojen ollessa normaalijakautuneita

Edellä vertailut perustettiin Keskeiseen raja-arvolauseeseen, joka takaa keskiarvon normaalisuuden suurilla otoskoilla. Seuraus oli, että pienillä havaintomäärillä luottamusvälien todellinen peittävyys ei välttämättä ollut tarkoitetunlainen.

Tässä jaksossa oletetaan, että havainnot ovat normaalijakautuneita. Tällöin luottamusvälit voidaan laskea niin, että niiden peittävyys on täsmälleen oikea kaikilla havaintomäärillä.

7.7.1 Normaalijakauman odotusarvon luottamusväli, jos σ^2 tunnetaan

Jos $X_i \sim \mathsf{N}(\mu,\sigma^2)$, niin

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1).$$

Yllä $\bar{X} = \sum_{i=1}^{n} X_i/n$. Tällöin pätee eksaktisti

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

Jos σ^2 tunnetaan, voidaan johtaa ja laskea jakson 7.2 tapaan μ:lle $100(1-\alpha)$ %:n luottamusväli

$$\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}. \tag{44}$$

Siinä $\hat{\mu} = \bar{X}$.

7.7.2 Normaalijakauman odotusarvon luottamusväli, jos σ^2:sta ei tunneta

Useimmiten varianssia σ^2 ei tunneta. Estimoidaan se kaavalla $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ (jaksot 6.1 ja 6.5). Voidaan osoittaa, että standardoitu tunnusluku

$$\frac{\hat{\mu} - \mu}{s/\sqrt{n}} \sim \mathsf{t}(n-1)$$

noudattaa t-jakaumaa $n-1$ vapausasteella. Jälleen saadaan eksakti yhtäsuuruus

$$\mathsf{P}\left(t_{\alpha/2}(n-1) < \frac{\hat{\mu} - \mu}{s/\sqrt{n}} < t_{1-\alpha/2}(n-1)\right) = 1 - \alpha.$$

Ero aiempiin todennäköisyyslaskuihin on, että yllä satunnaismuuttuja noudattaa $\mathsf{t}(n-1)$-jakaumaa ja että siksi epäyhtälöissä on sen $100(1-\alpha/2)$. persentiilit. Luottamusväliksi μ:lle luottamustasolla $(1-\alpha)$ saadaan

$$\hat{\mu} \pm t_{1-\alpha/2}(n-1)\sqrt{\frac{s^2}{n}}. \tag{45}$$

Esimerkki. Hinta varianssin estimoinnista. Jos varianssi estimoimaan, luottamusväli levenee. Olkoot $X_i \sim \mathsf{N}(\mu,1)$ ja $n = 20$. Jos varianssi tunnetaan, 95 %:n luottamusvälin odotusarvolle rajat ovat

$$\hat{\mu} \pm 1.960\sqrt{\frac{1}{20}} = \hat{\mu} \pm 0.4382613.$$

(kaava (44)). Luottamusvälin leveys on $2 \times 0.4382613 = 0.8765225$.

Estimoidaan varianssi, ja saadaan ihmeen kautta estimaattorin odotusarvo oikea arvo $s^2 = 1$ (jakso 6.1). 95 %:n luottamusvälin odotusarvolle rajat ovat nyt

$$\hat{\mu} \pm 2.093\sqrt{\frac{1}{20}} = \hat{\mu} \pm 0.4680144$$

(kaava (45)). 97.5. persentiili $\mathsf{t}(19)$-jakaumasta 2.093 on laskettu R.n käskyllä `qt(0.975,19)`. Luottamusvälin leveys on $2 \times 0.4680144 = 0.9360288$.

Jälkimmäinen luottamusväli on $0.9360288 - 0.8765225 = 0.0595063$ verran edellistä leveämpi. Se on hinta tietämättömyydestä varianssin suuruudesta eli sen estimoinnista, kun väliestimoidaan odotusarvoa. □

7.7.3 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit yhtäsuuria ja tunnetaan

Kiinnostuksen kohteena on kahden normaalijakautuneen satunnaismuuttujan X_1:n ja X_2:n odotusarvon erotus. Oletetaan idealisoitu tilanne, jossa jakaumien varianssit tunnetaan ja ne ovat samat: $X_{1i} \sim \mathsf{N}(\mu_1,\sigma^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma^2)$. Populaatioista on poimittu n_1:n ja n_2:n suuruiset riippumattomat otokset. Keskiarvojen erotuksen $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2 = \sum_{j=1}^{n} X_{1j}/n_1 - \sum_{j=1}^{n} X_{2j}/n_2$ varianssi on $\sigma^2/n_1 + \sigma^2/n_2 = \sigma^2(1/n_1 + 1/n_2)$. Tällöin

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \mathsf{N}(0,1),$$

ja

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} < z_{1-\alpha/2} \right) = 1 - \alpha.$$

$100(1-\alpha)$ %:n luottamusvälin erotukselle $\mu_1 - \mu_2$ rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

7.7.4 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit erisuuria ja tunnetaan

Oletetaan edellisen jakson tilanne paitsi, että jakaumien varianssit ovat erisuuria: $X_{1i} \sim \mathsf{N}(\mu_1,\sigma_1^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma_2^2)$. Varianssit tunnetaan. Nyt

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathsf{N}(0,1).$$

$100(1-\alpha)$ %:n luottamusvälin erotukselle $\mu_1 - \mu_2$ rajat ovat vastaavasti

$$\hat{\mu}_1 - \hat{\mu}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

7.7.5 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit yhtäsuuria ja tuntemattomia

Palataan yhtäsuurien varianssien tilanteeseen $X_{1i} \sim \mathsf{N}(\mu_1,\sigma^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma^2)$. Uusi realistinen piirre on, että varianssia σ^2 ei tunneta. Estimoidaan se molempien otosten avulla:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{\sum_{j=1}^{n}(X_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(X_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2}.$$

Yllä $s_i^2 = \sum_{j=1}^{n}(X_{ij} - \hat{\mu}_i)^2/(n_i - 1)$. Voidaan osoittaa, että tällöin

$$\mathsf{P}\left(t_{\alpha/2}(n_1 + n_2 - 2) < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{s\sqrt{1/n_1 + 1/n_2}} < t_{1-\alpha/2}(n_1 + n_2 - 2) \right) = 1 - \alpha.$$

Siinä $t_{1-\alpha/2}(n_1+n_2-2)$ on t-jakauman n_1+n_2-2 vapausasteella $100\times(1-\alpha/2)$. persentiili. Luottamustasolla $1-\alpha$ luottamusvälin erotukselle $\mu_1-\mu_2$ rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm t_{1-\alpha/2}(n_1+n_2-2)s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

7.7.6 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit erisuuria ja tuntemattomia

Olkoot $X_{1i} \sim \mathsf{N}(\mu_1,\sigma_1^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma_2^2)$. Realistisin tilanne on, että varianssit σ_1^2 ja σ_2^2 ovat tuntemattomia ja mahdollisesti erisuuria. Ilmeiset estimaatit niille ovat $s_i^2 = \sum_{j=1}^{n}(X_{ij}-\hat{\mu}_i)^2/(n_i-1)$, $i = 1,2$. Toisin kuin muualla jaksossa 7.7, nyt joudutaan tyytymään approksimatiiviseen luottamusväliin. Edellisten jaksojen tapaan muotoillun tunnusluvun

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1-\mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

jakauma riippuu varianssien suhteesta σ_1^2/σ_2^2 sekä otoskoista n_1 ja n_2 (ns. Behrens–Fisher-jakauma). Voidaan osoittaa, että tunnusluku yllä noudattaa approksimatiivisesti t-jakaumaa ν vapausasteella, jossa vapausasteet ν lasketaan kaavalla

$$\nu = \mathsf{int}\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]. \tag{46}$$

$\mathsf{int}[x]$ on argumentin x kokonaislukuosa (esim. $\mathsf{int}[36.51] = 36$).[67] Approksimatiivisen $100(1-\alpha)$ %:n luottamusvälin erotukselle $\mu_1-\mu_2$ rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm t_{1-\alpha/2}(\nu)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}.$$

Jos molemmissa otoksissa havaintoja on enemmän kuin 30, $t_{1-\alpha/2}$:n voi korvata luottamusvälissä $z_{1-\alpha/2}$:lla (Ramachandran ja Tsokos 2015, 290).

[67] Approksimaation nimitys vaihtelee. Dudewicz ja Mishra (1988, 502), Ugarte ym. (2016, 477) sekä Lomax ja Hahs-Vaughn (2012, 173) yhdistävät approksimaation Welchiin. Ramachandran ja Tsokos (2015, 348) puhuvat Smith–Satterthwaite-menettelystä. Armitage ym. (2002, 110) liittävät Welchin hieman erilaiseen jakaumaan ja Satterthwaiten approksimaatioon yllä. Wilcox (2012, 322-324) käyttää molempia nimityksiä ja toteaa, että tässä esitetty menettely on Welchin ja on erikoistapaus Satterthwaiten julkaisemasta.)

7.8 Odotusarvon luottamusväli, jos havaintojen jakaumaa ei tunneta

Olkoot havainnot X_i jatkuva-arvoisia mutteivät välttämättä normaalijakautuneita. Tukeudutaan taas keskeiseen raja-arvolauseeseen. Suurilla havaintomäärillä

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1),$$

vaikka havainnoista ei tiedettäisi juuri muuta kuin, että ne ovat riippumattomia ja että niillä on odotusarvo (μ) ja varianssi (σ^2). Estimoidaan odotusarvo ($\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i/n$) ja varianssi ($s^2$ tai $\hat{\sigma}^2$). Luottamustason $1-\alpha$ luottamusvälin rajat ovat

$$\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\frac{s^2}{n}}.$$

Siinä on käytetty s^2:sta varianssin estimaattorina.

Yllä esitettyyn tapaan muodostettava luottamusväli on kapeampi kuin kaavan (45) mukainen luottamusväli, koska $z_{1-\alpha/2} < t_{1-\alpha/2}(n-1)$. Nyt johdettu luottamusväli pätee, vaikka aineisto olisi normaalijakautunutta. Ei ole mielekästä, että tietämättömyydestä seuraisi kapeampi luottamusväli. Luottamusvälin laskeminen kaavalla (45) on suositeltavampaa, jos havaintojen jakaumaa ei tiedetä. Jaksossa 12.3 pohditaan, milloin tällainen menettely toimii.

8 Otoskoon määrääminen

Moni tutkimus perustuu itsetehtyyn otantaan. Tällöin tulee pohtia, kuinka suuri otos kerätään. Mitä suurempi otos, sitä tarkemmat estimaatit saadaan keskimäärin (ellei tilanne ole aivan poikkeuksellinen). Otanta maksaa (rahaa, aikaa, vaivaa jne.), joten tutkijan täytyy tasapainotella tutkimuksen tarkkuuden ja kustannusten välillä.

Alla pohditaan otoskoon määräämistä, kun etukäteen on päätetty otantavirheen itseisarvon

$$a = \mid \hat{\theta} - \theta \mid$$

suuruus, johon tyydytään tietyllä luottamustasolla. Tässä θ on estimoitava parametri ja $\hat{\theta}$ on sen estimaatti.

8.1 Otoskoon määrääminen odotusarvoa estimoitaessa

Olkoon $100(1-\alpha)$ %:n luottamusväli odotusarvolle μ suurilla havaintomäärillä edeltä tuttua muotoa

$$\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

($\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i/n$). Luottamusvälin pituus halutaan rajata $2a$:n pituiseksi

$$z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} = a$$

luottamustasolla $1-\alpha$. Järjestellään ja neliöidään yhtälö yllä, ja ratkaistaan tarvittava otoskoko n:

$$\frac{\sigma^2}{n} = \left(\frac{a}{z_{1-\alpha/2}}\right)^2 \Leftrightarrow$$
$$n = \left(\frac{z_{1-\alpha/2}\sigma}{a}\right)^2.$$

Otoskoko riippuu keskihajonnasta σ. Jos se tunnetaan (suurinpiirtein), kaavasta yllä voidaan laskea kerättävän otoksen suuruus. Vaikka keskihajontaa ei tunnettaisi, sille saatetaan voida asettaa yläraja. Yläraja sijoitetaan kaavaan, ja ratkaistaan otoskoko.

8.2 Otoskoon määrääminen suhteellista osuutta estimoitaessa

Suhteellista osuutta estimoitaessa $100(1-\alpha)$ %:n luottamusväli on

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

(kaava (40)). Luottamusvälin leveys riippuu nyt kiinnostuksen kohteena olevasta parametrista π:

$$z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = a \Leftrightarrow$$
$$n = \left(\frac{z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})}}{a}\right)^2.$$

Jos π:n suuruudesta ei ole käsitystä tai halutaan varautua estimointitarkkuuden kannalta hankalimpaan tilanteeseen, asetetaan $\pi = 0.5$:

$$n = \left(\frac{z_{1-\alpha/2}\sqrt{0.5 \times 0.5}}{a}\right)^2 = \left(\frac{z_{1-\alpha/2}}{2a}\right)^2.$$

Näin tehdään, koska tulo $\pi(1-\pi)$ on suurimmillaan ja luottamusväli leveimmillään, jos $\pi = 0.5$ (kuva 21). Jos π:n suuruudesta on käsitys, voidaan sen yläraja asettaa kaavaan yllä.

Esimerkki. Gallup (jatkoa). K-puolueen kannatuksen arvioidaan olevan noin 33 % mutta korkeintaan 40 %. Kannatus halutaan selvittää ± 2.5 %-yksikön tarkkuudella luottamustasolla 0.95. Tarvittava otoskoko on noin 1475:

$$n = \left(\frac{1.960\sqrt{0.4(1-0.4)}}{0.025}\right)^2 \approx 1475.$$

Jos tarkkuudeksi edellytetään ±1.25 %-yksikköä, vaaditaan noin 5900 haastattelua:

$$n = \left(\frac{1.960\sqrt{0.4(1-0.4)}}{0.0125} \right)^2 \approx 5900.$$

Luottamusvälin puolittaminen vaatii otoskoon nelinkertaistamisen — kuten jaksossa 7.2 selitettiin. Otannalla saadaan helpohkosti melko osuvia vastauksia, mutta tarkat estimaatit vaativat moninkertaisia otoskokoja. □

9 Populaation koon estimointi

Monista yhteiskunnallisesti merkittävistä tai muuten kiinnostavista asioista ei ole tilastoja. Otantatutkimuksilla voi olla mahdollista selvittää ilmiön yleisyyttä ja sitä kautta populaation kokoa. Joissain asioissa otantatutkimuskaan ei ole mahdollinen.

Esimerkki. Populaation koko. Kiinnostava kysymys voi olla tietyllä poliittisella, psykologisella, sosiaalisella tai taloudellisella tavalla toimivien määrä.

Konkreettisia kuvitteellisia esimerkkejä:

- Isisin kannattajien
- Kansallissosialistisen ideologian kannattajien
- Seksipalveluja myyvien
- Taudinkantajien
- Dopingia harrastavien
- Kodittomien
- Talousrikollisten
- Ylipäänsä sosiaaliselta tai lailliselta kannalta kyseenalaisesti tai arkaluontoisesti toimivien
- Tietynlaisen puhelimen tai muun tuotteen omistajien
- Tietynlaisen harrastuksen omaavien
- Vapaaehtoistyötä tekevien
- Mielenosoitukseen osallistujien lukumäärän
- Ylipäänsä tietyllä tavalla ideologisesti/seksuaalisesti/uskonnollisesti/kiinnostuneesti suuntautuneiden

lukumäärän estimointi.

Näistä joidenkin muttei kaikkien populaatioiden koko saattaa olla otantatutkimuksella selvitettävissä. Harva kertoisi olevansa Isis-mielinen terroristi tai prostituoitu gallup-haastattelijalle, josko haastattelija heitä otokseensa tavoittaisikaan. □

Merkintä-takaisinpyynti-menetelmällä (pyydystys-uudelleenpyydystys, *capture-recapture*, *contact-recontact*, *catch and release*, *mark and recapture*) estimoidaan populaation kokoa. Menetelmän idea esitettiin aikanaan kalakannan suuruuden

arvioimiseksi.[68] Onkiminen toimii edelleen sopivana metaforana menetelmän selittämisessä.

Ongitaan järvestä $K > 0$ kalaa, merkitään ja päästetään ne takaisin vapaaksi. Myöhemmin, kun merkityt kalat ovat sekoittuneet muiden kalojen sekaan, ongitaan $n > 0$ kalaa. Niistä $k > 0$ todetaan merkityiksi. Luonteva ajatus on, että merkittyjen kalojen osuus otoksessa olisi (karkeasti) merkittyjen kalojen osuus järvessä:

$$\frac{k}{n} = \frac{K}{N}.$$

Yhtälöstä saadaan estimaattori kalakannan koolle järvessä:

$$\hat{N} = \frac{Kn}{k}.$$

Estimaattorin $\hat{N}$ varianssi tunnetaan, ja populaation koolle voidaan laskea luottamusväli. Niihin perehtyminen sivuutetaan. Oleellista on tutustua mielenkiintoiseen ideaan ja nähdä sille sovellusmahdollisuuksia.

Esimerkki. Amfetamiinien ja opiaattien ongelmakäyttäjät. Partasen ym:iden (2007) artikkelissa selitetään menetelmän sovellusta:[69]

> – – huumeiden ongelmakäytöllä tarkoitetaan sosiaalisia tai terveydellisiä haittoja aiheuttanutta amfetamiinien tai opiaattien käyttöä, johon viranomaiset ovat puuttuneet – – ja josta on seurannut merkintä viranomaisrekisteriin. – – Tilastolliset arviot tehtiin merkintä-takaisinpyynti-menetelmällä – – , jota on käytetty 1990-luvulta lähtien EU-maiden huumeiden ongelmakäyttöä kartoittavissa tutkimuksissa – – . Aineisto koottiin henkilöistä, jotka oli kirjattu vuonna 2005 johonkin edellä mainituista rekistereistä amfetamiinien tai opiaattien vuoksi. – – Rekistereihin merkittyjen tapausten – – perusteella laadittiin matemaattinen malli, jolla arvioitiin tilastollisesti rekistereihin kirjaamattomien huumeiden ongelmakäyttäjien määrää – – . Yhdistämällä tämä arvio rekisterien sisältämiin tapauksiin saatiin arviot ongelmakäyttäjien kokonaismäärästä. – – Neljän rekisterin perusteella tehdyn tilastollisen arvion mukaan Suomessa oli vuonna 2005 noin 14 500–19 000 amfetamiinien ja opiaattien ongelmakäyttäjää.

Esimerkki. Konsertin yleisömäärä. Wikipedian mukaan maailman suurimmissa konserteissa on ollut 3 000 000 kuulijaa, Helsingissä syksyllä 2016 konsertoivan Jean-Michel Jarren konsertissa 1990 Pariisissa 2 500 000 kuulijaa ja Genesiksen Helsingistä alkaneen 2007-kiertueen Rooman konsertissa 2 000 000 kuulijaa.

[68] Menetelmän kehittäjiksi nimetään järjestään Petersen (1896) ja Lincoln (1930). Pierre-Simon Laplace käytti kuitenkin vastaavaa tekniikkaa jo 1786 Ranskan väkiluvun selvittämiseen (Everitt ja Palmer 2011, 56).

[69] P. Partanen, P. Hakkarainen, A. Hankilanoja, K. Kuussaari, S. Rönkä, M. Salminen, T. Seppälä ja Ari Virtanen (2007): Amfetamiinien ja opiaattien ongelmakäytön yleisyys Suomessa 2005. *Yhteiskuntapolitiikka*, 72, 553–560. Lisää aiheesta: M. Forsell, A. Virtanen, M. Jääskeläinen, H. Alho ja A. Partanen (2010): *Huumetilanne Suomessa 2010*. Raportti 40/2010. Terveyden ja hyvinvoinnin laitos. Yliopistopaino. P. Hakkarainen (2015): Miten tutkia huumetrendejä? Kirjassa A. Häkkinen ja M. Salasuo (toim.): *Salattu, hävetty, vaiettu. Miten tutkia piilossa olevia ilmiöitä*. Nuorisotutkimuseura. Julkaisuja 161. Vastapaino. Tampere. R. King, S.M. Bird, A.M. Overstall, G. Hay ja S.J. Hutchinson (2014): Estimating Prevalence of Injecting Drug Users and Associated Heroin-Related Death Rates in England by Using Regional Data and Incorporating Prior Information. *Journal of the Royal Statistical Society: Series A*, 177, 209–236.

Googlaus tuottaa monta linkkiä (mm. BBC), joissa Genesiksen Rooman konsertin 14.7.2007 kuulijamääräksi todetaan 500 000. Myös Wikipedian Genesiksen 2007-konserttikiertueelle omistetulla sivulla kuulijamääräksi todetaan 500 000.[70] Kuulijamäärän arvioinnissa on valtavia heittoja. Määrää on ollut vaikea arvioida.

Kuulijamäärän objektiivinen estimointi: Konserttiintulijoille jaetaan etukäteen kirkkaanvärisiä T-paitoja, jollaisen käyttämiseen kannustetaan (käyttäjät saavat konserttialueella ilmaisen aterian, paluubussilipun tms.). Otetaan osasta konserttiyleisöä tarkka kuva, ja lasketaan monta kirkasväristä T-paitaa ja ihmistä kuvassa on. Nykytekniikka ehkä mahdollistaisi iltakonserttiin modernimman menettelyn: Kuulijoille jaetaan kertakäyttöisiä digitaalisia lähettimiä. Niiden lukumäärä tunnistetaan konserttialueen tietyllä osalla. Molemmilla tavoilla saataisiin objektiivinen estimaatti yleisön määrästä. Estimaatin tarkkuutta voitaisiin myös arvioida. □

10 Testiteoriaa

10.1 Ilkka Mellinin opetusmonisteen jakso 4.1

- Nollahypoteesi(H_0) ja vastahypoteesi (H_1).
- Merkitsevyystesti ja -taso.
- Yksi- ja kaksisuuntaiset testit.
- Testin hylkäysalue.
- Testin voima.
- Virheet testauksessa.

Mellinin opetusmonisteen jaksosta 4.1 sivuutetaan p-arvo, joka selitetään alla.

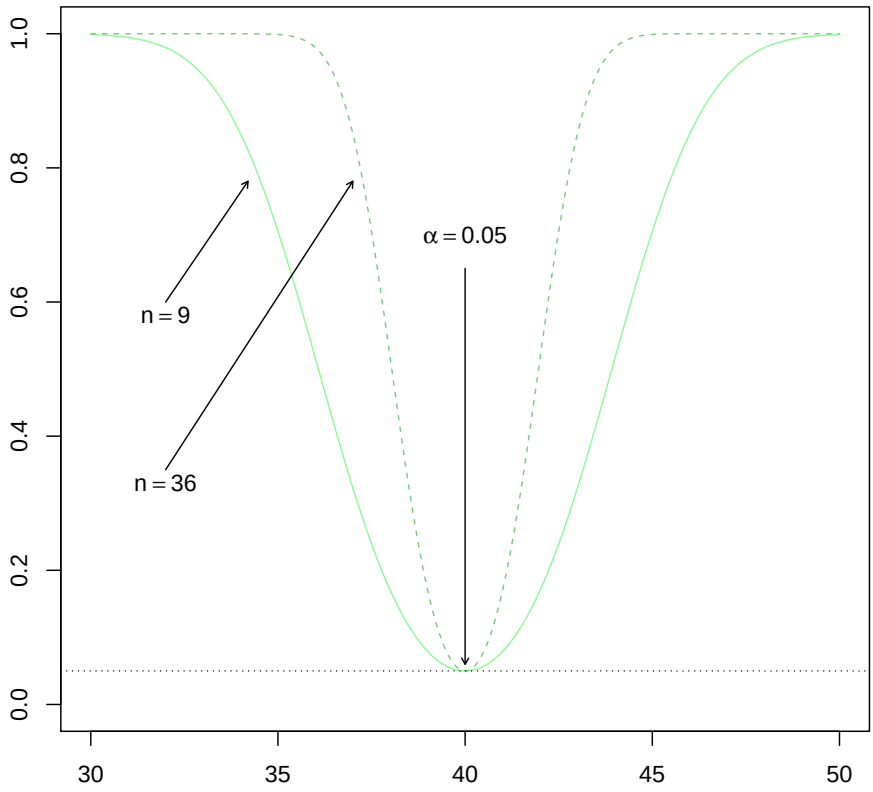
Esimerkki. Testin voima. Kuva 22 havainnollistaa testin voiman kasvamista havaintomäärän kasvaessa.[71] Lasketaan keskiarvo $\hat{\mu}$ Normaalijakaumaa $\mathsf{N}(40{,}36)$ noudattavista satunnaismuuttujista. Testisuure on $Z = (\hat{\mu} - 40)/(\sqrt{36/n})$. Nollahypoteesi $\mu_0 = 40$ hylätään merkitsevyystasolla 0.05, jos testisuureen havaitun arvon itseisarvo $|z|$ ylittää Standardinormaalijakauman 97.5. persentiilin 1.960.

