JOHDATUS TILASTOTIETEESEEN

2. KIRJA

TILASTOTIETEEN JATKOKURSSI

ILKKA MELLIN

HELSINGIN YLIOPISTO

TILASTOTIETEEN LAITOS

JOHDATUS TILASTOTIETEESEEN. 2. KIRJA.

TILASTOTIETEEN JATKOKURSSI

ILKKA MELLIN **KESÄ 1996**

Sisällys

1. TODENNÄKÖISYYSLASKENTA

1.1 TODENNÄKÖISYYS JA SEN TULKINTA

1.1.1 JOHDANTO

Empiiriset eli kokemusperäiset ilmiöt voidaan jakaa *determinististisiin* ja *satunnaisiin*.

DETERMINISTINEN ILMIÖ

Ilmiö on *deterministinen*, jos ilmiön alkutilan perusteella voidaan ennustaa tarkasti sen lopputila eli *tulos*.

Fysiikan ja erityisesti klassisen mekaniikan tutkimat luonnonilmiöt ovat esimerkkejä deterministisistä ilmiöistä. Esimerkiksi ammuksen lentorata voidaan ennustaa hyvin tarkasti ammuksen painon, lähtönopeuden, lähtökulman, ilman vastuksen ja ampumasuunnan perusteella. On kuitenkin hyvä tietää, että *kaaosteorian* mukaan on olemassa deterministisiä ilmiöitä, joihin liittyy ennustamattomuutta.

SATUNNAISILMIÖ

Ilmiö on *satunnainen*, jos sen alkutilasta ei voida tarkasti ennustaa sen lopputilaa eli *tulosta*, mutta ilmiön toistuessa nähdään toistumiskerroista määrättyjen *tulosvaihtoehtojen suhteellisten osuuksien* eli *frekvenssien* jakautuvan säännönmukaisesti.

Esimerkkejä satunnaisista ilmiöistä tarjoavat lähes kaikki elävään luontoon sekä ihmisten ja ihmisryhmien käyttäytymiseen liittyvät empiiriset ilmiöt, mutta myös monet fysiikan ilmiöt kuten kvanttimekaniikan ilmiöt ovat satunnaisia. Esimerkiksi radioaktiivisen aineen tietyn atomiytimen hajoamisajankohtaa on mahdotonta ennustaa. Radioaktiiviselle aineelle voidaan kuitenkin ilmoittaa ns. puoliintumisaika, joka kertoo kuinka kauan kestää, että puolet ko. aineen ytimistä on hajonnut. Se, että puoliintumisaika voidaan ilmoittaa hyvinkin tarkasti, on eräs radioaktiiviseen hajoamiseen liittyvistä säännönmukaisista piirteistä.

Satunnaisilmiön tulosvaihtoehtojen esiintymiseen pitää aina liittyä säännönmukaisuutta: *Satunnaisilmiön tulos ei saa vaihdella mielivaltaisella tavalla.* Kun satunnaisilmiö toistuu, tulosvaihtoehtojen ilmiön toistumiskerroista määrätyt suhteelliset frekvenssit vaihtelevat toistumiskerrasta toiseen, mutta vaihtelu ei saa olla miten mielivaltaisen suurta — vaihtelun pitää olla hallittua. Tällä tarkoitetaan seuraavaa: Satunnaisilmiön toistuessa pitää tulla yhä epätodennäköisemmäksi, että tulosvaihtoehtojen suhteelliset frekvenssit poikkeavat paljon tulosvaihtoehtojen

todennäköisyyksistä. Tästä ei saa tehdä sellaista johtopäätöstä, että suuret poikkeamat olisivat *mahdottomia*, ne tulevat vain yhä *epätodennäköisemmiksi*.

Esimerkiksi rahanheiton tulokseen liittyvän satunnaisvaihtelun säännönmukaisuudella tarkoitetaan seuraavaa: Kun rahaa heitetään toistuvasti, voidaan tarkkailla kruunien siihen astisista heittokerroista määrätyn suhteellisen frekvenssin käyttäytymistä. Tämä suhteellinen frekvenssi vaihtelee heittokerrasta toiseen, mutta vaihtelu ei ole kuitenkaan mielivaltaista. Esimerkiksi, jos 1,000 heitossa kruunan suhteelliseksi frekvenssiksi on saatu 0.9, niin on hyvin vähän todennäköistä, että tekemällä 9,000 heittoa lisää, kruunan suhteelliseksi frekvenssiksi saataisiin 0.1.

Säännönmukaisuus tulosvaihtoehtojen esiintymisessä eli tulosvaihtojen suhteellisen frekvenssin *stabiliteetti* ilmiön toistuessa mahdollistaa sen, että satunnaisilmiöiden käyttäytymistä voidaan tutkia mielekkäästi. Matemaattinen todennäköisyyden teoria eli *todennäköisyyslaskenta* muodostaa matemaattisen mallin satunnaisilmiöiden käyttäytymiselle.

Tilastotieteen päämääränä on tehdä *johtopäätöksiä* jostakin empiirisestä ilmiöstä siitä kerätyn *havaintoaineiston* perusteella. Aineisto kerätään *perusjoukosta*, jossa ilmiötä tutkitaan. Johtopäätökset perustuvat aineistosta määrättyihin otostunnuslukuihin kuten keskilukuihin, hajontalukuihin ja riippuvuuden mittoihin.

Jos havaintoaineisto saadaan satunnaisotoksesta tai satunnaistetusta kokeesta, havaintoaineisto vaihtelee satunnaisesti otoksesta toiseen. Tätä satunnaisvaihtelun muotoa kutsutaan *otosvaihteluksi*. Havaintoaineiston otosvaihtelu periytyy kaikkiin otoksesta laskettuihin tunnuslukuihin. Havaintoaineiston ja otostunnuslukujen satunnaisvaihtelun syynä on aineiston poimintaan liittyvä satunnaisuus. Todennäköisyyden teoria antaa *mallin* tunnuslukujen vaihtelulle otoksesta tai kokeesta toiseen, *jos ilmiön taustalla olevan perusjoukon olosuhteet pysyvät samana*. Se, että perusjoukon olosuhteet pysyvät samoina takaa sen, että tutkittavan satunnaisilmiön tulosvaihtoehdot eivät vaihtele mielivaltaisella tavalla.

TODENNÄKÖISYYDEN IDEA

Todennäköisyyden matemattinen mallittaminen lähtee liikkeelle siitä empiirisestä havainnosta, että satunnaisten ilmiöiden tulosvaihtoehtojen suhteelliset frekvenssit näyttävät lähestyvän ilmiön toistuessa kiinteitä lukuarvoja.

ESIMERKKI 1.

- Rahan heitossa sekä kruunien että klaavojen suhteelliset frekvenssit näyttävät lähestyvän pitkissä heittosarjoissa lukua 1/2.
- Poikien suhteellinen frekvenssi kaikkien syntyneiden lasten joukossa näyttää olevan kaikkialla hieman lukua 1/2 suurempi. ●

Satunnaisilmiön tulosvaihtoehdon *todennäköisyys* voidaan määritellä ko. tulosvaihtoehdon havaituksi suhteelliseksi frekvenssiksi, joka on määrätty ilmiön toistumiskerroista. Jotta todennäköisyyttä olisi mahdollista tutkia mielekkäästi, on vaadittava, että tulosvaihtoehdon suhteellinen frekvenssi käyttäytyy säännönmukaisesti, kun satunnaisilmiö toistuu. Tätä kutsutaan todennäköisyyden *empiiriseksi*

määritelmäksi, koska sen mukaan jonkin tulosvaihtoehdon todennäköisyys voidaan määrätä vain kokeellisesti, ts. keräämällä satunnaisilmiöstä havaintoja.

EMPIIRINEN TODENNÄKÖISYYS

Jos tarkasteltavan tapahtumavaihtoehdon *suhteellinen frekvenssi* lähestyy jotakin kiinteätä lukua satunnaisilmiön toistuessa, on tuo luku tapahtumavaihtoehdon *todennäköisyys*.

On syytä huomata, että empiiristä todennäköisyyttä ei käytännössä voida koskaan määrätä tarkasti, koska ilmiön havaitsemista voidaan ainakin periaatteessa jatkaa, jolloin suhteellinen frekvenssi jatkaa suhteellisen frekvenssin stabiilisuudesta huolimatta vaihteluaan havainnosta toiseen. Määritelmässä oletetaan kuitenkin, että vaihtelu on jossakin mielessä hallittua. On syytä huomata, että jos satunnaisilmiöön liittyvien harvinaistenkin tapahtumien todennäköisyydet pyritään selvittämään empiirisesti edes jonkinlaisella tarkkuudella, on satunnaisilmiöstä kerättävä paljon havaintoja.

Matemaattinen todennäköisyys on empiirisen todennäköisyyden idealisointi. Se kuvaa sitä, mitä tapahtuisi, jos ilmiö toistuisi äärettömän monta kertaa. Todennäköisyyttä tarkastellaan myöhemmin tässä luvussa myös matemaattisesti.

Empiirinen todennäköisyys ei salli todennäköisyyksien liittämistä kertaluonteisiin tapahtumiin. Kertaluonteisten tapahtumien tulosvaihtoehtoihin voidaan liittää ilmiön havaitsijan persoonalliset käsitykset tulosvaihtoehtojen todennäköisyyksistä. Tällaiset todennäköisyydet ovat luonteeltaan *subjektiivisia*.

1.1.2 TODENNÄKÖISYYSMALLI

Määritellään satunnaisilmiön *todennäköisyysmalli*:

TODENNÄKÖISYYSMALLI

Satunnaisilmiön *todennäköisyysmallissa* on kaksi osaa:

1. Mahdollisten tulosvaihtojen kuvaus.

2. Tulosvaihtoehtojen todennäköisyyksien kuvaus.

OTOSAVARUUS

Todennäköisyysmallin on sisällettävä kuvaus siitä mitkä tulosvaihtoehdot ovat mahdollisia. *Otosavaruus* sisältää tämän tiedon.

OTOSAVARUUS

Satunnaisilmiön *otosavaruus* S on kaikkien mahdollisten tulosvaihtoehtojen joukko.

Otosavaruus on siis se *perusjoukko*, jossa satunnaisilmiötä tarkastellaan: Tulosvaihtoehdot ovat tämän perusjoukon *alkioita*. Jokin niistä tulosvaihtoehdoista, joista otosavaruus koostuu, on tuloksena, kun satunnaisilmiö tapahtuu. Satunnaisilmiön tulosvaihtoehtoja kutsutaan usein *alkeistapahtumiksi*. Tällöin halutaan korostaa sitä, että satunnaisilmiötä ei voida "purkaa" alkeistapahtumia alkeellisempiin tulosvaihtoehtoihin.

Äärellisen otosavaruuden tapauksessa luettelemme usein otosavaruuden alkiot aaltosulkujen ympäröimänä. Siten merkintä

$$S = \{a_1, a_2, \ldots, a_n\}$$

tarkoittaa sitä, että otosavaruus S koostuu alkeistapahtumista $a_1, a_2, \ldots, a_n$.

ESIMERKKI 1.

- Rahanheitossa S = {kruuna,klaava}.
- Lapsen sukupuolen määrätymisessä S = {tyttö,poika}.
- Nopan heitossa S = {1,2,3,4,5,6}.
- Kahden nopan heitossa S = {(1,1),(1,2),(1,3),...,(6,5),(6,6)} (36 silmälukuparia). ●

TODENNÄKÖISYYDET

Todennäköisyysmallin pitää sisältää kuvaus myös siitä miten satunnaisilmiön tulosvaihtoehtoihin liitetään niiden todennäköisyyttä kuvaavat mittaluvut. Tämä ei kuitenkaan riitä, vaan myös tulosvaihtoehtojen yhdistelmiin on voitava liittää todennäköisyydet. Tämä tehdään todennäköisyyslaskennan lakien avulla. Tulosvaihtoehtojen yhdistelmiä kutsutaan *tapahtumiksi*.

TAPAHTUMA

Mikä tahansa satunnaisilmiön tulosvaihtoehtojen joukko on *tapahtuma*, ts. tapahtuma on otosavaruuden osajoukko.

Jos tapahtumaan liittyy äärellinen määrä tulosvaihtoehtoja, voimme määritellä tapahtuman luettelemalla tapahtumaan liittyvät tulosvaihtoehdot samaan tapaan kuin voimme määritellä otosavaruuden äärellisen otosavaruuden tapauksessa luettelemalla siihen liittyvät tulosvaihtoehdot. Olkoon A jokin otosavaruuden S tapahtuma, johon liittyvät alkeistapahtumat $a_1, a_2, \ldots, a_k$. Tällöin siis merkitsemme

$$A = \{a_1, a_2, \ldots, a_k\}.$$

Tällaisen merkinnän käyttäminen ei ole kuitenkaan aina kätevää. Käytämme usein myös merkintää, jossa määrittelemme tapahtumaan A liittyvät alkeistapahtumat lauseella, jonka S:n alkiot on toteutettava, jotta ne olisivat A:n alkioita. Jos merkitsemme tällaista lausetta symbolisesti $\mathscr{P}$:llä, voimme siis käyttää merkintää:

A = "Ne S:n alkiot, jotka toteuttavat lauseen $\mathscr{P}$".

ESIMERKKI 2.

Kuten esimerkissä 1 todettiin, kahden nopan heitossa otosavaruus eli perusjoukko S on silmälukujen 1,2,3,4,5,6 muodostamien parien

(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)

joukko. Nämä parit ovat alkeistapahtumia kahden nopan heiton muodostamassa satunnaisilmiössä.

Parien

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

muodostama joukko on perusjoukon osajoukko, jossa kummankin nopan heiton tuloksena on sama silmäluku. Se on perusjoukon osajoukkona siis tapahtuma. Tästä tapahtumasta voidaan käyttää merkintöjä

A = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}
= "Sama silmäluku kahden nopan heitossa". ●

Tapahtuman A todennäköisyyttä merkitään seuraavalla tavalla:

P(A).

P-kirjan tulee englanninkielen sanasta *probability*.

TODENNÄKÖISYYDEN PERUSLAIT

Tapahtuman A todennäköisyys P(A) on luku 0:n ja 1:n välissä:

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

Otosavaruuden eli kaikkien mahdollisten tulosvaihtoehtojen joukon S todennäköisyys on 1:

$P(S) = 1.$

Jos A ja B ovat kaksi tapahtumaa ja

$$P(A) > P(B),$$

sanotaan, että A on *todennäköisempi* kuin B. Tämä määritelmä mahdollistaa tapahtumien todennäköisyyksien vertailemisen.

Mitä lähempänä tapahtuman todenäköisyys on maksimiarvoaan 1, sitä *yleisempi* on tapahtuma. Mitä lähempänä tapahtuman todennäköisyys on minimiarvoaan 0, sitä *harvinaisempi* on tapahtuma. *Varman* tapahtuman todennäköisyys on 1 ja *mahdottoman* tapahtuman todennäköisyys on 0.

Jos otosavaruus S on äärellinen, ts. mahdollisia tulosvaihtoehtoja on äärellinen määrä, tulosvaihtojen todennäköisyyksien on toteutettava seuraava ehto: Yksittäisten tulosvaihtoehtojen todennäköisyyksien summan on oltava 1:

ÄÄRELLISEN OTOSAVARUUDEN ALKEISTAPAHTUMIEN TODENNÄKÖISYYDET

Olkoon $S = \{a_1, a_2, \ldots, a_n\}$, jossa a_i on tulosvaihtoehto ja olkoon $P(\{a_i\}) = p_i$ tulosvaihtoehdon a_i todennäköisyys. Tällöin

$$p_1 + p_2 + \ldots + p_n = 1.$$

Äärellisen otosavaruuden tapauksessa tapahtuman todennäköisyys saadaan laskemalla niiden alkeistapahtumien eli tulosvaihtoehtojen todennäköisyydet yhteen, jotka yhdessä muodostavat tapahtuman.

ÄÄRELLISEN PERUSJOUKON TAPAHTUMIEN TODENNÄKÖISYYDET

Jos A on jokin tapahtuma eli $A \subset S$, niin tapahtuman A todennäköisyys saadaan laskemalla A:han liittyvien otosavaruuden S tulosvaihtoehtojen todennäköisyydet yhteen. Toisin sanoen, jos $A = \{a_1, a_2, \ldots, a_k\}$ ja $P(\{a_i\}) = p_i$, niin

$$P(A) = p_1 + p_2 + \ldots + p_k.$$

1.1.3 TODENNÄKÖISYYDEN KÄSITTEEN TULKINNAT

Todennäköisyyden käsitteelle voidaan antaa seuraavat tulkinnat:

- Todennäköisyys suhteellisena frekvenssinä.
- Todennäköisyys subjektiivisena uskottavuuden asteena.

TODENNÄKÖISYYDEN FREKVENSSITULKINTA

Olkoon tapahtuman A *todennäköisyys* $P(A)$. Oletetaan, että satunnaisilmiö on toistunut n kertaa ja, että tapahtuma A on tällöin sattunut f kertaa. Tällöin tapahtuman A *frekvenssi* on f ja tapahtuman A *suhteellinen frekvenssi* on f/n. Todennäköisyyden

empiirisen määritelmän mukaan tapahtuman A todennäköisyys voidaan samaistaa tapahtuman A suhteellisen frekvenssin kanssa:

TODENNÄKÖISYYS SUHTEELLISENA FREKVENSSINÄ

Oletetaan, että satunnaisilmiö on toistunut n kertaa ja, että tapahtuma A on sattunut tällöin f kertaa. Tällöin tapahtuman A *todennäköisyys* on

$$P(A) = \frac{f}{n}.$$

Suhteelliset frekvenssit ilmaistaan usein *prosentteina*.

ESIMERKKI 1.

Vuoden 1991 eduskuntavaaleissa Keskustapuolue sai 24.8% annetuista äänistä. Tähän tietoon voidaan liittää todennäköisyystulkinta seuraavasti:

Todennäköisyys, että satunnaisesti valittu äänioikeutensa käyttänyt henkilö äänesti Keskustapuoluetta oli 0.248. ●

Edellä esitettyyn empiirisen todennäköisyyden määritelmään sisältyi ajatus siitä, että suhteellinen frekvenssi käyttäytyy stabiilisti, jos satunnaisilmiön toistumiskertojen lukumäärän n annetaan kasvaa rajatta. Tämä saa tarkan matemaattisen muotoilun ns. *suurten lukujen laissa*. Suurten lukujen lain mukaan tapahtuman A suhteellinen frekvenssi lähestyy tapahtuman A todennäköisyyttä, kun n kasvaa rajatta. Lähestyminen tapahtuu siten, että suurten poikkeamien todennäköisyys tulee pieneksi.

SUBJETIIVINEN TODENNÄKÖISYYS

Koska empiirinen todennäköisyys ei mahdollista ainutkertaisten tapahtumien todennäköisyyksien määräämistä, ainutkertaisten tapahtumien todennäköisyyksille ei voida antaa frekvenssitulkintaa. Ainutkertaisille tapahtumille on tarjottu tulkinnaksi *vedonlyöntisuhteiden* määrittelemiä *subjektiivisia todennäköisyyksiä*:

SUBJEKTIIVINEN TODENNÄKÖISYYS

Jos henkilö suostuu lyömään tapahtuman A sattumisesta vetoa käyttäen vedonlyöntisuhdetta K:L, hänen *subjektiivinen todennäköisyytensä* tapahtuman A sattumiselle on

$$P(A) = \frac{K}{K+L}.$$

Ainutkertaisen tapahtuman subjektiivinen todennäköisyys on siis henkilökohtainen ja realisoituu siitä vedonlyöntisuhteesta, johon henkilö suostuu. Tapahtuman subjektiivinen todennäköisyys mittaa henkilökohtaista uskottavuuden astetta tapahtuman toteutumiselle.

ESIMERKKI 2.

Englannissa on mahdollista lyödä vetoa esimerkiksi siitä eroavatko Charles ja Diana kuluvan vuoden aikana. Jos henkilö suostuu vedonlyöntisuhteeseen 3:1 eron puolesta, tarkoittaa se sitä, että ko. henkilön subjektiivinen todennäköisyys erolle on

$$\frac{3}{3+1}=\frac{3}{4}.$$

Jos henkilö sijoittaa vetoon 3 mk, hän saa 4 mk, jos C. ja D. eroavat kuluvan vuoden aikana. Jos C. ja D. eivät eroa kuluvan vuoden aikana, henkilö menettää rahansa. ●

1.1.4 KLASSINEN TODENNÄKÖISYYS

Klassinen todennäköisyyden teoria käsittelee sellaisia äärellisiä otosavaruuksia, joissa alkeistapahtumilla eli tulosvaihtoehdoilla on sama todennäköisyys. Tällaisessa tapauksessa sanotaan, että tulosvaihtoehdot ovat *symmetrisiä*. Useimpia uhkapelejä voidaan kuvata äärellisillä otosavaruuksilla, joissa tulosvaihtoehdot ovat symmetrisiä. Tulosvaihtoehtojen symmetrisyys liittyy tällöin rahojen, noppien, korttipakan, rulettipyörän tms. *fysikaaliseen symmetriaan*. Uhkapeliin osallistuvat olettavat, että *reilussa* pelissä pelivälineet ovat symmetrisiä. Tällöin pelivälineitä kutsutaan tavallisesti *harhattomiksi*. Symmetria-argumentti mahdollistaa todennäköisyyksien määräämisen päättelyn keinoin.

SYMMETRISET TULOSVAIHTOEHDOT

Jos satunnaisilmiöllä on n tulosvaihtoehtoa, joilla on sama todennäköisyys, tulosvaihtoehdot ovat *symmetrisiä*. Toisin sanoen, äärellisen otosavaruuden $S=\{a_1,a_2,\ldots,a_n\}$ tulosvaihtoehdot ovat *symmetrisiä*, jos

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

kaikille tulosvaihtoehdoille a_i.

Symmetriset alkeistapahtumat mahdollistavat tapahtuman klassisen todennäköisyyden määritelmän:

KLASSINEN TODENNÄKÖISYYS

Symmetristen tulosvaihtoehtojen tapauksessa tapahtuman A *klassinen todennäköisyys* saadaan lausekkeesta

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

jossa k on A:han liittyvien tulosvaihtoehtojen lukumäärä eli tapahtumalle A *suotuisien* tulosvaihtoehtojen lukumäärä ja n on kaikkien mahdollisten tulosvaihtoehtojen lukumäärä eli tulosvaihtoehtojen lukumäärä otosavaruudessa S.

ESIMERKKI 1.

Yhden nopan heitossa otosavaruus muodostuu silmäluvuista 1,2,3,4,5,6:

$$S=\{1,2,3,4,5,6\}$$

Olettamalla noppa fysikaalisesti symmetriseksi voidaan päätellä, että jokaisen silmäluvun todennäköisyys on 1/6. Tällöin noppa on siis harhaton.

Nopan heiton tuloksena on siis 6 erilaista tulosvaihtoehtoa. Jos

A = "Silmäluku on pariton"

$= \{1,3,5\}$,

niin tapahtumalle A suotuisia alkeistapahtumia on 3. Siten

$P(A) = 3/6 = 1/2$. ●

Klassiset todennäköisyydet lasketaan tavallisesti päättelemällä. Useissa tilanteissa otosavaruuden alkeistapahtumien lukumäärän ja tarkastelun kohteena olevaan tapahtumaan liittyvien alkeistapahtumien lukumäärän määrääminen luettelemalla alkeistapahtumat on käytännössä mahdotonta. Tällöin lukumäärien määräämisessä käytetään tavallisesti apuna matematiikan osa-aluetta, jota kutsutaan *kombinatoriikaksi*.

1.1.5 KOMBINATORIIKKA

Oletetaan, että kaupungista A pääsee kaupunkiin B 3:a tietä tai 3:lla eri tavalla ja kaupungista B pääsee kaupunkiin C 2:ta tietä tai 2:lla eri tavalla. Monellako eri tavalla A:sta voidaan mennä C:hen? Vastaus on ilmeisesti $3\times2 = 6$ eri tavalla.

Yleisemmin: Oletetaan, että jokin *operaatio* voidaan suorittaa usealla vaihtoehtoisella tavalla. Vaihtoehtojen lukumäärän laskeminen voidaan tehdä käyttämällä apuna matemaattista teoriaa, jota kutsutaan *kombinatoriikaksi*.

ESIMERKKI 1.

Tarkastellaan rintasyövän hoitoa.

Rintasyöpä voidaan leikata kolmella eri tavalla:

1. Säästäen rinta.
2. Poistamalla rinta.
3. Poistamalla rinta ja myös imusolmukkeet kainalosta.

Leikkauksen jälkeen potilaalle voidaan antaa jälkihoitona:

1. Sädehoitoa.
2. Sytostaatteja.

Hoito koostuu 2:sta erillisestä operaatiosta: leikkauksesta ja jälkihoidosta. Leikkaushoitoa voidaan antaa 3:lla eri tavalla ja jälkihoitoa voidaan antaa kahdella eri tavalla. Oletetaan, että jälkihoidon valinta voidaan tehdä riippumattomasti leikkaushoidon valinnasta.

On syytä huomata, että tilanne on analoginen kappaleen alussa mainitun tien valinnan kanssa. Voidaan siis päätellä, että potilaan hoito voidaan toteuttaa $3 \times 2 = 6$ eri tavalla.

Tilastollisen tutkimuksen kohteena voisi olla eri hoitomuotojen vertailu. Vertailu voitaisiin perustaa esimerkiksi 5:n vuoden jälkeen elossa olevien potilaiden lukumäärään. Tällainen suora vertailu on kuitenkin mielekästä vain siinä tapauksessa, että vaihtoehtoisia hoitomenetelmiä voidaan soveltaa samanlaisille potilaille (yhtä vakaville tapauksille). ●

Kuten edellä toedettiin, kombinatoriikkaa tarvitaan todennäköisyyslaskennassa laskettaessa otosavaruuden ja tarkastelun kohteena olevan tapahtuman tulosvaihtoehtojen lukumäärät sellaisten otosavaruuksien yhteydessä, joissa tapahtumavaihtoehdot ovat symmetrisiä.

KOMBINATORIIKAN PERUSONGELMAT

Kombinatoriikassa on kolme perusongelmaa:

KOMBINATORIIKAN PERUSONGELMAT

Olkoon $A = \{a_1,\ldots,a_n\}$ jokin äärellinen joukko.

1. Monellako tavalla joukon A alkiot voidaan järjestää *jonoon*?
2. Montako k:n alkion *jonoa* voidaan muodostaa joukon A alkioista?
3. Monellako tavalla joukon A alkioista voidaan valita k:n alkion *osajoukko*?

Huomaa, että *jono* on joukkoa, jonka alkiot on *järjestytty*, kun taas *osajoukko* on joukko, jonka alkioiden *järjestyksellä ei ole merkitystä*. Määrittelemme usein jonon luettelemalla sen alkiot *kaarisulkujen* sisällä. Kuten edellä on todettu, joukko määritellään usein luettelemalla sen alkiot *aaltosulkujen* sisällä.

ESIMERKKI 2.

Jonot (1,2,3) ja (2,1,3) ovat eri jonoja. Sen sijaan joukot {1,2,3} ja {2,1,3} ovat samat. ●

Jos A:ssa on vähän alkioita, saadaan vastaukset kombinatoriikan perusongelmiin luettelemalla eri vaihtoehdot.

ESIMERKKI 3.

Olkoon $A = \{1,2,3\}$.

A:n alkioiden järjestetyt jonot:

123, 132, 213, 231, 312, 321 (6 kpl)

2:n alkion muodostamat järjestetyt jonot:

12, 21, 13, 31, 23, 32 (6 kpl)

2:n alkion osajoukot:

{1,2},{1,3},{2,3} (3 kpl)

Kaikki A:n osajoukot:

{1,2,3},{1,2},{1,3},{2,3},{1},{2},{3},∅ (8 kpl)

Merkintä ∅ tarkoittaa ns. tyhjää joukkoa, jossa ei ole yhtään alkiota. ●

Jos A:ssa on paljon alkioita, alkioiden luetteloinnin ei ole käytännössä mahdollista. Tällöin on käytettävä kombinatoriikan tarjoamia apuvälineitä.

KOMBINATORIIKAN PERUSPERIAATTEET

Suurin osa kombinatoriikan laskutoimituksista perustuu seuraavien kahden periaatteen soveltamiseen:

1. *Yhteenlaskuperiaate*:

Oletetaan, että operaatioita M ja N ei voida suorittaa yhtäaikaa, ts. M ja N ovat *toisensa poissulkevia*. Oletetaan, että M voidaan suorittaa n_1 tavalla ja N voidaan suorittaa n_2 tavalla. Tällöin yhdistetty operaatio

"Suoritetaan M *tai* N"

voidaan suorittaa $n_1 + n_2$ tavalla.

2. *Kertolaskuperiaate*:

Oletetaan, että operaatiot M ja N voidaan suorittaa peräkkäin tai yhtä aikaa toisistaan *riippumattomasti*. Riippumattomuudella tarkoitetaan seuraavaa: Se, mikä vaihtoehdoista on valittu operaatiota M suoritettaessa ei vaikuta siihen, mikä vaihtoehdoista valitaan operaatiota N suoritettaessa. Oletetaan, että M voidaan suorittaa n_1 tavalla ja N voidaan suorittaa n_2 tavalla. Tällöin yhdistetty operaatio

"Suoritetaan M *ja* N"

voidaan suorittaa $n_1 n_2$ tavalla.

PERMUTAATIOT

PERMUTAATIO

Joukon alkioiden *permutaatio* on mikä tahansa alkioiden jono eli järjestetty joukko.

Olkoon $A = \{a_1,\ldots,a_n\}$. Merkitään

$n_A = \mathrm{n}(A) = n = A$:n alkioiden lukumäärä.

Seuraava lause ratkaisee kombinatoriikan 1. perusongelman:

LAUSE 1.

Olkoon $\mathrm{n}(A) = n$. Tällöin A:n alkioiden kaikkien permutaatioiden eli järjestettyjen jonojen lukumäärä on

$$n! = n\cdot(n-1)\cdot\ldots\cdot 2\cdot 1,$$

jossa $n!$ on ns. *n-kertoma*.

Todistus:

Kuvaamme seuraavassa A:n alkioiden järjestämistä jonoon ns. *lokeromallilla*.

Ajatellaan, että käytettävissä on n tyhjää lokeroa, joihin A:n alkiot asetetaan niin, että kuhunkin lokeroon tulee täsmälleen 1 alkio. Lokeroiden täyttäminen voidaan tehdä vaiheittain seuraavasti:

Vaihe 1: Valitaan A:sta alkio 1. lokeroon. Valinta voidaan tehdä n eri tavalla, koska A:ssa on n alkiota.

Vaihe 2: Valitaan A:sta alkio 2. lokeroon. Valinta voidaan tehdä $(n-1)$ eri tavalla, koska A:ssa on vaiheen 1 jälkeen $(n-1)$ alkiota jäljellä.

Vaihe 3: Valitaan A:sta alkio 3. lokeroon. Valinta voidaan tehdä $(n-2)$ eri tavalla, koska A:ssa on vaiheen 2 jälkeen $(n-2)$ alkiota jäljellä.

...

Vaihe $(n-1)$: Valitaan A:sta alkio $(n-1)$. lokeroon. Valinta voidaan tehdä 2:lla eri tavalla, koska A:ssa on vaiheen $(n-2)$ jälkeen 2 alkiota jäljellä.

Vaihe n: Valitaan A:sta alkio n. lokeroon. Valinta voidaan tehdä 1:llä tavalla, koska A:ssa on vaiheen $(n-1)$ jälkeen 1 alkio jäljellä.

Jokaisessa vaiheessa lokeroon pantavan A:n alkion valinta voidaan tehdä riippumattomasti edellisissä vaiheissa tehdyistä valinnoista. Kertolasku-periaatteen mukaan lokeroiden täyttäminen voidaan tehdä

$n\cdot(n-1)\cdot\ldots\cdot 2\cdot 1 = n!$ eri tavalla. ■

n-kertoma voidaan määritellä *rekursiivisesti* seuraavalla kaavalla:

N-KERTOMA

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

ESIMERKKI 4.

Määritellään

$$0! = 1.$$

Kertoman määritelmän mukaan

$$1! = 1$$

$$2! = 2\cdot 1 = 2$$

$$3! = 3\cdot 2\cdot 1 = 6$$

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

$$5! = 5\cdot 4\cdot 3\cdot 2\cdot 1 = 120$$

jne

Kertomat kasvavat hyvin nopeasti. ●

Seuraava lause ratkaisee kombinatoriikan 2. perusongelman:

LAUSE 2.

Olkoon $n(A) = n$. Tällöin A:n alkioiden k alkiota sisältävien permutaatioiden eli järjestettyjen jonojen lukumäärä on

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

Todistus:

Edellisen lauseen todistuksesta nähdään, että n:stä alkiosta voidaan valita k alkiota k lokeroon

$$n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$$

eri tavalla. Laventamalla saadaan

$$n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$$

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

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

■

KOMBINAATIOT

KOMBINAATIO.

Joukon alkioiden *kombinaatio* on mikä tahansa joukon alkioiden osajoukko.

Kombinaatio eroaa permutaatiosta siten, että kombinaatiossa ei alkioiden järjestyksellä ole merkitystä, kun taas permutaatiossa on.

Seuraava lause ratkaisee kombinatoriikan 3. perusongelman:

LAUSE 3.

Olkoon n(A) = n . Tällöin A:n alkioiden k alkiota sisältävien kombinaatioiden eli osajoukkojen lukumäärä on

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

Todistus:

Merkitään C(n,k):llä niiden tapojen lukumäärää, joilla n:n alkion joukosta voidaan valita k:n alkion osajoukko, ts. C(n,k) on kysytty (toistaiseksi tuntematon) kombinaatioiden lukumäärä.

Ratkaistaan kombinaatioiden lukumäärän ongelma määräämällä n:n alkion joukon k alkiota sisältävien *permutaatioiden* lukumäärä kahdella eri tavalla.

Lauseen 2 mukaan n:n alkion joukon k alkiota sisältävien permutaatioiden lukumäärä on

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

Permutaatioiden lukumäärä voidaan laskea myös seuraavalla tavalla, jossa permutaatiot tehdään kahdessa vaiheessa:

Valitaan n:n alkion joukosta k alkiota sisältävä osajoukko eli kombinaatio. Tämä voidaan tehdä $C(n,k)$ eri tavalla, jossa $C(n,k)$ on kysytty kombinaatioiden lukumäärä.

Järjestetään valitut k alkiota jonoon. Tämä voidaan tehdä lauseen 1 mukaan $k!$ eri tavalla.

On syytä huomata, että vaiheet 1 ja 2 voidaan tehdä toisistaan riippumattomasti.

Siten n:n alkion joukon k alkiota sisältävien permutaatioiden lukumäärä on kertolaskuperiaatteen nojalla $C(n,k)\cdot k!$, jonka pitää siis yhtyä lauseesta 2 saatavaan lukumäärään. Siten

$$C(n,k)\cdot k! = \frac{n!}{(n-k)!},$$

josta saadaan

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

Kombinaatioiden lukumäärän ilmaiseva lauseke

$$C(n,k) = \binom{n}{k}$$

luetaan "n yli k:n".

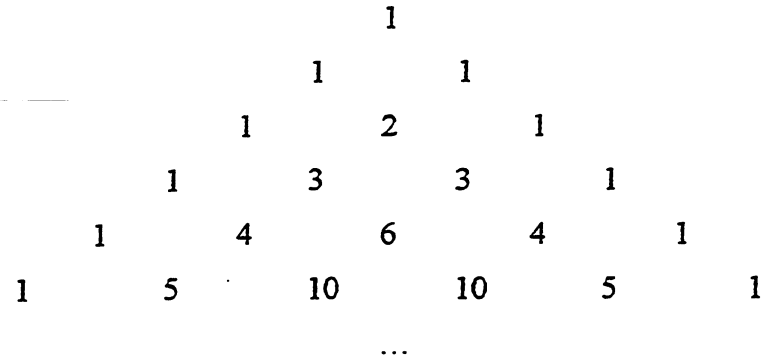
Koska $0! = 1$, niin

$$\binom{n}{0} = \frac{n!}{0!n!} = 1 = \frac{n!}{n!0!} = \binom{n}{n}.$$

Luvut $C(n,k)$ voidaan määrätä helposti ns. *Pascalin kolmion* avulla, jos n ei ole kovin suuri.

PASCALIN KOLMIO

Seuraavaa lukukaaviota kutsutaan *Pascalin kolmioksi*:



Jokaisen rivin luvut saadaan 3:sta rivistä alkaen (rivin 1. ja viimeistä lukua lukuun ottamatta) välittömästi luvun yläpuolella olevien kahden luvun summana. Esimerkiksi 6:lla rivillä

$$5 = 1 + 4$$

ja

$$10 = 4 + 6.$$

Miten Pascalin kolmio liittyy kombinaatioiden lukumäärään? Voidaan osoittaa, että (n+1). rivin luvut ovat samat kuin n:stä luvusta valittujen 0:n, 1:n, 2:n,..., n:n luvun kombinaatioiden eli osajoukkojen lukumäärät. Siten Pascalin kolmion (n+1). rivin luvut ovat

$$C(n,0),\ C(n,1),\ C(n,2),\ldots,\ C(n,n-1),\ C(n,n).$$

Pascalin kolmion lukujen muodostamisessa käytetty sääntö voidaan ilmaista kaavan muodossa seuraavasti:

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

Siinä (n+1). rivin k. luku on ilmaistu n. rivin ($k-1$). luvun ja k. luvun summana.

Pascalin kolmio on symmetrinen rivin keskikohdan suhteen, mikä voidaan ilmaista kaavan muodossa seuraavasti:

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

Pascalin kolmion lukuja sanotaan matematiikassa *binomikertoimiksi*, koska Pascalin kolmion luvut ovat binomin $(x+y)$ peräkkäisten potenssien kehitelmien kertoimia:

LAUSE 4. BINOMILAUSE

$$(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \ldots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n$$

$$= \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k.$$

Todistus:

Kun binomi $(x+y)$ korotetaan potenssiin n, saadaan summalauseke, jonka kaikki termit ovat muotoa $x^{n-k}y^k$, jossa $k = 0,1,2,\ldots,n$. Yhdistetään samaa muotoa olevat termit ja järjestetään termit x:n laskevien potenssien mukaan. Tällöin saadaan $(n+1)$ termiä sisältävä summalauseke, jonka $(k+1)$. termi on muotoa $x^{n-k}y^k$. Kaikki muotoa $x^{n-k}y^k$ olevat termit ovat tuloja, joissa on n tekijää. Tehtävänä on laskea, kuinka monella tavalla tällainen tulo syntyy potenssiin korotuksessa.

Käytetään tilanteen kuvaamiseen jälleen lokeromallia. Oletetaan, että käytössä on kahta eri lajia olevia olioita, joista toiset ovat tyyppiä x ja toiset tyyppiä y. Tehtävänä on täyttää n lokeroa $(n-k)$:lla x-tyypin olioilla, jolloin jäljelle jääneet k lokeroa tulevat automaattisesti täytetyksi y-tyypin olioilla, ja laskea kuinka monella tavalla tämä voidaan tehdä.

Koska tyypin x olioita ei voida erottaa toisistaan, niiden järjestyksellä ei ole mahdollisten sijoitusten lukumäärää laskettaessa merkitystä. Siksi ongelma voidaan muuntaa seuraavaan muotoon: Miten monella eri tavalla n:n alkion joukosta voidaan valita $(n-k)$ alkiota, kun alkioiden järjestykseen ei tarvitse kiinnittää huomiota? Tähän ongelmaan antaa vastauksen lause 3.

Lauseesta 3 seuraa, että x voi esiintyä muotoa $x^{n-k}y^k$ olevassa tulossa $C(n,k)$ eri paikassa y:n joutuessa jäljelle jääville paikoille. Termin $x^{n-k}y^k$ kertoimeksi saadaan siten $C(n,k)$. ■

ESIMERKKI 5.

Tarkastellaan binomilausetta tapauksessa $n = 4$ ja $k = 2$ ja erityisesti sitä, kuinka monella tavalla tulo x^2y^2 syntyy korotettaessa binomi $(x+y)$ potenssiin 4.

Kaikki mahdolliset muotoa x^2y^2 olevat tulot ovat

xxyy	*xyxy*	*xyyx*
yxxy	*yxyx*	*yyxx*

joita on 6 kappaletta.

Binomikertoimen kaavan mukaan

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{4\cdot 3\cdot 2\cdot 1}{2\cdot 1\cdot 2\cdot 1} = 6. \bullet$$

Binomikertoimilla on paljon sovelluksia todennäköisyyslaskennassa ja tilastotieteessä. Esimerkiksi ns. *binomitodennäköisyyden* kaava on suora sovellus lauseesta 4.

1.1.6 ESIMERKKEJÄ KOMBINATORISISTA LASKUTOIMITUKSISTA

ESIMERKKI 1.

Tarkastellaan kymmenjärjestelmän numeroita 0,1,2,3,4,5,6,8,9 (10 kpl). Kuinka monta 3-numeroista lukua voidaan muodostaa näistä luvuista?

Sovelletaan jälleen lokeromallia. Tehtävänä on täyttää 3 lokeroa 10:llä erilaisella esineellä. Lisäksi oletetaan, että jokaista erilaista esinettä on käytettävissä niin monta, että kaikki 3 lokeroa voidaan haluttaessa täyttää samanlaisilla esineillä. Tällöin

1. lokero voidaan täyttää 10 eri tavalla,
2. lokero voidaan täyttää 10 eri tavalla,
3. lokero voidaan täyttää 10 eri tavalla.

Kertolaskuperiaatteen mukaan lokerot voidaan täyttää

$$10 \cdot 10 \cdot 10 = 1{,}000$$

eri tavalla.

Tulos on tietysti muutenkin selvä, koska kolminumeroisia lukuja on 1,000 kpl: 000, 001, 002,..., 999. ●

ESIMERKKI 2.

Olkoon tilanne muuten sama kuin esimerkissä 1, mutta lisätään vaatimus, jonka mukaan sama numero saa esiintyä vain kerran kussakin luvussa.

