TILP150 Sanasto

Päivitetty 9. toukokuuta 2010

Johdanto

Tässä raportissa esitetään tilastomenetelmien peruskurssin (TILP150) oleellisten `termien ja käsitteiden sanasto`. Raportti perustuu kurssin luentomonisteeneen a´la Kärkkäinen & Högmander ja käyttää samoja merkintöjä tilastollisille suureille[1]. Käsitelistan tarkoituksena on toimia opiskelijan "nopeana referenssinä" kurssin laskuharjoituksia työstettäessä ja kokeisiin valmistautuessa. Kurssimonisteen lisäksi käsitteiden määrittelyyn on käytetty kirjoittajan omaa asiantuntemusta sekä hiukan kurssikirjaa, jota myös suosittelen opiskelijoille:

> **Moore, D.S. and McCabe, G.P.** *Introduction to the Practice of Statistics* (Fifth Edition. Freeman, 2008).

Sanasto on suunniteltu luettevaksi pdf-lukuohjelman (Adobe Reader, Foxit Reader, Ghostview,...) avulla, sillä käsitteet on järjestetty lukukohtaisesti monisteen esittelyjärjestyksessä ja jokaisesta käsitteestä on hyperlinkki raportin lopussa olevaan sanastoon (ja takaisin).

Käsitelistan mahdollisista virheistä ja epäjohdonmukaisuuksista voi lähettää sähköpostia osoitteeseen `riku.jarvinen@alkio.fi`.

[1] Suurelista on monisteen alussa.

Luku 1

Tilastotiede
Tilastollinen tutkimus

Luku 2

Perusjoukko
Havaintoyksikkö
Kokonaistutkimus
Otantatutkimus
Tilastollinen analysointi
Tilastollisen tutkimuksen työvaiheet

Luku 3

Satunnaisotos
Näyte
Harhattomuus
Otantavaihtelu
Otantavirhe
Otantamenetelmä
Otantayksikkö
Yksinkertainen satunnaisotanta
Systemaattinen otanta
Ositettu otanta
Osite
Tasainen kiintiöinti
Suhteellinen kiintiöinti
Ryväsotanta
Kyselytutkimus
Vastauskato
Kysymystyypit
Kokeellinen tutkimus
Koeyksilö
Vastemuuttuja
Riippumaton muuttuja
Taso
Käsittely
Kausaalisuus
Kontrollointi
Koesuunnitelma
Harha
Toistaminen
Kaksoissokkokoe

Luku 4

Kvantitatiivinen ominaisuus
Kvalitatiivinen ominaisuus
Latentti ominaisuus
Muuttuja
Jatkuva muuttuja
Diskreetti muuttuja
Mittaluku
Havaintoarvo
Mitta-asteikko
Nominaaliasteikko
Ordinaaliasteikko
Intervalliasteikko
Suhdeasteikko
Mittaaminen
Systemaattinen virhe
Satunnaisvirhe
Validiteetti
Reliabiliteetti
Laadullisten ominaisuuksien mittaaminen
Indikaattorimuuttuja
Poikkeava havainto
Robusti menetelmä
Havaintomatriisi
Empiirinen jakauma

Luku 5

Frekvenssijakauma
Frekvenssi
Frekvenssitaulu
Suhteellinen frekvenssi
Suhteellinen frekvenssijakauma
Prosentuaalinen frekvenssijakauma
Luokitus
Summafrekvenssi
Summafrekvenssijakauma
Suhteellinen summafrekvenssijakauma
Prosentuaalinen summafrekvenssijakauma
Graafinen esitys nominaaliast. muuttujalle
Histogrammi
Epäjatkuvan muuttujan tulkinta
Moodi
Mediaani
Keskiarvo (aritmeettinen)
Robusti tunnusluku
Symmetrinen jakauma
Oikealle vino jakauma

Vasemmalle vino jakauma
Alakvartiili
Yläkvartiili
Viiksilaatikko
Fraktiili
Hajontaluku
Vaihteluväli
Vaihteluvälin pituus
Kvartiiliväli
Kvartiilivälin pituus
Varianssi
Keskihajonta
Variaatiokerroin
Ehdollinen frekvenssijakauma
Erot jakaumien sijainneissa ja vaihteluissa

Luku 6

Kahden muuttujan yhteisjakauma
Ehdot riippuvuudelle
Ristiintaulukko
Solufrekvenssi
Rivisumma
Sarakesumma
Marginaalijakauma
Sarakeprosentti
Riviprosentti
Odotettu frekvenssi
Jäännös
Standardoitu jäännös
Riippuvuusluvut
Khii toiseen
Kontingenssikerroin
Dikotominen muuttuja
Riskisuhde
Sirontakuvio
Tilastollinen riippuvuus
Lineaarinen riippuvuus
Korrelaatiokerroin
Pearsonin korrelaatiokerroin
Kovarianssi
Spearmanin järjestyskorrelaatiokerroin
Korrelaatiomatriisi
Kovarianssimatriisi
Regressiosuora
Regressiokerroin
PNS-menetelmä
Selitysaste
Regressioanalyysi

Luku 7

Tunnusluvuilla estimointi
Satunnaisilmiö

Alkeistapaus
Tapahtuma
Todennäköisyys suhteellisena frekvenssinä
Klassinen todennäköisyys
Komplementtitapahtuma
Odotusarvo satunnaismuuttujalle
Varianssi satunnaismuuttujalle
Keskihajonta satunnaismuuttujalle
Tiheysfunktio
Kertymäfunktio
Normaalijakauma
Normaalijakauman tiheysfunktio
Standardoitu normaalijakauma
Normaalijakauman kertymäfunktio
Teoreettinen jakauma

Luku 8

Estimaattori
Estimaatti
Harhaton estimaattori
Keskiarvon odotusarvo
Keskiarvon keskivirhe
Keskiarvonotantajakauma
Piste-estimaatti
Luottamusväli
Odotusarvon luottamusväli tunnetulle varianssille
Odotusarvon luottamusväli tuntemattomalle varianssille
Suhteellinen osuus perusjoukossa
Suhteellisen osuuden estimointi
Suhteellisen osuuden luottamusväli
Tilastollinen hypoteesi
Parametrinen testi
Parametriton testi
Muunnos (tilastotestaaminen)
Nollahypoteesi
Vastahypoteesi
Merkitsevyystesti
P-arvo
Tilastollisen testin virhetyypit
Testin voimakkuus
Suhteellisiin osuuksiin liittyvät testit
Tilastollisen testin ja luottamusvälin yhteys

Yksisuuntainen vastahypoteesi
Kaksisuuntainen vastahypoteesi
Yhden otoksen suhteellisen osuuden testi
Odotusarvoihin liittyvät testit
Heteroskedastisuus
Toisistaan riippuvat mittaukset
t-jakauma
F-jakauma
Jakaumien sijaintiin liittyvät parametrittomat testit
Riippuvuuteen liittyvät testit

Sanasto

Alakvartiili Voidaan määritellä usealla eri tavalla. Alakvartiili on

1. sellainen muuttujan arvo, jota pienempiä havaintoja on aineistossa 1/4, tai
2. mediaania pienempien havaintojen mediaani, tai
3. sen luokan muuttujan arvo, jossa prosentuaalinen summafrekvenssi saavuttaa 25 %.

Tilasto-ohjelmat laskevat kvartiileita hieman eri tavoin, joten niiden antamat tulokset saattavat hiukan poiketa toisistaan (tällä ei usein ole tutkimuksen kannalta merkitystä, sillä erot ovat hyvin pieniä). 4

Alkeistapaus Satunnaisilmiön yksi mahdollinen tulos. 5

Dikotominen muuttuja Nominaaliasteikkoinen muuttuja, jolla on kaksi luokkaa (esimerkiksi sukupuoli, kyllä/ei jne.). 4

Diskreetti muuttuja Muuttuja, joka voi saada vain erillisiä lukuarvoja; mahdolliset arvot voidaan periaatteessa luetella. Nominaali- ja ordinaaliasteikkoiset muuttujat ovat aina epäjatkuvia. Välimatka-asteikkoisista muuttujista epäjatkuvia ovat ainakin lukumäärämuuttujat, esimerkiksi lasten lukumäärä perheessä. Tunnetaan myös nimellä *epäjatkuva muuttuja*. 3

Ehdollinen frekvenssijakauma Havainnot on jaettu jonkin taustatekijän mukaan ryhmiin (kahteen tai useampaan) ja tarkastellaan muuttujien arvoja (frekvenssejä) taustatekijäositteessa.. 4

Ehdot riippuvuudelle Tavallisesti asetetaan kolme vaatimusta, jotta voidaan sanoa, että ”X vaikuttaa Y:hyn”:

1. X:n ja Y:n välillä vallitsee tilastollinen riippuvuus,
2. X edeltää ajallisesti tai loogisesti Y:tä
3. mikään kolmas tekijä Z ei aiheuta X:n ja Y:n välistä riippuvuutta

. 4

Empiirinen jakauma Havaintomatriisin yhdestä sarakkeesta poimitut mittaluvut muodostavat tarkasteltavan muuttujan empiirisen jakauman. 3

Epäjatkuvan muuttujan tulkinta Joidenkin epäjatkuvien muuttujien (esimerkiksi opintopisteet) vaihteluväli voi olla hyvin laaja ja aineistoon voi muodostua useita hyvin pieniä luokkia. Tällöin epäjatkuva muuttuja voidaan tulkita jatkuvana, eli valitaan aineistolle sopiva luokitus ja jaotellaan havainnot sen mukaisiin ryhmiin. 3

Erot jakaumien sijainneissa ja vaihteluissa Erot jakaumissa voidaan karkeasti jakaa neljään ryhmään:

1. sijainti sama ja vaihtelu samanlaista,
2. sijainti eri, mutta vaihtelu samanlaista,

3. sijainti sama, mutta vaihtelu erilaista,
4. sijainti eri ja vaihtelu erilaista.