Merkinnällä $n = 9$ — havaintojen lukumäärä — osoitettu käyrä kuvaa, kuinka testin voima kasvaa μ:n etääntyessä 40:stä. Nollahypoteesin $\mu = 40$ ympäristössä voima on jonkin verran suurempi kuin testin koko 0.05. Testin koko lähenee yhtä, kun μ etääntyy 40:stä. Merkinnällä $n = 36$ osoitettu käyrä toistaa samat kuviot mutta kärjekkäämmin, kun havaintoja on 36. Odotusarvon μ poikkeama 40:stä johtaa nyt kauttaaltaan suurempaan testin voimaan kuin silloin, kun havaintoja oli vain 9. Samansuuruinen erotus $\hat{\mu} - 40$ johtaa useammin

[70] `Https://en.wikipedia.org/wiki/List_of_largest_concerts` ja `https://en.wikipedia.org/wiki/Turn_It_On_Again:_The_Tour` (viitattu 4.4.2016).

[71] Kuva pohjautuu R-koodiin 9.1 Ugarten ym:iden (2016) kirjassa.

nollahypoteesin hylkäämiseen, kun $n = 36$ kuin silloin, kun $n = 9$. Testin koko on sama 0.05 molemmilla havaintomäärillä. □



Kuva 22: Kaksisuuntaisen testin voima, kun testisuure on $(\hat{\mu} - 40)/(\sqrt{36/n})$, testin koko on 0.05, havainnot noudattavat jakaumaa $\mathsf{N}(40{,}36)$ ja nollahypoteesi on $\mu_0 = 40$.

10.2 Yksinkertainen ja yhdistetty hypoteesi

Hypoteesi on *yksinkertainen* (*simple*), jos se rajaa jakauman yhdeksi ainoaksi mahdolliseksi. Hypoteesi on *yhdistetty* (*composite*), jos useampi jakauma toteuttaa hypoteesin asettaman ehdon.

Esimerkki. Yksinkertaisia nollahypoteeseja.

- Poisson-jakauman $\mathsf{Poi}(\mu)$ määrittää yksi parametri μ. Nollahypoteesi "jakauma on $\mathsf{Poi}(\mu_0)$" on yksinkertainen (hypoteesi rajaa μ:n arvoksi μ_0:n). Ei ole olemassa muuta Poisson-jakaumaa kuin nollahypoteesin rajaama.
- Binomijakauman $\mathsf{Bin}(n,\pi)$ ainoa parametri on π (otoskokoa n ei lueta parametriksi). Nollahypoteesi "jakauma on $\mathsf{Bin}(n,\pi_0)$" on yksinkertainen.
- Normaalijakauman $\mathsf{N}(\mu,1)$ ainoa vapaa parametri on μ. Nollahypoteesi "jakauma on $\mathsf{N}(\mu_0,1)$" on yksinkertainen.
- Nollahypoteesi, joka rajaa Multinomijakauman $\mathsf{Mul}(n,\pi_1,\ldots,\pi_c)$ parametrit yhtäsuuriksi "$\pi_1 = \ldots = \pi_c = 1/c$" on yksinkertainen. Ei ole olemassa muita Multinomijakaumia, jotka toteuttaisivat hypoteesin asettaman ehdon. □

Esimerkki. Yhdistettyjä hypoteeseja.

- Normaalijakauman $\mathsf{N}(\mu,\sigma^2)$ määrittelee kaksi parametria μ ja σ. Nollahypoteesi "jakauma on $\mathsf{N}(\mu_0,\sigma^2)$" on yhdistetty: On olemassa lukemattomia $\mathsf{N}(\mu_0,\sigma^2)$ jakaumia, jotka ovat hypoteesinmukaisia ($\mathsf{N}(\mu_0,1)$, $\mathsf{N}(\mu_0,1.682)$ jne.).
- Vastahypoteesi "jakauma on $\mathsf{N}(\mu,\sigma^2)$" on yhdistetty: On lukemattomia parametrikombinaatioita (μ,σ^2), jotka voivat päteä Normaalijakaumalle.
- Nollahypoteesi "jakauma on $\mathsf{Mul}(n,\pi_1,\pi_2,\ldots,\pi_c)$, jossa $\pi_1 = \pi_2$" on yhdistetty. Parametrit $\pi_3,\ldots,\pi_c$ voivat saada erilaisia arvoja, vaikka pätisi $\pi_1 = \pi_2$. □

Sekä nolla- että vastahypoteesi voivat olla yksinkertaisia tai yhdistettyjä. Tyypillisesti ainakin vastahypoteesi on yhdistetty. Jos jakauman määrittelee monta parametria, niin usein hypoteesi liittyy tiettyyn parametriin. Sekä nolla- että vastahypoteesi ovat tällöin yhdistettyjä. Jos nollahypoteesi rajaa kaikkien parametrien arvot (ei lainkaan harvinainen tilanne), niin nollahypoteesi on yksinkertainen.

10.3 *P*-arvo

Empiirisessä kvantitatiivisessa tutkimuksessa tyypillisesti raportoidaan testisuureen *havaittu merkitsevyystaso* eli *p-arvo*. On tärkeää ymmärtää, mitä p-arvo tarkoittaa — ja mitä se ei tarkoita. Tutkimusten mukaan monet soveltajat tulkitsevat p-arvon väärin.

Olkoon nollahypoteesi yksinkertainen, testisuure sellainen, että sen suuret arvot ovat nollahypoteesia vastaan (esim. $|T| > 0$, jossa $|T|$ on testisuure) ja testisuureen jakauma symmetrinen. Tällöin p-arvo on todennäköisyys, että testisuure saa arvon, joka on nollahypoteesiin verrattuna yhtä poikkeava tai vielä poikkeavampi kuin havaittu arvo, kun nollahypoteesi pätee. Jos nollahypoteesi on yhdistetty, p-arvo on suurin mahdollinen todennäköisyys (yläraja) tälle tapahtumalle nollahypoteesin pätiessä. Yksinkertaisissa kurssilla esillä olevissa tilanteissa ylärajaa ei tarvitse hakea, vaikka nollahypoteesi olisi yhdistetty.

Pienen p-arvon tulkinta on, että havaitun tapahtuman todennäköisyys on pieni nollahypoteesin pätiessä, eli aineisto on ristiriidassa sen kanssa. Yleinen väärinkäsitys on, että p-arvo olisi todennäköisyys, että nollahypoteesi pätee. P-arvo on kuitenkin nimenomaan nollahypoteesin pätiessä laskettu todennäköisyys.

Esimerkki. P-arvo yhdistetyn nollahypoteesin tilanteessa. Olkoon nollahypoteesi, että normaalijakautuneen satunnaismuuttujan $\mathsf{N}(\mu,\sigma^2)$ odotusarvo on μ_0. Jakauman varianssia ei tunneta, mutta sen tiedetään olevan pienempi kuin 20. Nollahypoteesi on yhdistetty, koska se sallii erilaisia (μ_0,σ^2)-pareja. Olkoon vastahypoteesi yksisuuntainen $\mu > \mu_0$.

Pohditaan testisuuretta

$$\hat{\mu} - \mu_0,$$

jossa $\hat{\mu}$ on havaintojen keskiarvo $\sum_{i=1}^{n} X_i/n$. Erotuksen $\hat{\mu} - \mu_0$ suuret arvot ovat ristiriidassa nollahypoteesin kanssa. Testisuureen varianssi riippuu σ^2:sta: $\mathsf{V}(\hat{\mu}-\mu_0) = \sigma^2/n$ (jaksot 4.1 ja 4.4). Mitä suurempi σ^2, sitä todennäköisempi on suuri testisuureen arvo. P-arvo on yläraja tälle todennäköisyydelle. Se saataisiin laskemalla todennäköisyys $(\hat{\mu} - \mu_0)$:lle olettamalla, että $X_i \sim \mathsf{N}(\mu_0,20)$. Sen pätiessä todennäköisyys havaitulle erolle $\hat{\mu} - \mu_0$ olisi mahdollisimman suuri.

Esimerkin tilanteessa on käytettävissä paljon järkevämpi testisuure

$$\frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} \sim \mathsf{t}(n-1),$$

jonka jakauma ei riipu varianssista σ^2 ($s^2 = \sum_{i=1}^{n}(X_i - \hat{\mu})^2/(n-1)$). P-arvon laskeminen ei vaadi todennäköisyyden ylärajan hakemista, vaikka nollahypoteesi on yhdistetty. Testisuureen p-arvo on todennäköisyys

$$\mathsf{P}\left(T \geq \frac{\hat{\mu} - \mu}{s/\sqrt{n}}\right),$$

jossa T on $\mathsf{t}(n-1)$-jakaumaa noudattava satunnaismuuttuja.

Olkoon $\hat{\mu} = 101.5$, $\mu_0 = 100$, $s^2 = 16$ ja $n = 64$. Tällöin testisuureen

$$\frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} = \frac{101.5 - 100}{4/8} = 3$$

p-arvo on noin 0.002. Se on laskettu R:n käskyllä `1-pt(3,63)`. Testisuureen havaittu arvo on poikkeuksellinen suuri nollahypoteesin pätiessä. Nollahypoteesi hylätään tavanomaisimmilla merkitsevyystasoilla. □

10.4 Luottamusvälien ja testien yhteys

Tarkastellaan kaksisuuntaisen testin riskitasolla 100α % ja samasta testisuureesta muodostetun kaksisuuntaisen $100(1-\alpha)$ %:n luottamusvälin yhteyttä. Oletetaan yksinkertaisuuden ja konkreettisuuden vuoksi, että havainnot noudattavat Normaalijakaumaa $\mathsf{N}(\mu,\sigma^2)$, testataan nollahypoteesia "odotusarvo on μ", havaintojen varianssin estimaatti on $s^2 > 0$, $n > 0$ on otoskoko, testisuure on $(\hat{\mu}-\mu)/(s/\sqrt{n})$ ja sen kriittiset arvot ovat $t_{\alpha/2}(n-1) < 0$, $t_{1-\alpha/2}(n-1) > 0$ ja $t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$. Tällöin

$$
\begin{aligned}
1-\alpha &= \mathsf{P}\left(t_{\alpha/2}(n-1) < \frac{\hat{\mu}-\mu}{s/\sqrt{n}} < t_{1-\alpha/2}(n-1) \right) \\
&= \mathsf{P}\left(t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \hat{\mu}-\mu < t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right) \\
&= \mathsf{P}\left(-\hat{\mu}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < -\mu < -\hat{\mu}+t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right) \\
&= \mathsf{P}\left(\hat{\mu}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} > \mu > \hat{\mu}-t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right) \\
&= \mathsf{P}\left(\hat{\mu}-t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \hat{\mu}+t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right).
\end{aligned}
$$

Ensimmäisen rivin mukaan toisistaan riippumattomissa toistokokeissa testisuureen $(\hat{\mu}-\mu)/(s/\sqrt{n})$ arvo osuu todennäköisyydellä $1-\alpha$ välille $(t_{\alpha/2}(n-1), t_{1-\alpha/2}(n-1))$. Mikäli näin käy, nollahypoteesi jää voimaan. Jos testisuureen arvo sijoittuu välin $(t_{\alpha/2}(n-1), t_{1-\alpha/2}(n-1))$ ulkopuolelle, nollahypoteesi hylätään.

Viimeisen rivin muoto on tuttu luottamusvälin lausekkeista edellä: Toisistaan riippumattomissa toistokokeissa väli

$$
\left(\hat{\mu}-t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \hat{\mu}+t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \right)
$$

peittää odotusarvon μ todennäköisyydellä $1-\alpha$ (kaava (45)).

Jos testisuure sijoittuu kriittisten arvojen väliin, yhtälöketjun ensimmäisellä rivillä olevat epäyhtälöt pätevät, eikä nollahypoteesia hylätä. Tällöin yhtälöketjun viimeisellä rivillä olevat epäyhtälöt pätevät eli luottamusväli peittää μ:n. Jos nollahypoteesia, että odotusarvo on μ ei hylätä, testin riskitasoa 100α % vastaava $100(1-\alpha)$ %:n luottamusväli peittää μ:n.

Jos testisuure ei sijoitu kriittisten arvojen väliin, yhtälöketjun yllä ensimmäisellä rivillä olevat epäyhtälöt eivät päde. Nollahypoteesi hylätään. Tällöin myöskään yhtälöketjun viimeisellä rivillä olevat epäyhtälöt eivät ole voimassa eli luottamusväli ei peitä μ:tä. Jos nollahypoteesi "odotusarvo on μ" hylätään, testin riskitasoa vastaava luottamusväli ei peitä μ:tä.

Esimerkki. P-arvo yhdistetyn nollahypoteesin tilanteessa (jatkoa). Olkoon vastahypoteesi $\mu \neq \mu_0$, ja tehdään kaksisuuntainen testi nollahypoteesille $\mu_0 = 100$

riskitasolla 1 % edellisen esimerkin tilanteessa. Nollahypoteesia ei hylätä, jos

$$-2.656 < \frac{\hat{\mu} - \mu}{s/\sqrt{n}} < 2.656.$$

Siinä $2.656 = t_{0.995}(63)$ on laskettu R-komennolla `qt(0.995,63)`. Testisuure

$$\frac{101.5 - 100}{4/8} = 3$$

ei sijoitu välille $(-2.656, 2.656)$, joten nollahypoteesi hylätään 1 %:n riskitasolla. R-käsky `2*(1-pt(3,63))`laskee testisuureen p-arvoksi 0.004. Näin ollen 99 %:n luottamusväli ei kata nollahypoteesin mukaista odotusarvoa 100:

$$\left(101.5 - 2.656 \times \frac{4}{8}, 101.5 + 2.656 \times \frac{4}{8}\right) = (100.17,102.83)$$

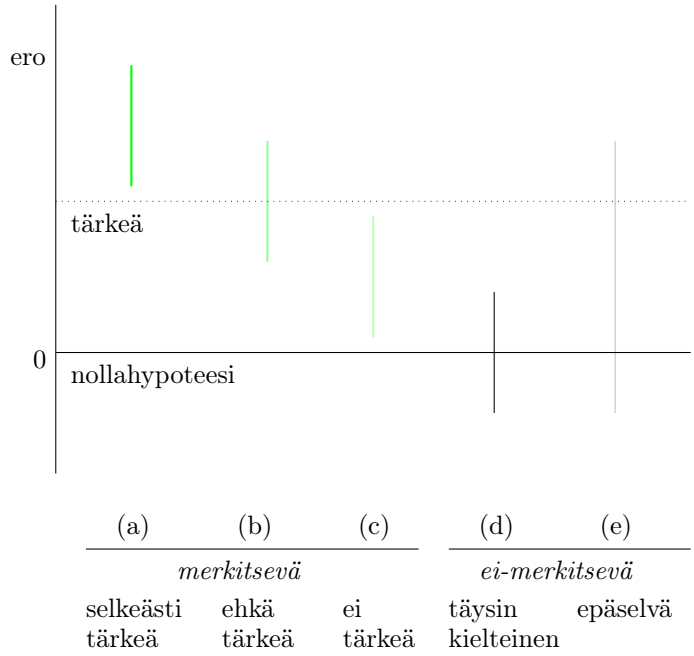
(kaava (40)). P-arvo ei ole kaukana 0.01:stä testin merkitsevyystasosta. Se heijastuu siinä, että luottamusvälin reuna ei ole kaukana nollahypoteesin mukaisesta odotusarvosta (100). □

10.5 Tilastollinen merkitsevyys ja käytännön merkitys

Tilastollinen merkitsevyys ja käytännön merkitys ovat eri asioita. Sovellusten kannalta merkityksetönkin poikkeama nollahypoteesista muodostuu suurilla havaintomäärillä tilastollisesti merkitseväksi, koska testien voima menee kohti yhtä otoskoon kasvaessa. Kuva 23 havainnollistaa. Pystyakselilla on estimoidun parametrin tai estimoitujen parametrien erotuksen ero nollahypoteesin mukaisesta arvosta (tyypillisesti 0).

Tilanteissa (a)–(c) luottamusväli ei peitä nollahypoteesin mukaista arvoa ja ero on tilastollisesti merkitsevä. Ainoastaan tilanteessa (a) on tehty selkeästi tärkeä löydös: Ero on tilastollisesti merkitsevä ja käytännön kannalta suuri. Tilanteessa (b) luottamusväli peittää sekä käytännön kannalta tärkeitä että toisarvoisia eroja. Löydös voi olla tärkeä tai toisarvoinen. Vaikka ero on tilastollisesti merkitsevä, kohdan c) löydös on mielenkiinnoton, koska luottamusväli peittää vain eron arvoja, joilla ei ole käytännön merkitystä. Luottamusväli on mahdollisesti estimoitu suuresta otoksesta, jolloin pienetkin poikkeamat nollahypoteesista tulevat tilastollisesti merkitseviksi.

Tilanteissa (d)–(e) ero ei ole tilastollisesti merkitsevä. Ero on mielenkiinnoton sekä tilastotieteen että sovellusalan näkökulmasta tilanteessa (d). Luottamusväli peittää nollahypoteesin mukaisen arvon eikä yllä käytännön merkitystä omaaviin arvoihin. Löydös on kaikinpuolin negatiivinen. Viimeinen tilanne (e) on moniselitteinen. Luottamusväli on niin leveä, että se peittää eron nollahypoteesiarvon mutta myös käytännön kannalta suuria arvoja. Otoskoko on mahdollisesti ollut pieni, jolloin parametrin tai parametrien estimaatit ovat epätarkkoja.



Kuva 23: Luottamusvälit, tilastollinen merkitsevyys ja käytännön merkitys (Armitage ym. 2002, 92).

11 Testejä

11.1 Testejä suhteellisille osuuksille

11.1.1 Suhteellisen osuuden testi

Estimoidaan suhteellista osuutta $\hat{\pi} = y/n$:llä (tapahtumien ja havaintojen lukumäärien suhde). Olkoon nollahypoteesin mukaan suhteellinen osuus $\pi = \pi_0$. Sen pätevyyttä voidaan koetella testisuureella

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}, \tag{47}$$

joka on suurilla havaintomäärillä standardinormaalijakautunut (jakso 6.3). Jos $\pi_0 = 0.5$, niin approksimaatio toimii, jos havaintoja on ainakin 20. Muuten pitää olla $n\pi_0 \geq 10$ ja $n(1-\pi_0) \geq 10$. (Agresti ja Finlay 2009, 156 ja 172.)

Esimerkki. Oikeuspoliittinen tutkimuslaitos tutki avioeroihin liittyviä riitoja lapsista käräjäoikeuksissa ajalla 14.11.2005–13.2.2006 (529 havaintoa).[72] Keskitytään tutkimaan päätöksiä, joissa lapset määrättiin asumaan vain jomman-

[72] E. Valkama ja M. Litmala (2006): Lasten huoltoriidat käräjäoikeuksissa. OPTL:n jul-

kumman vanhemman luona. Näissä päätöksissä lapsi määrättiin asumaan 35:ssä isän ja 83:ssa äidin luona (118 havaintoa).[73] Vastaavat prosenttiosuudet ovat noin 29.7 ja 70.3.

Testataan kaksisuuntaisesti 1 %:n riskitasolla nollahypoteesia, että käräjäoikeudet määräävät lapset asumaan eri sukupuolta olevien vanhempien luona yhtä todennäköisesti ($\pi_0 = 0.5$). Testisuure on

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.297 - 0.5}{\sqrt{0.5(1-0.5)/118}} = -4.410.$$

Havaintoja on yli 20, joten normaalisuusapproksimaatioehto tilanteessa $\pi_0 = 0.5$ täyttyy. Testisuureen p-arvo saadaan Standardinormaalijakaumasta ja on noin 0.00001 (`2*pnorm(-4.410)`). Nollahypoteesi hylätään 1 %:n riskitasolla. Isät voittavat huoltoriidan 0.5:ttä pienemmällä todennäköisyydellä.

Testin voi tehdä yhtä hyvin vertaamalla 0.5:teen äitien voitto-osuutta 0.703. Testisuureen arvo olisi 4.410. P-arvo, testin tulos ja johtopäätös olisivat samat. □

11.1.2 Suhteellisten osuuksien erotuksen testi, jos osuudet ovat riippumattomia

Testataan, eroavatko suhteelliset osuudet π_1 ja π_2 populaatioissa. Molemmista on n_1:n ja n_2:n kokoiset riippumattomat otokset. Niistä estimoidaan havaitut suhteelliset osuudet $\hat{\pi}_1 = y_1/n_1$ ja $\hat{\pi}_2 = y_2/n_2$, joissa y_i ja n_i ovat tapahtumien ja havaintojen lukumäärät i. otoksessa, $i = 1{,}2$. Testisuure on

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}}. \tag{48}$$

Siinä erotuksen $\hat{\pi}_1 - \hat{\pi}_2$ varianssi $\pi(1-\pi)(1/n_1 + 1/n_2)$ estimoidaan yhdistetystä otoksesta, koska nollahypoteesin mukaan suhteellinen osuus π on sama populaatioissa:

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}.$$

Testisuure on standardinormaalijakautunut, kun havaintoja on paljon. Peukalosääntö riittävälle approksimaatiolle on, että sekä $y_i > 10$ että $n_i - y_i > 10$. Jos

kaisuja 224. `Https://helda.helsinki.fi/handle/10138/152456` (viitattu 7.4.2016).

Tutkimuslaitos pyysi käräjäoikeuksia lähettämään tiedot ratkaisuistaan sitä mukaa kuin riidat ratkaistiin. Tutkimuslaitos sai 565:n riidan asiapaperit, joista puutteellisina tai muusta syystä hylättiin 36 tapausta (n. 6 %). Laitoksen tutkijat arvioivat olleen ilmeistä, että käräjäoikeudet eivät olleet systemaattisesti lähettäneet kaikkia ratkaistuja päätöksiä tutkimuslaitokselle. Vain osa riidoista koski lasten asumista. Aineiston keruutapa ei ole ongelmaton, mm. koska tutkimuksesta kerrottiin etukäteen käräjäoikeuksille. Tutkimuksen tekijät huomauttavat itse muista epävarmuustekijöistä aineiston edustavuudessa mutta pitävät aineistoa kuitenkin melko luotettavana.

[73] Lapset määrättiin asumaan isän luona 27.3 %:ssa, äidin luona 65.2 %:ssa ja 7.5 %:ssa päätöksistä (127 havaintoa) molemmilla (lapset ”jaettiin” tai määrättiin ”vuoroasumisesta”).

testi tehdään kaksisuuntaisena, riittää, että $y_i > 5$ ja $n_i - y_i > 5$. (Agresti ja Finlay 2009, 190.)

Esimerkki. Palo-Repo (2015) tutki Helsingin hovioikeuden 2003–2006 ratkaisemia riitoja lasten huoltajuudesta tai asumisesta.[74] Toisesta vanhemmasta tai hänen uudesta puolisostaan tehdyt syytökset väkivaltaan, päihteisiin tai huumeisiin tai mielenterveysongelmiin liittyen ovat yleisiä näissä riidoissa: Palo-Revon tutkimasta 198 riidasta 94:ssä eli 47.5 %:ssa oli tehty tällainen syytös.

Tutkitaan eroa syytösten toteennäyttämisosuuksissa osa-aineistossa, jossa vain toinen vanhempi tekee syytöksen ja osa-aineistossa, jossa molemmat vanhemmat tekevät syytöksen:

		toteennäytetty (lkm)			toteennäytetty (osuus)		
		kyllä	ei	Σ	kyllä	ei	Σ
syytös	vain toisesta	41	30	71	0.577	0.423	1
	molemmista	16	30	46	0.348	0.652	1
	Σ	57	60	117	0.487	0.513	1

Aineistossa, jossa vain toinen vanhempi teki syytöksen, toteennäytettyjen syytösten osuus on $100 \times 41/71 \approx 57.74648$ %. Aineistossa, jossa molemmat vanhemmat tekivät syytöksen, osuus on $100 \times 16/46 \approx 34.78261$ %. Testataan 5 %:n riskitasolla nollahypoteesia, että syytösten toteennäyttämisen todennäköisyydet eivät eroa, kun vain toinen vanhempi tekee syytöksen tai kun molemmat vanhemmat tekevät syytöksen.

