Tilastolliset menetelmät: Johdanto

1. **Tilastotiede tieteenalana**
2. **Tilastollisten aineistojen kerääminen ja mittaaminen**
3. **Tilastollisten aineistojen kuvaaminen**

Sisällys

1. Tilastotiede tieteenalana

1.1. Mitä tilastotiede on?

1.2. Tilastotieteen sovelluskohteet

Tässä luvussa yritämme vastata seuraaviin kysymyksiin:

- **Mitä tilastotiede on ja mitä se ei ole**?
- **Mihin tilastotiedettä käytetään**?

Tilastotiede *on* **yleinen menetelmätiede**, jota sovelletaan, jos **reaalimaailman ilmiöstä halutaan tehdä johtopäätöksiä ilmiötä kuvaavien kvantitatiivisten** tai **numeeristen tietojen perusteella** sellaisissa tilanteissa, joissa **tietoihin liittyy epävarmuutta** tai **satunnaisuutta**.

Tilastototiede *ei ole oppi* **tilastoista**, vaikka niiden menetelmien konstruointi, joilla tilastoja tuotetaan, jalostetaan ja analysoidaan onkin keskeinen osa tilastotiedettä. Tilastotiede *ei ole* myöskään **matematiikan** *osa-alue*, vaikka tilastotieteen menetelmät ja mallit ovatkin matemaattisia.

Tilastotiede kehtitää **matemaattisia malleja satunnaisilmiöitä kuvaavia kvantitatiivisia tietoja generoiville prosesseille**. Koska tietoihin liittyy epävarmuutta tai satunnaisuutta, **tilastolliset mallit** perustuvat **todennäköisyyslaskentaan**.

Tarkastelemme tässä luvussa myös **tilastotieteen sovelluskohteita** sekä selvennämme käsitteitä **tilastotiede** ja **tilastot**.

Avainsanat:

Arvonta, Empiirinen tutkimus, Epävarmuus, Havaintoaineisto, Johtopäätösten tekeminen, Koe, Kuvaileva tilastotiede, Kvantitatiivinen tieto, Kyselytutkimus, Laadunvalvonta, Lääketieteellinen koe, Matemaattinen malli, Matematiikka, Menetelmätiede, Numeerinen tieto, Otanta, Päätöksenteko, Reaalimaailman ilmiö, Satunnaisilmiö, Satunnaisuus, Soveltava tilastotiede, Teoreettinen tilastotiede, Tieto, Tilasto, Tilastoala, Tilastollinen aineisto, Tilastollinen päättely, Tilastollinen tutkimus, Tilastollinen tutkimusasetelma, Tilastollinen malli, Tilastollinen menetelmä, Tilastotiede, Tilastotieteen osa-alue,Tilastotieteen reuna-alue, Tilastotoimi, Todennäköisyyslaskenta, Tulosvaihtoehto, Tunnusluku, Tutkimus, Tutkimusaineisto, Yhteiskunta

1.1. Mitä tilastotiede on?

Mitä tilastotiede *ei ole*?

Tieteenalan määritteleminen on aina vaikeata. Esitämme kaiken uhallakin ensin pari *negatiivista* luonnehdintaa tilastotieteelle.

Tilastotiede **ei ole** − nimestään huolimatta − **oppi tilastoista** tai **tilastojen tuotannosta**! Mutta sen sijaan on totta, että *tilastojen tuotannon, jalostuksen ja analysoinnin menetelmien kehittäminen muodostaa keskeisen osan tilastotiedettä.*

Tilastotiede **ei ole** myöskään **matematiikan osa-alue**! Mutta sen sijaan on totta, että *tilastotieteen menetelmät ja mallit ovat matemaattisia ja perustuvat* **todennäköisyyslaskentaan**:

- Matematiikalla on tilastotieteessä *välineellinen rooli*.
- Tilastotiede käyttää *matematiikan kieltä*.

Mitä tilastotiede *on*?

Tehdään sitten muutama *positiivinen* määrittely-yritys:

Tilastotiede on **yleinen menetelmätiede**. Tilastotiede kehittää ja soveltaa menetelmiä ja malleja, joiden avulla *reaalimaailman ilmiöistä* voidaan tehdä **johtopäätöksiä** ilmiöitä kuvaavien **numeeristen** tai **kvantitatiivisten tietojen** perusteella tilanteissa, joissa tietoihin liittyy **epävarmuutta** ja **satunnaisuutta**.

Tilastollisten menetelmien avulla reaalimaailman ilmiöitä kuvaavat numeeriset tai kvantitatiiviset *tiedot jalostetaan* sellaiseen muotoon, että ilmiöitä koskevat *johtopäätökset tulevat mahdollisiksi*. Tietojen jalostaminen merkitsee *tietojen tiivistämistä* **graafisiksi esityksiksi** ja **tunnusluvuiksi** sekä **tilastollisten mallien** rakentamista *tiedot generoineille prosesseille* tai *mekanismeille*.

Tilastollisissa tutkimusasetelmissa reaalimaailman ilmiöitä kuvaaviin numeerisiin tai kvantitatiivisiin tietoihin liittyy aina *epävarmuutta* ja *satunnaisuutta*. Reaalimaailman ilmiötä kuvaavien tietojen **tilastollinen analyysi** perustuu siihen, että tietoihin liittyvän epävarmuuden ja satunnaisuuden ajatellaan johtuvan *tiedot generoineesta prosessista* tai *mekanismista*. Epävarmuuden ja satunnaisuuden generoijana voi olla *ilmiö itse* tai ne voivat olla seurausta *menetelmästä, jolla tutkimuksen kohteet valitaan*.

Satunnaisilmiöt ja todennäköisyyslaskenta

Reaalimaailman ilmiö on **satunnaisilmiö**, jos seuraavat ehdot pätevät:

(i) Ilmiöllä on useita erilaisia **tulosvaihtoehtoja**.

(ii) **Sattuma** määrää mikä tulosvaihtoehdoista toteutuu.

(iii) Vaikka *ilmiön tulos vaihtelee* ilmiön toistuessa *satunnaisesti*, ilmiön tulosvaihtoehtojen *suhteellisten osuuksien jakauma käyttäytyy* **tilastollisesti stabiilisti**, kun ilmiön toistokertojen lukumäärä kasvaa.

Todennäköisyyslaskennan tehtävänä on tuottaa *matemaattisia malleja* satunnaisilmiöissä havaittavalle *tilastolliselle stabiliteetille*. Todennäköisyyslaskentaa käsitellään monisteessa **Todennäköisyyslaskenta**.

Satunnaisilmiöihin liittyy aina *ennustamattomuutta*. **Satunnaisilmiön yksittäistä tulosta *ei voida tietää etukäteen***. Satunnaisilmiöihin on kuitenkin liityttävä *säännönmukaisuutta*, jonka on tultava esille ilmiön toistuessa: **Vaikka satunnaisilmiön tulos vaihtelee satunnaisesti ilmiön toisto-kerrasta toiseen**, **ilmiön tulosvaihtoehtojen *suhteellisten osuuksien jakauman on käyttäydyttävä stabiilisti,* kun toistokertojen lukumäärä kasvaa**.

Esimerkki 1: Satunnaisilmiöitä.

Satunnaisilmiöistä kohdataan esimerkkejä mitä moninaisimmilla alueilla:

- Kvanttimekaniikan ja hiukkasfysiikan ilmiöt ovat perusluonteeltaan satunnaisia.
- Luonnontieteellisiin mittauksiin liittyvien mittausvirheiden syntymekanismit ovat (ainakin osittain) satunnaisprosesseja.
- Uhkapeleissä kuten *arpajaisissa*, lotossa, ruletissa, korttipeleissä ja noppapeleissä sattumalla on keskeinen rooli.
- Perinnöllisyys noudattaa sattuman lakeja.
- Eliöiden ominaisuuksien jakautuminen populaatiossa on satunnaista.
- Ihmisten, ihmisryhmien ja ihmisten muodostamien organisaatioiden sosiaalisessa ja taloudellisessa käyttäytymisessä on monia satunnaisia elementtejä.
- Teknisten prosessien tuloksien ominaisuudet jakautuvat satunnaisesti.

Tilastollisen tutkimusaineiston kerääminen satunnaisilmiönä

Voimme ajatella, että tilastollisen tutkimuksen kohteet on aina valittu **arpomalla**. Arvonta on *satunnaisilmiö*:

(i) Arvontaan liittyy aina *ennustamattomuutta*, koska *yksittäisen arvonnan tulosta ei voida tietää etukäteen*.

(ii) Arvonta noudattaa kuitenkin *todennäköisyyden lakeja*.

Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös *tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen*. Tutkimuksen kohteita kuvaavien tietojen käyttäytymisessä havaitaan kuitenkin arvontaa toistettaessa juuri sitä *säännönmukaisuutta*, jota kutsutaan **tilastolliseksi stabiliteetiksi**. Tämä säännönmukaisuus on tilastollisen tutkimuksen kohde.

Esimerkki 2: Tilastollisten aineistojen kerääminen arvontana.

Esimerkkejä tilastollisten aineistojen keräämisen menetelmistä, jotka perustuvat *arvontaan*:

- **Satunnaistetut kokeet**
- **Satunnaisotanta**

Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsittelyiden vaikutuksia kokeen kohteisiin. Erilaisten virhelähteiden kontrolloimiseksi käsittelyt on syytä *arpoa* kohteille.

Otannalla tarkoitetaan laveasti *tutkimusaineistojen keräämisen menetelmiä*. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on syytä valita *arpomalla*.

Teoreettinen ja soveltava tilastotiede

Teoreettinen tilastotiede kehittää *matemaattisia malleja* prosesseille, jotka *generoivat* reaalimaailman ilmiöitä kuvaavia numeerisia tai kvantitatiivisia *tietoja, joihin liittyy epävarmuutta ja satunnaisuutta*. Teoreettisen tilastotieteen kehittämät mallit perustuvat *todennäköisyyslaskentaan* ja niitä kutsutaan **tilastollisiksi malleiksi**, **stokastisiksi malleiksi** tai **todennäköisyysmalleiksi**. Tilastollisten mallien avulla reaalimaailman ilmiöitä kuvaaviin tietoihin liittyvät *systemaattiset* ja *satunnaiset piirteet* voidaan *erottaa* ja *kuvata*.

Soveltava tilastotiede soveltaa teoreettisen tilastotieteen kehittämiä matemaattisia malleja reaalimaailman ilmiöitä kuvaavien numeeristen tai kvantitatiivisten *tietojen analysointiin*.

Teoreettinen ja soveltava tilastotiede kulkevat tilastollisessa tutkimuksessa käsi kädessä:

- Teoreettinen tilastotiede *kehittää tilastomatemaattisia malleja* soveltavan tilastotieteen empiiristen ongelmien ratkaisemiseksi.
- Soveltava tilastotiede *käyttää hyväkseen teoreettisen tilastotieteen kehittämiä malleja*.

Kuvaileva tilastotiede ja tilastollinen päättely

Deskriptiivinen eli **kuvaileva tilastotiede** kehittää ja soveltaa menetelmiä, joiden avulla tutkimuksen kohteena olevasta ilmiöstä kerättyjä numeerisia tai kvantitatiivisia tietoja voidaan *kuvailla* ja *esitellä*.

Kuvailevan tilastotieteen työkaluja:

- **Tilastografiikka**
- **Tilastolliset tunnusluvut**
- **Tilastolliset mallit**

Tilastollinen inferenssi eli **päättely** kehittää ja soveltaa menetelmiä, joiden avulla tutkimuksen kohteena olevasta ilmiöstä voidaan *tehdä johtopäätöksiä* ilmiöstä kerättyjen numeeristen tai kvantitatiivisten tietojen perusteella.

Tilastollisen päättelyn työkaluja:

- **Tilastolliset mallit**
- **Tilastollinen testaus**

Kuvaileva tilastotiede ja tilastollinen päättely kulkevat tilastollisessa tutkimuksessa käsi kädessä.

Tilastotieteen osa-alueita

Tilastotiede jakautuu moniin osa-alueisiin. Osa-alueita on niin paljon, että ammatti-tilastotieteilijäkään ei voi hallita niitä kaikkia.

Esimerkki 3: Tilastotieteen osa-alueita.

> Aikasarja-analyysi, Bayeslaiset menetelmät, Biometria, Demometria, Ei-parametriset menetelmät(1), Ekonometria, Estimointi(1), Kemometria, Koesuunnittelu, Kuvaileva tilastotiede(1), Laadunvalvonta, Lineaariset mallit(1), Matemaattinen tilastotiede, Monimuuttujamenetelmät, Otantateoria, Regressioanalyysi(1), Robustit menetelmät, Spatiaaliset menetelmät, Testaus(1), Tilastografiikka(1), Tilastollinen päättely(1), Tilastollinen tietojenkäsittely, Varianssianalyysi(1)

[(1)] Tässä monisteessa käsiteltäviä aiheita.

Tilastotieteen reuna-alueita

Monet tieteenalat ovat rikastaneet tilastotiedettä ja/tai niissä sovelletaan rutiininomaisesti tilastollisia menetelmiä.

Esimerkki 4: Tilastotieteen reuna-alueita.

> Finanssimatematiikka, Hahmontunnistus, Hermoverkot, Kaaosteoria, Katastrofiteoria, Kuvankäsittely, Kybernetiikka, Operaatioanalyysi, Peliteoria, Päätösteoria, Riskiteoria, Signaalinkäsittely, Stokastiset prosessit, Todennäköisyyslaskenta, Tulevaisuudentutkimus, Vakuutusmatematiikka

1.2. Tilastotieteen sovelluskohteet

Tilastotiedettä voidaan **yleisenä menetelmätieteenä** soveltaa – ja myös pitäisi soveltaa – kaikkialla, missä *tuotetaan* reaalimaailmaa ja sen ilmiöitä kuvaavaa *numeerista* tai *kvantitatiivista tietoa*.

Tilastollisia menetelmiä voidaan soveltaa tietojen *keruun*, *jalostuksen* ja *analysoinnin* jokaisessa vaiheessa. Tilastollisia menetelmiä sovellettaessa päämääränä on jalostaa tiedot muotoon, joka *mahdollistaa* reaalimaailmaa ja sen ilmiöitä koskevien *johtopäätösten tekemisen*.

Tilastotiedettä voidaan *yleisenä menetelmätieteenä* soveltaa kaikissa tieteissä, joiden **tutkimusaineistot** voidaan esittää *numeerisessa* tai *kvantitatiivisessa* muodossa. Jokainen tiede, jonka tutkimusaineistot voidaan esittää numeerisessa tai kvantitatiivisessa muodossa *voi soveltaa* / *voisi soveltaa* / *pitäisi soveltaa* tilastollisia menetelmiä sekä tutkimusaineistoja *kerättäessä* että niitä *analysoitaessa*. Siten jokainen **empiirisen tutkimuksen** *havaintoaineisto* on tilastollisen tutkimuksen mahdollinen kohde.