Jakaumien vertailu voidaan suorittaa sekä graafisesti että numeerisesti tunnuslukujen avulla. 4

Estimaatti Havaittu tilastollisen tunnusluvun arvo, jolla arvioidaan jotakin perusjoukon parametria.. 5

Estimaattori Otostunnusluku, jolla tiettyä mielenkiinnon kohteena olevaa perusjoukon jakauman parametria arvioidaan. Esimerkiksi keskiarvolla arvioidaan perusjoukon odotusarvoa. 5

F-jakauma Fisherin F-jakaumaa käytetään esimerkiksi varianssianalyysissä odotusarvojen yhtäsuuduuden testaamisessa. Jakauma on aina oikealle vino ja sen muodon määrää kaksi vapausastetta df_1 ja df_2, jolloin niitä vastaavaa F-jakaumaa merkitään $\mathrm{F}(df_1, df_2)$. F-jakauman taulukko on luentomonisteen liitteenä (taulukosta voidaan lukea F-testisuureen arvoja). 6

Fraktiili Esim. kvartiilit. p %:n fraktiili on muuttujan arvo, jota pienempiä havaintoja on p %. 4

Frekvenssi Yhdessä luokassa olevien havaintojen lukumäärä. Luokan i frekvenssiä merkitään f_i. 3

Frekvenssijakauma Laadullisen muuttujan tapauksessa (empiirisestä jakaumasta muodostettu) frekvenssijakauma kertoo, kuinka monta kertaa mitäkin muuttujan arvoa on havaittu. Jokaisesta muuttujan arvosta tulee frekvenssijakauman luokka (jollei käyttäjä yhdistele luokkia).

Määrällisen epäjatkuvan muuttujan frekvenssijakauma kertoo, kuinka monta kertaa mitäkin muuttujan arvoa ja mitä arvoja on havaittu.

Jatkuvan muuttujan tapauksessa empiirinen jakauma kertoo, mitä muuttujan arvoja on havaittu, kun taas varsinaisessa frekvenssijakaumassa muuttujan arvot on yleensä luokiteltu (jatkuvalle muuttujalle on yleensä tyypillistä, että useita havaintoarvoja on aineistossa vain yksi kappale). 3

Frekvenssitaulu frekvenssijakauman esitysmuoto, monisteesta ja laskuharjoituksista tuttu. 3

Graafinen esitys nominaaliast. muuttujalle Tavallisesti käytetään pylväsdiagrammia ja pylväät on järjestetty suurimmasta pienempään (nominaaliasteikkoisella muuttujalla ei ole järjestysrelaatiota luokkien suhteen). Pylväiden väliin jää tila, joka on noin puolet pylväiden leveydestä. Pylväsdiagrammi voidaan piirtää vaakatasossa tai pystytasossa sen mukaan, kumpi vaikuttaa järkevämmältä ratkaisulta. 3

Hajontaluku Numero, joka kuvaa havaintojen hajanaisuutta (tai, yhtä hyvin, keskittyneisyyttä), usein jonkin keskiluvun ympärillä (esimerkiksi keskiarvoa vastaa *keskihajonta*). 4

Harha Koeasetelma suosii systemaattisesti tiettyjä tuloksia (vrt. harhaisuus). 2

Harhaton estimaattori Hyvä estimaattori on *harhaton*, eli se antaa keskimäärin oikean tuloksen (ainakin suurilla otosmäärillä) ja harhattoman estimaattorin keskihajonta lähestyy nollaa otoskoon kasvaessa. Esimerkiksi keskiarvo $\overline{X}$ on perusjoukon odotusarvon μ harhaton estimaattori. 5

Harhattomuus Otoksesta lasketut tulokset ovat ”keskimäärin” oikeita. Yksittäiset otokset antavat kuitenkin erilaisia tuloksia riippuen siitä, mitkä tilastoyksiköt ovat valikoituneet otokseen. 2

Havaintoarvo Muuttujan havaittu arvo. Merkitään usein pienellä kirjaimella (vrt. muuttujaa merkitään isolla kirjaimella). 3

Havaintomatriisi Aineistoon keruun jälkeen havainnoista muodostettava taulukko, jota mm. tilastolliset ohjelmistot käyttävät aineiston tallennusmuotona. Havaintomatriisissa on kullakin rivillä yhteen tilastoyksikköön liittyvät tiedot jokaisesta muuttujasta ja yhdessä sarakkeessa yhden muuttujan havaitut arvot jokaiselle tilastoyksikölle. Havaintomatriisin ensimmäinen sarake on varattu tilastoyksiköiden ”nimille” ja on siten erityisasemassa (sitä ei pidä sekoittaa muihin sarakkeisiin). 3

Havaintoyksikkö Perusjoukon alkio. Kutsutaan myös nimellä *tilastoyksikkö*. 2

Heteroskedastisuus Kaksi riippumatonta tutkittavaa otosta ovat heteroskedastisia, mikäli niiden jakaumilla on erisuuret varianssit. 6

Histogrammi Määrällisen jatkuvan muuttujan frekvenssijakauma esitetään histogrammin avulla. Histogrammi tarkoittaa pylväsdiagrammia, jossa pylväät on piirretty kiinni toisiinsa ikään kuin kuvastamaan jatkuvuutta. Pylvään pinta-ala (leveys $\times$ korkeus) kuvaa frekvenssiä (usein kaikki pylväät ovat samanlevyisiä, jolloin korkeus kuvaa frekvenssiä käytännössä). 3

Indikaattorimuuttuja Mittaa epäsuorasti abstraktin käsitteen ominaisuutta, esimerkiksi kirkossa käyntien lukumäärä uskonnollisuutta joidenkin mielestä. 3

Intervalliasteikko Tunnetaan myös nimellä *Välimatka-asteikko*. Muuttujan mittaus kertoo mittauksen kohteen eron suuruuden toiseen kohteeseen verrattuna. Mittaluvuilla voidaan ilmaista ominaisuuden määrää, mittayksikkö on olemassa ja aritmeettisia laskutoimituksia saa tehdä. 3

Jakaumien sijaintiin liittyvät parametrittomat testit Kun t-testien ja varianssianalyysin oletukset eivät ole voimassa, voidaan käyttää parametrittomia testejä. Usein näin tehdään, kun tarkasteltava tilastollinen muuttuja on vain järjestysasteikollinen, ei-normaalinen tai jos varianssianalyysissä verrattavien ryhmien varianssit poikkeavat toisistaan. Parametrittömiä testejä ovat mm. seuraavat:

- Kahden riippumattoman otoksen Mann-Whitneyn U-testi

- Kahden riippuvan otoksen merkkitesti ja Wilcoxonin testi
- Usean riippumattoman otoksen Kruskal-Wallisin testi

. 6

Jatkuva muuttuja Muuttuja, joka voidaan (teoriassa) mitata kuinka tarkasti tahansa ja kahden mittausarvon välissä on aina muita arvoja (vrt. matematiikan käsite jatkuvuus). Muuttujan mahdollisia arvoja on tällöin ääretön määrä. 3

Jäännös Soluittain laskettava poikkeama eli havaitun solufrekvenssin ja odotetun frekvenssin erotus: $f_{ij} - e_{ij}$. Jos jäännös on positiivinen, niin havaintoja on liikaa verrattuna muuttujien riippumattomuustilanteeseen. Jos jäännös on negatiivinen, niin havaintoja on vastaavasti liian vähän. 4

Kahden muuttujan yhteisjakauma Tarkastellaan kahden muuttujan jakauman yhteisvaihtelua. Tavoitteena on tutkia, onko pareittain havaittujen muuttujien välillä tilastollista riippuvuutta. 4