Toteennäytettyjen syytösten osuus yhdistetyssä aineistossa on

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{71 \times 0.5774648 + 46 \times 0.3478261}{71 + 46} = \frac{41 + 16}{117} \approx 0.4871795.$$

Suhteellisten osuuksien erotukseen perustuva testisuure on

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.5774648 - 0.3478261}{\sqrt{0.4871795 \times (1 - 0.4871795) \times (\frac{1}{71} + \frac{1}{46})}} \approx 2.427354.$$

2×2 -aineistotaulukon jokaisessa solussa on yli 5 havaintoa. Testisuureen normaalisuusapproksimaation ehto täyttyy. Testisuureen p-arvo on 0.0152 (`2*(1-pnorm(2.427354))`). Nollahypoteesi hylätään 5 %:n riskitasolla. Oikeus katsoo syytöksen toteennäytetyksi todennäköisemmin silloin, kun vain toinen oikeudenkäynnin osapuolista on tehnyt syytöksen.

[74] M. Palo-Repo (2015): Lasten huolto- ja asumisriidat Helsingin hovioikeudessa 2003–2006. Pro gradu -tutkielma (tilastotiede). Valtiotieteellinen tiedekunta. Helsingin yliopisto. `Https://helda.helsinki.fi/handle/10138/155254` (viitattu 2.1.2016).

11.1.3 Suhteellisten osuuksien erotuksen testi, jos osuudet eivät ole riippumattomia

Palataan tilanteeseen, jossa havainnot noudattavat Multinomijakaumaa $\mathsf{Mul}(1,\pi_{11}, \pi_{12},\pi_{21},\pi_{22})$

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

ja havaittu aineisto on

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

(jakso 7.4). Tällaisen aineiston yhteydessä mielenkiintoinen nollahypoteesi on, päteekö *reunahomogeenisuus* $\pi_{1+} = \pi_{+1}$ ja $\pi_{2+} = \pi_{+2}$. Reunahomogeenisuutta voidaan testata tutkimalla vain toisen yhtälön pitävyyttä, koska reunatodennäköisyydet summautuvat 1:ksi ja yhtälöt seuraavat siksi toisistaan.

Reunahomogeenisuuden pätiessä $\pi_{12} = \pi_{21}$:

$$\pi_{1+} = \pi_{+1} \;\Leftrightarrow\; \pi_{11} + \pi_{12} = \pi_{11} + \pi_{21} \;\Leftrightarrow\; \pi_{12} = \pi_{21}.$$

Tällöin havaitussa aineistossa n_{12}:n ja n_{21}:n tulisi olla satunnaisvaihtelun puitteissa yhtäsuuret.

Ajatellaan $(1,2)$-solun havaintoja "tapahtumina", ja oletetaan, että $n_{12} + n_{21} > 0$. Havaintoja n_{12} vastaava satunnaismuuttuja N_{12} noudattaa Binomijakaumaa $\mathsf{Bin}(n_{12} + n_{21}, 0.5)$, jos nollahypoteesi reunahomogeenisuudesta pätee. Testisuure muodostetaan binomijakautuneen satunnaismuuttujan normaalisuusapproksimaatiosta:

$$\frac{n_{12} - (n_{12} + n_{21}) \times 0.5}{\sqrt{(n_{12} + n_{21}) \times 0.5 \times 0.5}} = \frac{0.5(n_{12} - n_{21})}{0.5\sqrt{n_{12} + n_{21}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

(jakso 4.5.3). Testisuuretta verrataan Standardinormaalijakauman kriittisiin arvoihin. Testiä kutsutaan *McNemarin testiksi*.

Agrestin (2013, 416) mukaan normaalisuusapproksimaatio toimii, jos $n_{12} + n_{21} > 10$. Agresti ja Finlay (2009, 202) antavat tiukemman ehdon $n_{12}+n_{21} > 20$. Fagerlandin. Lydersenin ja Laaken (2013 ja 2014) simulointikokeiden mukaan testin koko on varsin lähellä oikeaa, kun reunahomogeenisuushypoteesiehdon $\pi_{+1} = \pi_{1+}$ reunatodennäköisyydet ovat välillä $0.1 - 0.9$ ja $n_{12} + n_{21} \geq 15$.

Esimerkki. Parisuhdeväkivalta (jatkoa). Testataan 1 %:n riskitasolla, havaitsevatko lapset eri todennäköisyyksillä äitiinsä ja isäänsä kohdistuvaa parisuhdeväkivaltaa. Aineistossa havaitut todennäköisyydet ovat 0.088 ja 0.056 ja havaittujen frekvenssien $n_{12} = 674$ ja $n_{21} = 231$ summa on yli 20 (s. 102). Jakaumaapproksimaatio on käytettävissä. McNemarin testisuure on

$$\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{674 - 231}{\sqrt{674 - 231}} = 21.04757.$$

Sen p-arvo on 0 R:n raportointitarkkuudella (`2*(1-pnorm(21.04757))`). Ero osuuksissa 0.088 ja 0.056 on tilastollisesti merkitsevä. Lapset aistivat enemmän isän äitiin kuin äidin isään kohdistamaa väkivaltaa. □

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Agresti ja Finlay (2009, 202) neuvovat käyttämään Binomijakauman kriittisiä arvoja, jos ehto $n_{12} + n_{21} > 20$ ei täyty. Ohje ei ole hyvä. Fagerland ym. (2013) arvioivat, että tällaisen testin todellinen koko on niin paljon tarkoitettua pienempi ja testi niin heikko, että sitä ei tulisi käyttää koskaan.

Fagerlund ym. (2013) suosittelevat päätekstissä kuvattua normaalisuusapproksimaatioon perustuvan testin käyttöä kaikissa tilanteissa. Fagerlundin ym:iden (2014) tulokset ovat samantapaisia. Testin todellinen koko voi olla hieman suurempi tai selvästi pienempi kuin testin nimellinen koko. Heidän toteamansa suurin positiivinen poikkeama todellisen ja nimellisen koon välillä on 0.0037 (0.0537 − 0.05).

Jos vaaditaan, että testin todellinen koko ei saa ylittää nimellistä kokoa, Fagerlund ym. suosittelevat Binomitestiä keski-p-korjauksella. Sen todellinen koko ei vaikuta koskaan ylittävän nimellistä kokoa.

Jos havaintoja on vähän, todelliset koot voivat olla ovat selvästi liian pieniä käytettäessä normaalisuusapproksimaatiotestiä tai Binomitestiä keski-p-korjauksella. Edellisen koko on tällöinkin lähempänä oikeaa.

Myös Fagerlund ym. (2014) suosittelevat normaalisuusapproksimaatiotestiä tai Binomijakaumaan perustuvaa testiä keski-p-korjauksella.

11.2 χ^2-testejä

Vuonna 1900 Karl Pearson julkaisi χ^2-testin. Sitä on sanottu yhdeksi 20. vuosisadan tärkeimmistä keksinnöistä. Moni opiskelija pukertaa sellaisen kandidaatintutkielmaansa. Samaa testiä käytetään Euroopan hiukkasfysiikan tutkimuskeskus CERNissä vahvistamaan ydinfysiikan viimeisimpiä saavutuksia. Testiä voidaan käyttää monenlaisten hypoteesien testauksessa.

11.2.1 Empiirisen ja teoreettisen jakauman yhteensopivuustesti

Tutkitaan satunnaismuuttujia N_i, jotka noudattavat Multinomijakaumaa $\mathsf{Mul}(n,\pi_1,\ldots,\pi_c)$. Solutodennäköisyydet π_i määräytyvät teoriasta tai hypoteesista, jonka pitävyyttä halutaan koetella. Taustalla voi olla jatkuva jakauma, jonka arvot on luokiteltu. Nollahypoteesi asettaa parametreille arvot $\pi_1 = \pi_{10},\ldots,\pi_c = \pi_{c0}$.

Nollahypoteesin voimassaollessa i. solun odotettu frekvenssi on

$$e_i \equiv \mathsf{E}(N_i) = n\pi_{i0}$$

(jakso 4.2.4). Havaittujen frekvenssien n_i ja odotettujen frekvenssien $n\pi_{i0}$ ei tulisi poiketa suuresti toisistaan nollahypoteesin pätiessä. Testisuure

$$X^2 = \sum_{i=1}^{c} \frac{(N_i - e_i)^2}{e_i} \overset{n \text{ suuri}}{\sim} \chi^2(c-1) \tag{49}$$

perustuu tälle ajatukselle. Mikäli nollahypoteesi ei päde, havaittujen ja odotettujen frekvenssien poikkeamat ja X^2 paisuvat. Nollahypoteesin pätiessä X^2 noudattaa suurilla havaintomäärillä χ^2-jakaumaa $c-1$:llä vapausasteella. Suuret arvot johtavat nollahypoteesin hylkäämiseen. Testiä kutsutaan χ^2-*testiksi*.

Jakauma-approksimaation toimivuudelle on esitetty monia nyrkkisääntöjä kuten, että kaikki odotetut frekvenssit ovat vähintään yksi ($e_i \geq 1$) ja vähintään 80 % niistä on suurempia kuin viisi ($e_i > 5$). Lindgrenin (1976, 424) mukaan approksimaatio on melko hyvä, jos havaintojen lukumäärä on neljä–viisi kertaa solujen lukumäärä ($n > 4 \times c$) vaikka yksittäisiä yhtä pienempiä odotettuja frekvenssejä olisi ($e_i < 1$). Jos sovellettava sääntö ei toteudu, tulee luokkia yhdistää niin, että se toteutuu.

Esimerkki.[75] Poissaolot. Sairastamisen voisi olettaa jakautuvan tasaisesti kaikille viikonpäiville. Yrityksen johto epäilee, että työntekijät ilmoittautuvat sairaaksi muita päiviä useammin viikonloppua ympäröivinä maanantaina ja perjantaina. Johto keräsi poissaolotiedot seuraavilta neljältä viikolta:

ma	ti	ke	to	pe	Σ
49	35	32	39	45	200
40	40	40	40	40	200

Maanantaisin ja perjantaisin on muita päiviä enemmän poissaoloja. Alemmalla rivillä on nollahypoteesin tasaisesti jakautuneista sairaspäivistä mukaiset odotetut frekvenssit $n\pi_{i0} = 200 \times 1/5 = 40$. Ovatko poikkeamat odotetuista frekvenssistä tilastollisesti merkitseviä 5 %:n riskitasolla? Johto laskee salamannopeasti R-käskyillä

```
lkm <- c(49,35,32,39,45)
chisq.test(lkm, p = c(1/5,1/5,1/5,1/5,1/5))
qchisq(0.95,4)
```

testisuureen arvoksi

$$\frac{(49-40)^2}{40} + \cdots + \frac{(45-40)^2}{40} = 4.9$$

ja $\chi^2(4)$-jakauman 95. persentiiliksi 9.49. Jakauman vapausasteet ovat luokkien lukumäärä miinus yksi eli $5-1=4$. Nollahypoteesia ei ole syytä hylätä. Käsky `chisq.test(lkm, p = c(1/5,1/5,1/5,1/5,1/5))` tulostaa testisuureen p-arvoksi $0.298 = \mathsf{P}(X^2 > 4.9)$. Sen saa laskettua myös käskyllä `1-pchisq(4.9,4)`. □

[75] Mellin (1996, 176).

Luokittelu voi vaikuttaa testin tulokseen. Luokittelu kannattaa muodostaa mahdollisimman teräväksi kysymyksenasettelun kannalta, koska se kasvattaa testin voimaa.

Esimerkki. Poissaolot (jatkoa). Poissaolojen epäillään keskittyvän viikonlopun ympärille perjantaille ja maanantaille. Nollahypoteesi diskreetistä tasaisesta jakaumasta on luonteva, mutta aineisto kannattaa luokitella poissaoloihin tiistaista torstaihin ja perjantaista maanantaihin, koska näiden luokkien väliseen eroon yrityksen johdon epäilyt kohdistuvat. Uudelleenluokiteltu aineisto ja vastaavat odotetut frekvenssit ovat:

ti–to	ma ja pe	Σ
106	94	200
120	80	200

Yllä $n\pi_{10} = 200 \times 3/5 = 120$ ja $n\pi_{20} = 200 \times 2/5 = 80$.

χ^2-testisuureen arvo on

$$\frac{(106-120)^2}{120} + \frac{(94-80)^2}{80} = 4.083333.$$

Vapausasteita on $2 - 1 = 1$. Tällöin 5 %:n riskitasoa vastaava kriittinen arvo on 3.841 (`qchisq(0.95,1)`). Testisuureen arvo on sitä suurempi, joten nollahypoteesi hylätään 5 %:n riskitasolla. Testin p-arvo on $\mathsf{P}(X^2 > 4.083333) = 0.04330815$ (`1-pchisq(4.083333,1)`). Testin voi tehdä yhtälailla R-käskyillä alla:

```
lkm <- c(106,94)
chisq.test(lkm, p = c(3/5,2/5))
```

Poissaolojen määrässä on viikonloppuefekti. Mielekkäällä luokittelulla nollahypoteesi hylätään. □

Voidaan osoittaa, että jakauman yhteensopivuus -testisuure on kahden luokan tilanteessa sama testisuure kuin suhteellisen osuuden testisuure (47) neliöitynä:

$$X^2 = \left(\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \right)^2 = z^2$$

(esim. Dudewicz ja Mishra 1988, 529–530). Molemmat noudattavat $\chi^2(1)$-jakaumaa suurilla havaintomäärillä. X^2-testisuure hälyttää yhtälailla liian pienistä kuin suurista frekvensseistä odotettuun verrattuna. Jos testi on perusteltua tehdä yksisuuntaisena, testin voimaa voidaan kasvattaa tekemällä yksisuuntainen testi testisuureella z. Luopumalla voimasta yhteen suuntaan, voitetaan voimaa toiseen suuntaan.

Esimerkki. Poissaolot (jatkoa). Yrityksen johto epäilee ylimääräisiä — ei muita päiviä vähäisempiä — poissaoloja perjantaisin ja maanantaisin. Testi on järkevintä tehdä yksisuuntaisena testisuureella (47).

Olkoon π perjantai- ja maanantaipoissaolojen osuus poissaoloista. Nollahypoteesi on, että $\pi = 2/5 = 0.4$. Vastahypoteesi on, että $\pi > 2/5$. Havaittu osuus on nollahypoteesin määrittämää suurempi:

$$\hat{\pi} = \frac{49 + 45}{200} = 0.47.$$

Testisuure on

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0 \times (1 - \pi_0)/n}} = \frac{0.47 - 0.40}{\sqrt{0.40 \times (1 - 0.40)/200}} = 2.020726.$$

Testisuureen p-arvo on $\mathsf{P}(Z > 2.020726) = 0.0216$ (`1-pnorm(2.020726)`). Nollahypoteesi hylätään 5 % riskitasolla. Poissaolojen osuus 0.47 on suurempi kuin odotettu 0.4. Viikonlopun yhteydessä on keskimääräistä enemmän poissaoloja. Testisuureen p-arvo 0.02165407 on puolet vastaavan X^2-testisuureen p-arvosta 0.04331669. Puolittuminen saatiin aikaiseksi yksisuuntaisella testillä.

Todetaan lopuksi tarkistuksena, että $z^2 = 2.020726^2 = 4.083334 \approx 4.083333 = X^2$. □

Tyypillisempi on tilanne, että vertailujakauman parametrit estimoidaan sovittaen vertailujakauma aineistoon. Jos estimoitavia parametreja on s, niin suurilla havaintomäärillä X^2 noudattaa jakaumaa, jonka kriittiset arvot sijaitsevat $\chi^2(c-1-s)$- ja $\chi^2(c-1)$-jakaumien vastaavien kriittisten arvojen välissä. Estimoitu jakauma sopii aineistoon paremmin kuin todellinen jakauma. Vapausasteet ja kriittiset arvot pienenevät siksi vertailujakaumassa. Jos vertailujakaumana käyttää $\chi^2(c-1-s)$-jakaumaa, niin kriittiset arvot voivat olla liian pienet ja testin koko liian suuri, jolloin nollahypoteesi hylätään liian usein. Approksimaatio

$$X^2 \overset{n \text{ suuri}}{\sim} \chi^2(c-1-s)$$

voi silti olla hyvä. (Tang, He ja Tu 2012, 41 ja 151.)

Esimerkki. Perijättärien hedelmällisyys. Francis Galton pohti syitä, miksi suurmiesten suvut Iso-Britanniassa monesti kuolevat pois eli sukuun ei synny poikaa viemään sukunimeä eteenpäin. Taulukossa on syntyneiden poikien lukumäärä aatelisen miehen avioliitossa luokiteltuna sen mukaan, onko puolisonsa ollut perijätär vai ei. (Tyttöjen lukumäärät eivät ole esimerkissä merkityksellisiä.)[76]

Galton tulkitsi taulukkoa tähän tapaan: Avioliitosta perijättären kanssa syntyy vähemmän poikia kuin muista avioliitoista. Perijättäret ovat siten vähemmän hedelmällisiä kuin muut naiset. Syy hedelmättömyyteen on ilmeinen, koska naisesta tulee perijätär vain perheestä, jossa ei ole poikia (aikakautensa lainsäädännön mukaan), eli hedelmättömyys on perijättärien perinnöllinen ominaisuus.

[76] Taulukko ja tehtävän tiedot ovat Bulmerin (2013, 154–156) kirjasta. Alkuperäislähde olisi Galton (1869).

	avioliittojen lkm, joissa äiti	
syntyneitä poikia	oli perijätär	ei ollut perijätär
0	11	1
1	8	5
2	11	7
3	11	17
4	5	10
5	3	4
6	1	4
7	0	2
>7	0	0
avioliittojen lkm	50	50
poikien lkm (Σ)	104	168
tyttöjen lkm (Σ)	103	142

Sovitetaan $\mathsf{Poi}(\mu_j)$-jakauma poikien lukumäärille perijätär- ja ei-perijätär-avioliitoissa ($j = 1{,}2$). Odotusarvojen (μ_j) estimaatit ovat poikien lukumäärien keskiarvot $\hat{\mu}_1 = 2.08$ ja $\hat{\mu}_2 = 3.36$ (jakso 6.6). Havaitut sekä Poisson-jakaumasta estimoidut pistetodennäköisyydet ja odotetut frekvenssit luokille 1–7 saadaan kaavoista

$$\frac{n_{ij}}{n}, \quad e^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!} \quad \text{ja} \quad ne^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!}.$$

Yllä $n = 50$, n_{ij} on havaittu frekvenssi ja $i = 1, \ldots, 7$ ja $j = 1{,}2$. Esimerkiksi perijätär-aineistossa havaittu ja estimoitu pistetodennäköisyys ja estimoitu odotettu frekvenssi avioliitoille, joissa on täsmälleen yksi poika, ovat

$$\frac{8}{50} = 0.16, \quad e^{-2.08}\frac{2.08^1}{1} \approx 0.2598548 \quad \text{ja} \quad 50e^{-2.08}\frac{2.08^1}{1} \approx 12.99274.$$

R laskee estimoidut pistetodennäköisyydet kätevästi käskyillä `dpois(0:7,2.08)` ja `dpois(0:7,3.36)`. Luokan "> 7 " estimoitu todennäköisyys ryhmälle j on

$$1 - \sum_{i=0}^{7} e^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!}$$

(R-käskyt `1-ppois(7,2.08)` ja `1-ppois(7,3.36)`).

Näin lasketut havaitut ja estimoidut pistetodennäköisyydet sekä havaitut ja odotetut frekvenssit ovat alla:

	havaittu todennäköisyys		estimoitu todennäköisyys	
	perijätär	ei-perijätär	perijätär	ei-perijätär
0	0.22	0.02	0.124930212	0.03473526
1	0.16	0.10	0.259854841	0.11671047
2	0.22	0.14	0.270249035	0.19607359
3	0.22	0.34	0.187372664	0.21960242
4	0.10	0.20	0.097433785	0.18446603
5	0.06	0.08	0.040532455	0.12396117
6	0.02	0.08	0.014051251	0.06941826
7	0.00	0.04	0.004175229	0.03332076
>7	0.00	0.00	0.001400527	0.02171203

	havaittu frekvenssi		estimoitu frekvenssi	
	perijätär	ei-perijätär	perijätär	ei-perijätär
0	11	1	6.2465106	1.736763
1	8	5	12.9927421	5.835524
2	11	7	13.5124518	9.803679
3	11	17	9.3686332	10.980121
4	5	10	4.8716893	9.223302
5	3	4	2.0266227	6.198059
6	1	4	0.7025625	3.470913
7	0	2	0.2087614	1.666038
>7	0	0	0.07002636	1.085602

Testataan χ^2-testillä, kestääkö nollahypoteesi, että aineistot noudattavat estimoituja Poisson-jakaumia. Luokkia on 9, joista $4-5$:ssä havaittu frekvenssi on 5:ttä pienempi. Kummassakaan aineistossa χ^2-jakauma-approksimaation ehto, että 80 % havaituista frekvensseistä tulisi olla viittä suurempia, ei täyty. Yhdistetään perijätär-aineistossa viisi viimeistä luokkaa, jotta approksimaatioehto täyttyy:

	havaittu frekvenssi	estimoitu frekvenssi	estimoitu todennäköisyys
0	11	6.2465106	0.124930212
1	8	12.9927421	0.259854841
2	11	13.5124518	0.270249035
3	11	9.3686332	0.187372664
>3	5	7.879662	0.1575932

Luokkia on nyt 5, joista yhdessäkään odotettu frekvenssi ei ole alle 5. Jakauma-approksimaation ehto toteutuu. Estimoituja parametreja on 1 ($\hat{\mu}_1$). X^2-testisuure noudattaa siten jakaumaa, jonka kriittisten arvojen alaraja on $\chi^2(3)$-jakauman $(5-1-1=3)$ kriittiset arvot.

X^2-testisuure uudelleenluokitellulle perijätär-aineistolle on

$$\frac{(11-6.2465106)^2}{6.2465106}+\cdots+\frac{(5-7.879662)^2}{7.879662}=6.4464$$

Sen p-arvo on 0.092 (R-komento `1-pchisq(6.4464,3)`). Nollahypoteesia, että perijättärien poikien lukumäärä noudattaa Poi(2.08)-jakaumaa, ei hylätä. (P-arvo `1-pchisq(6.4464,4)` olisi 0.168, jos vertailujakaumana käytettäisiin $\chi^2(4)$-jakaumaa.)

Ei-perijätär aineistossa yhdistetään 2 ensimmäistä ja 3 viimeistä luokkaa, minkä jälkeen luokkia on 6. Tällöin $X^2 = 5.2816$. Sitä verrataan $\chi^2(4)$-jakaumaan $(6 - 1 - 1 = 4)$. Testisuureen p-arvo on 0.260 (`1-pchisq(5.2816,4)`). ($\chi^2(5)$-jakaumasta p-arvoksi `1-pchisq(5.2816,5)` saataisiin 0.382.) Nollahypoteesi Poi(3.36)-jakaumasta jää voimaan.

Testien mukaan Poisson-jakauma vaikuttaa kelvolliselta kuvaukselta poikien lukumäärille perijätär- ja ei-perijätär avioliitoissa. □

11.2.2 Empiiristen jakaumien yhteensopivuustesti ja riippumattomuustesti

Aineistona on I:hin luokkaan jaoteltuja havaintoja J-luokkaisesta satunnaismuuttujasta Y ($I \geq 2$ ja $J \geq 2$):

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	n_{11}	$\cdots$	n_{1J}	n_{1+}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	n_{I1}	$\cdots$	n_{IJ}	n_{I+}
	Σ	n_{+1}	$\cdots$	n_{+J}	n

Riviluokittelumuuttujaa merkitään yllä X:llä. Havaintoja on yhteensä n, joka on kiinteä luku.[77] Havaittujen frekvenssien n_{ij} taustalla on satunnaismuuttuja N_{ij}. Tällaista aineistoa kutsutaan frekvenssi- tai kontingenssitaulukoksi tai ristiintaulukoiduksi. On hyvä hahmottaa kaksi tapaa, joilla aineisto on voinut syntyä.

Aineisto on voitu koostaa I:stä riippumattomasta otoksesta multinomijakautuneesta satunnaismuuttujasta Y (*riippumaton multinomiaalinen otanta*). Otosten koot n_{i+} ovat kiinteitä. Taulukon kullakin rivillä Y on jakautunut tavalla, joka saattaa vaihdella luokittelumuuttujan X arvon mukaan. Merkitään Y:n ehdollisia solutodennäköisyyksiä $\pi_{j|i}$:llä. Y:n ehdolliset luokittaiset jakaumat ovat riveillä alla:

[77] Jos n on satunnainen, voidaan päättely ehdollistaa reunafrekvensseille n_{i+}. Näin aineistoa voidaan analysoida kuin se olisi saatu riippumattomalla multinomiaalisella otannalla. Jos kaikki reunafrekvenssit n_{i+} ja n_{+j} ovat kiinteitä tai päättely ehdollistetaan niille, solufrekvenssit noudattavat hypergeometrista jakaumaa. Pienillä havaintomäärillä voi käyttää siihen tukeutuvaa päättelyä kuten keski-p-korjattua Fisherin tarkkaa testiä.