Käytetään uudelleen esimerkin 1 lokeromallia:

1. lokero voidaan täyttää 10 eri tavalla,
2. lokero voidaan täyttää 9 eri tavalla, koska 1 luku on jo käytetty,
3. lokero voidaan täyttää 8 eri tavalla, koska 2 lukua on jo käytetty.

Kertolaskuperiaatteen mukaan lokerot voidaan täyttää nyt

$$10 \cdot 9 \cdot 8 = 720$$

eri tavalla.

Samaan tulokseen päästään myös seuraavassa esitettävällä vaihtoehtoisella tavalla muuntamalla ongelma seuraavaan muotoon: Kuinka monella tavalla numerot 0,1,2,3,4,5,6,8,9 voidaan asettaa 3 numeroa sisältävään järjestettyyn jonoon? Tähän antaa vastauksen lause 2, jonka mukaan 10:n alkion joukon 3 alkiota sisältäviä permutaatioita on

$$P(10,3) = \frac{10!}{(10-3)!} = \frac{10!}{7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{7!} = 10 \cdot 9 \cdot 8 = 720 \text{ kpl.} \bullet$$

Huomaa, että esimerkit 1 ja 2 kuvaavat olennaisesti 3:n *alkion otoksen poimintaa* 1,000:n alkion otoksesta, kun otanta tehdään *takaisinpanolla* (esimerkki 1) ja *ilman takaisinpanoa* (esimerkki 2).

ESIMERKKI 3.

Montako erilaista rekisterikilpeä voidaan muodostaa, jos kaikki rekisterikilvet ovat muotoa

XXnnn,

jossa

X tarkoittaa vokaaleja A, E, I, O, U, Y

ja

n tarkoittaa numeroita 0,1,2,3,4,5,6,8,9.

Oletetaan lisäksi, että sama numero tai kirjain saa esiintyä useamman kerran kilvessä.

Kertolaskuperiaatteen mukaan

XX-jonoja on 6·6 = 36 kpl ja

nnn-jonoja on 10·10·10 = 1,000 kpl.

Siten erilaisia rekisterikilpiä on *kertolaskuperiaatteen* mukaan kaikkiaan

36·1,000 = 36,000 kpl. ●

ESIMERKKI 4.

Montako erilaista lottoruudukkoa on olemassa?

Lotossa 1 ruudukko muodostuu valitsemalla 7 numeroa 39:stä. Huomaa, että samaa numeroa ei voi valita kuin kerran ja, että numeroiden järjestyksellä ei ole väliä.

Kyse on siis siitä, kuinka monta 7 alkion osajoukkoa voidaan valita 39:n alkion muodostamasta joukosta.

Vastauksen antaa lause 3. Sen mukaan erilaisia lottoruudukoita on

$$\binom{39}{7} = \frac{39!}{7!32!} = 15{,}380{,}937 \text{ kpl.} \bullet$$

ESIMERKKI 5.

Montako sellaista lottoruudukkoa on olemassa, joissa on 5 oikein?

Jotta saataisiin *täsmälleen* 5 oikein, on valittava 5 numeroa 7 *oikean* numeron joukosta ja 2 numeroa 32 *väärän* numeron joukosta.

5 numeroa voidaan valita 7 oikean numeron joukosta

$$\binom{7}{5}$$

eri tavalla.

2 numeroa voidaan valita 32 väärän numeron joukosta

$$\binom{32}{2}$$

eri tavalla.

Kertolaskuperiaatteen mukaan 5 oikein voidaan siis lotota

$$\binom{7}{5}\binom{32}{2} = \frac{7!}{5!2!} \cdot \frac{32!}{2!30!} = 21 \cdot 496 = 10{,}416$$

eri tavalla. ●

Yhdistämällä esimerkit 4 ja 5 voidaan määrätä todennäköisyys sille, että saadaan 5 oikein lotossa:

ESIMERKKI 6.

Määrätään todennäköisyys saada 5 oikein lotossa. Ratkaistaan tehtävä klassisen todennäköisyyden määritelmän avulla.

Otosavaruuden S muodostavat kaikki erilaiset lottoruudukot. Esimerkin 4 mukaan sellaisia on 15,380,927 kpl.

Haluamme siis tietää tapahtuman

A = "Saadaan 5 oikein lotossa"

todennäköisyys.

Tapahtumalle A suotuisien tulosvaihtoehtojen lukumäärä on esimerkin 5 mukaan 10,416 kpl.

Klassisen todennäköisyyden määritelmän mukaan

$P(A) = 10{,}416/15{,}380{,}927 \approx 0.000677$. ●

ESIMERKKI 7.

Monissa korttipeleissä pelaajalle jaetaan 5 korttia 52:n kortin pakasta (esim. pokerissa). Montako erilaista 5:n kortin "kättä" on olemassa?

Lauseen 3 mukaan erilaisia 5:n kortin käsiä on

$$\binom{52}{5} = 2{,}598{,}960 \text{ kpl.}$$ ●

ESIMERKKI 8.

Monissa korttipeleissä pelaajalle jaetaan 13 korttia 52:n kortin pakasta (esim. bridgessä). Montako erilaista 13:n kortin "kättä" on olemassa?

Lauseen 3 mukaan erilaisia 13:n kortin käsiä on

$$\binom{52}{13} = 635{,}013{,}559{,}600 \text{ kpl.} \bullet$$

ESIMERKKI 9.

Montako erilaista sanaa, joissa on 2, 3 tai 4 kirjainta voidaan muodostaa vokaaleista a, e, i, o, u, y?

Kertolaskuperiaatteen mukaan:

mahdollisten 2-kirjaimisten sanojen lukumäärä on $6^2 = 36$,

mahdollisten 3-kirjaimisten sanojen lukumäärä on $6^3 = 216$,

mahdollisten 4-kirjaimisten sanojen lukumäärä on $6^4 = 1296$.

Koska sanat eivät voi olla samanaikaisesti 2- ja 3-kirjaimisia, operaatiot

"Muodosta 2-kirjaiminen sana"

ja

"Muodosta 3-kirjaiminen sana"

ovat toisensa poissulkevia. Sama pätee 2- ja 4-kirjaimisia ja 3- ja 4-kirjaimisia sanoille.

Siten voimme soveltaa *yhteenlaskuperiaatetta*, jonka mukaan 2-, 3- tai 4-kirjaimisia sanoja on

$$36 + 216 + 1{,}296 = 1{,}548. \bullet$$

ESIMERKKI 10.

Tilastotieteen laitoksen johtoryhmässä on 6 jäsentä. Ne edustavat eri henkilöstöryhmiä seuraavasti (henkilökunta syksyllä 1995):

Professorit ja apulaisprofessorit (4): 2 jäsentä

- 2 professoria
- 2 apulaisprofessoria

Muu henkilökunta (9): 2 jäsentä

- 2 yliassistenttia
- 3 assistenttia
- 2 lehtoria
- 1 tuntiopettaja
- 1 toimistosihteeri

Opiskelijat (n. 80): 2 jäsentä

Montako erilaista johtoryhmää on mahdollista valita?

Kertolaskuperiaatteeen mukaan erilaisia johtoryhmiä on

$$\binom{4}{2}\binom{9}{2}\binom{80}{2} = 682{,}560 \text{ kpl.} \bullet$$

1.1.7 MULTINOMIKERROIN

Binomikerroin $C(n,k)$ ilmaisee n:n alkion joukon k alkiota sisältävien osajoukkojen lukumäärän. Huomaa, että osajoukon valinta voidaan ajatella tehtävän myös seuraavalla tavalla: Jaetaan n:n alkion joukko kahteen osaan, joissa ei ole yhteisiä alkioita siten, että toiseen osaan tulee k alkiota ja toiseen loput $(n-k)$ alkiota.

Binomikerroin voidaan yleistää ns. *multinomikertoimeksi*, joka ilmaisee kuinka monella tavalla n:n alkion joukko voidaan jakaa k osaan, joissa ei ole yhteisiä alkioita eli *k pisteviraaseen* osaan siten, että

osajoukkoon 1 valitaan n_1 alkiota,

osajoukkoon 2 valitaan n_2 alkiota,

...

osajoukkoon k valitaan n_k alkiota.

Koska osat eivät saa sisältää yhteisiä alkioita,

$$n_1 + n_2 + \ldots + n_k = n.$$

Huomaa, että vaiheessa k alkioita ei itse asiassa voida enää valita, vaan jäljelle jääneet n_k alkiota *joutuvat k.* osajoukkoon.

Seuraava lause antaa erilaisten jakojen lukumäärän, joissa n:n alkion joukko jaetaan k pistevieraaseen osajoukkoon siten, että osajoukoissa on $n_1, n_2, \ldots, n_k$ alkiota. On syytä huomata, että lause ratkaisee myös seuraavaan ongelman:

> Olkoon n:n alkion joukossa k keskenään samanlaisten alkioiden ryhmää ja olkoot ryhmäkoot $n_1, n_2, \ldots$ ja n_k. Kuinka monella tavalla joukon alkiot voidaan järjestää jonoon?

Tämä ongelma voidaan muotoilla myös seuraavaan yhtäpitävään muotoon:

> Kuinka monta erilaista permutaatiota voidaan muodostaa n:n alkion joukosta, jos alkiot muodostavat k keskenään samanlaisten alkioiden ryhmää, joiden ryhmäkoot ovat $n_1, n_2, \ldots$ ja n_k?

LAUSE 1.

n:n alkion joukosta, joka jaetaan k:hon pistevieraaseen osajoukkoon, joissa alkioiden lukumäärät ovat $n_1, n_2, \ldots, n_k$, voidaan muodostaa

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

erilaista osajoukkoa.

Todistus:

Tehdään osajoukkoihin jako vaiheittain:

Vaihe 1: Valitaan n:stä alkiosta n_1 alkiota.

Tämä voidaan tehdä $C(n,n_1)$ eri tavalla.

Vaihe 2: Valitaan jäljelle jääneistä $(n-n_1)$:stä alkiosta n_2 alkiota.

Tämä voidaan tehdä $C(n-n_1,n_2)$ eri tavalla.

Vaihe 3: Valitaan jäljelle jääneistä $(n-n_1-n_2)$:sta alkiosta n_3 alkiota.

Tämä voidaan tehdä $C(n-n_1-n_2,n_3)$ eri tavalla.

...

Vaihe $k-1$: Valitaan jäljelle jääneistä $(n-n_1-n_2-\ldots-n_{k-2})$:sta alkiosta n_{k-1} alkiota.

Tämä voidaan tehdä $C(n-n_1-n_2-\ldots-n_{k-2},n_{k-1})$ eri tavalla.

Vaihe k: Valitaan jäljelle jääneistä $(n-n_1-n_2-\ldots-n_{k-2}-n_{k-1})$:stä alkiosta n_k alkiota.

Tämä voidaan tehdä $C(n-n_1-n_2-\ldots-n_{k-2}-n_{k-1},n_k)$ eri tavalla.

Jokaisessa vaiheessa on sovellettu kappaleen 1.1.5 lausetta 3.

Kertolaskuperiatteen mukaan osajoukkoihin jako voidaan tehdä

$$C(n,n_1)\cdot C(n-n_1,n_2)\cdot C(n-n_1-n_2,n_3)\cdots \cdot C(n-n_1-n_2-\ldots-n_{k-2},n_{k-1})\cdot C(n-n_1-n_2-\ldots-n_{k-2}-n_{k-1},n_k)$$

$$=\binom{n}{n_1}\cdot\binom{n-n_1}{n_2}\cdot\binom{n-n_1-n_2}{n_3}\cdots \cdot\binom{n-n_1-n_2-\ldots-n_{k-2}}{n_{k-1}}\binom{n-n_1-n_2-\ldots-n_{k-2}-n_{k-1}}{n_k}$$

$$=\frac{n!}{n_1!(n-n_1)!}\cdot\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdot\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}\cdots \cdot\frac{(n-n_1-n_2-\ldots-n_{k-2})!}{n_{k-1}!(n-n_1-n_2-\ldots-n_{k-2}-n_{k-1})!}\cdot\frac{n_k!}{n_k!0!}$$

$$=\frac{n!}{n_1!n_2!n_3!\cdots n_{k-1}!n_k!}$$

eri tavalla.

Lausekkeen supistuminen yhtälöketjun viimeisessä lenkissä perustuu siihen, että tulon peräkkäisissä tekijöissä edellisen tekijän nimittäjässä on aina sama lauseke kuin seuraavan tekijän osoittajassa. Kahdessa viimeisessä tekijässä on lisäksi käytetty hyväksi sitä, että

$$n-n_1-n_2-\ldots-n_{k-2}-n_{k-1}=n_k. \blacksquare$$

ESIMERKKI 1.

Sanassa *kassa* on 5 kirjainta, joiden joukossa on 3 erilaista kirjainta:

a: 2 kpl

k: 1 kpl

s: 2 kpl

Lauseen 1 mukaan näistä kirjaimista voidaan muodostaa

$$\binom{5}{2\ 1\ 2} = \frac{5!}{2!1!2!} = 30$$

erilaista sanaa. ●

ESIMERKKI 2.

Monissa korttipeleissä (esim. pokerissa) jokaiselle pelaajalle jaetaan 5 korttia. Oletetaan, että pelaajia on 4. Kuinka monta erilaista "kättä" voidaan pelaajille jakaa?

Ratkaisu 1:

Vaihe 1: 1. pelaajalle jaetaan 5 korttia 52:sta.
Tämä voidaan tehdä C(52,5) eri tavalla.

Vaihe 2: 2. pelaajalle jaetaan 5 korttia jäljelle jääneistä 47:stä.
Tämä voidaan tehdä C(47,5) eri tavalla.

Vaihe 3: 3. pelaajalle jaetaan 5 korttia jäljelle jääneistä 42:stä.
Tämä voidaan tehdä C(42,5) eri tavalla.

Vaihe 4: 4. pelaajalle jaetaan 5 korttia jäljelle jääneistä 37:stä.
Tämä voidaan tehdä C(37,5) eri tavalla.

Kertolaskuperiaatteen mukaan kortit voidaan jakaa pelaajille

$$\binom{52}{5}\binom{47}{5}\binom{42}{5}\binom{37}{5} = \frac{52!47!42!37!}{5!47!5!42!5!37!5!32!} = \frac{52!}{5!5!5!5!32!} \approx 1.47 \times 10^{24}$$

eri tavalla.

Huomaa, että jakoa ei tavallisesti tehdä näin, vaan pelaajille jaetaan 1 kortti kerrallaan. Jos korttipakka on hyvin sekoitettu, niin jakamisen voi yhtä hyvin tehdä 5 korttia kerrallaan.

Ratkaisu 2:

Korttien jakoa voidaan ajatella myös seuraavasti: Kun 5 korttia jaetaan kullekin 4:stä pelaajasta, jää pakkaan 32 korttia. Siten 52:n kortin pakka tulee jaetuksi 5:een osaan, joissa on 5,5,5,5 ja 32 korttia. Tällaisten jakojen lukumäärän ilmaisee multinomikerroin

$$\binom{52}{5\ \ 5\ \ 5\ \ 5\ \ 32} = \frac{52!}{5!5!5!5!32!} \approx 1.47 \times 10^{24}.$$

Tulos on tietysti sama kuin ratkaisussa 1. ●

1.1.8 TODENNÄKÖISYYDEN LAIT

Käsittelemme tässä kappalessa todennäköisyyslaskennan sääntöjä. Sääntöjä havainnollistetaan pääasiassa klassista todennäköisyyttä käsittelevillä esimerkeillä.

TOISENSA POISSULKEVAT TAPAHTUMAT

Tapahtumat A ja B ovat *toisensa poissulkevia*, jos A:n tapahtuminen estää B:n tapahtumisen ja päinvastoin: B:n tapahtuminen estää A:n tapahtumisen. Tällöin ei ole sellaista tulosvaihtoehtoa, joka liittyisi sekä tapahtumaan A että B.

ESIMERKKI 1.

Oletetaan, että haluamme tutkia satunnaisilmiönä arpajaisia, joissa on 10 arpalippua. Arpajaisissa jokaisella arvalla pitää olla sama todennäköisyys voittaa.

Edellä on todettu, että satunnaisilmiön todennäköisyysmallin muodostavat otosavaruus ja otosavaruuden alkioihin liitetyt todennäköisyydet.

Otosavaruudeksi pitää tässä tapauksessa valita lukujen 0,1,2,...,9 muodostama joukko, ts.

$$S = \{0,1,2,3,4,5,6,7,8,9\}.$$

Jokaisen luvun todennäköisyyden pitää olla 1/10, jolloin alkeistapahtumat eli tulosvaihtoehdot $\{0\},\{1\},\{2\},\ldots,\{9\}$ ovat symmetrisiä.

Olkoon

$$A = \{7,8,9\}.$$

Tapahtumalle A suotuisia tulosvaihtoehtoja on 3 kappaletta.

Siten

$$P(A) = 3/10 = 0.3.$$

Olkoon

$$B = \{0,1\}.$$

Siten

$$P(B) = 2/10 = 0.2.$$

On helppoa todeta, että tapahtumat A ja B ovat toisensa poissulkevia, koska niissä ei ole yhteisiä alkioita. Lisäksi A on todennäköisempi kuin B, koska $P(A) > P(B)$. ●

ESIMERKKI 2.

Korttipakassa on 52 korttia, jotka jakautuvat 4 maahan:

♠ = pata, ♥ = hertta, ♦ = ruutu, ♣ = risti.

Kussakin maassa on 13 korttia:

1=A(ässä),2,3,4,5,6,7,8,9,10,
11=J(sotilas),12=Q(kuningatar),13=K(kuningas).

Yksikään kortti ei voi olla samanaikaisesti pata ja ruutu. Siten tapahtumat

A = "Satunnaisesti valittu kortti on pata",

B = "Satunnaisesti valittu kortti on ruutu"

ovat toisensa poissulkevia. ●

YHTEENLASKUSÄÄNTÖ TOISENSA POISSULKEVILLE TAPAHTUMILLE

Oletetaan, että tapahtumat A ja B ovat toisensa poissulkevia. Tällöin *yhdistetyn* tapahtuman

A *tai* B tapahtuu

todennäköisyys saadaan laskemalla tapahtumien A ja B todennäköisyydet yhteen:

YHTEENLASKUSÄÄNTÖ TOISENSA POISSULKEVILLE TAPAHTUMILLE

Jos tapahtumat A ja B ovat toisensa poissulkevia, niin

$$P(A \text{ tai } B) = P(A) + P(B).$$

Yhteenlaskusäännön mukaan tapahtuman todennäköisyys voidaan laskea tapahtumaan liittyvien tulosvaihtoehtojen eli alkeistapahtumien todennäköisyyksien summana, jos perusjoukko on äärellinen: Siis jos tapahtumaan A liittyvät tulosvaihtoehdot ovat

$$a_1, a_2, \ldots, a_k$$

ja niitä vastaavat todennäköisyydet ovat

$$p_1, p_2, \ldots, p_k,$$

niin

$$P(A) = p_1 + p_2 + \ldots + p_k.$$

ESIMERKKI 3.

Tarkastellaan esimerkin 1 todennäköisyysmallia. Olkoon

A = "Satunnaisesti valitun arvan numero on pariton",

B = "Satunnaisesti valitun arvan numero on 4:n monikerta".

Parittomia numeroita ovat 1,3,5,7,9. Siten

$$P(A) = 5/10 = 0.5.$$

Luvut 0, 4 ja 8 ovat luvun 4 monikertoja. Siten

$$P(B) = 3/10 = 0.3.$$

Koska luvut 0,4,8 ovat parillisia, A ja B ovat tapahtumina toisensa poissulkevia. Siten

$$P(A \text{ tai } B) = P(A) + P(B) = 0.5 + 0.3 = 0.8. \bullet$$

ESIMERKKI 4.

Tarkastellaan esimerkin 2 todennäköisyysmallia. Olkoon

A = "Satunnaisesti valittu kortti on ♦",

B = "Satunnaisesti valittu kortti on ♣".

Tällöin

$$P(A) = P(B) = 13/52 = 0.25.$$

Koska yksikään kortti ei voi olla samanaikaisesti ♦ ja ♣, A ja B ovat toisensa poissulkevia ja

$$P(A \text{ tai } B) = P(A) + P(B) = 13/52 + 13/52 = 26/52 = 0.5. \bullet$$

Oletetaan, että satunnaisilmiö on voidaan jakaa äärelliseen määrään toisensa poissulkevia tapahtumia

$$A_1, A_2, \ldots, A_k.$$

Tämä merkitsee sitä, että *täsmälleen* yksi tapahtumista $A_1, A_2, \ldots, A_k$ sattuu aina, kun satunnaisilmiö esiintyy. Jos näiden tapahtumien todennäköisyydet ovat

$$P(A_1) = p_1, P(A_2) = p_2, \ldots, P(A_k) = p_k,$$

niin

$$p_1 + p_2 + \ldots + p_k = 1.$$

Huomaa, että yhteenlaskusääntö *ei päde* edellä esitetyssä muodossa, jos yhdistetyn tapahtuman A *tai* B tapahtumat *eivät ole* toisensa poissulkevia.

YHDISTETYN TAPAHTUMAN *A JA B* TODENNÄKÖISYYS

Edellä tarkasteltiin toisensa poissulkevien tapahtumien A ja B yhdistetyn tapahtuman A *tai* B tapahtuu todennäköisyyden määräämistä. Kaksi tapahtumaa voidaan yhdistää myös muilla tavoilla. Tarkastellaan nyt tapahtuman

A *ja* B tapahtuu

todennäköisyyden määräämistä. Tapahtuma A *ja* B sattuu, jos sekä A että B sattuvat.

ESIMERKKI 5.

Oletetaan, että

A = "Satunnaisesti valittu kortti on ♠",

B = "Satunnaisesti valittu kortti on ässä".

Tällöin yhdistetty tapahtuma A *ja* B on

A ja B = "Satunnaisesti valittu kortti on pataässä".

Koska korttipakassa on täsmälleen 1 pataässä, tapahtuman *A ja B* todennäköisyys on

$P(A \text{ ja } B) = 1/52.$ ●

Jos tapahtumat *A* ja *B* ovat *toisensa poissulkevia*,

$P(A \text{ ja } B) = 0,$

koska tällöin ei ole olemassa yhtään tulosvaihtoehtoa, joka liittyisi sekä *A*:han että *B*:hen. Jos tapahtumat *A* ja *B* ovat toisensa poissulkevia, tapahtuma *A ja B* on itse asiassa *mahdoton*.

ESIMERKKI 6.

Olkoon

A = "Satunnaisesti valittu kortti on pata"

B = "Satunnaisesti valittu kortti on hertta".

Koska yksikään kortti ei voi olla samanaikaisesti pata ja hertta,

$P(A \text{ ja } B) = 0.$ ●

YLEINEN YHTEENLASKUSÄÄNTÖ

Edellä tarkasteltiin yhdistetyn tapahtuman *A tai B* todennäköisyyden määräämistä tapahtumien *A* ja *B* todennäköisyyksien avulla siinä tapauksessa, että *A* ja *B* ovat toisensa poissulkevia. Tällöin yhdistetyn tapahtuman *A tai B* todennäköisyys voitiin laskea tapahtumien *A* ja *B* todennäköisyyksien summana.

Voidaanko yhteenlaskusääntö toisensa poissulkeville tapahtumille muuntaa muotoon, joka soveltuu myös tapauksiin, joissa tapahtumat *A* ja *B* eivät ole toisensa poissulkevia? Vastaus on *kyllä*! Tämä tapahtuu ottamalla huomioon yhdistetyn tapahtuman *A ja B* todennäköisyys. Tarkastellaan ensin seuraavia esimerkkejä:

ESIMERKKI 7.

Japanissa ei ole tarkkoja tilastoja eri uskontojen harjoittajista. Erään arvion mukaan 80% japanilaista on shintolaisia ja 80% japanilaisista on buddhalaisia. Äkkinäinen saattaisi päätellä tästä, että 160% japanilaisista on shintolaisia tai buddhalaisia .... Tämä ei ole tietystikään mahdollista!

Mikä on ratkaisu tähän näennäiseen ristiriitaan? Ratkaisuna on tietysti se, että huomattava osa japanilaisista noudattaa kummankin uskonnon menoja: Japanissa häät pidetään tavallisesti shintolaisia menoja noudattaen, kun taas hautajaiset pidetään tavallisesti buddhalaisia menoja noudattaen.

Tämä merkitsee sitä, että olla shintolainen ja olla buddhalainen eivät ole Japanissa toisensa poissulkevia tapahtumia. Siksi ei ole oikein laskea shintolaisina tai buddhalaisina itseään pitävien suhteellisia osuuksia yhteen, kun

halutaan määrätä näiden kahden uskonnon tunnustajien yhteinen suhteellinen osuus.

Annettujen tietojen perusteella voimme päätellä seuraavaa: *Joko* buddhalaisina *tai* shintolaisia pitää itseään 80% — 100%:n japanilaisista. *Sekä* buddhalaisina *että* shintolaisina pitää itseään 60% — 80% japanilaisista. ●

ESIMERKKI 8.

Eräässä lukemistottumuksia koskeneessa tutkimuksessa saatiin selville, että erään kunnan kotitalouksissa luettiin Seuraa ja Apua seuravasti:

Seura	20%,
Apu	16%,
Seura ja Apu	1%.

Miten määrätään Seuraa *tai* Apua lukevien kotitalouksien osuus kaikista kotitalouksista ko. kunnassa?

On syytä huomata, että Seuraa *tai* Apua lukevien osuutta ei saa määrätä laskemalla Seuraa lukevien ja Apua lukevien osuudet yhteen. Tämä johtuu siitä, että ne Seuraa lukevat kotitaloudet, joissa luetaan Apua ovat samoja kotitalouksia, kuin ne Apua lukevat kotitaloudet, joissa luetaan Seuraa. Siten sekä Seuraa että Apua lukevat 1% kotitalouksista ovat mukana sekä Seuraa lukevien osuudessa että Apua lukevien osuudessa ja suorassa yhteenlaskussa ne tulisivat mukaan kaksi kertaa.

Oikea tapa määrätä Seuraa *tai* Apua lukevien kotitalouksien osuus on seuraava: Lasketaan yhteen Seuraa ja Apua lukevien kotitalouksien osuudet ja vähennetään summasta sekä Seuraa että Apua lukevien kotitalouksien osuus, joka muuten tulisi mukaan kaksi kertaa: Siten Seuraa *tai* Apua lukee

$$20\% + 16\% - 1\% = 35\%$$

kotitalouksista.

Tästä saamme helposti mallin vastaavan todennköisyyden määräämiseksi:

Määritellään tapahtumat

A = "Satunnaisesti valitussa kotitaloudessa luetaan Seuraa"

B = "Satunnaisesti valitussa kotitaloudessa luetaan Apua",

A *ja* B = "Satunnaisesti valitussa kotitaloudessa luetaan Seuraa *ja* Apua",

A *tai* B = "Satunnaisesti valitussa kotitaloudessa luetaan Seuraa *tai* Apua".

Tehtävänä on määrätä tapahtuman A *tai* B todennäköisyys.

Tapahtumien A, B sekä A *ja* B todennäköisyydet ovat

$P(A) = 20/100$,

$P(B) = 16/100$,

$P(A \text{ ja } B) = 1/100.$

Tapahtuman *A tai B* todennäköisyys on yllä olevan mukaan

$$\begin{aligned} P(A \text{ tai } B) &= P(A) + P(B) - P(A \text{ ja } B) \\ &= 20/100 + 16/100 - 1/100 \\ &= 35/100. \bullet \end{aligned}$$

Edellisessä esimerkissä sovellettua kaavaa kutsutaan todennäköisyyslaskennan *yleiseksi yhteenlaskusäännöksi*.

YLEINEN YHTEENLASKUSÄÄNTÖ

$P(A \text{ tai } B) = P(A) + P(B) - P(A \text{ ja } B).$

Jos tapahtumat *A* ja *B* ovat toisensa poissulkevia,

$P(A \text{ ja } B) = 0.$

Tällöin saadaan yhteenlaskusääntö toisensa poissulkeville tapahtumille.

ESIMERKKI 9.

Tarkastellaan kahden nopan heittoa satunnaisilmiönä. Kuten edellä on todettu, ilmiön tulosvaihtoehtoina ovat silmälukujen 1,2,3,4,5,6 muodostamat parit

(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)

Olkoot

A = "Saadaan sama silmäluku kummallakin nopalla",

B = "1. nopan silmäluku on 5 tai enemmän".

Tällöin

$A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\},$

$B = \{(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$
$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\},$

$A \text{ ja } B = \{(5,5),(6,6)\},$

$A \text{ tai } B = \{(1,1),(2,2),(3,3),(4,4),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$
$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.$

Siten

$$P(A \text{ tai } B) = P(A) + P(B) - P(A \text{ ja } B)$$
$$= 6/36 + 12/36 - 2/36$$
$$= 16/36.$$

Tulos saadaan myös suoraan laskemalla tapahtumaan *A tai B* liittyvien tapahtumavaihtojen lukumäärä. ●

KOMPLEMENTTITAPAHTUMA

Oletetaan, että tapahtuma *A ei* satu tarkastellussa satunnaisilmiössä. Koska oletamme, että ilmiö kuitenkin esiintyy, sattuu *A*:n *komplementtitapahtuma* eli *ei-A*, jonka merkintä on A^C.

KOMPLEMENTTITAPAHTUMAN TODENNÄKÖISYYS

Komplementtitapahtuman todennäköisyys saadaan seuraavasta kaavasta:

KOMPLEMENTTITAPAHTUMAN TODENNÄKÖISYYS

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

Tapahtumat A ja A^C muodostavat aina otosavaruuden S jaon toisensa poissulkeviin tapahtumiin, ts. aina joko A tai *ei-A* eli A^C sattuu. Siksi

$$P(S) = P(A \text{ tai } A^C) = P(A) + P(A^C) = 1.$$

ESIMERKKI 10.

Eräs sosiologi tutki sosiaalista liikkuvuutta Englannissa. Tutkimus perustui aineistoon, jossa oli tiedot poikien ja heidän isiensä sosiaaliluokasta. Sosiaaliluokka määrättiin sellaisten tekijöiden kuten koulutus ja ammatti perusteella. Sosiaaliluokka ilmaistiin 5-portaisella asteikolla, jossa

1 = alin luokka,

5 = ylin luokka.

Sosiaaliluokka on siis järjestyasteikollinen muuttuja.

Seuraava taulukossa on esitetty todennäköisyydet, jolla poika, jonka isä kuului luokkaan 1 (= alin luokka), päätyi kuhunkin 5:stä luokasta:

Pojan luokka	1	2	3	4	5
Todennäköisyys	0.48	0.38	0.08	0.05	0.01

Olkoon

A = "Poika jää luokkaan 1",

B = "Poika saavuttaa kaksi korkeinta luokkaa".

Tällöin

$$P(A) = 0.48,$$

$$P(B) = 0.05 + 0.01 = 0.06.$$

Mikä on todennäköisyys, että poika kokee sosiaalista nousua eli ei jää luokkaan 1? Tämä tapahtuma on selvästi tapahtuman A komplementti. Siten kysytty todennäköisyys on

$$P(A^C) = 1 - P(A) = 1 - 0.48 = 0.52.$$

Tapahtumat A ja B ovat toisensa poissulkevat, joten todennäköisyys, että poika jää isänsä sosiaaliluokkaan *tai* saavuttaa kaksi korkeinta luokkaa on

$$P(A \textit{ tai } B) = P(A) + P(B) = 0.48 + 0.06 = 0.54. \bullet$$

EHDOLLINEN TODENNÄKÖISYYS

Tarkastellaan kahta tapahtumaa A ja B. Kysymme nyt, mikä on A:n todennäköisyys sillä ehdolla, että B on tapahtunut? Kutsumme tätä todennäköisyyttä A:n *ehdolliseksi todennäköisyydeksi ehdolla* B ja merkitsemme sitä seuraavasti:

$$P(A|B).$$

Kutsumme tapahtumaa B *ehtotapahtumaksi*.

Tapahtuman ehdollinen todennäköisyys voi olla suurempi, yhtäsuuri tai pienempi kuin tapahtuman ehdoton todennäköisyys. Siten kaikki seuraavat tapaukset ovat mahdollisia:

(1) $P(A|B) > P(A)$,

(2) $P(A|B) = P(A)$,

(3) $P(A|B) < P(A)$.

Siten tapahtuman ehdollinen todennäköisyys ei ole välttämättä sama kuin tapahtuman ehdoton todennäköisyys. Tämä merkitsee seuraavaa: Jos *tieto* tapahtuman B sattumisesta otetaan huomioon tapahtuman A todennäköisyyttä määrättäessä, saatetaan saada eri tulos eri, kuin silloin, kun tapahtuman B sattumista ei oteta huomioon. Tieto tapahtuman B sattumisesta voi siis vaikuttaa arvioon tapahtuman A todennäköisyydestä.

Tapauksessa (2) tieto tapahtuman B sattumisesta ei muuta arviota tapahtuman A todennäköisyydestä millään tavalla. Tämä tilanne johtaa *riippumattomuuden* käsitteeseen, jota käsitellään myöhemmin.

ESIMERKKI 11.

Nostetaan korttipakasta 2 korttia.

Todennäköisyys, että ensimmäisenä nostettu kortti on pata, on

$$\frac{13}{52} = \frac{1}{4},$$

koska 52:n kortin pakassa on 13 pataa.

Oletetaan, että ensimmäisenä nostettu kortti on todellakin pata. Tällöin todennäköisyys, että myös toisena nostettu kortti on pata, on

$$\frac{12}{51}.$$

Tämä seuraa seuraavista kahdesta seikasta:

- Toista korttia nostettaessa pakassa on jäljellä 51 korttia.
- Toista korttia nostettaessa pakassa on jäljellä 12 pataa.

Laskettu todennäköisyys on ehdollinen todennäköisyys

P(2. kortti on pata|1. kortti on pata). ●

EHDOLLINEN KESKIARVO JA EHDOLLINEN TODENNÄKÖISYYS

Palautetaan mieleen 1. kirjassa käsitelty *ehdollisen keskiarvon* käsite: Ehdolliset keskiarvot liittyvät kahden muuttujan x ja y havaintoarvojen välisen riippuvuuden kuvaamiseen. Muuttujan y ehdollinen keskiarvo määrätään sellaisista y:n arvoista, joissa muuttujan x arvot on rajoitettu johonkin x:n arvojen luokkaan. Tällöin ehtona on se, että muuttujan x arvot kuuluvat ko. luokkaan. Mielenkiinnon kohteena on erityisesti se miten muuttujan y ehdolliset keskiarvot riippuvat muuttujan x luokista. Jos muuttujan x luokka vaikuttaa muuttujan y ehdollisiin keskiarvoihin, muuttujien välillä on tilastollista riippuvuutta. Ehdolliset keskiarvot johtavat kätevästi *regression* käsitteeseen.

Ehdollisessa keskiarvossa ja ehdollisessa todennäköisyydessä on hyvin samankaltainen idea. Ehdollinen keskiarvo mahdollistaa muuttujan x muuttujasta y sisältämän informaation hyväksikäytön ennustettaessa muuttujan y arvoja. Ehdollinen todennäköisyys mahdollistaa tapahtuman B tapahtumasta A sisältämän informaation hyväksikäytön tapahtuman A todennäköisyyden arvioimisessa. Tämä samankaltaisuus ei ole sattumaa; tällä kurssilla sivuutamme kuitenkin yhteyden lähemmän tarkastelun.

TULOSÄÄNTÖ

Yhdistetyn tapahtuman

A ja B tapahtuu

todennäköisyys voidaan lausua ehdollisten todennäköisyyksien avulla seuraavilla tavoilla:

TULOSÄÄNTÖ

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

ESIMERKKI 12.

Palataan esimerkkiin 11.

Nostetaan korttipakasta 2 korttia. Todennäköisyys, että molemmat kortit ovat patoja, voidaan laskea tulosääntöä käyttäen seuraavalla tavalla:

Olkoot

A = "1. kortti on pata",

B = "2. kortti on pata".

Kysytty todennäköisyys on tapahtuman *A ja B* todennäköisyys.

Tulosäännöstä ja esimerkistä 11 seuraa, että

$$P(A\ ja\ B) = P(B|A)P(A) = \frac{12}{51}\cdot\frac{13}{52} = \frac{3}{51} \approx 0.0588. \bullet$$

TULOSÄÄNTÖ JA EHDOLLINEN TODENNÄKÖISYYS

Tulosääntö mahdollistaa tapahtuman A ehdollisen todennäköisyyden määräämisen seuraavalla kaavalla, joka on hyödyllinen silloin, kun yhdistetyn tapahtuman *A ja B* todennäköisyys on helppo määrätä:

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

Tämä kaava voidaan perustella todennäköisyyden frekvenssitulkintaa ja empiirisen todennäköisyyden määritelmää käyttäen. Tarkastellaan kahden tapahtuman A ja B sekä niiden komplementtitapahtumien A^C ja B^C sattumista pitkässä jonossa *riippumattomia* saman satunnaisilmiön toistoja[1]. Huomaa, että satunnaisilmiö on sellainen, että joko A tai *ei-A* sattuu ja joko B tai *ei-B* sattuu.

Saatamme tällöin nähdä esimerkiksi seuraavan tapahtumaparien jonon:

Toisto	1	2	3	4	5	6	7	8
	A^CB^C	A^CB	A^CB^C	AB^C	AB	AB^C	A^CB^C	A^CB

Tapahtuman B todennäköisyys on todennäköisyyden frekvenssitulkinnan mukaan näiden 8 tapahtumaparin jonossa $P(B) = 3/8$ ja yhdistetyn tapahtuman *A ja B* todennäköisyys on $P(A\ ja\ B) = 1/8$. Muodostetaan karsittu eli *ehdollinen tapahtumajono* niistä tapahtumapareista, joissa B on sattunut:

Toisto	1	2	3	4	5	6	7	8
		A^C			A			A^C

Niiden tapahtumaparien lukumäärä, joissa B on sattunut on 3. Lasketaan niistä tapahtumapareista, joissa B on sattunut ne, joissa myös A on sattunut. Niitä on 1 kappale. Siten tapahtuman A todennäköisyys on karsitussa eli *ehdollisessa tapahtumajonossa* 1/3.

---

[1] Riippumattomuuden määritelmä annetaan alla. Riippumattomuus voidaan ymmärtää siten, että se mitä on tapahtunut aikaisemmilla ilmiön toistokerroilla ei vaikuta tapahtumiin myöhemmillä toistokerroilla millään tavalla. Tapahtumien A ja B todennäköisyydet pysyvät siis samoina toistokerrasta toiseen.

Edellä esitetyn kaavan mukaan tapahtuman A ehdollinen todennäköisyys ehdolla B on

$$P(A|B) = \frac{P(A \text{ ja } B)}{P(B)} = \frac{1/8}{3/8} = \frac{1}{3},$$

mikä on sama tulos, kuin saatiin tarkastelemalla tapahtuman A sattumista ehdollisessa tapahtumajonossa.

Formaalisti tämä voidaan ilmaista seuraavasti: Olkoon

n = toistojen lukumäärä,

n(A ja B) = niiden toistojen lukumäärä, joissa sekä A että B sattuu,

n(B) = niiden toistojen lukumäärä, joissa B sattuu.

Tarkastellaan tapahtuman A suhteellista frekvenssiä siinä karsitussa eli ehdollisessa koesarjassa, jonka määrää se, että tapahtuma B on sattunut. Tämä suhteellinen frekvenssi on

$$\frac{n(A \text{ ja } B)}{n(B)}.$$

Kun toistoja jatketaan, lähestyy n(A ja$B$)/n empiirisen todennäköisyyden määritelmän perusteella todennäköisyyttä P(A ja B) ja n(B)/n todennäköisyyttä P(B). Koska karsitussa eli ehdollisessa koesarjassa tapahtuman A suhteellisen frekvenssin voidaan olettaa lähestyvän empiirisen todennäköisyyden määritelmän mukaan tapahtuman A ehdollista todennäköisyyttä, voidaan kirjoittaa

$$\frac{n(A \text{ ja } B)}{n(B)} = \frac{n(A \text{ ja } B)/n}{n(B)/n} \approx \frac{P(A \text{ ja } B)}{P(B)} = P(A|B).$$

ESIMERKKI 13.

Palataan esimerkin 8 tilanteeseen.

Mikä on todennäköisyys, että kotitaloudessa, jossa luetaan Seuraa, luetaan myös Apua? Kysytty todennäköisyys on tapahtuman B ehdollinen todennäköisyys, kun ehtona on tapahtuma A.

Tapahtumien A, B sekä A ja B todennäköisyydet ovat

P(A) = 20/100,

P(B) = 16/100,

P(A ja B) = 1/100.

Edellä esitetyn ehdollisen todennäköisyyden kaavan mukaan

$$P(B|A) = \frac{P(A \text{ ja } B)}{P(A)} = \frac{1/100}{20/100} = \frac{1}{20} \neq P(B).$$

Tässä tapauksessa tapahtumat B ja A eivät siis ole riippumattomia: Tiedosta, että kotitaloudessa luetaan Seuraa on hyötyä arvioitaessa todennäköisyyttä, että kotitaloudessa luetaan Apua. ●

RIIPPUMATTOMUUS

Tarkastellaan ennen riippumattomuuden määrittelyä seuraavaa esimerkkiä:

ESIMERKKI 14.

Oletetaan, että Rutenian tasavallan parlamentissa on 200 edustajanpaikkaa. Ruteniassa on vain kaksi puoluetta Repijät ja Säilyttäjät. Parlamentin paikkajakauma on seuraava:

Puolue	Edustajan-paikat	Miehiä	Naisia
Repijät	50	20	30
Säilyttäjät	150	60	90
Yhteensä	200	80	120

Olkoot

A = "Satunnaisesti valittu edustaja on mies",

B = "Satunnaisesti valittu edustaja kuuluu Repijöihin".