Kaksisuuntainen vastahypoteesi Kun vastahypoteesi on muotoa "tarkasteltava asia eroaa nollahypoteesin H_0 mukaisesta tuloksesta", niin vaihtoehtoja on kaksi: vastahypoteesin mukainen tulos voi olla suurempi tai pienempi kuin nollahypoteesin mukainen tulos. Tällaista vastahypoteesia kutsutaan kaksisuuntaiseksi vastahypoteesiksi ja sitä on perusteltua käyttää silloin, kun ei tarkasti tiedetä, mihin suuntaan tutkittavan perusjoukon ominaisuus saattaa poiketa nollahypoteesin mukaisesta tilanteesta. 6

Kaksoissokkokoe Esimerkiksi lääkkeen testaus siten, että edes lääkärit eivät tiedä, kuka saa oikeaa lääkettä ja kuka lumelääkettä. 2

Kausaalisuus Syy-seuraussuhde vastemuuttujan ja selittäjän välillä, jota kokeessa etsitään tai siinä ilmenee. 2

Kertymäfunktio Satunnaismuuttujan X kertymäfunktio on $F(a) = \mathcal{P}(X \leq a)$. Se kertoo todennäköisyyden sille, että satunnaismuuttujan X arvo on korkeintaan a. Kertymäfunktion arvo on aina väliltä $[0, 1]$. Kertymäfunktiolla on seuraavat ominaisuudet:

- $\mathcal{P}(X > a) = 1 - \mathcal{P}(X \leq a) = 1 - F(a)$
- $\mathcal{P}(a < X < b) = F(b) - F(a)$
- Jos X on jatkuva, niin $\mathcal{P}(X \leq a) = \mathcal{P}(X < a)$ (yksittäisen arvon todennäköisyys $\mathcal{P}(X = a) = 0$)
- Jatkuvalle tiheysfunktiolle $f(x)$ pätee $F'(x) = f(x)$.

. 5

Keskiarvo (aritmeettinen) Muuttujaa X vastaavien havaintoarvojen keskiarvoa merkitään $\overline{x}$. Jos havaintoja on n kpl ja havaintoarvot ovat $x_1, ..., x_n$, niin keskiarvo lasketaan yhtälöstä

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

Keskiarvo kuvaa havaintojen keskimääräistä arvoa ja sitä voidaan käyttää välimatka-asteikoisille muuttujille. 3

Keskiarvon keskivirhe Keskiarvon keskihajontaa $SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ kutsutaan keskiarvon keskivirheeksi. Huomaa, että otoskoon n kasvaessa keskivirhe pienenee eli odotusarvon estimoinnin tarkkuus paranee. Keskiarvon keskivirhettä estimoidaan otoskeskihajonnan S avulla eli $SE(\overline{X}) = \frac{S}{\sqrt{n}}$. 5

Keskiarvon odotusarvo Tiedämme, että keskiarvo antaa keskimäärin oikean tuloksen (odotusarvon μ, ts. keskiarvon odotusarvo on $E(\overline{X}) = \mu$) (vrt. keskiarvo on odotusarvon harhaton estimaattori.). 5

Keskiarvonotantajakauma Olkoon satunnaisotos $X_1, X_2, \ldots, X_n$, missä X_i:t ovat riippumattomia ja olkoon $n \geq 30$. Tällöin

$$\overline{X} \simeq N(\mu, \frac{\sigma^2}{n}),$$

ts. keskiarvo noudattaa likimain normaalijakaumaa X_i:den jakaumasta riippumatta. Jos havainnot X_i noudattavat normaalijakaumaa $N(\mu, \sigma^2)$, niin keskiarvon $\overline{X}$ jakauma on täsmälleen normaalijakauma: $X \sim N(\mu, \frac{\sigma^2}{n})$. 5

Keskihajonta Varianssin s^2 neliöjuuri, eli

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

. Keskihajonnalla on sama mittayksikkö kuin keskiarvolla, joten se soveltuu hyvin ”luonnolliseksi” hajontaluvuksi. Keskihajontaa voi käyttää vain välimatka-asteikkoisille muuttujille. 4

Keskihajonta satunnaismuuttujalle Merkitään $SD(X) = \sqrt{\sigma^2} = \sigma$. Keskihajonta σ kuvaa satunnaismuuttujan arvojen vaihtelua odotusarvon μ ympärillä vastaavasti kuin otoskeskihajonta kuvaa otoksessa olevaa vaihtelua otoskeskiarvon $\overline{x}$ ympärillä.. 5

Khii toiseen Riippuvuusluku χ^2 on kaikkien standardoitujen jäännösten neliöiden summa ja se lasketaan yhtälöstä

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}\right]^2,$$

missä r ja s ovat rivien ja sarakkeiden määrät. Riippuvuusluvun arvo on aina vähintään nolla ja mitä suurempi se on, sitä vahvempaa on muuttujien välinen riippuvuus (samankokoisia ristiintaulukoita verrattaessa). χ^2-arvolla ei ole ylärajaa. Kun tulos halutaan yleistää perusjoukkoon, niin χ^2 suuruus yhdessä rivien ja sarakkeiden lukumäärän kanssa ratkaisee tuloksen tilastollisen merkitsevyyden. 4

Koesuunnitelma Selvitys mm. seuraavista asioista: tutkimuksen kohde, käsittelyt, vaste ja kontrollin muodot. 2

Koeyksilö kokeen kohde, ts. yksilö jolle koe tehdään (vrt. havaintoyksikkö). 2

Kokeellinen tutkimus Tehdään jokin koe (lääketieteellinen, luonnontieteellinen) ja analysoidaan siitä saatua aineistoa tilastollisilla menetelmillä. Tutkimus on koe, jos tutkimuksen kohteiden olosuhteita muutetaan tahallisesti (esimerkiksi annetaan lääkettä / ei anneta lääkettä). 2

Kokonaistutkimus Tutkitaan kaikki perusjoukon havaintoyksiköt. 2

Komplementtitapahtuma Tapahtuman vastatapahtuma. Jos A on tapahtuma, niin A:n komplementtitapahtumaa merkitään symbolilla A^c. Vastatapahtuman todennäköisyys $\mathcal{P}(A^c)$ on $1 - \mathcal{P}(A)$.. 5

Kontingenssikerroin Kontingenssikerroin saadaan χ^2-testisuureesta seuraavasti:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

missä n on otoksen koko. Kontingenssikertoimen maksimiarvo riippuu ristiintaulukon koosta kaavan

$$C_{max} = \sqrt{\frac{k-1}{k}},$$

missä $k = min\{r, s\}$. Havaitaan, että kontingenssikerroin on aina pienempi kuin 1 ja sen maksimiarvo lähestyy ykköstä, kun taulukon koko kasvaa. 4

Kontrollointi Kokeen varioiminen, vapausasteina mm. satunnaistaminen, toistaminen, vertailu ja lohkominen. 2

Korrelaatiokerroin Vakioluku, joka mittaa kahden muuttujan välisen lineaarisen riippuvuuden voimakkuutta ja laatua. Tällä kurssilla käytettyjä korrelaatiokertoimia ovat mm. Pearsonin tulomomenttikorrelaatiokerroin ja Spearmanin järjestyskorrelaatiokerrroin. 4

Korrelaatiomatriisi Jos tarkasteltavia muuttujia on enemmän kuin kaksi, niin eri muuttujaparien väliset korrelaatiokertoimet voidaan esittää taulukkona, jota kutsutaan korrelaatiomatriisiksi. Siihen laitetaan muuttujien välisistä ristiintaulukoista määritetyt korrelaatiokertoimien arvot. Diagonaalille tulee ykköset (muuttujan korrelaatio itsensä kanssa) ja yläkolmiossa samat korrelaatiot kuin alakolmiossa (esim. $r_{xy} = r_{yx}$). 4

Kovarianssi Muuttujien X ja Y kovarianssi s_{xy} on muuttujien välisen riippuvuuden mittari, joka riippuu aineistosta. Kovarianssi saadaan yhtälöstä

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Kovarianssit eri aineistoista eivät ole keskenään vertailukelpoisia ennen kuin ne on standardisoitu eli normitettu jakamalla muuttujia vastaavien keskihajontojen tulolla, ks. Pearsonin korrelaatiokerroin.. 4

Kovarianssimatriisi Kuten korrelaatiomatriisi, mutta taulukkoon merkitään korrelaatiokertoimien sijaan muuttujapareja vastaavat kovarianssit. Voidaankin ajatella, että kovarianssimatriisi on ns. normittamaton korrelaatiomatriisi, jonka normitus hoituu luonnollisesti kertomalla sitä sopivalla normitusmatriisilla (joka sisältää keskihajontojen käänteisluvut sopivasti aseteltuina). 4

Kvalitatiivinen ominaisuus Kvalitatiiviset eli *laadulliset* ominaisuudet jakavat tutkimuksen kohteet eri luokkiin, esimerkiksi sukupuoli, koulutus, älykkyys,.... 3

Kvantitatiivinen ominaisuus Ominaisuus, jota voidaan mitata reaalisesti (*määrällisesti*), kuten esimerkiksi ikä, pituus, lasten lukumäärä, lämpötila,.... 3