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	$\pi_{1\vert 1}$	$\cdots$	$\pi_{J\vert 1}$	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	$\pi_{1\vert I}$	$\cdots$	$\pi_{J\vert I}$	1

Ehdollisten solutodennäköisyyksien estimaatit ovat $\hat{\pi}_{j|i} = n_{ij}/n_{i+}$ (jakso 6.4).

Vaihtoehtoisesti aineisto on voinut muodostua yhtenä multinomiaalisena otoksena (*multinomiaalinen otanta*), jossa yhden havainnon todennäköisyys osua (i,j)-soluun on π_{ij}:

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	π_{11}	$\cdots$	π_{1J}	π_{1+}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	π_{I1}	$\cdots$	π_{IJ}	π_{I+}
	Σ	π_{+1}	$\cdots$	π_{+J}	1

Tässä molemmat luokittelumuuttujat X ja Y ajatellaan satunnaismuuttujiksi. Kaikki reunafrekvenssit n_{i+} ja n_{+j} ovat satunnaisia. Solutodennäköisyyksien estimaatit ovat $\hat{\pi}_{ij} = n_{ij}/n$ (jakso 6.4).

Tutkijaa kutkuttava hypoteesi voi olla, että riveittäiset jakaumat ovat identtisiä ehdollisten jakaumien taulukossa: $\pi_{j|1} = \cdots = \pi_{j|I}$. Tällöin j. sarakkeen ehdollisten todennäköisyyksien estimaatit ja estimoidut odotetut frekvenssit ovat

$$\hat{\pi}_{j|i0} = \frac{n_{+j}}{n} \quad \text{ja} \quad n_{i+}\hat{\pi}_{j|i0} = \frac{n_{i+}n_{+j}}{n}.$$

Niissä on merkitty alaindeksillä 0 nollahypoteesin pätiessä estimoitua solutodennäköisyyttä.

Jos $J = 2$, X^2 testaa I:n suhteellisen osuuden yhtäsuuruutta. Tällöin X^2 on kahden suhteellisen osuuden yhtäsuuruutta testaavan testisuureen (48) yleistys.

Multinomijakautuneen yhden otoksen tilanteessa monesti kiinnostava hypoteesi on, että satunnaismuuttujat X ja Y ovat riippumattomia. Tällöin olisi $\pi_{ij} = \pi_{i+}\pi_{+j}$ (kaava (8)). Hypoteesin pätiessä estimoitu solutodennäköisyys ja estimoitu odotettu solufrekvenssi ovat

$$\hat{\pi}_{ij0} = \hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n^2} \quad \text{ja} \quad n\hat{\pi}_{ij0} = n\frac{n_{i+}n_{+j}}{n^2} = \frac{n_{i+}n_{+j}}{n}.$$

Odotetut estimoidut solutodennäköisyydet ovat $n_{i+}n_{+j}/n$ molemmilla aineiston muodostumistavoilla. Riveittäisten jakaumien samuutta ja satunnaismuuttujien X ja Y riippumattomuutta voidaan testata samalla χ^2-testisuureella

$$X^2 = \sum_{i=1}^{I}\sum_{i=1}^{J}\frac{(N_{ij} - e_{ij})^2}{e_{ij}} \overset{n \text{ suuri}}{\sim} \chi^2((I-1)(J-1)). \tag{50}$$

Siinä odotetut solufrekvenssit on laskettu kumman vaan edellä kuvatun nollahypoteesin mukaisesti:

$$e_{ij} = \frac{n_{i+}n_{+j}}{n} = n_{i+}\hat{\pi}_{j|i0} = n\hat{\pi}_{ij0}.$$

Vapausasteiden lukumäärän kaavassa (50) voi hahmottaa näin: Riippumattomassa multinomiaalisessa otannassa taulukon kullakin rivillä on $I(J-1)$ vapaata parametria, sillä yksi parametreista määräytyy muista ehdosta $\sum_{j=1}^{J} \pi_{j|i} = 1$. Kukin estimoitava parametri vie yhden vapausasteen. Nollahypoteesin pätiessä on vain yksi estimoitava jakauma (saraketodennäköisyydet) ja estimoitavia parametreja siinä $J-1$. Vapausasteita on

$$I(J-1) - (J-1) = (I-1)(J-1).$$

Multinomiaalisessa otannassa vapaita parametreja on $IJ-1$ — viimeinen parametri määräytyy ehdosta $\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1$. Riippumattomuushypoteesin pätiessä estimoituja parametreja on $I-1+J-1 = I+J-2$, sillä reunatodennäköisyydet π_{i+} ja π_{+j} summautuvat molemmat 1:ksi. Vapausasteita on tällöinkin

$$IJ-1-(I+J-2) = (I-1)(J-1).$$

Vanha ohjenuora (Cochran 1954) on, että jakauma-approksimaatio on toimiva, vaikka yksi odotettu frekvenssi olisi noin 1, jos korkeintaan 20 % muista odotetuista frekvensseistä on alle 5 ja vapausasteita on enemmän kuin 1. Toinen ohje on, että kaikkien sarakkeiden ja rivien odotettujen frekvenssien keskiarvon tulisi olla suurempia kuin 5 ja kaikkien odotettujen frekvenssien olla vähintään 1. Jos $I = J = 2$, niin kaikkien odottujen frekvenssien pitää olla suurempia kuin 5. Agresti (2013, 78) varoittaa, että jos taulukossa on sekä hyvin pieniä että kohtuullisen suuria odotettuja frekvenssejä, niin jakauma-approksimaatio saattaa olla huono. Kaikkia tilanteita kattavaa ohjetta on vaikea luoda.

Esimerkki. Lääketeollisuuden tutkimukset. *Nature*-lehdessä kerrottiin 2013 tilastotieteen väärinkäytöksestä lääketieteellisessä tutkimuksessa, joka johti lääkeyhtiön toimitusjohtajan tuomioon. Myös arkisemmissa medioissa on kyseenalaistettu lääketieteellisten tutkimusten luotettavuutta. Alla on otteita Iltalehden 8.9.2014 ja Helsingin Sanomien 25.5.2015 artikkeleista.[78]

> Professori vertaa lääketeollisuutta järjestäytyneeseen rikollisuuteen. Professori Peter C. Gøtzsche on kohauttanut kirjallaan – – .[79] – – Gøtzsche puhui – – Helsingin yliopistolla – – . – – lääkeyritykset pyrkivät ja useimmiten myös pystyvät kontrolloimaan lääkkeiden kehitystä ja testausta alusta loppuun. – – Joskus julkiset ja riippumattomat rinnakkaistestit on saatu tehtyä, ja ne ovat osoittaneet, ettei lääketehtaiden omiin kokeisiin voi luottaa. – – Niillä on valtava intressi

[78] E. Callaway (2013): Uncertainty on Trial. *Nature*, 502, 3, 17–18. `Http://www.bloomberg.com/news/articles/2013-03-05/ex-intermune-ceo-harkonen-s-wire-fraud-conviction-upheld`, `http://www.iltalehti.fi/ulkomaat/2014090718638829_ul.shtml` ja `http://www.hs.fi/tiede/a1432608772406` (viittaukset 16.4.2016).

[79] P. Gøtzsche (2014): *Tappavat lääkkeet ja järjestäytynyt rikollisuus. Näin lääketeollisuus on turmellut terveydenhoidon.* Sitruuna-kustannus. Kerava. Kirja on Britannian lääkäriliiton palkitsema.

manipuloida tuloksia niin, että tuotteella voitaisiin osoittaa olevan positiivista vaikutusta tai peitellä uuden lääkkeen sivuvaikutuksia. Houkutus vilppiin on liian suuri, kun kukaan ulkopuolinen ei pysty tarkistamaan kokeiden tuloksia ja kun pelissä on helposti miljardien voitot.

Luukatoa lääkitään turhaan — moderneihin sairauksiin tuhlataan aikaa ja rahaa. – – Lääkkeiden vaikutus lonkkamurtumien ehkäisemisessä on niin vähäinen, ettei ehkäisevä lääkehoito ole perusteltua, todetaan British Medical Journalissa tänään keskiviikkona julkaistavassa katsaustutkimuksessa. Sen on tehnyt Helsingin yliopiston professorin Teppo Järvisen johtama kansainvälinen ryhmä. – – katsaus antaa lääketutkimuksista huolestuttavan kuvan. Yli 30 tutkimuksesta vain yksi sai puhtaat paperit kaikkien luotettavuutta mittaavien kriteerien osalta. "Yleistrendi oli, että mitä enemmän tutkimuksessa oli puutteita, sitä varmemmin lääkkeellä havaittiin positiivisia vaikutuksia", sanoo Järvinen. – – BMJ:n katsauksen suurin arvo lienee siinä, että se tekee näkyväksi, miten moderneja sairauksia tehdään. "Niitä tehdään tarkoitushakuisilla tutkimuksilla", sanoo Järvinen. Potilasryhmiä ja aineistoa käsitellään niin, että tulokset saadaan lääkkeen kannalta positiivisiksi.

Ylen uutinen 11.8.2010 on esimerkin varsinainen aihe:

Lääketeollisuuden omat tutkimukset päätyvät positiivisiin tuloksiin selvästi useammin kuin julkisten tai muiden tahojen rahoittamat. – – nyt vinouma havaittiin myös ClinicalTrials-tutkimusrekisterissä, joka perustettiin julkaisuharhan vähentämiseksi.

Julkaisuharha syntyy, kun negatiiviset tutkimustulokset pimitetään ja vain suotuisat havainnot julkaistaan. Näin tutkittava lääke voi näyttää tehokkaammalta ja turvallisemmalta kuin se todellisuudessa on. – – Annals of Internal Medicine -lehdessä julkaistu 546 lääketutkimuksen selvitys paljasti, että 85 prosenttia lääketeollisuuden tutkimuksista päätyi tutkitun lääkkeen kannalta positiiviseen tulokseen, kun niin kävi noin 50 prosentissa viranomaisten rahoittamista. Järjestöjen tai muiden tahojen tutkimuksista 72 prosenttia päätyi suotuisaan tulokseen, mutta osuus suureni selvästi, jos yhtenä rahoittajana oli lääkeyhtiö.

Tutkijat muistuttavat, että julkaisuharha on vain yksi monista seikoista, jotka selittävät rahoittajan ja tulosten yhteyttä. Tuloksia voi muokata itselleen suotuisaksi muun muassa viilaamalla tutkimusasetelmaa tai valitsemalla sopivia potilaita tutkittavaksi – – . Lisäksi lääkeyhtiöt ovat tarkkoja siitä mitä tutkimuksia ne rahoittavat, mikä osaltaan selittää positiivisten tulosten määrää.

Taulukossa alla on osa-aineisto uutisessa viitatusta tutkimuksesta.[80] Lääketeollisuuden tutkimukset päätyivät positiiviseen tulokseen lääkkeen vaikutuksesta $100 \times 188/220 \approx 85.4$ %:ssa tutkimuksista: Viranomaisten tutkimuksissa osuus oli vain $100 \times 18/36 = 50.0$ %. Tutkitaan χ^2-testillä, eroavatko osuudet 0.854 ja 0.5 tilastollisesti merkitsevästi.

Muodostetaan taulukot havaituista (n_{ij}) ja estimoiduista odotetuista (e_{ij}) solufrekvensseistä:

	tulos		
rahoitus	+	−	Σ
lääketeollisuudelta	188	32	220
viranomaisilta	18	18	36
Σ	206	50	256

[80] `Http://yle.fi/uutiset/laakefirmojen_tutkimukset_rahoittajilleen_myonteisia/1894597`. F.T. Bourgeois, S. Murthy ja K.D. Mandl (2010): Outcome Reporting Among Drug Trials Registered in ClinicalTrials.gov. *Annals of Internal Medicine*, 153, 158–167.

	tulos		
rahoitus	+	−	Σ
lääketeollisuudelta	177.03125	42.96875	220
viranomaisilta	28.96875	7.03125	36
Σ	206	50	256

Estimoidut odotetut frekvenssit on laskettu kaavalla

$$e_{ij} = \frac{n_{i+}n_{+j}}{n}.$$

Esimerkiksi $e_{11} = 220 \times 206/256 = 177.03125$. Testisuure on

$$\begin{aligned} X^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx \frac{(188 - 177.03125)^2}{177.03125} + \cdots + \frac{(18 - 7.03125)^2}{7.03125} \\ &\approx 24.744. \end{aligned}$$

Testisuure noudattaa suurissa otoksissa $\chi^2((2-1)(2-1))$- eli $\chi^2(1)$-jakaumaa. Sen kriittinen arvo 0.1 %:n riskitasolla on 10.828 (`qchisq(0.999,1)`). Testisuureen p-arvo on noin 6.54×10^{-7} (`1-pchisq(24.744,1)`). Nollahypoteesi hylätään 0.1 %:n riskitasolla, sillä $24.744 > 10.828$. Testi on nopeasti tehty R-komennoilla alla.

```
taulukko <- matrix(c(188,18,32,18),nrow=2)
chisq.test(taulukko, correct=F)
chisq.test(taulukko) $expected
```

Kolmas rivi raportoi ylle kirjatut estimoidut odotetut frekvenssit.

Ero on tilastollisesti merkitsevä. Lääketeollisuuden tutkimukset päätyvät positiiviseen tulokseen useammin kuin viranomaisten rahoittamat tutkimukset. □

Voidaan osoittaa, että 2×2-taulukon tilanteessa X^2-testisuure on sama kuin testisuure (48) neliöitynä:

$$X^2 = \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})} = z^2.$$

Molemmat noudattavat $\chi^2(1)$-jakaumaa suurissa otoksissa. X^2-testisuureella on voimaa, poikkeavat suhteelliset osuudet suuntaan tai toiseen toisistaan. Jos poikkeama vain toiseen suuntaan on mielekäs, testi kannattaa tehdä yksisuuntaisena testisuureella z.

Esimerkki. Lääketeollisuuden tutkimukset (jatkoa). Lääketeollisuuden rahoittamissa tutkimuksissa ei ole syytä epäillä, että lääkkeet osoittautuisivat keskimääräistä harvemmin toimiviksi. Testi on perusteltua tehdä yksisuuntaisena.

Mielletään aineisto kerätyksi erikseen lääketeollisuuden tutkimuksista ja viranomaisten tekemistä tutkimuksista (riippumaton multinomiaalinen otanta tai ehdollistetaan analyysi rivifrekvensseille), eli verrataan ehdollisia jakaumia:

	tulos		
rahoitus	+	−	Σ
lääketeollisuudelta	0.8545455	0.1454545	1
viranomaisilta	0.5	0.5	1
Σ	0.8046875	0.1953125	1

Nollahypoteesin mukainen positiivisen tutkimustuloksen tuottaneiden tutkimusten osuus on

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{188 + 18}{220 + 36} = 0.8046875.$$

Testisuure on

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.8545455 - 0.5}{\sqrt{0.8046875 \times 0.1953125 \times (\frac{1}{220} + \frac{1}{36})}} \approx 4.974345.$$

Testisuure noudattaa standardinormaalijakaumaa suurilla havaintomäärillä. Kriittinen arvo yksisuuntaisessa testauksessa 1 %:n riskitasolla on 2.326 (`qnorm(0.99)`). Havaittu testisuure on kriittistä arvoa suurempi, joten nollahypoteesi hylätään 1 %:n riskitasolla. Testisuureen p-arvo on puolittunut noin 3.27×10^{-7}:ksi (`1-pnorm(4.974345)`). Lääketeollisuus saa lääkkeen toimivuudesta positiivisen tuloksen useammin kuin viranomaiset.

Tarkistus: $z^2 = 4.974345^2 \approx 24.744 = X^2$. Tässä lasketun testisuureen neliö on edellisessä esimerkissä laskettu X^2-testisuure. □

11.3 Testejä havaintojen ollessa normaalijakautuneita

Testit seuraavat suoraviivaisesti vastaavista luottamusväleistä jaksossa 7.7. Alla kuvataan kaksisuuntaiset testit $100 \times \alpha$ %:n riskitasolla. Yksisuuntaiset testit voidaan muodostaa vastaavalla tavalla kuin aiemmissa yhteyksissä. Yksinkertaisuuden vuoksi varienssien vertailutesti kuvataan alla yksisuuntaisena (jakso 11.3.9).

11.3.1 Testi Normaalijakauman odotusarvolle, jos σ^2 tunnetaan

Jos $X_i \sim \mathsf{N}(\mu,\sigma^2)$, σ^2 tunnetaan ja nollahypoteesin mukaan $\mu = \mu_0$, niin

$$\frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \tag{51}$$

ja

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

Yllä $\hat{\mu} = \sum_{i=1}^{n} X_i/n$. Nollahypoteesi hylätään, jos testisuure (51) osuu hylkäysalueelle eli $\mid (\hat{\mu} - \mu_0)/(\sigma/\sqrt{n} \mid > z_{1-\alpha/2}$.

Esimerkki. R:n `rnorm`-käskyn luotettavuus. R:llä voi tuottaa näennäisesti satunnaisia lukuja Standardinormaalijakaumasta `rnorm`-käskyllä. Luvut ovat näennäisesti satunnaisia, koska antamalla sama siemenluku (*seed*), R tuottaa aina samat luvut. Alla oleva koodi tuottaa 100 000 näennäisesti satunnaista lukua Normaalijakaumasta. Testataan, imitoiko `rnorm`-käsky hyvin Normaalijakaumaa, jonka odotusarvo on $\mu_0 = 1$ (nollahypoteesi). Koodi alla asettaa lisäksi Normaalijakauman keskihajonnaksi 1, tuottaa näennäiset satunnaisluvut ja laskee niiden keskiarvoksi 0.9968141:

```
set.seed(21042016) # Luku suluissa on mielivaltainen siemenluku.
x <- rnorm(n=100000, mean=1, sd=1)
x
mean(x)
```

Testisuureen arvo on

$$\frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.9968141 - 1}{1/\sqrt{100000}} = -1.007481.$$

Se on $\mathsf{N}(0,1)$-jakauman 15.68518. persentiili (`pnorm(-1.007481)`). Kaksisuuntaisessa testauksessa p-arvo on noin 0.314. Ei ole syytä hylätä nollahypoteesia, että `rnorm`-komennolla tuotetut luvut imitoivat satunnaismuuttujaa, jonka odotusarvo on 1. □

11.3.2 Testi Normaalijakauman odotusarvolle, jos σ^2:sta ei tunneta

Jos varianssia σ^2 ei tunneta, kaavaa (51) vastaava testisuure on

$$\frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} \sim \mathsf{t}(n-1)$$

($s^2 = \sum_{i=1}^{n}(X_i - \hat{\mu})^2/(n-1)$). Pätee

$$\mathsf{P}\left(t_{\alpha/2}(n-1) < \frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} < t_{1-\alpha/2}(n-1)\right) = 1 - \alpha.$$

Nollahypoteesi H_0: $\mu = \mu_0$ hylätään, jos $|\ (\hat{\mu} - \mu_0)/(s/\sqrt{n}\ | > t_{1-\alpha/2}(n-1)$.

11.3.3 Testi Normaalijakaumien odotusarvojen erotukselle, jos vaarianssit yhtäsuuria ja tunnetaan

Oletetaan, että $X_{1i} \sim \mathsf{N}(\mu_1,\sigma^2)$, $X_{2i} \sim \mathsf{N}(\mu_2,\sigma^2)$ ja että niiden yhteinen varianssi σ^2 on tiedossa. Nollahypoteesi on H_0: $\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$, jossa alaindeksillä on merkitty nollahypoteesin mukaista arvoa odotusarvojen erotukselle (tyypillisesti 0). Koska

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sigma\sqrt{1/n_1 + 1/n_2}} < z_{1-\alpha/2}\right) = 1 - \alpha,$$

niin nollahypoteesi hylätään, jos $|\ [\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0]/(\sigma\sqrt{1/n_1 + 1/n_2})\ | > z_{1-\alpha/2}$.

11.3.4 Testi Normaalijakaumien odotusarvojen erotukselle, jos varianssit erisuuria ja tunnetaan

Olkoot satunnaismuuttujien $X_{1i} \sim \mathsf{N}(\mu_1,\sigma_1^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma_2^2)$ varianssit erisuuria ja tiedossa. Yhtälöstä

$$\mathsf{P}\left(z_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

seuraa, että H_0: $\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ hylätään, jos $|\ [\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0]/(\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})\ | > z_{1-\alpha/2}$.

11.3.5 Testi Normaalijakaumien odotusarvojen erotukselle, jos varianssit yhtäsuuria ja tuntemattomia

Jos tiedetään, että satunnaismuuttujien $X_{1i} \sim \mathsf{N}(\mu_1,\sigma^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma^2)$ varianssit ovat yhtäsuuria, niin varianssi estimoidaan kaavalla

$$s^2 = \frac{\sum_{j=1}^{n}(X_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n}(X_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2}.$$

Jos testisuureen

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{s\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2) \tag{52}$$

itseisarvoksi tulee $t_{1-\alpha/2}(n_1 + n_2 - 2)$:tä suurempi arvo, H_0: $\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ hylätään.

Testiä tulee käyttää vain, jos on selkeä peruste, että varianssit ovat yhtäsuuria. Ei ole harvinaista, että ensin testataan, ovatko varianssit yhtäsuuria (jakso 11.3.9) ja jos nollahypoteesi varianssien yhtäsuuruudesta jää voimaan, jatketaan testaamaan odotusarvojen yhtäsuuruutta tässä esitetyllä tavalla. Tällainen menettely ei ole suositeltava (Wilcox 2012, 319).

11.3.6 Testi Normaalijakaumien odotusarvojen erotukselle, jos varianssit erisuuria ja tuntemattomia

Jos $X_{1i} \sim \mathsf{N}(\mu_1,\sigma_1^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma_2^2)$ ja varianssit ovat (mahdollisesti) erisuuria ja tuntemattomia, testin nollahypoteesille "$\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$" koko on vain osapuilleen $1 - \alpha$. Testisuureen

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

jakauma riippuu suhteesta σ_1^2/σ_2^2 sekä otoskoista n_1 ja n_2. Jakauma on approksimatiivisesti t-jakauma ν vapausasteella, jossa vapausasteet ν määritellään kaavalla (46). Testiä kutsutaan *Welchin* tai *Satterthwaiten testiksi.* Jos molemmissa

otoksissa havaintoja on enemmän kuin 30, vertailujakaumana voidaan käyttää Standardinormaalijakaumaa (Ramachandran ja Tsopos 2015, 345–346).

Esimerkki. Kännykkään puhuminen ja reaktioaika.[81] Koehenkilöt ($n_1 = n_2 = 32$) ajoivat autosimulaattoria. Simulaattorissa välähti sattumanvaraisesti välillä punainen ja välillä vihreä valo. Koehenkilöiden tuli painaa jarrupoljinta heti nähtyään punaisen valon. Ensimmäisessä kokeessa koehenkilöt puhuivat puhelimessa ajaessaan. Toisessa kokeessa he kuuntelivat radio-ohjelmaa tai äänitallennekirjaa kun ajoivat. Kokeissa kirjattiin koehenkilöiden reaktioajat punaisen valon välähdykseen millisekunneissa. Reaktioaikojen ero kullakin henkilöllä laskettiin. Keskimääräinen reaktioaika ensimmäisessä kokeessa 585.1875 oli pidempi kuin toisessa kokeessa 534.5625. Vastaavat varianssit olivat 8036.415 ja 4415.286. Testattava nollahypoteesi on H_0: $\mu_1 - \mu_2 = 0$, ja testisuure on

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{585.1875 - 534.5625}{\sqrt{8036.415/32 + 4415.286/32}} = 2.566408.$$

Vertailujakauma on t-jakauma vapausasteilla

$$\nu = \mathsf{int}\left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}\right] = \mathsf{int}\left[\frac{\left(\frac{8036.415}{15} + \frac{4415.286/32}{15}\right)^2}{\frac{\left(\frac{8036.415}{32}\right)^2}{32 - 1} + \frac{\left(\frac{4415.286/32}{15}\right)^2}{32 - 1}}\right]$$
$$= \mathsf{int}[57.16537] = 57.$$

Testisuureen p-arvo noin 0.013 saadaan t(57)-jakaumasta (`2*(1-pt(2.566408, 57))`). Jakson 11.3.6 peukalosäännön mukaan vertailujakaumaksi voi ottaa Standardinormaalijakauman, jos otosten koot ovat yli 30. Sääntö täyttyy esimerkissä. Standardinormaalijakaumasta laskettu p-arvo 0.010 on varsin sama kuin juuri laskettu. Sääntö vaikuttaa toimivan kohtuullisesti. P-arvot ovat evidenssiä nollahypoteesia vastaan. Kännykkään puhuminen vaikuttaa pidentävän reaktioaikaa.