Esimerkki 1: Tilastotieteen sovelluskohteita eri tieteenaloilla.

- *Biotieteet*:

  Biokemia, Biologia, Ekologia, Eläinlääketiede, Eläintiede, Farmakologia, Kasvitiede, Lääketiede, Perinnöllisyystiede

- *Ihmistieteet*:

  Arkeologia, Kielitiede, Psykologia

- *Luonnontieteet*:

  Fysiikka, Kemia, Tähtitiede

- *Maatalous-* ja *metsätieteet*:

  Kasvinviljelytiede, Kotieläinten jalostustiede, Metsänarviointitiede, Metsänviljelytiede

- *Tekniset tieteet*:

  Hahmontunnistus, Kalibrointi, Koesuunnittelu, Kuvankäsittely, Laadunvalvonta, Laskennallinen tekniikka, Lääketieteellinen tekniikka, Neuroverkot, Päätöksenteko-menetelmät, Prosessinvalvonta, Signaalinkäsittely, Spektroskopia, Tietoliikenne-tekniikka

- *Yhteiskuntatieteet*:

  Sosiaalitieteet, Taloustiede

Esimerkki 2: Tilastotieteen eksoottisia sovelluksia: Dendrokronologia.

Arkeologiassa puuesineiden *ajoituksessa* voidaan käyttää apuna mm. puiden vuosilustojen muodostamia *aikasarjoja*. Myös *ilmastonmuutoksien tutkimuksessa* voidaan käyttää apuna puiden vuosilustojen muodostamia *aikasarjoja*.

Kummassakin esimerkissä puiden vuosilustosarjojen analysoinnissa voidaan soveltaa **dendrokronologiaksi** kutsuttua tilastollisten menetelmien kokonaisuutta. Dendrokronologian menetelmät ovat läheistä sukua tilastollisen *aikasarja-analyysin* menetelmille.

Esimerkki 3: Tilastotieteen eksoottisia sovelluksia: Tietokonetomografia.

Lääketieteellisissä tutkimuksissa käytetään (esim. syöpäkasvaimia etsittäessä) apuna **tietokonetomografiaa** eli **viipalekuvausta**. Siinä ihmisen kudoksista tai elimistä tuotetaan ns. *viipalekuvia* laitteella, joka mittaa sähkömagneettisen tai hiukkassäteilyn muuttumista säteilyn kulkiessa kudosten tai elinten läpi.

Kuvaa muodostavaan laitteeseen on ohjelmoitu algoritmi, joka ratkaisee *inversio-ongelmaksi* kutsutun matemaattisen ongelman. Inversio-ongelman ratkaisumenetelmät voidaan luontevimmin tulkita *bayeslaisten* tilastollisten menetelmien muodostamassa kehikossa.

Tilastotieteellä on monta nimeä

Tilastotieteestä käytetään monessa tieteessä omaa erityistä nimeä.

Esimerkki 4: Tilastotieteen nimiä eri tieteenaloilla.

Biometria tai *Biostatistiikka*	=	Bio- ja lääketieteiden tilastotiede
Demometria	=	Väestötiede
Ekonometria	=	Taloustieteen tilastotiede
Epidemiologia	=	Tautien leviämismekanismeja koskeva lääketieteen osa-alue
Kemometria	=	Kemian tilastotiede

Tilastotiede ja tilastot

Sana **tilasto** tuo useimmille ensimmäisenä mieleen *yhteiskuntaa* ja *sen toimintaa* kuvaavat *numeeristen tietojen järjestelmälliset kokoelmat*. Yhteiskuntaa ja sen toimintaa kuvaavien *tilastojen tuotannossa* ja *analysoinnissa tarvittavien menetelmien kehittäminen* on keskeinen osa tilasto-tiedettä, mutta on syytä huomata, että tilastotieteen sovellusalue on paljon tätä laajempi.

Tilastotieteen kannalta *mikä tahansa* reaalimaailman ilmiötä kuvaava *numeeristen* tai *kvantitatiivisten tietojen järjestelmällinen kokoelma* muodostaa **tilastollisen aineiston** ja siten **tilastollisen tutkimuksen** mahdollisen kohteen. Esimerkiksi kaikki empiirisen tai kvantitatiivisen tutkimuksen tutkimus- tai havaintoaineistot ovat tilastotieteen kannalta tilastollisia aineistoja.

Terminologiaa:

Tilastoala	=	Tilastotiede + Tilastotoimi
Tilastotiede	=	Teoreettinen tilastotiede + Soveltava tilastotiede
Tilastotoimi	=	Tilastojen tuotanto + Tilastojen hyödyntäminen

Tilastotiede, tilastot ja yhteiskunta

Ihminen ei voi toimia maailmassa järkevästi, ellei hän pysty muodostamaan *oikeata kuvaa maailmasta* ja sen *tilasta*. Nykyaikana oikeata kuvaa varten tarvitaan maailmaa ja sen tilaa *merkityksellisesti ja oikein kuvaavia, ajantasaisia* (*tilasto-*) *tietoja*. Merkityksellisesti ja oikein todellisuutta kuvaavat, ajantasaiset (*tilasto-*) *tiedot* ovat *välttämättömiä* modernin yhteiskunnan toiminnalle ja niiden saatavuutta voidaan pitää jopa toimivan *demokratian edellytyksenä*.

Yhteiskunnan kaikilla sektoreilla toiminnan seuranta, päätöksenteko ja *ennakointi* perustuvat sekä yhteiskunnan eri sektoreita kuvaaviin (*tilasto-*) *tietoihin* että *tilastollisiin menetelmiin*. *Päätöksenteko* sekä *julkisella* että *yksityisellä sektorilla* (*elinkeinoelämässä*) perustuu hyvin pitkälti yhteiskuntaa ja elinkeinoelämää kuvaaviin (*tilasto-*) *tietoihin* ja *tilastollisiin menetelmiin*. Esimerkiksi *tuotantoprosessien ohjaus* ja *laadunvalvonta* teollisuudessa sekä *markkinatutkimus* kaupan alalla perustuvat tilastollisiin menetelmiin.

Koska todellisuutta kuvaaviin (*tilasto-*) *tietoihin* sisältyy (lähes) aina *epävarmuutta* ja *satunnaisuutta*, tilastotiede ja tilastolliset menetelmät luovat perustan *tilastojen tuotannolle, jalostukselle* ja *analysoinnille*. Niinpä tilastojen tuotannon, jalostuksen ja analysoinnin *menetelmien kehittäminen* on keskeinen osa tilastotieteen tehtäväkenttää.

Esimerkkejä tilastollisista tutkimusasetelmista

Esimerkki 5: Kyselytutkimukset.

Päätöksentekijät ja tiedotusvälineet kartoittavat säännöllisien välein suomalaisten mielipiteet erilaisista yhteiskuntaa koskevista kysymyksistä.

Esimerkkejä:

- Miten suomalaiset suhtautuvat mahdolliseen NATO-jäsenyyteen?
- Miten suomalaiset suhtautuvat ydinvoiman lisärakentamiseen?
- Mitkä ovat poliittisten puolueiden kannatusosuudet?

Mielipiteet selvitetään *kyselytutkimuksilla*, joiden kohteeksi poimitaan tyypillisesti 1000-2000 suomalaista. Kyselytutkimuksen **tavoitteena on tehdä kyselyn tulosten perusteella johtopäätöksiä mielipiteiden jakautumisesta *kaikkien* suomalaisten joukossa**.

Miten 1000-2000 suomalaiseen kohdistetun kyselyn tulokset voidaan *yleistää* koskemaan kaikkia suomalaisia?

- Kyselyn tulokset voidaan yleistää, jos kyselyn kohteiksi poimittujen suomalaisten joukko muodostaa *edustavan pienoiskuvan* Suomen kansasta.
- Pienoiskuva on *edustava*, jos mielipiteet jakautuvat kyselyn kohteiksi poimittujen joukossa *samalla tavalla* kuin kaikkien suomalaisten muodostamassa *perusjoukossa*.
- Kyselyn kohteiden *poiminta arpomalla* on ainoa menetelmä, joka mahdollistaa edustavan pienoiskuvan saamisen.
- Kyselyn kohteiden *poimintaa* kaikkien suomalaisten muodostamasta *perusjoukosta arpomalla* kutsutaan tilastotieteessä (**satunnais-**) **otannaksi** ja tutkimuksen kohteeksi poimittua perusjoukon osaa kutsutaan (**satunnais-**) **otokseksi**.

Arvonnan käyttö kyselyn kohteiden poiminnassa merkitsee sitä, että *kyselyn tulokset ovat* **satunnaisia** seuraavassa mielessä: *Jos arvontaa toistettaisiin, kysely tuottaisi* (*suurella toden-*

näköisyydellä) *joka kerran* (*ainakin jonkin verran*) *erilaiset tulokset, koska eri arvonnoissa kyselyyn poimittaisiin* (*suurella todennäköisyydellä*) *eri henkilöt*.

Kysymyksiä:

- Miten *yhdestä otoksesta saadut* ja *satunnaiset* kyselytulokset voidaan *yleistää* koskemaan koko sitä *perusjoukkoa*, josta *otos* poimitaan?
- Miten *luotettava* tällainen yleistys on?

Vastauksia:

- Jos kyselyn kohteiden poiminnassa on käytetty *satunnaisotantaa*, kyselyn tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa *tilastollinen malli*, joka mahdollistaa sekä *kyselyn tulosten yleistämisen* että *yleistyksen luotettavuuden arvioimisen*.
- Yleistyksen luotettavuutta ei pystytä arvioimaan, *ellei otoksen poiminnassa ole käytetty satunnaisotantaa*.
- Kyselytutkimusten *suunnittelussa*, *toteutuksessa* ja *tulosten analysoinnissa* sovelletaan mm. seuraavia tilastollisia menetelmiä: **otanta**, **estimointi** ja **testaus**.

Esimerkki 6: Lääketieteelliset kokeet.

Erään tappavan taudin hoitoon on kehitetty *uusi lääke*, jonka toivotaan parantavan enemmän potilaita kuin kauan käytössä ollut *vanha lääke*. Miten saadaan *varmuus* siitä, että uusi lääke on parempi kuin vanha lääke?

Paranemistulosten *vertailemiseksi* järjestetään *tilastollinen koe*:

(i) Jaetaan joukko potilaita *arpomalla* kahteen ryhmään:

Ryhmälle 1 annetaan *uutta* lääkettä.

Ryhmälle 2 annetaan *vanhaa* lääkettä.

(ii) *Verrataan* parantuneiden *suhteellisia osuuksia* ryhmissä 1 ja 2.

Kokeen **tavoitteena on tehdä kokeen tulosten perusteella *yleisiä* johtopäätöksiä uuden lääkkeen tehokkuudesta**.

Miten yhdestä kokeesta saadut tulokset voidaan *yleistää* koskemaan kaikkia tautia sairastavia potilaita?

- Kokeen tulokset voidaan yleistää, jos kokeessa uutta ja vanhaa lääkettä saavien potilaiden ryhmät ovat *samankaltaisia* kaikissa muissa suhteissa paitsi siinä, että niihin kohdistetaan kokeessa *erilainen käsittely*.
- Tällöin mahdolliset *erot* parantuneiden suhteellisissa osuuksissa *on oltava seurausta erilaisista käsittelyistä*.
- Kokeen kohteiden *jakaminen ryhmiin arpomalla* on ainoa menetelmä, joka mahdollistaa samankaltaisten ryhmien saamisen.
- Kokeen kohteiden *jakamista* erilaisen käsittelyn kohteiksi joutuviin *ryhmiin arpomalla* kutsutaan tilastotieteessä **satunnaistamiseksi**.

Arvonnan käyttö ryhmiin jaossa merkitsee sitä, että *koetulokset ovat* **satunnaisia** seuraavassa mielessä: *Jos arvontaa toistettaisiin, kokeesta saataisiin* (*suurella todennäköisyydellä*) *joka*

kerran (*ainakin jonkin verran*) *erilaiset tulokset, koska eri arvonnoissa saataisiin* (*suurella todennäköisyydellä*) *erilaiset ryhmäjaot.*

Kysymyksiä:

- Miten *yhdestä kokeesta saadut* ja *satunnaiset* koetulokset voidaan *yleistää* koskemaan kaikkia ko. tautia sairastavia potilaita?
- Miten *luotettava* tällainen yleistys on?

Vastauksia:

- Jos potilaiden jaossa ryhmiin on käytetty *satunnaistamista,* kokeen tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa *tilastollinen malli,* joka mahdollistaa sekä *koetulosten yleistämisen* että *yleistyksen luotettavuuden arvioimisen.*
- Yleistyksen luotettavuutta ei pystytä arvioimaan, *ellei ryhmiin jaossa ole käytetty satunnaistamista.*
- Tilastollisen kokeen *suunnittelussa, toteutuksessa* ja *tulosten analysoinnissa* sovelletaan mm. seuraavia tilastollisia menetelmiä: **koesuunnittelu**, **estimointi** ja **testaus**.

Esimerkki 7: Laadunvalvonta.

Tehdas valmistaa korkealuokkaisia sulkimia kameroihin. Tehdas pyrkii siihen, että yli 90 % sulkimista kestää vähintään 100 000 kameran laukaisua. Sulkimien *laadun valvonta* on toteutettu seuraavalla tavalla:

(i) Tuotantolinjalta *poimitaan arpomalla* joukko sulkimia rasituskokeeseen.

(ii) Rasituskokeessa *määrätään* vähintään 100 000 laukaisua kestävien sulkimien *suhteellinen osuus*.

Kokeen **tavoitteena on tehdä kokeen tulosten perusteella *yleisiä* johtopäätöksiä sulkimien kestävyydestä**.

Miten vain osaan sulkimista kohdistetun rasituskokeen tulokset voidaan *yleistää* koskemaan kaikkia sulkimia?

- Kokeen tulokset voidaan yleistää, jos rasituskokeen kohteiksi poimittujen sulkimien joukko muodostaa *edustavan pienoiskuvan* kaikista valmistetuista sulkimista.
- Pienoiskuva on *edustava,* jos sulkimien kesto jakautuu rasituskokeeseen poimittujen sulkimien joukossa *samalla tavalla* kuin kaikkien valmistettujen sulkimien muodostamassa *perusjoukossa*.
- Rasituskokeen kohteiden *poiminta arpomalla* on ainoa menetelmä, joka mahdollistaa edustavan pienoiskuvan saamisen.
- Rasituskokeen kohteiden *poimintaa* kaikkien valmistettujen sulkimien muodostamasta *perusjoukosta arpomalla* kutsutaan tilastotieteessä (**satunnais**-) **otannaksi** ja tutkimuksen kohteeksi poimittua perusjoukon osaa kutsutaan (**satunnais**-) **otokseksi**.