Tällöin

$$P(A|B) = 20/50 = 0.4.$$

Huomaa, että myös

$$P(A) = 80/200 = 0.4.$$

Siten ehtotapahtuma B:n ei sisällä sellaista tietoa, joka muuttaisi A:n todennäköisyyden toiseksi, kun B:n tapahtuminen otetaan huomioon. Tämä johtuu siitä, että miesedustajien suhteellinen osuus Repijöiden joukossa on sama kuin koko parlamentissa. Miesedustajien suhteellinen osuus on myös Säilyttäjien joukossa sama kuin koko parlamentissa. Huomaa, että vastaava pätee myös naisedustajien suhteelliselle osuudelle kummassakin puolueessa erikseen ja koko parlamentissa. ●

Edellinen esimerkki johtaa seuraavaan määritelmään:

RIIPPUMATTOMUUS

Tapahtumat A ja B ovat *riippumattomia*, jos

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

Määritelmän mukaan se, että Rutenian parlamentin edustaja on nainen ja se, että hän kuuluu Säilyttäjiin ovat riippumattomia tapahtumia. Samoin se, että edustaja on mies ja se, että hän kuuluu Säilyttäjiin ovat riippumattomia tapahtumia. Itse asiassa Rutenian parlamentissa edustajan sukupuoli ja se mihin puolueeseen hän kuuluu ovat riippumattomia tapahtumia.

Voidaan osoittaa, että *riippumattomuus on symmetrinen ominaisuus*: Jos siis A on riippumaton B:stä, niin myös B on riippumaton A:sta.

TULOSÄÄNTÖ RIIPPUMATTOMILLE TAPAHTUMILLE

Tulosääntö riippumattomille tapahtumille A ja B kuuluu edellä esitetyn perusteella seuraavasti: Jos tapahtumat A ja B ovat riippumattomia,

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

Voidaan osoittaa, että myös käänteinen pätee: Jos

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

niin tapahtumat A ja B ovat riippumattomia. Siten tätä kaavaa voidaan pitää riippumattomuuden määritelmänä:

RIIPPUMATTOMUUS

Tapahtumat A ja B ovat *riippumattomia*, jos ja vain, jos

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

TOISTOKOKEET JA RIIPPUMATTOMUUS

Edellä esitetty määritelmä kahdelle riippumattomalle tapahtumalle voidaan yleistää useammalle tapahtumalle $A_1, A_2, \ldots, A_n$ seuraavalla tavalla:

Tapahtumat $A_1, A_2, \ldots, A_n$ ovat *riippumattomia*, jos ja vain jos

$$P(A_{i_1} \text{ ja} \ldots \text{ja } A_{i_k}) = P(A_{i_1}) \times \ldots \times P(A_{i_k})$$

kaikille indeksijoukon $\{1,2,\ldots,n\}$ osajoukoille $\{i_1,\ldots,i_k\}$, joissa $k = 2,\ldots,n$. Erityisesti

$$P(A_1 \text{ ja } A_2 \text{ ja} \ldots \text{ja } A_n) = P(A_1) \times P(A_2) \times \ldots \times P(A_n).$$

Jos $n = 2$, saadaan edellä esitetty kaava kahden tapahtuman riippumattomuudelle.

Tämä määritelmä mahdollistaa *riippumattomien toistokokeiden* määritelmän antamisen. Riippumattomat toistokokeet ovat keskeisessä asemassa tilastollisessa tutkimuksessa. Huomaa, että tässä sanalla koe on laajempi sovellusalue, kuin sanalla koe on koesuunnittelussa. Sana toistokoe voi viitata mihin tahansa seuraavassa esitettävät ehdot täyttävään satunnaisilmiöön.

Toistokokeella tarkoitetaan sitä, että samaa satunnaisilmiötä *toistetaan* tai se *toistuu* useita kertoja samoissa olosuhteissa. Jos toistokokeita tehdään n kappaletta, puhutaan *n-kertaisesta toistokokeesta*. Jos ilmiön olosuhteet eivät muutu toistokokeiden välillä tai niiden vaikutuksesta, satunnaisilmiöön liittyvien tapahtumien todennäköisyydet pysyvät muuttumattomina toistokokeesta toiseen. Tämä merkitsee sitä, että toistokokeet ovat riippumattomia. Riippumattomuus voidaan ilmaista seuraavalla tavalla: Oletetaan, että A_i on *mikä tahansa* i. toistonkokeen tapahtuma n-kertaisessa toistokokeessa. Toistokokeet ovat riippumattomia, jos tapahtumat $A_1, A_2, \ldots, A_n$ ovat riippumattomia. Tällöin tapahtumat aikaisemmissa toistokokeissa eivät vaikuta myöhempien tapahtumien todennäköisyyteen.

Tulemme myöhemmin näkemään, että *otoksen poiminta silloin, kun se tehdään palauttamalla jo poimittu otosyksikkö takaisin perusjoukkoon on esimerkki riippumattomasta toistokokeesta*. Sen sijaan, *jos otos poimitaan palauttamatta jo poimittua otosyksikköä takaisin perusjoukkoon, toistokokeet eivät ole riippumattomia.*

Seuraava esimerkki havainnollistaa millä tavalla otoksen poiminta palauttaen eroaa poiminnasta palauttamatta.

ESIMERKKI 15.

Oletetaan, että uurnassa on 3 arpalippua, jotka on numeroitu seuraavasti:

1 2 3

Olkoon tarkasteltava satunnaisilmiö arpalipun satunnainen nostaminen uurnasta. Oletetaan, että uurnasta on nostettu arpalippu, joka osoittautuu numeroksi 3.

Ennen toisen arpalipun nostamista, käytettävissä on kaksi erilaista toimintatapaa:

1. Palautetaan jo nostettu arpalippu uurnaan. Toista arpalippua nostettaessa esimerkiksi numeron 1 todennäköisyys on tällöin 1/3, koska kaikki kolme arpalippua ovat uurnassa. Siten ehdollinen todennäköisyys

   P(Nostetaan 1|Nostetaan 3) = 1/3 = P(Nostetaan 1).

2. Jätetään jo nostettu arpalippu palauttamatta uurnaan. Toista arpalippua nostettaessa numeron 1 todennäköisyys on tällöin 1/2, koska uurnassa on nyt jäljellä vain kaksi lippua. Siten ehdollinen todennäköisyys

   P(Nostetaan 1|Nostetaan 3) = 1/2 ≠ P(Nostetaan 1).

Tapauksessa 1 nostot ovat siis riippumattomia tapahtumia, tapauksessa 2 nostot eivät ole riippumattomia. ●

On syytä huomata, että peräkkäisten satunnaislukujen poiminta satunnaislukujen taulukosta muodostaa jonon riippumattomia toistokekeita olettaen, että satunnaisluvut on määrätty oikein. Tällöin numeroiden tunteminen taulukon jossakin osassa ei auta ennustamaan numeroita taulukon muissa osissa.

Lopuksi on syytä huomata, että monissa tilastotieteen osa-alueissa *riippuvuudet* ovat analyysin kohteena: Regressioanalyysissa mallitetaan muuttujan riippuvuutta yhdestä tai useammasta muuttujasta. Aikasarjojen analyysissa taas mallitetaan muuttujan aikariippuvuutta muuttujan omasta historiasta ja/tai muista muuttujista ja niiden historiasta.

TOISTOKOKEET JA BINOMITODENNÄKÖISYYDET

Tarkastellaan jonkin satunnaisilmiön tapahtumaa A. Oletetaan, että tapahtuman A todennäköisyys

$$P(A) = p$$

jolloin tapahtuman A komplementtitapahtuman A^C (*ei-A*) todennäköisyys

$$P(A^C) = 1 - P(A) = 1 - p = q.$$

Toistetaan ko. satunnaisilmiötä (tai annetaan ko. satunnaisilmiön toistua) samoissa olosuhteissa n kertaa ja tarkastellaan tapahtuman A sattumista tässä toistokoesarjassa. Otetaan tehtäväksi määrätä todennäköisyys sille, että A tapahtuu k kertaa.

Tyypillinen tulosjono n:n toistokokeen sarjassa, jossa A tapahtuu k kertaa, on seuraava ($n = 8$, $k = 5$):

Toistokoe	1	2	3	4	5	6	7	8
Tapahtuma	A	A	A^C	A	A^C	A^C	A	A
Todennäköisyys	p	p	q	p	q	q	p	p

Koska toistokokeet oletettiin riippumattomiksi, voidaan soveltaa riippumattomien tapahtumien tulosääntöä. Mainitun tulosjonon todennäköisyydeksi saadaan kertomalla tapahtumien todennäköisyydet toistokokeiden sarjassa keskenään

$$ppqpqqpp,$$

joka voidaan kirjoittaa lyhyemmin muotoon

$$p^5q^3.$$

Yleisesti sellaisen tulosjonon, jossa A tapahtuu k kertaa ja *ei-A* tapahtuu $(n-k)$ kertaa, todennäköisyys on

$$p^kq^{n-k}.$$

Sellaisten tulosjonojen lukumäärä, joissa tapahtuma A sattuu täsmälleen k kertaa, on

$$C(n,k)=\binom{n}{k}.$$

Tämä seuraa siitä binomikertoimen $C(n,k)$ tulkinnasta, joka mainitaan multinomikerrointa käsittelevässä kohdassa (kts. kombinatoriikkaa käsittelevää kappaletta): Jos n:n alkion joukossa on k kappaletta alkioita a ja $(n-k)$ kappaletta alkioita b, niin alkiot a ja b voidaan järjestää $C(n,k)$ eri tavalla jonoon.

Olkoon

$$A_k^n$$

se yhdistetty tapahtuma, jossa A sattuu k kertaa n:n toistokokeen sarjassa. Koska erilaiset tulosjonot ovat tapahtumina toisensa poissulkevia, saadaan tapahtuman A_k^n todennäköisyys p_k^n laskemalla yhteen sellaisten yksittäisten tulosjonojen todennäköisyydet, joissa A on sattunut k kertaa. Edellä esitetyn mukaan jokaisen tällaisen tulosjonon todennäköisyys on p^kq^{n-k} ja erilaisten jonojen lukumäärä on $C(n,k)$. Siten

$$P(A_k^n)=p_k^n=\binom{n}{k}p^kq^{n-k}.$$

Todennäköisyyttä p_k^n kutsutaan *binomitodennäköisyydeksi*, koska binomilauseen mukaan

$$\begin{aligned}1&=(p+q)^n\\&=\sum_{k=0}^{n}\binom{n}{k}p^kq^{n-k}\\&=\sum_{k=0}^{n}p_k^n.\end{aligned}$$

ESIMERKKI 16.

Heitetään rahaa 5 kertaa. Mikä on todennäköisyys, että saadaan täsmälleen 5 kruunaa?

Rahanheitolle satunnaisilmiönä voidaan muodostaa seuraava todennäköisyysmalli:

Otosavaruus on muotoa

$$S=\{\text{kruuna,klaava}\}.$$

Tapahtumavaihtoehdot oletetaan symmetrisiksi, jolloin niiden todennäköisyydet ovat

$$P(\text{kruuna}) = P(\text{klaava}) = 1/2.$$

Merkitään

$$A_k^5 = \text{"}k \text{ kruunaa 5:ssä rahanheitossa"}.$$

Binomitodennäköisyyden kaavan mukaan

$$P(A_k^5) = \binom{5}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{5-k}$$

$$= \binom{5}{k}\frac{1}{2^5}.$$

Todennäköisyyksistä voidaan muodostaa seuraava taulukko:

Kruunien lkm 5:ssä heitossa	Todennäköisyys
0	0.03125
1	0.15625
2	0.3125
3	0.3125
4	0.15625
5	0.03125

ESIMERKKI 17.

Määrätään todennäköisyys sille, että 10:ssä rahanheitossa saadaan vähintään 1 klaava. Koska tapahtumat "Saadaan k klaavaa 10:ssä rahaheitossa" ovat toisensa poissulkevia, olisi kysytty todennäköisyys mahdollista määrätä yhteenlaskusäännön avulla. Tämä olisi kuitenkin työlästä. On paljon helpompaa käyttää hyväksi tapahtuman

"Vähintään 1 klaava 10:ssä rahanheitossa"

komplementtitapahtuman

"Ei yhtään klaavaa 10:ssä rahanheitossa"

= "Täsmälleen 10 kruunaa 10:ssä rahanheitossa"

todennäköisyyttä.

Binomitodennäköisyyden kaavasta saadaan

$$P(\text{10 kruunaa}) = \binom{10}{10}\frac{1}{2^{10}} = \frac{1}{2^{10}} \approx 0.000977,$$

koska C(10,10) = 1.

Siten

$$P(1,2,\ldots \textit{ tai } 10 \text{ klaavaa}) = 1 - P(10 \text{ kruunaa}) \approx 0.99902. \bullet$$

1.1.9 TODENNÄKÖISYYS MATEMAATTISENA KÄSITTEENÄ

Edellä on tarkasteltu todennäköisyyttä sekä suhteellisena frekvenssinä että klassisena todennäköisyytenä. Nämä tarkastelukehikot eivät kuitenkaan täytä matemaattisen täsmällisyyden vaatimuksia. Tämän vuosisadan suuria edistysaskelia matematiikassa on ollut todennäköisyyslaskennan teorian esittäminen *aksiomaattisena* järjestelmänä. Aksiomaattisen lähestymistavan merkitys on siinä, että kaikki aksioomien matemaattiset seuraukset ovat voimassa kaikille sellaisille struktuureille, jotka toteuttavat ko. aksioomajärjestelmän. Esimerkiksi kaikki edellä esitetyt ja myös jatkossa esitettävät todennäköisyyslaskennan säännöt voidaan todistaa seuraavassa esitettävistä 3:sta aksioomasta.

Tarkastellaan ensin todennäköisyyslaskennan aksioomia *äärellisissä otosavaruuksissa*.

ÄÄRELLISET OTOSAVARUUDET

Olkoon

$$S = \{s_1, s_2, \ldots, s_n\}$$

äärellinen *perusjoukko* eli *otosavaruus*, jonka alkioita s_i, $i = 1,2,\ldots,n$ kutsutaan *alkeistapahtumiksi* eli *tulosvaihtoehdoiksi*. Kutsutaan perusjoukon S osajoukkoja *tapahtumiksi*. Olkoon

$$\mathcal{F} = \{A \mid A \subset S\}$$

kaikkien perusjoukon S osajoukkojen muodostama joukko eli kaikkien otosvaruuden S *tapahtumien* muodostama joukko. Joukkoa $\mathcal{F}$ kutsutaan usein perusjoukon S osajoukkojen muodostamaksi *joukkoperheeksi*. Joukkoperheen sen alkioina ovat perusjoukon S osajoukot. Voidaan osoittaa, että joukkoperheessä $\mathcal{F}$ on 2^n alkiota, jos perusjoukossa S on n alkiota.

Palautetaan mieleen seuraavat joukko-opin merkinnät:

Jos perusjoukon S alkio s on *kuuluu* joukkoon A eli s on joukon A alkio, niin merkitään

$$s \in A.$$

Vastaavasti, jos perusjoukon S alkio s ei *kuulu* joukkoon A eli s ei ole joukon A alkio, niin merkitään

$$s \notin A.$$

Jos joukko A on joukon B *osajoukko*, niin merkitään

$$A \subset B.$$

Joukko A on joukon B *osajoukko*, jos jokainen joukon A alkio kuuluu joukkoon B, ts. $A \subset B$, jos siitä, että $s \in A$ seuraa, että $s \in B$.

Olkoon $\varnothing$ *tyhjä joukko*, jossa ei ole yhtään alkiota. Tyhjä joukko $\varnothing$ vastaa *mahdotonta* tapahtumaa. Otosavaruus S taas vastaa *varmaa* tapahtumaa: Otosavaruus S sisältää *kaikki* tulosvaihtoehdot, jotka ovat mahdollisia.

Perusjoukossa S voidaan määritellä seuraavat *joukko-operaatiot*:

Osajoukon A *komplementti* A^C on niiden perusjoukon alkioiden joukko, jotka eivät kuulu A:han eli

$$A^C = \{s \in S | s \notin A\}.$$

A^C vastaa tapahtuman A komplementtitapahtumaa.

Osajoukkojen A ja B *yhdiste* $A \cup B$ on niiden perusjoukon alkioiden joukko, jotka kuuluvat A:han *tai* B:hen tai molempiin eli

$$A \cup B = \{s \in S | s \in A \text{ tai } s \in B\}.$$

$A \cup B$ vastaa yhdistettyä tapahtumaa A *tai* B.

Osajoukkojen A ja B *leikkaus* $A \cap B$ on niiden perusjoukon alkioiden joukko, jotka kuuluvat A:han *ja* B:hen eli

$$A \cap B = \{s \in S | s \in A \text{ ja } s \in B\}.$$

$A \cap B$ vastaa yhdistettyä tapahtumaa A *ja* B.

Jos $A \cap B = \varnothing$, niin joukot A ja B ovat *pistevieraita* eli ne ovat tapahtumina *toisensa poissulkevia*.

Olkoon P *joukkofunktio*, joka liittää jokaiseen perusjoukon S osajoukkoon A reaaliluvun $P(A)$. Joukkofunktio P on *todennäköisyys*, jos seuraavat *aksioomat* eli perusoletukset ovat voimassa:

(1) $P(S) = 1$.

(2) $0 \le P(A) \le 1$ kaikille $A \in \mathcal{F}$.

(3) Jos $A \in \mathcal{F}$, $B \in \mathcal{F}$ ja $A \cap B = \varnothing$, niin $P(A \cup B) = P(A) + P(B)$.

Esimerkiksi todennäköisyys suhteellisena frekvenssinä, klassinen todennäköisyys ja ehdollinen todennäköisyys toteuttavat aksioomat (1), (2) ja (3). Tämä todistetaan myöhemmin malliksi siinä tapauksessa, että todennäköisyys määritellään suhteellisena frekvenssinä. Klassisen todennäköisyyden tapaus voidaan käsitellä vastaavalla tavalla. Todistus on helppo myös ehdollisen todennäköisyyden tapauksessa, mutta sivuutetaan.

Kuten tässä kappaleessa on nähty, perusjoukon osajoukot voidaan rinnastaa satunnaisilmiöön liittyviin tapahtumiin. Siten joukko-opin puhetavoilla on vastineet todennäköisyyslaskennassa. Seuraavassa kohdassa on lueteltu nämä vastaavuudet.

JOUKKO-OPIN KÄSITTEET JA NIIDEN VASTINEET TODENNÄKÖISYYSLASKENNASSA

Joukko-oppi	Todennäköisyyslaskenta	Merkintä
Perusjoukko	Otosavaruus	S
Tyhjä joukko	Mahdoton tapahtuma	$\varnothing$
Alkio	Alkeistapahtuma s	$s \in S$
Osajoukko	Tapahtuma A	$A \subset S$
Komplementti	Komplementtitapahtuma *ei-A*	A^C
Yhdiste	Yhdistetty tapahtuma *A tai B*	$A \cup B$
Leikkaus	Yhdistetty tapahtuma *A ja B*	$A \cap B$

TODENNÄKÖISYYS SUHTEELLISENA FREKVENSSINÄ JA TODENNÄKÖISYYDEN AKSIOOMAT

Olkoon S johonkin satunnaisilmiöön liittyvä otosavaruus. Toistetaan satunnaisilmiötä samoissa olosuhteissa toisistaan riippumattomasti n kertaa. Muodostetaan uusi otosavaruus S^n toistokokeiden kaikista mahdollisista tulosjonoista. Olkoon

f_A = tapahtuman A frekvenssi n:n toistokokeen sarjassa.

Määritellään tapahtuman A *frekventistinen todennäköisyys* $P_f(A)$ tapahtuman A suhteellisena frekvenssinä

$$P_f(A) = f_A/n.$$

On helppo nähdä, että frekventistinen todennäköisyys P_f toteuttaa todennäköisyyden aksioomat:

Aksiooma (1):

$$P_f(S) = 1.$$

Todistus:

Koska tapahtuma S sattuu joka kerran, kun satunnaisilmiö esiintyy, niin $f_S = n$. Siten

$$P_f(S) = \frac{n}{n} = 1. \blacksquare$$

Aksiooma (2):

$$0 \le P_f(A) \le 1.$$

Todistus:

Koska $0 \le f_A \le n$, niin

$$0 \le \mathrm{P}_f(A) = \frac{f_A}{n} \le 1. \blacksquare$$

Aksiooma (3):

Jos $A \subset S$ ja $B \subset S$ ja $A \cap B = \varnothing$, niin $\mathrm{P}_f(A \cup B) = \mathrm{P}_f(A) + \mathrm{P}_f(B)$.

Todistus:

Olkoot A ja B toisensa poissulkevia eli $A \cap B = \varnothing$. Tällöin $f_{A \cup B} = f_A + f_B$, koska A ja B eivät voi sattua samanaikaisesti. Siten

$$\mathrm{P}_f(A \cup B) = \frac{f_{A \cup B}}{n} = \frac{f_A + f_B}{n} = \frac{f_A}{n} + \frac{f_B}{n} = \mathrm{P}_f(A) + \mathrm{P}_f(B). \blacksquare$$

Siten kaikki äärelliseen perusjoukkoon liittyvien aksioomien (1), (2) ja (3) seuraukset pätevät myös frekventistiselle todennäköisyydelle.

LAUSEIDEN TODISTAMINEN TODENNÄKÖISYYDEN AKSIOOMISTA

Kuten edellä todettiin kaikki aksioomien (1), (2) ja (3) seuraukset pätevät kaikille struktuureille, jotka toteuttavat nämä aksioomat. Seuraavassa komplementti-todennäköisyyden kaava ja yleinen yhteenlaskusääntö todistetaan esimerkkeinä käytettävästä tekniikasta.

LAUSE 1. KOMPLEMENTTITODENNÄKÖISYYS

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

Todistus:

Joukot A ja A^{C} muodostavat perusjoukon S jaon pistevieraisiin osajoukkoihin, ts.

$$A \cup A^{\mathrm{C}} = S$$

ja

$$A \cap A^{\mathrm{C}} = \varnothing.$$

Aksioomien (1) ja (3) mukaan

$$1 = \mathrm{P}(S) = \mathrm{P}(A \cup A^{\mathrm{C}}) = \mathrm{P}(A) + \mathrm{P}(A^{\mathrm{C}}),$$

josta väite seuraa. ■

LAUSE 2. YLEINEN YHTEENLASKUSÄÄNTÖ

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

Todistus:

Jaetaan joukko $A \cup B$ pistevieraisiin osajoukkoihin A ja $A^{\mathrm{C}} \cap B$: Siten

$$A \cup B = A \cup (A^{\mathrm{C}} \cap B).$$

Joukot A ja $A^C \cap B$ ovat pistevieraita, koska

$$A \cap (A^C \cap B) = \varnothing.$$

Jaetaan myös joukko B pistevieraisiin osajoukkoihin $A \cap B$ ja $A^C \cap B$: Siten

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

Joukot $A \cap B$ ja $A^C \cap B$ ovat pistevieraita, koska

$$(A \cap B) \cap (A^C \cap B) = \varnothing.$$

Aksioomasta (3) seuraa, että

$$P(A \cup B) = P(A) + P(A^C \cap B)$$

ja

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

Väite seuraa vähentämällä nämä kaksi yhtälöä puolittain toisistaan. ■

ÄÄRETTÖMÄT OTOSAVARUUDET

Jos otosavaruus on ääretön, eivät aksioomat (1), (2) ja (3) kelpaa määrittelemään todennäköisyyttä, vaan aksioomajärjestelmää joudutaan korjaamaan seuravassa esitettävällä tavalla. Tällöin saadaan ns. *Kolmogorovin aksioomat* todennäköisyydelle.

Edellisessä määritelmässä todennäköisyys voitiin määritellä kaikille äärellisen perusjoukon S osajoukoille. Jos otosavaruus on (ylinumeroituvasti) ääretön, tämä ei ole aina mahdollista. Joukkoperheeseen $\mathcal{F}$ voidaan ottaa vain ne perusjoukon S osajoukot, jotka muodostavat ns. *σ-algebran.*

Joukon S osajoukkojen perhe $\mathcal{F}$ muodostaa σ-algebran, jos

(1) $S \in \mathcal{F}$,

(2) Jos $A \in \mathcal{F}$, niin $A^C \in \mathcal{F}$.

(3) Jos $A_1, A_2, \ldots \in \mathcal{F}$, niin $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

σ-algebran aksiooma (3) tarkoittaa seuraavaa: Jos $A_1, A_2, \ldots$ on numeroitava joukko perusjoukon S osajoukkoja, niiden yhdiste kuuluu perheeseen $\mathcal{F}$.

σ-algebran $\mathcal{F}$ alkioille määritelty joukkofunktio P on *todennäköisyys*, jos se toteuttaa seuraavat aksioomat:

(1) $P(S) = 1$.

(2) $0 \le P(A) \le 1$ kaikille $A \in \mathcal{F}$.

(3)' Jos $A_1, A_2, \ldots \in \mathcal{F}$ ja $A_i \cap A_j = \varnothing$ aina, kun $i \neq j$, niin

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

Kolmogorovin aksioomat (1) ja (2) äärettömälle otosavaruudelle ovat samat kuin äärelliselle otosavaruudelle. Sen sijaan äärellisen otosvaruuden aksiooma (3) on jouduttu korvaamaan aksioomalla (3)', joka sallii todennäköisyyden määräämisen pareittain toisensa poissulkeville tapahtumille yksittäisten tapahtumien todennäköisyyksien summana myös silloin, kun tapahtumia on numeroituvasti ääretön määrä. Aksiooma (3) on aksiooman (3)' erikoistapaus.

Matematiikan kannalta Kolmogorovin aksioomien sisältönä on se, että *todennäköisyys on normeerattu täydellisesti additiivinen mitta*:

- Aksiooman (3) mukaan todennäköisyysmitta käyttäytyy kuten tavanomaiset mitat paino, pituus ja pinta-ala.
- Aksiooman (1) mukaan mitta on normeerattu.
- Aksiooman (3)' mukaan mitta on täydellisesti additiivinen.

σ-algebran määrittely on tarpeen Kolmogorovin aksioomien yhteydessä, koska tällä tavalla tulee määritellyksi mitkä otosavaruuden S osajoukot kelpaavat tapahtumiksi. Yleisessä tapauksessa σ-algebran määrittely todellakin rajoittaa tapahtumiksi kelpaavien perusjoukon osajoukkojen perhettä. Tämä merkitsee sitä, että äärettömissä otosavaruuksissa saattaa olla osajoukkoja, joita ei voi "mitata" todennäköisyysmitalla; niitä kutsutaan *epämitallisiksi* joukoiksi.

Emme tule jatkossa käyttämään Kolmogorovin aksioomia äärettömälle perusjoukolle annetussa muodossa.

1.1.10 BAYESIN KAAVA

KOKONAISTODENNÄKÖISYYS

Olkoon $B_1, B_2, \ldots, B_n$ otosavaruuden S *jako* toisensa poissulkeviin tapahtumiin, ts.

$$S = B_1 \cup B_2 \cup \ldots \cup B_n$$

ja

$$B_i \cap B_j = \varnothing, \text{ jos } i \neq j.$$

Tämä merkitsee sitä, että satunnaisilmiön esiintyessä täsmälleen yksi tapahtumista $B_1, B_2, \ldots, B_n$ sattuu.

Olkoon A jokin ko. satunnaisilmiön liittyvä tapahtuma, ts. $A \subset S$. Otosavaruuden S jako toisensa poissulkeviin tapahtumiin *indusoi* A:n *jaon* toisensa poissulkeviin tapahtumiin $A \cap B_1, A \cap B_2, \ldots, A \cap B_n$. Tämä tarkoittaa siitä, että

$$A = A \cap S$$
$$= A \cap \left(\bigcup_{i=1}^{n} B_i \right)$$
$$= \bigcup_{i=1}^{n} (A \cap B_i)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)$$

ja

$$(A \cap B_i) \cap (A \cap B_j) = \varnothing, \text{ kun } i \neq j.$$

Soveltamalla yhteenlaskusääntöä toisensa poissulkeviin tapahtumiin ja kertolaskusääntöä saadaan seuraava kaava:

KOKONAISTODENNÄKÖISYYDEN KAAVA

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n)$$
$$= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n).$$

Kaavassa tapahtuman A todennäköisyys on lausuttu todennäköisyyksien $P(B_i)$ ja ehdollisten todennäköisyyksien $P(A|B_i)$ avulla. Kaavaa voidaan soveltaa tapahtuman A todennäköisyyden määräämiseen sellaisissa tilanteissa, joissa nämä ehdolliset todennäköisyydet ovat tunnettuja.

ESIMERKKI 1.

Eräs rautakaupan tukkuliike ostaa ruuveja 3:lta tehtaalta A, B ja C. Ruuvit toimitetaan tehtailta samanlaisissa 100:n ruuvin pakkauksissa. Osa ruuveista on viallisia. Tehtaitten toimitusosuudet ja viallisten ruuvien keskimääräiset lukumäärät eri tehtaiden pakkauksissa on ilmoitettu seuraavassa taulukossa:

Tehdas	Osuus toimituksista %	Viallisten ruuvien lkm per pakkaus
A	60	5
B	30	7
C	10	10

Mikä on todennäköisyys, että satunnaisesti tukkuliikkeestä ostettu ruuvi on viallinen?

Olkoon tapahtuma

D = "Ruuvi on viallinen",

A = "Ruuvi on tehtaalta A",

B = "Ruuvi on tehtaalta B",

C = "Ruuvi on tehtaalta C".

Tehtävänä on laskea tapahtuman D todennäköisyys. kun käytettävissä on seuraavat ym. taulukosta saatavat tiedot:

$$P(A) = 0.6,\ P(B) = 0.4,\ P(C) = 0.1,$$

$$P(D|A) = 0.05,\ P(D|B) = 0.07,\ P(D|C) = 0.1.$$

Tapahtuman D todennäköisyys saadaan kokonaistodennäköisyyden kaavasta:

$$\begin{aligned} P(D) &= P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) \\ &= 0.6 \cdot 0.05 + 0.3 \cdot 0.07 + 0.1 \cdot 0.1 \\ &= 0.061. \bullet \end{aligned}$$

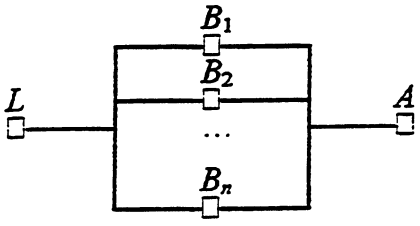
Kokonaistodennäköisyyden kaavalle voidaan esittää seuraava *systeemiteoreettinen* tulkinta: Olkoot tapahtumat A ja B_i, $i = 1,2,\ldots,n$ jonkin *systeemin tiloja*. Oletetaan, että lopputilaan A voidaan päästä lähtötilasta L vain tilojen B_i kautta. Kokonaistodennäköisyyden kaavassa esiintyvillä todennäköisyyksille voidaan antaa seuraava tulkinta:

$P(B_i)$ = P(Päästään lähtötilasta L tilaan B_i),

$P(A|B_i)$ = P(Päästään lopputilaan A tilasta B_i),

$P(A)$ = P(Päästään lähtötilasta L lopputilaan A).

Kokonaistodennäköisyyden tulkinnassa käytettyä systeemiä voidaan havainnollistaa seuraavalla kaaviolla:



BAYESIN KAAVA

Tarkastellaan sama tilannetta, jota käytettiin kokonaistodennäköisyyden kaavaa johdettaessa: Olkoon $B_1, B_2,\ldots, B_n$ otosavaruuden S jako toisensa poissulkeviin tapahtumiin, ts.

$$S = B_1 \cup B_2 \cup \ldots \cup B_n$$

ja

$$B_i \cap B_j = \varnothing, \text{ jos } i \neq j.$$

Olkoon lisäksi A jokin ko. satunnaisilmiöön liittyvä tapahtuma, ts. $A \subset S$.

Ehdollinen todennäköisyys $P(B_i|A)$ voidaan lausua tulosäännön mukaan muodossa

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

Soveltamalla tulosääntöä uudelleen voidaan yhtälön oikean puolen *osoittaja* kirjoittaa muotoon

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

Kokonaistodennäköisyyden kaavan mukaan yhtälön oikean puolen *nimittäjä* voidaan kirjoittaa muotoon

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n).$$

Yhdistämällä nämä tulokset saadaan seuraava kaava:

BAYESIN KAAVA

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

Kaava on hyödyllinen sellaisissa tilanteissa, joissa halutaan laskea ehdollinen todennäköisyys $P(B_i|A)$, kun *prioritodennäköisyydet*[1] $P(B_i)$ ja ehdolliset todennäköisyydet $P(A|B_i)$ tunnetaan. Todennäköisyyttä $P(B_i|A)$ on tapana kutsua *posteriori-todennäköisyydeksi*[2].

Priori- ja posterioritodennäköisyyden nimitykset selittyvät parhaiten soveltamalla kokonaistodennäköisyyden kaavan yhteydessä esitettyä systeemiteoreettista tulkintaa. Bayesin kaava ilmaisee todennäköisyyden sille, että tilaan A on tultu tilan B_i kautta. Posterioritodennäköisyyttä $P(B_i|A)$ kutsutaan usein tapahtuman A *käänteistodennäköisyydeksi*. Bayesin kaava muuntaa tapahtumien B_i prioritodennäköisyydet tapahtumien B_i posterioritodennäköisyyksiksi: Tiedämme ennen tapahtuman A sattumista, että tapahtuman B_i prioritodennäköisyys on $P(B_i)$. Sen jälkeen kun A on sattunut, voimme määrätä tapahtuman B_i posterioritodennäköisyyden $P(B_i|A)$.

ESIMERKKI 2.

Sairaalassa tutkitaan keuhkosyövästä epäiltyjä potilaita. Potilaille tehdyn kyselyn mukaan heistä 45% tupakoi säännöllisesti. Tutkimuksissa 90%:lla tupakoitsevista potilaista todetaan keuhkosyöpä, kun vain 5%:lla tupakoitsemattomista potilaista todetaan keuhkosyöpä. Mikä on todennäköisyys, että satunnaisesti valittu syöpäpotilas X on säännöllinen tupakoitsija?

Olkoot

B_1 = "X on tupakoitsija",

---

[1] prior (*lat.*), edeltävä, aikaisempi.

[2] posterior (*lat.*), jälkeen tuleva, myöhempi.

B_2 = "X ei ole tupakoitsija",

A = "X:llä on keuhkosyöpä".

Annettujen tietojen mukaan

$$\mathrm{P}(B_1) = 0.45,$$

$$\mathrm{P}(B_2) = 0.55,$$

$$\mathrm{P}(A|B_1) = 0.9,$$

$$\mathrm{P}(A|B_2) = 0.05.$$

Bayesin kaavan mukaan

$$\begin{aligned}\mathrm{P}(B_1|A) &= \frac{\mathrm{P}(B_1)\mathrm{P}(A|B_1)}{\mathrm{P}(B_1)\mathrm{P}(A|B_1)+\mathrm{P}(B_2)\mathrm{P}(A|B_2)} \\ &= \frac{0.45 \cdot 0.9}{0.45 \cdot 0.9 + 0.55 \cdot 0.05} \\ &\approx 0.936.\end{aligned}$$

1.1.11 TODENNÄKÖISYYDET JA VERKOT

Monimutkaisia satunnaisilmiöitä on usein hyödyllistä kuvata verkkodiagrammien avulla. Seuraavassa tarkastellaan *puudiagrammeja* ja *toimintaverkkoja*. Kummallakin verkkotyypillä on monia käytännöllisiä sovelluksia.

PUUDIAGRAMMIT JA TODENNÄKÖISYYDET

Puudiagrammi on *puumainen verkko*, jolla voidaan kuvata monivaiheisia satunnaisilmiöitä.

Jotta satunnaisilmiötä voitaisiin kuvata puudiagrammilla, se on voitava jakaa toisiaan seuraaviin vaiheisiin siten, että jokainen vaihe koostuu toisensa poissulkevista tapahtumavaihtoehdoista. Jokaisessa vaiheessa toteutuu jokin vaihtoehdoista. Tämä johtaa uusiin vaihtoehtoihin, joista vuorostaan jokin toteutuu. Tätä jatkuu kunnes saavutetaan satunnaisilmiön lopullinen tulos. Lopullisia tapahtumavaihtoehtoja on tällaisessa asetelmassa aina useita kappaleita. Se mikä lopullisista vaihtoehdoista toteutuu, riippuu siitä, mitkä vaihtoehdoista ovat toteutuneet ilmiön edeltävissä vaiheissa.

Tehtävänä on laskea lopullisten tapahtumavaihtoehtojen todennäköisyydet, kun otetaan huomioon kaikki edeltävät vaiheet ja niissä toteutuneiden vaihtoehtojen todennäköisyydet. Jos kaikkien tapahtumavaihtoehtojen todennäköisyydet tunnetaan, voidaan lopullisten tapahtumavaihtoehtojen todennäköisyys määrätä soveltamalla todennäköisyyslaskennan sääntöjä.

Satunnaisilmiöön liittyvien tapahtumien vaiheittaisuutta voidaan kuvata puudiagrammilla, jossa tapahtumavaihtoehtoja vastaa puun haarautuminen oksiksi, jotka saattavat haarautua uudelleen, jne. Kuhunkin haarautumiseen liitetään vastaavan vaihtoehdon toteutumista kuvaava todennäköisyys.

Valaistaan tällaisia tilanteita seuraavien esimerkkien avulla:

ESIMERKKI 1.

Kirjassa 1 käsitellään seuraavaa esimerkkiä.

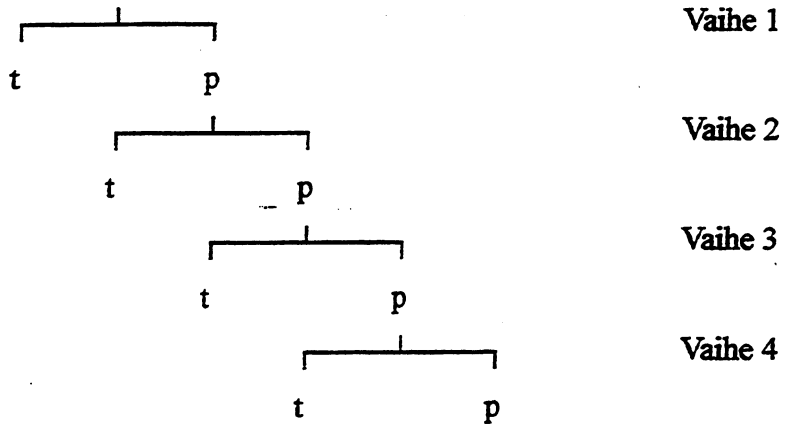
Eräs pariskunta haluaa saada lapsekseen tytön. Siksi pariskunta suunnittelee lasten hankkimista seuraavalla strategialla: Lapsia päätetään hankkia, kunnes saadaan tyttö, mutta lapsia ei kuitenkaan missään tapauksessa hankita enemmän kuin 4. Mikä on strategian onnistumisen eli tytön saamisen todennäköisyys?

Kirjassa 1 esimerkillä havainnollistettiin *simuloinnin* käyttöä todennäköisyyksien laskemisessa.

Ratkaistaan ongelma rakentamalla puumainen verkko. Puu koostuu yhä uusista haarautumisista kahdeksi oksaksi: Liitetään vasemmanpuoleinen oksa tytön saamiseen (t) ja oikeanpuoleinen oksa pojan saamiseen (p). Oletetaan, että kummankin vaihtoehdon todennäköisyys on aina 1/2. Aina, kun saadaan tyttö joudutaan puussa siis vasemmalle oksalle ja samalla oksa katkeaa. Oikeakin

oksa katkeaa vaiheessa 4, koska silloin on saatu 4 poikaa ja strategia on epäonnistunut.

Puu näyttää siis seuraavalta:



Kaikki vasemmanpuoleiset kirjaimeen t päättyvät oksat liittyvät onnistuneeseen strategiaan. Jokaisessa haarautumisessa todennäköisyys joutua kummalle tahansa kahdesta haarasta on 1/2 riippumatta aikaisemmin syntyneiden lasten sukupuolesta. Tämä merkitsee sitä, että lapsen sukupuolen oletetaan määräytyvän riippumattomasti kaikissa edellisissä vaiheissa syntyneiden lasten sukupuolista.

Miten määrätään todennäköisyys, että käytetyllä strategialla saadaan tyttö?

Olkoon

B = "Pariskunta saa käytetyllä strategialla tytön",

A = "Pariskunta saa tytön",

A^C = "Pariskunta saa pojan".

Oletusten mukaan

$$P(A) = P(A^C) = 1/2.$$

Olkoon

A_i = "Pariskunta saa vaiheessa i tytön",

A_i^C = "Pariskunta saa vaiheessa i pojan".

Pariskunnan strategia onnistuu eli B tapahtuu, jos pariskunta saa tytön vaiheessa 1 *tai* vaiheessa 2 *tai* vaiheessa 3 *tai* vaiheessa 4 eli, jos tapahtuu A_1 *tai* A_2 *tai* A_3 *tai* A_4. Muussa tapauksessa strategia epäonnistuu. Tapahtumat A_i ovat toisensa poissulkevia, joten kysytty todennäköisyys on

$$P(B) = P(A_1) + P(A_2) + P(A_3) + P(A_4).$$

Tarkastellaan tapahtumien A_1, A_2, A_3 ja A_4 todennäköisyyksiä erikseen.

Selvästi

$$P(A_1) = P(A) = P(A^C) = 1/2.$$

Tapahtuma A_2 voidaan lausua yhdistettynä tapahtumana muodossa

$A_2 = A \cap A_1^C$.

Siten tapahtumalle A_2 voidaan antaa vaihtoehtoinen sanallinen tulkinta

A_2 = "Vaiheessa 2 syntyy tyttö, kun vaiheessa 1 on syntynyt poika".

Tapahtuman A_1^C todennäköisyys ja tapahtuman A ehdollinen todennäköisyys ehdolla A_1^C tunnetaan:

$P(A_1^C) = 1/2$,

$P(A|A_1^C) = P(A) = 1/2$.