Huom! Otosten koehenkilöt olivat samoja, eli otokset eivät olleet riippumattomia eikä käytetty testi sopiva. Laskut edellä vain havainnollistavat testin käyttöä. Jaksossa 11.3.7 sovelletaan tällaiseen testausasetelmaan sopivaa testiä esimerkin aineistoon. □

Esimerkki. Poikien ja tyttöjen suorituserot. Lapsiasiavaltuutetun vuosikirjasta 2014:[82]

[81] Agresti ja Finlay (2009, 195–196).

[82] Lapsiasiavaltuutetun vuosikirja 2014. Eriarvoistuva lapsuus. Lasten hyvinvointi kansallisten indikaattorien valossa. Lapsiasiavaltuutetun toimiston julkaisuja 2014:3. `Http://lapsiasia.fi/wp-content/uploads/2014/12/Vuosikirja-2014.pdf` (viitattu 28.3.2015). S. 81.

> Valtaosa suomalaislapsista selviytyy PISA-lukutaitotestistä hyvin, mutta viime vuosina aiempaa suurempi osa 15-vuotiaista on saanut testistä heikkoa lukutaitoa ilmentävän pistemäärän. Etenkin pojista yhä suurempi osa suoriutuu testistä heikosti. Vuonna 2012 pojista 18 prosenttia oli luokiteltavissa heikkoihin lukijoihin. Tytöistä heikosti luki viisi prosenttia. Vuosituhannen vaihteessa vastaavat prosenttiosuudet olivat pojilla yksitoista ja tytöillä kolme. Myös muissa maissa tytöt suoriutuvat lukutaitotestistä poikia paremmin, mutta Suomessa sukupuolten välinen ero on OECD-maiden suurin. Lukutaidon eriarvoisuus näyttää lisääntyneen, sillä lukutaitotestipistemäärän aiemmin alhainen keskihajonta vastaa Suomessa nyt OECD-maiden keskiarvoa.

PISA-tutkimuksessa (*Programme for International Student Assessment*) verrataan 15-vuotiaiden koulutaitoja OECD-maissa. Taulukossa on suomalaisten poikien ($n_1 = 2954$) ja tyttöjen ($n_2 = 2856$) matematiikan ja lukemisen koepistemäärien keskiarvot ja -hajonnat vuoden 2012 tutkimuksessa.[83] Oletetaan, että pistemäärät ovat normaalijakautuneita.

	keskiarvo		otoskeskihajonta	
aine/sukupuoli	poika	tyttö	poika	tyttö
matematiikka	541.79	539.21	80.00	73.02
lukeminen	508.39	563.48	83.96	72.45

Testataan 0.5 %:n riskitasolla (kaksisuuntainen testi) nollahypoteeseja, että matematiikan ja lukemisen kokeiden pistemäärien odotusarvot ovat pojilla ja tytöillä samat (kummankin aineen kohdalla nollahypoteesi on H_0: $\mu_1 - \mu_2 = 0$).

Matematiikan pistemäärille testisuure on

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{541.79 - 539.21}{\sqrt{\frac{80.00^2}{2954} + \frac{73.02^2}{2856}}} \approx 1.284636.$$

Otoskoot ovat niin suuria, että vertailujakaumana voidaan käyttää Standardinormaalijakaumaa. Kaksisuuntaisessa testissä 0.5 %:n riskitasolla kriittiset arvot Standardinormaalijakaumasta ovat -2.807 (`qnorm(0.0025)`) ja 2.807. Koska $1.285 < 2.807$, niin nollahypoteesia ei hylätä. Testisuureen p-arvo on noin 0.199 (`2*(1-pnorm(1.284636))`). Matematiikan kokeen pistemäärän ero on pieni aineistossa eikä ole tilastollisesti merkitsevä. Nollahypoteesia samoista odotusarvoista ei hylätä. Tämä on esimerkki kuvan 23 täysin kielteisestä tuloksesta.

Lukemispistemääriin sovellettuna testisuure saa arvon -26.804:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{508.39 - 563.48}{\sqrt{\frac{83.96^2}{2954} + \frac{72.45^2}{2856}}} \approx -26.804.$$

Sen p-arvo on 3×10^{-158} (`2*pnorm(-26.804)`). Lukemiskokeen pistemäärän odotusarvo on pojilla pienempi kuin tytöillä. Ero on suuri sekä poikien ja tyttöjen

[83] Kiitän professori emeritus Seppo Laaksosta PISA-tunnuslukujen laskemisesta 24.4.2013. Professori emeritus Laaksonen on korjannut aineistoa niin, että sitä voidaan pitkälti käsitellä niin kuin se olisi tehty yksinkertaisella satunnaisotannalla, vaikka PISA-tutkimus tehdään monimutkaisemmalla otantamenetelmällä.

osaamisen kannalta että tilastollisesti. Tämä on esimerkki kuvan 23 selkeästi tärkeästä tuloksesta. □

11.3.7 Testi Normaalijakaumien odotusarvojen erotukselle, jos havainnot pareittaisia

Edellisissä jaksoissa verrattiin kahden ryhmän odotusarvoja. Kummassakin otoksessa oli satunnaisvaihtelua. Se johtui osin siitä, mitä havaintoa otokseen oli tullut. Suurissa ryhmissä sattuman vaikutus häviää ja toisen havainnon poikkeuksellisuus kumoutuu toisen havainnon poikkeuksellisuudella toiseen suuntaan. Pienillä havaintomäärillä niin ei välttämättä käy.

Jos aineisto voidaan muodostaa kaltaisista pareista (*matched pairs*), sattuman vaikutusta voidaan yleensä pienentää ja testin voimaa kasvattaa. Edelleen oletetaan havaintojen normaalisuus molemmissa tutkittavissa populaatioissa: $X_{1i} \sim \mathsf{N}(\mu_1,\sigma_1^2)$ ja $X_{2i} \sim \mathsf{N}(\mu_2,\sigma_2^2)$, $i = 1,\ldots,n$. Kukin i. havainto ryhmissä 1 ja 2 kytkeytyvät toisiinsa, niin, että niitä on luontevaa ajatella pareina (X_{1i},X_{2i}). Odotusarvojen vertailutesti muodostetaan erotuksille

$$D_i = X_{1i} - X_{2i}.$$

Nollahypoteesin pätiessä erotusten odotusarvo on μ_{D0} — tyypillisesti 0. Testaustilanne on jaksossa 11.3.2 kuvatunlainen mutta pitäen tutkittavina satunnaismuuttujina D_i:tä. Testisuure on

$$\frac{\hat{\mu}_D - \mu_{D0}}{s_D/\sqrt{n}} \sim \mathsf{t}(n-1), \tag{53}$$

jossa $s_D^2 = \sum_{i=1}^n (D_i - \hat{\mu}_D)^2/(n-1)$ ja $\hat{\mu}_D = \sum_{i=1}^n D_i/n$. Testiä kutsutaan *parittaisten erojen t-testiksi.*

Havaintojen X_{1i} ja X_{2i} ollessa riippumattomia erotuksen $X_{1i} - X_{2i}$ varianssi on $\mathsf{V}(X_1) + \mathsf{V}(X_2)$ (jakso 4.1). Voidaan osoittaa, että muulloin erotuksen varianssi on

$$\mathsf{V}(X_1) + \mathsf{V}(X_2) - 2\mathsf{C}(X_1,X_2), \tag{54}$$

jossa $\mathsf{C}(X_1,X_2)$ on X_1:n ja X_2:n kovarianssi eli korrelaatio kerrottuna muuttujien keskihajonnoilla. Kaltaisille pareille havaintojen korrelaatio on yleensä positiivinen. Tällöin erotuksen D_i:n varianssi on pienempi kuin riippumattomien havaintojen tilanteessa. Pienempi varianssi johtuu havaintoja sitovien yhteisten tekijöiden kumoutumisesta erotuksessa. Intuitiivisesti pienemmän varianssin tulisi johtaa voimakkaampaan testiin.

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Parittaisten erojen t-testiä voi käyttää, vaikka otokset olisivat riippumattomia, jos niissä on yhtä paljon havaintoja. Nollahypoteesin odotusarvojen yhtäsuuruudesta pätiessä erotuksen odotusarvo olisi 0 ja testi tulokseen (53) perustuen mahdollinen. Järkevämpää on käyttää otosten riippumattomuuden olettavaa testisuuretta (52) (tai versiota, joka sallii erisuuret varianssit populaatioissa). Se noudattaa nollahypoteesin pätiessä $\mathsf{t}(n+n-2)$-jakaumaa. Sen

vapausasteet ovat kaksinkertaiset parittaisten erojen t-testisuureen jakaumaan $\mathrm{t}(n-1)$ verrattuna. Pienemmät vapausasteet tarkoittavat, että testisuureen varianssi on suurempi, joten parittaisten erojen t-testi on ilmeisesti heikompi.

Jos otokset on mahdollista tuottaa omalla koejärjestelyllä, riittäisikö parien komponenttien pienikin positiivinen korrelaatio motivoimaan parittaisen koejärjestelyn ja parittaisten erojen t-testin käytön? Jos parien havaintojen korrelaatio olisi hyvin pieni, oltaisiin lähellä edellä kuvattua tilannetta, jossa parittaisten erojen t-testi on heikompi. Parittaiseen koejärjestelyyn ei ilmeisesti kannattaisi lähteä. Karkea sääntö on, että parittaisten erojen t-testi on voimakkaampi, jos parien havaintojen korrelaatio on suurempi kuin 0.25 (Wilcox 2012, 402). Tällöin kannattaisi koejärjestely muotoilla kaltaisiksi pareiksi.

Parittaisten erojen t-testi voi olla heikompi kuin vastaava riippumattomuuden olettava testi. Niin käy, jos havaintojen korrelaatio on negatiivinen, jolloin erotuksen $X_{1i}-X_{2i}$ varianssi on suurempi kuin riippumattomassa tilanteessa (kaava (54)). Tällaiseen koejärjestelyyn ei kannata ryhtyä. Jos jo olemassaoleva aineisto tiedetään tämäntapaiseksi, täytyy pareittaisten erojen t-testiä käyttää, koska riippumattomat otokset olettava testi ei ole käyttökelpoinen.

Esimerkki. Lääkkeen teho. Uuden lääkkeen tehoa voitaisiin tutkia muodostamalla kaksi ryhmää (mieluusti arpomalla henkilöt ryhmiin), joista toinen saisi uutta lääkettä ja toinen vanhaa lääkettä. Lääkkeellä voitaisiin esimerkiksi pyrkiä pienentämään verenpainetautia sairastavien verenpainetta. Tutkimusjakson jälkeen mitattaisiin ryhmistä, kuinka hyvin lääke on toiminut. Keskiarvojen vertailutestillä (jakso 11.3.6) pääteltäisiin, onko lääkkeiden tehossa eroa.

Voimakkaampaan testiin päästäisiin, jos lääkkeen tehoa verrattaisiin samojen ihmisten välillä. Koehenkilöt käyttäisivät ensin vanhaa lääkettä ja sen jälkeen uutta lääkettä. Mikäli uuden lääkkeen käytön jälkeen verenpaine olisi keskimäärin laskenut, se viittaisi uuden lääkkeen olevan vanhaa tehokkaampi. Tällaisen testin pitäisi olla voimakkaampi kuin edellä kuvatun, koska verenpaineiden koehenkilökohtaisessa vertailussa eliminoituu monia sekoittavia tekijöitä ensimmäiseen tutkimusasetelmaan verrattuna. Jos koehenkilö syö paljon verenpainetta nostavaa suolaa, testin tulokseen vaikuttaa, kumpaan koeryhmään hän päätyy ensimmäisessä tutkimusjärjestelyssä (runsas suolan käyttö kenties kumoaa lääkkeen vaikutuksen), mikä kasvattaa tutkittavan erotuksen varianssia. Toisessa tutkimusjärjestelyssä ero lääkkeiden toimivuuden välillä selviää, vaikka koehenkilö olisi erityisen suolaisen ruoan ystävä. □

Esimerkki. Kaksoskokeet. Tekemällä toiselle identtiselle kaksoselle toinen tutkimus/operaatio tms. ja toiselle toinen, voidaan eliminoida geneettisten tekijöiden sattumanvarainen vaikutus tutkimuksessa. Tutkimus perustuu vertailuun tuloksessa kunkin kaksosparin välillä.

Esimerkki. Kännykkään puhuminen ja reaktioaika (jatkoa). Koehenkilöt olivat samoja jakson 11.3.6 esimerkin molemmissa kokeissa. Aineisto tulisi analysoida pareittaisena. Reaktioaika ja niiden erotus kullakin koehenkilöllä on taulukoitu alla. Koehenkilö 28:n reaktioajat ovat olleet ylivoimaisesti hitaimmat. Reaktioaikojen erotus on hänellä silti vasta 3. suurin. Pareittainen vertailu on palauttanut koehenkilön tulokset samaan suuruusluokkaan muiden kanssa.

koehenk.	1	2	3	4	5	6	7	8	9	10	11	12
känny	636	623	615	672	601	600	542	554	543	520	609	559
radio	604	556	540	522	459	544	513	470	556	531	599	537
erotus	32	67	75	150	142	56	29	84	-13	-11	10	22
koehenk.	13	14	15	16	17	18	19	20	21	22	23	24
känny	595	565	573	554	626	501	574	468	578	560	525	647
radio	619	536	554	467	525	508	529	470	512	487	515	499
erotus	-24	29	19	87	101	-7	45	-2	66	73	10	148
koehenk.	25	26	27	28	29	30	31	32				
känny	456	688	679	960	558	482	527	536				
radio	448	558	589	814	519	462	521	543				
erotus	8	130	90	146	39	20	6	-7				

R-koodi alla laskee erotuksen keskiarvon 50.625 ja otoskeskihajonnan 52.48579:

```
x1 <- c(636,623,615,672,601,600,542,554,543,520,609,559,595,565,573,554,
        626,501,574,468,578,560,525,647,456,688,679,960,558,482,527,536)
x2 <- c(604,556,540,522,459,544,513,470,556,531,599,537,619,536,554,467,
        525,508,529,470,512,487,515,499,448,558,589,814,519,462,521,543)
d <- x1-x2
mean(d)
sqrt(var(d))
```

Testisuureen arvo on

$$\frac{\hat{\mu}_D}{s_D/\sqrt{n}} = \frac{50.625}{52.48579/\sqrt{32}} = 5.456301.$$

Sitä verrataan t(31)-jakaumaan. Testisuureen p-arvo on noin 6×10^{-6} (`2*(1-pt(5.5,31))`). Nollahypoteesi hylätään kaikilla tavanomaisilla riskitasoilla. Puhelimeen puhuminen hidastaa reaktioaikaa enemmän kuin radion kuuntelu.

P-arvo pareittaisten erotusten t-testissä on pienempi kuin samasta aineistosta (oletusten vastaisesti) laskettu odotusarvojen erotuksen testisuureen p-arvo. Se on intuitiivista, sillä pareittaisessa vertailussa karsiutuu satunnaistekijöitä pois ja testin voiman kasvaminen on odotettua.

Testi on nyt tehty oikein. Agresti ja Finlay (2009, 196) varoittavat, että aineistosta voidaan silti tehdä vain tunnustelevia päätelmiä, koska se on ilmeisesti kerätty itsevalikoituneella otannalla. □

11.3.8 Varianssin testaus

Jaksossa 6.5 todettiin, että

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

Nollahypoteesi H_0: $\sigma^2 = \sigma_0^2$ hylätään, jos $(n-1)s^2/\sigma_0^2 < \chi^2_{\alpha/2}(n-1)$ tai $(n-1)s^2/\sigma_0^2 > \chi^2_{1-\alpha/2}(n-1)$.

Oletus normaalisuudesta on tärkeä. Jos havainnot eivät ole normaalijakautuneita, testin koko voi poiketa paljon α:sta.

Esimerkki. R:n `rnorm`-käskyn luotettavuus (jatkoa). Luodaan 100 000 näennäissatunnaislukua koodilla alla. Tutkitaan, kestääkö nollahypoteesi, että luvut olisivat (Normaali)jakaumasta, jonka varianssi $\sigma_0^2 = 1$. Koodi tulostaa otosvarianssiksi 1.007285:

```
set.seed(21042016)
x <- rnorm(n=100000, mean=1, sd=1)
var(x)
```

Testisuureen arvo on

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{99999 \times 1.007285}{1} = 100727.5.$$

Se on $\chi^2(99\ 999)$-jakauman 94.8082. persentiili (`pchisq(100727.5,99999)`).

Yksisuuntaisessa testauksessa (pidettäessä 1:stä suurempia arvoja mahdollisina) testisuureen p-arvo olisi 0.051918 (`(1-pchisq(100727.5,99999))`). Jos jakauma on epäsymmetrinen, ei ole yksikäsitteistä, miten p-arvo tulisi määritellä kaksisuuntaisessa testauksessa. χ^2-jakauma on näin suurilla vapausasteilla hyvin symmetrinen, joten kerrotaan yksisuuntaisen testin testisuureen p-arvo kahdella. P-arvoksi saadaan näin 0.103836.

Jakauman symmetrisyyden tarkistusta: $\chi^2(99999)$-jakauman 5. ja 95. persentiili ovat 99264.54 ja 100735.7 (`qchisq(0.05,99999)` ja `qchisq(0.95,99999)`). Ne ovat yhtä kaukana 99999:stä ($99999 - 99264.54 = 736.7325 = 100735.7 - 99999$). Jakauma vaikuttaa symmetriseltä.

Otosvarianssin poikkeama teoreettisesta varianssista ei anna aihetta hylätä nollahypoteesia. R:n `rnorm`-käsky vaikuttaa toimivan tarkoitetulla tavalla varianssilla mitattuna. □

11.3.9 Kahden varianssin testaus

Edellistä tyypillisempi testaustilanne on kahden varianssin yhtäsuuruuden testaus. Edelleen oletetaan havaintojen normaalisuus. Se on oleellinen oletus alla olevan jakaumateorian pätemiselle.

On laskettu n_1:n ja n_2:n kokoisista otoksista otosvarianssit $s_1^2 > s_2^2$. Nollahypoteesi on, että niitä vastaavat varianssit populaatiossa ovat samat (H_0: $\sigma_1^2 = \sigma_2^2$). Testisuure s_1^2/s_2^2 noudattaa F-jakaumaa vapausasteilla $n_1 - 1$ ja $n_2 - 1$:

$$\frac{s_1^2}{s_2^2} \sim \mathsf{F}(n_1 - 1, n_2 - 1).$$

Suuret testisuureen arvot johtavat nollahypoteesin hylkäämiseen.

Esimerkki. Poikien ja tyttöjen suorituserot (jatkoa). Harvardin yliopiston taloustieteen professori (1983–2007), Maailman pankin pääekonomisti (1991–1993), Yhdysvaltojen valtiovarainministeri (1999–2001), Harvardin yliopiston rehtori

(2001–2006), presidentti Barack Obaman nimittämän kansallisen talousneuvoston johtaja 2009–2010 Lawrence Summers käyttää ajoittain terävämpää kieltä kuin tieteessä on tavallista (esim. Summers 1991). Vuonna 2005 hän pohti tieteellisessä seminaarissa, miksi naisia on vähän huippuyliopistoissa ja -tutkimuslaitoksissa. Summers esitti selityksenä, että monet ominaisuudet kuten pituus, paino, taipumus rikollisuuteen, älykkyysosamäärä, matemaattinen lahjakkuus ja tieteellinen kyvykkyys vaihtelevat miehillä enemmän kuin naisilla ja että pienetkin erot keskihajonnoissa sukupuolten välillä johtavat suuriin eroihin sukupuolten välillä poikkeuksellisen lahjakkaiden yksilöiden lukumäärissä. Summersin argumentointi johti kohuun, syytöksiin seksismistä ja tutkijan huolimattomuudesta ja hän joutui pyytämään anteeksi sanomaansa. Summers erosi rehtorin tehtävästään seuraavana vuonna jouduttuaan konfliktiin yliopiston muun henkilökunnan kanssa ja mahdollisesti myös seksismisyytösten johdosta. Summers on sittemmin kritisoinut ”absurdia poliittista korrektiutta” ja ”totalitarismin hivuttautumista yliopistoihin” mielessä, mistä yliopistoissa on sallittua keskustella.[84]

Eroavatko suomalaisten poikien ja tyttöjen matematiikan ja lukemisen koepistemäärien varianssit vuoden 2012 PISA-tutkimuksessa? Testataan nollahypoteesia varianssien yhtäsuuruudesta H_0: $\sigma_1^2 = \sigma_2^2$. Testisuureen arvo koepistemäärien variansseille matematiikassa on 1.200:

$$\frac{s_1^2}{s_2^2} = \frac{80.00^2}{73.02^2} \approx 1.200318.$$

F-jakauman vapausasteilla $n_1 - 1 = 2953$ ja $n_2 - 1 = 2855$ kriittinen arvo 0.1 %:n riskitasolla on 1.122 (`qf(0.999,2953,2855)`). Testisuureen p-arvo on noin kuuden desimaalin tarkkuudella nolla (`1-pf(1.1200318,2953,2855)`). Koska $1.200 > 1.122$, niin nollahypoteesi hylätään 0.1 %:n riskitasolla. Matematiikan koepistemäärän varianssi on suurempi pojilla kuin tytöillä.

Lukemisen otosvariansseista laskettuna testisuure saa arvon 1.343:

$$\frac{s_1^2}{s_2^2} = \frac{83.96^2}{72.45^2} \approx 1.342975.$$

Sen p-arvo on 1×10^{-15} (`1-pf(1.342975,2953,2855)`). Nollahypoteesi varianssien yhtäsuuruudesta kaatuu 0.1 %:n riskitasolla. Myös lukemisen koepistemäärän varianssi on suurempi pojilla kuin tytöillä.

Summersin johtopäätöksiin ei ole vielä syytä edetä. On mahdollista, että poikien pistemäärien suurempi varianssi johtuisi erityisen heikosti pärjäävien poikien osapopulaatiosta. Normaalisuusoletus ei tällöin pätisi pojille. Aineisto ei ole käytettävissä, joten asiaa ei voida tässä tutkia enempää. □

11.3.10 Jos havainnot eivät ole normaalijakautuneita

Jos havainnot ovat jatkuva-arvoisia mutteivät noudata Normaalijakaumaa, monia edellä esitetyistä testeistä voidaan käyttää. Keskeisen raja-arvolauseen (jak-

[84] `Https://en.wikipedia.org/wiki/Lawrence_Summers` (viitattu 21.4.2016) ja Wainer (2007).

so 4.4) perusteella

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1),$$

vaikka havainnoista ei tiedettäisi juuri muuta kuin, että ne ovat riippumattomia ja että niillä on odotusarvo ja varianssi. Yllä voidaan varianssi korvata estimaatilla s^2. Agresti ja Finlay (2009, 122, 155 ja 196) mainitsevat otoskoon 30 karkeana rajana, jota suuremmilla havaintomäärillä kaksisuuntainen testaaminen on mahdollista Keskeiseen raja-arvolauseen perustuen. Jos havaintojen jakauma on hyvin vino tai aineistossa on erityisen poikkeavia havaintoja (*outliers*), niin yksisuuntainen testi on epäluotetettava. Kahden odotusarvon erotusta testattaessa pätevät samantapaiset huomiot.

Tämäntapaiset huomautukset koskevat myös luottamusvälejä. Esimerkiksi kaksisuuntaisten luottamusvälien peittävyys voi olla lähellä tarkoitettua, vaikka havainnot eivät olisi normaalijakautuneita. Agrestin ja Finlayn (2009, 122) mukaan $n > 15$ voi riittää siihen.

Vaikka tässä esityksessä ei ole mahdollista tutustua yksityiskohtaisesti esimerkkiaineistoihin, todellisessa tutkimuksessa aineistoihin tulisi perehtyä huolellisesti ennen testaamista tai muuta tilastollista päättelyä. Yksi ilmeinen perehdyttävä asia on havaintojen jakauma ja mahdollinen normaalisuus.

Viivojen välinen jakso ei kuulu kurssivaatimuksiin.