Kvartiiliväli Merkitään (Q_1, Q_3). Kvartiiliväli kertoo, millä "alueella" keskimmäiset 50 % havainnoista sijaitsevat. Ei ole yhtä herkkä äärimmäisille havainnoille kuin vaihteluväli. 4

Kvartiilivälin pituus Saadaan vähentämällä ala- ja yläkvartiili eli $Q_3 - Q_1 =: IQR$; vaatii muuttujalta vähintään välimatka-asteikollisuutta. 4

Kyselytutkimus Otantatutkimus, joka perustuu tutkijan tekemään lomakkeeseen, jossa on valmiit kysymykset. 2

Kysymystyypit Avoimet kysymykset, puoliavoimet kysymykset, monivalintakysymykset. 2

Käsittely Tasojen yhdistelmä eli "tietyt olosuhteet voimassa", tarkastellaan tilannetta ja verrataan muihin vastaaviin tilanteisiin). 2

Laadullisten ominaisuuksien mittaaminen Abstrakteille käsitteille, kuten älykkyys, uskonnollisuus jne. ei välttämättä ole olemassa yleisesti hyväksyttyä mittaria. Tilastollista tutkimusta varten tällainen tulee kuitenkin määritellä, jolloin mittarin validiteetin arviointi voi olla hankalaa. 3

Latentti ominaisuus Piilevä, vaikeasti mitattava asia (esimerkiksi älykkyys, kauneus,...). Mitataan epäsuorasti muiden mitattavien ominaisuuksien avulla (esimerkiksi missikisat ja numeerinen arviointi). 3

Lineaarinen riippuvuus Lineaarinen eli suoraviivainen riippuvuus tarkoittaa sirontakuvion näkökulmasta sitä, että kuvion pisteen näyttävät asettuvan jollekin koordinaatiston suoralle (edes suurin piirtein). Matemaattisesti voidaan kirjoittaa $y = a + bx$, missä y on muuttujan Y arvo ja x vastaavasti muuttujan X arvo (jokin piste suoralla). Vakio b on suoran kulmakerroin, joka kuvaa suoran jyrkkyyttä, ts. "kuinka paljon x muuttuu, jos y muuttuu jonkin annetun määrän" eli y on x:n monikerta tai päinvastoin. Numero a on vakio, joka kuvaa suoran ja y-akselin leikkauspisteen koordinaattia (usein tilastollisen tarkastelun kannalta ei ole mielenkiintoinen, kannattaa ajatella vakioterminä, joka nostaa tai laskee suoraa annetulla määrällä eli ei vaikuta kulmakertoimeen). Lisää lineaarisesta riippuvuudesta on regressiosuoran yhteydessä. 4

Luokitus Empiirisen jakauman havaitut arvot jaetaan sopiviin luokkiin siten, että jokainen havaintoarvo kuuluu täsmälleen yhteen luokkaan. Luokituksen valinta vaikuttaa jakauman muotoon ja näin ollen jatkoanalyysimenetelmiin joten sen valinta on aineistokohtaista. 3

Luottamusväli Satunnaisväli, joka peittää estimoitavan parametrin (usein odotusarvo) halutulla todennäköisyydellä. Tavallisesti käytetään 95 %:n ja 99 %:n luottamusvälejä (tarkoittaa sitä, että estimoidessa on 5 %:n tai vastaavasti 1 %:n todennäköisyys sille, että estimoitava parametri ei ole ks. välillä). 5

Marginaalijakauma Ristiintaulukon reunoille muodostuvat summajakaumat (sarakesummien ja rivisummien jakaumat). Huomaa, että molempien reunajakaumien frekvenssit ja solufrekvenssit summautuvat otoskooksi n. Kutsutaan myös *reunajakaumaksi*. 4

Mediaani Merkitään M_d. Se muuttujan arvo, jota pienempiä ja suurempia havaintoarvoja on yhtä paljon. Voidaan käyttää, jos muuttuja on vähintään järjestysasteikollinen. Kun havaintoja on parillinen määrä, niin mediaani on kahden keskimmäisen arvon keskiarvo (jos muuttuja on vähintään välimatka-asteikollinen). Kun muuttuja on järjestysasteikollinen, niin mediaaneja voi olla kaksi (jos ne ovat erisuuria) ja nämä ovat havaittuja arvoja. Mediaania voidaan kutsua myös *keskikvarttiiliksi*. 3

Merkitsevyystesti testin avulla tutkitaan sitä, kuinka harvinainen havaittu testisuureen arvo on nollahypoteesin mukaisessa tilanteessa. Mitä harvinaisempi testisuureen arvo on, sitä (tilastollisesti) merkitsevämpi tulos on saatu ja sitä enemmän uskotaan H_0:n vastahypoteesiin H_1.. 5

Mitta-asteikko Kertoo mitattavan muuttujan luonteen. Mitta-asteikon eri tyyppejä ovat *nominaaliasteikko*, *ordinaaliasteikko*, *intervalliasteikko* ja *suhdeasteikko*, joista viimeinen on intervalliasteikon erikoistapaus. 3

Mittaaminen Yritetään määrittää tutkittavan kohteen haluttu ominaisuus mahdollisimman suurella tarkkuudella. Mittaamiseen liittyy aina mittausvirhe.

$$\text{Havaittu muuttujan arvo} = \text{”oikea” arvo} + \text{systemaattinen virhe} + \text{satunnaisvirhe}$$

. 3

Mittaluku Muuttujaan mittaamisessa liitettävä arvo, joka riippuu tilastoyksilön/koeyksilön mitattavasta ominaisuudesta sekä mittaustavasta.. 3

Moodi Merkitään M_o, kutsutaan myös nimellä *tyyppiarvo*. Tarkoittaa sitä muuttujan arvo(luokka)a, jossa on eniten havaintoja. Toimii jo luokitteluasteikolla. Jos usealla luokalla on suurin frekvenssi, niin moodeja on useita. Ei sovi kovin hyvin jatkuville muuttujille. 3

Muunnos (tilastotestaaminen) Jos parametrisen testin oletukset eivät ole voimassa, voidaan parametrittömän testin käyttämisen sijaan tehdä sellainen muunnos alkuperäisiin havaintoihin, jolla esim. jakauman normaalisuus saavutetaan ja voidaan käyttää parametrista testiä. Muunnos tarkoittaa käytännössä matemaattista funktiota, jolla kuvataan havaintoarvojen empiirinen jakauma toiseksi jakaumaksi, jolle tehdään halutut tilastolliset testit. Muunnosta tehdessä tulee huomioida, että tulosten varsinaiset tulkinnat (esimerkiksi keskiarvoon liittyen) tehdään alkuperäisiä havaintoja käyttäen. 5

Muuttuja Kuvaa mielenkiinnon kohteena olevaa ominaisuutta, sen määrää tai laatua. Muuttujan nimi kirjoitetaan tilastotieteessä usein isolla kirjaimella. 3

Nollahypoteesi on oletus, joka on usein muotoa ”eroa ei ole” tai ”riippuvuutta ei ole” ja sitä merkitään symbolilla H_0. Nollahypoteesi antaa tilastolliselle analyysille kenties tylsän, mutta teoreettisesti ja laskennallisesti yksinkertaisen lähtökohdan. 5

Nominaaliasteikko Tunnetaan myös nimellä *luokitteluasteikko*. Kun muuttuja on mitattu nominaaliasteikolla, mittaus kertoo, mihin luokkaan havainto kuuluu. Luokkien välillä ei ole järjestystä ja luokat ovat keskenään samanarvoisia. Mittaluvut ovat luokkien tunnuksia, mittaluvuille ei voi suorittaa laskutoimituksia. 3

Normaalijakauma Tilastotieteissä ja todennäköisyyslaskennassa keskeisessä roolissa oleva jakauma. Normaalijakauman kuvaaja on tuttu kellokäyrä, joka on symmetrinen odotusarvokohtansa suhteen. Normaalijakaumaa merkitään $N(\mu, \sigma^2)$, missä μ on odotusarvo ja σ^2 on varianssi. Parametri μ kertoo jakauman sijainnin reaaliakselilla ja σ^2 ”kellon” leveyden. Kun satunnaismuuttuja X noudattaa normaalijakaumaa parametreilla μ ja σ, merkitään $X \sim N(\mu, \sigma^2)$.. 5

Normaalijakauman kertymäfunktio Kun satunnaismuuttuja X on standardoitu noudattamaan normaalijakaumaa $N(0, 1)$ (muuttujanvaihto $X \to Z$), niin standardoitua normaalijakaumaa noudattavan satunnaismuuttujan Z kertymäfunktiolle $\Phi(z)$ pätee

$$\Phi(z) = \mathcal{P}(Z \leq z) = \int_{-\infty}^{z} \phi(x)\, dx$$

Standardoidun normaalijakauman kertymäfunktion arvot voidaan lukea taulukosta. Negatiivisilla z:n arvoilla kertymäfunktion arvo saadaan kaavalla $\Phi(-z) = 1 - \Phi(z)$.. 5