Koska tapahtumat A ja A_1^C ovat riippumattomia, riippumattomien tapahtumien tulosäännön mukaan

$$P(A_2) = P(A|A_1^C)P(A_1^C)$$
$$= (1/2)\cdot(1/2) = 1/4.$$

Vastaavalla tavalla

$$P(A_2^C) = P(A^C|A_1^C)P(A_1^C)$$
$$= (1/2)\cdot(1/2) = 1/4.$$

Tapahtuma A_3 voidaan lausua yhdistettynä tapahtumana muodossa

$A_3 = A \cap A_2^C$.

Siten tapahtumalle A_3 voidaan antaa vaihtoehtoinen sanallinen tulkinta

A_3 = "Vaiheessa 3 syntyy tyttö, kun vaiheissa 1 ja 2 on syntynyt poika".

Tapahtuman A_2^C todennäköisyys ja tapahtuman A ehdollinen todennäköisyys ehdolla A_2^C tunnetaan:

$P(A_2^C) = 1/4$,

$P(A|A_2^C) = P(A) = 1/2$.

Koska tapahtumat A ja A_2^C ovat riippumattomia, riippumattomien tapahtumien tulosäännön mukaan

$$P(A_3) = P(A|A_2^C)P(A_2^C)$$
$$= (1/2)\cdot(1/4) = 1/8.$$

Vastaavalla tavalla

$$P(A_3^C) = P(A^C|A_2^C)P(A_2^C)$$
$$= (1/2)\cdot(1/4) = 1/8.$$

Tapahtuma A_4 voidaan lausua yhdistettynä tapahtumana muodossa

$A_4 = A \cap A_3^C$.

Siten tapahtumalle A_4 voidaan antaa vaihtoehtoinen sanallinen tulkinta

A_4 = "Vaiheessa 4 syntyy tyttö, kun vaiheissa 1,2 ja 3 on syntynyt poika".

Tapahtuman A_3^C todennäköisyys ja tapahtuman A ehdollinen todennäköisyys ehdolla A_3^C tunnetaan:

$$P(A_3^C) = 1/8,$$

$$P(A|A_3^C) = P(A) = 1/2.$$

Koska tapahtumat A ja A_3^C ovat riippumattomia, riippumattomien tapahtumien tulosäännön mukaan

$$\begin{aligned} P(A_4) &= P(A|A_3^C)P(A_3^C) \\ &= (1/2)\cdot(1/8) = 1/16. \end{aligned}$$

Siten

$$\begin{aligned} P(B) &= P(A_1) + P(A_2) + P(A_3) + P(A_4) \\ &= 1/2 + 1/4 + 1/8 + 1/16 \\ &= 15/16. \end{aligned}$$

Huomaa seuraavat kaksi seikkaa:

Kuhunkin tapahtumavaihtoehtoon A_i päästään vain yhtä puun haaraa pitkin. Tapahtumavaihtoehdon A_i todennäköisyys saadaan siihen johtaviin puun haaroihin liittyvien todennäköisyyksien tulona.

Tapahtumaan B päästään useaa haaraa pitkin. Nämä haarat vastaavat vaihtoehtoisia eli toisensa poissulkevia tapoja päästä onnistuneeseen lopputulokseen eli saadaan tyttö. Tapahtuman B todennäköisyys saadaan tapahtumavaihtoehtoihin A_i johtavien puun haaroihin liittyvien todennäköisyyksien summana. ●

Voimme yleistää edellisessä esimerkissä esitetyn tavan määrätä sellaisiin satunnaisilmiöihin liittyvien tapahtumien todennäköisyyksiä, joita voidaan kuvata puudiagrammeilla seuraaviksi säännöiksi:

- *Summasääntö puutodennäköisyyksille*:

  Jos tapahtumaan johtaa useampi puun haara, tapahtuman todennäköisyys saadaan laskemalla eri haaroja vastaavat todennäköisyydet yhteen.

- *Tulosääntö puutodennäköisyyksille*:

  Yhtä puun haaraa vastaavan tapahtuman todennäköisyys saadaan laskemalla haaraan liittyvien todennäköisyyksien tulo.

ESIMERKKI 2.

Munuaistaudissa potilaan munaiset lopettavat toimintansa ja potilas kuolee. Oletetaan, että hoitona käytetään kahta menetelmää: *dialyysia* ja *munuaisensiirtoa*.

Erään tutkimuksen mukaan hoitotulokset ovat seuraavat:

- Dialyysipotilaat:

  68% potilaista on hengissä 5:n vuoden kuluttua.

32% potilaista on kuollut 5:n vuoden kuluttua.

- Munuaisensiirtopotilaat:

  48%:lla potilaista siirretty munuainen toimii normaalisti ja he ovat hengissä 5:n vuoden kuluttua.

  43%:lla potilaista siirretty munuainen ei ryhdy toimimaan normaalisti ja heille on tämän jälkeen annettava dialyysihoitoa. Näistä potilaista 42% on hengissä ja 58% on kuollut 5:n vuoden kuluttua.

  9% potilaista ei selviä hengissä siirrosta.

Kumpi hoidoista munuaistautipotilaan kannattaa valita, jos hän haluaa maksimoida todennäköisyytensä olla hengissä 5:n vuoden kuluttua?

Hoitovaihtoehdoista ja niihin liittyvistä todennäköisyyksistä voidaan rakentaa kaksi puuta, joissa käytetään seuraavia merkintöjä hoitotapahtumille ja niiden seurauksille:

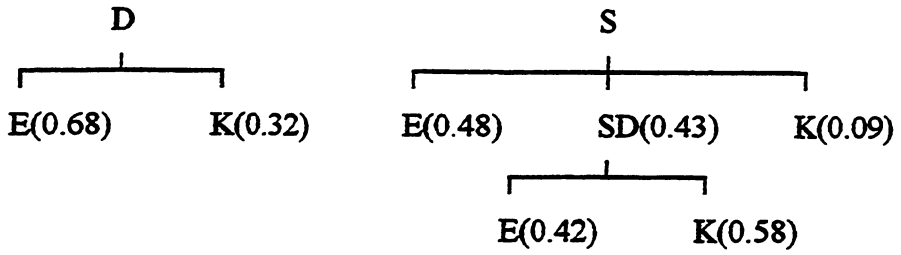
D = "Dialyysi",

S = "Munuaisensiirto",

SD = "Siirtopotilaan joutuminen dialyysiin",

E = "Elää 5:n vuoden kuluttua",

K = "On kuollut 5:n vuoden kuluttua".



Vaaditun vertailun tekemiseksi joudutaan määräämään todennäköisyys, että siirtopotilas, joka on joutunut dialyysiin on elossa 5:n vuoden kuluttua. Tämä todennäköisyys voidaan ilmaista symbolisesti muodossa

$$P(E \textit{ ja } SD|S).$$

Puutodennäköisyyksien tulosäännön mukaan tämä todennäköisyys saadaan määräämällä vastaavan oksankärkeen johtavien todennäköisyyksien tulo. Ko. todennäköisyys voidaan myös määrätä tulosääntöä käyttämällä:

$$\begin{aligned} P(E \textit{ ja } SD|S) &= P(SD|S)P(E|SD \textit{ ja } S) \\ &= 0.43 \cdot 0.42 \\ &\approx 0.18. \end{aligned}$$

5:n vuoden kuluttua elossa olevan siirtopotilaan munuainen joko toimii tai, jos se ei toimi, hän saa dialyysihoitoa. Koska nämä ovat tapahtumina toisensa

poissulkevia, kysytyksi todennäköisyydeksi saadaan puutodennäköisyyksien yhteenlaskusäännön mukaan

$$P(E|S) = 0.48 + 0.18 = 0.66$$

$$< P(E|D) = 0.68 .$$

Tämän mukaan potilaan siis kannattaa valintatilanteessa valita dialyysihoito, jos hän haluaa maksimoida eloonjäämisensä todennäköisyyden.

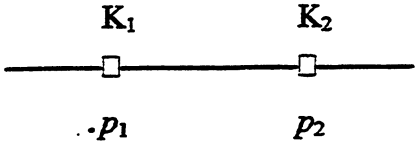
Todellisuudessa valinta ei olisi ollenkaan helppo. Tämä johtuu siitä, että onistuneen munuaisensiirron jälkeen ei tarvita muita hoitotoimenpiteitä, mutta dialyysihoito vaatii useita monituntisia hoitojaksoja joka viikko. ●

TOIMINTAVERKOT

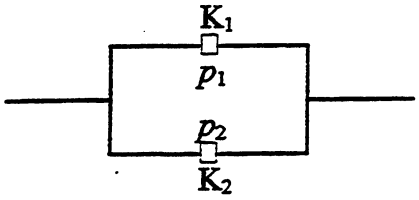
Toisena esimerkkinä verkkodiagrammien soveltamisesta tarkastellaan *toiminta-verkkoja*. Toimintaverkot kuvaavat useista komponenteista koostuvan *systeemin* toimintaa. Tehtävänä on laskea systeemin *toimintatodennäköisyys*, kun komponenttien toimintatodennäköisyydet tunnetaan.

Toimintaverkkojen komponentit voidaan kytkeä kahdella erilaisella tavalla toisiinsa. Kytkentöjä kutsutaan *sarjaan kytkennäksi* ja *rinnan kytkennäksi*. Nimitykset ovat peräisin sähköverkkojen kuvaamisessa käytetystä kielestä. Kytkentöjä havainnollistavat kahden komponentin tapauksessa seuraavat verkot, joissa K_i tarkoittaa komponenttia i ja p_i on komponentin i toimintatodennäköisyys ($i = 1,2$).

Sarjaan kytkentä:



Rinnan kytkentä:



Systeemin toiminnan määräävät seuraavat säännöt:

- Sarjaan kytkentä toimii, jos *kaikki* sen komponentit toimivat.
- Rinnan kytkentä toimii, jos *ainakin yksi* sen komponenteista toimii.

Oletetaan, että systeemiä vastaava toimintaverkko ja komponenttien toiminta-todennäköisyydet tunnetaan. Oletetaan lisäksi, että yhdenkään komponentin toiminta tai toimimattomuus ei vaikuta systeemin muiden komponenttien toiminta-todennäköisyyksiin.[1] Miten määrätään systeemin toimintatodennäköisyys?

Seuraavat säännöt määräävät systeemin toimintatodennäköisyyden:

Olkoot *sarjaan* kytkettyjen komponenttien K_1 ja K_2 toimintatodennäköisyydet p_1 ja p_2. Koska jokaisen komponentin toimiminen oletetaan riippumattomaksi toisten komponenttien toimimisesta, komponenttien K_1 ja K_2 muodostaman systeemin toimintatodennäköisyys on riippumattomien tapahtumien tulosäännön mukaan

$$\begin{aligned} &P(K_1 \text{ toimii } \textit{ja}\ K_2 \text{ toimii}) \\ &\quad = P(K_1 \text{ toimii})P(K_2 \text{ toimii}) \\ &\quad = p_1p_2. \end{aligned}$$

Olkoot *rinnan* kytkettyjen komponenttien K_1 ja K_2 toimintatodennäköisyydet p_1 ja p_2. Niiden muodostaman systeemin toimintatodennäköisyys on yleisen yhteenlaskusäännön mukaan

$$\begin{aligned} &P(K_1 \text{ toimii } \textit{tai}\ K_2 \text{ toimii}) \\ &\quad = P(K_1 \text{ toimii}) + P(K_2 \text{ toimii}) - P(K_1 \text{ toimii } \textit{ja}\ K_2 \text{ toimii}) \\ &\quad = p_1 + p_2 - p_1p_2. \end{aligned}$$

Huomaa, että samaan tulokseen päästään myös seuraavalla laskutoimituksella:

Komplementtisäännön mukaan

$$P(K_1 \text{ toimii } \textit{tai}\ K_2 \text{ toimii}) = 1 - P(K_1 \text{ ei toimii } \textit{ja}\ K_2 \text{ ei toimi}).$$

Riippumattomien tapahtumien kertosäännön mukaan

$$P(K_1 \text{ ei toimi } \textit{ja}\ K_2 \text{ ei toimi}) = P(K_1 \text{ ei toimi})P(K_2 \text{ ei toimi}).$$

Komplementtisäännön mukaan

$$P(K_1 \text{ ei toimi}) = 1 - p_1$$

ja

$$P(K_2 \text{ ei toimi}) = 1 - p_2.$$

Yhdistämällä nämä tulokset ja sieventämällä saadaan

$$\begin{aligned} P(K_1 \text{ toimii } \textit{tai}\ K_2 \text{ toimii}) \quad &= 1 - (1 - p_1)(1 - p_2) \\ &= p_1 + p_2 - p_1p_2. \end{aligned}$$

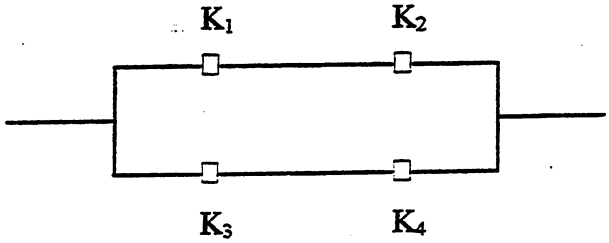
Tulos on tietysti sama kuin edellä.

ESIMERKKI 3.

Tarkastellaan seuraavaa verkkoa, jossa jokaisen komponentin toiminta-todennäköisyys on p:

---

[1] Tämä oletus saattaa olla käytännössä epärealistinen.



Koska komponentit K_1 ja K_2 on kytketty sarjaan, niiden muodostaman osasysteemin toimintatodennäköisyys on edellä sarjaan kytkennälle esitetyn kaavan mukaan ($p_1 = p_2 = p$)

$$p^2.$$

Myös komponenttien K_3 ja K_4 muodostaman osasysteemin toimintatodennäköisyys on edellä sarjaan kytkennälle esitetyn kaavan mukaan ($p_3 = p_4 = p$)

$$p^2.$$

Systeemi muodostuu näistä kahdesta osasysteemistä rinnan kytkettynä, joten sen toimintatodennäköisyys on edellä rinnan kytkennälle esitetyn kaavan mukaan ($p_1 = p_2 = p^2$)

$$p^2 + p^2 - p^2 \cdot p^2 = 2p^2 - p^4.$$

Seuraavassa taulukossa systeemin toimintatodennäköisyys on ilmoitettu muutamalle erilaiselle p:n arvolle:

p	Systeemin toiminta-todennäköisyys
0.1	0.0199
0.2	0.0784
0.3	0.1719
0.4	0.2944
0.5	0.4375
0.6	0.5904
0.7	0.7399
0.8	0.8704
0.9	0.9639

Huomaa, että tarkastelun kohteena olevan systeemin toimintatodennäköisyys kasvaa komponenttien toimintatodennäköisyyden kasvaessa. Jos komponenttien toimintatodennäköisyys kasvaa tarpeeksi suureksi, tarkasteltavan systeemin toimintatodennäköisyys tulee jopa suuremmaksi kuin yksittäisen komponentin toimintatodennäköisyys. ●

Huomaa, että samanlaisten komponenttien kytkeminen rinnan lisää systeemin toimintavarmuutta. Sen sijaan sarjaan kytkennät vähentävät systeemin toiminta-

varmuutta. Siksi samanlaisten komponenttien kytkemistä rinnan saatetaan käyttää laitteissa ja koneissa toimintavarmuuden kasvattamiseksi. Monimutkaisten työsuoritusten yhteydessä saman operaation toistamista voidaan perustella samalla tavalla. Tällaiset pohdinnat ovat keskeisessä asemassa esimerkiksi teollisuustuotteiden laadun varmistuksessa.

1.2 SATUNNAISMUUTTUJAT

1.2.1 JOHDANTO

Otosavaruus määriteltiin kappaleessa 1.1.2 satunnaisilmiön tulosvaihtoehtojen joukoksi. Jos heitämme rahaa 3 kertaa, mahdollisina tulosvaihtoehtoina eli alkeistapahtumina ovat seuraavat kruunien ja klaavojen jonot:

HHH

HHT HTH THH

HTT THT TTH

TTT

Jonoissa on merkitty H = kruuna (*engl.* head) ja T = klaava (*engl.* tail).

Tilastollinen tutkimus kohdistuu satunnaisilmiöiden *numeerisiin* tuloksiin. Kun satunnaisilmiönä on rahan heittäminen 3 kertaa, kruunien lukumäärä heitoissa on tällainen numeerinen tulos. Merkitään kruunien lukumäärää heitoissa symbolilla X. Muuttujan X arvot eri tulosvaihtoehdoille voidaan järjestää seuraavaksi taulukoksi:

Tulosvaihtoehto	X:n arvo
HHH	3
HHT, HTH, THH	2
HTT, THT, TTH	1
TTT	0

Kun 3:n rahanheiton heittosarjoja toistetaan, kruunien lukumäärä X vaihtelee heittosarjasta toiseen. Tämä vaihtelu on ilmiöön liittyvää *satunnaisvaihtelua*. Muuttujaa X kutsutaan *satunnaismuuttujaksi*, koska sen arvot vaihtelevat satunnaisesti heittosarjasta toiseen. Nimityksen taustalla on seuraava ajatus: Vaikka satunnaismuuttujan arvot tunnetaan, emme tiedä mikä arvoista *realisoituu* eli *toteutuu* ennenkuin satunnaisilmiö sattuu. Satunnaismuuttujan toteutunut arvo riippuu siis siitä, mikä tulosvaihtoehdoista on satunnaisilmiön tulos. Todennäköisyyden frekvenssitulkinnan mukaan tulosvaihtoehtojen esiintymisfrekvenssit riippuvat niiden todennäköisyyksistä. Siten todennäköisyyden lait määräävät myös satunnaismuuttujan arvojen esiintymisen.

Edellä esitetystä taulukosta nähdään, että satunnaismuuttuja X arvoihin voidaan liittää todennäköisyydet seuraavan taulukon mukaan:

X	Toden-näköisyys
0	1/8
1	3/8
2	3/8
3	1/8

Määrittelemme seuraavaksi satunnaismuuttujan käsitteen:

SATUNNAISMUUTTUJA

Satunnaismuuttuja on muuttuja, jonka arvot määrää satunnaisilmiön numeerinen tulos.

Matemaattiselta kannalta satunnaismuuttuja on *funktio* eli *kuvaus*, joka kuvaa satunnaisilmiöön liittyvät tapahtumat reaalilukujen joukkoon.[1]

Merkitsemme satunnaismuuttujia tavallisesti aakkosten loppupään isoilla kursivoiduilla kirjaimilla X, Y ja Z. On syytä huomata, että satunnaismuuttujan arvo on *täysin määrätty* välittömästi sen jälkeen, kun satunnaisilmiö on sattunut. Sen sijaan ennen ilmiön sattumista tiedämme vain millä todennäköisyyksillä satunnaismuuttuja saa arvonsa.

Minkä tahansa satunnaisilmiön tulosvaihtoehtoihin voidaan liittää numeeriset arvot. Useissa satunnaisilmiöissä tulosvaihtoehdot itse ovat numeerisia. Näin on asian laita esimerkiksi silloin, kun tarkastellaan kvantitatiivisia ilmiöitä (kts. K1: 2.5). Esimerkkejä kvantitatiivisista muuttujista ovat lukumäärä, pituus, paino, pinta-ala ja hinta. Vaikka satunnaisilmiön tulos ei olisikaan suoraan numeerinen, voidaan tulos aina kuitenkin *koodata* numeeriseksi. Tämä vastaa sitä koodausta, joka tehdään silloin, kun kvalitatiivisen muuttujan arvot koodataan mittaustilanteessa (kts. K1: 2.5.4).

Tilastotieteen kannalta on tärkeätä se, että *tilastollisen muuttujan havaitut arvot voidaan tulkita satunnaismuuttujan arvoiksi*. Havaintoarvoihin liittyvä satunnaisvaihtelu on seurausta havaintojen keräämisessä käytetystä satunnaistuksesta (kts. kokeita ja otantaa käsitteleviä kappaleita, K1: 2.2 ja 2.4). Tästä tulkinnasta seuraa se, että havaintoarvojen *jakauma* sekä kaikki jakaumaa kuvaavat *tunnusluvut* kuten keskiarvo, hajonta ja korrelaatiokerroin ovat havaintoarvojen funktioina satunnaismuuttujia.

ESIMERKKI 1.

Olkoon satunnaisilmiönä lapsen sukupuolen määräytyminen. Tällöin otosavaruus

---

[1] Satunnaismuuttuja on funktio, joka kuvaa sellaiset otosavaruuden osajoukot, jotka muodostavat σ-algebran reaalilukujen joukkoon. Satunnaismuuttuja on siis *mitallinen* kuvaus.

$S = \{\text{tyttö,poika}\}$.

Voimme liittää tähän satunnaisilmiöön satunnaismuuttujan seuraavasti:

$$X = \begin{cases} 1, \text{ jos lapsi on tyttö} \\ 0, \text{ jos lapsi on poika} \end{cases}$$

Huomaa, että satunnaismuuttuja X:n koodaus on täysin mielivaltainen; olennaista on vain se, että tytöt ja pojat saavat eri koodin. ●

Kuten edellä on todettu, satunnaisilmiötä kuvaava *todennäköisyysmalli* muodostuu otosavaruudesta S ja otosavaruuden osajoukkoihin eli tapahtumiin liittyvistä todennäköisyyksistä. Kun satunnaisilmiön tuloksia kuvataan satunnaismuuttujalla, otosavaruus S muodostuu satunnaismuuttujan mahdollisista arvoista. Otosavaruutta S ei useinkaan määritellä eksplisiittisesti, koska se on asiayhteydestä selvä. Toinen osa todennäköisyysmallia liittyy satunnaisilmiön tapahtumien todennäköisyyksien määrittelyyn. Tämä tapahtuu määrittelemällä satunnaismuuttujan saamien arvojen todennäköisyyksien jakauma eli *todennäköisyysjakauma*. Todennäköisyysjakaumia tarkastellaan seuraavassa erikseen *diskreeteille* ja *jatkuville* satunnaismuuttujille (vrt. diskreetin ja jatkuvan muuttujan käsitteitä mittaamista käsittelevässä kappaleessa 1. Kirjassa, kts. K1: 2.5.5).

1.2.2 DISKREETIT SATUNNAISMUUTTUJAT JA NIIDEN TODENNÄKÖISYYSJAKAUMAT

Annetaan ensin diskreetin satunnaismuuttujan määritelmä:[1]

DISKREETTI SATUNNAISMUUTTUJA

Satunnaismuuttuja on *diskreetti*, jos se saa äärellisen määrän erillisiä arvoja.

Olkoon diskreetin satunnaismuuttujan X arvot $x_1, x_2, \ldots, x_k$. Satunnaismuuttujaan X voidaan liittää *todennäköisyysmalli* seuraavassa esitettävällä tavalla. Perusjoukko S muodostuu X:n saamista arvoista:

$$S = \{x_1, x_2, \ldots, x_k\}.$$

Oletamme yleensä, että perusjoukon S alkiot on järjestetty seuraavalla tavalla:

$$x_1 < x_2 < \ldots < x_k.$$

Tämä oletus ei ole mitenkään välttämätön, mutta oletus yksinkertaistaa tavallisesti esitystä niin, että oletus kannattaa tehdä.

Diskreetin satunnaismuuttujan arvoihin voidaan liittää todennäköisyydet luettelemalla tulosvaihtoehtojen x_i todennäköisyydet. Seuraavassa tarkastellaan diskreetin satunnaismuuttujan todennäköisyysjakauman määrittelemistä.

---

[1] Itse asiassa diskreetti satunnaismuuttuja voidaan määritellä siten, että se saa *numeroituvasti* äärettömän määrän erillisiä arvoja. Sivuutamme kuitenkin tämän yleisemmän määritelmän tässä esityksessä.

DISKREETIN SATUNNAISMUUTTUJAN TODENNÄKÖISYYS-JAKAUMA

Olkoon diskreetin satunnaismuuttujan numeeriset tulosvaihtoehdot $x_1, x_2, \ldots, x_k$. Tulosvaihtoehdot x_i ja niiden todennäköisyydet

$$P(X = x_i) = p_i$$

muodostavat *diskreetin todennäköisyysjakauman*, jos todennäköisyydet p_i toteutettavat ehdot

1. $0 \le p_i \le 1.$

2. $\sum_{i=1}^{k} p_i = p_1 + p_2 + \ldots + p_k = 1.$

Diskreetin satunnaismuuttujan X saamien arvojen x_i ja niihin liittyvien todennäköisyyksien p_i määrittelemiä lukupareja (x_i, p_i), $i = 1,2,\ldots,k$ kutsutaan usein diskreetin jakauman *pistetodennäköisyysfunktioksi*.

Huomaa, että diskreetin todennäköisyysjakauman määritelmä sisältää sen, että

$$P(X = x) = 0,$$

jos $x \notin \{x_1, x_2, \ldots, x_k\}$ eli, jos x ei kuulu otosavaruuteen S.

Jos A on jokin perusjoukon S osajoukko, niin *minkä tahansa* muotoa $X \in A$ olevan tapahtuman todennäköisyys saadaan laskemalla yhteen niiden tulosvaihtoehtojen x_i todennäköisyydet p_i, jotka yhdessä muodostavat joukon A.

Jos diskreetti satunnaismuuttuja saa vain muutamia arvoja, sen todennäköisyysjakauma voidaan antaa taulukkona. Diskreetin jakauman pistetodennäköisyysfunktiota voidaan kuvata graafisesti piirtämällä todennäköisyyksistä *pylväsdiagrammi* (kts. K1: 3.2.3).

ESIMERKKI 1.

Erään kurssin loppukokeessa arvosanojen jakauma oli seuraava:

Arvosana	0	1	2	3
Osuus (%)	20	30	40	10

Tarkastellaan satunnaisesti valitun loppukokeeseen osallistuneen opiskelijan arvosanaa. Arvosana on diskreetti satunnaismuuttuja, jonka jakauma (pistetodennäköisyysfunktio) voidaan antaa seuraavana taulukkona:

Arvosana x_i	0	1	2	3
Todennäköisyys p_i	0.2	0.3	0.4	0.1

Olkoon satunnaisesti valitun opiskelijan arvosana satunnaismuuttuja X. Siten tapahtuma "Satunnaisesti valittu opiskelija sai vähintään arvosanan 2" voidaan ilmaista seuraavassa muodossa:

$$X \geq 2.$$

Tämän tapahtuman todennäköisyys on

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 0.4 + 0.1 = 0.5.$$

Huomaa, että laskutoimituksessa on itse asiassa käytetty toisensa poissulkevien tapahtumien yhteenlaskusääntöä. ●

ESIMERKKI 2.

Satunnaislukujen taulukko muodostuu numeroiden 0,1,2,3,4,5,6,7,8,9 muodostamasta jonosta, jossa jokaisella numerolla on sama todennäköisyys esiintyä millä tahansa paikalla jonossa (kts. K1: 2.2.2).[1]

Poimitaan taulukosta satunnaisesti 1-numeroinen luku. Voimme liittää poimintaan satunnaisilmiönä satunnaismuuttujan X, joka kuvaa poimitun luvun numeroarvoa. Vähimmäisvaatimus satunnaislukujen taulukolle on se, että satunnaismuuttujan X todennäköisyysjakauma liittää tapahtumiin

$$X = k,\ k = 0,1,2,\ldots,9$$

yhtä suuret todennäköisyydet. Siten

$$P(X = k) = 1/10,\ k = 0,1,2,\ldots,9.$$

Tapahtumavaihtoehdot $X = k$ ovat siten *symmetrisiä*. Satunnaismuuttujaan X liittyvän pistetodennäköisyysfunktion määrittelee lukuparien

$$(k,\ 1/10),\ k = 0,1,2,\ldots,9$$

joukko. ●

Esimerkin 2 jakauma on erikoistapaus ns. *diskreetistä tasaisesta jakaumasta*.

ESIMERKKI 3.

Millainen on kruunien lukumäärän jakauma 4:n rahan heitossa?

Oletetaan, että rahat ovat harhattomia, ts., että yhden rahan heiton tuloksia (H = kruuna, T = klaava) voidaan pitää symmetrisinä tulosvaihtoehtoina. Tällöin

---

[1] Lisäksi vaaditaan, että numerot ovat toisistaan *riippumattomia*, ts. tieto siitä millaisia numeroita on taulukon jossakin osassa ei sisällä tietoa siitä millaisia numeroita on taulukon muissa osissa. Tämä tarkoittaa sitä, että numeroiden 0,1,2,3,4,5,6,7,8,9 on oltava taulukossa *sekaisessa* järjestyksessä.

$$P(H) = P(T) = 1/2.$$

Kruunien lukumäärä 4:n rahan heitossa on diskreetti satunnaismuuttuja, jonka mahdolliset arvot ovat 0,1,2,3,4. Merkitään tätä satunnaismuuttujaa X:llä. Mitkä ovat eri kruunien lukumääriä vastaavat todennäköisyydet? Vastaus saadaan todennäköisyyslaskennan avulla.

Voimme pitää tarkasteltavaa satunnaisilmiötä toistokokeena, jossa rahaa heitetään 4 kertaa. Tarkastelun kohteena on kuinka usein tapahtuma H sattuu tässä 4:n toistokokeen sarjassa. Kuten edellä on todettu tällaisen tapahtuman todennäköisyydet saadaan binomikaavasta.

Siten

$$P(X = k) = \binom{4}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{4-k} = \binom{4}{k}\frac{1}{2^4}.$$

Todennäkäisyydet eri k:n arvoille ovat seuraavat:

$$P(X = 0) = 1/16 = 0.0625,$$

$$P(X = 1) = 4/16 = 0.25,$$

$$P(X = 2) = 6/16 = 0.375,$$

$$P(X = 3) = 4/16 = 0.25,$$

$$P(X = 4) = 1/16 = 0.0625.$$

Siten satunnaismuuttujan X todennäköisyysjakauma (pistetodennäköisyysfunktio) voidaan ilmaista seuraavana taulukkona:

Tulos x_i	0	1	2	3	4
Todennäköisyys p_i	0.0625	0.25	0.375	0.25	0.0625

DISKREETIN MUUTTUJAN EMPIIRISET JAKAUMAT

On syytä huomata, että mihin tahansa diskreetin muuttujan *havaittujen arvojen jakaumaan* voidaan liittää diskreetti todennäköisyysjakauma empiirisen todennäköisyyden määritelmää soveltamalla.

Olkoot tarkasteltavan diskreetin muuttujan mahdolliset arvot

$$x_1, x_2, \ldots, x_k$$

ja oletetaan, että muuttujasta kerätty n havaintoa. Oletetaan, että muuttujan arvojen $x_1, x_2, \ldots, x_k$ havaitut frekvenssit ovat

$$f_1, f_2, \ldots, f_k,$$

jolloin $f_1 + f_2 + \ldots + f_k = n$. Soveltamalla empiirisen todennäköisyyden määritelmää voidaan määritellä satunnaismuuttuja X, jonka arvoihin $x_1, x_2, \ldots, x_k$ liitetään todennäköisyydet niiden *suhteellisina frekvensseinä*:

$$P(X = x_i) = f_i/n,\ i = 1,2,\ldots,k.$$

Siten mitä tahansa diskreetin muuttujan havaittujen arvojen jakaumaa kuvaavaa *pylväsdiagrammia* vastaa jonkin diskreetin satunnaismuuttujan *empiirinen* pistetodennäköisyysfunktio (kts. K1: 3.2.3).

1.2.3 JATKUVAT SATUNNAISMUUTTUJAT

Edellisen kappaleen esimerkissä 2 tarkasteltiin 1-numeroisen satunnaisluvun poimintaan liittyvää diskreettiä satunnaismuuttujaa. Tällöin tulosvaihtoehtoina olivat erilliset luvut 0,1,2,3,4,5,6,7,8,9. Oletetaan nyt, että haluamme poimia luvun satunnaisesti 0:n ja 1:n väliltä niin, että *mikä tahansa* luku ko. väliltä kelpaa tulokseksi. Poimintaa voidaan havainnollistaa *onnenpyörän* avulla: Oletamme, että onnenpyörä pyörii vapaasti keskipisteensä ympäri ja, että se voi pysähtyä pyöräytyksen jälkeen mihin asentoon tahansa. Pyörän kehä asetetaan vastaamaan lukuja 0:n ja 1:n väliltä[1]. Pyörän ulkopuolelle kiinnitetään osoitin, jonka osoittamasta kohdasta poimittava luku luetaan[2].

Voimme liittää tähän satunnaislukujen poimintamenetelmään satunnaismuuttujan X, joka voi saada mitä tahansa arvoja 0:n ja 1:n väliltä. Kutsumme satunnaismuuttujaa X *jatkuvaksi*, koska X saa arvoja jatkuvasti joltakin väliltä. Jonkinlaisena jatkuvien satunnaismuuttujien vastakohtana *diskreetit* satunnaismuuttujat voivat saada vain erillisiä arvoja. Väli [0,1) muodostaa tarkastelun kohteena olevan satunnaisilmiön otosavaruuden[3]. Jotta todennäköisyysmalli satunnaismuuttujalle X olisi täydellinen, meidän on annettava sääntö, joka liittää ko. satunnaisilmiön erilaisiin tapahtumiin todennäköisyydet. Esimerkissämme haluaisimme kaikkien tulosvaihtoehtojen todennäköisyyksien olevan yhtä suuria. Emme kuitenkaan voi liittää todennäköisyyksiä satunnaismuuttujan X yksittäisiin arvoihin ja laskea todennäköisyyksiä yhteen yhdistettyjen tapahtumien todennäköisyyksien määräämiseksi. Tämä johtuu siitä, että onnenpyörä voi pysähtyä äärettömän moneen eri asentoon ja *todennäköisyys, että pyörä pysähtyy johonkin tiettyyn asentoon on 0.*

Joudumme siksi kehittämään uuden tavan liittää todennäköisyydet jatkuviin satunnaismuuttujiin liittyviin tapahtumiin. Tämä tapahtuu kuvaamalla jatkuvan satunnaismuuttujan arvojen todennäköisyysjakaumaa jatkuvalla käyrällä. Jatkuvien satunnaismuuttujien kuvaamissa satunnaisilmiöissä tapahtumat voidaan samaistaa reaalilukujen väleihin tai ne voidaan panna kokoon reaalilukujen väleistä joukko-opin

---

[1] Monikäsitteisyyden välttämiseksi sovitaan, että luku 0 on poimittavien lukujen joukossa, mutta luku 1 ei ole.

[2] Joudumme olettamaan, että lukeminen voidaan tehdä *miten tarkasti tahansa*.

[3] Käytämme seuraavia merkintöjä:

$$[a,b] = \{a \le x \le b\},$$

$$[a,b) = \{a \le x < b\},$$

$$(a,b] = \{a < x \le b\},$$

$$(a,b) = \{a < x < b\}.$$

Välit eroavat toisistaan vain päätepisteittensä suhteen. Huomaa, että $x \in [a,b]$ tarkoittaa, että

$$a \le x \le b.$$

sääntöjen mukaan. Tapahtuman todennäköisyys saadaan käyrän ja vaaka-akselin välisen alueen *pinta-alana*, kun aluetta rajoittaa tapahtumaan liittyvä reaalilukujen väli.

Käyrän ja vaaka-akselin välistä aluetta kutsutaan jatkossa tavallisesti *käyrän alapuoliseksi alueeksi*. Koska todennäköisyydet ovat ei-negatiivisia, käyrä ei saa saada muita kuin ei-negatiivisia arvoja. Koska koko otosavaruuden todennäköisyyden pitää olla 1, käyrän alapuolisen alueen pinta-alan pitää olla 1 sillä välillä, jonka koko otosavaruus määrää.

JATKUVAT SATUNNAISMUUTTUJAT

Satunnaismuuttuja on *jatkuva*, jos se saa kaikki arvot joltakin reaalilukujen väliltä.

Jatkuvaan satunnaismuuttujaan X liittyvä otosavaruus S muodostuu siis jostakin reaalilukujen välistä $[a, b]$. Tämä väli voi ulottua reaalilukujen muodostamalla lukusuoralla jompaan kumpaan suuntaan tai molempiin suuntiin äärettömyyteen.

Olkoon A jokin otosavaruuden S osaväli. Tämä merkitsee sitä, että A on muotoa $[c, d]$ ja $A \subset S$. Jatkuvaan satunnaismuuttujaan X liittyvä todennäköisyysmalli liittää jokaiseen tapahtumaan $X \in A$ todennäköisyyden $P(X \in A)$, joka saadaan määräämällä sen alueen pinta-ala, jota rajoittavat satunnaismuuttujan X jakaumaa kuvaava jatkuva käyrä $f(x)$ ja x-akselin pisteiden osajoukko A.

JATKUVAN SATUNNAISMUUTTUJAN TODENNÄKÖISYYS-JAKAUMA

Olkoon jatkuvaan satunnaismuuttujaan X liittyvä otosavaruus S. Funktio $f(x)$ määrittelee satunnaismuuttujan X *jatkuvan todennäköisyysjakauman*, jos $f(x)$ toteuttaa seuraavat ehdot:

1. $f(x)$ on jatkuva.
2. $f(x) \geq 0$, kaikille x.
3. Käyrän $f(x)$ ja vaaka-akselin välisen alueen pinta-ala on 1 sillä välillä, joka määrittelee otosavaruuden S.

Jatkuvan satunnaismuuttujan X saamia arvoja satunnaismuuttujaan liittyvine jatkuvine käyrineen kutsutaan siis *jatkuvaksi todenäköisyysjakaumaksi*. Jatkuvaan satunnaismuuttujaan liittyvien tapahtumien todennäköisyydet määrittelevää jatkuvaa käyrää $f(x)$ kutsutaan satunnaismuuttujan X *tiheysfunktioksi*. Mikä tahansa ehdot 1, 2 ja 3 täyttävä jatkuva käyrä on jonkun jatkuvan satunnaismuuttujan tiheysfunktio.

Nimitys tiheysfunktio perustuu seuraavaan ajatukseen: Olkoot $[c,d]$ ja $[e,f]$ kaksi x-akselin väliä, jotka ovat otosavaruuden S osajoukkoja. Oletetaan, että tiheysfunktion $f(x)$ alapuolisen alueen pinta-ala välillä $[c,d]$ on suurempi kuin välillä $[e,f]$. Koska tämä merkitsee sitä, että tapahtuma $X \in [c,d]$ on todennäköisempi kuin tapahtuma $X \in [e,f]$, voidaan ajatella, että todennäköisyys on jakautunut "tiheämmin"

välille $[c,d]$ kuin välille $[e,f]$. Siten jaktuvan satunnaismuuttujan tiheysfunktio kuvaa *todennäköisyysmassan* jakautumista otosavaruuden S osaväleille.

Huomaa, että tiheysfunktion $f(x)$ arvot *eivät ole* todennäköisyyksiä. Todennäköisyyksiä ovat käyrän $f(x)$ alapuolisten alueiden pinta-alat. Yksittäisten pisteiden muodostamien tapahtumien todennäköisyydet ovat jatkuvan satunnaismuuttujan tapauksessa aina nollia:

$$P(X = x) = 0.$$

Jatkuva jakauma eroaa tässä dikreetistä jakaumasta, jossa

$$P(X = x_i) = p_i > 0,$$

jos $x_i \in S$ eli x_i on otosavaruuden piste.

ESIMERKKI 1.

Luvun poimintaa satunnaisesti väliltä $[0,1)$ voidaan kuvata tiheysfunktiolla

$$f(x) = 1, \text{ kun } 0 \leq x < 1.$$

Tiheysfunktio $f(x)$ on siis vakio 1 välillä $[0,1)$. Tapahtuman

$$0.2 < x < 0.35$$

todennäköisyydeksi saadaan helposti

$$P(0.2 < x < 0.35) = 0.15. \bullet$$

Esimerkin 1 todennäköisyysjakauma on erikoistapaus *jatkuvasta tasaisesta jakaumasta*. Vertaa esimerkissä 1 esitettyä jatkuvaa tasaista jakaumaa kappaleessa 1.2.2 esitettyyn esimerkkiin 2 diskreetistä tasaisesta jakaumasta. Diskreetin tasaisen jakauman tapauksessa jokaisella perusjoukon alkeistapahtumalla oli sama todennäköisyys. Tätä vastaa jatkuvan tasaisen jakauman tapauksessa se, että kaikkien samanmittaisten välin $[0,1)$ osavälien todennäköisyydet ovat samoja.

JATKUVIIN SATUNNAISMUUTTUJIIN LIITTYVÄT TODENNÄKÖISYYDET JA INTEGROINTI

Jatkuvan satunnaismuuttujan tiheysfunktio voidaan määritellä matemaattisesti seuraavalla tavalla:

JATKUVAN SATUNNAISMUUTTUJAN TIHEYSFUNKTIO

Saakoon satunnaismuuttuja X arvoja jatkuvasti väliltä $[a,b]$, jossa voi olla $a = -\infty$ tai $b = +\infty$. Jos on olemassa jatkuva funktio $f(x)$ siten, että

1. $f(x) \geq 0$, kaikille $x \in [a,b]$,

2. $\int_a^b f(x)dx = 1$,

niin $f(x)$ on satunnaismuuttujan X *tiheysfunktio*.

Tapahtuman $c \leq X \leq d$ todennäköisyys voidaan määrätä integroimalla funktio $f(x)$:

$$P(c \leq X \leq d) = \int_c^d f(x)dx.$$

Integraali määrittelee sen alueen pinta-alan, jota rajoittavat käyrä $f(x)$, x-akseli ja x-akselin väli $[c,d]$ (tai oikeammin pisteistä c ja d käyrälle $f(x)$ piirretyt pystysuorat viivat).

Käytännössä esiintyvissä tilanteissa integrointia ei tarvitse yleensä suorittaa eskplisiittisesti, koska tavanomaisiin jakaumiin liittyviä pinta-aloja on taulukoitu valmiiksi taulukkokokoelmiin tai tarvittavat pinta-alat laskee käytetty tilasto-ohjelma.

ESIMERKKI 2.

Tilastotieteen tärkein jatkuva jakauma on *normaalijakauma*, jonka tiheysfunktio on muotoa

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

Tiheysfunktion lausekkeessa exp tarkoittaa eksponenttifunktiota $\exp(z) = e^z$.