Wilcox (2012, jaksot 5.5 ja 8.2) tekee varoittavia simulointeja. Jos havainnot tulevat Normaalijakaumaa pitkähäntäisemmästä symmetrisestä jakaumasta, odotusarvon t-testin (jakso 11.3.2) koko voi olla melko suurillakin havaintomäärillä tarkoitettua pienempi, vaikka jakauma saattaa silmämääräisesti vaikuttaa normaaliselta. Jos jakauma on vino, testin koko voi olla pienempi tai suurempi kuin tarkoitettu. Molemmissa tilanteissa odotusarvon t-testin voima voi heikentyä suuresti. Odotusarvojen erotusta testattaessa t-testin (jakso 11.3.5) koko on varsin oikea, vaikka havainnot eivät olisi normaalijakautuneita, jos ne ovat peräisin samasta vaikka vinosta jakaumasta. Melko pienikin havaintojen varianssia kasvattava poikkeama normaalisuudesta voi heikentää testiä suuresti symmetristenkin jakaumien tilanteessa. Welchin testin (jakso 11.3.6) koko kestää melko hyvin, vaikka jakaumat eivät olisi normaalisia mutta ovat identtisiä vaikka vinojakin. Jos jakaumat eroavat, Welchin testi voi toimia huonosti. (Mt:n jakso 8.3.3.) Wilcox (mt:n jaksot 4.6 ja 5.5) osoittaa huolestuttavia esimerkkejä myös normaalisuuteen perustuvista luottamusväleistä, jos normaalisuus ei päde.

---

12 Regressio

12.1 Regressio kohti odotusarvoa

Francis Galtonin ensimmäinen regressio vuodelta 1877 on kuvassa 24 (”herneen siemen -vanhemmat” ja ”herneen siemen -jälkipolvi”).[85] Oleellisesti sama ilmiö on kuvassa 25 — ilmeisesti toisessa koskaan tehdyssä regressiossa (Pearson 1930,

[85] Kuvio on Gillhamin (2009) artikkelista. Kuvion alkuperäislähde on Pearson (1920). Kuvio löytyy myös Pearsonin (1930) kirjasta. Regressiosuora on Pearsonin uusiksi laskema ja ilmeisesti hänen apulaisensa (A. Davinin) piirtämä Galtonin muistiinpanojen vuodelta 1875 avulla. (Pearson 1920, 34 ja 1930, 4.)

13) — jossa on Galtonin vuonna 1886 havaitsema vastaava yhteys vanhempien pituuksien painotetun keskiarvon (*mid-parent*; ”keskivanhempi”) ja heidän lastensa pituuden välillä.[86] Kuviosta nähdään, että vanhempien keskipituus on ollut keskimäärin runsas 68 tuumaa. Keskivanhempi-suora kuvaa vanhempien keskipituuden poikkeamaa keskimääräisestä pituudesta (suoran kulmakerroin on yksi). Kuvion mukaan

- keskimääräistä pidempien vanhempien lapsi on myös keskimääräistä pidempi muttei yhtä paljon kuin vanhempansa (suoran ”children” kulmakerroin on 0:n ja 1:n välillä).
- keskimääräistä lyhyempien vanhempien lapsi on myös keskimääräistä lyhyempi muttei yhtä paljon kuin vanhempansa.
- pituus regressoituu (palautuu, taantuu) eli pyrkii palaamaan kohti odotusarvoansa (yllä runsas 68 tuumaa). (*Regression toward the mean* tai *regression to the mean.*) Pitkien vanhempien lapset ovat keskimääräistä pidempiä ja lyhyempien vanhempien lapset keskimääräistä lyhyempiä, mutteivät yhtä paljon pidempiä tai lyhyempiä keskipituuteen nähden kuin vanhempansa.

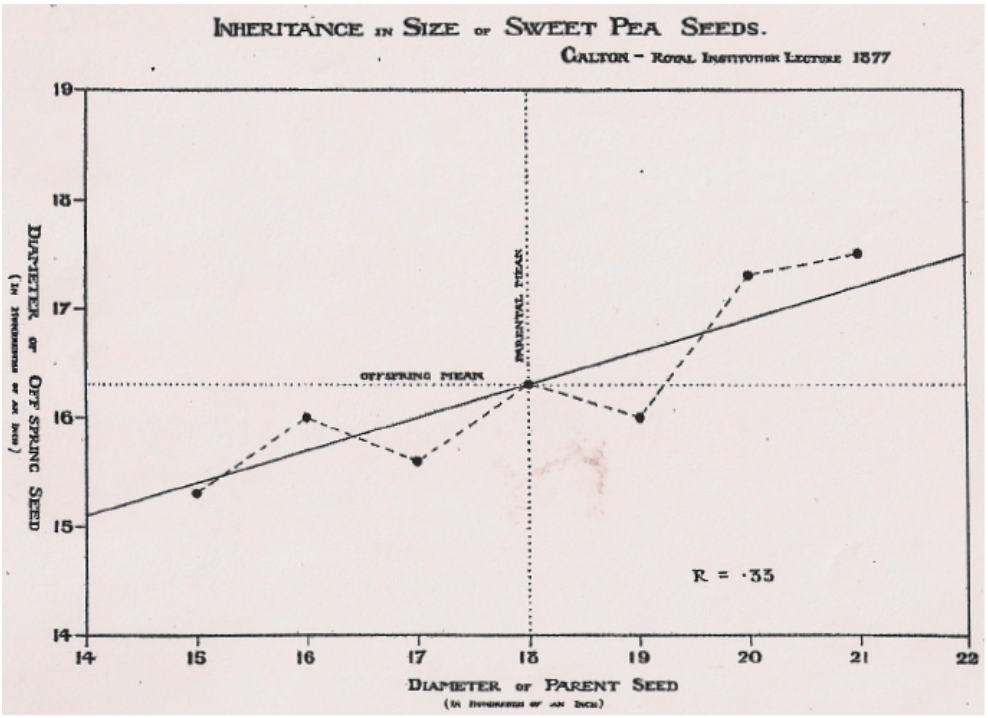
Mieleen saattaisi tulla — kuten Galtonille aikoinaan — että regressiosta keskipituutta kohti seuraisi sukupolvi sukupolvelta pituuden vaihtelun pieneneminen niin, että lopulta kaikki olisivat keskipituisia. Niin ei käy, koska lasten pituuksissa on aina sattumanvaraisuutta, vaikka lasten pituus keskimäärin regressoituukin vanhempiensa pituudesta. Galton havainnollisti asiaa kuvalla 26 vuonna 1901.[87] Galtonin keskivanhempi-käsitettä käytetään kasvututkimuksessa edelleen (esim. Saari ym. 2012).

Regressiossa odotusarvoa kohti on yleisemmin kyse kahdesta samoinjakautuneesta muuttujasta, joiden yhteyden summeeraavan suoran kulmakerroin on alle yhden. Ääritilanteessa muuttujien välillä ei ole mitään yhteyttä: Kuvitteelliseen kuvioon piirretyn summeeraavan sovitteen kulmakerroin on nolla, ja poikkeamat odotusarvosta pyrkivät keskimäärin ”korjaantumaan” täysin seuraavassa havainnossa. Nykypäivään ja yhteiskuntatieteisiin liittyviä esimerkkejä on helppo keksiä.

Esimerkki. Sosiaalitukien määrä. Verrataan sosiaalitukia saavien (tai rikosten, avioliittojen, syntyneiden lasten jne.) lukumäärää suomalaisissa kaupungeissa vuosina 2017 (y-akseli) ja 2016 (x-akseli). Tällöin poikkeuksellisen suuri sosiaalitukea saavien määrä tietyssä kaupungissa tasoittuu lähemmäksi odotusarvoa

[86] Keskivanhempikäsitteen määritelmä löytyy esimerkiksi Wikipediasta (`http://en.wikipedia.org/wiki/Midparent`; viitattu 23.4.2016). Kuvio on Pearsonin (1930, 16) kirjasta. Alunperin kuvio on julkaistu Galtonin (1886) artikkelissa. Tällaisia kuvia on myös Galtonin (1889, 96 ja 107) kirjassa.

[87] Kuvio on artikkelista W. Gilchrist (2012): Galton — A Victorian Worth Celebrating, Significance Web Exclusive. `Http://www.significancemagazine.org/details/webexclusive/1497449/Galton---A-Victorian-worth-celebrating.html` (viitattu 26.1.2014).



Kuva 24: Galtonin ensimmäinen regressio 1877.

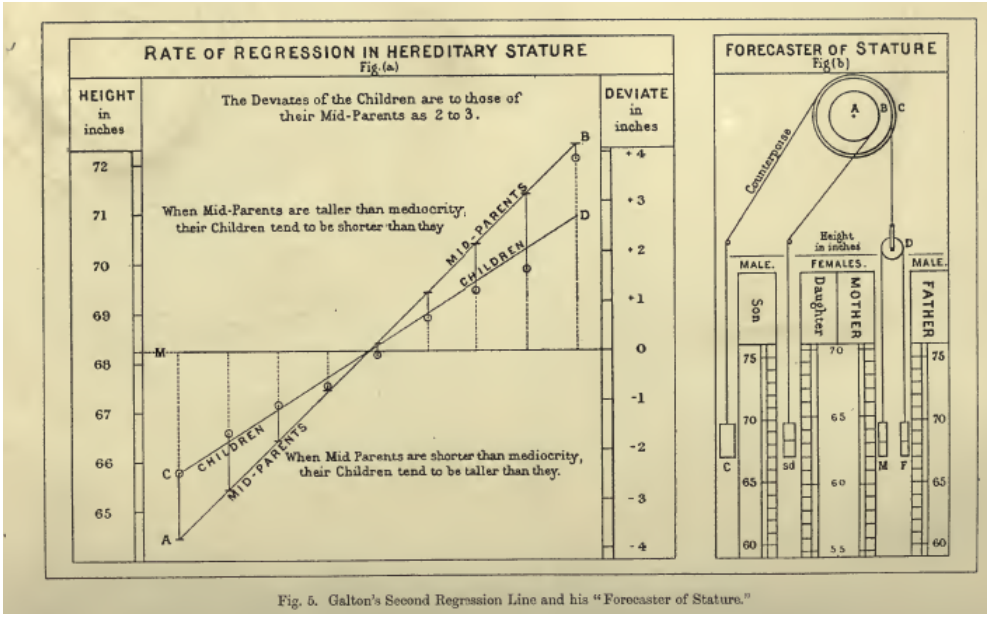
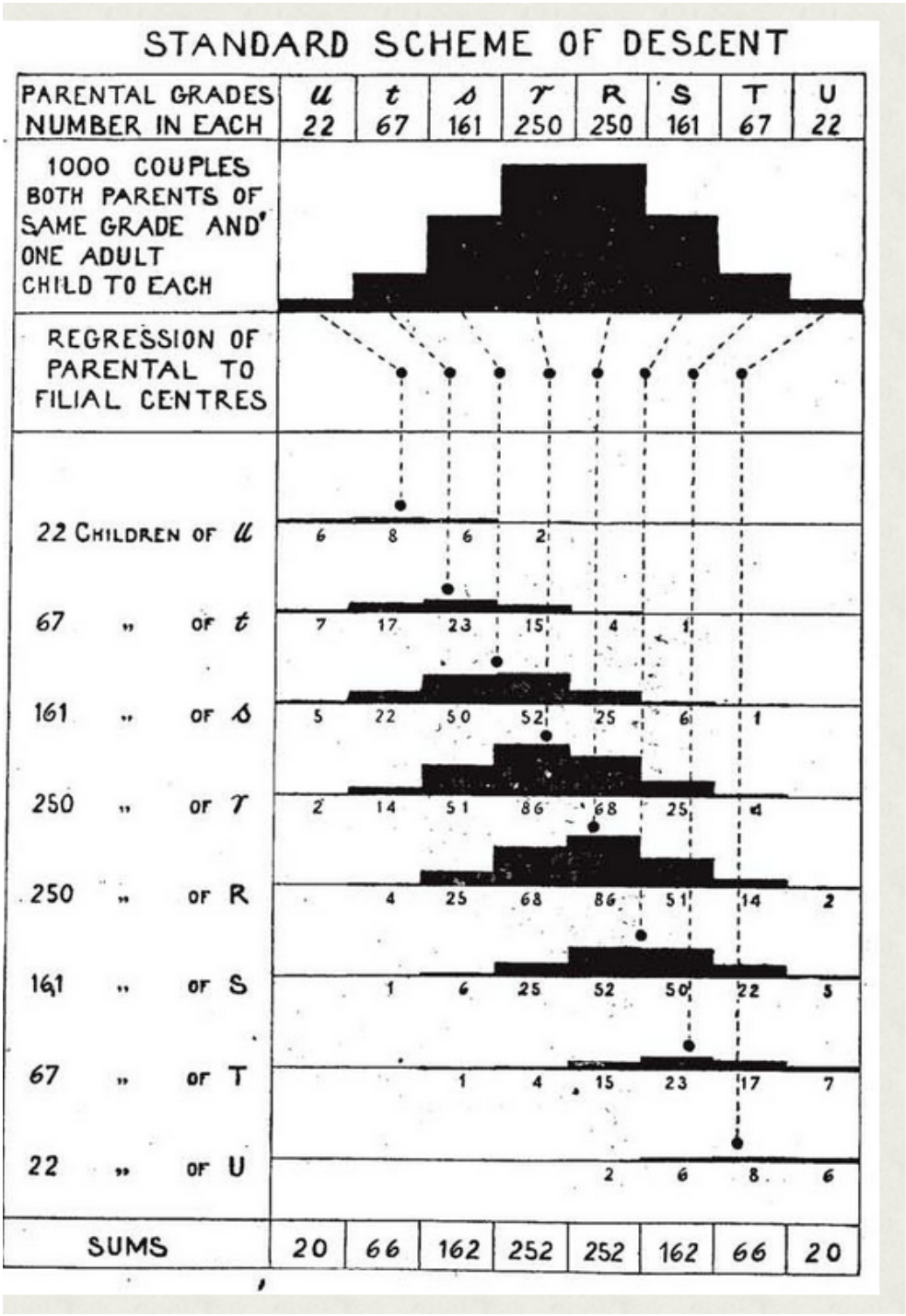


Fig. 5. Galton's Second Regression Line and his "Forecaster of Stature."

Kuva 25: Galtonin toinen regressio 1886.



Kuva 26: Galtonin havainnollistus vanhempien ja lasten pituuden jakaumista 1901.

seuraavana vuonna. Vastaavasti tavanomaista pienemmästä sosiaalitukea nauttivien lukumäärästä vuonna 2016 ilahtuneet kaupunginjohtajat joutuvat tyypillisesti pettymään, kun sosiaalitukea haetaan vuonna 2017 edellistä vuotta enemmän. □

Esimerkki. Voittajan kirous. Edelläkuvatunkaltainen ilmiö on ”voittajan kirous” (*winner's curse*): Kun suuresta joukosta esimerkiksi työpaikan tai urheilujoukkueen jäsenyyden hakijoista poimitaan suorituksiltaan paras, ei valittu yllä aivan aiempien suoritustensa mukaiseen tulokseen. □

Edellä vertailtiin kahden satunnaismuuttujan yhteyttä, kun ne ovat samoinjakautuneita ja niiden hajontakuvioon piirretyn muuttujien välisen systemaattisen komponentin summeeraavan suoran kulmakerroin on (itseisarvoltaan) alle yksi (tyypillisesti oleellisesti sama satunnaismuuttuja jossain mielessä kahdesti mitattuna). Regressioanalyysissa (jakso 12.3) voidaan sallia useampia muuttujia, jotka voivat olla erilailla jakautuneita tai kiinteitä, eikä systemaattisen vaikutuksen suuruutta tarvitse rajoittaa edelliseen tapaan. (Tällöin ei voida puhua regressiosta odotusarvoa kohti aivan samassa mielessä kuin edellä.) Regressioanalyysillä pyritään selvittämään systemaattisuus yhden muuttujan ja muiden muuttujien välillä. Aina regressioanalyysissa on kuitenkin kyse pohjimmiltaan samasta ilmiöstä kuin edellä eli että osa havaintojen käyttäytymisestä on systemaattista ja osa sattumaa. Sattuman vaikutus tulisi regressoida ”pois” muuttujien välistä yhteyttä arvioitaessa. Esimerkiksi Galtonin tutkimusaineistossa lasten ja vanhempien pituuksien suhteella on geneettinen (systemaattinen) selitys, mutta osin lasten pituudet johtuvat (tutkijan näkökulmasta) sattumanvaraisista seikoista kuten lapsen perimistä geeneistä, lapsen saaman ruoan ravinnepitoisuudesta tai lapsen sairastamista taudeista, kellonajasta, jolloin lapsi on mitattu (aamulla lapsi on pidempi) ja niin edelleen. On vain hieman liioiteltua sanoa, että lähes asiassa kuin asiassa on regressiota. Campbellin ja Kennyn (1999, ix) mukaan regressio odotusarvoa kohti on yhtä väistämätön asia kuin verot tai kuolema.

12.2 Regressiovirhepäätelmä

Regressiovirhepäätelmä (*regression fallacy*) tehdään, kun satunnaisvaihtelussa regressiota odotusarvoa kohti kuvaavan suoran ympärillä kuvitellaan kausaalisuutta kuten että regressio johtaisi jakauman typistymiseen. Yhteiskuntatieteilijät ovat joskus hahmottavinaan kausaalisuutta tilanteista, joissa sitä ei ole. Vaikka ongelma on tunnettu, edelleen tehdään virheellisiä tulkintoja.

Esimerkki. Yritysten keskiarvoistuminen. Kuuluisa esimerkki on tilastotieteen(!) professori Horace Secrist. Hän julkaisi 1933 massiivisen empiirisen tutkimuksen amerikkalaisten yritysten liikevoittojen kehityksestä 1920–1930. Hän havaitsi, että yritysten, jotka pärjäsivät parhaimmin tai huonoimmin 1920, liikevoitot olivat lähestyneet 1930 kaikkien yritysten liikevoittojen keskiarvoa. Secrist päätteli, että taloudellinen kilpailu pakotti yritykset ”keskiarvoistumaan” ajan myötä. Löytönsä korostamiseksi Secrist antoi kirjalleen nimeksi *The Triumph of Mediocrity in Business*. Todellisuudessa yritysten liikevoittojen jakauma ei ollut

muuttunut, ja Secristin havainnot selittyvät regressiolla odotusarvoa kohti.[88] □

Esimerkki. Business-kirjallisuus. Kahneman (2011, 204–208) kritisoi business-kirjallisuutta, jossa perehdytään menestyneiden yhtiöiden strategioihin, yrityskulttuureihin ja johtamistapoihin. Esimerkkinä hän mainitsee Collinsin ja Porrasin (2000) kirjan. Sen viesti on, että jokaisen toimitusjohtajan, johtajan ja yrittäjän tulisi lukea se, jotta muutkin yritykset osaisivat noudattaa menestyneiden yritysten toimintamalleja ja pärjäisivät. Kahnemanin mukaan Collinsin ja Porrasin ylistämät yritykset eivät pian tutkimuksen julkaisemisen jälkeen enää pärjänneet juurikaan kilpailijoitansa paremmin. Kahneman viittaa muihin vastaaviin tapauksiin, joissa tutkimuksessa hehkutettujen yritysten kukoistus lopahtaa tutkimuksen julkaisemisen jälkeen. Regressio odotusarvoa kohti on luonteva tulkinta tällaisille tapahtumille. Ihaillut yritykset olivat erityisen menestyviä tutkimushetkellä sattumalta. □

Esimerkki. Hävittäjälentäjät. Kahneman (mts. 174) kertoo mainion esimerkin, kuinka ihmiset voivat kuvitella kausaalisuutta siellä, missä on pelkkää sattumaa (lyhennetty käännös luennoitsijan):

> Sain yhden elämäni tyydyttävimmistä eureka-kokemuksistani opettaessani Israelin ilmavoimien lentokouluttajille tehokkaan opettamisen psykologiaa. Olin kertonut kouluttajille, kuinka hyvän suorituksen palkitseminen toimii paremmin kuin virheistä rangaitseminen. Yksi vanhemmista kouluttajista arveli, että hyvästä suorituksesta palkitseminen sopii ehkä linnuille muttei hävittäjälentäjäkadeteille: ”Olen monesti kehunut kadetteja puhtaasta suorituksesta vaikeassa lentoliikkeessä. Seuraavalla kerralla he järjestään suoriutuvat samasta liikkeestä huonommin. Toisaalta olen monesti huutanut kadetin korvakuulokkeeseen haukkuen häntä huonosta suorituksesta. Ylipäänsä haukkumani kadetit pärjäävät seuraavalla yrityksellä paremmin. Olkaa siis hyvä, älkääkä kertoko meille, että kehuminen toimii ja rangaistus ei, koska asia on juuri päinvastoin.”

Edellä opitun perusteella on helppo hahmottaa, että vanhemman kouluttajan kokemukset selittyvät sattumalla: Erityisen hyvin pärjänneen kadetin suoritus regressoitui seuraavalla lennolla kohti odotusarvosuoritustaan ja erityisen heikosti suoriutuneen kadetin suoritus samoin. Kouluttaja virheellisesti liitti muutoksiin kuvittelemansa syy-seuraus -suhteen kehuistaan ja karjumisistaan.[89] □

12.3 Regressioanalyysi

Regressioanalyysi on käytetyimpiä tilastotieteellisiä menetelmiä. Ei liene olemassa ainakaan kaupallista tilasto-ohjelmistoa, joka ei sisältäisi regressioanalyysiä. Yksi syy lienee, että se mahdollistaa muuttujan vaikutuksen suuruuden toiseen muuttujaan tai vaikutuksen olemassaolon ylipäätään arvioinnin ja testaamisen (tiettyjen oletusten pätiessä). Ne ovat polttavia kysymyksiä monen tutkijan mielessä. Regressioanalyysi on tässä mielessä usein hyvin antoisaa ja tuloksellista.

[88] Stigler (1999) kertoo Secristin tutkimuksesta tarkemmin. Ks. myös Wallis ja Roberts (1956, luku 8).

[89] Lisää esimerkkejä on Wainerin (2005) kirjan luvussa 10) sekä Wallisin ja Robertsin (1956) kirjan luvussa 8. Ks. myös Friedman (1992), Jerrim ja Vignoles (2013) sekä Goldstein (2015).

12.4 Yhden selittäjän lineaarinen regressiomalli

Tarkastellaan kahta muuttujaa y ja x. Edellisen pitää olla välimatka-asteikollinen; jälkimmäinen voi olla myös luokitteluasteikollinen. Muuttuja y määräytyy lineaarisen regressiomallin

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{55}$$

mukaisesti x:n arvoista. Muuttujaa y kutsutaan selitettäväksi muuttujaksi ja muuttujaa x selittäväksi muuttujaksi, selittäjäksi tai regressoriksi. Viimeinen termi ε on mallin jäännös eli satunnaistermi, jonka odotusarvo on 0 ja varianssi on σ^2. Mallin parametrit ovat kiinteitä lukuja (esim. $\beta_0 = 90.5$ ja $\beta_1 = 0.5$), joiden suuruudet (tyypillisesti) ovat tuntemattomia ja joiden selvittämiseen regressioanalyysillä pyritään. Parametria β_0 kutsutaan usein mallin vakioksi ja parametria β_1 (regressio)kertoimeksi.

Jaksossa oletetaan yksinkertaisuuden vuoksi, että selittäjä x on kiinteä — ja jaksossa 12.5, että selittäjät x_i ovat kiinteitä. Empiirisessä tutkimuksessa oletus ei usein ole uskottava. Analyysit luvussa pätevät, jos selittäjät ja jäännökset ovat toisistaan riippumattomia satunnaismuuttujia ja tehdään sopivia oletuksia ja merkintöjä tuunataan. Regressiomallia voi soveltaa, vaikka selittäjät olisivat satunnaismuuttujia.

Jaksossa 12.1 viitattu systemaattinen komponentti on yhden selittäjän regressiossa selitettävän (selittäjän x arvosta riippuva) odotusarvo

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x.$$

Yllä $\mathsf{E}(\cdot)$ on odotusarvon symboli ja on käytetty oletusta $\mathsf{E}(\varepsilon) = 0$.

Tavalliseen x,y-koordinaatistoon funktio $y = \beta_0 + \beta_1 x$ piirtyy suorana, jonka kulmakerroin on β_1 ja joka leikkaa y-akselin kohdassa β_0. Periaatteessa β_0 kertoo siten selitettävän odotusarvon, kun selitettävän arvo on 0:

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \varepsilon) = \beta_0.$$

Empiirisessä analyysissa tämä tulkinta ei ole aina järkevä, mihin palataan myöhemmin (jaksot 12.4.1 ja **??**).

Mallin mukaan y:n suuruus riippuu x:n suuruudesta lineaarisesti parametrin β_1 välityksellä: Jos x muuttuu yksikön verran, niin y muuttuu β_1:n verran. Esimerkiksi jos y on lapsen pituus, x on isän pituus ja $\beta_1 = 0.5$, niin mallin mukaan lapsen pituus tapaa olla 0.5 senttimetriä pidempi, jos isä on senttimetrin pidempi. Vakio β_0 asettaa mallin kuvaaman suoran sopivalle korkeudelle. Jäännös ε kuvaa y:n vaihtelua, joka ei selity x:n vaihtelulla. Esimerkiksi lapsen pituuteen vaikuttaa muitakin tekijöitä kuin isän pituus (jakso 12.1). Ne jäävät mallissa huomioimatta ja puristetaan ε:iin.