Arvonnan käyttö rasituskokeen kohteiden poiminnassa merkitsee sitä, että *koetulokset ovat* **satunnaisia** seuraavassa mielessä: *Jos arvontaa toistettaisiin, kokeesta saataisiin* (*suurella todennäköisyydellä*) *joka kerran* (*ainakin jonkin verran*) *erilaiset tulokset, koska eri arvonnoissa kokeeseen poimittaisiin* (*suurella todennäköisyydellä*) *eri sulkimet.*

Kysymyksiä:

- Miten *yhdestä kokeesta saadut* ja *satunnaiset* koetulokset voidaan *yleistää* koskemaan kaikkia sulkimia?
- Miten *luotettava* tällainen yleistys on?

Vastauksia:

- Jos rasituskokeen kohteiden poiminnassa on käytetty *satunnaisotantaa*, kokeen tuloksiin sisältyvälle epävarmuudelle ja satunnaisuudelle voidaan muodostaa *tilastollinen malli*, joka mahdollistaa sekä *koetulosten yleistämisen* että *yleistyksen luotettavuuden arvioimisen*.
- Yleistyksen luotettavuutta ei pystytä arvioimaan, *ellei kokeen kohteiden poiminnassa ole käytetty satunnaisotantaa*.
- Kokeen *suunnittelussa*, *toteutuksessa* ja *tulosten analysoinnissa* sovelletaan mm. seuraavia tilastollisia menetelmiä: **koesuunnittelu**, **otanta**, **estimointi** ja **testaus**.

2. Tilastolliset aineistot, niiden kerääminen ja mittaaminen

2.1. Tilastolliset aineistot ja niiden todennäköisyysmallit

2.2. Tilastolliset kokeet

2.3. Suorat havainnot

2.4. Kokonaistutkimus

2.5. Otantatutkimus

2.6. Satunnaistamisen merkitys

2.7. Mittaaminen, mitta-asteikot ja tilastolliset muuttujat

Tarkastelemme tässä luvussa **tilastollisia aineistoja** sekä niiden *keräämistä*.

Tilastollinen aineisto koostuu tutkimuksen kohteita ja niiden olosuhteita kuvaavien muuttujien **havaituista arvoista**.

Tilastollinen aineisto voi syntyä **tilastollisen kokeen** tuloksena tai se voi olla peräisin **suorista havainnoista**. Tutkimusta kutsutaan **kokonaistutkimukseksi**, jos tutkimuksen kohteeksi otetaan *kaikki* tutkimuksen mahdolliset kohteet. **Otantatutkimuksessa** tutkimuksen kohteiksi poimitaan mahdollisimman *edustava osa* tutkimuksen mahdollisten kohteiden joukosta.

Tilastollisten aineistoja kerättäessä tutkimuksen kohteet pitää poimia **satunnaisesti** eli kohteet on **arvottava** kaikkien mahdollisten kohteiden joukosta. Se, että tutkimuksen kohteet arvotaan, mahdollistaa **tilastollisten**, so. **todennäköisyyslaskentaan perustuvien matemaattisten mallien** soveltamisen tilastollisiin aineistoihin.

Tilastollinen aineisto kerätään **mittaamalla** tutkimuksen kohteiden ominaisuudet. Mittaustapahtumassa kohteiden ominaisuuksia vastaaville tilastollisille muuttujille annetaan niiden arvot. Tilastollisten muuttujien **mitta-asteikollisilla ominaisuuksilla** on syvällinen vaikutus tutkimusongelman ratkaisemiseen valittavaan tekniikkaan.

Avainsanat:

Binomijakauma, Diskreetti muuttuja, Havainto, Havaintoarvo, Havaintoyksikkö, Hypergeometrinen jakauma, Intervalliasteikko, Jatkuva muuttuja, Järjestysasteikko, Koe, Kokonaistutkimus, Kontrolloitu koe, Kvalitatiivinen muuttuja, Kvantitatiivinen muuttuja, Laatueroasteikko, Mittaaminen, Mitta-asteikko,
Mittari, Moniasteinen otanta, Nominaaliasteikko, Ordinaaliasteikko, Ositettu otanta, Otanta, Otantamenetelmä, Otanta palauttaen, Otanta palauttamatta, Otantatutkimus, Perusjoukko, Reliabiliteetti, Ryväsotanta, Satunnaisotanta, Satunnaistaminen, Suhdeasteikko, Suora havainto, Tarkkuus, Tilastollinen aineisto, Tilastollinen koe, Tilastollinen menetelmä, Tilastollisen aineiston kerääminen, Validiteetti, Välimatka-asteikko, Yksinkertainen satunnaisotanta

2.1. Tilastolliset aineistot ja niiden todennäköisyysmallit

Tilastolliset aineistot

Tilastollisen tutkimuksen *kaikki mahdolliset kohteet* muodostavat tutkimuksen (*kohde*-) **perusjoukon**. Tutkimuksen kohteita tarkastellaan aina jonkin *perusjoukon muodostamassa kehikossa*. Tutkimuksen kohteiksi valittuja *perusjoukon alkioita* kutsutaan **havaintoyksiköiksi**.

Tilastollinen aineisto koostuu havaintoyksiköiden *ominaisuuksia* ja *olosuhteita* kuvaavista *numeerisista* tai *kvantitatiivisista tiedoista*. Havaintoyksiköitä koskevia numeerisia tai kvantitatiivisia tietoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi**.

Tilastolliset tutkimusasetelmat

Tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina *epävarmuutta* ja *satunnaisuutta*.

Seurauksia:

(i) Tilastollisissa tutkimusasetelmissa ajatellaan, että *havaintoarvot on generoinut ilmiö, joka on luonteeltaan satunnainen.*

(ii) Tilastollisen tutkimuksen kohteita kuvaavat muuttujat tulkitaan tilastollisissa tutkimus-asetelmissa *satunnaismuuttujiksi* ja havaintoarvot tulkitaan näiden *satunnaismuuttujien realisoituneiksi arvoiksi*.

Tilastolliset mallit

Tilastollisella mallilla tarkoitetaan tutkimuksen kohteita kuvaavien satunnaismuuttujien *todennäköisyysjakaumaa, jonka ajatellaan generoineen ko. satunnaismuuttujien havaitut arvot*. Voimme myös ajatella, että havaintoarvot ovat syntyneet *arpomalla* tilastollisena mallina käytetystä todennäköisyysjakaumasta saatavin todennäköisyyksin.

Huomautus:

- Todennäköisyysjakaumat riippuvat tavallisesti *parametreista* eli *vakioista*, joiden arvoja ei yleensä tunneta.

Kun tilastollista mallia sovelletaan jotakin reaalimaailman ilmiötä kuvaavan havaintoaineiston analysointiin, kohdataan tavallisesti seuraavat mallin **parametreja** koskevat ongelmat:

(i) Parametrien arvoja *ei tunneta* ja ne on **estimoitava** eli *arvioitava* havaintoaineistosta.

Estimointia käsitellään luvuissa **Estimointi**, **Estimointimenetelmät** ja **Väliestimointi**.

(ii) Parametrien arvoista on esitetty *oletuksia* tai *väitteitä*, joita halutaan **testata** eli asettaa koetteelle havaintoaineistosta saatua informaatiota vastaan.

Testausta käsitellään luvuissa **Tilastollinen testaus**, **Testit suhdeasteikollisille muuttujille**, **Testit järjestysasteikollisille muuttujille**, **Testit laatueroasteikollisille muuttujille**, **Yhteensopivuuden, homogeenisuuden ja riippuvuuden testaaminen**.

Tilastollisten mallien *parametrien estimointi* ja *testaus* muodostavat keskeisen osan **tilastollista päättelyä**.

Tilastollisten aineistojen kerääminen

Muutetaanko tutkimuksessa tutkimuksen kohteiden olosuhteita *aktiivisesti*?

(i) Tutkimus on **koe**, jos tutkimuksen tavoitteena on selvittää, miten kohteiden olosuhteiden *aktiivinen muuttaminen* vaikuttaa tutkimuksen kohteisiin.

(ii) Tutkimus perustuu **suoriin havaintoihin**, jos tutkimuksen tavoitteena on *vain seurata*, miten kohteiden *olosuhteet* ja *niissä tapahtuvat muutokset* vaikuttavat kohteisiin.

Kohdistuuko tutkimus *kaikkiin* perusjoukon alkioihin vai johonkin perusjoukon *osaan*?

(i) Tutkimusta kutsutaan **kokonaistutkimukseksi**, jos *kaikki perusjoukon alkiot tutkitaan*.

(ii) Tutkimusta kutsutaan **otantatutkimukseksi**, jos *tutkimus kohdistuu johonkin perusjoukon osajoukkoon*.

2.2. Tilastolliset kokeet

Kokeellisen tutkimuksen tavoitteena on selvittää, *millaisia vaikutuksia erilaisilla käsittelyillä on kohteisiin*. **Käsittelyllä** tarkoitetaan tutkimuksen kohteiden olosuhteiden aktiivista, suunnitelmallista ja järjestelmällistä muuttamista. Tiukasti ottaen **vain kokeiden perusteella voidaan tehdä kausaalisia eli syy-yhteyksiä koskevia päätelmiä**.

Huomautus:

- Tutkimus perustuu *suoriin havaintoihin*, jos tutkimuksen kohteiden olosuhteita ei muuteta aktiivisesti; ks. kappaletta **Suorat havainnot**.

Koeasetelmat

Koeasetelmalla tarkoitetaan kokeen tekemiseen liittyviä periaatteita ja sääntöjä:

(i) **Mitä *käsittelyitä* kokeen kohteisiin sovelletaan**?

(ii) **Miten kokeen kohteet *valitaan***?

(iii) **Mikä on tehtävien *koetoistojen lukumäärä***?

Kontrolloidut kokeet

Kokeesta voidaan tehdä *luotettavia johtopäätöksiä* vain, jos koe on **kontrolloitu**:

(i) Koetuloksiin vaikuttavien **ulkopuolisten sekoittavien tekijöiden kontrolloimiseksi** kokeessa on **vertailtava** vähintään kahden erilaisen käsittelyn vaikutuksia.

(ii) Erilaisten käsittelyiden kohteiksi valittavien perusjoukon alkioiden välisten **systemaattisten erojen kontrolloimiseksi** käsittelyiden kohdistamisessa on käytettävä **satunnaistusta**.

(iii) Koetuloksiin liittyvän **satunnaisvaihtelun kontrolloimiseksi** kokeessa on tehtävä riittävästi **koetoistoja**.

Kutsumme kontrolloituja kokeita tavallisesti **tilastollisiksi kokeiksi**.

Huomautus:

- Tilastollisten kokeiden suunnittelua ja analysointia käsitellään monisteessa **Koesuunnittelu ja tilastolliset mallit**.

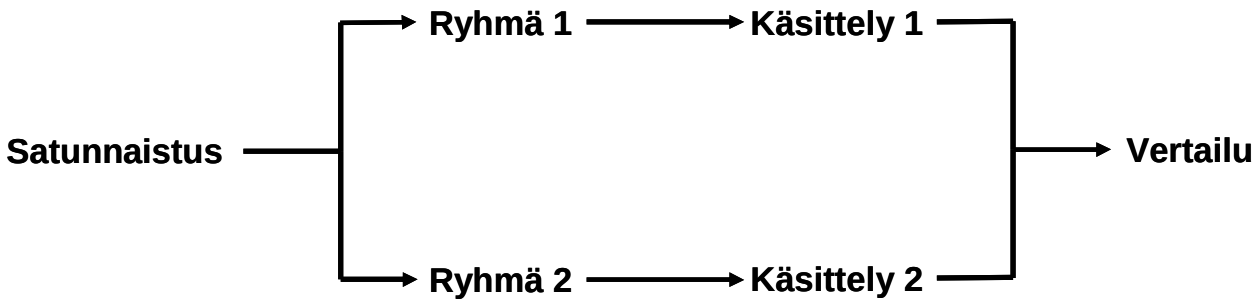
Esimerkki 1: Yksinkertainen kontrolloitu koe.

Ehkä yksinkertaisin mahdollinen kontrolloidun kokeen asetelma on seuraava:

(1) Jaetaan kokeen kohteet *satunnaisesti* kahteen ryhmään.

(2) Kohdistetaan ryhmiin *erilaiset* käsittelyt.

(3) *Vertaillaan* käsittelyiden vaikutuksia.



Kontrolloidut kokeet: Kommentteja

Oletetaan, että *koe on kontrolloitu* eli *kokeessa on sovellettu* suunnitelmallisesti ja järjestelmällisesti *vertailua*, *satunnaistusta* ja *koetoistoja*.

Tällöin:

(i) Koetuloksien analysointi tilastotieteen keinoin *on mahdollista*.

(ii) Koetuloksiin liittyvät systemaattiset ja satunnaiset tekijät *voidaan erottaa ja kuvata* ja kuvauksen luotettavuus *voidaan arvioida*.

(iii) Käsittelyiden vaikutuksista kokeen kohteisiin *voidaan tehdä luotettavia johtopäätöksiä*.

Oletetaan, että koe *ei ole kontrolloitu* eli kokeessa ei ole käytetty suunnitelmallisesti ja järjestelmällisesti *vertailua*, *satunnaistusta* ja *koetoistoja*.

Tällöin:

(i) Koetuloksien analysointi tilastotieteen keinoin ei ole mahdollista.

(ii) Koetuloksiin liittyviä systemaattisia ja satunnaisia tekijöitä *ei voida erottaa ja kuvata* ja kuvauksen luotettavuutta *ei voida arvioida*.

(iii) Käsittelyiden vaikutuksista kokeen kohteisiin *ei voida tehdä luotettavia johtopäätöksiä*.

Jos koe *ei ole kontrolloitu*, koeasetelma saattaa systemaattisesti suosia joitakin tulosvaihtoehtoja. Jos koeasetelma suosii systemaattisesti joitakin tulosvaihtoehtoja, asetelmaa sanotaan **harhaiseksi**. Harhaisten koeasetelmien perusteella *ei voida tehdä luotettavia johtopäätöksiä*.

Satunnaistaminen

Kokeen **satunnaistamisen** tarkoittaa sitä, että käsittelyiden kohdistamisessa käytetään **arvontaa**.

Arvonta on ainoa *puolueeton* tapa kohdistaa käsittelyitä, koska arpominen *ei suosi* mitään perusjoukon osaa. Se, että satunnaistettujen kokeiden tulosten analysointiin voidaan soveltaa tilastollisia menetelmiä perustuu siihen, että *arvonnassa noudatetaan todennäköisyyslaskennan lakeja*.

Satunnaistus takaa suurella varmuudella sen, että kokeessa erilaisten käsittelyiden kohteiksi joutuvat perusjoukon osajoukot ovat ennen käsittelyiden soveltamista ominaisuuksiltaan *samankaltaisia*. Juuri sen takia kokeen tuloksista voidaan tehdä kausaalipäätelmiä: **Jos koe on satunnaistettu**, **kokeen tuloksissa havaitut systemaattisten erojen on johduttava erilaisista käsittelyistä**.