Normaalijakauman tiheysfunktio $f(x)$ on muotoa

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

Tiheysfunktio on käytännön laskujen kannalta hankala, sillä sen integrointi (todennäköisyyden laskeminen) antaa tulokseksi erikoisfunktion. Tästä syystä normaalijakaumaan liittyvät todennäköisyydet luetaan useimmiten taulukoista.. 5

Näyte Tutkijan valitsema osajoukko perusjoukosta, josta ei voida tehdä koko perusjoukkoa koskevia päätelmiä. Valinta on usein tehty esimerkiksi jonkin halutun ominaisuuden perusteella. 2

Odotettu frekvenssi solun ij odotettu frekvenssi e_{ij} saadaan yhtälöstä

$$e_{ij} = f_{i.} \cdot \frac{f_{.j}}{n},$$

missä $f_{i.}$ on (i:nnen rivin rivisumma ja $f_{j.}$ on j:nnen sarakkeen sarakesumma. Odotettu frekvenssi saadaan siis jakamalla kaikki vaihtoehdon i vastanneet kaikkien vastanneiden määrällä ja kertomalla tulosta sarakesummalla $f_{.j}$, ks. luentomonisteen esimerkki. Odotetut frekvenssit ovat keskimääräisiä solufrekvenssejä tilanteessa, jossa rivi- ja sarakemuuttujat ovat toisistaan riippumattomia. Jos havaittu ja odotettu frekvenssi jossakin solussa poikkeavat selvästi toisistaan, niin muuttujien välillä voi olla riippuvuutta. 4

Odotusarvo satunnaismuuttujalle Merkitään μ. Odotusarvo on satunnaismuuttujan kaikkien mahdollisten arvojen painotettu keskiarvo, jossa painoina ovat arvojen todennäköisyydet. Se kertoo, millaisia arvoja satunnaismuuttujasta saadaan keskimäärin.
Epäjatkuvalle satunnaismuuttujalle odotusarvo määritellään yhtälöllä

$$\mu = \sum_{i=1}^{\infty} x_i \cdot p_i,$$

missä x_i on havaintoarvo ja p_i arvoa vastaava todennäköisyys. Jatkuvalle muuttujalle odotusarvo saadaan integraalina

$$\mu = \int_{-\infty}^{+\infty} x f(x)\ dx$$

. 5

Odotusarvoihin liittyvät testit Odotusarvoihin liittyvät neljä keskeistä parametrista tilastollista testiä ovat

- Yhden otoksen keskiarvotesti
- Kahden riippumattoman otoksen odotusarvojen yhtäsuuruuden testaus t-testillä
- Kahden riippuvan otoksen odotusarvojen yhtäsuuruuden testaus t-testillä eli parittainen t-testi
- Useamman kuin kahden riippumattoman otoksen odotusarvon yhtäsuuruuden testaus varianssianalyysillä

. 6

Odotusarvon luottamusväli tunnetulle varianssille Mikäli tutkittavan perusjoukon varianssi σ tunnetaan ja satunnaisotoksen muuttujat X_i noudattavat normaalijakaumaa, niin odotusarvon $100(1-\alpha)\%$:n luottamusväli voidaan kirjoittaa muodossa

$$\overline{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},$$

missä $z_{\frac{\alpha}{2}}$ on $N(0,1)$-jakauman $100(1-\frac{\alpha}{2})$ %:n fraktiili, esimerkiksi $z_{\frac{0,05}{2}} = 1,96$. Huomaa, että $\overline{x}$ on aina luottamusvälin keskipiste. Tätä luottamusvälin määrittelyä voidaan käyttää myös silloin, kun n on riittävän suuri, vaikka X_i:t eivät olisi normaalijakautuneet. 5

Odotusarvon luottamusväli tuntemattomalle varianssille Usein perusjoukon varianssia σ^2 ei tunneta, vaan se joudutaan estimoimaan. Tällöin käytetään luottamusvälin laskemiseen $t(n-1)$-jakaumaa (kun X_i:t noudattavat normaalijakaumaa). luottamusväli voidaan ilmaista muodossa

$$\overline{x} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}}$$

Luottamusväliä muodostettaessa voidaan t-jakauman sijaan käyttää normaalijakaumaa, kun otoskoko n on riittävän suuri. 5

Oikealle vino jakauma Jakauma, jonka oikea häntä on pidempi kuin vasen. Tämäntyyppiselle jakaumalle mediaani on pienempi kuin keskiarvo, $M_d < \overline{x}$. Mediaani sopii keskiarvoa paremmin tunnusluvuksi. 3

Ordinaaliasteikko Tunnetaan myös nimellä *järjestysasteikko*. Mittaus kertoo mittauksen kohteen järjestyksen (esim. paremmuus, suuruus) suhteessa toiseen kohteeseen. Mittaluvut ovat luokkien tunnuksia, joilla on lisäksi looginen järjestys. Mittaluvuilla ei ole määritelty mittayksikköä, eli suuruusjärjestyksessä etäisyydet eivät ole vakioita tai joskus edes järkevästi määritettävissä. Aritmeettisia laskutoimituksia ei voi suorittaa luokkien välillä.. 3

Osite Yksi ositetun otannan ryhmistä valitun taustatekijän jollakin ehdolla (esimerkiksi *Aku Ankkaa* lukevat miehet, jolloin taustatekijänä on lehden lukeminen). 2

Ositettu otanta Perusjoukko jaetaan ryhmiin eli osiitteisiin jonkin taustatekijän suhteen, josta on olemassa ennakkotietoa. Tutkittava ominaisuus on tavallisesti yhteydessä taustatekijään, jolloin ositetulla otannalla voidaan pienentää otantavirhettä. Ositetun otannan yhteydessä käytetään tasaista kiintiöintiä (jokaisesta osiitteesta poimitaan yhtä monta alkiota) tai suhteellista kiintiöintiä (poimitaan jokaisesta osiitteesta suhteellisesti sama määrä alkioita). 2

Otantamenetelmä Tapa, jolla perusjoukosta valitaan (satunnaisesti) tutkittava aineisto. 2

Otantatutkimus Tutkitaan perusjoukkoa edustava osajoukko, joka on hankittu otantamenetelmällä. 2

Otantavaihtelu Otoksen ja perusjoukon välinen ero on peräisin otantavaihtelusta, sillä otokseen valitut yksilöt vaihtelevat eri otoksissa. 2

Otantavirhe Otoksen ja perusjoukon alkioiden ominaisuuksissa oleva ero (seurausta siitä, että tilastoyksiköistä on tutkittu vain satunnaisesti valittu osajoukko ja otos ei täysin pysty kuvaamaan perusjoukkoa); aiheuttaa epävarmuutta tilastolliseen päättelyyn. 2

Otantayksikkö Yleensä tilastoyksikkö (esim. yksinkertainen satunnaisotanta) tai tilastoyksiköiden joukko (ryväsotanta). 2

Parametrinen testi Tilastolliset testit vaativat tietyt matemaattiset oletukset, jotta niitä voidaan käyttää. Yleensä havaintojen oletetaan olevan satunnaisia ja keskenään riippumattomia. Lisäksi oletukset voivat liittyä muuttujan mitta-asteikkoon tai jakaumaan. Testejä, jotka olettavat tutkittavilta muuttujilta normaalijakaumaa, kutsutaan usein *parametrisiksi testeiksi.*. 5

Parametriton testi Jolleivät valitun (parametrisen) testin oletukset ole voimassa, voidaan testi vaihtaa vastaavaan ns. *parametrittomaan* testiin, jolloin havaintojen taustalla olevista jakaumista ei tehdä oletuksia. Parametrittomat testit ovat vastaavia parametrisia testejä heikompia, ts. ne vaativat vahvemman näytön mahdollisesta erosta tai riippuvuudesta, jotta testin tulos olisi tilastollisesti merkittävä. 5

Pearsonin korrelaatiokerroin Olkoon tarkasteltavana muuttujista X ja Y tehdyt havainnot $(x_1, y_1), \ldots, (x_n, y_n)$. Pearsonin korrelaatiokerroin r_{xy} saadaan yhtälöstä

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x \cdot s_y},$$

missä $\overline{x}$ ja $\overline{y}$ ovat keskiarvot ja s_x ja s_y niitä vastaavat keskihajonnat. Yhtälön osoittajaa kutsutaan muuttujien X ja Y kovarianssiksi ja sitä merkitään s_{xy}. 4

Perusjoukko Tutkimuskohteiden joukko, josta halutaan tehdä päätelmiä. Tutkimus perustuu yleensä perusjoukon osajoukkoon (ellei kyseessä ole kokonaistutkimus). Perusjoukosta käytetään myös nimeä *populaatio*. 2