Erikoistapaus on $\beta_1 = 0$. Tällöin malli (55) typistyy niin, että y on satunnaisesti jakautunut vakion β_0 ympärillä:

$$y = \beta_0 + 0 \times x + \varepsilon = \beta_0 + \varepsilon. \tag{56}$$

Kiinnostavin asia mallissa (55) onkin tyypillisimmin parametrin β_1 suuruus — esimerkiksi poikkeaako se nollasta eli päteekö malli (55) vai (56).

Regressioanalyysi on keino arvioida parametrien suuruutta ja systemaattista komponenttia, kun mallin kuvaamasta ilmiöstä on havaintoaineisto. Mallin (55) kohdalla aineisto koostuisi havaintopareista $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. Tässä $n \geq 2$ on havaintojen lukumäärä.

Galtonin herneen siemen -vanhemmat ja niiden jälkipolvi sekä vanhempien ja lasten pituudet -esimerkit (jakso 12.1) havainnollistavat regressiota odotusarvoa kohti mallin (55) mukaisesti, kun $0 < \beta_1 < 1$. Kuvitteellinen sosiaalitukien saajat -esimerkki (jakso 12.1) vastaa mallia (56): Sosiaalitukien saajien lukumäärä edellisenä vuonna (x) ei auta ennustamaan heidän lukumääräänsä kuluvana vuonna (y): Kerroin $\beta_1 = 0$, ja lukumäärät pyrkivät palautumaan kohti odotusarvoaan β_0.

Kuvaan 27 on piirretty keinotekoinen aineisto ($n = 25$) — vaikkapa helsinkiläisten isien (x) ja heidän poikiensa (y) pituuksista aikuisina.[90] Kukin piste vastaa yhtä havaintoparia (x_i,y_i). Mitä pidempi isä, sitä pidempi tapaa poika olla. Mutta kuinka paljon? Voitaisiinko yhteys tiivistää suoraksi, jonka parametreista voitaisiin päätellä vaikutuksen keskimääräinen suuruus?

12.4.1 Yhden selittäjän lineaarisen regressiomallin estimointi

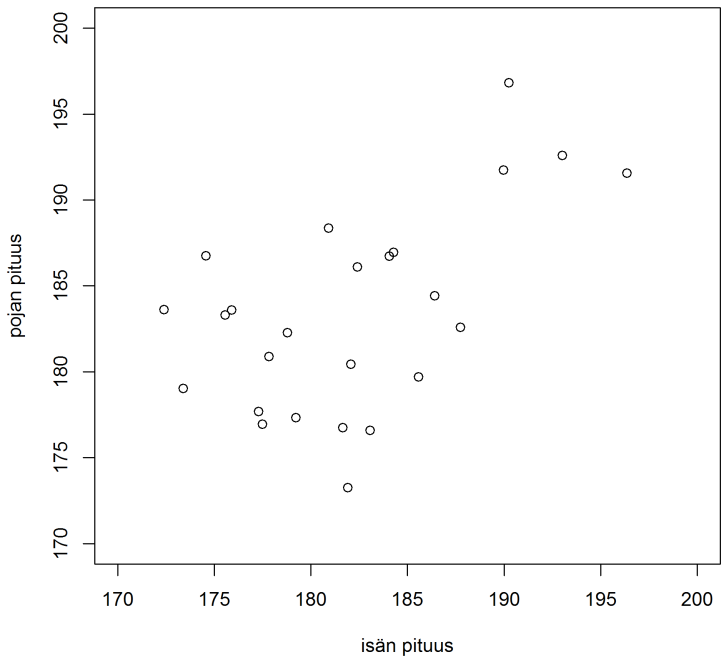
Yhden selittäjän regressiossa aineistoon sovitetaan regressiosuora, joka summeeraa muuttujien välisen riippuvuuden eli systemaattisen osan. Regressiosuoran parametriarvot ovat vastaus edellä esitetyn tapaisiin kysymyksiin.

Sovittaminen voidaan tehdä periaatteessa monella tavalla. Ylivoimaisesti käytetyin tapa on pienimmän neliösumman (PNS) menetelmä. Siinä parametrit β_0 ja β_1 valitaan niin, että y_i -havaintojen poikkeamat sovitettavasta suorasta neliöidään ja neliöiden summa minimoidaan:

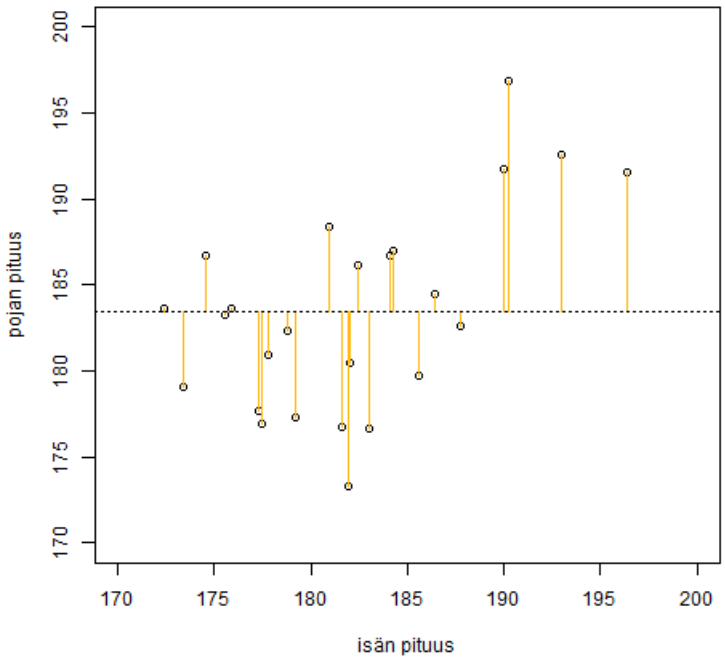
$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Merkintä ”min” tarkoittaa, että sen oikealla puolella oleva lauseke minimoidaan min-merkinnän alapuolelle merkittyjen suureiden suhteen. Minimoinnin voi ajatella tapahtuvan ikään kuin kokeilemalla eri lukuarvoja β_0:lle ja β_1:lle ja valitsemalla sellainen β_0,β_1-pari, että lauseke $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ ei voi saada pienempiä arvoja. (Todellisuudessa tilasto-ohjelmisto ratkaisee minimointitehtävän yhdellä laskutoimituksella eikä kokeile eri arvoja.) Poikkeamien suoralta $y_i - \beta_0 - \beta_1 x_i$ kasvaessa (itseisarvoltaan) kasvaa neliösumma $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ nopeasti. PNS-menetelmä pyrkii siten tuottamaan regressiosuoran, joka ei koskaan sijoittuisi kovin kauas yhdestäkään havaintopisteestä. Termien $(y_i - \beta_0 - \beta_1 x_i)$ neliöinnin takia minimoinnin kannalta ei ole väliä, onko y_i suu-

[90] Aineisto on luotu olettaen sekä isien että poikien keskipituudeksi ja -hajonnaksi 181.0 cm ja 6.06 cm. Nämä ovat uusimpien suomalaisten miesten pituustietojen mukaiset luvut (professori Leo Dunkell, henkilökohtainen tiedonanto 16.3.2010). Isien ja poikien pituuden korrelaatioksi on oletettu 0.5, joka vastaa melko tarkasti todellista korrelaatiota (esim. Pearson ja Lee 1903). Aineisto, kuviot ja analyysit alla tehtiin R-ohjelmiston MASS-paketin avulla (Venables ja Ripley 2002).



Kuva 27: Isien ja poikien pituudet.



Kuva 28: Poikien pituuksien poikkeamat poikien keskipituudesta.

rempi tai pienempi kuin mallin mukainen arvo $\beta_0 - \beta_1 x_i$. Kaikkia poikkeamia kohdellaan tässä mielessä samanarvoisesti.

Neliösumman minimoivia parametriarvoja kutsutaan PNS-estimaateiksi ja niitä merkitään $\hat{\beta}_0$:lla ja $\hat{\beta}_1$:lla. Suureita

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

kutsutaan sovitteiksi ja suureita

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

residuaaleiksi ($i = 1, \ldots, n$). Regressiosuora $\hat{\beta}_0 + \hat{\beta}_1 x_i$ saadaan piirtämällä hajontakuvioon suora sovitteiden ($\hat{y}_i$) kautta. Residuaalit ($\hat{\varepsilon}_i$) ovat jäännösten (ε_i) estimaatteja.

Sovite $\hat{y}_i$ voidaan ilmaista muuttujien keskiarvojen $\bar{y} = \sum_{i=1}^n y_i/n$ ja $\bar{x} = \sum_{i=1}^n x_i/n$ avulla:

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}). \tag{57}$$

Sovite poikkeaa $\bar{y}$ -keskiarvosta $\hat{\beta}_1$ kertaa x_i:n poikkeaman omasta keskiarvostaan verran. Regressiosuora kulkee pisteen ($\bar{x}$,$\bar{y}$) kautta.

Tärkeä käsite on residuaalineliösumma

$$\mathsf{RNS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Nimityksensä mukaisesti se on summa residuaalien neliöistä. Se on sitä suurempi, mitä enemmän y_i-havainnot poikkeavat sovitteista $\hat{y}_i$. Sen avulla lasketaan estimaatti jäännöksen varianssille:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Jäännösvarianssin estimaatti $\hat{\sigma}^2$ mittaa residuaalin neliön keskimääräistä suuruutta eli vaihtelevuutta aineistossa.[91] Yleensä toivotaan, että $\hat{\sigma}^2$ olisi pieni, koska silloin malli selittää hyvin y:n vaihtelun. Monesti raportoidaan jäännöksen estimoitu keskihajonta $\mathsf{SD}(\hat{\sigma}^2) = \sqrt{\hat{\sigma}^2} = \hat{\sigma}$ (*standard deviation*), koska se on samassa mittayksikössä kuin selitettävä muuttuja ja on siksi helpompi hahmottaa.

Määritellään vastaavasti kokonaisneliösumma

$$\mathsf{KNS} = \sum_{i=1}^n (y_i - \bar{y})^2. \tag{58}$$

Se kuvaa, kuinka suurta on y_i-havaintojen vaihtelu keskiarvonsa ympärillä.

[91] Syy jakaa RNS $n-2$:lla eikä n:llä liittyy näin saadun σ^2:n estimaatin teoreettisiin ominaisuuksiin ja jakson 12.4.2 jakaumateoriaan. Muistisääntö on, että estimoinnissa havaintoja "menetetään" yksi kutakin estimoitua parametria kohti. Vrt. $\hat{\sigma}^2$:n kaava (64) monen selittäjän regressiomallissa (62).

Neliösummista saadaan mallin selityskyvylle mittari selitysosuus

$$\mathsf{R}^2 = \frac{\mathsf{KNS} - \mathsf{RNS}}{\mathsf{KNS}} = 1 - \frac{\mathsf{RNS}}{\mathsf{KNS}}. \tag{59}$$

Selitysosuus saa lähellä yhtä olevia arvoja, mikäli residuaalineliösumma on pieni suhteessa selitettävän kokonaisneliösummaan ($\mathsf{RNS}/\mathsf{KNS} \approx 0$). Tällöin y selittyy hyvin x:llä. Mikäli x:llä ei ole selityskykyä, residuaalineliösumma ei eroa paljoa kokonaisneliösummasta ($\mathsf{RNS}/\mathsf{KNS} \approx 1$). Tällöin selitysosuus on lähellä nollaa.

Selitysosuus on hyvin intuitiivinen mittari mallin hyvyydelle. Nyt esillä olevassa yhden selittäjän regression tilanteessa se onkin aiemmista opinnoista tutun otoskorrelaatiokertoimen (r) neliö:

$$\mathsf{R}^2 = r^2. \tag{60}$$

Kuvat 28 ja 29 havainnollistavat käsitteitä isä-poika -aineiston avulla. Kuvaan 28 on piirretty poikien pituuksien poikkeamat poikien pituuksien keskiarvosta ($(y_i - \bar{y})$:t). Poikkeamien neliöiden summa on kokonaisneliösumma (58). Kuvassa 29 regressiosuora määrittää sovitteen kunkin x_i-havainnon kohdalla. Residuaalit ovat (x_i,y_i)-havainnoista regressiosuoraan pystysuorasti meneviä viivoja ($(y_i - \hat{y}_i)$:t). Toinen suora tuottaisi toiset residuaalit. Kuvion residuaalien neliöiden summa on pienin mahdollinen. Mallin (55) PNS-estimointi tuotti tästä aineistosta tulokset

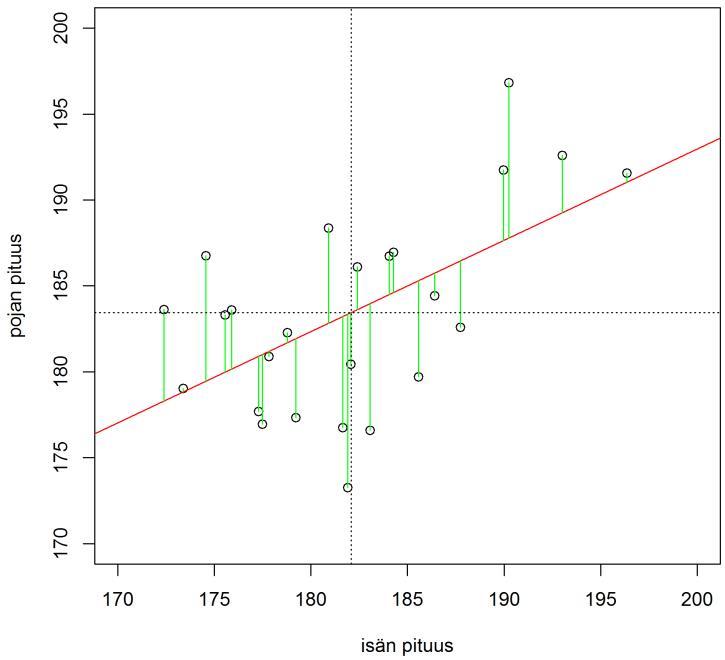
$$\begin{aligned} y &= 86.79 + 0.531x + \hat{\varepsilon} \\ &= 183.4 + 0.531(x - 182.1) + \hat{\varepsilon}, \\ \hat{\sigma} &= 4.96, \quad \mathsf{R}^2 = 0.313. \end{aligned}$$

Malli ennustaa pojalle lisää pituutta 0.531 eli noin 0.5 senttimetriä isän pituuden kasvaessa senttimetrillä ja selittää noin 31 % poikien pituuden vaihtelusta aineistossa. Toinen rivi yllä esittää mallin kaavan (57) muodossa isien ja poikien pituuksien keskiarvojen ($\bar{x} = 182.1$ ja $\bar{y} = 183.4$) avulla ($86.79 \approx 183.4 - 0.531 \times 182.1$). Kaavasta (60) seuraa, että otoskorrelaatio on selitysosuuden neliöjuuri: $r = \sqrt{\mathsf{R}^2}$. Pituuksien otoskorrelaatio on siten $\sqrt{0.313} \approx 0.559$. Jäännöksen estimoitu keskihajonta on 4.96.

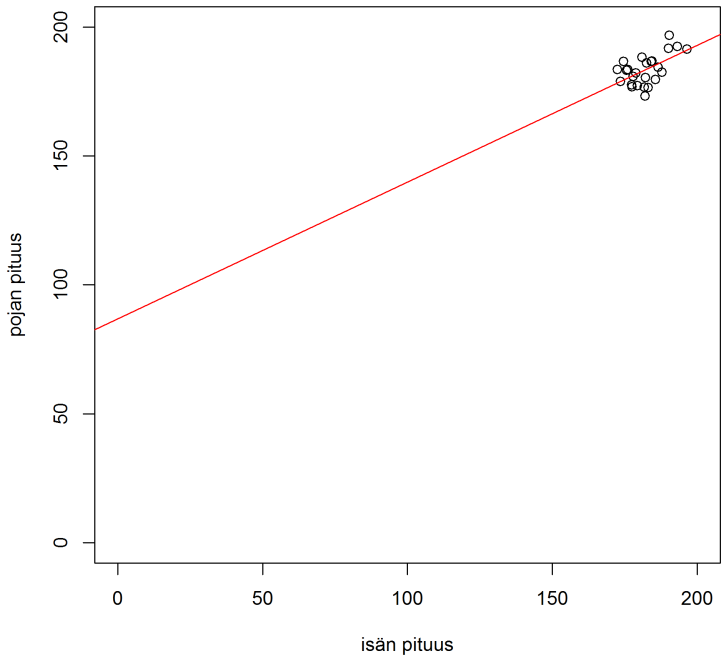
Koska aineisto oli keinotekoinen, estimointituloksia voidaan verrata aineiston tuottaneeseen todelliseen malliin. Suureiden todelliset arvot ovat $\beta_0 = 90.5$, $\beta_1 = 0.5$, $\sigma \approx 5.25$ (seuraa tehdyistä oletuksista tavalla, jota ei tässä selitetä), korrelaatio populaatiossa $\rho = 0.5$ ja selitysosuus populaatiossa $\mathsf{R}^2 = \rho^2 = (0.5)^2 = 0.25$. Kaikki suureet tulivat estimoiduksi varsin hyvin.

Kuten usein on, estimoidulla vakiolla ei ole järkevää tulkintaa. Estimoidun mallin ja vakion mukaan pojan pituus olisi noin 87 senttimetriä, jos isän pituus olisi 0 senttimetriä (kuvio 30), mikä on järjetön ajatus. Malli ei välttämättä antaisi luotettavaa ennustetta edes periaatteessa mahdollisen mutta poikkeuksellisen lyhyen isän (esim. $x = 155$) pojan pituudelle. Regressiomalleja ei ylipäänsä kannata yrittää soveltaa aineiston vaihteluvälin ulkopuolella (ekstrapoloida).

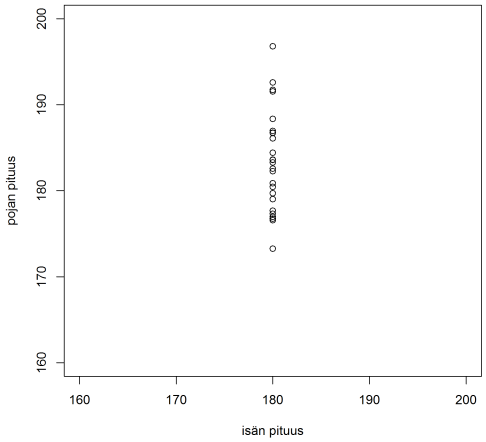
Edellä implisiittisesti oletettiin, että kaikki x_i-havainnot eivät ole yhtäsuuria. Jos ne olisivat, β_1-parametria ei voisi estimoida, mitä kuva 31 havainnollistaa. Seuraavassa jaksossa tarvitaan muitakin oletuksia.



Kuva 29: Isien ja poikien pituuden regressiosuora ja residuaalit.



Kuva 30: Regressiosuoralla ekstrapolointi.



Kuva 31: Kaikki isät samanpituisia.

12.4.2 Yhden selittäjän lineaarisen regressiomallin testaus

Ei ole harvinaista, että tutkijan päämielenkiinto on estimoinnissa. Vaikka se ei olisi, pääsääntöisesti regressiomallin tarkastelussa ei tulisi rajoittua estimointituloksiin. Mallia tulisi aina testata. Filosofia on sama kuin muutenkin tilastotieteessä: Pelkkä estimaatin tai yleisemmin tilastollisen tunnusluvun subjektiivinen arviointi ei ole riittävää; tulee myös laskea väliestimaatti tai testata, poikkeaako tunnusluku nollasta tai muusta oleelliseksi katsotusta arvosta tilastollisesti merkitsevästi. Hedelmällisen tilastotieteen soveltamisen tunnusmerkkejä on, että on arvioitu sekä tunnuslukujen merkittävyyttä sovellusalan kannalta että niiden tilastollista merkitsevyyttä luottamusvälien tai testien avulla. Alla keskitytään testaukseen, koska niin tehdään empiirisessä kirjallisuudessa.

Regressioanalyysillä voidaan testata parametreihin liittyviä nollahypoteeseja (H_0). Sellaisia ovat esimerkiksi H_0: $\beta_1 = 0$ tai H_0: $\beta_1 = 1$. Vakion suuruutta testataan harvoin, koska sillä ei ole usein selkeää sovellukseen liittyvää merkityksellistä tulkintaa (vrt. isä-poika -malli edellä).

Hypoteesien testaukseen tarvitaan lisäoletuksia:

- Jäännös noudattaa Normaalijakaumaa odotusarvolla 0 ja varianssilla σ^2: $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$, $i = 1, \ldots, n$.
- Jäännökset ε_i eivät korreloi keskenään eli ne ovat riippumattomia toisistaan (Normaalijakauman tilanteessa korreloimattomuudesta seuraa riippumattomuus).

Oleellista on ymmärtää, että $\hat{\beta}_1$ on satunnaismuuttuja. Se saa yhden tietyn arvon tutkittavana olevassa aineistossa. Jos tutkittavana olisi toinen — esimerkiksi espoolainen 25 havainnon aineisto isien ja poikien pituuksista — saataisiin

toisensuuruinen $\hat{\beta}_1$. Samoin estimaatti muuttuisi, jos tutkittaisiin 25 vantaalaisen, 25 kaunialaisen jne. aineistoa isien ja poikien pituuksista. Koska $\hat{\beta}_1$ on satunnaismuuttuja, on sillä (ilmeisesti) myös keskihajonta, jota kutsutaan tässä yhteydessä keskivirheeksi (jakso 6.1).

Edellä lueteltujen oletuksien pätiessä estimaattien jakaumat tunnetaan. Tavalla, jota tässä ei selitetä, voidaan laskea estimoitu keskivirhe $\hat{\beta}_1$:lle ($\mathsf{SD}(\hat{\beta}_1)$) ja muodostaa t-testisuure eli t-arvo

$$t_{\beta_1=\beta_{10}} = \frac{\hat{\beta}_1 - \beta_{10}}{\mathsf{SD}(\hat{\beta}_1)} \sim \mathsf{t}_{n-2}.$$

Nollahypoteesin H_0: $\beta_1 = \beta_{10}$ pätiessä se noudattaa t-jakaumaa vapausasteilla $n-2$. Huomionarvoista on, että jakauma riippuu havaintojen lukumäärästä mutta tunnetaan kaikilla havaintomäärillä. Testisuure on hyvin intuitiivinen. Estimaatin $\hat{\beta}_1$ poikkeama nollahypoteesin mukaisesta arvosta β_{10} suhteutetaan estimaatin estimoituun keskivirheeseen. Suurikaan poikkeama ei ole tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on suuri. Toisaalta pienikin poikkeama on tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on hyvin pieni. Keskivirhe pienenee havaintojen lukumäärän kasvaessa. (Muutkin tekijät vaikuttavat keskivirheen suuruuteen.)

Tyypillisimmin testataan nollahypoteesia $\beta_1 = 0$. Tällöin testisuure on yksinkertaisesti β_1:n estimaatti jaettuna estimoidulla keskivirheellään:

$$t_{\beta_1=0} = \frac{\hat{\beta}_1}{\mathsf{SD}(\hat{\beta}_1)} \sim \mathsf{t}_{n-2}. \tag{61}$$

Monet tilasto-ohjelmistot tulostavat tämän testisuureen regressoitaessa selitettävää yhdellä selittävällä muuttujalla. Toiset ohjelmistot raportoivat PNS-estimaatin ja sen keskivirheen, jolloin käyttäjän tehtävä on muodostaa osamäärä $\hat{\beta}_1/\mathsf{SD}(\hat{\beta}_1)$. Tieteellisissä artikkeleissa käytäntö vaihtelee: Joissain raportoidaan estimaatti ja t-arvo ja toisissa estimaatti ja sen estimoitu keskivirhe. Jälkimmäisessä tilanteessa lukijan tulee osata itse muodostaa t-arvo, jos haluaa tietää sen suuruuden.

Testaaminen etenee tämän jälkeen tavanomaiseen tapaan eli valitaan sopivaksi katsottu riskitaso, ja katsotaan, onko testisuureen itseisarvo suurempi kuin riskitasoon liittyvä kriittinen arvo (kaksisuuntainen testaus). Esimerkiksi 5 %:n riskitasoa käytettäessä kriittiset arvot olisivat isä-poika -esimerkissä t-jakauman $25-2=23$:lla vapausasteella 2.5. tai 97.5. persentiilit.