2.3. Suorat havainnot

Suoriin havaintoihin perustuvassa tutkimuksessa tavoitteena on saada selville tutkimuksen kohteiden olosuhteisiin puuttumatta, mitä vaikutuksia kohteiden olosuhteilla ja niissä tapahtuvilla muutoksilla on kohteisiin. On syytä huomata, että tiukasti ottaen **suoriin havaintoihin perustuvien tutkimusten perusteella ei voida tehdä kausaalisia eli syy-yhteyksiä koskevia johtopäätöksiä**.

Huomautus:

- Tutkimus on koe, jos kohteiden olosuhteita muutetaan tutkimuksessa aktiivisesti; ks. kappaletta **Tilastolliset kokeet**.

Suorat havainnot: Kommentteja

Suoria havaintoja tehtäessä havaintojen tulokset saattavat olla *harhaisia*. Havaintojen tulokset ovat **harhaisia**, jos havaintoja tehtäessä suositaan systemaattisesti joitakin tulosvaihtoehtoja. Harhaisten havaintotulosten perusteella ei voida tehdä luotettavia johtopäätöksiä.

Harhan syntyminen saatetaan pystyä välttämään havaintoja tehtäessä, jos havaintojen kohteet valitaan perusjoukosta **satunnaisesti** (ellei tavoitteena ole tutkia kaikkia perusjoukon alkioita). Tämä merkitsee *satunnaisotannan* soveltamista havaintojen kohteiden valintaan; ks. kappaletta **Satunnaisotanta**.

2.4. Kokonaistutkimus

Tutkimus on **kokonaistutkimus**, jos se kohdistuu *kaikkiin* (kohde-) *perusjoukon alkioihin*.

Huomautuksia:

- Kokonaistutkimuksen tekeminen *on vain harvoin mahdollista*.
- Jos perusjoukko on *ääretön*, kokonaistutkimuksen tekeminen *on jopa periaatteessa mahdotonta*.
- Äärelliseen perusjoukkoon kohdistuvat kokonaistutkimukset *voidaan tulkita otanta-tutkimuksiksi*: Tällöin tutkimuksen kohteena oleva äärellinen perusjoukko tulkitaan otokseksi hypoteettisesta äärettömästä perusjoukosta.

2.5. Otantatutkimus

Tutkimus on **otantatutkimus**, jos se *kohdistuu johonkin perusjoukon osajoukkoon*. Otanta-tutkimuksessa perusjoukon osajoukosta tehdyt johtopäätökset pyritään *yleistämään* koko perusjoukkoon.

Tutkimuksen kohteeksi valittua perusjoukon osajoukkoa kutsutaan **otokseksi**. Otoksen valitsemista eli poimimista kutsutaan **otannaksi**. Otoksen poiminnassa käytettyjä menetelmiä kutsutaan **otantamenetelmiksi**.

Perusjoukosta voidaan tehdä luotettavia johtopäätöksiä otoksen perustella vain, jos otos muodostaa perusjoukon *edustavan* pienoiskuvan. Otoksen poimiminen perusjoukosta **satunnaisesti** takaa suurella todennäköisyydellä sen, että otos muodostaa perusjoukon edustavan pienoiskuvan.

Otoksen poiminta satunnaisesti merkitsee otokseen poimittavien havaintoyksiköiden **arpomista** perusjoukon alkioiden joukosta. Arvonta on ainoa *puolueeton tapa* poimia otos, koska arpominen *ei suosi* mitään perusjoukon osaa. Arvonnassa noudatetaan todennäköisyyslaskennan lakeja.

Otantamenetelmiä

Tilastollisessa tutkimuksessa sovelletaan tutkimusasetelmasta riippuen erilaisia otantamenetelmiä.

Otannan perusmuoto:

- **Yksinkertainen satunnaisotanta**

Muita otantamenetelmiä:

- **Systemaattinen otanta**
- **Ositettu otanta**
- **Ryväsotanta**
- **Moniasteinen otanta**

Yksinkertainen satunnaisotanta

Yksinkertainen satunnaisotanta on otannan perusmuoto, jossa *jokaisella perusjoukon alkiolla on yhtä suuri todennäköisyys tulla valituksi otokseen*. Jos otos poimitaan yksinkertaisella satunnaisotannalla, myös jokaisella perusjoukon samankokoisella osajoukolla on sama todennäköisyys tulla valituksi otokseksi.

Yksinkertainen satunnaisotanta voidaan aina tulkita **arvonnaksi**, joka on toteutettu seuraavalla tavalla:

(i) Alkiot arvotaan perusjoukosta otokseen yksi alkio kerrallaan.

(ii) Perusjoukkoon kuuluvilla alkioilla on jokaisessa arvonnassa yhtä suuri todennäköisyys tulla valituksi otokseen.

Yksinkertaisen satunnaisotannan perusmuodossa alkiot poimitaan perusjoukosta otokseen **palauttaen**: Poimittu alkio palautetaan aina ennen uuden alkion arpomista takaisin perusjoukkoon, jolloin sama alkio voi tulla poimituksi otokseen useita kertoja. Otannassa palauttaen arvonnat *ovat riippumattomia*: Alkion todennäköisyys tulla poimituksi otokseen ei riipu siitä mitä alkiota otokseen on jo poimittu. Otantaan palauttaen liittyviä todennäköisyyksiä hallitaan **binomijakauman** avulla; ks. monisteen **Todennäköisyyslaskenta** lukua **Diskreettejä jakaumia**.

Yksinkertaiseksi satunnaisotannaksi kutsutaan tavallisesti myös menetelmää, jossa alkiot poimitaan perusjoukosta otokseen **palauttamatta**: Poimittua alkiota ei palauteta ennen uuden alkion arpomista takaisin perusjoukkoon, jolloin smam alkio ei voi tulla poimituksi otokseen kuin kerran. Otannassa palauttamatta arvonnat *eivät ole riippumattomia*: Alkion todennäköisyys tulla poimituksi otokseen muuttuu arvonnan edistyessä. Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hypergeometrisen jakauman** avulla; ks. monisteen **Todennäköisyyslaskenta** lukua **Diskreettejä jakaumia**.

Systemaattinen otanta

Systemaattisessa otannassa otokseen poimitaan joka *k*. alkio perusjoukon alkioiden järjestetystä jonosta. Systemaattista otantaa sovelletaan usein yksinkertaisen satunnaisotannan sijasta, jos perusjoukon alkioista on käytettävissä tietorekisteri tai luettelo tai havaintoja kerätään ajassa tai tilassa.

Huomautuksia:

- Systemaattinen otanta *ei tiukasti ottaen kuulu satunnaisotannan menetelmiin*, koska siinä ei sovelleta arvontaa.

- Systemaattinen otanta tuottaa kuitenkin täysin samat tulokset kuin yksinkertainen satunnaisotanta, jos perusjoukon alkioiden järjestys on tutkittavan ilmiön kannalta satunnainen.

Ositettu otanta

Ositettua otantaa voidaan soveltaa tilanteissa, joissa perusjoukko koostuu jonkin perusjoukon alkioiden ominaisuuden suhteen *homogeenisista ryhmistä*. Tällöin otos kerätään siten, että jokaisesta ryhmästä eli *ositteesta* poimitaan **osaotos**, jotka yhdistetään yhdeksi otokseksi.

Esimerkki 1: Edustavuuden takaaminen ositetulla otannalla.

Oletetaan, että maassa on useita erikokoisia kieliryhmiä ja tavoitteena on vertailla eri kieliryhmiin kuuluvien taloudellista asemaa. Jokaisesta ryhmästä saadaan otokseen *riittävä edustus* poimimalla jokaisesta ryhmästä samankokoinen osaotos.

Ryväsotanta

Ryväsotantaa voidaan soveltaa tilanteissa, joissa perusjoukko voidaan jakaa **ryppäisiin** eli **ryhmiin**. Tällöin otos kerätään kahdessa vaiheessa:

(1) Poimitaan ensin joukko ryppäitä kaikkien ryppäiden joukosta.

(2) Poimitaan jokaisesta vaiheessa (1) poimitusta ryppäästä joukko perusjoukon alkioita ja yhdistetään alkiot yhdeksi otokseksi.

Huomautus:

- Vaiheissa (1) ja (2) voidaan soveltaa yksinkertaista satunnaisotantaa tai systemaattista otantaa.

Moniasteinen otanta

Moniasteista otantaa voidaan soveltaa tilanteissa, joissa perusjoukko voidaan jakaa **ryppäisiin** eli **ryhmiin hierarkkisesti** eli perusjoukko voidaan jakaa ryppäisiin, jotka puolestaan voidaan jakaa aliryppäisiin jne. Otos kerätään vaiheittain poimimalla 1. asteen ryppäiden joukosta joukko ryppäitä, joista jokaisesta poimitaan joukko aliryppäitä jne. kunnes päästään poimimaan perusjoukon alkioita.

Huomautus:

- Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaisotantaa tai systemaattista otantaa.

2.6. Satunnaistamisen merkitys tilastollisten aineistojen keräämisessä

Edellä on kuvattu seuraavia tilastollisten aineistojen keräämisen menetelmiä:

(i) **Kontrolloidut kokeet**

(ii) **Satunnaisotanta**

Kummassakin tapauksessa aineiston keräämisessä sovelletaan *arvontaa*. Arvonnan soveltaminen merkitsee seuraavaa:

Kaikki *tutkimuksen kohteita kuvaavat* (numeeriset tai **kvantitatiiviset) *tiedot* ja *myös niistä johdetut suureet ovat satunnaisia*. Tilastollisten – todennäköisyyslaskentaan perustuvien – mallien soveltaminen tilastollisten aineistojen analyysiin perustuu juuri tähän tosiseikkaan**.

2.7. Mittaaminen, mitta-asteikot ja tilastolliset muuttujat

Tilastollisen tutkimuksen *kohteiden ominaisuuksia* ja *olosuhteita* sekä niiden muutoksia kuvaavat *numeeriset* tai *kvantitatiiviset tiedot* saadaan selville *mittaamalla*. **Mittaaminen** tarkoittaa *numeeristen arvojen liittämistä* tutkimuksen kohteiden ominaisuuksiin ja olosuhteisiin.

Mittaria voidaan pitää *funktiona*, joka *liittää numeeriset arvot* tutkimuksen kohteiden ominaisuuksiin ja olosuhteisiin.

Mittari

Ominaisuus $\longrightarrow$ **Numeerinen arvo**

Mittauksen *tulos* voidaan aina ilmaista jonkin tutkimuksen kohteen ominaisuutta tai olosuhdetta kuvaavan *muuttujan arvona*.

Ominaisuus $\longleftrightarrow$ **Muuttuja**

Tilastollisessa tutkimuksessa tutkimuksen kohteiden ominaisuuksia ja olosuhteita kuvataan mittaustapahtumassa aina *numeerisilla muuttujilla*.

Mittarin validiteetti ja tarkkuus

Mittari on **validi** eli *oikea*, jos se esittää mittauksen kohteena olevaa ominaisuutta *oikein*, *merkityksellisesti* ja *tarkoituksenmukaisesti*.

Mittari on **tarkka**, jos se on *harhaton* ja *reliaabeli*:

(i) Mittari on **harhaton**, jos se *ei systemaattisesti ali- tai yliarvioi* mitattavan ominaisuuden määrää.

(ii) Mittari on **reliaabeli** eli *luotettava*, jos mittaustulos *ei muutu*, kun mittausta toistetaan.

Mitta-asteikot

(i) Mittaus on tehty **nominaali**- eli **laatueroasteikolla**, jos mittaus kertoo *mihin luokkaan* mittauksen kohde kuuluu.

Esimerkkejä:

Sukupuoli, Asuinpaikka, Väri, Viallisuus.

(ii) Mittaus on tehty **ordinaali**- eli **järjestysasteikolla**, jos mittaus kertoo onko mittauksen kohteella mitattavaa ominaisuutta *enemmän* tai *vähemmän* kuin jollakin toisella kohteella.

Esimerkkejä:

Kouluarvosanat, Aineen kovuus.

(iii) Mittaus on tehty **intervalli**- eli **välimatka-asteikolla**, jos mittaus kertoo *kuinka paljon* kahden mitattavan kohteen ominaisuudet *eroavat* toisistaan.

Esimerkkejä:

Lämpötila Celsius-asteissa.

(iv) Mittaus on tehty **suhdeasteikolla**, jos mittaus kertoo *kuinka monta kertaa enemmän* tai *vähemmän* mittauksen kohteella on mitattavaa ominaisuutta kuin jollakin toisella kohteella.

Esimerkkejä:

Lukumäärä, Pituus, Pinta-ala, Tilavuus, Paino, Aika, Nopeus, Paine, Rahamäärä, Korko

Huomautus:

- Jos ominaisuutta voidaan mitata kaikilla neljällä mitta-asteikoilla, *mittaustuloksen informatiivisuus*, mutta samalla myös *mittauksen vaativuus* kasvaa seuraavassa järjestyksessä:

  Nominaaliasteikko → Ordinaaliasteikko → Intervalliasteikko → Suhdeasteikko

Kvalitatiiviset ja kvantitatiiviset muuttujat

(i) Ominaisuutta ja sitä kuvaavaa muuttujaa kutsutaan **kvalitatiiviseksi**, jos mittauksen kohteet voidaan *luokitella* mittauksen perusteella toisistaan eroaviin *kategorioihin* tai *luokkiin*. Kvalitatiivisia ominaisuuksia kuvataan *laatueroasteikollisilla muuttujilla*.

(ii) Ominaisuutta ja sitä kuvaavaa muuttujaa kutsutaan **kvantitatiiviseksi**, jos mittaus tuottaa ominaisuuden *määrällisen arvon*. Kvantitatiivisia ominaisuuksia kuvataan *välimatka-* tai *suhde-asteikollisilla muuttujilla*.

Diskreetit ja jatkuvat muuttujat

(i) Mitattavaa ominaisuutta vastaava muuttuja on **diskreetti**, jos se voi saada vain erillisiä arvoja. Diskreettejä muuttujia ovat esimerkiksi kaikki laatueroasteikollisten ja järjestysasteikollisten muuttujien lisäksi myös sellaiset kvantitatiiviset muuttujat kuten lukumäärämuuttujat.

(ii) Mitattavaa ominaisuutta vastaava muuttuja on jatkuva, jos se voi saada kaikki arvot joltakin väliltä. Jatkuvia muuttujia ovat esimerkiksi useimmat fysikaaliset suureet kuten pituus, pinta-ala, tilavuus, paino, aika, nopeus ja paine sekä myös monet talouselämää kuvaavat suureet kuten rahamäärä ja korko.