Piste-estimaatti Esimerkiksi otoksesta laskettu keskiarvo. Piste-estimaatti ei yksin kerro arvioinnin tarkkuudesta mitään, ja tästä syystä sen sijaan käytetään usein väliä (luottamusväliä), jonka peittää tarkasteltavan parametrin arvon annetulla (suurella) todennäköisyydellä. 5

PNS-menetelmä Sirontakuvioon sovitetaan suora, joka sopii siihen ”mahdollisimman hyvin.” PNS-menetelmässä lasketaan suoralle kertoimet, ks. regressiosuora . 4

Poikkeava havainto Poikkeava havainto eroaa muista ”tavallisista” havainnoista joko poikkeuksellisen suuren/pienen arvonsa vuoksi tai kuuluu laadullisen muuttujan tapauksessa eri luokkaan kuin muut vastaukset. Syynä havaintoon voivat olla mm. mittausvirheet tai aineiston käsittelyssä tapahtuneet virheet, vieraasta perusjoukosta peräisin oleva havainto (esimerkiksi jonkin taustatekijän suhteen tarkasteltuna) tai tilastoyksikön erikoislaatuinen luonne. Poikkeava havainto voidaan poistaa, voidaan käyttää robusteja menetelmiä (joissa havainto ei merkittävästi muuta mm. tilastollisia tunnuslukuja) tai sitten hyväksyä havainto sellaisenaan. 3

Prosentuaalinen frekvenssijakauma Muodostuu luokista i ja suhteellisten frekvenssien prosenttiosuuksista $100 \cdot p_i$ %. 3

Prosentuaalinen summafrekvenssijakauma Kuten suhteellinen summafrekvenssijakauma, jossa suhteelliset osuudet on kerrottu sadalla (saadaan prosenttiosuus). 3

Regressioanalyysi Yritetään löytää tutkittavalle ilmiölle malli, joka on riittävän yksinkertainen ja tulkinnallinen ja samalla mahdollisimman selitysvoimainen. 4

Regressiokerroin Regressiosuoran yhtälössä regressiokerroin kuvaa x-muuttujan vaikutusta y-muuttujaan. 4

Regressiosuora Kun muuttujien välinen tilastollinen riippuvuus on lineaarista, niin sitä voidaan mallintaa ja analysoida ns. regressiosuoralla, joka sovitetaan muuttujien X ja Y pisteparien (x_i, y_i) joukkoon. Sovitettu käyrä on muotoa $y = bx + a$ (vrt. lineaarinen riippuvuus), missä a ja b ovat vakioita, jotka määräytyvät pisteparien avulla seuraavasti:

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x}^2)}$$
$$a = \overline{y} - b\overline{x}$$

Edellisillä ehdoilla tehtyä suoran sovitusta kutsutaan *pienimmän neliösumman suoraksi* (PNS-suora). 4

Reliabiliteetti Toistettavuus, mittarin luotettavuuus satunnaisvirheen näkökulmasta. Hyvä reliabiliteetti tarkoittaa, että mittaus on toistettava, ts. useat samoissa olosuhteissa tehdyt mittaukset antavat saman tuloksen (tämä tulos ei välttämättä ole ns. ”oikea” arvo eli systemaattista virhettä voi esiintyä). Jos reliabiliteetti on hyvä, niin tulosten hajonta on verraten pieni. 3

Riippumaton muuttuja kuvaa kohteen olosuhdetta, jota muuttamalla saadaan mahdollisesti muutosta vastemuuttujaan (selittää muutosta vastemuuttujassa). Tunnetaan myös nimillä *selittäjä* ja *faktori*. 2

Riippuvuusluvut Lukuja, jotka mittaavat ristiintaulukoinnissa ilmenneen riippuvuuden voimakkuutta. Riippuvuuslukuja on useita, mm. χ^2 (khii toiseen), C (kontingenssikerroin), RR (riskisuhde),. . . . 4

Riippuvuuteen liittyvät testit Kun halutaan tutkia kahden muuttujan välistä tilastollista riippuvuutta, asetetaan tutkimuksen nollahypoteesiksi se, että riippuvuutta ei ole. Tutkimushypoteesina on, että riippuvuus on olemassa. Jos muuttujat ovat luokitteluasteikkoisia (laadullisia), käytetään χ^2-riippumattomuustestiä. Vastaavasti voidaan kahdelle välimatka-asteikkoiselle muuttujalle testata korrelaatiokerrointa. Myös regressiosuoran parametreja voidaan testata, ts. onko tutkittavien muuttujien välillä lineaarista riippuvuutta. 6

Riskisuhde Dikotomiselle muuttujalle riskisuhde RR lasketaan yhtälöstä

$$RR = \frac{f_{11}/f_{.1}}{f_{12}/f_{.2}},$$

ks. luentomonisteen esimerkit rokotuksen saaneista ja ei-saaneista sekä koleraan/polioon sairastuneista. Jos riskisuhde on $RR = 1$, niin taulukossa ei ole riippuvuutta. Jos RR poikkeaa merkittävästi ykkösestä, niin riippuvuus on olemassa (riippuvuuden mielekkyys riippuu aineistosta). 4

Ristiintaulukko Kaksiulotteinen frekvenssitaulu, jossa on yksi muuttuja sarakemuuttujana ja toinen rivimuuttujana (valinta riippuu aineistosta; yleensä sarakemuuttujana on selittäjä ja rivimuuttujana selitettävä muuttuja). Tunnetaan myös nimellä *kontingenssitaulu*. 4

Riviprosentti Ristiintaulukoinnin yhteydessä; kuten sarakeprosentti, mutta ehtona on rivillä oleva muuttujan arvo. 4

Rivisumma Sarakkeessa ”yht.” on rivimuuttujan frekvenssit luokittain. Yksittäinen frekvenssi $f_{i\cdot}$ on nimeltään rivisumma, joka saadaan laskemalla yhteen rivimuuttujan luokan i frekvenssit (kaikille j). 4

Robusti menetelmä Tilastollinen menetelmä tai tunnusluku, joka ei muutu kovin paljon, jos aineistoon lisätään muusta joukosta selvästi poikkeava yksittäinen havainto (esimerkiksi mediaani). 3

Robusti tunnusluku Tilastollinen tunnusluku, joka ei ole herkkä yksittäisille poikkeaville havainnoille (esimerkiksi mediaani). Keskiarvo ei ole robusti. 3

Ryväsotanta Käytetään suurissa otoksissa. Esimerkiksi tutkittaessa kaikkia Suomen koululaisia jaetaan perusjoukko (opiskelijat) ryppäisiin koulukohtaisesti, jolloin otantayksiköksi tulee ryväs. Poimitaan jollakin valitulla otantamenetelmällä sopiva joukko rypäitä otannan ensimmäisessä vaiheessa, ja sen jälkeen toisessa vaiheessa jo poimituista rypäistä arvotaan haluttu määrä opiskelijoita lopulliseen otokseen. Ryväsotanta on käytännöllisempi vaihtoehto suurissa tutkimuksissa kuin suora satunnaisotanta tai systemaattinen otanta.. 2

Sarakeprosentti Ristiintaulukoinnin yhteydessä; ehdollinen prosentuaalinen frekvenssijakauma, ehtona on sarakkeessa oleva muuttujan arvo. 4

Sarakesumma Kuten rivisumma, merkitään $f_{\cdot j}$ ja lasketaan yhteen sarakemuuttujan luokan j frekvenssit (kaikille i). 4

Satunnaisilmiö Ilmiö, johon liittyy useita eri tulosvaihtoehtoja, esimerkiksi kolikon heitto, lapsen syntymä,.... 5

Satunnaisotos Tutkittavat tilastoyksiköt on valittu perusjoukosta satunnaisesti (oleellista on, että tutkittavat yksilöt muodostavat pienoiskuvan otoksesta). 2

Satunnaisvirhe Virhe, jonka suunnasta ei ole tietoa, esimerkiksi mittarin arvot vaihtelevat molemmin puolin ”oikeaa” arvoa. Myös esimerkiksi inhimilliset, satunnaiset huolimattomuusvirheet. 3

Selitysaste Mittaa selittävän muuttujan X selitysvoimaa: $R^2 = r_{xy}^2 \cdot 100\ \%$. Maksimiarvo saavutetaan täydellisen riippuvuuden tilanteessa. Selitysaste kertoo, kuinka monta prosenttia muuttuja X selittää muuttujan Y vaihtelusta. Jos jäännökset ovat suuria, niin silloin korrelaatio ja selitysaste ovat pieniä (lähellä nollaa). 4

Sirontakuvio Piirretään tavalliseen xy-koordinaatistoon kahden muuttujan X ja Y vastaavat arvot pistepareina (x, y). Sirontakuvio tunnetaan myös nimillä *pisteparvi*, *hajontakuvio* ja *korrelaatiodiagrammi*. Se kuvaa kahden muuttujan välistä (tilastollista) riippuvuutta ja yhteisjakaumaa. 4