Voidaan osoittaa, että $f(x)$ on x:n funktiona jatkuva ja, että

1. $f(x) > 0$, kaikille x,

2. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Siten $f(x)$ on todellakin jatkuvan jakauman tiheysfunktio.

Jos satunnaismuuttujalla X on ym. tiheysfunktio, sanotaan, että X on *normaalinen* $N(\mu,\sigma^2)$ tai, että X *noudattaa normaalijakaumaa parametrein* μ ja σ^2.[1]

Jos $\mu = 0$ ja $\sigma = 1$, sanotaan normaalijakaumaa *standardoiduksi*. Olkoon Z standardoitua normaalijakaumaa noudattava satunnaismuuttuja. Muotoa

$$Z \leq z$$

olevien tapahtumien todennäköisyyksiä on taulukoitu taulukkokokoelmiin.

Useat tilasto-ohjelmat tulostavat muotoa $X \leq x$ olevien tapahtumien todennäköisyyksiä. Kaikkien muiden tapahtumien todennäköisyydet saadaan tätä muotoa olevien tapahtumien todennäköisyyksistä käyttämällä hyväksi todennäköisyyden laskusääntöjä.

Esimerkiksi tapahtuman

$$a \leq X \leq b$$

todennäköisyys saadaan laskutoimituksella

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a),$$

jossa on sovellettu toisensa poissulkevien tapahtumien yhteenlaskusääntöä: Tapahtumat $a \leq X \leq b$ ja $X < a$ ovat toisensa poissulkevia ja niiden yhdiste on tapahtuma $X \leq b$.[2]

Normaalijakaumaa $N(\mu,\sigma^2)$ noudattaviin satunnaismuuttujiin X liittyvien muotoa

$$X \leq x$$

olevien tapahtumien todennäköisyydet voidaan määrätä *standardoinnilla*. Olkoon

$$z = \frac{x - \mu}{\sigma}$$

muuttujan x *standardoitu arvo*. Jos satunnaismuuttuja Z noudattaa standardoitua normaalijakaumaa $N(0,1)$, niin

$$P(X \leq x) = P(Z \leq z),$$

Tämä johtuu siitä, että standardoiduissa yksiköissä kaikki normaalijakaumat ovat samanmuotoisia.

Olkoon esimerkiksi X normaalinen $N(1,0.5^2)$. Tällöin tapahtuman

$$-0.5 \leq X \leq 2$$

---

[1] Parametrien μ ja σ^2 tulkintaan palataan myöhemmin.

[2] Jotta vältyttäisiin monikäsitteisyysongelmalta, piste a on sijoitettu täsmälleen yhteen väleistä (tapahtumista) $a \leq X \leq b$ ja $X < a$. Koska jatkuvan jakauman tapauksessa $P(X = a) = 0$ kaikille a, päätepisteen kuuluminen väliin ei muuta välin todennäköisyyttä. On syytä muistaa, että diskreettien jakaumien tapauksessa yksittäisille pisteille on määritelty positiiviset todennäköisyydet, jolloin välin päätepisteen kuuluminen väliin saattaa vaikuttaa välin todennäköisyyteen.

todennäköisyys saadaan seuraavalla tavalla:

Ensinnäkin

$$P(-0.5 \le X \le 2) = P(X \le 2) - P(X < -0.5).$$

Standardoidun normaalijakauman taulukoista saadaan

$$P(X \le 2) = P(Z \le \frac{2-1}{0.5}) = P(Z \le 2) \approx 0.9772,$$

$$\begin{aligned} P(X < -0.5) &= P(Z \le \frac{-0.5-1}{0.5}) \\ &= P(Z \le -3) \\ &= 1 - P(Z \le 3) \\ &\approx 0.0013. \end{aligned}$$

Siten

$$P(-0.5 \le X \le 2) \approx 0.9772 - 0.0013 = 0.9759. \bullet$$

Johdantokurssilla jatkuvista jakaumista oli esillä normaalijakauman lisäksi χ^2- ja Studentin t-jakaumat.

JATKUVAN MUUTTUJAN EMPIIRISET JAKAUMAT

On syytä huomata, että mihin tahansa jatkuvan muuttujan *havaittujen* arvojen jakaumaan voidaan liittää todennäköisyysjakauma empiirisen todennäköisyyden määritelmää soveltamalla.

Oletetaan, että tarkasteltava jatkuva muuttuja saa arvoja väliltä

$$[a,b],$$

jossa voi olla $a = -\infty$ tai $b = +\infty$. Oletetaan, että muuttujasta on kerätty n havaintoa. Oletetaan, että väli $[a,b]$ on jaettu k toisensa poissulkevaan väliin pisteillä $x_0 = a < x_1 < x_2 < \ldots < x_{k-1} < x_k = b$. Oletetaan, että muuttujan arvojen havaitut frekvenssit näiden välien määrittelemissä luokissa ovat

$$f_1, f_2, \ldots, f_k,$$

jolloin $f_1 + f_2 + \ldots + f_k = n$.

Soveltamalla empiirisen todennäköisyyden määritelmää voidaan määritellä satunnaismuuttuja X, johon liitetään todennäköisyydet väleihin

$$[x_{i-1}, x_i),\ i = 1,2,\ldots,k^1$$

liittyvien havaittujen suhteellisten frekvenssien avulla, ts. määritellään

$$P(x_{i-1} \le X < x_i) = f_i/n,\ i = 1,2,\ldots,k.$$

---

[1] Käytämmme *päätepistesopimusta*, jonka mukaan välin vasemmanpuoleinen päätepiste kuuluu väliin, mutta oikeanpuoleinen ei kuulu.

Siten mitä tahansa jatkuvan muuttujan havaittujen arvojen luokiteltua frekvenssijakaumaa kuvaavaa *histogrammia* vastaa jonkin satunnaismuuttujaan *empiirinen* tiheysfunktio (kts. K1: 3.2.2).

Huomaa, että empiirinen tiheysfunktio ei ole minkään diskreetin satunnaismuuttujan pistetodennäköisyysfunktio eikä myöskään minkään jatkuvan satunnaismuuttujan tiheysfunktio. Edellinen väite on triviaalisti totta. Jälkimmäinen väite nähdään todeksi, koska empiirisellä tiheysfunktiolla on (yleensä) *epäjatkuvuus-kohdat* pisteissä $x_0, x_1, x_2, \ldots, x_k$. Sen sijaan empiirinen tiheysfunktio on kyllä jatkuva pisteiden $x_0, x_1, x_2, \ldots, x_k$ välissä. Tämä merkitsee sitä, että jatkuvan muuttujan havaittujen arvojen jakaumaan liittyvä satunnaismuuttuja on esimerkki satunnaismuuttujasta, joka ei ole diskreetti eikä jatkuva.

1.2.4 KERTYMÄFUNKTIO

Teoreettisessa tilastotieteessä todennäköisyysjakaumia tarkastellaan tavallisesti niiden *kertymäfunktioiden* avulla olivatpa jakaumat diskreettejä, jatkuvia tai jotakin muuta tyyppiä. Tämä johtuu siitä, että *kaikki ko. satunnaisilmiöön liittyvät todennäköisyydet voidaan ilmaista kertymäfunktion avulla.*

KERTYMÄFUNKTIO

Satunnaismuuttujan X *kertymäfunktio* $F(x)$ määritellään kaavalla

$$P(X \leq x) = F(x).$$

Kertymäfunktio kuvaa muotoa $X \leq x$ olevien tapahtumien todennäköisyyttä x:n funktiona. Kertymäfunktion $F(x)$ nimitys tulee siitä, että $F(x)$ kuvaa paljonko *todennäköisyysmassaa* on kertynyt vasemmalta pisteeseen x saakka.

Kertymäfunktion tärkeimmät ominaisuudet on koottu seuraavaan luetteloon:

KERTYMÄFUNKTION OMINAISUUDET

Jos funktio $F(x)$ toteuttaa seuraavat ehdot, se on jonkin satunnaismuuttujan *kertymäfunktio*.

1. $F(-\infty) = 0$.
2. $F(+\infty) = 1$.
3. Funktio $F(x)$ on *ei-vähenevä*:

   $F(x_1) \leq F(x_2)$, jos $x_1 \leq x_2$.
4. Funktio $F(x)$ on *oikeanpuoleisesti jatkuva*:

   $F(x+h) \longrightarrow F(x)$, jos $h \longrightarrow 0$ oikealta.

Kertymäfunktion keskeinen asema teoreettisessa tilastotieteessä perustuu siihen, että satunnaismuuttujan X kertymäfunktion $F(x)$ avulla voidaan määrätä kaikki satunnaismuuttujaan X liittyvät todennäköisyydet: kertymäfunktio määrittelee ko. satunnaismuuttujan todennäköisyysjakauman. Tämä johtuu siitä, että mikä tahansa tapahtuma voidaan ilmaista muotoa

$$X \leq x$$

olevien tapahtumien avulla käyttäen apuna joukko-opin sääntöjä. Tapahtuman todennäköisyys saadaan määrätyksi soveltamalla todennäköisyyslaskennan sääntöjä näihin yhdistettyihin tapahtumiin.

DISKREETIN JAKAUMAN KERTYMÄFUNKTIO

Diskreetin jakauman kertymäfunktio on epäjatkuva funktio, jossa on *epäjatkuvuuskohta* eli *hyppäys* niissä pisteissä x_i, joihin liittyy positiivinen todennäköisyys p_i. Hyppäyksen suuruus pisteessä x_i on p_i. Kertymäfunktio on *vakio* niiden otosavaruuden pisteiden x_i *välissä*, joihin liittyy liittyy positiivinen todennäköisyys p_i. Siten diskreetin jakauman kertymäfunktio on muodoltaan *porrasfunktio*, joissa todennäköisyydet p_i, $i = 1,2,\ldots,k$ määräävät portaiden korkeudet ja vastavien otosavaruuden pisteiden välimatkat $x_{i+1} - x_i$, $i = 1,2,\ldots,k-1$ määräävät askelmien pituudet. Portaat lähtevät tasolta 0 ja nousevat oikealle tasolle 1.

Edellä kuvattu porrasfunktio toteuttaa kertymäfunktiolle asetetut ehdot 1–4. Huomaa, että niissä pisteissä x_i, joissa porrasfunktiolla on epäjatkuvuuskohta, pätee seuraava:

$$F(x_i+h) \longrightarrow F(x_i),$$

jos $h \longrightarrow 0$ oikealta.

$$F(x_i+h) \longrightarrow F(x_{i-1}),$$

jos $h \longrightarrow 0$ vasemmalta. Hyppäyksen suuruus pisteessä x_i on

$$F(x_i) - F(x_{i-1}) = p_i.$$

Diskreetin jakauman kertymäfunktion portaat ja pistetodennäköisyydet vastaavat toisiaan seuraavalla tavalla:

DISKREETIN SATUNNAISMUUTTUJAN KERTYMÄFUNKTIO JA PISTETODENNÄKÖISYYSFUNKTIO

$$P(X = x_i) = p_i = F(x_i) - F(x_{i-1}),\ i = 1,2,\ldots,k.$$

Tässä $F(x_0) = 0$. Yhtälöistä nähdään se, että kaikki diskreettiin satunnaismuuttujaan liittyvät todennäköisyydet voidaan määrätä kertymäfunktion avulla.

ESIMERKKI 1.

Olkoon satunnaisilmiönä kolme rahanheittoa ja olkoon satunnaismuuttuja

X = "Klaavojen lukumäärä".

Otosavaruus on tällöin muotoa

$$S = \{0,1,2,3\}.$$

Binomikaavasta seuraa, että

$$P(X = k) = \binom{3}{k}\frac{1}{2^3},\ k = 0,1,2,3.$$

Satunnaismuuttujan X pistetodennäköisyysfunktio on annattu seuraavassa taulukossa:

k	0	1	2	3
p_k	1/8	3/8	3/8	1/8

Satunnaismuuttujan X kertymäfunktio on porrasfunktio

$$F(x)=\begin{cases}0, & x<0\\ \frac{1}{8}, & 0\le x<1\\ \frac{4}{8}, & 1\le x<2\\ \frac{7}{8}, & 2\le x<3\\ 1, & 3\le x\end{cases}$$

Kertymäfunktiosta saadaan esimerkiksi seuraavat todennäköisyydet:

$$P(X\le 1.6)=F(1.6)=p_0+p_1=4/8,$$

$$P(X\le 2)=F(2)=p_0+p_1+p_2=7/8,$$

$$P(X\le 3.1)=F(3.1)=p_0+p_1+p_2+p_3=1.\ \bullet$$

JATKUVAN SATUNNAISMUUTTUJAN KERTYMÄFUNKTIO

Jatkuvan satunnaismuuttujan kertymäfunktio voidaan määritellä tiheysfunktion *integraalin* avulla:

JATKUVAN SATUNNAISMUUTTUJAN KERTYMÄFUNKTIO JA TIHEYSFUNKTIO

$$F(x)=P(X\le x)=\int_{-\infty}^{x} f(t)\,dt.$$

Jatkuvan jakauman tapauksessa kertymäfunktio $F(x)$ kuvaa siis tiheysfunktion $f(x)$ alapuolisen alueen *pinta-alaa* pisteestä x vasemmalle.

Edellä esitetyt kertymäfunktion ominaisuudet 1–4 tulevat ilmeisiksi, kun käytetään kertymäfunktion arvojen pinta-alatulkintaa:

1. $F(-\infty)=0$, koska tiheysfunktion $f(x)$ alapuolisen alueen pinta-ala pisteestä $-\infty$ vasemmalle on 0.
2. $F(+\infty)=1$, koska tiheysfunktion $f(x)$ koko alapuolisen alueen pinta-ala on 1.
3. Funktio $F(x)$ on ei-vähenevä, koska tiheysfunktion alapuolisen alueen pinta-ala pisteestä x vasemmalle kasvaa (tai ei ainakaan vähene), kun x kasvaa.
4. Funktio $F(x)$ on jatkuvana funktiona oikeanpuoleisesti jatkuva.

Jatkuvan satunnaismuuttujan tapauksessa tiheysfunktio saadaan kertymäfunktiosta *derivoimalla*:

$$f(x) = \frac{d}{dx} F(x).$$

ESIMERKKI 1.

Tarkastellaan sellaista onnenpyörää, jossa onnenpyörän keskustaan on asennettu vapaasti pyörivä osoitin. Pyöräytetään osoitinta ja tarkastellaan osoittimen pysähdyttyä sitä kulmaa, jonka osoitin muodostaa lähtöasentoonsa nähden. Olkoon tämä kulma X. X on jatkuva satunnaismuuttuja, joka on jakautunut tasaisesti välille $[0°,360°)$. Satunnaismuuttujan X tiheysfunktio on muotoa

$$f(x) = \begin{cases} \frac{1}{360}, x \in [0,360) \\ 0, \text{muulloin} \end{cases}.$$

Vastaava kertymäfunktio on

$$\begin{aligned} F(x) &= \int_0^x \frac{1}{360} dt \\ &= \frac{1}{360} t \Big|_0^x \\ &= \frac{1}{360} x, \end{aligned}$$

kun $x \in [0,360)$. Kun $x < 0$, $F(x) = 0$ ja, kun $x \geq 360$, $F(x) = 1$.

Määrätään tapahtuman

$$90 \leq X \leq 180$$

todennäköisyys.

Jos todennäköisyyden määräämiseen käytetään satunnaismuuttujan X tiheysfunktiota, saadaan integroimalla:

$$\begin{aligned} P(90 \leq X \leq 180) &= \int_{90}^{180} f(x)dx \\ &= \int_{90}^{180} \frac{1}{360} dx = \frac{1}{360} x \Big|_{90}^{180} = \frac{1}{360} 180 - \frac{1}{360} 90 = \frac{1}{4}. \end{aligned}$$

Jos todennäköisyyden määräämiseen käytetään satunnaismuuttujan X kertymäfunktiota saadaan:

$$P(90 \leq X \leq 180) = F(180) - F(90) = \frac{1}{3} 2 - \frac{1}{3} 1 = \frac{1}{3}. \bullet$$

Edellisen esimerkin lopussa esitetyt kaksi tapaa määrätä johonkin väliin liittyvä todennäköisyys jatkuvan satunnaismuuttujan tapauksessa voidaan esittää yleisessä muodossa seuraavalla tavalla:

$$P(a \le X \le b) = \int_a^b f(t)dt = F(b) - F(a).$$

Huomaa, että normaalijakauman (ja useimpien muiden) jakaumien taulukoissa on taulukoitu nimenomaan kertymäfunktion

$$F(x) = P(X \le x)$$

arvoja eri x:n arvoilla. Myös useimmat tilasto-ohjelmat tulostavat todennäköisyysjakaumille nimenomaan kertymäfunktion arvoja. Todettakoon lopuksi, että *normaalijakauman* kertymäfunktion funktionaalista muotoa ei tunneta, vaan sen arvot määrätään aina *numeerisesti*.

1.2.5 SATUNNAISMUUTTUJAN ODOTUSARVO JA VARIANSSI

Todennäköisyyslaskennassa tarkastellaan teoreettisia (matemaattisia) malleja havaintoaineiston tarkasteluun käytetyille apuvälineille: Todennäköisyys on todennäköisyyden frekvenssitulkinnan mukaan idealisoitu kuvaus tarkastelun kohteena olevaan satunnaisilmiöön liittyvän tapahtuman suhteellisen frekvenssin säännönmukaisesta pitkän ajan käyttäytymisestä. Todennäköisyysjakaumaa voidaan pitää suhteellisten frekvenssien jakauman idealisoituna muotona. Pistetodennäköisyysfunktio on idealisoitu muoto pylväsdiagrammista ja tiheysfunktio on idealisoitu muoto histogrammista. Kuten Kirjassa 1 on esitetty, havaintoaineiston kuvaamiseen käytetään usein *otostunnuslukuja* (kts. K1: 3). Tässä kappaleessa tarkastellaan havaintoaineiston kuvaamiseen käytettyjen tavanomaisten otostunnuslukujen teoreettisia vastineita. Niiden tehtävänä on kuvata todennäköisyysjakaumia samaan tapaan, kuin otostunnusluvut kuvaavat havaintoaineistojen jakaumia.

ODOTUSARVO

Havaintoarvojen keskimääräistä arvoa tai aineiston paikkaa kuvataan tavallisesti havaintoarvojen *aritmeettisella keskiarvolla*. Sen teoreettinen vastine on *odotusarvo*.

ESIMERKKI 1.

Vermontin osavaltio USA:ssa järjestää päivittäin arpajaiset, joiden säännöt ovat seuraavat: Arvat on numeroitu luvuilla 1,2,...,1000. Arpajaisissa on siten 1000 *erilaista* arpaa. Pelaaja valitsee yhden tai useamman arvan. Yhden arvan hinta on 1$. Pelaaja voi halutessaan valita useita samalla numerolla varustettuja arpoja. Osavaltio arpoo joka päivä uuden voittonumeron lukujen 1,2,...,1000 joukosta. Jokaista voittonumerolla varustettua arpaa kohden jaetaan 500 $:n

voitto. Jos pelaaja ostaa useita arpoja, paljonko hän voittaa *keskimäärin* yhtä ostamaansa arpaa kohden?

Arpanumeroita on siis 1000, joista voittonumeroita on 1. Pelaaja voittaa *varmasti*, jos hän ostaa 1000 arpaa, joissa on numerot 1,2,...,1000. Tällöin voitto on yhtä ostettua arpaa kohden

$$\frac{500\$}{1000} = 0.5\$.$$

Tämä tulos on *keskimääräinen voitto* yhtä ostettua arpaa kohden. Koska arvat maksavat 1$ kappaleelta, pelaaja itse asiassa kärsii 1/2$:n tappion yhtä ostettua arpaa kohden.

Määrätään saatu tulos hieman toisella tavalla, joka johtaa odotusarvon käsitteeseen.

Ostetuista 1000 arvasta siis 1 voittaa 500$ ja 999 voittaa 0$. Siten *kokonais-voitto* ostetuista arvoista on

$$500\$ \times 1 + 0\$ \times 999.$$

Yhtä arpaa kohden saatava voitto on siten

$$\frac{500\$ \times 1 + 0\$ \times 999}{1000} = 0.5\$.$$

Huomaa, että tämä laskutoimitus voidaan kirjoittaa myös seuraavaan muotoon:

$$500\$\frac{1}{1000} + 0\$\frac{999}{1000} = 0.5\$.$$

Siinä keskimääräinen voitto on esitetty voittojen 500$ ja 0$ *painotettuna keskiarvona*, kuin painoina ovat voittonumeroiden *suhteelliset osuudet* lukujen 1,2,...,1000 joukossa (kts. K1: kappale 3.3.2).

Tarkastellaan tilannetta, jossa pelaaja valitsee useita arpoja *satunnaisesti*.

Olkoon X satunnaismuuttuja, joka kuvaa voittoa yhtä arpaa kohden. X on diskreetti satunnaismuuttuja, jonka jakauma on annettu seuraavassa taulukossa:

Voitto ($)	0	500
Todennäköisyys	0.999	0.001

Todennäköisyyden frekvenssitulkinnan mukaan keskimäärin 1/1000 ostetuista arvoista voittaa 500 $ ja keskimäärin 999/1000 voittaa 0 $. Siten keskimäärinen voitto yhtä ostettua arpaa kohden on

$$500\$\frac{1}{1000} + 0\$\frac{999}{1000} = 0.5\$.$$

Tulos on tietysti sama kuin edellä, mutta nyt voittojen painotetussa keskiarvossa painot tulkitaan voittojen *todennäköisyyksiksi*. ●

Esimerkissä 1 mainittua painotettua keskiarvoa kutsutaan voittoa kuvaavan satunnaismuuttujan X *odotusarvoksi*, koska pelaaja voi *odottaa*, että hän saa *keskimäärin* 1/2 $:n voitto jokaista ostamaansa arpaa kohden.

Huomaa, että termi odostusarvo on hämäävä siinä mielessä, että *yhden* pelin tulos ei ole välttämättä odostusarvon suuruinen: Todennäköinen tulos yhden arvan ostamisesta on, että voittaa 0 $, koska todennäköisyys, että ei voita on 0.999. Suunnilleen 1 arpa 1000:sta saa 500 $:n voiton. Yksikään pelaaja ei voita voiton odotusarvon ilmoittamaa rahasummaa 1/2 $. Voiton odotusarvolle onkin syytä antaa *frekvenssitulkinta*: Jos ostaa useita arpoja, saa voittona keskimäärin 50 c arpaa kohden.

Edellinen tarkastelu johtaa diskreetin satunnaismuuttujan odotusarvon määriltelmään:

DISKREETIN SATUNNAISMUUTTUJAN ODOTUSARVO

Olkoon X *diskreetti* satunnaismuuttuja, joka saa arvot $x_1, x_2, \ldots, x_k$ todennäköisyyksillä $p_1, p_2, \ldots, p_k$. Tällöin satunnaismuuttujan X *odotusarvo* on vakio

$$\begin{aligned} \mathrm{E}(X) &= \mu_X \\ &= \sum_{i=1}^{k} x_i p_i \\ &= x_1 p_1 + x_2 p_2 + \cdots + x_k p_k. \end{aligned}$$

E tulee englannin sanasta *expectation* (odotusarvo) ja μ (myy) tulee englannin sanan *mean* (keskiarvo) ensimmäisen kirjaimen kreikankielisestä vastineesta.

Jos todennäköisyydet p_i on määrätty empiirisen todennäköisyyden määritelmän perusteella suhteellisina frekvensseinä f_i/n, satunnaismuuttujan X odotusarvo yhtyy muuttujanarvojen $x_1, x_2, \ldots, x_k$ *painotettuun keskiarvoon*, jossa painoina käytetään ko. arvojen frekvenssejä:

$$\begin{aligned} \mathrm{E}(X) &= \mu_X \\ &= \frac{1}{n} \sum_{i=1}^{k} x_i f_i. \end{aligned}$$

ESIMERKKI 2.

Heitettäessä harhatonta rahaa 4 kertaa klaavojen lukumäärän todennäköisyysjakaumaksi saadaan

Tulos x_i	0	1	2	3	4
Todennä-köisyys	0.0625	0.25	0.375	0.25	0.0625

Klaavojen lukumäärän odotusarvo on siten

$$E(X) = 0\times0.0625 + 1\times0.25 + 2\times0.375 + 3\times0.25 + 4\times0.0625 = 2.$$

Huomaa, että jakauma on tässä tapauksessa symmetrinen pisteen 2 suhteen ja odotusarvo yhtyy jakauman symmetriapisteeseen, joka on samalla jakauman painopiste. ●

ESIMERKKI 3.

Mikä on amerikkalaisperheen keskimääräinen koko? Amerikkalaisperheiden koot jakautuvat tilastojen mukaan seuraavalla tavalla:

Perheen koko	2	3	4	5	6	7
Osuus	0.413	0.236	0.211	0.090	0.032	0.018

Tähän perheiden koon havaittu jakauma voidaan tulkita satunnaisesti valitun perheen koon ilmaisevan diskreetin satunnaismuuttujan empiirisenä todennäköisyysjakaumana.

Satunnaisesti valitun perheen odotettavissa oleva koko X on

$$E(X)$$
$$= 2\times0.413 + 3\times0.236 + 4\times0.211 + 5\times0.090 + 6\times0.032 + 7\times0.018$$
$$= 3.146. \bullet$$

Jatkuvan jakauman tapauksessa odotusarvo määrätään integraalin avulla:

JATKUVAN SATUNNAISMUUTTUJAN ODOTUSARVO

Olkoon $f(x)$ *jatkuvan* satunnaismuuttujan X tiheysfunktio. Tällöin satunnaismuuttujan X *odotusarvo* on vakio

$$E(X) = \mu_X$$
$$= \int_{-\infty}^{+\infty} x f(x)dx.$$

Huomaa, että jatkuvan satunnaismuuttujan odotusarvon määritelmässä esiintyvä integraali on approksimatiivisesti satunnaismuuttujan saamien arvojen painotettu summa, jossa painoina ovat tiheysfunktion arvot. Tämä nähdään suoraan integraalin määritelmän perusteella. Jatkuvien ja diskreettien satunnaismuuttujien odotusarvot ovat siten saman matemaattisen idean muunnelmia.

ESIMERKKI 4.

Tarkastellaan jälleen luvun valitsemista satunnaisesti väliltä [0,1). Satunnaisesti valitun luvun X tiheysfunktio on

$$f(x)=\begin{cases}1, \text{ jos } 0\le x<1\\ 0, \text{ muulloin}\end{cases}$$

Siten satunnaismuuttujan X odotusarvo on

$$E(X)=\int_0^1 x\,dx=\frac{1}{2}x^2\Big|_0^1=0.5.$$

Huomaa, että jakauma on tässä tapauksessa symmetrinen pisteen 0.5 suhteen ja odotusarvo yhtyy jakauman symmetriapisteeseen, joka on samalla jakauman painopiste. ●

ODOTUSARVON OMINAISUUKSIA

ODOTUSARVO ON JAKAUMAN PAINOPISTE

Esimerkeissä 2 ja 4 todettiin, että odotusarvo yhtyi jakauman painopisteeseen. Tämä on totta yleisesti:

Odotusarvo sijoittuu jakauman painopisteeseen.

VAKION ODOTUSARVO

Odotusarvon painopistetulkinta tekee ymmärrettäväksi seuraavan säännön: Jos $X=a$ (vakio), niin

$$E(X)=a.$$

LINEAARIMUUNNOKSEN ODOTUSARVO

Odotusarvon painopistetulkinta tekee ymmärrettäväksi seuraavan säännön: Olkoot a ja b vakiota ja olkoon

$$Y=a+bX.$$

Satunnaismuuttujan X *lineaarimuunnoksen*[1] Y odotusarvo saadaan seuraavalla kaavalla:

$$E(Y)=a+bE(X).$$

Tämä kaava voidaan kirjoittaa myös seuraavaan muotoon:

$$\mu_{a+bX}=a+b\mu_X.$$

---

[1] Muunnosta $Y=a+bX$ kutsutaan *lineaariseksi*, koska muunnoksen määrittelevä yhtälö on *suoran* yhtälö ja matemaatiikassa suoran yhtälöä kutsutaan *lineaariseksi* eli suoraviivaiseksi.

ESIMERKKI 5.

Oletetaan, että eräästä miesopiskelijoiden ryhmästä satunnaisesti valitun opiskelijan pituuden odotusarvo on 180 cm. Miten saadaan heidän pituutensa odotusarvo tuumina?

Olkoon

X = pituus cm:nä,

Y = pituus tuumina.

Koska 1 tuuma = 2.54 cm,

$Y = bX$,

jossa b = 1/2.54 (tuuma/cm). Siten

$$\begin{aligned} E(Y) &= bE(X) \\ &= 180/2.54 \text{ tuumaa} \\ &= 70.87 \text{ tuumaa.} \bullet \end{aligned}$$

SUMMAN JA EROTUKSEN ODOTUSARVOT

Olkoot satunnaismuuttujien X ja Y odotusarvot $E(X) = \mu_X$ ja $E(Y) = \mu_Y$. Tällöin summan $X + Y$ odotusarvo on

$$E(X+Y) = E(X) + E(Y).$$

Tämä kaava voidaan kirjoittaa myös seuraavaan muotoon:

$$\mu_{X+Y} = \mu_X + \mu_Y.$$

Vastaavasti erotuksen $X - Y$ odotusarvo on

$$E(X-Y) = E(X) - E(Y)$$

Tämä kaava voidaan kirjoittaa myös seuraavaan muotoon:

$$\mu_{X-Y} = \mu_X - \mu_Y.$$

ESIMERKKI 6.

Matkapuhelin Oy myy valmistamiaan GSM-puhelimia tukkuliikkeille, mutta myös tehtaan myymälän kautta suoraan asiakkaille. Matkapuhelin Oy haluaa arvioida seuraavan vuoden voittoaan käyttäen apuna myyntiosaston johtajilta saatuja subjektiivisia todennäköisyysarvioita myynnin kappalemäärälle seuraavana vuonna.

Arviot myynnistä tukkuliikkeille voidaan esittää seuraavana taulukkona:

Myytävä määrä	1000	3000	5000	10,000
Toden-näköisyys	0.1	0.3	0.4	0.2

Arviot myynnistä tehtaan myymälän kautta voidaan esittää seuraavana taulukkona:

Myytävä määrä	300	500	750
Toden-näköisyys	0.4	0.5	0.1

Olkoon X satunnaismuuttuja, joka kuvaa tukkuliikkeille myytävien puhelinten lukumäärää ja Y satunnaismuuttuja, joka kuvaa tehtaan myymälän kautta myytävien puhelinten lukumäärää. Tällöin

$$\begin{aligned} E(X) &= 1000\times0.1 + 3000\times0.3 + 5000\times0.4 + 10{,}000\times0.2 \\ &= 5000, \\ E(Y) &= 300\times0.4 + 500\times0.5 + 750\times0.1 \\ &= 445. \end{aligned}$$

Jos voitto tukkuliikkeille myytävistä puhelimista on 1000 mk/puhelin ja tehtaan myymälästä myytävistä puhelimista 1500 mk/puhelin, niin voitto tukkuliikkeille myytävistä puhelimista seuraavan vuonna on $1000X$ mk ja tehtaan myymälästä myytävistä puhelimista $1500Y$ mk.

Siten odotettavissa oleva voitto tukkuliikkeille myydyistä puhelimista on

$$\begin{aligned} E(1000X) &= 1000E(X) \\ &= 1000\times5000 \text{ mk} \\ &= 5{,}000{,}000 \text{ mk} \end{aligned}$$

ja tehtaan myymälän kautta myydyistä puhelimista

$$\begin{aligned} E(1500Y) &= 1500E(Y) \\ &= 1500\times445 \text{ mk} \\ &= 667{,}500 \text{ mk}. \end{aligned}$$

Kokonaisvoitto seuraavana vuonna on satunnaismuuttuja

$$Z = 1000X + 1500Y.$$

Kokonaisvoiton Z odotusarvo on siten

$$\begin{aligned} E(Z) &= E(1000X) + E(1500Y) \\ &= 1000E(X) + 1500E(Y) \\ &= 5{,}000{,}000 \text{ mk} + 667{,}500 \text{ mk} \\ &= 5{,}667{,}500 \text{ mk}. \end{aligned}$$

Tämä on paras arvio seuraavan vuoden kokonaisvoitosta. ●

SUURTEN LUKUJEN LAKI

Odotusarvo on edellä määritelty satunnaismuuttujan keskimääräiseksi arvoksi. Se määrätään muuttujan arvojen todennäköisyysjakauman perusteella[1]. Olemme viitanneet esimerkin 1 lopussa toisenlaiseen odotusarvon tulkintaan: Oletetaan, että tarkasteltava satunnaisilmiö toistuu hyvin suuren määrän kertoja. Ilmiön toistuessa tarkkaillaan satunnaisilmiötä kuvaavan satunnaismuuttujan havaittujen arvojen aritmeettisen keskiarvon vaihtelua, kun keskiarvo määrätään ilmiön kaikista toistokerroista. Tällöin havaitaan tavallisesti, että keskiarvo vaihtelee satunnaismuuttujan jonkin kiinteän arvon ympärillä lähestyen sitä toistojen lukumäärän kasvaessa. Tämä kiinteä arvo on satunnaismuuttujan odotusarvo. Seuraava esimerkki kuvaa tällaista tilannetta.

ESIMERKKI 7.

Tarkastellaan 4:n rahan heittoa satunnaisilmiönä. Olkoon X satunnaismuuttuja, joka kuvaa kruunien lukumäärää yhdessä heitossa.

Kuten aikaisemmin on todettu, satunnaismuuttujaan X voidaan liittää todennäköisyydet binomikaavan avulla. Todennäköisyys saada k kruunaa 4:n rahan heitossa on

$$P(X = k) = \binom{4}{k}\frac{1}{2^k}, \quad k = 0,1,2,3,4.$$

Satunnaismuuttujan X odotusarvo on

$$E(X) = \sum_{k=0}^{4} k\,P(X = k) = 2.$$

Oletetaan, että heitämme rahoja toistuvasti ja otamme tarkkailun kohteeksi, miten kruunien lukumäärän keskiarvo kehittyy heittokertojen lisääntyessä. Tulokset 6:sta ensimmäisestä heittokerrasta voisivat olla esimerkiksi seuraavat:

---

[1] Diskreetin jakauman tapauksessa odotusarvo on satunnaismuuttujan saamien arvojen painotettu keskiarvo, kun painoina käytetään muuttujan arvojen todennäköisyyksiä. Myös jatkuvan jakauman tapauksessa odotusarvo voidaan tulkita painotetuksi keskiarvoksi satunnaismuuttujan saamista arvoista, kun painoina käytetään tiheysfunktion arvoja ja yhteenlaskun korvaa integrointi.

Heitto-kerta	Kruunien lkm	Kruunien keskimääräinen lkm per heitto-kerta
1	3	$\bar{x} = 3/1 = 3.00$
2	0	$\bar{x} = 3/2 = 1.50$
3	0	$\bar{x} = 3/3 = 1.00$
4	2	$\bar{x} = 5/4 = 1.25$
5	4	$\bar{x} = 9/5 = 1.80$
6	2	$\bar{x} = 11/6 = 1.83$

Kolmannen sarakkeen keskiarvot saadaan kruunien lukumäärien keskiarvoina siihen asti tehdyistä heitoista. Koska esimerkiksi 5. heittokerran jälkeen on saatu yhteensä 9 kruunaa, kruunia on saatu keskimäärin 9/5 = 1.8 kappaletta heittokertaa kohden.

Heittoja jatkettaessa kruunien lukumäärän keskiarvo heittokertaa kohden vaihtelee satunnaisvaihtelun vaikutuksesta heittokerrasta toiseen. Voidaan kuitenkin olettaa, että heittokertojen lisääntyessä suuret poikkeamat kruunien lukumäärän odotusarvosta 2 tulevat yhä harvinaisemmeksi. ●

Esimerkissä kuvattu käyttäytyminen on esimerkki tilastollisesta säännönmukaisuudesta, jota kutsutaan *suurten lukujen laiksi*. Suurten lukujen laki voidaan ilmaista seuraavassa muodossa:

SUURTEN LUKUJEN LAKI

Olkoot satunnaismuuttujan X riippumattomien[1] havaintoarvojen $X_1, X_2, \ldots, X_n$ odotusarvo $E(X_i) = \mu$ kaikille i.[2] Tällöin ko. havaintoarvojen aritmeettinen keskiarvo

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

lähestyy havaintojen lukumäärän n kasvaessa rajatta havaintoarvojen odotusarvoa μ siten, että todennäköisyys sille, että keskiarvo $\overline{X}$ poikkeaa odotusarvosta μ tulee yhä pienemmäksi ja häviää lopulta kokonaan.

Huomaa, että olemme vedonneet samantapaiseen tilastolliseen säännönmukaisuuteen *empiiristä todennäköisyyttä* määriteltäessä: Satunnaisilmiön toistuessa tapahtumien suhteelliset frekvenssit lähestyvät tapahtumien todennäköisyyksiä. Satunnaisilmiöön liittyvien tapahtumien todennäköisyydet muodostavat muuttujan todennäköisyysjakauman. Nyt olemme todenneet, että myös satunnaismuuttujan havaittujen arvojen aritmeettinen keskiarvo käyttäytymisessä voidaan havaita samantapaista säännönmukaisuutta.

Suurten lukujen laki on esimerkiksi vakuutusyhtiöiden ja pelikasinoiden toiminnan perusta. Yksittäisen pelaajan voitot tai tappiot ovat yksittäisessä pelissä ennustamattomia. Juuri se tekee pelaamisesta mielenkiintoista. Sen sijaan, kun suuret ihmisjoukot pelaavat, pelaajien keskimääräiset voitot ovat lähellä voiton odotusarvoa. Juuri tähän säännönmukaisuuteen pelikasinot perustavat oman voittonsa. Myös vakuutustoiminta on eräänlaista peliä: Henkivakuutuksen ottaja lyö vakuutusyhtiön kanssa vetoa omasta kuolemastaan. Jos hän voittaa (eli hän siis kuolee), vakuutusyhtiö joutuu maksamaan vakuutussumman omaisille.

Suurten lukujen lain toimintaan liittyy paljon *virheellisiä käsityksiä*. Useimmat niistä perustuvat seuraavaan väärinkäsitykseen: Suurten lukujen lain mukaan keskiarvo lähestyy havaintojen odotusarvoa (tai oikeammin, keskiarvon suuret poikkeamat odotusarvosta tulevat yhä epätodennäköisemmiksi), kun havaintojen lukumäärä kasvaa rajatta. Tämä merkitsee sitä, että suurten lukujen laki eräässä mielessä *korjaa* keskiarvon suuret poikkeamat odotusarvosta havaintojen lukumäärän kasvaessa. Tähän liittyy virheellinen kuvitelma, että korjauksen pitäisi näkyä jo pienillä havaintomäärillä. Tätä kuvitelmaa on leikillisesti kuvattu sanomalla, että ihmiset uskovat yleisesti *pienten*

---

[1] Riippumattomuudella tarkoitetaan tässä sitä, että se mitä tapahtumia on sattunut satunnaisilmiön aikaisemmilla tapahtumiskerroilla ei vaikuta tapahtumiin satunnaisilmiön myöhemmillä tapahtumiskerroilla.

[2] Huomaa, että olemme merkinneet havaintoarvoja isoilla kirjaimilla, mikä merkitsee sitä, että pidämme niitä satunnaismuuttujien arvoina. Tulemme jatkossa usein tulkitsemaan havainnot satunnaismuuttujiksi. Tämä on perusteltua, koska voimme aina olettaa, että havaintoarvot on kerätty käyttäen satunnaisotantaa tai satunnaistettua koetta. Siten ne vaihtelevat satunnaisesti otoksesta tai kokeesta toiseen. Palaamme tämän tulkinnan merkitykseen otantajakaumia koskevassa luvussa.

lukujen lakiin. Pienten lukujen laki ei kuitenkaan päde. Seuraavat esimerkit liittyvät virheelliseen pienten lukujen lakiin:

- Saman tuloksen esiintymistä peräkkäisissä satunnaisilmiön toistoissa kuvitellaan paljon harvinaisemmaksi tapahtumaksi, kuin se onkaan todennäköisyyden lakien mukaan.
- Koripallovalmentajat uskovat pelaajien "kuumaan käteen". Tällä tarkoitetaan seuraavaa: Jos pelaaja on tehnyt useita koreja peräkkäin, uskotaan, että hänen seuraavakin heittonsa uppoaa koriin todennäköisyydellä, joka on suurempi kuin siinä tapauksessa, että heittojen tulokset ovat toisistaan riippumattomia.
- Ruletin pelaajat yrittävät pitää kirjaa eri numeroiden esiintymisestä. Jos esimerkiksi numero 13 ei ole esiintynyt pitkään aikaan, he kuvittelevat, että sen on pakko tulla hyvin pian.

Kaikkiin näihin esimerkkeihin liittyy virheellinen mielikuva suurten lukujen lain toiminnasta. Mielikuviin yhdistyy vielä usein virheellinen kuvitelma siitä, että riippumattomissa satunnaisilmiöissä olisi sisäänrakennettu *muisti*, joka pitäisi huolta siitä, että keskiarvon havaitut poikkeamat odotusarvosta tasoittuvat lyhyellä aikavälillä. Suurten lukujen laki puhuu kyllä vaihtelun tasoittumisesta, mutta vasta *hyvin* pitkällä aikavälillä[1].

[1] Itse asiassa suurten lukujen laissa puhutaan tasoittumisesta, joka tapahtuu *äärettömän* monessa satunnaisilmiön toistokerrassa.

SATUNNAISMUUTTUJAN VARIANSSI

Odotusarvo kuvaa satunnaismuuttujan todennnäköisyysjakauman paikkaa samaan tapaan, kuin aritmeettinen keskiarvo kuvaa muuttujan havaittujen arvojen jakauman paikkaa. Jakauman paikan lisäksi kiinnostuksen kohteena on usein jakauman keskittyneisyys tai, mikä on sama asia, jakauman hajaantuneisuus paikkaa kuvaavan tunnusluvun ympärille. Jakauman hajaantuneisuutta voidaan kuvata jakauman *varianssin* tai paremminkin sen *standardipoikkeaman* avulla.