Tilastollisten muuttujien mitta-asteikot ja tilastolliset menetelmät

Tilastollisten muuttujien mitta-asteikollisilla ominaisuuksilla tai kvalitatiivisuudella/ kvantitatiivisuudella tai diskreettiydellä/jatkuvuudella on syvällinen vaikutus tutkimus-ongelman ratkaisemisessa käytettävien tilastollisten menetelmien valintaan. **Esimerkiksi aineistojen kuvaamisen menetelmät ja tilastolliset testit on tässä esityksessä luokiteltu muuttujien mitta-asteikon mukaan**.

3. Tilastollisten aineistojen kuvaaminen

Tilastollinen aineisto koostuu tutkimuksen kohteita ja niiden olosuhteita kuvaavien muuttujien **havaituista arvoista**.

Käsittelemme tässä luvussa **tilastollisten aineistojen kuvaamista**. Luvussa esitellään sekä **graafisten menetelmien** että **tunnuslukujen** käyttöä *yksiulotteisten* tilastollisten aineistojen kuvaamisessa. Käsittely on jaettu osiin tarkasteltavan *muuttujan tyypin* tai *mitta-asteikollisten ominaisuuksien* mukaan. *Kaksiulotteisten* tilastollisten aineistojen kuvaamista käsitellään monisteen **Regressio- ja varianssi-analyysi** luvussa **Tilastollinen riippuvuus ja korrelaatio**.

Tilastollisia ainiestoja kuvattaessa on tärkeintä antaa kuva **havaintoarvojen jakaumasta**. Tämä tapahtuu parhaiten piirtämällä jakaumasta **kuva**. **Diskreetin muuttujan** tapauksessa havaintoarvojen jakaumaa kuvataan numeerisesti **frekvenssijakaumalla**, jota vastaavaa graafista esitystä kutsutaan **pylväsdiagrammiksi**. **Jatkuvan muuttujan** tapauksessa havaintoarvojen jakaumaa kuvataan numerisesti **luokitellulla frekvenssijakaumalla**, jota vastaavaa graafista esitystä kutsutaan **histogrammiksi**.

Tunnusluvuista tärkeimmät ovat havaintojen *keskimääräisiä, tyypillisiä* tai *yleisiä arvoja kuvaavat* **keskiluvut** ja havaintoarvojen *hajaantuneisuutta kuvaavat* **hajontaluvut**.

Avainsanat:

Aritmeettinen keskiarvo, Box and Whisker -kuvio, Frekvenssi, Frekvenssijakauma, Geometrinen
keskiarvo, Graafinen esitys, Hajontaluku, Harmoninen keskiarvo, Havaintoarvo, Havaintoarvojen jakauma, Histogrammi, Huipukkuus, Järjestystunnusluku, Keskihajonta, Keskiluku, Keskusmomentti, Kvartiili, Kvartiilipoikkeama, Kvartiiliväli ja kvartiilivälin pituus, Luokiteltu frekvenssijakauma, Luokka-frekvenssi, Maksimi, Mediaani, Mimimi, Mitta-asteikko, Moodi, Origomomentti, Prosenttipiste, Pylväs-diagrammi, Robustisuus, Standardointi, Suhteellinen frekvenssi, Tilastollinen aineisto, Tilastollinen etäisyys, Tunnusluku, Varianssi, Vinous, Vaihteluväli ja vaihteluvälin pituus

3.1. Tilastolliset aineistot

Kutsumme tilastollisen tutkimuksen *kaikkien mahdollisten kohteiden* muodostamaa joukkoa tutkimuksen (**kohde-**) **perusjoukoksi** ja kutsumme tutkimuksen kohteiksi valittuja perusjoukon alkioita **havaintoyksiköiksi**. **Tilastollinen aineisto** koostuu havaintoyksiköitä kuvaavien muuttujien **havaituista arvoista**.

Olkoon tutkimuksen kohteiksi valittujen **havaintoyksiköiden lukumäärä** n ja olkoon

$$x_i \, , \; i = 1, 2, \ldots , n$$

kohdeperusjoukon alkioiden ominaisuutta kuvaavan muuttujan x **havaittu arvo** havaintoyksikössä i. Kutsumme muuttujan x havaittuja arvoja

$$x_1 \, , \, x_2 \, , \, \ldots \, , \, x_n$$

havaintoarvoiksi tai **havainnoiksi**. Havaintoarvo x_i saadaan *mittaamalla* muuttujan x arvo havaintoyksikölle i.

Huomautuksia:

- Tilastollinen aineisto voi syntyä *tilastollisen kokeen* tuloksena tai tekemällä *suoria havaintoja*.
- Jos tutkimuksen kohteena on koko perusjoukko, tutkimusta kutsutaan *kokonais-tutkimukseksi*, jos vain satunnaisesti valittu osa perusjoukosta tutkitaan, kutsutaan tutkimusta *otantatutkimukseksi*.

Lisätietoja: Ks. lukua **Tilastolliset aineistot, niiden kerääminen ja mittaaminen**.

3.2. Havaintoarvojen jakauma

Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan *x havaittujen arvojen*

$$x_1 \, , \, x_2 \, , \, \ldots \, , \, x_n$$

vaihtelua havaintoyksiköiden joukossa kuvaa parhaiten havaintoarvojen **jakauma**.

Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan x havaittujen arvojen jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:

- Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valitulla **graafisella esityksellä**.
- *Jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **tunnus-luvuilla**.

Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan x (mitta-asteikolliset) ominaisuudet (ks. lukua **Tilastollisten aineistot, niiden kerääminen ja mittaaminen**) määräävät muuttujan x havaittujen arvojen jakaumalle parhaiten sopivan kuvaustavan:

- Jos muuttuja x on *diskreetti*, sen havaittujen arvojen jakaumaa voidaan kuvata **frekvenssijakaumalla** ja sitä vastaavalla graafisella esityksellä **pylväsdiagrammilla**.
- Jos muuttuja x on *jatkuva*, sen havaittujen arvojen jakaumaa voidaan kuvata **luokitellulla frekvenssijakaumalla** ja sitä vastaavalla graafisella esityksellä **histogrammilla**.

Frekvenssit ja frekvenssien jakauma

Olkoon muuttuja *x diskreetti* ja oletetaan, että sen *mahdolliset arvot* ovat

$$y_1, y_2, \ldots, y_m$$

Olkoot

$$x_1, x_2, \ldots, x_n$$

muuttujan *x havaitut arvot*. Muuttujan *x* mahdollisen arvon y_k , $k = 1, 2, \ldots, m$ **frekvenssi** f_k kertoo *kuinka monta kertaa* y_k esiintyy havaintoarvojen $x_1, x_2, \ldots, x_n$ joukossa. Muuttujan *x mahdolliset arvot*

$$y_1, y_2, \ldots, y_m$$

yhdessä niiden *frekvenssien*

$$f_1, f_2, \ldots, f_m$$

kanssa muodostavat muuttujan *x havaittujen arvojen* $x_1, x_2, \ldots, x_n$ **frekvenssijakauman**.

Huomaa, että

$$f_1 + f_2 + \cdots + f_m = n$$

jossa *n* on havaintojen kokonaislukumäärä.

Pylväsdiagrammi

Frekvenssijakaumaa

$$(y_k , f_k) , k = 1, 2, \ldots, m$$

voidaan kuvata graafisesti **pylväsdiagrammilla**, jossa muuttujan x mahdollisen arvon y_k havainto-arvojen $x_1, x_2, \ldots, x_n$ joukossa esittää pylväs, jonka korkeus vastaa frekvenssiä f_k .

Huomautus:

- Pylväsdiagrammin tulkinta on analoginen *diskreetin todennäköisyysjakauman pistetodennäköisyysfunktion* tulkinnan kanssa.

Pylväsdiagrammin piirtäminen

Olkoot

$$y_1, y_2, \ldots, y_m$$

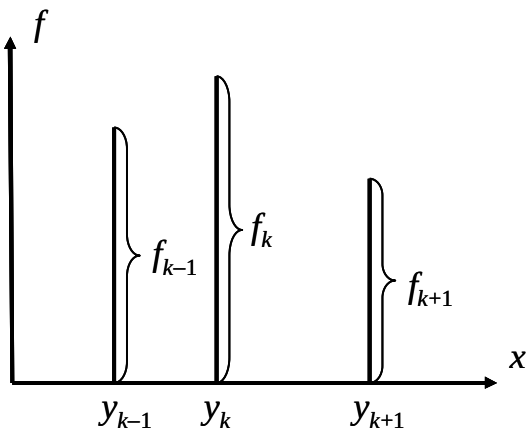
muuttujan *x* mahdolliset arvot ja olkoon

$$(y_k , f_k) , k = 1, 2, \ldots, m$$

muuttujan *x* havaittujen arvojen

$$x_1, x_2, \ldots, x_n$$

frekvenssijakauma. *Frekvenssi* f_k kertoo kuinka monta kertaa y_k esiintyy muuttujan *x* havaittujen arvojen joukossa.



Tarkastellaan muuttujan *x* mahdollista arvoa y_k vastaavan *pylvään* piirtämistä pylväsdiagrammiin.

Muuttujan x mahdolliset arvot y_k määräävät pylväiden *paikat*. Pylvään *korkeus* valitaan suhteessa arvon y_k frekvenssiin f_k .

Esimerkki 1: Pylväsdigrammin konstruointi.

Matemaattisen tilastotieteen kurssille osallistui 20 opiskelijaa.

Kurssin loppukokeen tehtävän 4 arvosteluasteikkona oli 0-6 pistettä niin, että

0 = huonoin pistemäärä

6 = paras pistemäärä

Opiskelijoiden saamat pisteet on annettu ylemmässä taulukossa oikealla.

Alemmassa taulukossa on annettu pisteiden *frekvenssijakauma*.

Kuva oikealla esittää pisteiden frekvenssijakaumaa vastaavaa *pylväsdiagrammia*.

Muuttujan

x = pistemäärä

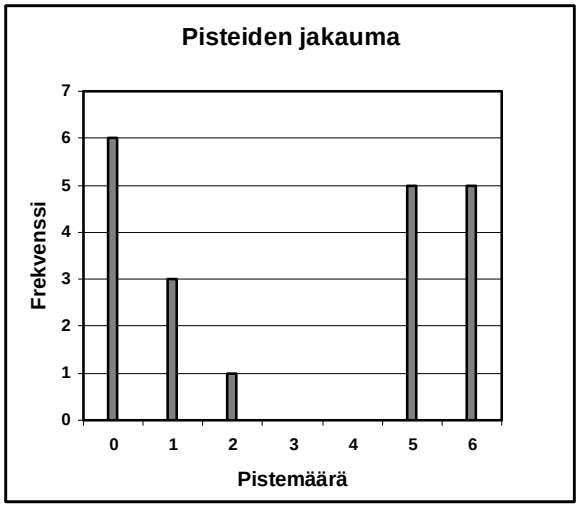
mahdolliset arvot määräävät pylväiden *paikan*.

Pylväät on piirretty niin, että niiden *korkeudet* vastaavat muuttujan x mahdollisten arvojen *frekvenssejä*.

Pisteet; n = 20

0	0	0	0	0
0	1	1	1	2
5	5	5	5	5
6	6	6	6	6

Pisteet	Frekvenssi
0	6
1	3
2	1
3	0
4	0
5	5
6	5



Luokkafrekvenssit ja luokkafrekvenssien jakauma

Olkoon muuttuja x *jatkuva* ja oletetaan, että sen *mahdolliset arvot* ovat *välillä*

(a, b)

jossa voi olla $a = -\infty$, $b = +\infty$. Jaetaan väli (a, b) pisteillä

$$a = a_0 < a_1 < a_2 < \quad \text{L} \quad < a_{m-1} < a_m = b$$

pistevieraisiin *osaväleihin*

$$(a_{k-1}\ ,\ a_k]\ ,\ k = 1, 2, \ldots , m$$

Olkoot

$$x_1, x_2, \ldots , x_n$$

muuttujan x *havaitut arvot*. Muuttujan x havaittujen arvojen **frekvenssi** f_k *luokassa* k kertoo niiden havaintoarvojen $x_1, x_2, \ldots , x_n$ *lukumäärän*, jotka kuuluvat väliin

$$(a_{k-1}\ ,\ a_k]\ ,\ k = 1, 2, \ldots , m$$

Luokkavälit

$$(a_{k-1}\ ,\ a_k]\ ,\ k = 1, 2, \ldots , m$$

yhdessä vastaavien *luokkafrekvenssien*

$$f_1, f_2, \ldots , f_m$$

kanssa muodostavat muuttujan x *havaittujen arvojen* $x_1, x_2, \ldots , x_n$ **luokitellun frekvenssi-jakauman**.

Huomaa, että

$$f_1 + f_2 + \cdots + f_m = n$$

jossa n on havaintojen kokonaislukumäärä.

Histogrammi

Luokiteltua frekvenssijakaumaa

$$((a_{k-1}\ ,\ a_k],\ f_k)\ ,\ k = 1, 2, \ldots , m$$

voidaan kuvata graafisesti **histogrammilla**, jossa muuttujan x havaittujen arvojen

$$x_1, x_2, \ldots , x_n$$

frekvenssiä f_k luokassa $(a_{k-1}\ ,\ a_k]$, esittää suorakaide, jonka *kantana* on väli

$$(a_{k-1}\ ,\ a_k]$$

ja jonka *korkeus* määrätään niin, että *suorakaiteen pinta-ala vastaa luokkafrekvenssiä* f_k .

Huomautus:

- Histogrammin tulkinta on analoginen *jatkuvan todennäköisyysjakauman tiheysfunktion* tulkinnan kanssa.

Histogrammin piirtäminen

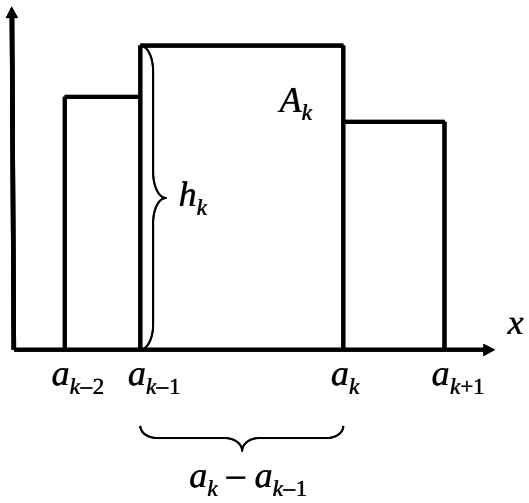
Olkoon

$$((a_{k-1}\ ,\ a_k],\ f_k)\ ,\ k = 1, 2, \ldots , m$$

muuttujan x havaittujen arvojen

$$x_1, x_2, \ldots , x_n$$

luokiteltu frekvenssijakauma. *Luokkafrekvenssi* f_k kertoo niiden havaintoarvojen lukumäärän, jotka kuuluvat *luokkaväliin* $(a_{k-1}\ ,\ a_k]$.



Tarkastellaan k. luokkaa vastaavan *suorakaiteen* piirtämistä histogrammiin.