Isä-poika -esimerkissä $\hat{\beta}_1$:n keskivirhe on 0.164, joten t-arvo on $0.531/0.164 \approx 3.238$. Esimerkin laskussa käytetty R-ohjelmisto raportoi sekä estimoidun keskivirheen että t-arvon, joka täsmää juuri lasketun kanssa. Ohjelmiston mukaan p-arvo on noin 0.004, joten nollahypoteesi $\beta_1 = 0$ hylätään 5 %:n riskitasolla ja paljon pienemmilläkin riskitasoilla. Samaan tulokseen päädytään vertaamalla t-arvoa 3.238 t-jakauman 23. vapausasteella 97.5. persentiiliin 2.069 (`qt(0.975,23)`).

Testin mukaan isien ja lasten pituus ovat yhteydessä ($\beta_1 \neq 0$). Tulos on odotettu. Mikäli tutkittava ilmiö olisi tuntemattomampi. keskeinen osa regres-

sioanalyysia olisi testata, poikkeaako parametri β_1 nollasta. Mikäli nollahypoteesia ei hylättäisi (t-arvo olisi itseisarvoltaan pienempi kuin kriittiset arvot), pääteltäisiin, että muuttujien välillä ei ole yhteyttä tai että aineisto ei ainakaan ole ristiriidassa oletuksen yhteyden puuttumisesta kanssa. Mallin selitysosuus olisi tällöin lähellä nollaa (mieti miksi!), ja kaavan (60) perusteella muuttujien välinen otoskorrelaatio olisi samoin lähellä nollaa.

Esimerkki. Itsemurhat. Daly ym. (2011) estimoivat PNS-menetelmällä yhtälön

$$\begin{array}{rcccccc} y & = & 24.912 & + & 8.255x & + & \hat{\varepsilon}, \\ & & (2.311) & & (3.992) & & \end{array}$$

$$\mathsf{R}^2 = 0.248,\ n = 15.$$

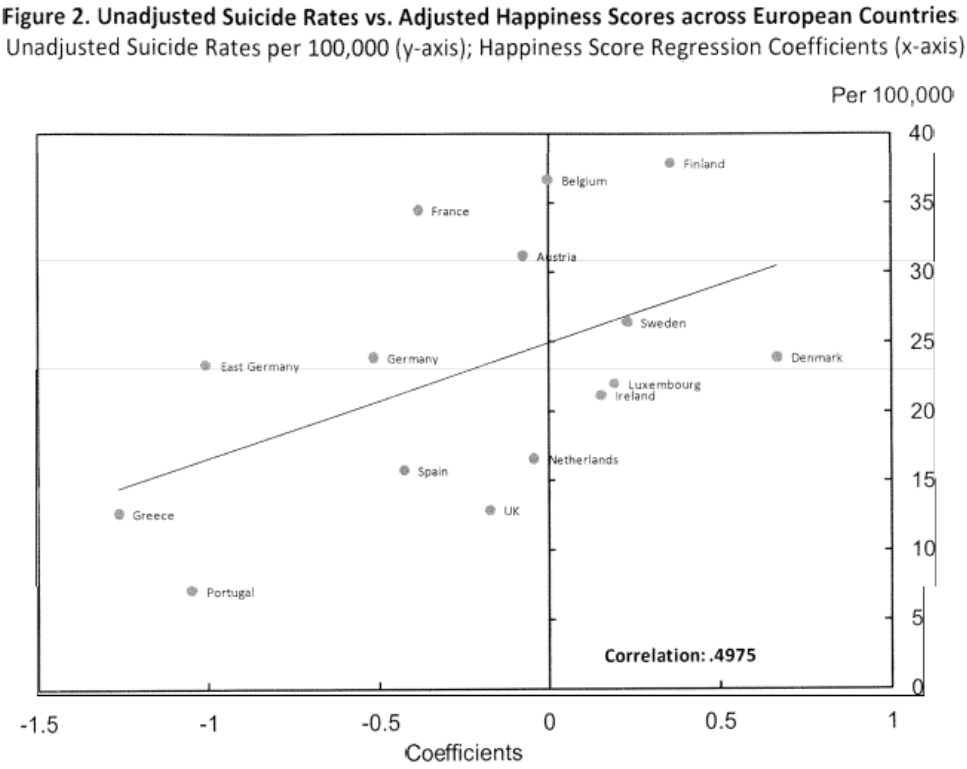
Yllä y on itsemurhien lukumäärä 100 000 kansalaista kohti, x on kansakunnan onnellisuutta mittaava indeksi, $\hat{\varepsilon}$ on residuaali, luvut suluissa ovat estimoituja keskivirheitä ja n on havaintojen lukumäärä. Jäännösten ε oletetaan noudattavan Normaalijakaumaa $\mathsf{N}(0,\sigma^2)$ ja olevan keskenään korreloimattomia. Kukin havaintopari (x_i,y_i) liittyy eurooppalaiseen valtioon ($i = 1,\ldots,15$). Havainnot ja niihin sovitettu regressiosuora ovat kuvassa 32. Mallin mukaan

- itsemurhaintensiteetti (itsemurhien lukumäärä 100 000 kansalaista kohti) on 24.912 (estimoitu vakio), kun onnellisuusindeksi saa arvon 0.
- itsemurhaintensiteetti kasvaa onnellisuusindeksin kasvaessa. Kun jälkimmäinen suurenee yksiköllä, edellinen kasvaa 8.255:llä (estimoitu kerroin onnellisuusindeksille).
- 24.8 % itsemurhaintensiteetin vaihtelusta selittyy onnellisuusindeksin vaihtelulla (selitysasteen R^2 suuruus).

Onnellisuusindeksi saa toiseksi ja itsemurhaintensiteetti suurimman arvon Suomen kohdalla. Suomessa tehdään itsemurhia vielä enemmän kuin malli ennustaa — eniten koko aineistossa.

Kuviossa raportoidun korrelaatiokertoimen 0.4975:n neliö on mallin selitysaste 0.248, koska mallissa on vain yksi selittäjä (kaava (60)). Koska jäännökset ε ovat normaalijakautuneita ja keskenään korreloimattomia, estimoidut kertoimet jaettuna estimoiduilla keskivirheillään ovat t-jakautuneita. Koska mallissa on vain yksi selittäjä ja havaintoja on 15, on jakauma t_{15-1-1} eli t_{13} (kaava (61)). Testisuure on $8.255/3.992 \approx 2.068$. Jakauman t_{13} 97.5. persentiili on 2.160 (oheisesta taulukosta). Koska $|2.068| < |2.160|$, niin nollahypoteesi ei tule aivan hylätyksi 5 %:n riskitasolla kaksisuuntaisessa testauksessa. Onnellisuusindeksi ei ole tilastollisesti merkitsevä selittäjä eikä aineiston perusteella ole syytä luopua oletuksesta, että itsemurhaintensiteetti ja onnellisuusindeksi eivät korreloi.

Kuvion perusteella saattaisi veikata, että muuttujien välillä olisi todellinen yhteys. Selitys tilastolliselle merkitsettömyydelle saattaa olla aineiston pieni koko: Tilastollisesti merkitsevä positiivinen suhde itsemurhaintensiteetin ja osavaltioiden onnellisuusindeksien välillä pätee Yhdysvalloissa (mt.), ja osavaltioita on



Kuva 32: Itsemurhien ja onnellisuuden yhteys 15 eurooppalaisessa valtiossa.

enemmän kuin eurooppalaisia valtioita regressiossa edellä. Mahdollisesti osavaltiot ovat myös homogeenisempia kuin eurooppalaiset valtiot, jolloin tutkittu suhde tulee selvemmin esiin osavaltioaineistossa (jäännös sisältää vähemmän vaihtelevia tekijöitä). □

12.5 Monen selittäjän lineaarinen regressiomalli

Selitettävä muuttuja y määräytyy nyt monen selittävän muuttujan x_i ($i = 1,\ldots,k$) lineaarisesta regressiomallista

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon. \tag{62}$$

Mallin tulkinta on samantapainen kuin mallin (55). Selitettävä y on jatkuvaarvoinen, selittäjät x_i voivat olla myös luokitteluasteikollisia ja jäännös ε on satunnaistermi odotusarvolla 0 ja varianssilla σ^2. Siihen tiivistyy y:n vaihtelu, joka ei selity x_i:den vaihtelulla. Parametrit $\beta_0,\ldots,\beta_k$ ovat kiinteitä yleensä tuntemattomia lukuja, joiden suuruudet pyritään selvittämään regressioanalyysilla (eritoten β_1:stä β_k:hon). Parametria β_0 kutsutaan vakioksi ja parametreja $\beta_1,\ldots,\beta_k$ (regressio)kertoimiksi.

Mallin (62) systemaattinen komponentti on selitettävän (selittäjien x_i arvoista riippuva) odotusarvo

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{63}$$

Vakio kuvaa nyt selitettävän odotusarvoa, kun kaikki selittäjät saavat arvon 0:

$$\mathsf{E}(y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \cdots + \beta_k \times 0 + \varepsilon) = \beta_0.$$

Kuten yhden selittäjän regressiossa (jakso 12.4.1), tämä tulkinta ei ole aina järkevä.

Kerroin β_i kuvaa x_i:n yksikön suuruisen muutoksen vaikutuksen y:hyn, kun muut selittäjät eivät muutu. Monesti mielenkiintoisin kysymys on, ovatko β_i-kertoimet nollia eli selittääkö x_i:den vaihtelu lainkaan y:n vaihtelua.

Monen selittäjän regressiomallin (62) systemaattisen komponentin ja parametrien selvittäminen edellyttää n:stä havaintovektorista $[x_{11}\ \ldots\ x_{1k}\ y_1],\ldots,$ $[x_{n1}\ \ldots\ x_{nk}\ y_n]$ koostuvaa aineistoa ($n \geq k$). Muuttujien ensimmäinen indeksi on havainnon numero ($i = 1,\ldots,n$) ja jälkimmäinen indeksi kertoo, mistä selittäjästä havaintoarvo x_{ij} on ($j = 1,\ldots,k$).

12.5.1 Monen selittäjän lineaarisen regressiomallin estimointi

Monen selittäjän regressiossa aineistoon sovitetaan selitettävän ja selittäjien välisen riippuvuuden summeeraava lineaarinen funktio eli mallin (62) systemaattinen osa (63).[92] Sovittaminen tehdään yleisimmin PNS-menetelmällä jaksossa 12.4.1 esitettyyn tapaan. Parametrien $\beta_0,\ldots,\beta_k$ lukuarvot valitaan minimoimaan y_i-havaintojen poikkeamien systemaattisesta komponentista neliöiden

[92] Systemaattisen osan geometrinen tulkinta ei ole yhtä helppo kuin yhden selittäjän regressiossa, jossa aineistoon sovitettiin suora. Mikäli selittäjiä on kaksi, sovitetaan aineistoon kaksiulotteinen taso.

summa:

$$\begin{aligned} & \min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2 \\ = \quad & \min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2. \end{aligned}$$

Neliösumman minimoivat parametriarvot ovat PNS-estimaatteja $\hat{\beta}_0, \ldots, \hat{\beta}_k$. Jaksossa 12.4.1 selitetyt käsitteet yleistyvät muutenkin suoraviivaisesti k:n selittäjän tilanteeseen. Sovitteet ($\hat{y}_i$), residuaalit ($\hat{\varepsilon}_i$), residuaalineliösumma ja jäännöksen varianssin estimaatti ovat nyt

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik},$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik},$$

$$\mathsf{RNS} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2$$

ja

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2. \tag{64}$$

Kokonaisneliösumman ja selitysosuuden kaavat (58) ja (59) eivät muutu. Merkittävä ero on, että selitysosuuden ja otoskorrelaation neliöt sitova kaava (60) pätee nyt, kun otoskorrelaatiokerroin r on laskettu selitettävän muuttujan y_i ja sen sovitteen $\hat{y}_i$ välille. (Tämä tulkinta on mahdollinen myös yhden selittäjän regression mallin kohdalla.)

12.5.2 Monen selittäjän lineaarisen regressiomallin testaus

Testaamista varten jaksossa 12.4.2 tehtyjä oletuksia pitää täydentää olettamalla nyt, että kaikki selittävät muuttujat x_i ovat kiinteitä (niissä ei ole satunnaisuutta). Lisäksi yhdenkään selittäjän x_i arvot eivät saa riippua täydellisesti lineaarisesti muiden selittäjien x_j, $j \neq i$, arvoista.[93]

Ehkä tärkein ja useimmin testattu monen selittäjän regressiomallin (62) β_i-kertoimia koskeva nollahypoteesi on, että ne ovat kaikki nollia (H_0: $\beta_1 = \cdots = \beta_k = 0$). Nollahypoteesin mukaan selittäjillä x_i ei ole tällöin lainkaan selityskykyä selitettävän muuttujan y suhteen. Tämän hypoteesin päteminen tai pätemättömyys on tutkijalle usein keskeisimpiä kysymyksiä. Nollahypoteesia testaava F-testisuure on hyvin yksinkertainen:

$$F = \frac{\mathsf{R}^2/k}{(1-\mathsf{R}^2)/(n-k-1)} \sim \mathsf{F}_{k,n-k-1} \tag{65}$$

[93] Yhden selittäjän tilanteessa jälkimmäinen ehto merkitsee, että kaikki ainoan selittäjän havainnot eivät saa olla samoja. Tätä tilannetta sivuttiin jakson 12.4.1 lopussa. Useamman selittäjän tilanteessa ei esimerkiksi ole sallittua, että yhden selittäjän arvot olisivat toisen selittäjän arvoja kerrottuna jollain luvulla.

Nollahypoteesin H_0: $\beta_1 = \cdots = \beta_k = 0$ pätiessä se noudattaa F-jakaumaa k:lla ja $n-k-1$:llä vapausasteella. Jälleen (vrt. jakso 12.4.2) jakauma riippuu havaintojen lukumäärästä ja on tunnettu kaikilla havaintomäärillä.

Useimmat tilasto-ohjelmistot laskevat F-testisuureen automaattisesti regression yhteydessä. Testisuureen suuret arvot ovat testin kannalta hälyttäviä. Mikäli tilasto-ohjelmisto ei ilmoita F-testisuureen p-arvoa, voidaan testisuureen arvon tilastollinen merkitsevyys arvioida F-jakauman kriittisten arvojen avulla.

F-testisuureella on selkeä intuitio. Mikäli selittäjät x_i kykenevät selittämään suuren osan selitettävän y vaihtelusta (KNS), niin jäljelle jäävä selittämätön vaihtelu (RNS) muodostuu pieneksi ja selitysosuus R^2 suureksi (kaava (59)). Kaavasta (65) nähdään, että mitä suurempi R^2 on, sitä suurempi on F-testisuure. F-testi siis hälyttää, kun selittäjillä on aineistossa hyvä selityskyky. Myös havaintojen lukumäärän kasvattaminen pyrkii kasvattamaan F-testisuuretta ja todennäköisyyttä hylätä nollahypoteesi, kun se ei päde. Mikäli nollahypoteesi pätee, R^2 tapaa jäädä pieneksi ja F-testisuure samoin.

Yleisiä mallin (62) parametreja koskevia nollahypoteeseja ovat, että i:nnen selittäjän kerroin on nolla (H_0: $\beta_i = 0$) tai että se on tietyn suuruinen (H_0: $\beta_i = \beta_i^0$). Edellisessä tilanteessa i:nnettä selittäjää ei tarvittaisi regressiossa (62). Näitä nollahypoteeseja voidaan testata jakson 12.4.2 tapaisilla t-testisuureilla:

$$t_{\beta_i=\beta_{i0}} = \frac{\hat{\beta}_i - \beta_{i0}}{\mathsf{SD}(\hat{\beta}_i)} \sim \mathsf{t}_{n-k-1}.$$

ja

$$t_{\beta_i=0} = \frac{\hat{\beta}_i}{\mathsf{SD}(\hat{\beta}_i)} \sim \mathsf{t}_{n-k-1}. \tag{66}$$

Vastaavan nollahypoteesin pätiessä ne noudattavat t-jakaumaa vapausasteilla $n-k-1$. Jakauma riippuu havaintojen lukumäärästä mutta tunnetaan kaikilla havaintomäärillä. Tilasto-ohjelmisto raportoi yleensä automaattisesti jälkimmäisen t-arvon kaikkien selittäjien estimoiduille kertoimille tai niiden estimoidut keskivirheet $\mathsf{SD}(\hat{\beta}_i)$, $i = 1, \ldots, n$. Testaus tapahtuu käytännössä jaksossa 12.4.2 selitetyllä tavalla. Siellä kuvattiin myös t-testisuureiden intuitio.

Monesti on kiinnostavaa testata, olisikovatko mallin (62) d ($0 < d \leq k$) oikeanpuoleisinta selittäjää tarpeettomia eli päteekö $\beta_{k_r+1} = \cdots = \beta_k = 0$, jossa $k_r = k - d > 0$. (Oletus tarpettomien selittäjien sijoittumisesta mallin oikeanpuoleisimmiksi tehdään merkintöjen yksinkertaistamiseksi.) Edellä "r" viittaa rajoitettuun. Näin rajoitettu malli olisi

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k_r} x_{k_r} + \varepsilon. \tag{67}$$

Se saadaan mallista (62) erikoistapauksena asettamalla d kappaletta β_i-kertoimia nollaksi ($k_r + 1$). selittäjästä lähtien.

Nollahypoteesia H_0: $\beta_{k_r+1} = \cdots = \beta_k = 0$ voidaan testata testisuureella

$$\frac{(\mathsf{R}^2 - \mathsf{R}_r^2)/d}{(1-\mathsf{R}^2)/(n-k-1)} \sim F_{d,n-k-1}.$$

Testisuure vaatii sekä regression (62) että regression (67) laskemisen. Jälkimmäisen regression selitysastetta on merkitty yllä R_r^2:lla. Nollahypoteesin pätiessä testisuure noudattaa F-jakaumaa d:llä ja $n-k-1$:llä vapausasteella. Testisuureen suuret arvot ovat hälyttäviä.

Tämänkin testisuureen toimintaperiaate on hyvin ymmärrettävä. Mikäli kertoimet $\beta_{k_r+1},\ldots,\beta_k$ poikkeavat tai osa niistä poikkeaa nollasta, mallien selitysasteiden tulisi erota selvästi. Tällöin erotus $\mathsf{R}^2-\mathsf{R}_r^2$ testisuureen osoittajassa muodostuu suureksi ja testisuure samoin. Mikäli d. viimeisellä selittäjällä ei ole selitysvoimaa (nollahypoteesi pätee), erotus ja testisuure jäävät pieniksi.

Muunkinlaisia rajoituksia (esim. $\beta_1=\beta_2$ tai $\beta_1+\cdots+\beta_k=1$) mallin (62) parametreille voidaan testata. Asia jätetään maininnan varaan.

Esimerkki. Siivoojien tuntipalkat. Keinänen ja Pakarinen (2009) tutkivat siivoojien tuntipalkkoja ja mahdollista palkkasyrjintää suomalaisessa siivousyrityksessä vuonna 2007. He estimoivat vaihtoehtoisia malleja, jotka ovat kaikki palkkasyrjintää koskevalta tulokseltaan yhtäpitäviä. Yksi heidän (PNS-menetelmällä) estimoimistaan malleista on

$$\begin{array}{rcccccccccc} y & = & 8.430 & + & 0.114x_1 & - & 0.001x_2 & + & 0.169x_3 & + & 0.339x_4 & + & \hat{\varepsilon}. \\ & & (0.000) & & (0.840) & & (0.983) & & (0.747) & & (0.000) & & \end{array} \tag{68}$$

$$\mathsf{R}^2=0.256,\ F_{4,132}=11.269,\ n=137.$$

Yhtälössä y on tuntipalkka, x_1 on indeksi, joka saa arvon 1, kun siivooja on mies ja 0 muuten, x_2 on siivoojan ikä, x_3 on indikaattori työsuhteen laadulle, joka saa arvon 1, kun työsuhde on toistaiseksi voimassa oleva ja 0 muutoin[94] ja x_4 on työsuhteen kesto vuosina. Muut merkinnät (F-testisuureella täydennettynä) ja oletukset ovat kuten edellisessä esimerkissä. Kukin havaintovektori $[x_{i1}\ \ldots\ x_{i4}\ y_i]$ liittyy yhteen siivojaan ($i=1,\ldots,137$).

Jäännöksen normaalisuusoletuksen perusteella testisuureet noudattavat t- ja F-jakaumia. Muuttujien selityskykyä yhdessä testaan F-testisuurella $F=11.269$. Nollahypoteesin pätiessä se noudattaa jakaumaa $F_{4,132}$ (kaava (65)). Sen 95. persentiili on 2.440 (`qf(0.95,4,132)`).[95] Koska $11.269>2.440$, niin nollahypoteesi hylätään. Mallin selittäjillä on yhdessä selityskykyä.

Sukupuoli-indikaattorin t-arvo on $0.114/0.840\approx 0.136$. Nollahypoteesin (kerroin on 0) pätiessä se noudattaa $\mathsf{t}_{137-4-1}$ eli t_{132}-jakaumaa (kaava (66)). Sen 95. persentiili on 1.656 (`qt(0.95,132)`).[96] Koska $|0.136|<|1.660|$, niin nollahypoteesia ei hylätä 10 %:n riskitasolla. Aineiston mukaan ei ole syytä luopua oletuksesta, että miehet ja naiset saavat samaa palkkaa (kun muut palkkaan

[94] Keinänen ja Pakarinen eivät selitä, miten tämä muuttuja on luotu. Selitys yllä on tämän tekstin kirjoittajan päättelemä. Keinänen ja Pakarinen eivät raportoi F-testisuuretta, mutta se on laskettavissa artikkelin tietojen perusteella.

[95] Taulukkokirjaa käytettäessä joudutaan tyytymään esimerkiksi jakauman $\mathsf{F}_{4,100}$ kriittisiin arvoihin. Sen 95. persentiili on 2.463.

[96] Taulukoita käytettäessä testisuuretta voi verrata t_{100}-jakaumaan. Sen 95. persentiili on 1.660. Myös Standardinormaalijakauman 95. persentiiliä 1.645 voisi käyttää kriittisenä arvona. Suurehkon havaintomäärän eli vapausasteiden suuruuden johdosta t- ja Standardinormaalijakaumien persentiilit poikkeavat vain vähän toisistaan.

vaikuttavat tekijät on huomioitu) eli että palkkasyrjintää ei ole. Ikämuuttujan estimoitu kerroin on 0.01, ja sen t-arvo on $0.001/0.983 \approx 0.001$. Testisuureen arvon perusteella on selvää, että ikämuuttuja ei voi olla tilastollisesti merkitsevä selittäjä millään järkevällä riskitasolla. Aineiston mukaan ikä ei vaikuta siivojan palkkaan. Samoin voidaan päätellä, että työsuhteen laatu ei näytä olevan yhteydessä palkkaan ($0.169/0.747 \approx 0.226$). Työsuhteen kesto on tilastollisesti merkitsevä selittäjä palkalle: mallin raportointitarkkuuden yllä mukaan $0.339/0.000 \approx \infty$ (kertoimen estimaatin keskihajonta todellisuudessa on varmasti hieman nollaa suurempi).

Selittäjistä ainoastaan työsuhteen kesto näyttää olevan yhteydessä siivoojien palkkaan. Kukin työvuosi nostaa palkkaa 0.339 euroa.

Tutkimuksessa seuraava vaihe voisi olla estimoida malli, jossa ainoa selittäjä on työsuhteen kesto. Tällöin saataisiin luultavasti hieman eri ja hieman tarkempi estimaatti lisätyövuoden palkkaa nostavalle vaikutukselle. Tällaisen mallin vakio olisi palkka siivoojalle, joka on juuri aloittanut siivoojan työuransa.

Mallin (68) vakiolla ei ole järkevää tulkintaa. Kirjaimellisesti tulkiten se olisi palkka 0-vuotiaalle naissiivoojalle, jonka työsuhde ei ole toistaiseksi voimassa oleva ja jonka työsuhde on vasta alkanut. □

12.6 Lopuksi

Viimeinen esimerkki. Elämänfilosofiaa. Oxfordin yliopisto koostuu 38 college'ista. Somerville College on nimetty skotlantilaisen matemaatikko Mary Somervillen (1780–1872) mukaan. Ensi vuodesta lähtien Royal Bank of Scotlandin 10 punnan setelin kuva-aihe on Mary Somerville. Hän paljasti 91-vuotiaana yhden menestymisensä selityksen:[97]

Jollen onnistu tänään, pureudun ongelmaan uudestaan huomenna.

Paina Somervillen ohje mieleesi. Menestystä opinnoillesi! Hyvää kesää! □

[97] `Http://blog.oup.com/2016/03/mary-somerville-royal-bank-of-scotland/` (viitattu 3.5.2016).