Olkoon satunnaismuuttujan X odotusarvo μ. Satunnaismuuttujan X arvojen poikkeama niiden odotusarvosta on satunnaismuuttuja

$$X - \mu.$$

Koska kaikkiin poikkeamiin suhtaudutaan symmetrisesti, ts. negatiiviset ja positiiviset poikkeamat ovat jakauman hajaantuneisuutta mitattaessa yhtä merkityksellisiä, tapana on tarkastella odotettavissa olevaa poikkeaman neliötä eli poikkeaman neliön odotusarvoa, jota kutsutaan satunnaismuuttujan X *varianssiksi*.

VARIANSSI

Satunnaismuuttujan X *varianssi* on satunnaismuuttujan omasta odotusarvostaan määrätyn poikkeaman neliön odotusarvo:

$$D^2(X) = \mathrm{var}(X) = \sigma_X^2 = E(X - \mu)^2.$$

Määritelmä voidaan esittää diskreetin satunnaismuuttujan tapauksessa seuraavassa muodossa:

DISKREETIN SATUNNAISMUUTTUJAN VARIANSSI

Olkoon X *diskreetti* satunnaismuuttuja, joka saa arvot $x_1, x_2, \ldots, x_k$ todennäköisyyksillä $p_1, p_2, \ldots, p_k$. Tällöin X:n *varianssi* on

$$D^2(X) = \sum_{i=1}^{k} (x_i - \mu)^2 p_i$$
$$= (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots + (x_k - \mu)^2 p_k.$$

Jatkuvan satunnaismuuttujan tapauksessa joudutaan varianssin määritelmässä käyttämään integraalia:

JATKUVAN SATUNNAISMUUTTUJAN VARIANSSI

Olkoon X *jatkuva* satunnaismuuttuja, jonka tiheysfunktio on $f(x)$. Tällöin X:n *varianssi* on

$$D^2(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx.$$

Satunnaismuuttujan jakauman hajaantuneisuuden kuvaajana käytetään odotusarvon rinnalla tavallisesti varianssin neliöjuurta, jota kutsutaan *standardipoikkeamaksi* tai *teoreettiseksi hajonnaksi*. Sana teoreettinen jätetään usein pois ja puhutaan vain pelkästään satunnaismuuttujan hajonnasta. Tämä saattaa aiheuttaa sekaannusta, koska myös havaintoarvojen jakauman hajaantuneisuuden erästä mittaria on tapana kutsua hajonnaksi. Useimmissa tapauksissa asiayhteydestä käy kuitenkin ilmi tarkoitetaanko satunnaismuuttujan jakauman hajaantuneisuutta kuvaavaa teoreettista hajontaa vai havaintoarvojen jakauman hajaantuneisuutta kuvaavaa otoshajontaa.

STANDARDIPOIKKEAMA

Satunnaismuuttujan X *standardipoikkeama* eli (teoreettinen) *hajonta* on varianssin $D^2(X)$ neliöjuuri

$$D(X) = \sigma_X = \sqrt{E(X-\mu)^2}.$$

Se, että odotusarvon rinnalla on parempi käyttää hajontaa eikä varianssia johtuu siitä, että standardipoikkeamalla ja odotusarvolla on sama *laatu*, kun taas varianssin laatu on neliöllinen. Oletetaan esimerkiksi, että tarkastelun kohteena on pituutta kuvaava satunnaismuuttuja. Satunnaismuuttujan odotusarvon ja hajonnan laatu voidaan ilmaista esimerkiksi cm:nä. Sen sijaan satunnaismuuttujan varianssin laatu on tällöin cm^2.

ESIMERKKI 8.

Edellä tarkasteltiin esimerkkiä, jonka aiheena oli Matkapuhelin Oy:n seuraavan vuoden myynnin ennakointi. Ennakoitua myyntiä pidettiin satunnaismuuttujana, jonka odotusarvosta oltiin kiinnostuneita. Tukkuliikkeille myytävien matkapuhelinten ennakoidun myynnin jakauma oli seuraava:

Myytävä määrä	1000	3000	5000	10,000
Toden-näköisyys	0.1	0.3	0.4	0.2

Sekä odotusarvo että varianssi saadaan järjestämällä laskut seuraavan taulukon muotoon:

x_i	p_i	$x_i p_i$	$(x_i - \mu)^2 p_i$
1000	0.1	100	$(1000 - 5000)^2 \times 0.1 = 1{,}600{,}000$
3000	0.3	900	$(3000 - 5000)^2 \times 0.3 = 1{,}200{,}000$
5000	0.4	2000	$(5000 - 5000)^2 \times 0.4 = 0$
10,000	0.2	2000	$(10{,}000 - 5000)^2 \times 0.2 = 5{,}000{,}000$
Summa	1.0	$E(X) = 5000$	$D^2(X) = 7{,}800{,}000$

Siten X:n standardipoikkeama on

$$D(X) = \sqrt{7{,}800{,}000} \approx 2792.8.$$

Standardipoikkeama kuvaa myyntiarvion vaihtelua. ●

ESIMERKKI 9.

Tarkastellaan jälleen satunnaisluvun X valintaa väliltä $[0,1)$. Edellä on jo todettu, että $E(X) = 0.5$. Varianssiksi saadaan

$$\begin{aligned} D^2(X) &= \int_0^1 (x - 0.5)^2 \cdot 1 dx \\ &= \int_0^1 (x^2 - x + 0.25) dx \\ &= \left(\tfrac{1}{3}x^3 - \tfrac{1}{2}x^2 + 0.25x\right)\Big|_0^1 \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{4} \\ &= \frac{1}{12}. \end{aligned}$$

●

VARIANSSIN OMINAISUUKSIA

VAKION VARIANSSI

Jos $X = a$ (vakio), niin

$$D^2(X) = 0.$$

Kääntäen, jos $D^2(X) = 0$, niin X on vakio (todennäköisyydellä 1).

LINEAARIMUUNNOKSEN VARIANSSI

Olkoot a ja b vakioita ja olkoon

$$Y = a + bX.$$

Satunnaismuuttujan X *lineaarimuunnoksen* Y odotusarvo saadaan kaavalla:

$$D^2(Y) = b^2D^2(X).$$

On ymmärrettävää, että vakio a ei vaikuta varianssiin, koska a:n lisääminen vastaa *siirtoa*, joka ei vaikuta satunnaismuuttujan jakauman keskittyneisyyteen tai hajaantuneisuuteen sen oman odotusarvon ympärille. Sen sijaan siirto vaikuttaa kyllä odotusarvoon, koska jakauman painopiste siirtyy siirrossa. Kertominen b:llä vastaa mittakaavan muutosta, joka vaikuttaa sekä odotusarvoon että varianssiin.

SUMMAN JA EROTUKSEN VARIANSSI

Tarkastellaan vielä kahden satunnaismuuttujan summan varianssia. Edellä todettiin, että summan odotusarvo on odotusarvojen summa. Sen sijaan summan varianssi ei ole välttämättä varianssien summa. Tämän perustelemiseksi tarkastellaan seuraavaa esimerkkiä:

ESIMERKKI 10.

Olkoon X se *osuus*, jonka kotitalous kuluttaa ja Y se *osuus*, jonka kotitalous säästää käytettävissä olevista tuloistaan. Vaikka X ja Y vaihtelevat vuodesta toiseen, summa $X + Y$ on aina 100% tuloista ja ei vaihtele ollenkaan. Tällöin siis

$$D^2(X + Y) = 0,$$

vaikka ilmeisesti sekä $D^2(X) > 0$ että $D^2(Y) > 0$. ●

Muuttujien X ja Y *riippuvuus* estää varianssien yhteenlaskun. Sen sijaan, jos satunnaismuuttujat ovat *riippumattomia*, varianssit saa laskea yhteen.

SATUNNAISMUUTTUJIEN RIIPPUMATTOMUUS

Satunnaismuuttujat X ja Y ovat *riippumattomia*, jos jokainen tapahtuma, joka liittyy vain X:ään on riippumaton jokaisesta tapahtumasta, joka liittyy vain Y:hyn.

Todennäköisyysmallit olettavat usein, että niihin liittyvät satunnaismuuttujat ovat riippumattomia. Riippumattomuus on oletus, jonka järkevyyttä on aina tarkasteltava huolellisesti. On syytä huomata, että monissa tilastollisissa menetelmissä juuri riippuvuudet ovat tutkimuksen kohteena. Tällaisia ovat esimerkiksi *regressioanalyysi* monine muunnelmineen, *aikasarja-analyysi* ja useat *monimuuttujamenetelmät*.

Jos X ja Y ovat *riippumattomia* satunnaismuuttujia, niin seuraavat yhtälöt pätevät:

$$D^2(X+Y) = D^2(X) + D^2(Y),$$
$$D^2(X-Y) = D^2(X) + D^2(Y).$$

ESIMERKKI 11.

Vermontin osavaltion arpajaisissa arvasta maksetun voiton odotusarvo ja teoreettinen hajonta voidaan laskea seuraavan taulukon avulla:

x_i	p_i	$x_i p_i$	$(x_i - \mu)^2 p_i$
0	0.999	0	$(0-0.5)^2 \times 0.999 = 0.24975$
500	0.001	0.5	$(500-0.5)^2 \times 0.001 = 249.50025$
Summa	1.000	$E(X) = 0.5$	$D^2(X) = 249.75000$

Siten voiton standardipoikkeama on

$$D(X) = \sqrt{249.75} \approx 15.80\$.$$

Koska arvan hinta on 1\$, *pelaajan voitto* on satunnaismuuttuja

$$V = X - 1.$$

Satunnaismuuttujan V odotusarvo on

$$E(V) = E(X) - 1 = -0.5\$.$$

Siten on odotettavissa, että arvan ostaja menettää 0.5\$ jokaista ostamaansa arpaa kohden.

Arpajaisia pidetään päivittäin. Jos ostaa arpoja peräkkäisinä päivinä, voidaan voittoja X ja Y pitää riippumattomina. Siten arvasta maksetun voiton odotusarvo on

$$E(X+Y) = E(X) + E(Y) = 0.50\$ + 0.50\$ = 1.00\$$$

ja varianssi on

$$D^2(X+Y) = D^2(X) + D^2(Y) = 249.75 + 249.75 = 499.50.$$

Siten summan $X+Y$ standardipoikkeama on

$$D(X+Y) = \sqrt{499.5} \approx 22.35\$.$$

Huomaa, että tämä ei ole muuttujien X ja Y standardipoikkeamien summa. ●

Edellisestä esimerkistä nähdään seuraava seikka:

Vaikka riippumattomien muuttujien varianssit voidaan laskea yhteen, niin standardipoikkeamia ei voi!

ESIMERKKI 12.

Eräs college USA:ssa käyttää ns. SAT-testiä[1] pääsykokeen osana. Kokemus on osoittanut, että SAT-testin kahden osion pistemäärien jakaumilla on seuraavat odotusarvot ja varianssit:

SAT-testin matematiikan osion pistemäärä X:

$$\mu_X = 625 \qquad \sigma_X = 90$$

SAT-testin verbaalisen osion pistemäärä Y:

$$\mu_Y = 590 \qquad \sigma_Y = 100$$

Mitkä ovat pistemäärien summan $X + Y$ odotusarvo ja standardipoikkema?

Summan odotusarvo on

$$\mu_{X+Y} = \mu_X + \mu_Y = 625 + 590 = 1215.$$

Sen sijaan summan $X + Y$ standardipoikkeamaa ei voi laskea annettujen tietojen perusteella, koska matematiikan osiosta ja verbaalisesta osiosta saatuja pisteitä ei voi pitää riippumattomina muuttujina. ●

[1] SAT-testi tarkoittaa ns. scholastic aptitude testiä, joka mittaa kykyä menestyä opinnoissa (kts. K1: 2.5.2).

1.3 TODENNÄKÖISYYSJAKAUMIA

1.3.1 DISKREETTEJÄ TODENNÄKÖISYYSJAKAUMIA

BINOMIJAKAUMA

Edellä on johdettu binomitodennäköisyyden kaava. Kertaamme tässä määritelmän olennaiset kohdat.

Tarkastellaan jonkin satunnaisilmiön tapahtumaa A. Olkoon tapahtuman A todennäköisyys

$$P(A) = p$$

ja tapahtuman A komplementtitapahtuman A^C (*ei-A*) todennäköisyys

$$P(A^C) = q.$$

Tällöin

$$P(A) + P(A^C) = p + q = 1.$$

Toistetaan mainittua satunnaisilmiötä (tai annetaan sen toistua) samoissa olosuhteissa toisistaan riippumatta n kertaa ja tarkastellaan tapahtuman A sattumista tässä toistokoesarjassa. Otetaan tehtäväksi määrätä todennäköisyys sille, että A tapahtuu tällöin k kertaa.

Sellaisen tulosjonon todennäköisyys, jossa A tapahtuu k kertaa ja *ei-A* tapahtuu $(n-k)$ kertaa, on riippumattomien tapahtumien tulosäännön mukaan

$$p^k q^{n-k}.$$

Erilaisten tulosjonojen lukumäärä, joissa tapahtuma A sattuu täsmälleen k kertaa, on

$$C(n,k) = \binom{n}{k}.$$

Tämä seuraa siitä, että binomikerroin $C(n,k)$ ratkaisee seuraavan kombinatorisen ongelman: Oletetaan, että n:n alkion joukossa on k kappaletta alkioita p ja $(n-k)$ kappaletta alkioita q. Kuinka monella tavalla alkiot p ja q voidaan järjestää jonoon?

Olkoon

$$A_k^n$$

se yhdistetty tapahtuma, jossa A sattuu n:n toistokokeen sarjassa k kertaa ja olkoon p_k^n tämän yhdistetyn tapahtuman todennäköisyys.

Koska erilaiset tulosjonot ovat tapahtumina toisensa poissulkevia, saadaan tapahtuman A_k^n todennäköisyys laskemalla yhteen sellaisten yksittäisten tulosjonojen todennäköisyydet, joissa A on sattunut k kertaa. Edellä esitetyn mukaan jokaisen tällaisen tulosjonon todennäköisyys on $p^k q^{n-k}$ ja erilaisten tulosjonojen (ts. tulojen,

joissa p ja q ovat eri järjestyksissä) lukumäärä on $C(n,k)$. Siten tapahtuman A_k^n todennäköisyydeksi saadaan

$$P(A_k^n) = p_k^n = \binom{n}{k} p^k q^{n-k}.$$

Edellä esitettyyn nojaten ns. *binomijakauma* voidaan määritellä seuraavalla tavalla:

BINOMIJAKAUMA

Olkoon A jokin tapahtuma ja X satunnaismuuttuja, joka kuvaa kuinka usein A sattuu n-kertaisessa riippumattomien toistokokeiden sarjassa. Tällöin satunnaismuuttujan X jakaumaan liityvät todenäköisyydet saadaan kaavasta

$$P(X = k) = p_k^n = \binom{n}{k} p^k q^{n-k}, \quad k = 0,1,2,\ldots,k.$$

Sanomme, että X on *jakautunut binomijakauman mukaan* (on *binomijakautunut*) parametrein n ja p ja käytämme lyhyttä merkintää

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

Binomijakautuneen satunnaismuuttujan X odotusarvo, varianssi ja standardipoikkeama ovat

$$E(X) = \mu_X = np,$$

$$D^2(X) = \sigma_X^2 = npq,$$

$$D(X) = \sigma_X = \sqrt{npq}.$$

Huomaa, että jos määritellään *standardoitu muuttuja*

$$Z = \frac{X - np}{\sqrt{npq}},$$

niin

$$E(Z) = 0,$$

$$D^2(Z) = 1.$$

Tämä seuraa siitä, että Z on satunnaismuuttujan X lineaarimuunnos:

$$Z = \frac{1}{\sqrt{npq}} X - \frac{np}{\sqrt{npq}}.$$

Oletetaan, että tapahtuman A todennäköisyys on määrätty *empiirisesti* eli

$$P(A) = p = f/n,$$

jossa f on tapahtuman A frekvenssi n-kertaisessa toistokokeessa. Tällöin

$$E(X) = \mu_X = f,$$

$$D^2(X) = \sigma_X^2 = \frac{f(n-f)}{n}.$$

ESIMERKKI 1.

Karnakin tehtaat valmistavat TV-vastaanottimia. Tehtaan tuotannosta joka kahdeskymmenes vastaanotin osoittautuu normaaleissa oloissa vialliseksi. Valmistettujen vastaanottimien laadun tarkkailemiseksi tehtaan laadunvalvontaosasto valitsee tuotantolinjalta satunnaisesti 15 vastaanotinta testattaviksi päivittäin. Jos testattavien 15 vastaanottimen joukosta löydetään enemmän kuin yksi viallinen, tuotantoprosessi keskeytetään. Seuraavassa tätä testausasetelmaa tarkastellaan satunnaisilmiönä.

Olkoon

A = "Testattava vastaanotin on viallinen".

Tällöin tehtaan toimiessa normaalisti

$$P(A) = p = 0.05.$$

Oletetaan, että vastaanottimet voidaan poimia tuotantolinjalta toisistaan riippumattomasti. Tällöin viallisten vastaanottimien lukumäärä testattavaksi valittujen 15 vastaanottimen joukossa on satunnaismuuttuja X, joka voidaan olettaa binomijakautuneeksi parametrein 15 ja 0.05, ts. $X \sim \text{Bin}(15,0.05)$.

Odotettavissa oleva viallisten vastaanottimien määrä on

$$E(X) = np = 15 \times 0.05 = 0.75.$$

Tästä voidaan päätellä, että 100 päivän aikana löydetään noin 75 viallista vastaanotinta.

Määrätään nyt todennäköisyys, että tuotantoprosessi joudutaan keskeyttämään. Kyseessä on tapahtuman

$$X > 1$$

todennäköisyys. Soveltamalla ensin komplementtitodennäköisyyden kaavaa ja sitten toisensa poissulkevien tapahtumien yhteenlaskusääntöä saadaan

$$\begin{aligned} P(X > 1) &= 1 - P(X \le 1) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{15}{0} p^0 (1-p)^{15} - \binom{15}{1} p^1 (1-p)^{14} \\ &\approx 0.171. \bullet \end{aligned}$$

HYPERGEOMETRINEN JAKAUMA

Miten määrätään jonkin järven kalakannan suuruus? Kannan suuruuden arviointi ei onnistu suoraan laskemalla, koska on vaikeata saada selville milloin kaikki kalat on laskettu. Kannan määrääminen voidaan tehdä käyttämällä ns. *merkintä-takaisinpyynti-menetelmää.*

Menetelmä käytetään seuraavalla tavalla: Järvestä pyydystetään joukko kaloja, jotka merkitään esimerkiksi kiinnittämällä pyrstöevään jokin merkki. Olkoon merkittyjen kalojen lukumäärä K. Merkityt kalat päästetään takaisin järveen. Jonkin ajan kuluttua järvestä pyydystetään uudelleen kaloja. Olkoon niiden lukumäärä n ja olkoon merkittyjen kalojen lukumäärä pyydystettyjen joukossa k. Oletetaan, että kalojen pyydystäminen voidaan järjestää kummallakin kerralla siten, että voidaan olettaa, että pyydystetyt kalat tulevat poimituksi satunnaisesti järven kaikkien kalojen joukosta. Oletetaan lisäksi, että merkityt kalat ovat sekoittuneet merkitsemättömien kalojen joukkoon ennen takaisinpyyntiä. Tällöin voidaan olettaa, että *merkittyjen kalojen suhteelliset osuudet koko järvessä ja takaisinpyynnissä saatujen kalojen joukossa ovat yhtäsuuret.* Jos siis järvessä on N kalaa, niin

$$\frac{K}{N} = \frac{k}{n}.$$

Tästä yhtälöstä järven kalakannan kooksi N saadaan

$$N = \frac{nK}{k}.$$ [1]

Tarkastellaan merkintä-takaisinpyynti-menetelmää satunnaisilmiönä. Merkittyjen kalojen lukumäärä toisella pyydystyskerralla on satunnaismuuttuja. Mikä on tämän satunnaismuuttujan jakauma?

Oletetaan siis, että otosavaruudessa S on N alkiota. Oletetaan, että S jakautuu kahteen toisensa poissulkevaan osaan (merkityt ja merkitsemättömät kalat) A ja A^C, joissa on K ja $N-K$ alkiota, ts.

$$\mathrm{n}(A) = K,$$
$$\mathrm{n}(A^C) = N-K.$$

Poimitaan S:stä satunnaisesti otos (pyydystetyt kalat takaisinpyynnissä) eli osajoukko B, jonka koko on n, ts.

$$\mathrm{n}(B) = n.$$

Olkoon X satunnaismuuttuja, joka kuvaa joukkoon A kuuluvien B:n alkioiden lukumäärää (merkittyjen kalojen lukumäärä takaisinpyynnissä). Mikä on X:n jakauma? Haluamme siis määrätä todennäköisyydet

$$\mathrm{P}(X = k),$$

eri k:n arvoille. Huomaa, että

---

[1] Koska nK/k ei ole välttämättä kokonaisluku, voidaan sopia, että N:ksi valitaan pienin kokonaisluku, joka on suurempi (tai yhtäsuuri) kuin luku nK/k.

$$n(B \cap A) = k$$

$$n(B \cap A^C) = n - k.$$

Kysytty todennäköisyys voidaan määrätä klassisen todennäköisyyden määritelmää käyttäen. Lasketaan ensin, kuinka monella tavalla voidaan poimia n:n alkion otos perusjoukosta, jonka koko on N. Tämä voidaan tehdä

$$\binom{N}{n}$$

eri tavalla. Lasketaan sitten, kuinka monella tavalla voidaan poimia n:n alkion otos niin, että joukosta A (merkityt kalat) saadaan k alkiota ja joukosta A^C (merkitsemättömät kalat) saadaan $n - k$ alkiota. Joukosta A voidaan poimia k alkiota

$$\binom{K}{k}$$

eri tavalla. Joukosta A^C voidaan poimia $n - k$ alkiota

$$\binom{N-K}{n-k}$$

eri tavalla. Koska nämä poiminnat voidaan tehdä toisistaan riippumattomasti, voidaan soveltaa kertolaskusääntöä riippumattomille tapahtumille. Sen mukaan n:n alkion otos voidaan poimia niin, että joukosta A saadaan k alkiota ja joukosta A^C saadaan $n - k$ alkiota

$$\binom{K}{k}\binom{N-K}{n-k}$$

eri tavalla.

Soveltamalla klassisen todennäköisyyden määritelmää saadaan

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

Edellä esitetyn nojalla voimme määritellä ns. *hypergeometrisen jakauman* seuraavalla tavalla:

HYPERGEOMETRINEN JAKAUMA

Olkoon perusjoukon S koko $n(S) = N$. Oletetaan, että S voidaan jakaa kahteen osajoukkoon A ja A^C, joiden koot ovat $n(A) = K$ ja $n(A^C) = N - K$. Poimitaan perusjoukosta S satunnaisesti osajoukko B, jonka koko on $n(B) = n$. Olkoon satunnaismuuttuja X osajoukkoon B tulleiden A:n alkioiden lukumäärä. Tällöin satunnaismuuttujan X jakaumaan liittyvät todennäköisyydet saadaan kaavasta

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0,1,2,\ldots,n.$$

Sanomme, että X noudattaa *hypergeometrista jakaumaa*, jonka parametrit ovat N, K ja n. Merkitsemme tätä lyhyesti

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

Hypergeometrisesti jakautuneen satunnaismuuttujan X odostusarvo, varianssi ja hajonta ovat

$$E(X) = np,$$

$$D^2(X) = npq\frac{N-n}{N-1},$$

$$D(X) = \sqrt{npq\frac{N-n}{N-1}},$$

jossa

$$p = \frac{K}{N},$$

$$q = 1 - p.$$

Huomaa, että hypergeometrista jakaumaa ja binomijakautumaa noudattavien satunnaismuuttujien odotusarvot ovat samat ja varianssit ovat samat lukuunottamatta hypergeometrisen jakauman kaavassa esiintyvää tekijää

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

jota sanotaan *äärellisen perusjoukon korjaustekijäksi*. Aina pätee

$$\frac{N-n}{N-1} \leq 1.$$

Korjaustekijän arvo on sitä lähempänä arvoa 1, mitä pienempi on *otantasuhde* n/N. Otantasuhde on pieni, jos otos (joukko B) muodostaa vain pienen osan perusjoukosta, niinkuin tavallisesti on asian laita. Hypergeometrinen jakauma ja binomijakauma eivät

eroa toisistaan muutenkaan kovin paljon. Ero on merkityksellinen vain silloin, kun otantasuhde n/N on suuri.

Binomijakaumalla ja hypergeometrisella jakaumalla on ehkä tärkeimmät sovelluksensa tilastotieteessä *otannassa*. Palaamme jakaumien käyttöön otannassa sekä jakaumien välisiin eroihin ja yhtäläisyyksiin seuraavassa kappaleessa.

ESIMERKKI 2.

Erään seuran pikkujoulujuhliin on painettu 100 arpaa, joista 20 voittaa. Oletetaan, että ostat 5 arpaa. Olkoon

X = Voittoarpojen lukumäärä ostettujen 5 arvan joukossa.

Tällöin

$X \sim \text{Hyperg}(100,20,5)$.

Siten

$$P(X=k)=\frac{\binom{20}{k}\binom{80}{5-k}}{\binom{100}{5}}.$$

Siten esimerkiksi

$P(X=0) = 0.32$,

$P(X=5) = 0.0002$,

Koska 100 arvan joukossa on 20 voittoarpaa, todennäköisyys, että satunnaisesti valittu arpa voittaa on

$p = 20/100 = 0.2$.

Siten odotettavissa oleva voittoarpojen lukumäärä 5:n satunnaisesti valitun arvan joukossa on

$E(X) = np = 5 \times 0.2 = 1$. ●

ESIMERKKI 3.

Pieni urheiluautoja valmistava tehdas ostaa autoihinsa moottorit suurelta autotehtaalta. Tehtaaseen on tullut 40 moottorin erä, josta valitaan 8 moottoria erityisen huolellisen testauksen kohteeksi. Jos yksikin testattavista moottoreista ei täytä erittäin vaativia testikriteereitä, koko erä palautetaan. Oletetaan, että erässä on todellisuudessa 2 moottoria, jotka eivät täytä testikriteereitä. Mikä on todennäköisyys, että erä tulee hyväksytyksi?

Olkoon

X = Testikriteereitä täyttämättömien moottoreiden lukumäärä testattavien 8 moottorin joukossa.

Tällöin

$X \sim \text{Hyperg}(40,2,8)$

ja siis

$$P(X=0)=\frac{\binom{2}{0}\binom{38}{8}}{\binom{40}{8}}\approx 0.64,$$

On siis melko todennäköistä, että tehtaassa hyväksytään sellainen 40 moottorin erä, jossa on mukana 2 moottoria, jotka eivät täytä testikriteereitä. Ainoa tapa pienentää tätä todennäköisyyttä on ottaa enemmän moottoreita testaukseen.

Koska erän 40 moottorin joukossa oli 2 moottoria, jotka eivät täytä testikriteereitä, todennäköisyys, että satunnaisesti valittu moottori ei täytä testikriteereitä on

$$p = 2/40 = 0.05.$$

Siten odotettavissa oleva viallisten moottoreiden lukumäärä testattavaksi valittujen 8 moottorin joukossa on

$$E(X) = np = 8\times 0.05 = 0.4.$$

Tämä merkitsee odotusarvon frekvenssitulkinnan mukaan seuraavaa:

Oletetaan, että tehtaaseen tulee 10 sellaista 40 moottorin erää, joissa on 2 viallista moottoria ja jokaisesta erästä valitaan 8 moottoria testattaviksi. Tällöin on odotettavissa, että testauksessa löydetään

$$10\times 0.4 = 4$$

viallista moottoria. ●

1.3.2 JATKUVIA JAKAUMIA

NORMAALIJAKAUMA

Tilastotieteen tärkein jatkuva jakauma on *normaalijakauma*.

NORMAALIJAKAUMA

Satunnaismuuttuja X on jakautunut *normaalijakauman* mukaan parametrein μ *ja* σ^2, jos X:llä on tiheysfunktio

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

Määritelmässä exp tarkoittaa exponenttifunktiota $\exp(z) = e^z$. Jos satunnaismuuttuja X on normaalijakautunut, merkitään lyhyesti

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

Jos $\mu = 0$ ja $\sigma = 1$, sanotaan normaalijakaumaa *standardoiduksi*.

Parametreilla μ ja σ^2 on seuraavat tulkinnat: μ on X:n *odotusarvo* eli

$$E(X) = \mu$$

ja σ^2 on X:n *varianssi* eli

$$D^2(X) = \sigma^2.$$

Siten X:n *standardipoikkeama* eli *teoreettinen hajonta* on

$$D(X) = \sigma.$$

Olkoon Z standardoitua normaalijakaumaa noudattava satunnaismuuttuja. Muotoa

$$Z \leq z$$

olevien tapahtumien todennäköisyyksiä on taulukoitu taulukkokokoelmiin. Useat tilasto-ohjelmat tulostavat muotoa

$$X \leq x$$

olevien tapahtumien todennäköisyyksiä.

Normaalijakaumaa $N(\mu,\sigma^2)$ noudattava satunnaismuuttuja X voidaan aina *standardoida*: Satunnaismuuttuja

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

noudattaa standardoitua normaalijakaumaa $N(0,1)$. Siten

$$E(Z) = 0,$$

$D^2(Z) = 1.$

NORMAALIJAKAUMAN OMINAISUUKSIA

1. Jos X on $N(\mu,\sigma^2)$ ja $Y = a + bX$, jossa a ja b ovat vakiota, niin

$$Y \sim N(a + b\mu, b^2\sigma^2).$$

Huomaa, että satunnaismuuttujan X lineaarimuunnoksen Y odotusarvoa ja varianssia koskevat tulokset on mainittu jo aikaisemmin. Tässä on uutena piirteenä se, että X:n normaalisuudesta seuraa lineaarimuunnoksen Y normaalisuus.

2. Jos $X \sim N(\mu_X,\sigma_X^2)$ ja $Y \sim N(\mu_Y,\sigma_Y^2)$ ja lisäksi X ja Y ovat *riippumattomia*, niin

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Huomaa, että satunnaismuuttujien X ja Y summan odotusarvoa ja varianssia koskevat tulokset on mainittu jo aikaisemmin. Tässä on uutena piirteenä se, että X:n ja Y:n normaalisuudesta seuraa summan $X + Y$ normaalisuus.

Ominaisuus 2 voidaan yleistää ilmeisellä tavalla koskemaan useamman kuin kahden normaalijaukaumaa noudattavan riippumattoman satunnaismuuttujan summan jakaumaa. Erityisesti, jos ko. satunnaismuuttujat noudattavat *samaa* normaalijakaumaa, saadaan seuraava tulos:

3. Jos $X_1, X_2, \ldots, X_n$ ovat riippumattomia ja noudattavat samaa normaalijakaumaa $N(\mu,\sigma^2)$, niin summa

$$X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2).$$

Ominaisuudesta 3 ja ominaisuudesta 1 seuraa tärkeä aritmeettista keskiarvoa koskeva tulos:

4. Olkoot $X_1, X_2, \ldots, X_n$ ovat riippumattomia samaa normaalijakaumaa $N(\mu,\sigma^2)$ noudattavia havaintoja ja olkoon

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

havaintojen aritmeettinen keskiarvo. Tällöin

$$\overline{X} \sim N(\mu,\sigma^2/n).$$

ESIMERKKI 1.

Naulatehdas pakkaa myyntiä varten 50 naulan rasioita. Oletetaan, että naulat painavat keskimäärin 20g/kpl ja niiden painon standardipoikkeama on 0.1g. Mikä on rasiallisen painon odotusarvo ja standardipoikkeama?

Jos tehtaan valmistamien naulojen painoa voidaan pitää normaalijakautuneena, niin naulan i paino

$$X_i \sim N(20g, (0.1g)^2).$$

Oletetaan lisäksi, että naulojen painot määräytyvät valmistusprosessissa toisistaan riippumattomasti. Rasiallisen paino on satunnaismuuttuja

$$Y = X_1 + X_2 + \ldots + X_{50}.$$

Ominaisuuden 3 mukaan laatikollisen paino

$$Y \sim N(\mu_Y, \sigma_Y^2),$$

jossa

$$\mu_Y = E(Y) = 50 \times 20g = 1kg,$$

$$\sigma_Y^2 = D^2(Y) = 50 \times (0.1g)^2 = 0.5g^2 \approx (0.707g)^2.$$

Siten

$$D(Y) \approx 0.707g. \bullet$$

KESKEINEN RAJA-ARVOLAUSE

Normaalijakauman keskeinen asema tilastotieteessä johtuu siitä, että monien satunnaismuuttujien on havaittu noudattavan normaalijakaumaa empiirisesti. Tämä on totta varsinkin silloin, kun satunnaismuuttujaa voidaan pitää usean satunnaisen tekijän summana ja yksikään summan tekijä ei ole summassa hallitsevassa asemassa.

Esimerkiksi monet ihmisten, eläinten ja kasvien ominaisuuksiin liittyviä muuttujia voidaan pitää satunnaismuuttujina, jotka ovat usean satunnaisen tekijän summia. Näitä tekijöitä ovat eri perintötekijöiden sekä ympäristön vaikutukset. Myös havaintovirheitä voidaan pitää tällaisina satunnaismuuttujina, koska havaintovirheet koostuvat tavallisesti monesta pienemmästä virhetekijästä. Havaintovirheiden tapaan käyttäytyy tavallisesti se vaihtelu, jota esiintyy erilaisten koneiden valmistamien tuotteiden ominaisuuksissa.

Kaikki tässä kuvatut empiiriset havainnot voidaan perustella ns. *keskeisellä raja-arvolauseella*, joka on tilastotieteen teorian ja käytännön kulmakiviä. Keskeisestä raja-arvolauseesta on useita erilaisia muotoja. Niistä useimmat ovat seuraavan muodon muunnelmia:

KESKEINEN RAJA-ARVOLAUSE

Olkoot satunnaismuuttujasta X tehtyjen riippumattomien havaintojen $X_1, X_2, \ldots, X_n$ odotusarvo $E(X_i) = \mu$ ja varianssi $D^2(X_i) = \sigma^2$ kaikille i. Tällöin havaintoarvojen aritmeettisen keskiarvo

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

on suurille havaintojen lukumäärille n *approksimatiivisesti normaalinen* $N(\mu, \sigma^2/n)$.

Merkitsemme keskeisessä raja-arvolauseessa esitettyä approksimatiivista jakaumatulosta tavallisesti seuraavasti:

$$\overline{X} \sim_a N(\mu, \sigma^2/n).$$

Keskeinen raja-arvolause ja suurten lukujen laki ovat esimerkkejä *asymptoottisista* tuloksista (suurten lukujen laki: kts. kappale 1.2.5). Asymptoottisessa tilastotieteen teoriassa tutkitaan miten satunnaismuuttujat käyttäytyvät, kun havaintojen lukumäärän n annetaan kasvaa rajatta. Tässä yhteydessä puhutaan usein myös *suurten otosten* teoriasta. Suurten otosten teoriaa tarvitaan usein sellaisissa tilanteissa, joissa satunnaismuuttujien käyttäytymistä ei hallita äärellisellä otoskoolla. Vaikka ääretöntä määrää havaintoja ei käytännössä voida kerätä, on suurten otosten teorialla kuitenkin suuri käytännöllinen merkitys.

Keskeisen raja-arvolauseen tärkeimpiä sovelluksia on seuraava, binomijakautunutta satunnaismuuttujaa koskeva tulos:

KESKEINEN RAJA-ARVOLAUSE JA BINOMIJAKAUMA

Oletetaan, että satunnaismuuttuja X on Bin(n,p). Oletetaan, että n on "tarpeeksi suuri" ja p on "tarpeeksi lähellä" arvoa 1/2 sekä $q = 1-p$. Tällöin X on *approksimatiivisesti normaalinen* N(np, npq).

Approksimatiivisuudella tarkoitetaan tässä seuraavaa: Olkoon Y satunnaismuuttuja, jonka jakauma on N(np, npq). Tällöin X:n ja Y:n jakaumat muistuttavat toisiaan siinä mielessä, että

$$P(a - 1/2 \le Y \le b + 1/2) \approx P(a \le X \le b).$$

Tämä merkitsee sitä, että binomitodennäköisyydet voidaan määrätä normaalijakaumasta, jos n on kyllin suuri ja p ei ole kovin kaukana arvosta 1/2.

Huomaa, että satunnaismuuttujat X ja Y muistuttavat aina toisiaan siinä mielessä, että niillä sama odotusarvo ja varianssi:

$$E(X) = E(Y) = np,$$
$$D^2(X) = D^2(Y) = npq.$$

Tapahtumassa

$$a - 1/2 \le Y \le b + 1/2$$

on käytössä ns. *jatkuvuuskorjaus*, jota tarvitaan, koska X on *diskreetti* ja Y on *jatkuva*. Korjauksen tarve tulee parhaiten esiin, jos tarkastellaan tapahtuman $Y = k$ todennäköisyyttä. Koska Y on jatkuva,

$$P(Y = k) = 0.$$

Sen sijaan

$$P(X = k) = \binom{n}{k} p^k q^{n-k} > 0.$$

Jatkuvuuskorjaus liittää tapahtumaan $Y = k$ välin

$$k - 1/2,\ k + 1/2,$$

jota vastaava todennäköisyys

$$P(k - 1/2 \le Y \le k + 1/2) > 0.$$

Keskeisen raja-arvolauseen mukaan

$$P(X = k) \approx P(k - 1/2 \leq Y \leq k + 1/2),$$

jos n on "tarpeeksi suuri" ja p on "tarpeeksi lähellä" arvoa 1/2. Jatkuvuuskorjaus jätetään usein tekemättä, ellei tarkoituksena ole laskea normaaliapproksimaation avulla yhden pisteen todennäköisyyttä.

Edellä on todettu, että binomijakauman normaaliapproksimaatio on käytännössä riittävän hyvä, kun n on "tarpeeksi suuri" ja p on "tarpeeksi lähellä" arvoa 1/2. Koska n on "tarpeeksi suuri" ja koska p on "tarpeeksi lähellä" arvoa 1/2. Käytännössä on osoittautunut riittäväksi, että seuraavat ehdot pätevät:

$$np > 5,$$

$$nq > 5,$$

jossa siis $q = 1 - p$.

On syytä ymmärtää, että kyseessä on *karkea nyrkkisääntö*. Seuraava taulukko antaa jonkinlaisen mielikuvan siitä, miten tarvittava havaintojen lukumäärä riippuu p:stä:

p	$n >$
0.05	100
0.1	50
0.2	25
0.5	10
0.7	25
0.9	50
0.95	100

Taulukosta nähdään, että mitä lähempänä p on ääriarvojaan 0 ja 1, sitä suurempi on tarvittava n.

ESIMERKKI 2.

Tarkastellaan otantatutkimusta, jonka päämääränä on selvittää Suomen presidentinvaalien toisen kierroksen ehdokkaiden A:n ja B:n kannatus äänioikeutettujen joukossa. Oletetaan, että ehdokkaiden kannatus jakaantuu todellisuudessa tasan. Mikä on todennäköisyys, että 1000 äänioikeutetun satunnaisotoksessa yli 550 kannattaa ehdokasta A?

Oletetaan, että satunnaismuuttuja X kuvaa ehdokas A:n kannattajien lukumäärää otoksessa. Satunnaismuuttuja X noudattaa hypergeometrista jakaumaa. Koska otantasuhde on kuitenkin hyvin pieni (perusjoukossa on n. 4

miljoonaa äänioikeutettua), voidaan hypergeometrista jakaumaa approksimoida binomijakaumalla. Siten

$$X \sim_a \text{Bin}(n, p),$$

jossa

$$n = 1000, p = 0.5, q = 1 - p.$$

Näiden tietojen perusteella

$$\text{E}(X) = np = 1000 \times 0.5 = 500,$$

$$\text{D}^2(X) = npq = 1000 \times 0.5 \times 0.5 = 250,$$

$$\text{D}(X) \approx 15.81.$$

Koska myös $nq = 500$, voidaan soveltaa normaaliapproksimaatiota. Siten approksimatiivisesti

$$\begin{aligned} \text{P}(X > 550) &= \text{P}(Z > \frac{550 - 500}{15.81}) \\ &= \text{P}(Z > 3.13) \\ &= 0.0009, \end{aligned}$$

jossa siis Z noudattaa standardoitua normaalijakaumaa N(0,1).

Siis on vain noin 9 mahdollisuutta 10,000:sta havaita 550 ehdokas A:n kannattajaa 1000 äänioikeutetun otoksessa, jos $p = 0.5$. Tästä voidaan tehdä seuraava johtopäätös: Jos otoksessa havaitaan 550 ehdokas A:n kannattajaa, kyseessä on erittäin harvinainen tapahtuma, jos p on todellakin 0.5. Siten on järkevää asettaa oletus $p = 0.5$ kyseenalaiseksi, jos 1000 äänioikeutetun otoksessa havaitaan 550 A:n kannattajaa. Tällaiset tarkastelut johtavat *tilastollisen testin* käsitteeseen. ●

2. OTOSJAKAUMAT

2.1 SATUNNAISOTANTA

2.1.1 YKSINKERTAINEN SATUNNAISOTANTA

Tutkimuksen kohteeksi tulevien alkioiden poimimista perusjoukosta sanotaan *otannaksi*.[1] Sitä perusjoukon osajoukkoa, joka poimitaan sanotaan *otokseksi*. Tilastollisen tutkimuksen päämääränä on tehdä otoksen perusteella perusjoukkoa koskevia johtopäätöksiä. Johtopäätösten tekeminen on mahdollista, jos otos *edustaa* perusjoukkoa riittävän hyvin. Edustavuuden takaa se, että otokseen tulevat perusjoukon alkiot poimitaan käyttäen *satunnaisotantaa*, ts. käyttäen apuna arvontaa, jossa sovelletaan todennäköisyyden lakeja.