Luokkaväli $(a_{k-1}\ ,\ a_k]$ muodostaa suorakaiteen *kannan*.

Suorakaiteen *korkeus* h_k saadaan ehdosta

$$\begin{aligned} A_k &= \text{ } k\text{. luokkaa vastaavan suorakaiteen pinta-ala} \\ &= (a_k - a_{k-1}) \times h_k \\ &= f_k \end{aligned}$$

Esimerkki 2: Histogrammin konstruointi.

Kone tekee *ruuveja*, joiden *pituudet vaihtelevat satunnaisesti*.

Poimitaan ruuvien joukosta satunnaisotos, jonka koko

$$n = 30$$

ja mitataan otokseen poimittujen ruuvien pituudet.

Otokseen poimittujen 30:n ruuvin pituudet (yksikkö = cm) on annettu ylimmässä taulukossa oikealla.

Muodostetaan otokseen poimittujen ruuvien pituuksien *luokiteltu frekvenssijakauma*.

Järjestetään sitä varten havaintoarvot *suuruus-järjestykseen*; ks. keskimmäistä taulukkoa oikealla.

Pituuksien luokiteltu frekvenssijakauma on annettu alimmassa taulukossa oikealla.

Esimerkiksi luokkaan, jonka määrää puoliavoin väli

$$(10.10, 10.15]$$

kuuluu 4 ruuvia.

Kuva oikealla esittää otokseen poimittujen ruuvien pituuksien luokiteltua frekvenssi-jakaumaa vastaavaa histogrammia.

Luokkavälit määräävät histogrammin suora-kaiteiden kannat.

Suorakaiteiden korkeudet saadaan ehdosta, jonka mukaan suorakaiteiden pinta-alojen on vastattava luokkafrekvenssejä.

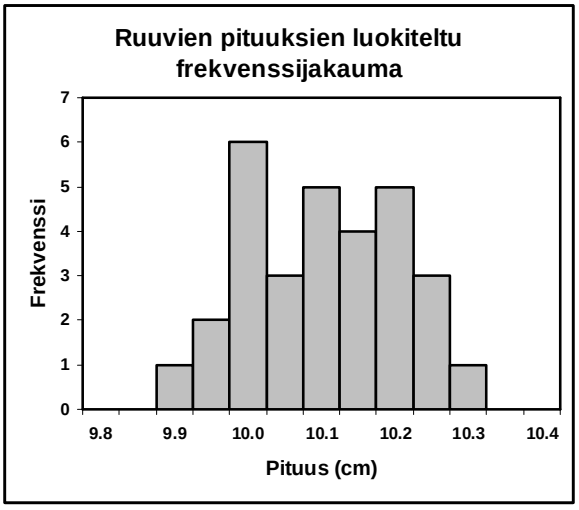
Ruuvien pituudet; *n* = 30

10.05	10.23	10.02	10.24	10.14
10.06	10.07	10.09	10.00	10.09
10.30	10.17	10.18	10.00	10.01
10.00	9.93	10.16	10.21	10.20
9.99	10.13	9.88	9.99	10.12
10.20	9.93	10.00	10.07	10.13

Ruuvien pituudet; *n* = 30

9.88	9.93	9.93	9.99	9.99
10.00	10.00	10.00	10.00	10.01
10.02	10.05	10.06	10.07	10.07
10.09	10.09	10.12	10.13	10.13
10.14	10.16	10.17	10.18	10.20
10.20	10.21	10.23	10.24	10.30

Luokkavälit	Luokkafrekvenssit
(9.85,9.90]	1
(9.90,9.95]	2
(9.95,10.00]	6
(10.00,10.05]	3
(10.05,10.10]	5
(10.10,10.15]	4
(10.15,10.20]	5
(10.20,10.25]	3
(10.25,10.30]	1



Mitta-asteikot ja havaintoarvojen jakauman kuvaaminen

Laatuero- tai järjestysasteikollisten muuttujien havaittujen arvojen jakauman kuvaamiseen käytettävät välineet:

- **Frekvenssijakauma**

- **Pylväsdiagrammi**

Välimatka- tai suhdeasteikollisten muuttujien havaittujen arvojen jakauman kuvaamiseen käytettävät välineet:

- **Luokiteltu frekvenssijakauma**
- **Histogrammi**

Lisätietoja mitta-asteikoista: ks. lukua **Tilastolliset aineistot, niiden kerääminen ja mittaaminen**.

3.3. Tunnusluvut

Olkoot

$$x_1, x_2, \ldots, x_n$$

muuttujan *x havaittuja arvoja*.

Muuttujan *x* havaittujen arvojen *jakaumaa* voidaan kuvailla ja esitellä *tiivistämällä* havainto-arvoihin sisältyvä *informaatio* sopivaan muotoon:

- *Jakaumaa kokonaisuutena* voidaan kuvata sopivasti valituilla **graafisilla esityksillä**.
- *Jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **tunnusluvuilla**.

Tunnuslukujen tehtävänä on kuvata havaintoarvojen jakauman keskeisiä *karakteristisia ominaisuuksia*:

- *Keskimääräisten*, *tyypillisten* tai *yleisten* havaintoarvojen *sijaintia* kuvataan **keskiluvuilla**.
- Havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* kuvataan **hajontaluvuilla**.
- Havaintoarvojen jakauman **vinoutta** ja **huipukkuutta** voidaan kuvata sopivasti valituilla tunnusluvuilla.

Havaintoarvojen jakauman *karakteristisia ominaisuuksia* on aina syytä kuvata usealla erilaisella tavalla. Havaintoaineiston *jakauma* ja *kuvauksen tavoitteet* määräävät mitä tunnuslukuja havainto-aineistosta *kannattaa* laskea. Tutkittavan muuttujan *mitta-asteikolliset ominaisuudet* määräävät mitä tunnuslukuja havaintoaineistosta *saa* laskea.

Huomautus:

- Tunnuslukujen antama kuvaus havaintoarvojen jakaumasta jää puutteelliseksi ja saattaa olla jopa harhaanjohtava, ellei sitä täydennetä sopivilla jakaumaa kuvaavilla graafisilla esityksillä kuten pylväsdiagrammilla tai histogrammilla.

Tunnusluvut ja mitta-asteikot

Tarkasteltavan muuttujan *mitta-asteikolliset ominaisuudet ohjaavat* havaintoaineiston kuvaamisessa käytettävien *tunnuslukujen valintaa*

Tavalliseimmat tunnusluvut voidaan ryhmitellä tarkastelun kohteena olevien muuttujien mitta-asteikollisten ominaisuuksien perusteella seuraavalla tavalla:

(i) Tunnusluvut **välimatka- ja suhdeasteikollisten muuttujien** havaituille arvoille:

Aritmeettinen keskiarvo keskilukuna, *Varianssi* ja *keskihajonta* hajontalukuina, *Origo-momentit, Keskusmomentit, Vinous* ja *huipukkuus, Harmoninen keskiarvo, Geometrinen keskiarvo*

(ii) Tunnusluvut **järjestysasteikollisten muuttujien** havaituille arvoille:

Järjestystunnusluvut, Minimi ja *maksimi, Vaihteluväli* ja *vaihteluvälin pituus, Prosentti-pisteet, Mediaani* keskilukuna, *Kvartiilit, Kvartiiliväli* ja *kvartiilivälin pituus, Kvartiili-poikkeama* hajontalukuna

(iii) Tunnusluvut **laatueroasteikollisten muuttujien** havaituille arvoille:

Suhteellinen frekvenssi, Moodi keskilukuna

Tarkastellaan vielä sitä mitä tunnuslukuja missäkin tilanteessa *saa laskea*.

Välimatka- ja suhdeasteikollisille muuttujille sallitut tunnusluvut:

- **Origo-** ja **keskusmomentit** ja niistä johdetut tunnusluvut
- *Keskilukuna* käytetään tavallisesti **aritmeettista keskiarvoa**, mutta monissa tilanteissa keskilukuna on syytä käyttää **mediaania** tai **moodia**
- *Hajontalukuna* käytetään tavallisesti **keskihajontaa** tai **varianssia**
- Kaikki *laatuero-* ja *järjestysasteikollisten muuttujien tunnusluvut*

Järjestysasteikollisille muuttujille sallitut tunnusluvut:

- **Järjestystunnusluvut** ja niistä johdetut tunnusluvut
- *Keskilukuna* käytetään tavallisesti **mediaania**, mutta monissa tilanteissa keskilukuna on syytä käyttää **moodia**
- *Hajontalukuna* käytetään usein **kvartiilipoikkeamaa**
- Kaikki *laatueroasteikollisten muuttujien tunnusluvut*

Huomautus:

- Välimatka- tai suhdeasteikollisten muuttujien tunnuslukuja *ei ole* mielekästä laskea järjestysasteikollisten muuttujien havaituille arvoille.

Laatueroasteikollisille muuttujille sallitut tunnusluvut:

- Suhteelliset frekvenssit
- *Keskilukuna* käytetään moodia

Huomautus:

- Järjestys-, välimatka- tai suhdeasteikollisten muuttujien tunnuslukuja *ei ole* mielekästä laskea laatueroasteikollisten muuttujien havaituille arvoille.

3.4. Suhdeasteikollisten muuttujien tunnusluvut

Aritmeettinen keskiarvo

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ **aritmeettinen keskiarvo** M saadaan kaavalla

$$M = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Aritmeettinen keskiarvo on havaintoarvojen *painopiste* ja kuvaa havaintoarvojen *keskimääräistä* arvoa.

Esimerkki 1. Aritmeettinen keskiarvo.

Lasketaan lukujen

$$1, -2, 0.5, 4, 5, -10$$

aritmeettinen keskiarvo. Lukujen summa on

$$\sum_{i=1}^{n} x_i = 1 + (-2) + 0.5 + 4 + 5 + (-10) = -1.5$$

joten

$$M = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(-1.5) = -0.25$$

Luokitellun aineiston aritmeettinen keskiarvo

Oletetaan, että *jatkuvan* muuttujan x havaituista arvoista on muodostettu *luokiteltu frekvenssi-jakauma* ja olkoon käytetty *luokkien lukumäärä k*. Oletetaan, että *luokkakeskuksina* ovat luvut

$$z_1, z_2, \ldots, z_k$$

ja että vastaavat *luokkafrekvenssit* ovat

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

Tällöin **luokitellun aineiston aritmeettinen keskiarvo** on

$$M = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i$$

jossa $n = \sum f_i$.

Aritmeettinen keskiarvo havaintoarvojen jakauman kuvaajana

Aritmeettinen keskiarvo kuvaa havaintoarvojen *keskimääräistä* arvoa. Havaintoarvojen aritmeettinen keskiarvo sijoittuu havaintoarvojen jakauman **painopisteeseen**.

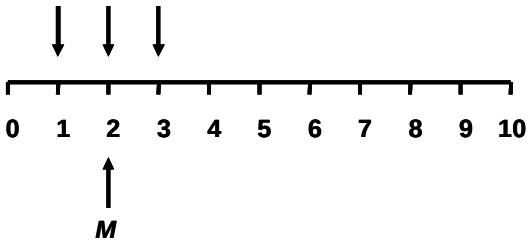
Jos havaintoarvojen jakauma on *vino* tai *monihuippuinen*, aritmeettinen keskiarvo *ei välttämättä ole tyypillinen* tai *yleinen havaintoarvo*.

Aritmeettinen keskiarvo *ei ole* **robusti** eli *se on herkkä poikkeaville havaintoarvoille*, koska *jokainen havainto-arvo vetää aritmeettista keskiarvoa puoleensa*.

Esimerkki 2: Aritmeettisen keskiarvon herkkyys poikkeaville havainnoille.

Havaintoarvojen 1, 2, 3 aritmeettinen keskiarvo on

$$M = \frac{1+2+3}{3} = 2$$

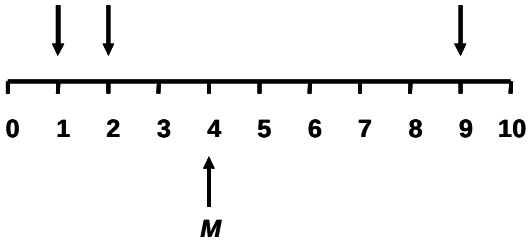


Muutetaan havaintoarvo 3 havaintoarvoksi 9 ja *pidetään muut havaintoarvot ennallaan*.

Tällöin *uudeksi* aritmeettiseksi keskiarvoksi tulee

$$M = \frac{1+2+9}{3} = 4$$

Ks. kuvaa oikealla.



Varianssi

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ (*otos-*) **varianssi** saadaan kaavoilla

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

jossa

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

on lukujen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo*. Otosvarianssi kuvaa havaintoarvojen *hajaantuneisuutta* (tai *keskittyneisyyttä*) niiden aritmeettisen keskiarvon (painopisteen) ympärillä.

Otosvarianssi s_x^2 voidaan laskea myös seuraavilla kaavoilla:

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

Perustelu:

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen muuttujan x* havaittuja arvoja ja olkoon

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

havaintoarvojen *aritmeettinen keskiarvo*. Tällöin

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\bar{x}^2) \\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\cdot n\bar{x}+n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2
\end{aligned}
$$

■

Lukujen $x_1, x_2, \ldots, x_n$ otosvarianssi lasketaan usein myös kaavalla

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{n-1}{n}s_x^2$$

jossa summalausekkeen jakajana on siis n.

Huomautus:

- Otosvarianssin kaksi erilaista kaavaa liittyvät erilaisiin tapoihin *estimoida* normaalijakauman $N(\mu, \sigma^2)$ *varianssiparametri* σ^2 :

  (i) s^2 on *harhaton estimaattori* parametrille σ^2.

  (ii) $\hat{\sigma}_x^2$ on parametrin σ^2 *suurimman uskottavuuden estimaattori*.

Ks. lukuja **Estimointi** ja **Estimointimenetelmät**.

Aritmeettisen keskiarvon ja varianssi laskeminen

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Jos havaintoarvojen $x_1, x_2, \ldots, x_n$ aritmeettinen keskiarvo ja varianssi joudutaan laskemaan *käsin* tai *laskinta* käyttäen, kannattaa laskut järjestää alla olevan taulukon muotoon ja käyttää niiden vieressä esitettyjä kaavoja.

i	x_i	x_i^2
1	x_1	x_1^2
2	x_2	x_2^2
M	M	M
n	x_n	x_n^2
Summa	$\sum_{i=1}^{n}x_i$	$\sum_{i=1}^{n}x_i^2$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)$$

Esimerkki 3: Aritmeettisen keskiarvon ja varianssin laskeminen.

Tarkastellaam aritmeettisen keskiarvon ja varianssin laskemista esimerkin 1 aineistosta.