Solufrekvenssi Yhdessä solussa ij olevien havaintojen lukumäärä, merkitään f_{ij}. 4

Standardoitu jäännös Jäännös jaetaan odotetun frekvenssin neliöjuurella

$$\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

. 4

Standardoitu normaalijakauma Satunnaismuuttuja X, joka noudattaa normaalijakaumaa, voidaan standardoida muunnoskaavalla

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

Standardointi tarkoittaa käytännössä sitä, että uusi muuttuja Z noudattaa normaalijakaumaa $N(0, 1)$, ts. $Z \sim N(0, 1)$, ja jakaumaa $N(0, 1)$ vastaavat normaalijakauman kertymäfunktion arvot eli todennäköisyydet voidaan lukea taulukosta. 5

Suhdeasteikko Intervalliasteikon erikoistapaus, jolla on absoluuttinen nollapiste, jossa mitattava ominaisuus häviää (esimerkiksi kohteen massa). Voidaan sanoa, kuinka monta kertaa suurempi tai pienempi kohde on verrattuna toiseen kohteeseen.. 3

Suhteellinen frekvenssi Merkitään $p_i = f_i/n$, missä n on kaikkien havaintojen määrä (ts. luokan i havaintojen lukumäärän osuus kaikista havainnoista). 3

Suhteellinen frekvenssijakauma Muodostuu luokista i ja suhteellisista frekvensseistä p_i. 3

Suhteellinen kiintiöinti Ositetun otannan yhteydessä poimitaan jokaisesta ositteesta suhteellisesti (prosentuaalisesti) sama määrä alkioita. 2

Suhteellinen osuus perusjoukossa Tarkasteltavan joukon suhteellista osuutta perusjoukossa (esimerkiksi lohien osuus kaikista Suomen kaloista) merkitään π:llä. Selvästi nähdään, että π on todennäköisyys, $0 \leq \pi \leq 1$.. 5

Suhteellinen summafrekvenssijakauma Kuten suhteellinen frekvenssijakauma, mutta lasketaan luokan i ja sitä edeltävien luokkien suhteellinen osuus (näiden summa siis...) luokan i suhteelliseksi osuudeksi. Tunnetaan myös nimellä *otoskertymäfunktio*. 3

Suhteellisen osuuden estimointi Suhteellista osuutta perusjoukossa arvioidaan estimaattorilla $\mathcal{P}$, joka saa eri arvoja p (suhteellisen osuuden estimaatteja) otoksesta riippuen. Estimaattori antaa keskimäärin oikean tuloksen eli $\mathcal{P}$:n odotusarvo on π. Estimaattorin varianssi on $Var(\mathcal{P}) =$

$\frac{\pi(1-\pi)}{n}$ ja keskihajonta eli keskivirhe $SD(\mathcal{P}) = \sqrt{\pi(1-\pi)/n}$. Kuten arvata saattaa, keskivirheen estimaattori on muotoa $\sqrt{\mathcal{P}(1-\mathcal{P}/n)}$. Estimaattorin $\mathcal{P}$ otantajakauma on likimain normaalinen silloin, kun otoskoko n on suuri, ts. $n\pi > 5$ ja $n(1-\pi) > 5$.. 5

Suhteellisen osuuden luottamusväli Standardoidun tunnusluvun

$$Z = \frac{\mathcal{P} - \pi}{SD(\mathcal{P})} = \frac{\mathcal{P} - \pi}{\sqrt{\pi(1-\pi)/n}}$$

otantajakauma on likimain normaalinen, ts. $Z \simeq N(0,1)$, kun otoskoko n on suuri eli $n\pi > 5$ ja $n(1-\pi) > 5$. Luottamusvälille saadaan näin ollen esitys

$$\mathcal{P} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\mathcal{P}(1-\mathcal{P})}{n}},$$

missä $z_{\frac{\alpha}{2}}$ on $N(0,1)$-jakauman $100(1-\frac{\alpha}{2})$ %:n fraktiili ja keskivirheen estimaattori on $\sqrt{\mathcal{P}(1-\mathcal{P})/n}$. Käytännössä luottamusväli lasketaan lausekkeesta

$$p \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}},$$

missä p on otoksesta laskettu tunnusluku. 5

Suhteellisiin osuuksiin liittyvät testit Testejä on kaksi kappaletta:

- Yhden otoksen suhteellisen osuuden testi, jota käytetään yhden muuttujan (ja jakauman) suhteellisen osuuden testaukseen.
- Suhteellisten osuuksien yhtäsuuruuden testi kahdelle riippumattomalle otokselle, jota käytetään nimensä mukaisesti kahden riippumattoman muuttujan (ja jakauman) suhteellisiin osuuksiin.

Testit ovat periaatteiltaan hyvin samanlaisia. 5

Summafrekvenssi Tunnetaan myös nimellä *kumulatiivinen frekvenssi.* Kun muuttuja on vähintään järjestysasteikkoinen, voidaan laskea, kuinka monta havaintoarvoa kuuluu luokkaan i ja sitä edeltäviin luokkiin. Summafrekvenssi on näiden havaintojen summa (luokassa i). Summafrekvenssit muodostavat summafrekvenssijakauman. 3

Summafrekvenssijakauma Summafrekvensseistä ja luokista muodostettu jakauma. Käytetään data-analyysissä siihen, kun arvioidaan teoreettisen jakauman sopivuutta empiiriseen jakaumaan. 3

Symmetrinen jakauma Keskiarvon oikealla ja vasemmalla puolella on (suurin piirtein) yhtä paljon havaintoja. Tällöin mediaani ja keskiarvo ovat likimain yhtäsuuria, $M_d \approx \overline{x}$. Keskiarvo sopii hyvin tunnusluvuksi, jos muuttuja on välimatka-asteikollinen. 3

Systemaattinen otanta Kirjoitetaan luettelon havaintoyksiköistä. Arvotaan yksi alkio ja poimitaan siitä lähtien systemaattisesti joka k. alkio eteen- ja taaksepäin (poimintaväli $k = N/n$, missä n on poimittavien alkioiden lukumäärä; pyöristetään tarvittaessa). Vaihtoehtoisesti voidaan

- jakaa tilastoyksiköt k:n alkion ryhmiin (n kappaletta ryhmiä),
- arpoa satunnaisluku m väliltä $[1, k]$ ja
- poimia ensimmäisestä ryhmästä m. alkio ja siitä eteenpäin joka k. alkio.

Täytyy pitää huoli siitä, että luettelossa ei ole jaksollisuutta mielenkiinnon kohteena olevan ominaisuuden suhteen (esimerkiksi sunnuntain lehden mainosten määrä tms). 2

Systemaattinen virhe Virhe, jonka suunta tunnetaan. Esimerkiksi lämpömittari, joka näyttää aina kolme astetta enemmän kuin ”oikein” näyttävä lämpömittari. 3

t-jakauma Studentin t-jakauma on jatkuva todennäköisyysjakauma, jota käytetään normaalijakautuneen aineiston analysointiin silloin, kun otoskoko n on pieni. Kun otoskoko kasvaa, lähestyy t-jakauma normaalijakaumaa. t-jakaumaa käytetään luottamusvälin laskemisessa silloin, kun tutkittavan perusjoukon varianssia ei tunneta. Jakauman fraktiilit eri otoskoille on taulukoitu monisteen loppuun. 6

Tapahtuma Mielivaltainen ϵ:n (satunnaisilmiöön liittyvä perusjoukko) osajoukko, joka koostuu yhdestä tai useammasta alkeistapauksesta. 5

Tasainen kiintiöinti Ositetun otannan yhteydessä poimitaan jokaisesta ositteesta absoluuttisesti sama määrä alkioita riippumatta ositteen koosta. 2

Taso Riippumattoman muuttujan (selittäjän) valittu arvo. 2

Teoreettinen jakauma Usein perusjoukon kiinnostavaa ominaisuutta kuvaavan muuttujan jakaumaa mallinnetaan jokin satunnaismuuttujan jakauman, esim. normaalijakauman, ja sen parametrien avulla. Tätä perusjoukon jakaumaa kutsutaan *teoreettiseksi jakaumaksi,* kun vastaava havaintoaineistossa oleva muuttujan jakauma on *empiirinen jakauma.*. 5

Testin voimakkuus Todennäköisyyttä tehdä tilastotestissä toisen luokan virhe merkitään β. Tämän tapahtuman vastatapahtuman todennäköisyys on

$$1 - \beta = \mathcal{P}(H_0 \text{ hylätään} | H_0 \text{ on epätosi})$$

ja sitä kutsutaan testin *voimakkuudeksi.* Voimakkuuden avulla voidaan etukäteen arvioida mm. sitä, kuinka suuri otoskoko tarvitaan määrätyn suuruisen keskiarvoeron havaitsemiseen testissä. 5