Teoreettisesti tärkein satunnaisotannan menetelmistä on *yksinkertainen satunnaisotanta*. Tavallisesti tutkimusaineistoa kerättäessä sovelletaan kuitenkin muita satunnaisotannan menetelmiä. Sellaisia ovat *ositettu otanta*, *moniasteinen otanta* ja *ryväsotanta*. Paljon sovellettu *systemaattinen otanta* ei ole itse asiassa satunnaisotantaa, mutta se toimii kuten satunnaisotanta, jos ne perusjoukon alkioiden ominaisuudet, joita otoksesta on tarkoitus tutkia, jakautuvat satunnaisesti siihen tietorekisteriin, johon systemaattinen poiminta kohdistetaan. Yksinkertainen satunnaisotanta muodostaa kuitenkin perustan erilaisten otannan menetelmien ymmärtämiselle.[2]

YKSINKERTAINEN SATUNNAISOTANTA

Oletetaan, että perusjoukosta halutaan poimia otos, jonka koko on n. Jos jokaisella perusjoukon osajoukolla, jonka koko on n, on sama todennäköisyys tulla valituksi otokseksi, poimintamenettelyä kutsutaan *yksinkertaiseksi satunnaisotannaksi*.

Yksinkertaisessa satunnaisotannassa jokaisella perusjoukon samankokoisella osajoukolla on sama todennäköisyys tulla poimituksi otokseksi.

Olkoon X tutkimuksen kohteena olevaa perusjoukon ominaisuutta kuvaava satunnaismuuttuja. Oletamme siis, että tutkittava ominaisuus jakautuu satunnaisesti perusjoukkoon. Kun perusjoukosta poimitaan otos, tutkimus kohdistetaan satunnaismuuttujan X käyttäytymiseen otoksessa. Olkoon poimitun otoksen koko n. Merkitään tutkimuksen kohteena olevaa ominaisuutta kuvaavaa satunnaismuuttujaa otokseen poimituille perusjoukon alkioille (ts. otos- eli havaintoyksiköille) seuraavalla tavalla:

---

[1] Myös perusjoukon alkioiden poimimista johonkin *kokeeseen* voidaan pitää otannan piiriin kuuluvana toimintana.

[2] Otannan menetelmiä selostetaan 1. kirjassa 1; kts. K1: 2.4.

$X_1, X_2, \ldots, X_n$.

Tällä merkinnällä halutaan korostaa sitä, että muuttujan X saamat arvot otoksessa määräytyvät satunnaisesti. Satunnaisotannassa sovellettu arvonta saa aikaan sen, että sattuma määrää, mitkä perusjoukon alkiot tulevat poimituiksi otokseen. Tästä seuraa se, että sattuma määrää, mitkä muuttujan X arvot havaitaan. Se, että muuttujan X havaintokohtaisia arvoja pidetään satunnaismuuttujan arvoina, on *tilastollisen päättelyn* kannalta keskeinen oletus. Oletuksesta seuraa se, että havaintoarvojen jakaumaa ja jakaumaa kuvaavia tunnusluvukuja voidaan pitää satunnaismuuttujina.

2.1.2 OTANTA TAKAISINPANOLLA JA ILMAN TAKAISINPANOA

Kun perusjoukon alkioita poimitaan otokseen, alkiot voidaan jokaisen poiminnan jälkeen palauttaa perusjoukkoon tai jättää palauttamatta. Jos poimittu alkio palautetaan perusjoukkoon, jolloin sillä on mahdollisuus tulla poimituksi uudelleen, kutsutaan poimintamenettelyä *otannaksi takaisinpanolla*. Jos poimittua alkiota ei palauteta perusjoukkoon, kutsutaan poimintamenettelyä *otannaksi ilman takaisinpanoa*.

ESIMERKKI 1.

Olkoon uurnassa valkoinen, musta ja punainen kuula. Poimitaan uurnasta satunnaisesti kuula, joka osoittautuu punaiseksi. Poimitaan uurnasta tämän jälkeen myös toinen kuula. Mikä on todennäköisyys, että se on musta?

Vastaus riippuu siitä, palautetaanko 1. kerralla nostettu kuula ennen 2. kuulan nostamista takaisin uurnaan vai ei:

(1) 1. kerralla nostettu kuula palautetaan uurnaan.

Tällöin

P(2. kuula on musta|1. kuula oli punainen) = 1/3.

(2) 1. kerralla nostettua kuulaa ei palauteta uurnaan.

Tällöin

P(2. kuula on musta|1. kuula oli punainen) = 1/2.

Tämä merkitsee sitä, että se palautetaanko 1. kuula uurnaan vai ei, vaikuttaa 2. kuulan poimintatodennäköisyyteen. ●

Esimerkissä käsitelty tilanne voidaan yleistää seuraavalla tavalla: Verrataan toisiinsa satunnaisesti valitun perusjoukon alkion poimintatodennäköisyyttä otannassa takaisinpanolla ja ilman takaisinpanoa. Olkoon perusjoukon koko N ja olkoon otoskoko n.

Jaetaan tarkastelu kahteen osaan:

(1) Oletetaan, että tarkasteltava perusjoukon alkio ei tule poimituksi otokseen.

Voimme muodostaa ko. alkion poimintatodennäköisyyksistä seuraavan taulukon:

	Poimintatodennäköisyys	
Poiminnan nro	Otanta takaisinpanolla	Otanta ilman takaisinpanoa
1	$1/N$	$1/N$
2	$1/N$	$1/(N-1)$
3	$1/N$	$1/(N-2)$
.....		
n	$1/N$	$1/(N-n+1)$

Tarkasteltavan alkion poimintatodennäköisyys ei siis muutu otannassa takaisinpanolla. Tämä johtuu siitä, että perusjoukko, josta poiminta tehdään pysyy poiminnan aikana samana. Sen sijaan tarkasteltavan perusjoukon alkion poimintatodennäköisyys kasvaa poiminnan edistyessä otannassa ilman takaisinpanoa. Tämä johtuu siitä, että perusjoukosta, josta poiminta tehdään poistuu yksi alkio jokaisessa poiminnan vaiheessa.

(2) Oletetaan, että tarkasteltava perusjoukon alkio tulee poimituksi vaiheessa k otokseen.

Voimme muodostaa ko. alkion poimintatodennäköisyyksistä seuraavan taulukon:

	Poimintatodennäköisyys	
Poiminnan nro	Otanta takaisinpanolla	Otanta ilman takaisinpanoa
1	$1/N$	$1/N$
2	$1/N$	$1/(N-1)$
.....		
k	$1/N$	$1/(N-k+1)$
$k+1$	$1/N$	0
.....		
n	$1/N$	0

Tarkasteltavan alkion poimintatodennäköisyys muuttuu otannassa ilman takaisinpanoa välittömästi 0:ksi sen jälkeen, kun ko. alkio tulee poimituksi otokseen. Sen sijaan otannassa takaisinpanolla alkion poimintatodennäköisyys pysyy muuttumattomana, vaikka se olisi jo poimittu otokseen.

Edellä esitetystä seuraa, että poiminnat ovat *riippumattomia* tapahtumia, jos poiminta tapahtuu takaisinpanolla. Sen sijaan poiminnat *eivät ole riippumattomia* (poimintatodennäköisyydet riippuvat siitä, mitä alkioita on poimittu aikaisemmin), jos poiminta tapahtuu ilman takaisinpanoa.

Poimintatodennäköisyyksien riippuvuus aikaisemmista poiminnoista voidaan ilmaista otannassa ilman takaisinpanoa ehdollisten todennäköisyyksien avulla seuraavalla tavalla:

$$P(\text{Alkio tulee poimituksi j. kerralla}|\text{Alkiota ei ole vielä poimittu}) = \frac{1}{N-j+1}.$$

$$P(\text{Alkio tulee poimituksi j. kerralla}|\text{Alkio on jo poimittu}) = 0.$$

Jos poiminta tehdään ilman takaisinpanoa, poimintatodennäköisyydet riippuvat siis siitä, mitä poiminnan aikana on aikaisemmin tapahtunut. Tällöin poiminnat eivät siis ole tapahtumina riippumattomia. Tämä riippuvuus tekee niistä tilastollisen päättelyn kaavoista, jotka liittyvät poimintaan ilman takaisinpanoa monimutkaisempia kuin ne kaavat, jotka liittyvät otantaan takaisinpanolla.

Jos perusjoukko on ääretön, yksittäisen havaintoyksikön poimintatodennäköisyys on aina nolla. Tällöin sillä, että poimittua alkiota ei palauteta poiminnan jälkeen perusjoukkoon, ei ole vaikutusta poimintatodennäköisyyksiin: Todennäköisyys poimia perusjoukosta sama alkio uudelleen on edelleen nolla. Tällöin otannalla ilman takaisinpanoa ja takaisinpanolla ei ole eroa. Sama pätee käytännössä myös silloin, kun perusjoukon koko N on hyvin suuri otoskokoon n verrattuna.

2.1.3 TODENNÄKÖISYYSMALLI YKSINKERTAISELLE SATUNNAISOTANNALLE

Tarkastellaan yksinkertaisen satunnaisotoksen poimimista äärellisestä perusjoukosta S, jonka koko on N. Olkoon tutkimuksen kohteena oleva ominaisuus $\mathscr{P}$ K:lla perusjoukon alkiolla. Olkoon tämän ominaisuuden määrämä perusjoukon S osajoukko A. Poimitaan perusjoukosta S otos, jonka koko on n ja määritellään satunnaismuuttuja

X = Niiden otokseen tulevien alkioiden lukumäärä, joilla on ominaisuus $\mathscr{P}$.

Tarkastellaan satunnaismuuttujan X jakaumaa, kun otos poimitaan palauttaen ja otos poimitaan palauttamatta.

(1) Otanta palauttaen:

1. Otokseen tulevien alkioiden poimiminen muodostaa n-kertaisen toistokokeen, jossa toistona on alkion poimiminen perusjoukosta.
2. Toistot ovat riippumattomia.
3. Jokaisella alkiolla on sama todennäköisyys K/N tulla poimituksi otokseen jokaisessa poiminnan vaiheessa.
4. Satunnaismuuttuja X noudattaa binomijakaumaa:

$$X \sim \text{Bin}(n,p),$$

jossa $p = K/N$.

(2) Otanta palauttamatta:

1. Otokseen tulevien alkioiden poimiminen muodostaa n-kertaisen toistokokeen, jossa toistona on alkion poimiminen perusjoukosta.
2. Toistot eivät ole riippumattomia.
3. Poimintatodennäköisyydet muuttuvat poiminnan aikana.
4. Satunnaismuuttuja X noudattaa hypergeometrista jakaumaa:

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

Ero näiden kahden otoksen poiminnan menetelmän välillä häviää, jos perusjoukko S on ääretön. Tämä johtuu siitä, että tällöin

2.' Toistot ovat riippumattomia tehdäänpä poiminta takaisinpanolla tai ilman takaisinpanoa.

3.' Poimintatodennäköisyys p ei muutu toistosta toiseen.

Otannan takaisinpanolla ja ilman takaisinpanoa välisen eron katoaminen näkyy myös seuravaasta *asymptoottisesta* tuloksesta: Jos perusjoukon S koko N kasvaa rajatta, hypergeometrinen jakauma lähestyy binomijakaumaa. Tämä voidaan ilmaista seuraavalla tavalla: Olkoon X_B binomijakaumaa $\text{Bin}(n,p)$ noudattava satunnaismuuttuja ja X_H hypergeometrista jakaumaa $\text{Hyperg}(N,K,n)$ noudattava satunnaismuuttuja. Tällöin

$$\text{P}(X_H = k) \approx \text{P}(X_B = k) = \binom{n}{k} p^k q^{n-k},$$

kun N on "suuri".

Tämä tulos näkyy myös siinä, että otoskokoon n nähden suurelle perusjoukon S koolle N hypergeometrisen jakauman varianssin ja standardipoikkeaman kaavoissa esiintyvä äärellisen perusjoukon korjaustekijä

$$\frac{N-n}{N-1} \approx 1,$$

jolloin

$$\text{D}^2(X_H) \approx \text{D}^2(X_B)$$

ja

$$D(X_H) \approx D(X_B).$$

Yhteenvetona voidaan sanoa seuraavaa: Kuvatkoon satunnaismuuttuja X niiden otokseen poimittujen perusjoukon S alkioiden lukumäärää, joilla on osajoukon A määräävä ominaisuus $\mathscr{P}$. Jos otos poimitaan äärettömästä tai otokseen nähden "hyvin suuresta" perusjoukosta, satunnaismuuttujaa X *voidaan käsitellä* binomijakautuneena tehtiinpä otanta takaisinpanolla tai ilman takaisinpanoa. Sen sijaan satunnaismuuttujaa X *pitää käsitellä* hypergeometrista jakaumaa noudattavana, jos otos poimitaan ilman takaisinpanoa äärellisestä perusjoukosta ja otantasuhde n/N on suuri. Satunnaismuuttuja X *on* binomijakautunut vain, jos otos poimitaan takaisinpanolla.

Koska otantasuhde n/N on niin pieni, että binomijakaumaan perustuvat tulokset ovat käyttökelpoisia? *Nyrkkisäännöksi* kelpaa

$$\frac{n}{N} < \frac{1}{10}.$$

Jos perusjoukko on äärellinen, se sovelletaanko otantaa palauttaen tai palauttamatta vaikuttaa lähes kaikkiin tilastollisen päättelyn kaavoihin. Siten esimerkiksi odotusarvon luottamusvälin tai odotusarvoa koskevan tavanomaisen t-testisuureen kaavat riippuvat siitä poimitaanko otos palauttaen tai palauttamatta. Tästä erosta ei kuitenkaan aina välitetä, vaan *käytännössä toimitaan usein ikään kuin otanta olisi tehty takaisinpanolla tai, että perusjoukko olisi ääretön*, vaikka kumpikaan ei olisi totta. Tämä johtuu siitä, että tilastollisen päättelyn kaavat, jotka liittyvät otantaan takaisinpanolla ovat yksinkertaisempia. Tästä käytännöstä johtuva virhe onkin yleensä pieni, ellei otantasuhde ole kovin iso. Muutamaa huomautusta lukuunottamatta jatkossa toimitaan ikään kuin otos olisi poimittu takaisinpanolla tai, että perusjoukko olisi ääretön. On syytä kuitenkin muistaa, että tarpeen tullen saattaa olla syytä käyttää äärellisestä perusjoukosta ilman takaisinpanoa tehtyyn otantaan perustuvia kaavoja.[1]

ESIMERKKI 1.

Erään tuotteen valmistaja väittää, että tuotteista korkeintaan 1% on viallisia. Tukkuliike poimii sille toimitetusta 1000 kappaleen erästä 25 kappaleen otoksen, jota testattaessa havaitaan 2 viallista tuotetta. Mikä on tämän tapahtuman todennäköisyys, jos valmistajan väite on tosi?

Perusjoukon S koko: $n(S) = N = 1000$

Tarkasteltava ominaisuus: Tuotteen viallisuus

Todennäköisyys, että satunnaisesti valittu tuote on viallinen:

$p = 0.01$

Viallisten tuotteiden joukon A koko: $n(A) = K = Np = 10$

Otoskoko: $n = 25$

Otantasuhde: $n/N = 0.025$

---

[1] On syytä lisäksi tietää, että myös jonkin muun otantamenetelmän kuin yksinkertaisen satunnaisotannan soveltaminen johtaa toisenlaisiin tilastollisen päättelyn kaavoihin.

Olkoon satunnaismuuttuja X viallisten lukumäärä otoksessa. Satunnaismuuttuja

$$X \sim \text{Hyperg}(1000,10,25).$$

Siten

$$\text{P}(X=2) = \frac{\binom{10}{2}\binom{990}{23}}{\binom{1000}{25}} \approx 0.0224.$$

Koska $n/N < 0.1$, binomijakauma-approksimaation pitäisi toimia hyvin. Oletetaan, että $X_B \sim \text{Bin}(25,0.01)$. Tällöin

$$\text{P}(X_B = 2) = \binom{10}{2} 0.01^2\ 0.99^{23} \approx 0.0238.$$

Ainakin tässä tapauksessa binomijakauma-approksimaatio toimii aivan riittävän hyvin.

Tarkastellaan vielä saadusta tuloksesta tehtävää johtopäätöstä. Odotettavissa oleva viallisten lukumäärä 25 kappaleen otoksessa on

$$\text{E}(X) = np = 0.25.$$

Otoksessa havaittiin 2 viallista ja tämän tapahtuman todennäköisyydeksi saatiin noin 0.02. Tämä merkitsee sitä, että 25 kappaleen otoksia poimittaessa havaitaan siis vain 2 kertaa 100:sta, että joukossa on 2 viallista, *jos oletus viallisten tuotteiden osuudesta* $p = 1/100$ *pätee*. Koska havaittu viallisten lukumäärä on siis melko harvinainen tapahtuma, mainittu oletus on ehkä syytä asettaa epäilyksen alaiseksi.

Esimerkin tarkastelut johtavat tilastollisen testaukseen: Tällöin oletus $p = 1/100$ asetetaan otoksesta saatujen tietojen koetteeseen. ●

2.1.4 OTOSJAKAUMAT

Oletetaan, että tutkimuksen kohteena olevasta perusjoukon S ominaisuutta kuvaavasta muuttujasta X on kerätty yksinkertaisella satunnaisotannalla havainnot

$$X_1, X_2, \ldots, X_n.$$

Koska havaintoarvot vaihtelevat satunnaisesti otoksesta toiseen, niitä voidaan pitää satunnaismuuttujien arvoina, jotka määräytyvät käsillä olevasta otoksesta. Tästä tulkinnasta seuraa se, että myös kaikki havainnoista lasketut *otostunnusluvut* kuten esimerkiksi aritmeettinen keskiarvo ja keskihajonta vaihtelevat satunnaisesti otoksesta toiseen, ts. ne ovat satunnaismuuttujia, joiden arvot määräytyvät käsillä olevasta otoksesta. Satunnaismuuttujina otostunnusluvuilla on todennäköisyysjakauma. Tätä jakaumaa kutsutaan ko. tunnusluvun *otosjakaumaksi*. Tunnusluvun otosjakauma muodostaa perustan tunnuslukua vastaavan perusjoukon parametrin luottamusvälin konstruoinnille ja parametria koskeville testeille.

Seuraavassa tarkastellaan lukumäärätietojen ja keskiarvon otosjakaumia.

2.2 LUKUMÄÄRÄTIETOJEN JA SUHTEELLISTEN OSUUKSIEN OTOSJAKAUMAT

2.2.1 JOHDANTO

Tässä kappaleessa tarkastellaan frekvenssin eli lukumäärän ja suhteellisen frekvenssin eli suhteellisen lukumäärän otosjakaumia.

2.2.2 FREKVENSSIN OTOSJAKAUMA

Olkoon S perusjoukko, jonka alkioista osalla on ominaisuus $\mathscr{P}$ ja osalla ei ole. Oletetaan, että ominaisuus $\mathscr{P}$ on mitattavissa ja, että se jakautuu satunnaisesti perusjoukon alkioiden keskuuteen. Olkoon A niiden perusjoukon S alkioiden joukko, joilla on ominaisuus $\mathscr{P}$. Siten

$$A = \{s \in S | s\text{:llä on ominaisuus } \mathscr{P}\},$$

$$A^C = \{s \in S | s\text{:llä ei ole ominaisuutta } \mathscr{P}\}.$$

Olkoon

$$P(A) = p$$

todennäköisyys, että satunnaisesti valittu perusjoukon S alkiolla s on ominaisuus $\mathscr{P}$. Merkitään

$$P(A^C) = 1 - P(A) = q.$$

Oletetaan, että perusjoukko S on äärellinen ja sen koko on N. Oletetaan lisäksi, että perusjoukon S osajoukon A koko on K. Siten

$$n(S) = N,$$

$$n(A) = K.$$

Soveltamalla klassisen todennäköisyyden määritelmää, saadaan

$$P(A) = N/K.$$

Poimitaan perusjoukosta S yksinkertainen satunnaisotos, jonka koko on n. Olkoon X satunnaismuuttuja, joka kuvaa niiden niiden havaintosyksiköiden *frekvenssiä* eli lukumäärää otoksessa, joilla on ominaisuus $\mathscr{P}$.[1]

Edellä on todettu, että lukumäärän X jakauma riippuu siitä tehdäänkö otanta takaisinpanolla vai ilman takaisinpanoa. Jos otanta tehdään takaisinpanolla, niin riippumatta siitä tehdäänkö otanta palauttaen tai palauttamatta

$$X \sim \text{Bin}(n,p).$$

---

[1] Merkitsemällä frekvenssiä isolla kirjaimella X on haluttu korostaa sitä, että frekvenssi on satunnaismuuttuja, jonka arvo määräytyy otokseen tulleista havainnoista. Aikaisemmin, kun tätä ei välitetty korostaa, frekvenssiä merkittiin kirjaimella f.

Jos otanta tehdään ilman takaisinpanoa,

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

Kummassakin tapauksessa

$$\text{E}(X) = np.$$

Jos otanta tehdään takaisinpanolla,

$$\text{D}^2(X) = npq.$$

Jos otanta tehdään ilman takaisinpanoa,

$$\text{D}^2(X) = npq\frac{N-n}{N-1}.$$

Palutetaan mieliin seuraavat kaksi suurten otosten tulosta: Jos perusjoukko S on ääretön (tai hyvin suuri otoskokoon verrattuna),

$$X \sim \text{Bin}(n,p).$$

Jos otoskoko n on tarpeeksi suuri ja p ei ole kovin kaukana arvosta 1/2, niin

$$X \sim_a \text{N}(np,npq).$$

Huomaa, että esitetyt jakaumatulokset ovat siinä mielessä *epäoperationaalisia*, että otosjakauma riippuu tuntemattomasta parametrista p. Se voidaan kuitenkin *estimoida* aineiston perusteella (kts. seuraavaa lukua).

Jatkossa emme tule muutamaa poikkeusta lukuunottamatta kiinnittämään huomiota siihen, poimitaanko otos äärellisestä perusjoukosta takaisinpanolla vai ilman takaisinpanoa. Kaikki tarkastelut tehdään ikään kuin otos olisi poimittu joko äärettömästä perusjoukosta tai äärellisestä perusjoukosta takaisinpanolla. Kuten edellä on todettu tällä seikalla ei kovin paljon merkitystä, ellei otantasuhde n/N ole hyvin iso.

2.2.3 SUHTEELLISEN FREKVENSSIN OTOSJAKAUMA

Olkoon X satunnaismuuttuja, joka kuvaa niiden niiden havaintosyksiköiden *frekvenssiä* eli lukumäärää otoksessa, joilla on ominaisuus $\mathscr{P}$. Tällöin

$$P = X/n$$

on niiden alkioiden *suhteellinen frekvenssi* eli suhteellinen lukumäärä tai suhteellinen osuus, joilla on ominaisuus $\mathscr{P}$.[1]

Koska otoskoko n on vakio, saadaan frekvenssin X odotusarvoa koskevasta tuloksesta

$$\text{E}(P) = p,$$

$$\text{D}^2(P) = \frac{pq}{n}.$$

---

[1] Merkitsemällä suhteellista frekvenssiä isolla kirjaimella P on haluttu korostaa sitä, että P on satunnaismuuttuja, jonka arvo määräytyy otokseen tulleista havainnoista. Aikaisemmin tätä seikkaa ei ole välitetty korostaa.

Tulokset seuraavat siitä, että $P = X/n$ on satunnaismuuttujan X lineaarimuunnos.

Huomaa, että edellisestä tuloksesta seuraa, että otoksesta määrätty ominaisuuden $\mathscr{E}$ omaavien havaintoyksiköiden suhteellisen osuuden odotettavissa oleva arvo yhtyy todennäköisyyteen poimia satunnaisesti ko. ominaisuuden omaava alkio perusjoukosta. Tämä on selvästi miellyttävä ominaisuus suhteelliselle frekvenssille.

Jos otoskoko on tarpeeksi suuri ja p ei ole kovin kaukana arvosta 1/2, niin

$$P \sim_a \mathrm{N}(p, \frac{1}{n}pq).$$

Tämä tulos on siis luonteeltaan asymptoottinen eli se on ns. suurten otosten tulos.

Huomaa, että esitetyt jakaumatulokset ovat siinä mielessä *epäoperationaalisia*, että otosjakauma riippuu tuntemattomasta parametrista p. Se voidaan kuitenkin *estimoida* aineiston perusteella (kts. seuraavaa lukua).

Suhteellisen osuuden P approksimatiiviseen otosjakaumaan on usein syytä liittää *jatkuvuuskorjaus*. Olkoon Y satunnaismuuttuja, jonka jakauma on $\mathrm{N}(p, pq/n)$. Tällöin P:n ja Y:n jakaumat muistuttavat toisiaan sillä tavalla, että

$$\mathrm{P}(a - 1/(2n) \le Y \le b + 1/(2n)) \approx \mathrm{P}(a \le P \le b).$$

Erityisesti

$$\mathrm{P}(P = p) \approx \mathrm{P}(p - 1/(2n) \le Y \le p + 1/(2n)).$$

Korjaus jätetään usein tekemättä.

ESIMERKKI 1.

Tarkastellaan Suomen presidentin vaalin toisen kierroksen tuloksia vuonna 1994. Rehn sai 46% annetuista äänistä. Mikä on todennäköisyys, että ennen vaalia tehdyssä kyselyssä yli puolet olisi kannattanut Rehniä,[1] jos otoskoko olisi ollut

(a) 200,

(b) 1000?

Olkoon P Rehniä kannattaneiden äänioikeutettujen suhteellinen osuus *otoksessa*. Jos otos edusti hyvin perusjoukkoa, oletuksista seuraa, että

$$\mathrm{E}(P) = p = 0.46$$

ja

$$\mathrm{D}^2(P) = \frac{pq}{n} = \frac{0.46 \cdot 0.54}{n},$$

jossa $n = 200$ tai 1000.

Haluamme saada selville todennäköisyyden

[1] Huomaa, että olemme tehneet seuraavan oletuksen: Rehnin kannattajien osuus äänioikeutetuista oli vaaleissa ja ajankohtana, jolloin kysely toteutettiin, sama.

$$P(P > 0.5).$$

Jos käytetämme normaaliapproksimaatiota, tämä todennäköisyys saadaan määräämällä todennäköisyys

$$P(Y > 0.5),$$

jossa $Y \sim N(p, pq/n)$. Standardoinnilla saadaan

$$P(Y > 0.5) = P\left(Z > \frac{0.5 - p}{\sqrt{pq/n}}\right),$$

jossa $Z \sim N(0,1)$.

(a) $n = 200$.

Tällöin

$$D(Y) = \sqrt{pq/n} = 0.0352$$

ja

$$P\left(Z > \frac{0.5 - 0.46}{0.0352}\right) = P(Z > 1.21) = 0.1131.$$

(b) $n = 1000$.

Tällöin

$$D(Y) = \sqrt{pq/n} = 0.0158$$

ja

$$P\left(Z > \frac{0.5 - 0.46}{0.0158}\right) = P(Z > 2.57) = 0.0051.$$

Siten otoskoon kasvattaminen tekee epätodennäköisemmäksi saada otos, jonka perusteella tehtäisiin väärä johtopäätös Rehnin kannatuksesta vaalissa. ●

2.3 KESKIARVON OTOSJAKAUMA

2.3.1 OTOSKESKIARVON OMINAISUUKSIA

Oletetaan, että tutkimuksen kohteena olevasta perusjoukon S ominaisuutta kuvaavasta muuttujasta X on kerätty yksinkertaisella satunnaisotannalla n havaintoa, joita vastaavat muuttujan X havaitut arvot ovat

$$X_1, X_2, \ldots, X_n.$$

Oletetaan lisäksi, että perusjoukko on ääretön tai, että otanta on tehty palauttaen.

Määritellään *otoskeskiarvo* kaavalla

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Oletetaan, että muuttujan X odotusarvo ja varianssi ovat

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

$$D^2(X) = \sigma^2.$$

Tällöin

$$E(X_i) = \mu,\ i = 1,2,\ldots,n,$$

$$D^2(X_i) = \sigma^2,\ i = 1,2,\ldots,n.$$

Tehdyistä oletuksista seuraa, että otoskeskiarvon odotusarvo ja varianssi ovat

$$E(\overline{X}) = \mu,$$

$$D^2(\overline{X}) = \frac{\sigma^2}{n}.$$

Varianssia koskeva tulos pätee, koska otokseen tulleita havaintoja voidaan pitää riippumattomina. Tämä seuraa seuraa siitä, että otoksen poiminta oletettiin tehdyn palauttaen tai, että perusjoukko on ääretön.

Tuloksista nähdään seuraavat seikat:

1. Otoksesta määrätyn otoskeskiarvon odotettavissa oleva arvo yhtyy ominaisuuden odotettavissa olevaan arvoon perusjoukossa.
2. Otoskeskiarvon vaihtelu on pienempää kuin ominaisuuden vaihtelu perusjoukossa.

Huomaa, että otoskeskiarvon odotusarvoa ja varianssia koskevat tulokset ovat siinä mielessä *epäoperationaalisia*, että kumpikin riippuu tuntemattomista parametreista μ ja σ. Ne voidaan kuitenkin *estimoida* havaintoaineiston perusteella (kts. seuraavaa lukua).

2.3.2 OTOSKESKIARVON OTOSJAKAUMA

Jos kappaleessa 1 tehtyjen oletusten lisäksi kaikkien havaintojen X_i oletetaan noudattavan normaalijakaumaa eli

$$X_i \sim N(\mu,\sigma^2),\ i = 1,2,\ldots,n$$

niin otoskeskiarvon otosjakauma on normaalinen:

$$\overline{X} \sim N(\mu,\frac{\sigma^2}{n}).$$

On syytä muistaa, että tämä tulos liittyy äärettömään perusjoukkoon tai otantaan palauttaen. Jos otanta tehdään palauttamatta, niin otoskeskiarvon varianssin kaavaa on korjattava äärellisen perusjoukon korjaustekijällä. Korjaustekijä jätetään kuitenkin tavallisesti huomioimatta, ellei otos muodosta huomattavaa osaa perusjoukosta.

Vaikka havainnot eivät noudattaisikaan normaalijakaumaa, saadaan keskeisestä raja-arvolauseesta tulos

$$\overline{X} \sim_a N(\mu,\frac{\sigma^2}{n}),$$

olettaen, että n on tarpeeksi suuri. Tämä tulos on luonteeltaan asymptoottinen eli se on ns. suurten otosten tulos.

Huomaa, että otoskeskiarvon jakaumaa koskevat tulokset ovat siinä mielessä *epäoperationaalisia*, että kumpikin riippuu tuntemattomista parametreista μ ja σ. Ne voidaan kuitenkin *estimoida* havaintoaineiston perusteella (kts. seuraavaa lukua).

ESIMERKKI 1.

Kone täyttää puolen litran (1l = 1000ml) pulloja viinalla. Täyttömäärä X oletetaan satunnaismuuttujaksi, jonka jakauma on normaalinen:

$$X \sim N(500ml,(20ml)^2).$$

Koneen toiminnan valvomiseksi täyttölinjalta poimitaan kerran tunnissa 25 pulloa 30 sekunnin välein pullojen sisällön mittaamiseksi. Kone pysäytetään, jos mitattujen pullojen sisällön keskiarvo on välin

[490ml,510ml]

ulkopuolella. Mikä on koneen pysäytystodennäköisyys, jos em. oletus pullojen täyttömäärästä pätee?

Oletetaan, että käytetty poimintamenettely takaa havaintojen riippumattomuuden. Tällöin

$$\overline{X} \sim N(500ml,\frac{(20ml)^2}{25}).$$

Siten

$$E(\overline{X}) = 500\text{ml},$$

$$D^2(\overline{X}) = (4\text{ml})^2.$$

Käyttämällä komplementtitapahtuman todennäköisyyden kaavaa, standardoimalla ja käyttämällä standardoidun normaalijakauman taulukoita saadaan:

$$\begin{aligned}
&P(\overline{X} < 490 \textit{ tai } \overline{X} > 510) \\
&= 1 - P(490 \le \overline{X} \le 510) \\
&= 1 - P(\frac{490-500}{4} \le Z \le \frac{510-500}{4}) \\
&= 1 - P(-2.5 \le Z \le +2.5) \\
&= 0.0124.
\end{aligned}$$

Siis hieman useammin kuin 1 kerran 100:sta kone joudutaan pysäyttämään virheellisesti, vaikka se täyttäisikin pulloja kuten pitää.

Tästä voidaan päätellä seuraavaa: Oletetaan, että kone joudutaan pysäyttämään sen takia, että tarkastettujen pullojen sisällön keskiarvo on välin [490,510] ulkopuolella. Tällöin on siis varsin todennäköistä, että oletus koneen toiminnasta ei päde, vaan kone täyttää pulloihin keskimäärin liian vähän tai liian paljon viinaa.

Tällaiset tarkastelut johtavat tilastolliseen testiteoriaan: Tällöin oletus $\mu = 500$ asetetaan otoksesta saatujen tietojen koetteeseen. ●

3. ESTIMOINTI

3.1 PISTE-ESTIMOINTI

3.1.1 JOHDANTO

Tilastotieteen tehtävänä on kehittää empiirisille ilmiöille *malleja*, joiden avulla ilmiöitä voidaan *kuvata*, *selittää*, *ennustaa* ja *kontrolloida*. Tilastollinen malli on ilmiön taustateoriaan[1] ja tilastotieteeseen perustuva kuvaus ilmiön *systemaattisista* ja *satunnaisista piirteistä*. Tilastollisen mallin satunnaisten piirteiden kuvaus perustuu johonkin *todennäköisyysmalliin*. Kuten luvussa 1 on todettu, todennäköisyysmalli koostuu *otosavaruuden* (eli perusjoukon) ja *otosavaruuden tapahtumien* (eli perusjoukon osajoukkojen) *todennäköisyyksien* kuvauksista.

Tilastollinen tutkimus voi kohdistua ainoastaan empiiristen ilmiöiden *mitattaviin* eli *numeerisiin ominaisuuksiin*. Jokaiseen mitattavaan ominaisuuteen voidaan liittää muuttuja, joka kertoo miten ominaisuus vaihtelee tutkimuksen kohteiden joukossa (kts. K1: 2.5). Tilastolliselle tutkimukselle ominainen piirre on se, että tämä vaihtelu voidaan tulkita *satunnaisvaihteluksi*. Siksi tilastollisessa tutkimuksessa ominaisuuksien vaihtelua kuvataan *satunnaismuuttujilla*. Satunnaismuuttujien mahdolliset arvot muodostavat tutkimuksen kohteena olevan ilmiön todennäköisyysmallin otosavaruuden. Otosavaruuden tapahtumien todennäköisyydet saadaan liittämällä ko. satunnaismuuttujiin jotkin *todennäköisyysjakaumat*. Kuten luvussa 1 on todettu todennäköisyysjakaumat riippuvat *parametreista*, joilla on tavallisesti tutkittavan ilmiön ominaisuuksiin liittyvät tulkinnat. Näiden tulkintojen tekeminen on keskeinen osa tilastollista tutkimusta. Ongelmallista on se, että parametrien arvoja ei yleensä tunneta.

Empiiristä ilmiötä tutkitaan keräämällä ilmiötä koskevia *havaintoja*. Havaintoarvot ovat jonkin ilmiön ominaisuutta kuvaavan satunnaismuuttujan arvoja. Itse asiassa havaintoarvo on tutkittavan ominaisuuden mitattu arvo jollekin perusjoukon alkiolle. Jos havaintojen keräämisessä käytetään satunnaisotantaa, havaintoarvot vaihtelevat satunnaisesti otoksesta toiseen. Havainnot ovat satunnaisia, koska havaintoyksiköiden poiminnassa on käytetty arvontaa. Myös *kokonaistutkimuksen* kohteena oleva joukko (esimerkiksi kaikkien suomalaisten joukko) voidaan ja on usein hyödyllistä tulkita satunnaisotokseksi, joka on poimittu *äärettömästä* perusjoukosta.

Tilastollinen tutkimus perustuu oletukseen siitä, että havainnot noudattavat tutkimuksen kohteena olevan ilmiön satunnaisia piirteitä kuvaavassa todennäköisyys-

[1] Tällaisia taustateorioita tuottavat reaalitieteet kuten esimerkiksi fysiikka ja kemia, bio- ja lääketieteet, taloustiede tai sosiaali- ja käyttäymistieteet.

mallissa määriteltyä todennäköisyysjakaumaa. Tämä oletus tekee mahdolliseksi *estimoida* eli *arvioida* jakauman parametrien arvot havaintojen perusteella.

Tutkittavaan ilmiöön liittyy tavallisesti sitä koskevia *hypoteeseja* eli *oletuksia*. Tällaiset hypoteesit pyritään ilmaisemaan ilmiötä kuvaavan todennäköisyysmallin parametrejä koskevina oletuksina. Tilastollisen tutkimuksen päämääränä saattaa olla selvittää pätevätkö asetetut oletukset. Ilmiöstä kerätyt havainnot mahdollistavat asetettujen oletuksien *testaamisen*. Testauksella pyritään selvittämään ovatko oletukset sopusoinnussa ilmiöstä kerättyjen havaintojen kanssa. Testauksen kohteeksi voidaan ottaa myös ilmiötä kuvaavaa todennäköisyysmallia koskevat oletetukset.

Estimointi ja testaus ovat siis tilastollisen tutkimuksen apuvälineitä, joiden avulla tutkittavasta ilmiöstä pyritään tekemään johtopäätöksiä siitä kerätyn havaintoaineiston perusteella.

Tilastollisessa tutkimuksessa on tyypillisesti seuraavat työvaiheet:

1. Muodostetaan tutkimuksen kohteena olevan ilmiön ominaisuuksia koskeva todennäköisyysmalli.
2. Kerätään ilmiötä koskeva havaintoaineisto.
3. Estimoidaan todennäköisyysmallin parametrit havaintojen perusteella.
4. Testataan todennäköisyysmallin parametreja koskevat hypoteesit.
5. Testataan todennäköisyysmallista tehdyt oletukset.

Jos vaiheessa 5 todetaan, että todennäköisyysmallista tehdyt oletukset eivät ole sopusoinnussa havaintojen kanssa, palataan vaiheeseen 1. Tällöin vaiheessa 1 muodostetaan saatujen kokemusten perusteella *korjattu* todennäköisyysmalli. Jos havaintoaineistoa kerättäessä on otettu huomioon tällainen mahdollisuus, saattaa olla riittävää estimoida korjatun todennäköisyysmallin parametrit ja testata korjatun mallin parametrejä koskevat hypoteesit ilman uusien havaintojen keräämistä.

ESIMERKKI 1.

Oletetaan, että haluamme tutkia tilastollisesti miesten pituutta. Pituuksia on tarkasteltava jossakin *hyvin määritellyssä* joukossa. Tällainen hyvin määritelty joukko on esimerkiksi kaikkien vuonna 1995 varusmiespalvelustaan suorittamaan astuneiden miesten joukko. Koska pituus on ominaisuus, joka määräytyy monimutkaisella tavalla monista perintö- ja ympäristötekijöistä, pituutta voidaan pitää satunnaismuuttujana, joka noudattaa normaalijakaumaa. Varusmiespalvelustaan vuonna 1995 astuneiden miesten pituuksien tilastolliseksi malliksi voidaan siksi ottaa seuraava todennäköisyysmalli:

> Satunnaisesti valitun miehen pituus X on satunnaismuuttuja, joka noudattaa normaalijakaumaa $N(\mu,\sigma^2)$.

Malli riippuu (tuntemattomista) parametreista μ ja σ^2, joilla on seuraava tulkinnat:

$$\mu = E(X)$$

on satunnaisesti valitun miehen pituuden odotettavissa oleva arvo. Parametri μ on myös pituuksien jakauman painopiste.

$$\sigma^2 = D^2(X) = E(X - \mu)^2$$

on satunnaisesti valitun miehen pituuden omasta odostusarvostaan määrätyn poikkeaman neliön odotusarvo ja kuvaa pituuksien jakauman keskittyneisyyttä tai hajaantuneisuutta pituuksien odotusarvon ympärille.

Kerätään miesten pituuksista havaintoja. Havainnot kerätään siitä joukosta, joka on kiinnostuksen kohteena. Olkoon kerättyjen havaintojen lukumäärä n. Oletetaan, että havaintojen kerääminen toteutetaan käyttämällä yksinkertaista satunnaisotantaa (palauttaen). Tällöin havaintoarvoja

$$X_1, X_2, \ldots, X_n$$

voidaan pitää riippumattomina satunnaismuuttujina, joista jokainen noudattaa normaalijakaumaa $N(\mu,\sigma^2)$.

Estimoinnin tehtävänä on muodostaa parhaat mahdolliset arviot parametreille μ ja σ^2. Jos parametrista μ on muodostettu jokin hypoteesi, voidaan sen sopusointua havaintojen kanssa tutkia muodostamalla hypoteesille jokin tilastollinen testi.

Miesten pituuksien tilastolliseen malliin liittyy *jakaumaoletus*, jonka mukaan satunnaisesti valitun miehen pituus X noudattaa normaalijakaumaa $N(\mu,\sigma^2)$. Myös tätä oletusta voidaan testata aineiston avulla. ●

Tämän kappaleen tarkoituksena on kuvata *estimoinnin* peruskäsitteet. Sovelluksina tarkastellaan suhteellisen osuuden ja odotusarvon estimointia. Testausta käsitellään seuraavassa luvussa.

3.1.2 ESTIMAATTORIT JA NIIDEN OMINAISUUDET

Oletetaan, että kiinnostuksen kohteena oleva perusjoukon ominaisuutta kuvataan satunnaismuuttujalla X. Oletetaan, että perusjoukosta S on poimittu yksinkertainen satunnaisotos (palauttaen) ja oletetaan, että satunnaismuuttujan X arvot (eli havaintokohtaiset arvot) otoksessa ovat

$$X_1, X_2, \ldots, X_n.$$

Otantamenetelmästä johtuen havaintoja voidaan pitää riippumattomina.