Muodostetaan esimerkin 1 luvuista seuraava taulukko:

i	x_i	x_i^2
1	1	1
2	–2	4
3	0.5	0.25
4	4	16
5	5	25
6	–10	100
Summa	–1.5	146.25

Taulukosta saadaan:

$$\sum_{i=1}^{n} x_i = -1.5$$

$$\sum_{i=1}^{n} x_i^2 = 146.25$$

Siten

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(-1.5) = -0.25$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{6-1}\left(146.25 - \frac{1}{6}(-1.5)^2\right) = \frac{1}{5}\times 145.875 = 29.175$$

Keskihajonta

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ (*otos-*) **keskihajonta** on

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{s_x^2}$$

jossa

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

on lukujen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 on lukujen $x_1, x_2, \ldots, x_n$ (*otos-*) *varianssi*.

Otoskeskihajonta kuvaa (kuten otosvarianssi) havaintoarvojen *hajaantuneisuutta* (tai *keskittyneisyyttä*) niiden aritmeettisen keskiarvon (painopisteen) ympärillä.

Varianssi ja keskihajonta havaintoarvojen jakauman kuvaajana

Keskihajonta ja varianssi ovat *havaintoarvojen vaihtelun mittoja*.

Varianssi on havaintoarvojen keskimääräinen neliöllinen poikkeama niiden aritmeettisesta keskiarvosta. Havaintoarvojen keskihajonta on varianssin neliöjuuri.

"*Pieni*" keskihajonta (varianssi) merkitsee sitä, että havaintoarvot *keskittyvät* niiden painopisteen (aritmeettisen keskiarvon) ympärille. "*Suuri*" keskihajonta (varianssi) merkitsee sitä, että havaintoarvot ovat *hajaantuneet* niiden painopisteen (aritmeettisen keskiarvon) ympärille.

Jos havaintoarvojen jakaumaa kuvaavana keskilukuna on käytetty aritmeettista keskiarvoa, hajontalukuna on luontevaa käyttää keskihajontaa:

(i) Keskihajonnalla ja aritmeettisella keskiarvolla on sama dimensio (laatu).

(ii) Varianssin ja aritmeettisen keskiarvon dimensio (laatu) ei ole sama.

Varianssi ja keskihajonta *eivät ole* **robusteja** eli *ne ovat herkkiä poikkeaville havaintoarvoille*.

Standardointi

Olkoon $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*. Tällöin **standardoitujen** *havaintoarvojen*

$$z_i = \frac{x_i - \bar{x}}{s_x}, \; i = 1, 2, \mathrm{K}, n$$

aritmeettinen keskiarvo ja varianssi ovat

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = 0$$

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \bar{z})^2 = 1$$

Tilastollinen etäisyys

Olkoot $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*. Tällöin havaintoarvojen x_k ja x_l **tilastollinen etäisyys** on

$$d_{kl} = \frac{x_k - x_l}{s_x}$$

Tilastollinen etäisyys suhteuttaa kahden havaintopisteen etäisyyden lukusuoralla kaikkien käytettävissä olevien havaintoarvojen keskihajontaan.

On hyvä tietää, että monien parametristen testien *testisuureet* voidaan tulkita *etäisyysmitoiksi*, jotka mittaavat testattavan hypoteesin todennäköisyysjakauman parametrille antaman arvon ja havainnoista lasketun parametrin estimaatin tilastollista etäisyyttä.; ks. luvun **Testit suhdeasteikollisille muuttujille**.

Origomomentit

Olkoot

$x_1, x_2, \ldots, x_n$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ ***k.* origo-momentti** on

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \ , k = 1, 2, \mathrm{K}$$

Keskusmomentit

Olkoot

$x_1, x_2, \ldots, x_n$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ ***k.* keskusmomentti** on

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^k$$

jossa

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

on lukujen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo*.

Vinous

Olkoot

$x_1, x_2, \ldots, x_n$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Havaintoarvojen $x_1, x_2, \ldots, x_n$ jakauman **vinoutta** voidaan kuvata otostunnusluvulla

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

jossa

m_2 = 2. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

m_3 = 3. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

(i) Jos

$c_1 > 0$

havaintojen jakauma on **positiivisesti vino** eli **vino oikealle**.

(iii) Jos

$c_1 = 0$

havaintojen jakauma on **symmetrinen**.

(iii) Jos

$c_1 < 0$

havaintojen jakauma on **negatiivisesti vino** eli **vino vasemmalle**.

Huipukkuus

Olkoot

$$x_1, x_2, \ldots, x_n$$

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Havaintoarvojen $x_1, x_2, \ldots, x_n$ jakauman **huipukkuutta** voidaan kuvata otostunnusluvulla

$$c_2 = \frac{m_4}{m_2^2} - 3$$

jossa

m_2 = 2. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

m_4 = 4. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

(i) Jos

$$c_2 > 0$$

havaintojen jakauma on **huipukas** normaalijakautuneeseen aineistoon verrattuna.

(ii) Jos

$$c_2 > 0$$

havaintojen jakauma on **yhtä huipukas** kuin normaalijakautunut aineisto.

(iii) Jos

$$c_1 < 0$$

havaintojen jakauma on **laakea** normaalijakautuneeseen aineistoon verrattuna.

Esimerkki 4: Havaintoarvojen 2. , 3. ja 4. keskusmomentti sekä vinous ja huipukkuus.

Määrätään esimerkkíen 1 ja 2 lukujen

1, –2, 0.5, 4, 5, –10

2. , 3. ja 4. keskusmomentti sekä vinous ja huipukkuus.

Esimerkissä 2 todettiin, että lukujen aritmeettinen keskiarvo ja otosvarianssi ovat:

$$\bar{x} = -0.25$$

ja

$$s_x^2 = 29.175$$

Pienellä laskutyöllä saadaan seuraavat summat:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 145.875$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^3 = -708.375$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^4 = 10134.96$$

Siten 2. , 3. ja 4. keskusmomentti ovat:

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{6}\times 145.875 = 24.3125$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3 = \frac{1}{6}\times(-708.375) = -118.063$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4 = \frac{1}{6}\times 10134.96 = 1689.16$$

Lukujen vinoudeksi ja huipukkuudeksi saadaan:

$$c_1 = \frac{m_3}{m_2^{3/2}} = \frac{-118.063}{24.3125^{3/2}} = -0.994844$$

$$c_2 = \frac{m_4}{m_2^2} - 3 = \frac{1689.16}{24.3125^2} - 3 = 0.142333$$

Harmoninen keskiarvo

Olkoot

$$x_1, x_2, \ldots, x_n$$

positiivisia lukuja. Lukujen $x_1, x_2, \ldots, x_n$ **harmoninen keskiarvo** on

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Lukujen $x_1, x_2, \ldots, x_n$ harmonisen keskiarvon käänteisluku on lukujen $x_1, x_2, \ldots, x_n$ käänteislukujen aritmeettinen keskiarvo:

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

Esimerkki 5: Harmoninen keskiarvo.

Esimerkki osoittaa, että *aritmeettinen keskiarvo ei ole aina sopiva keskiluku*.

Olkoon kahden kaupungin A ja B välimatka 120 km. Oletetaan, että ajat A:sta B:hen 60 km/h ja B:stä A:han 120 km/h. Mikä on ollut *keskinopeus* edestakaisella matkalla?

Matka A:sta B:hen ja takaisin	=	240 km
Ajoaika A:sta B:hen	=	2 h
Ajoaika B:stä A:han	=	1 h
Ajoaika yhteensä	=	3 h
Keskinopeus edestakaisella matkalla	=	240/3 = 80 km/h

Nopeuksien *aritmeettinen keskiarvo*

$$M = \frac{60+120}{2} = 90 \text{ km/h}$$

antaa väärän keskinopeuden. Sen sijaan nopeuksien *harmoninen keskiarvo*

$$H = \frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{120}\right)} = 80 \text{ km/h}$$

antaa oikean keskinopeuden.

Geometrinen keskiarvo

Olkoot

$$x_1, x_2, \ldots, x_n$$

positiivisia lukuja. Lukujen $x_1, x_2, \ldots, x_n$ **geometrinen keskiarvo** on

$$G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Lukujen $x_1, x_2, \ldots, x_n$ geometrisen keskiarvon logaritmi on lukujen $x_1, x_2, \ldots, x_n$ logaritmien aritmeettinen keskiarvo:

$$\log(G) = \frac{\log(x_1)+\log(x_2)+\cdots+\log(x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n}\log(x_i)$$

Esimerkki 6: Geometrinen keskiarvo.

Esimerkki osoittaa, että *aritmeettinen keskiarvo ei ole aina sopiva keskiluku*.

Olkoon lainan suuruus 100 € ja olkoon korkoprosentti 1. vuotena 10 % ja 2. vuotena 20 %. Jos lainaa ei lyhennetä, lainapääoma karttuu seuraavalla tavalla:

Pääoma 1. vuoden lopussa = 1.1×100 = 110 €

Pääoma 2. vuoden lopussa = 1.2×110 = 132 €

Lainapääoma karttuu siis kahdessa vuodessa 32 %. Jos kumpanakin vuotena käytettäisiin samaa korkoprosenttia, miten se pitäisi valita, jotta lainapääoma olisi 2. vuoden lopussa 132 € ?

Korkoprosenttien aritmeettinen keskiarvo

$$M = \frac{10+20}{2} = 15 \text{ \%}$$

tuottaa väärän lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = 1.15×100 = 115 €

Pääoma 2. vuoden lopussa = 1.15×115 = 132.25 €

Myös korkoprosentti

$$\frac{32}{2} = 16 \text{ \%}$$

tuottaa väärän lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = 1.16×100 = 116 €

Pääoma 2. vuoden lopussa = 1.16×116 = 134.56 €

Sen sijaan geometrinen keskiarvo

$$G = \sqrt{1.1 \times 1.2} = 1.1489125$$

antaa korkoprosentiksi

$$G - 1 = 0.1489125 = 14.89125\ \%$$

joka tuottaa oikean lainapääoman 2. vuoden lopussa:

Pääoma 1. vuoden lopussa = 1.1489125×100 = 114.89125 €

Pääoma 2. vuoden lopussa = 1.1489125×114.89125 = 132.00 €

Aritmeettinen, harmoninen ja geometrinen keskiarvo

Oletetaan, että *aritmeettinen keskiarvo M, harmoninen keskiarvo H* ja *geometrinen keskiarvo G* määrätään samoista positiiviluvuista $x_1, x_2, \ldots, x_n$. Tällöin

$$H \le G \le M$$

jos ja vain, jos

$$x_1 = x_2 = \cdots = x_n$$

3.5. Järjestysasteikollisten muuttujien tunnusluvut

Järjestystunnusluvut

Olkoot

$$x_1, x_2, \ldots, x_n$$

järjestys-, välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. *Järjestetään* havainto-arvot $x_1, x_2, \ldots, x_n$ suuruusjärjestykseen pienimmästä suurimpaan ja olkoot

$$z_1, z_2, \ldots, z_n$$

järjestykseen asetetut havaintoarvot. Suuruusjärjestyksessä *k*. havaintoarvoa z_k kutsutaan ***k.* järjestystunnusluvuksi**.

Minimi, maksimi, vaihteluväli

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, välimatka- tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruus-järjestykseen* pienimmästä suurimpaan.

Tällöin

z_1 = **minimiarvo**

z_n = **maksimiarvo**

(z_1 , z_n) = **vaihteluväli**

$z_n - z_1$ = **vaihteluvälin pituus**

Prosenttipisteet

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

Havaintoarvojen ***p. prosenttipiste***

$$z_{(p)}, p = 1, 2, \ldots, 99$$

on piste, joka jakaa havaintoaineiston *kahteen osaan*:

(i) p % havaintoarvoista on lukua $z_{(p)}$ *pienempiä* tai korkeintaan yhtä suuria kuin $z_{(p)}$.

(ii) $(100 - p)$ % havaintoarvoista on lukua $z_{(p)}$ *suurempia*.

Mediaani

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

Mediaani *Me* on havaintoarvojen 50. *prosenttipiste*:

$$Me = z_{(50)}$$

Mediaani jakaa havaintoaineiston *kahteen yhtä suureen osaan* niin, että toisessa *kaikki* havaintoarvot ovat mediaania *pienempiä*, toisessa *kaikki* havaintoarvot ovat mediaania *suurempia*.

Havaintoarvojen mediaani *Me* voidaan määrätä seuraavalla tavalla:

(1) Järjestetään havaintoarvot *suuruusjärjestykseen* pienimmästä suurimpaan.

(2a) Jos havaintoarvojen lukumäärä on *pariton*, mediaani on suuruusjärjestekseen asetetuista havaintoarvoista *keskimmäinen*.

(2b) Jos havaintoarvojen lukumäärä on *parillinen*, mediaani on suuruusjärjestekseen asetetuista havaintoarvoista *kahden keskimmäisen aritmeettinen keskiarvo*.

Esimerkki 1. Minimi, maksimi, vaihteluväli, vaihteluvälin pituus ja mediaani.

Määrätään kappaleen **Suhdeasteikollisten muuttujien tunnusluvut** esimerkin 1 lukujen

$$1, -2, 0.5, 4, 5, -10$$

minimi, maksimi, vaihteluväli, vaihteluvälin pituus ja mediaani.

Järjestetään luvut ensin suuruusjärjestykseen pienimmästä suurimpaan. Saamme tulokseksi lukusarjan

$$-10, -2, 0.5, 1, 4, 5$$

Saamme tästä lukusarjasta helposti seuraavat tunnusluvut:

minimi = –10

maksimi = 5

vaihteluväli = (–10, 5)

vaihteluvälin pituus = maksimi – minimi = – 5 – (–10) = –15

Koska lukuja on parillinen määrä, lukujen mediaani on suuruusjärjestekseen asetetuista havaintoarvoista kahden keskimmäisen aritmeettinen keskiarvo:

$$Me = \frac{0.5+1}{2} = 0.75$$

Mediaani havaintoarvojen jakauman kuvaajana

Mediaani on suuruusjärjestykseen asetettujen havainto-arvojen *keskimmäinen* havaintoarvo (tai kahden keskimmäisen aritmeettinen keskiarvo). Jos havaintoarvojen jakauma on *symmetrinen*, havaintoarvojen mediaani ja aritmeettinen keskiarvo yhtyvät.

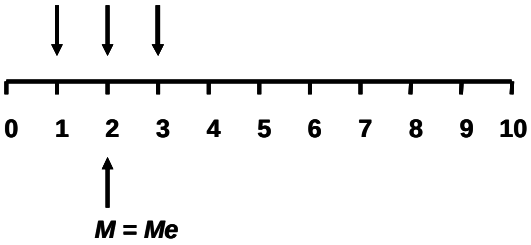
Jos havaintoarvojen jakauma on *yksihuippuinen*, mutta *vino*, havaintoarvojen mediaani kuvaa *tyypillisiä* havaintoarvoja usein paremmin kuin niiden aritmeettinen keskiarvo. Jos havaintoarvojen jakauma on *monihuippuinen*, mediaani *ei välttämättä ole yleinen havaintoarvo*.