Tiheysfunktio Jatkuvan satunnaismuuttujan todennäköisyysjakauma määritellään satunnaismuuttujan tiheysfunktiolla. Tiheysfunktio $f(x)$ on (yleensä) jatkuva ja sillä on ominaisuudet

$$f(x) \geq 0 \qquad \text{ja} \qquad \int_{-\infty}^{+\infty} f(x)\, dx = 1,$$

ts. funktion f ja x-akselin rajoittama pinta-ala on 1. Tiheysfunktion avulla todennäköisyys voidaan laskea mille tahansa välille $[a, b]$ integraalina $\int_a^b xf(x)\, dx$.. 5

Tilastollinen analysointi Tilastollinen analysointi jakautuu kuvailevaan analyysiin ja tilastolliseen päättelyyn. Kuvailevassa analyysissä selvennetään tutkittavan ilmiön luonnetta sekä yritetään löytää poikkeavia havaintoja. Tilastollisessa päättelyssä tehdään koko perusjoukkoa koskevia päätelmiä osajoukon perusteella. Tutkimus on induktiivista (yksittäistapauksista yleiseen etenevää), joten tulokset ovat epävarmoja ja pitävät paikkansa tietyllä todennäköisyydellä. 2

Tilastollinen hypoteesi on ennakkoon asetettu oletus koskien perusjoukkoa, jakaumaa tai jakauman parametreja. Sen paikkansapitävyyttä testataan (tilastollisesti) käytettävissä olevan aineiston perusteella. 5

Tilastollinen riippuvuus Jos kahden muuttujan välillä on tilastollinen riippuvuus, voidaan tutkitun aineiston asettamissa rajoissa ennustaa toisella muuttujalla toisen käyttäytymistä. Mikäli toisistaan riippuvat muuttujat ovat X ja Y, niin voidaan (matemaattisessa mielessä) sanoa, että Y on X:n funktio tai päinvastoin. Tämä riippuvuusfunktio voi olla periaatteessa mikä tahansa, mutta tällä kurssilla keskitytään erityisesti lineaariseen riippuvuuteen. 4

Tilastollinen tutkimus Reaalimaailmaan liittyvää tutkimusongelmaa yritetään tilastollisilla menetelmillä kuvailla ja pienellä otoksella tehdä suurempaa joukkoa koskevia päätelmiä. Pyritään erottamaan toisistaan ilmiöiden säännönmukaisuudet ja satunnaiset piirteet. 2

Tilastollisen testin virhetyypit Tilastollisen testin soveltaminen voi johtaa väärään päätelmään. Virhetyyppejä on kaksi:

- Hylätään H_0, vaikka se on tosi (ns. 1. lajin virhe)
- Ei hylätä H_0:aa, viakka se on epätosi (ns. 2. lajin virhe)

Molempien virhetyyppien mahdollisuuttta voidaan pienentää kasvattamalla otoskokoa n. 5

Tilastollisen tutkimuksen työvaiheet Suunnittelu, aineiston hankintamenetelmät, aineiston kuvaamismenetelmät ja tilastollinen päättely. 2

Tilastotiede Menetelmätiede, jonka tavoitteena on sellaisten menetelmien kehittäminen, joilla reaalimaailman ilmiöitä ja niihin liittyviä aineistoja voidaan kuvailla, selittää, ennustaa ja hallita. 2

Toisistaan riippuvat mittaukset Esimerkiksi kokeellisessa tutkimukseessa (lääkekoe tms.) voidaan käyttää samoja henkilöitä useaan kertaan (mittausarvo ennen ja jälkeen lääkkeen määräämistä). Tällöin mitattavat muuttujat muodostavat parin (X_{A_i}, X_{B_i}), missä X_{A_i} on mittaus ennen käsittelyä ja X_{B_i} mittaus käsittelyn jälkeen, $i = 1, \ldots n$. Sanotaan, että mittaukset ovat parittaisia eli riippuvat toisistaan. 6

Toistaminen Koe tehdään samoissa olosuhteissa usealle koeyksilölle. 2

Tunnusluvuilla estimointi Otoksesta lasketuilla tunnusluvuilla pyritään arvioimaan perusjoukon vastaavia tunnuslukuja, esimerkiksi keskiarvoa, keskihajontaa tai porosenttiosuutta perusjoukossa. 5

Vaihteluväli Merkitään (min, max), missä min on aineiston pienin arvo ja max on suurin arvo. 4

Vaihteluvälin pituus Merkitään $R = max - min$, aineiston pienimmän ja suurimman arvon erotus. 4

Validiteetti Mittarin kyky mitata haluttua ominaisuutta systemaattisen virheen mielessä: mittari on validi, jos mittausta toistettaessa saadaan keskimäärin oikea arvo (vastausten hajonta oikean arvon ympärillä voi olla merkittävä). 3

Variaatiokerroin Määritellään yhtälöllä $CV = s/\overline{x}$, eli keskihajonnan suhteena keskiarvoon. Variaatiokerroin edellyttää muuttujalta suhdeasteikkoa. Variaatiokertoimen avulla voidaan vertailla eri mittayksiköillä mitattujen muuttujien suhteellisia hajontoja toisiinsa. 4

Varianssi Hajontaluku, joka lasketaan yhtälöstä

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

missä n on otoskoko ja $\overline{x}$ on muuttujaa X vastaavien havaintojen keskiarvo. Huomaa, että aina pätee $s^2 \geq 0$. Varianssia voi käyttää vain välimatka-asteikkoisille muuttujille. Joskus varianssi kutsutaan myös *otosvarianssiksi*. 4

Varianssi satunnaismuuttujalle Merkitään $\sigma^2 = Var(X) = E(X - \mu)^2$. Varianssi on satunnaismuuttujan arvon keskimääräinen neliöpoikkeama odotusarvostaan (vrt. otosvariassi s^2).. 5

Vasemmalle vino jakauma Jakauma, jonka vasen häntä on pidempi kuin oikea. Tämäntyyppiselle jakaumalle mediaani on suurempi kuin keskiarvo, $M_d > \overline{x}$. Mediaani sopii keskiarvoa paremmin tunnusluvuksi. 4

Vastahypoteesi on yleensä varsinainen sisällöllinen tutkimushypoteesi, joka (tavallisesti) on nollahypoteesin antiteesi, ts. H_0 ei ole voimassa. Vastahypoteesia merkitään H_1. Vastahypoteesi voi olla esimerkiksi muotoa ”eroa on” tai ”riippuvuutta on”. Tilanteesta riippuen valitaan joko kaksi- tai yksisuuntainen vastahypoteesi. 5

Vastauskato Kyselytutkimuksessa: Kohderyhmän henkilöt eivät vastaa kyselyyn tai kyselyyn annetut vastaukset katoavat jostakin syystä. Jos kato on merkittävä, niin joudutaan suorittamaan jälkitoimenpiteitä, mm. lähettää uusia kyselyjä tai korvata henkilöitä samantyyppisillä henkilöillä. 2

Vastemuuttuja Kuvaa koeyksilön tutkittavaa ominaisuutta. Tunnetaan myös nimellä *selittävä muuttuja*. 2

Viiksilaatikko Empiirisen jakauman esittäminen graafisesti tiivistetyssä muodossa seuraavien tunnuslukujen avulla:

$$\{min, Q_1, M_d, Q_3, max\}$$

Mallikuvia löytyy luentomonisteesta. 4

Yksinkertainen satunnaisotanta Kirjoitetaan luettelo, joka kattaa kaikki perusjoukon alkiot. Alkiot numeroidaan järjestyksessä $1,\ldots,N$ ja numeroidusta joukosta poimitaan arpomalla haluttu määrä numeroita väliltä 1–N. 2

Yksisuuntainen vastahypoteesi Jos vastahypoteesi on muotoa ”tutkittava asia on suurempi” tai vaihtoehtoisesti ”tutkittava asia on pienempi” kuin nollahypoteesin H_0 mukainen tulos, niin puhutaan yksisuuntaisesta vastahypoteesista. Tällöin toinen suunta on perusteltua jättää tarkistamatta. Jos esimerkiksi tutkitaan, vähentääkö rokotus sairastuvuutta, on mielekästä käyttää yksisuuntaista vastahypoteesia ”rokotuksen saaneista sairastuneiden suhteellinen osuus on pienempi kuin vertailuryhmän sairastuneiden osuus” tai jotakin vastaavaa tilanteeseen sopivaa. 6

Yläkvartiili Voidaan määritellä usealla eri tavalla. Alakvartiili on

1. sellainen muuttujan arvo, jota pienempiä havaintoja aineistossa on 3/4, tai
2. mediaania suurempien havaintojen mediaani, tai
3. sen luokan muuttujan arvo, jossa prosentuaalinen summafrekvenssi saavuttaa 75 %.

Tilasto-ohjelmat saattavat laskevat kvartiileja hieman eri tavoin, joten niiden antamat tulokset voivat hieman poiketa toisistaan. 4