Havaintojen $X_1, X_2, \ldots, X_n$ mitä tahansa funktiota

$$T = T(X_1, X_2, \ldots, X_n)$$

kutsutaan *otossuureeksi*. Koska havainnot $X_1, X_2, \ldots, X_n$ ovat *satunnaismuuttujia*, myös otossuure T on satunnaismuuttuja: sen arvo vaihtelee satunnaisesti otoksesta toiseen. Kaikki *otostunnusluvut*, kuten esimerkiksi havaintojen keskiarvo ja hajonta, ovat otossuureita.

Jos otossuuretta T käytetään perusjoukon ominaisuudelle muodostetun tilastollisen mallin parametrien estimointiin, sanotaan otossuuretta parametrin

estimaattoriksi. Estimaattori on otossuureena satunnaismuuttuja. Estimaattorin otoskohtaisia arvoja sanotaan *estimaateiksi*.

Oletetaan, että satunnaismuuttujalle X muodostettu tilastollinen malli eli siitä tehty jakaumaoletus on muotoa

$$f(x;\theta),$$

jossa f on satunnaismuuttujan X jakauman pistetodennäköisyysfunktio tai tiheysfunktio ja θ on jokin jakauman muodon määräävä parametri. Parametrin θ arvoa ei yleensä tunneta. Merkitään parametrin estimaattoria asettamalla parametria tarkoittavan (kreikkalaisen) kirjaimen päälle "hattu". Siten parametrin θ estimaattori on satunnaismuuttuja

$$\hat{\theta} = \theta(X_1, X_2, \cdots, X_n).$$

Estimaattori $\hat{\theta}$ on havaintojen $X_1, X_2, \ldots, X_n$ funktiona satunnaismuuttuja.[1] Satunnaismuuttujana parametrin θ estimaattorilla $\hat{\theta}$ on todennäköisyysjakauma, johon tilastollinen päättely suurelta osin perustuu.

Mikä tahansa havaintojen funktio ei ole järkevä estimaattori tilastollisen mallin parametrille. Siksi tilastotieteessä on tapana tarkastella *hyvälle* estimaattorille esitettäviä vaatimuksia. On syytä panna mieleensä, että seuraavassa esitetyt vaatimukset ovat niin rajoittavia, että niitä kaikkia ei voida täyttää muuta kuin sellaisissa yksinkertaisissa tilanteissa, joita käsitellään tällaisessa tilastotieteen alkeisesityksissä.

HYVÄN ESTIMAATTORIN OMINAISUUKSIA

Olkoon θ jonkin todennäköisyysjakauman parametri ja $\hat{\theta}$ sen estimaattori.

HARHATTOMUUS

Jos

$$E(\hat{\theta}) = \theta,$$

niin estimaattori $\hat{\theta}$ on *harhaton*.

Jos estimaattori $\hat{\theta}$ on harhaton, estimaattorin *odotettavissa oleva arvo* yhtyy tuntemattoman parametrin θ todelliseen arvoon. Tällä tarkoitetaan seuraavaa: Oletetaan, että parametri θ estimoidaan sen harhatonta estimaattoria $\hat{\theta}$ käyttäen useista toisistaan riippumattomista satunnaisotoksista. Tällöin estimaattori $\hat{\theta}$ saa eri otoksissa arvoja, jotka keskittyvät parametrin θ todellisen arvon ympärille.

---

[1] Poikkeamme tässä sopimuksesta, jonka mukaan satunnaismuuttujia merkitään kursivoiduilla (latinalaisten) aakkosten loppupään kirjaimilla.

TYHJENTÄVYYS

$\hat{\theta}$ on *tyhjentävä*, jos se käyttää kaiken otokseen sisältyvän parametria θ koskevan informaation.

Tyhjentävä estimaattori $\hat{\theta}$ ei siis jätä mitään otokseen sisältyvää parametria θ koskevaa informaatiota käyttämättä.

TARKENTUVUUS

$\hat{\theta}$ on *tarkentuva*, jos estimaattorin $\hat{\theta}$ arvot lähestyvät parametrin θ todellista arvoa siinä mielessä, että suuret poikkeamat todellisesta arvosta tulevat yhä epätodennäköisemmiksi otoskoon kasvaessa.

Tarkentuvan estimaattorin $\hat{\theta}$ arvot lähestyvät tuntematonta parametrin θ arvoa otoskoon kasvaessa. Huomaa, että *suurten lukujen laki* sanoo, että otoskeskiarvo on (tietyin ehdoin) havaintojen odostusarvon tarkentuva estimaattori (kts. kappale 1.2.5).

TEHOKKUUS

Olkoot $\hat{\theta}_1$ ja $\hat{\theta}_2$ kaksi parametrin θ harhatonta estimaattoria. Tällöin $\hat{\theta}_1$ on *tehokkaampi* kuin $\hat{\theta}_2$, jos

$$D^2(\hat{\theta}_1) \leq D^2(\hat{\theta}_2).$$

Kahdesta saman parametrin estimaattorista on siis tehokkaampi se, jonka varianssi on pienempi. Jos parametrin θ harhattoman estimaattorin $\hat{\theta}$ varianssi on pienempi kuin minkä tahansa muun harhattoman estimaattorin varianssi, estimaattoria $\hat{\theta}$ kutsutaan *täystehokkaaksi*.

Seuraavassa käy ilmi, että yksinkertaisten satunnaisotoksen tapauksessa tavanomaisilla binomi- ja normaalijakauman parametrien estimaattoreilla on kaikki edellä mainitut hyvyysominaisuudet. Tästä *ei saa* kuitenkaan tehdä sellaista johtopäätöstä, että sama pätee myös monimutkaisempien otosasetelmien tai tilastollisten mallien yhteydessä. Tilastollisen mallin parametreille saattaa olla tarjolla useita vaihtoehtoisia estimaattoreita, joista jokaisella on omat hyvät puolensa. Tällaisissa tapauksissa hyvän estimaattorin valinta saattaa olla niin vaativa tehtävä, että on syytä kääntyä asiantuntijan puoleen.

3.1.3 ESTIMOINTIMENETELMÄT

Miten tilastollisen mallin parametrit estimoidaan?

Oletetaan, että tutkimuksen kohteena oleva perusjoukon S ominaisuutta kuvataan satunnaismuuttujalla X, jonka todennäköisyysjakauman pistetodennäköisyysfunktio tai tiheysfunktio on funktio $f(x;\theta)$. Funktio $f(x;\theta)$ riippuu parametrista θ, jonka arvo ei tunneta. Poimitaan perusjoukosta S yksinkertainen satunnaisotos (palauttaen) ja oletetaan, että satunnaismuuttujan X arvot otoksessa ovat

$$X_1, X_2, \ldots, X_n.$$

Jokainen havainto X_i noudattaa siis jakaumaa $f(x_i;\theta)$.

Parametrien estimointiin on kehitetty monia erilaisia menetelmiä. Niistä ehdottomasti tärkein on *suurimman uskottavuuden menetelmä*.[1]

SUURIMMAN USKOTTAVUUDEN ESTIMAATTORI

$\hat{\theta}$ on parametrin θ suurimman uskottavuuden estimaattori, jos se maksimoi otoksen $X_1, X_2, \ldots, X_n$ todennäköisyyden.

Mikään muu parametrin arvo ei siis tuota yhtä suurella todennäköisyydellä sitä otosta, joka on saatu. Käytämme suurimman uskottavuuden estimaattorista usein lyhennettä SU-estimaattori.

Suurimman uskottavuuden menetelmää ei kehitetä tässä matemaattisesti täsmällisellä tavalla, koska tällä kurssilla ei ole määritelty usean satunnaismuuttujan *yhteisjakauman* käsitettä. Voidaan osoittaa, että riippumattomien samaa jakaumaa noudattavien havaintojen yhteisjakauma on havaintojen pistetodennäköisyysfunktioiden tai tiheysfunktioiden tulo. Siten havaintojen $X_1, X_2, \ldots, X_n$ yhteisjakauma on muotoa

$$f(x_1;\theta)\cdot f(x_2;\theta)\cdot \ldots \cdot f(x_n;\theta).$$

Yhteisjakauma riippuu estimoitavasta parametrista θ ja sitä kutsutaan θ:n funktiona parametrin θ *uskottavuusfunktioksi*. Suurimman uskottavuuden estimaattori $\hat{\theta}$ maksimoi uskottavuusfunktion parametrin θ suhteen. SU-estimaattori voidaan tavallisesti määrätä koulun matematiikan kurssilta tutulla menetelmällä, jolla etsitään funktion (paikalliset) *ääriarvot*:

1. Derivoidaan uskottavuusfunktio θ:n suhteen.
2. Asetetaan derivaatta 0:ksi.
3. Ratkaistaan näin saatu yhtälö θ:n suhteen.
4. Varmistaudutaan siitä, että saatu derivaatan nollakohta vastaa maksimia.

Esimerkkinä suurimman uskottavuuden menetelmän käytöstä seuraavassa esitetään normaalijakauman $N(\mu,\sigma^2)$ odotusarvoparametrin μ estimaattorin johto:

---

[1] *Engl.* maximum likelihood method, lyh. ML-method.

ESIMERKKI 1.

Olkoot havainnot $X_1, X_2, \ldots, X_n$ riippumattomia ja oletetaan, että jokainen X_i noudattaa normaalijakaumaa $N(\mu,\sigma^2)$. Otoksen $X_1, X_2, \ldots, X_n$ yhteisjakauma on havaintojen $X_1, X_2, \ldots, X_n$ riippumattomuuden takia muotoa

$$f(x_1;\mu,\sigma^2)\cdot f(x_2;\mu,\sigma^2)\cdot\ldots\cdot f(x_n;\mu,\sigma^2),$$

jossa

$$f(x_i;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\},\ i=1,2,\ldots,n$$

on normaalijakauman tiheysfunktio.

Havaintojen $X_1, X_2, \ldots, X_n$ yhteisjakauma voidaan kirjoittaa eksponenttifunktion laskusääntöjä käyttämällä muotoon

$$\sigma^{-n}(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

Tämä lauseke on maksimoitava parametrin μ suhteen. Tätä maksimointitehtävää voidaan yksinkertaistaa ottamalla lausekkeesta logaritmi ja jättämällä pois maksimoinnin kannalta sellaiset epäolennaiset termit, jotka eivät sisällä μ:tä. Tämä yksinkertaistus on luvallinen, sillä logaritmointi ei vaikuta maksimoinnin tulokseen, koska logaritmifunktio on monotonisesti kasvava funktio.

Maksimointi voidaan siten kohdistaa μ:n funktioon

$$g(\mu)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Derivoimalla funktio $g(\mu)$ μ:n suhteen ja merkitsemällä tulos 0:ksi saadaan μ:n SU-estimaattorin määräämiseksi yhtälö

$$\frac{dg(\mu)}{d\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0.$$

Tästä saadaan helposti yhtälö

$$\sum_{i=1}^{n}x_i-n\mu=0,$$

josta saadaan parametrin μ suurimman uskottavuuden estimaattoriksi havaintojen $X_1, X_2, \ldots, X_n$ aritmeettinen keskiarvo

$$\hat{\mu}=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Parametrin μ arvo $\hat{\mu}$ vastaa uskottavuusfunktion maksimia, koska funktion $g(\mu)$ toinen derivaatta μ:n suhteen on negatiivinen kaikille μ:n arvoille:

$$\frac{d^2g(\mu)}{d\mu^2}=-\frac{n}{\sigma^2}<0.\ \bullet$$

Emme tule jatkossa käyttämään suurimman uskottavuuden menetelmää eksplisiittisesti, vaikka tulemme kyllä kertomaan ovatko käsiteltävät estimaattorit suurimman uskottavuuden estimaattoreita. On kuitenkin hyvä tietää jo nyt, että regressiomallin parametrien *pienimmän neliösumman menetelmä* tuottaa tietyin edellytyksin saman tuloksen kuin suurimman uskottavuuden menetelmä.

Suurimman uskottavuuden estimaattorilla on hyvin yleisin ehdoin seuraavat ominaisuudet:

- SU-estimaattori on *tarkentuva*.
- SU-estimaattori on suurilla otoskoilla *approksimatiivisesti normaalinen*.

Tämä merkitsee sitä, että SU-estimaattori toteuttaa kaikissa tavallisissa tapauksissa *suurten lukujen lain* ja *keskeisen raja-arvolauseen* (kts. kappaleet 1.2.5 ja 1.3.2). Kumpikin mainituista ominaisuuksista on sekä teoreettisesti että käytännöllisesti tärkeä. Jälkimmäistä ominaisuutta voidaan pitää lisäperusteluna normaalijakauman keskeiselle asemalle tilastotieteessä.

3.1.4 NORMAALIJAKAUMAN PARAMETRIEN ESTIMOINTI

Oletetaan, että tutkimuksen kohteena olevan satunnaisilmiön ominaisuutta kuvaavasta satunnaismuuttujasta X on käytettävissä havainnot $X_1, X_2, \ldots, X_n$, jotka ovat riippumattomia ja noudattavat samaa jakaumaa $N(\mu,\sigma^2)$.

Edellä todettiin, että odotusarvo- eli paikkaparametrin μ SU-estimaattori on havaintojen $X_1, X_2, \ldots, X_n$ aritmeettinen keskiarvo

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Voidaan osoittaa, että varianssiparametrin σ^2 suurimman uskottavuuden estimaattori on havaintojen $X_1, X_2, \ldots, X_n$ otosvarianssi

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 .$$

Huomaa, että varianssin σ^2 suurimman uskottavuuden estimaattorin $\hat{\sigma}^2$ kaavassa on jakajana n, eikä $n-1$ kuten tavanomaisessa otosvarianssin s^2 kaavassa.

Normaalijakauman odotusarvoparametrin μ suurimman uskottavuuden estimaattorina otoskeskiarvolla $\overline{X}$ on seuraavat ominaisuudet: $\overline{X}$ on

- harhaton,
- tyhjentävä,
- tarkentuva,
- täystehokas,
- normaalinen $N(\mu,\sigma^2/n)$.

Nämä otoskeskiarvon ominaisuudet normaalijakauman paikkaparametrin estimaattorina merkitsevät sitä, että otoskeskiarvo kuvaa hyvin *normaalijakautuneen* tai lähes normaalijakautuneen aineiston keskimääräisten arvojen sijaintia. Jos otos ei ole normaalijakautunut, otoskeskiarvolla ei ole välttämättä mainittuja hyviä ominaisuuksia odotusarvon estimaattorina. Otoskeskiarvo on siis parhaimmillaan kuvattaessa normaalijakautuneen perusjoukon paikkaa. Jos perusjoukko ei ole jakautunut normaalijakauman mukaan, saattaa olla syytä harkita jotakin muuta paikan estimointiin sopivan tunnusluvun, kuten mediaanin tai moodin, käyttöä.

Normaalijakauman paikkaparametri μ voidaan estimoida myös määräämällä havaintojen *mediaani Md*.[1] Myös mediaani *Md* on parametrin μ estimaattorina harhaton. Sen sijaan mediaani ei ole tällaisessa tilanteessa yhtä tehokas kuin aritmeettinen keskiarvo. Voidaan osoittaa, että normaalijakautuneen perusjoukon tapauksessa

---

[1] Mediaani on suuruusjärjestykseen asetettujen havaintojen keskimmäinen havainto. Se jakaa aineiston kahteen yhtä suureen osaan, joista toisessa kaikki havaintoarvot ovat mediaaniarvoa suurempia, toisessa pienempiä.

$$D^2(Md) = \frac{\pi}{2} \cdot \frac{\sigma^2}{n} > \frac{\sigma^2}{n} = D^2(\bar{X}).$$

Vaikka mediaani ei siis ole normaalijakautuneen perusjoukon tapauksessa yhtä hyvä odotusarvon estimaattori kuin otoskeskiarvo, se on usein suositeltava ei-normaalisen perusjoukon tapauksessa. Näin on varsinkin silloin, kun perusjoukon jakauma on sellainen *vino* jakauma kuten esimerkiksi tulojakauma.

Varianssin σ^2 suurimman uskottavuuden estimaattori $\hat{\sigma}^2$ ei ole harhaton. Sen sijaan varianssiestimaattori

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

on harhaton. Huomaa, että

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2.$$

Siten estimaattoreiden välinen ero on käytännössä merkityksetön. Varianssin suurimman uskottavuuden estimaattorin $\hat{\sigma}^2$ ja harhattoman estimaattorin s^2 ero häviää kokonaan, kun otoskoon n annetaan kasvaa rajatta. Kumpikin estimaattoreista on siis tarkentuva. Harhatonta estimaattoria s^2 käytetään ehkä yleisemmin kuin SU-estimaattoria $\hat{\sigma}^2$.

ODOTUSARVON JA VARIANSSIN ESTIMOINTI ILMAN NORMAALISUUSOLETUSTA

Oletetaan, että tutkimuksen kohteena olevan satunnaisilmiön ominaisuutta kuvaavan satunnaismuuttujan X odotusarvo on μ ja varianssi on σ^2. Oletetaan, että ilmiöstä on käytettävissä havainnot $X_1, X_2, \ldots, X_n$, jotka ovat riippumattomia ja, että kaikilla havainnoilla sama odotusarvo μ ja varianssi σ^2. Nyt satunnaismuuttujasta X ei siis tehdä normaalisuusoletusta.

Olkoon $\bar{X}$ havaintojen $X_1, X_2, \ldots, X_n$ aritmeettinen keskiarvo. Otoskeskiarvolla $\bar{X}$ on havaintojen odotusarvon estimaattorina seuraavat ominaisuudet: $\bar{X}$ on

- harhaton,
- (tietyin ehdoin) suurissa otoksissa approksimatiivisesti normaalinen $N(\mu,\sigma^2/n)$.

Nämä otoskeskiarvon $\bar{X}$ ominaisuudet on jo mainittu otosjakaumaa käsitelleessä kohdassa (kts. kappale 2.3.2).

Olkoon s^2 havaintojen $X_1, X_2, \ldots, X_n$ otosvarianssi. Otosvarianssilla s^2 on havaintojen teoreettisen varianssin estimaattorina seuraavat ominaisuudet: s^2 on

- harhaton.

Tässä kappaleessa esitettyjen tulosten mukaan aritmeettinen keskiarvo $\bar{X}$ ja otosvarianssi s^2 ovat usein käyttökelpoisia estimaattoreita odotusarvolle μ ja varianssille σ^2 myös silloin, kuin havainnot eivät ole normaalijakautuneita.

3.1.5 BINOMIJAKAUMAN PARAMETRIEN ESTIMOINTI

Oletetaan, että tutkimuksen kohteena on tapahtuman A esiintyminen n-kertaisessa toistokokeessa. Olkoon X satunnaismuuttuja, joka kuvaa tapahtuman A frekvenssiä näissä toistoissa. Tällöin X noudattaa binomijakaumaa Bin(n,p). Tässä tapauksessa uskottavuusfunktiona on satunnaismuuttujan X pistetodennäköisyysfunktio.

Vaikka binomijakauma riippuu otoskoosta n, sitä ei voi pitää parametrina tavanomaisessa mielessä, koska n yleensä tunnetaan. Voidaan osoittaa, että tapahtuman A todennäköisyyttä kuvaavan parametrin p SU-estimaattori on

$$\hat{p} = \frac{X}{n}.$$

Tämä merkitsee sitä, että *suhteellinen frekvenssi X/n* on todennäköisyyden p SU-estimaattori.

Binomijakauman parametrin p suurimman uskottavuuden estimaattorina suhteellisella frekvenssillä $\hat{p}$ on seuraavat ominaisuudet: $\hat{p}$ on

- harhaton,
- tyhjentävä,
- tarkentuva,
- täystehokas,
- approksimatiivisesti normaalinen N(p, pq/n).

Nämä suhteellisen frekvenssin $\hat{p}$ ominaisuudet binomijakauman parametrin p estimaattorina toimivat lisäperusteluina suhteellisen frekvenssin yleiselle käytölle kuvattaessa jonkin tapahtuman todennäköisyyttä.

3.2 VÄLIESTIMOINTI

3.2.1 JOHDANTO

Edellisessä kappaleessa tarkasteltiin tutkimuksen kohteena olevan perusjoukon ominaisuutta kuvaavan satunnaismuuttujan todennäköisyysjakauman parametrien *estimointia* eli arviointia perusjoukosta kerätyn havaintoaineiston perusteella. Parametreihin voidaan tavallisesti liittää tulkinta perusjoukon ominaisuuksina. Esimerkiksi odotusarvo kuvaa perusjoukon jakauman paikkaa.

Palautetaan mieleen, että parametrin arvon määräävää havaintoarvojen funktiota kutsutaan parametrin *estimaattoriksi* ja ko. funktion havaintoarvoista määrättyä arvoa kutsutaan parametrin *estimaatiksi*.

Koska päämääränä oli saada arvio parametrin todelliselle, mutta tuntemattomalle arvolle, käytettyä menetelmää kutsutaan usein *piste-estimoinniksi*. On kuitenkin osoittautunut hyödylliseksi liittää estimaattorin arvon määräämiseen tarkastelu, jota kutsutaan *väliestimoinniksi*. Väliestimoinnissa parametrille määrätään havainnoista riippuva väli, joka peittää tietyllä, tutkijan valittavissa olevalla todennäköisyydellä tuntemattoman parametrin arvon. Ko. väliä kutsutaan *luottamusväliksi* ja tutkijan valitsemaa todennäköisyyttä kutsutaan *luottamustasoksi*. Luottamusväli määrätään tavallisesti niin, että se on symmetrinen kiinnostuksen kohteena olevan parametrin estimaattorin suhteen.

Luottamusvälille voidaan antaa todennäköisyyden frekevenssitulkinnan mukaan seuraava tulkinta:

> *Oletetaan, että tutkimuksen kohteena olevasta perusjoukosta poimitaan useita toisistaan riippumattomia satunnaisotoksia. Määrätään niiden kerätyistä otoksista muodostettujen luottamusvälien suhteellinen frekvenssi, jotka peittävät tuntemattoman parametrin arvon. Tämä suhteellinen frekvenssi yhtyy valittuun luottamustasoon.*

Luottamustaso kuvaa eräässä mielessä sitä *varmuutta*, jonka voimme havaintojen perusteella saada siitä, että tuntematon parametrin arvo sijaitsee luottamusvälillä.

Olkoon θ parametri, jonka luottamusväli halutaan määrätä ja olkoon $\hat{\theta}$ parametrin θ estimaattori. Estimaattorin $\hat{\theta}$ suhteen *symmetrinen* luottamusväli parametrille θ on muotoa

$$\hat{\theta} - a,\ \hat{\theta} + a,$$

jossa $-a$ ja $+a$ riippuvat yleensä sekä havainnoista että valitusta luottamustasosta. Jos luottamustasoksi on valittu $1 - \alpha$, niin $-a$ ja $+a$ pyritään määräämään siten, että ehto

$$P(\hat{\theta} - a \le \theta \le \hat{\theta} + a) = 1 - \alpha$$

pätee. Ehdossa mainittu todennäköisyyden määräämisessä käytetään tavallisesti apuna estimaattorin $\hat{\theta}$ otosjakaumaa.

Olkoon parametrin θ estimaattori $\hat{\theta}$ harhaton, ts. olkoon $E(\hat{\theta}) = \theta$ ja olkoon estimaattorin $\hat{\theta}$ varianssi $D^2(\hat{\theta})$. Olkoon $\hat{D}^2(\hat{\theta})$ on estimaattorin $\hat{\theta}$ varianssin estimaattori. *Standardoidaan* $\hat{\theta}$ määrittelemällä satunnaismuuttuja

$$U = \frac{\hat{\theta} - E(\hat{\theta})}{\hat{D}(\hat{\theta})}$$
$$= \frac{\hat{\theta} - \theta}{\hat{D}(\hat{\theta})}.$$

Oletetaan, että satunnaismuuttujan U jakauma on symmetrinen pisteen 0 suhteen ja olkoon satunnaismuuttujan U jakauman kertymäfunktio G.

Määrätään satunnaismuuttujan U kertymäfunktion G avulla piste $+u_{\alpha/2}$, joka toteuttaa ehdon

$$P(U \geq +u_{\alpha/2}) = \alpha/2.$$

Satunnaismuuttujan U jakauman symmetrisyyden takia piste $-u_{\alpha/2}$ toteuttaa ehdon

$$P(U \leq -u_{\alpha/2}) = \alpha/2.$$

Siten

$$P(-u_{\alpha/2} \leq U \leq +u_{\alpha/2}) = 1 - \alpha.$$

Ottamalla huomioon satunnaismuuttujan U määritelmä, voidaan epäyhtälöketju

$$-u_{\alpha/2} \leq U \leq +u_{\alpha/2}$$

muokata sen kanssa yhtäpitävään muotoon

$$\hat{\theta} - u_{\alpha/2}\,\hat{D}(\hat{\theta}) \leq \theta \leq \hat{\theta} + u_{\alpha/2}\,\hat{D}(\hat{\theta}).$$

Koska nämä epäyhtälöketjut ovat yhtäpitäviä,

$$P(\hat{\theta} - u_{\alpha/2}\,\hat{D}(\hat{\theta}) \leq \theta \leq \hat{\theta} + u_{\alpha/2}\,\hat{D}(\hat{\theta})) = 1 - \alpha.$$

Siten parametrin θ symmetrinen luottamusväli on muotoa

$$\hat{\theta} - u_{\alpha/2}\,\hat{D}(\hat{\theta}),\ \hat{\theta} + u_{\alpha/2}\,\hat{D}(\hat{\theta}),$$

jossa $+u_{\alpha/2}$ on *luottamustasoon* $1 - \alpha$ liittyvä *luottamuskerroin*.

Seuraavassa tarkastellaan odotusarvon ja suhteellisen osuuden luottamusvälejä.

3.2.2 ODOTUSARVON LUOTTAMUSVÄLI

Oletetaan, että tutkimuksen kohteena oleva perusjoukon ominaisuutta kuvaava satunnaismuuttujan X odotusarvo $E(X) = \mu$ ja varianssi $D^2(X) = \sigma^2$. Tarkastelu jaetaan seuraavassa kolmeen osaan:

1. Oletetaan, että X on normaalinen ja, että σ^2 on tunnettu.
2. Oletetaan, että X on normaalinen ja, että μ ja σ^2 ovat tuntemattomia.
3. Oletetaan, että X ei ole välttämättä normaalinen ja, että μ ja σ^2 ovat tuntemattomia.

ODOTUSARVON LUOTTAMUSVÄLI, KUN σ^2 TUNNETAAN

Oletetaan, että tutkittavaa perusjoukon ominaisuutta kuvaava satunnaismuuttuja $X \sim N(\mu,\sigma^2)$. Oletetaan, että σ^2 on tunnettu. Tämä oletus on epärealisitinen, mutta konstruoitava luottamusväli auttaa ymmärtämään yleisempää tapausta. Olkoot $X_1, X_2, \ldots, X_n$ yksinkertainen satunnaisotos jakaumasta $N(\mu,\sigma^2)$ ja olkoon

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

havaintojen aritmeettinen keskiarvo. Kappaleesta 3.1.2 tiedetään, että $\overline{X}$ on parametrin μ suurimman uskottavuuden estimaattori.

Luottamusvälin konstruktio perustuu siihen, että otoksen normaalisuuden takia otoskeskiarvo $\overline{X} \sim N(\mu,\sigma^2/n)$. Siten standardoitu muuttuja

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Määritellään väli

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

jossa $-z_{\alpha/2}$ on piste, joka erottaa standardoidun normaalijakauman vasemmalle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Symmetrian takia piste $+z_{\alpha/2}$ erottaa standardoidun normaalijakauman oikealle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Siten $+z_{\alpha/2}$ ja $-z_{\alpha/2}$ ovat luottamustasoon $1-\alpha$ liittyvät luottamuskertoimet. Jos $\Phi(z)$ on standardoidun normaalijakauman kertymäfunktio niin,

$$\Phi(-z_{\alpha/2}) = \alpha/2,$$

$$\Phi(+z_{\alpha/2}) = 1-\alpha/2.$$

Tällöin ko. väli on parametrin μ luottamusväli luottamustasolla $1-\alpha$. Väli siis peittää tuntemattoman parametrin arvon μ todennäköisyydellä $1-\alpha$:

$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

Luottamusväli on symmetrinen keskipisteensä $\overline{X}$ suhteen ja välin keskipiste vaihtelee satunnaisesti otoksesta toiseen. Eri otoksista lasketuista väleistä $100\times(1-\alpha)\%$ kuitenkin peittää tuntemattoman μ:n arvon.

Todettakoon lopuksi, että luottamusvälin pituudella on seuraavat ominaisuudet:

Luottamusväli *kapenee*, jos

- havaintojen lukumäärä n kasvaa,
- luottamustasoa $1-\alpha$ pienennetään,
- standardipoikkeama σ pienenee.

Luottamusväli *levenee*, jos

- luottamustasoa $1-\alpha$ kasvatetaan,
- standardipoikkeama σ kasvaa.

ODOTUSARVON LUOTTAMUSVÄLI, KUN σ^2 EI OLE TUNNETTU

Oletetaan, että tutkittavaa perusjoukon ominaisuutta kuvaava satunnaismuuttuja $X \sim N(\mu,\sigma^2)$. Oletetaan, että σ^2 ei ole tunnettu. Olkoot $X_1, X_2, \ldots, X_n$ yksinkertainen satunnaisotos jakaumasta $N(\mu,\sigma^2)$ ja olkoon

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

havaintojen aritmeettinen keskiarvo ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

havaintojen otosvarianssi. Kappaleesta 3.1.2 tiedetään, että $\overline{X}$ on parametrin μ suurimman uskottavuuden estimaattori ja s^2 on parametrin σ^2 harhaton estimaattori.

Luottamusvälin konstruktio perustuu siihen, että seuraava tulos pätee: Jos perusjoukko on normaalijakautunut, standardoitu muuttuja

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

on jakautunut kuten ns. *Studentin t-jakauma*, jossa vapausasteiden luku on $n-1$. Tähän tulokseen perustuu myös ns. t-testisuure (kts. testausta käsittelevää lukua).

Määritellään väli

$$\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

jossa $-t_{\alpha/2}$ on piste, joka erottaa t-jakauman, jossa vapausasteiden luku on $n-1$, vasemmalle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Symmetrian takia piste $+t_{\alpha/2}$ erottaa t-jakauman, jossa vapausasteiden luku on $n-1$, oikealle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Siten $t_{\alpha/2}$ on luottamustasoon $1-\alpha$ liittyvä luottamuskerroin. Olkoon $T_{n-1}(x)$ t-jakauman, jossa vapausasteiden luku on $n-1$, kertymäfunktio. Tällöin

$$T_{n-1}(-t_{\alpha/2}) = \alpha/2,$$

$$T_{n-1}(+t_{\alpha/2}) = 1 - \alpha/2.$$

Siten ym. väli on parametrin μ luottamusväli luottamustasolla $1-\alpha$. Väli siis peittää tuntemattoman parametrin arvon μ todennäköisyydellä $1-\alpha$.

Luottamusväli on symmetrinen keskipisteensä $\overline{X}$ suhteen ja sekä välin keskipiste että pituus vaihtelevat satunnaisesti otoksesta toiseen. Eri otoksista lasketuista väleistä $(1-\alpha)\%$ kuitenkin peittää tuntemattoman μ:n arvon:

$$P(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha.$$

Siitä, että parametri σ^2 joudutaan estimoimaan havainnoista, seuraa siis se, että normaalijakaumasta määrätyt pisteet $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ on korvattava t-jakaumasta määrätyillä pisteillä $-t_{\alpha/2}$ ja $+t_{\alpha/2}$. Miten tämä vaikuttaa luottamusväliin?

> *Jos $\sigma^2 = s^2$, t-jakaumasta määrätty luottamusväli on leveämpi kuin normaalijakaumasta määrätty luottamusväli.*

Tämä on ymmärrettävää, koska siitä, että parametria σ^2 ei tunneta seuraa se, että myös luottamusvälin päätepisteet vaihtelevat otoksesta toiseen. Siten siitä, että σ^2 ei ole tunnettu, maksetaan se hinta, että tuntemattoman parametrin μ sijaintiin liittyy suurempi epävarmuus.

Todettakoon lopuksi, että luottamusvälin pituudella on seuraavat ominaisuudet:

Luottamusväli *kapenee*, jos

- havaintojen lukumäärä n kasvaa,
- luottamustasoa $1 - \alpha$ pienennetään,
- otoshajonta s pienenee.

Luottamusväli *levenee*, jos

- luottamustasoa $1 - \alpha$ kasvatetaan,
- otoshajonta s kasvaa.

STUDENTIN t-JAKAUMA

Studentin t-jakauma on jatkuva todennäköisyysjakauma, joka silmämääräisesti muistuttaa standardoitua normaalijakaumaa. Esimerkiksi t-jakauma on symmetrinen pisteen 0 suhteen ja sen odotusrvo, mediaani ja moodi sijaitsevat pisteessä 0. Se on kuitenkin *paksuhäntäisempi* kuin normaalijakauma ja tuottaa siten leveämpiä luottamusvälejä kuin normaalijakauma. t-jakauman muoto riippuu ns. *vapausasteiden lukumäärästä f*. Kun f kasvaa, t-jakauma alkaa muistuttaa yhä enemmän normaalijakaumaa. Voidaan osoittaa, että f:n kasvaessa rajatta t-jakauma lähestyy normaalijakaumaa siinä mielessä, että hyvin suurilla vapausasteiden lukumäärillä jakaumat tuottavat käytännössä samat todennäköisyydet. Tämä näkyy myös t-jakauman taulukoista: Jo f:n arvolla 30 jakaumien tuottamat todennäköisyydet ovat hyvin lähellä toisiaan.

ODOTUSARVON LUOTTAMUSVÄLI, KUN σ^2 EI OLE TUNNETTU JA OTOS EI OLE NORMAALIJAKAUMASTA

Oletetaan, että tutkimuksen kohteena olevaa perusjoukon ominaisuutta kuvaavan satunnaismuuttujan X odotusarvo $E(X) = \mu$ ja varianssi $D^2(X) = \sigma^2$. Nyt ei siis tehdä normaalisuusoletusta. Olkoot $X_1, X_2, \ldots, X_n$ yksinkertainen ko. perusjoukosta poimittu satunnaisotos ja olkoon

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

havaintojen aritmeettinen keskiarvo sekä

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

havaintojen otosvarianssi. Kappaleesta 3.1.2 tiedetään, että otoskeskiarvo $\overline{X}$ on parametrin μ ja otosvarianssi s^2 on parametrin σ^2 harhaton estimaattori.

Tällaisessa tilanteessa luottamusväli voidaan konstruoida keskeisen raja-arvolauseen avulla (kts. kappale 2.3.2): Otoskeskiarvo $\overline{X}$ on tiettyjen hyvin yleisten ehtojen pätiessä suurissa otoksissa approksimatiivisesti jakautunut kuten normaalijakauma $N(\mu,\sigma^2/n)$, vaikka havainnot eivät olisikaan normaalijakautuneita. Tällöin standardoitu muuttuja

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim_a N(0,1).$$

On syytä kuitenkin muistaa, että pienissä, ei-normaalisista perusjoukoista poimituissa otoksissa saattaa olla parempi käyttää jotakin muuta tunnuslukua kuin otoskeskiarvoa jakauman paikan estimointiin.

Määritellään väli

$$\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}},$$

jossa $-z_{\alpha/2}$ on piste, joka erottaa standardoidun normaalijakauman vasemmalle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Symmetrian takia $+z_{\alpha/2}$ on piste, joka erottaa standardoidun normaalijakauman oikealle hännälle todennäköisyysmassan, jonka suuruus on $\alpha/2$. Siten $z_{\alpha/2}$ on luottamustasoon $1-\alpha$ liittyvä luottamuskerroin. Jos siis $\Phi(z)$ on standardoidun normaalijakauman kertymäfunktio niin,

$$\Phi(-z_{\alpha/2}) = \alpha/2,$$

$$\Phi(+z_{\alpha/2}) = 1 - \alpha/2.$$

Tällöin ko. väli on parametrin μ *approksimatiivinen* luottamusväli luottamustasolla $1-\alpha$. Väli siis peittää tuntemattoman parametrin arvon μ *likimäärin* todennäköisyydellä $1-\alpha$: *Approksimatiivisesti* pätee

$$P(\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha.$$

Luottamusväli on symmetrinen keskipisteensä $\bar{X}$ suhteen ja sekä välin keskipiste että pituus vaihtelevat satunnaisesti otoksesta toiseen. Eri otoksista lasketuista väleistä approksimatiivisesti $100\times(1-\alpha)\%$ kuitenkin peittää tuntemattoman μ:n arvon.

Todettakoon lopuksi, että luottamusvälin pituudella on seuraavat ominaisuudet:

Luottamusväli *kapenee*, jos

- havaintojen lukumäärä n kasvaa,
- luottamustasoa $1-\alpha$ pienennetään,
- otoshajonta s pienenee.

Luottamusväli *levenee*, jos

- luottamustasoa $1-\alpha$ kasvatetaan,
- otoshajonta s kasvaa.

ESIMERKKI 1.

Eräässä markkinatutkimuksessa haluttiin selvittää paljonko pienyritykset käyttävät keskimäärin rahaa mainontaan. Tutkimus suoritettiin kysymällä 30:ltä satunnaisesti valitulta pienyritykseltä mainonnan vuotuisia kustannuksia.

Vastauksista voitiin laskea seuraavat tunnusluvut:

$$\bar{X} = 45{,}000 \text{ mk},$$

$$s^2 = (9{,}000 \text{ mk})^2.$$

Mikä on 95%:n luottamusväli mainonnan todellisille vuosikustannuksille pienyrityksissä?

Määritellään satunnaismuuttuja

$$\frac{\bar{X}-\mu}{s/\sqrt{n}},$$

jossa μ on satunnaisesti valitun yrityksen odotettavissa olevat mainonnan odotettavissa olevat vuosikustannukset. Jos oletamme, että kustannukset jakaantuvat normaalijakauman mukaan, tämä satunnaismuuttuja on jakaantunut kuten Studentin t-jakauma vapausastein, joiden luku on $n-1$.

Siten voimme määrätä t-jakaumasta luottamuskertoimet $-t_{2.5}$ ja $+t_{2.5}$ siten, että

$$P(-t_{2.5} \le \frac{\bar{X}-\mu}{s/\sqrt{n}} \le +t_{2.5}) = 0.95,$$

josta saadaan 95%:n luottamusväli parametrille μ:

$$\bar{X} \pm t_{2.5}\frac{s}{\sqrt{n}} = 45000 \pm 2.045\frac{9000}{\sqrt{30}} = 45000 \pm 3360$$

eli

41,640 mk, 48,360 mk.

Jos otos on ollut edustava ja mainonnan vuosikustannukset jakautuvat normaalijakauman mukaan perusjoukossa, tämä väli peittää satunnaisesti valitun pienyrityksen mainonnan vuosikustannusten odotusarvon μ todennäköisyydellä 0.95.

Jos perusjoukkoa ei voida olettaa normaaliseksi, luottamusväli joudutaan konstruoimaan vetoamalla keskeiseen raja-arvolauseeseen. Jos keskeistä raja-arvolausetta voidaan soveltaa, approksimatiivinen 95%:n luottamusväli parametrille μ on

$$\overline{X} \pm z_{.25} \frac{s}{\sqrt{n}} = 45000 \pm 1.960 \frac{9000}{\sqrt{30}} = 45000 \pm 3220$$

eli

41,780 mk, 48,220 mk.

Tämä väli on hieman kapeampi kuin t-jakaumaan perustuva väli. Tämä väli peittää todelliset mainonnan kustannukset μ perusjoukossa approksimatiivisesti todennäköisyydellä 0.95. Sama väli saataisiin myös, jos varianssi olisi tunnettu ja se yhtyisi havaittuun varianssiin. ●

ESIMERKKI 2.

Kuulalaakereita valmistettaessa on tärkeätä ottaa huomioon kuulien halkaisijoiden *toleranssi*. Toleranssilla tarkoitetaan seuraavaa: Kuulien halkaisijoiden on oltava tietyllä välillä, jotta laakerit toimisivat kunnolla. Siksi laakereita valmistavan tehtaan laaadunvalvontaosasto on mitannut 100 satunnaisesti valitun kuulan halkaisijan.

Oletetaan, että otoskeskiarvo

$$\overline{X} = 0.8420 \text{ cm}$$

ja otoshajonta

$$s = 0.0012 \text{ cm}.$$

Tehdään oletus, että kuulan halkaisija on (ainakin approksimatiivisesti) normaalijakaumaa noudattava satunnaismuuttuja. Teollisuustuotteiden ominaisuuksista voidaan usein tehdä perustellusti normaalisuusoletus. Tämä johtuu siitä, että tuotteiden ominaisuudet vaihtelevat valmistusprosessissa satunnaisesti tuotteesta toiseen ja vaihtelun voidaan olettaa kumuloituvan useasta osatekijästä. Normaalisuusoletus nojaa siten keskeiseen raja-arvolauseeseen (kts. kappale 1.3.2).

Jos kuulien halkaisijoiden normaalisuusoletus on järkevä, 99.9%:n luottamusväli satunnaisesti valitun kuulan odotettavissa olevalle halkaisijalle μ on

$$\overline{X} \pm t_{0.0005} \frac{s}{\sqrt{n}} = 0.8420 \pm 3.291 \frac{0.0012}{\sqrt{100}} = 0.8420 \pm 0.000395$$

eli

0.841605 cm, 0.842395 cm.

Tämä väli siis peittää satunnaisesti valitun kuulan odotettavissa olevan halkaisijan μ todennäköisyydellä 0.999. ●

Täydennyksenä edelliseen esimerkkiin todettakoon, että tekniikan sovelluksissa on usein tapana ilmoittaa muotoa

$$\bar{X} \pm \frac{s}{\sqrt{n}}$$

oleva luottamusväli. Normaalijakautuneen aineiston tapauksessa tiedetään, että tämä väli peittää todennäköisyydellä 0.68 tutkittavan satunnaismuuttujan odotettavissa olevan arvon. Suuretta $s / \sqrt{n}$ kutsutaan usein *odotettavissa olevaksi virheeksi*.