Mediaani *on* **robusti** eli se *ei ole* – toisin kuin aritmeettinen keskiarvo – *herkkä poikkeaville havaintoarvoille*.

Esimerkki 2: Mediaanin robustisuus

Havaintoarvojen 1, 2, 3 aritmeettinen keskiarvo on

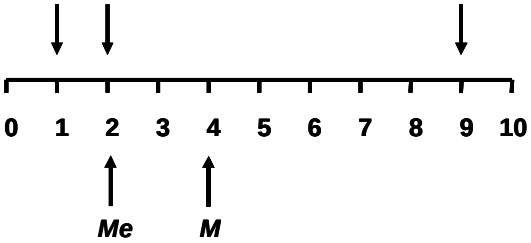
$$M = \frac{1+2+3}{3} = 2$$



Muutetaan havaintoarvo 3 havaintoarvoksi 9 ja *pidetään muut havaintoarvot samoina*.

Tällöin *uudeksi* aritmeettiseksi keskiarvoksi tulee

$$M = \frac{1+2+9}{3} = 4$$



Sen sijaan havaintoarvojen mediaani

$Me = 2$

ei ole muuttunut.

Ks. kuvaa oikealla.

Mediaani, aritmeettinen keskiarvo ja vinous

Oletetaan, että *aritmeettinen keskiarvo M* ja *mediaani Me* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen luokitellusta *frekvenssijakaumasta*. Tällöin pätee seuraava:

(i) *Vasemmalle vinoilla jakaumilla*

$M < Me$

(ii) *Symmetrisillä jakaumilla*

$$M \approx Me$$

(iii) *Oikealle vinoilla jakaumilla*

$$Me > M$$

Esimerkki 3: Mediaani, aritmeettinen keskiarvo ja vinous.

Alla esitetyt histogrammikuviot perustuvat sataan satunnaislukugeneraattorin avulla generoituun havaintoarvoon.

(i) Oikeanpuoleinen histogrammi.

$$X \sim \chi^2(5)$$

Jakauma on vino *oikealle*:

Vinous = 1.25

Aritmeettinen keskiarvo = 5.19

Mediaani = 4.41

(ii) Vasemmanpuoleinen histogrammi.

$$Y = 20 - X$$

jossa

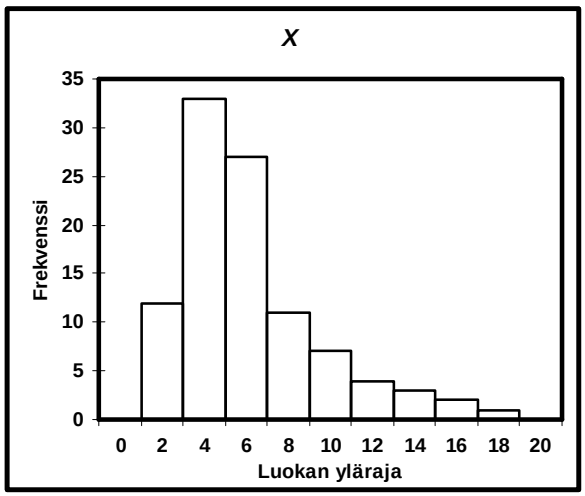
$$X \sim \chi^2(5)$$

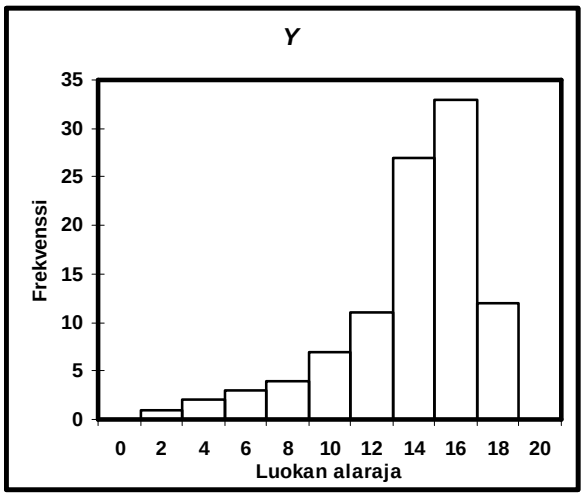
Jakauma on vino *vasemmalle*:

Vinous = –1.25

Aritmeettinen keskiarvo = 14.81

Mediaani = 15.59





Luokitellun aineiston mediaani

Luokitellun aineiston mediaani voidaan laskea kaavalla

$$Me = L_i + \frac{\frac{1}{2}n - \Sigma f_j}{f_i} \times c_i$$

jossa

L_i = mediaaniluokan alaraja

$\sum f_j$ = kaikkien mediaaniluokan alapuolella oleviin luokkiin kuuluvien havaintoarvojen frekvenssi

f_i = mediaaniluokkaan kuuluvien havaintoarvojen frekvenssi

c_i = mediaaniluokan pituus

n = havaintoarvojen lukumäärä

Kvartiilit

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruus-järjestykseen* pienimmästä suurimpaan.

Tällöin

Q_1 = **Alakvartiili** = 25. prosenttipiste = $z_{(25)}$

Q_2 = **Keskikvartiili** = 50. prosenttipiste = $z_{(50)}$

Q_3 = **Yläkvartiili** = 75. prosenttipiste = $z_{(75)}$

Kvartiilit Q_1, Q_2, Q_3 jakavat suuruusjärjestykseen asetetun havaintoaineiston *neljään yhtä suureen osaan*. Erityisesti:

Alakvartiili Q_1 = Havaintoarvojen mediaania *Me pienempien havaintoarvojen mediaani*

Keskikvartiili Q_2 = Havaintoarvojen *mediaani Me*

Yläkvartiili Q_3 = Havaintoarvojen mediaania *Me suurempien havaintoarvojen mediaani*

Kvartiilit, kvartiiliväli, kvartiilipoikkeama

Olkoot havaintoarvojen *kvartiilit* Q_1, Q_2, Q_3.

Tällöin

(Q_1 , Q_3) = **kvartiiliväli**

$Q_3 - Q_1$ = IQR = **kvartiilivälin pituus**

$(Q_3 - Q_1)/2$ = $IQR/2$ = **kvartiilipoikkeama**

Kvartiiliväliä, kvartiilivälin pituutta (IQR = interquartile range) ja kvartiilipoikkeamaa voidaan käyttää kuvaamaan havaintoarvojen *hajaantuneisuutta* (*keskittyneisyyttä*).

Jos havaintoarvojen jakaumaa kuvaavana *keskilukuna* on käytetty *mediaania*, *hajontalukuna* käytetään usein *kvartiilipoikkeamaa*.

Box-Whisker-kuvio

Havaintoarvojen jakaumaa voidaan usein kätevästi havainnollistaa ns. **Box and Whisker -kuviolla**.

Olkoon

$$(Q_1 , Q_3)$$

havaintoarvojen *kvartiiliväli* ja

$$Me = Q_2$$

havaintoarvojen *mediaani* ja olkoon

$$IQR = Q_3 - Q_1$$

havaintoarvojen *kvartiilivälin* (Q_1 , Q_3) *pituus*.

Määritellään *sisäaidat* f_1 ja f_3 kaavoilla

$$f_1 = Q_1 - 1.5 \times IQR$$

$$f_3 = Q_3 + 1.5 \times IQR$$

Olkoon a_1 *pienin* havaintoarvo, joka toteuttaa ehdon

$$a_1 \geq f_1$$

Olkoon a_3 *suurin* havaintoarvo, joka toteuttaa ehdon

$$a_3 \leq f_3$$

Määritellään *ulkoaidat* F_1 ja F_3 kaavoilla

$$F_1 = Q_1 - 3 \times IQR$$

$$F_3 = Q_3 + 3 \times IQR$$

Box and Whisker -kuvio koostuu *laatikosta, viiksistä* ja kuvioon merkityistä *poikkeuksellisista havainnoista*.

Box and Whisker -kuvion piirtäminen:

(i) Piirretään suorakaiteen muotoinen **laatikko** kuvaamaan havaintoarvojen *kvartiiliväliä*

$$(Q_1 , Q_3)$$

Merkitään havaintoarvojen *mediaani*

$$Me = Q_2$$

laatikkoon poikkiviivalla.

(ii) Piirretään jananmuotoiset **viikset** laatikon molemmille puolille kuvaamaan välejä

$$(a_1 , Q_1) \text{ ja } (Q_3 , a_3)$$

(iii) Merkitään väleihin

$$(F_1 , a_1) \text{ ja } (a_3 , F_3)$$

kuuluvat havaintoarvoja *tähdillä* ja väleihin

$$(-\infty , F_1) \text{ ja } (F_3 , +\infty)$$

kuuluvat havaintoarvoja *ympyröillä*.

Laatikon ja viiksien määrittelemään väliin

$$(a_1 , a_3)$$

kuuluvat havaintoarvoja pidetään *tavallisina*. Erityisesti kvartiilivälin

$$(Q_1 , Q_3)$$

määrittelemä laatikko sulkee sisäänsä *keskimmäiset* 50 % havaintoarvoista.

Tähdillä ja ympyröillä merkityt, välin (a_1 , a_3) ulkopuolelle jäävät havaintoarvot ovat *poikkeuksellisia*:

(i) Tähdillä merkityt, väleihin

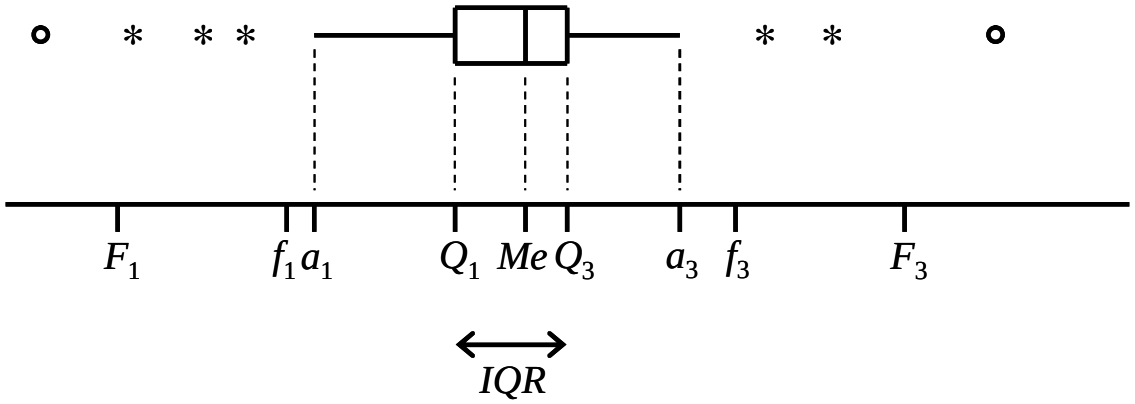
$$(F_1 , a_1) \text{ ja } (a_3 , F_3)$$

kuuluvat havaintoarvot ovat vain *lievästi* poikkeuksellisia.

(ii) Ympyröillä merkityt, väleihin

$$(-\infty , F_1) \text{ ja } (F_3 , +\infty)$$

kuuluvat havaintoarvot ovat *voimakkaasti* poikkeuksellisia.



3.6. Laatueroasteikollisten muuttujien tunnusluvut

Frekvenssi

Olkoon *otoskoko* eli kerättyjen *havaintoarvojen lukumäärä n*. Olkoon *A* jokin perusjoukon osajoukko ja olkoon *f* otokseen kuuluvien *A*-tyyppisten havaintoarvojen *frekvenssi* eli *lukumäärä*.

Tällöin *A*-tyyppisten havaintoarvojen **suhteellinen frekvenssi** eli **osuus** otoksessa on

$$\frac{f}{n}$$

Moodi

Frekvenssijakauman **moodi** eli tyyppiarvo *Mo* on *yleisin havaintoarvo*.

Luokitellun aineiston moodi

Luokitellun frekvenssijakauman **moodi** eli tyyppiarvo Mo on siinä luokassa, jossa luokiteltua frekvenssijakaumaa vastaava histogrammi saavuttaa maksiminsa.

Huomautuksia:

- Jos käytetty luokitus *on tasavälinen,* luokitellun frekvenssijakauman moodi *on* siinä luokassa, jota vastaava frekvenssi on suurin.

- Jos käytetty luokitus *ei ole tasavälinen*, luokitellun frekvenssi jakauman moodi *ei välttämättä* ole siinä luokassa, jota vastaava frekvenssi on suurin.

Luokitellun aineiston moodi voidaan laskea kaavalla

$$Mo = L_i + \frac{d_{i-1}}{d_{i-1} + d_{i+1}} \times c_i$$

jossa

L_i = moodiluokan alaraja

d_{i-1} = moodiluokan ja sitä alemman luokan suorakaiteiden korkeuksien erotus

d_{i+1} = moodiluokan ja sitä ylemmän luokan suorakaiteiden korkeuksien erotus

c_i = moodiluokan pituus

Moodi havaintoarvojen jakauman kuvaajana

Moodi kuvaa *yleisimpien* havaintoarvojen sijoittumista havaintoarvojen jakaumassa.

Jos havaintoarvojen jakauma *on yksihuippuinen* ja *symmetrinen*, havaintoarvojen moodi, mediaani ja aritmeettinen keskiarvo yhtyvät. Jos havaintoarvojen jakauma on *monihuippuinen*, jakaumalla on useita *lokaaleja moodeja*. Jos havaintoarvojen jakauma on *monihuippuinen*, jakauman *lokaalit moodit* antavat usein paremman kuvan jakaumasta kuin mediaani tai aritmeettinen keskiarvo.

Moodi, mediaani, aritmeettinen keskiarvo ja vinous

Oletetaan, että *aritmeettinen keskiarvo M*, *mediaani Me* ja *moodi Mo* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen *luokitellusta frekvenssijakaumasta*. Tällöin pätee seuraava:

(i) *Vasemmalle vinoilla jakaumilla*

$$M < Me < Mo$$

(ii) *Symmetrisillä jakaumilla*

$$M \approx Me \approx Mo$$

(iii) *Oikealle vinoilla jakaumilla*

$$Mo > Me > M$$

Esimerkki 1: Moodi, mediaani, aritmeettinen keskiarvo ja vinous.

Alla esitetyt histogrammikuviot perustuvat sataan satunnaislukugeneraattorin avulla generoituun havaintoarvoon.

(i) Oikeanpuoleinen histogrammi:

$X \sim \chi^2(5)$

Jakauma on vino *oikealle*.

Vinous = 1.25

Aritmeettinen keskiarvo = 5.19

Mediaani = 4.41

Moodi $\in$ (2, 4]

(ii) Vasemmanpuoleinen histogrammi:

$$Y = 20 - X$$

jossa

$$X \sim \chi^2(5)$$

Jakauma on vino *vasemmalle*.

Vinous = –1.25

Aritmeettinen keskiarvo = 14.81

Mediaani = 15.59

Moodi $\in$ (16, 18]