



Standardointi tekee eri asteikkolisista muuttujista vertailukelpoisia

- Standardointi (normeeraus): eli muuttujan arvosta vähennetään sen keskiarvo ja jaetaan hajonnalla

→ arvot verrattavissa minkä muun normeeratun muuttujan kanssa

→ puhutaan Z-pisteistä

→ voidaan muodostaa esim. summamuuttujia

→ miten z-pisteet tehdään

Analyze → Descriptive statistics

→ Descriptives

1



Mitta-asteikon vaikutus analyysimenetelmän ja merkitsevyydestin valintaan

- Peruskysymys: onko asteikko luonteeltaan luokitteleva vai jatkuva?
- Keskiarvon laskeminen joko on mahdollista (ikä, kuukausitulot jne.) tai sitten ei (siviilisääty, sukupuoli jne.)
- Laatuasteikkolisien muuttujien lisäksi luokitteleviksi luetaan järjestysasteikkoliset muuttujat, joilla kylläkin on suunta, mutta luokkaväleistä ei vallitse yksimielisyyttä (esim. ammatit)
- Luokitteleville muuttujille soveltuvat analyysit ja tilastolliset testit pohjautuvat joko ristiintaulukoon tai non-parametriin menetelmiin

2



Tilastollisen menetelmän valinta

		Selitettävä -, riippuva -, y-muuttuja	
		Luokitteleva tai järjestysasteikko	Välimatka tai suhteasteikko
Selittävä -, riippumaton -, X -muuttuja	Luokitteleva tai järjestysasteikko	Ristiintaulukointi, Loglineaariset mallit	T-testi, Manova, Anova
	Välimatka tai suhteasteikko	Logistinen ja multinomiaalinen regressioanalyysi	Regressioanalyysi, polku- ja rakenneyhtälömallit

3



- Jatkuville muuttujille (välimatka- ja suhteasteikkolisille muuttujille, joissa mitattu ominaisuus kasvaa tasaisesti muuttujien arvojen kasvaessa) ovat luovallisia kutakuinkin kaikki keskiarvoon perustuvat monimuuttujaiset korrelaatio- ja varianssipohjaiset menetelmät

Ristiintaulukointi eli kontingenssitalukointi

Käyttö:

- Aineiston kuvaaminen
- Kahden luokittelevan muuttujan välisen yhteyden selvittäminen
- Muuttujien jakaumien vertaaminen eri ryhmissä
- Soveltuu kategorisille (laatu- tai järjestysasteikko) muuttujille ja luokitelluille numeerisille (välimatka- tai suhteasteikko) muuttujille

4



Ristiintaulukointi

Päätely:

- Jakaumien silmämääräinen ja numeerinen tarkastelu
- Miten havainnot sirontuvat ja painottuvat soluihin, missä soluissa suurimmat frekvenssit?
- Rivi- ja sarakeprosentit (ks. spss)
- Vertailu odotusfrekvensseihin

Todellinen aineisto:
VSM-taso x 8. lk
2. aste
koulutustaso
KOE = 2 tai useampi virhe
KONTROLLI = 0-1
virhetta

	1=lukio	2=amat tikoulu	3=en tiedä	YHT.
Koe	1	4	4	9
Kontrolli	7	1	1	9
YHT.	8	5	5	18

5



Riippuvuuden tilastollisen merkitsevyyden testaaminen

- X²-testi (Khiin neliön testi, riippumattomuustesti)

Hypoteesit ja niiden testaus:

- H₀= muuttujat ovat riippumattomia
- H₁= muuttujat eivät ole riippumattomia
- Vapausaste (df) lasketaan taulukon rivien ja sarakkeiden lukumäärästä, $df = (r-1)(s-1)$
- Testin havaittu merkitsevyytaso eli p-arvo riippuu testisuureen arvosta ja vapausasteesta
- P-arvo kertoo erehtymisriskin suuruuden, kun testattava nollahypoteesi hylätään (eli mikä on riski sille, että riippuvuus johtuu sattumasta)
- Yleensä hylätään nollahypoteesi (eli todetaan, ettei aineisto tue nollahypoteesia), jos erehtymisriski on korkeintaan 5% eli p-arvo < 0,05)

6



Ristiintaulukoinnin edellytykset

Edellytykset:

enintään (max.) 20% odotetuista frekvensseistä saa olla pienempiä kuin 5, ja

pienin odotettu frekvenssi > 1 eli ei saa olla odotusarvoltaan tyhjiä soluja

tai
odotetulta frekvenssiltään tyhjiä soluja saa olla siellä täällä, ei kuitenkaan kokonais nollarivejä tai -sarakkeita

7

Parametriset ja Ei-parametriset testit

Tilastolliset testit voidaan jakaa parametrisiin ja ei-parametrisiin testeihin sen mukaan, minkälaisia jakaumia testit käyttävät.

Parametrisilla testeillä on jakaumaoletuksia.

- vähintään välimatka-asteikko
- jakaumien normalisuus

Ei-parametrisillä testeillä ei ole jakaumaoletuksia.

- testit eivät kuitenkaan yhtä voimakkaita kuin parametriset testit, jonka vuoksi kannattaa käyttää parametrisiä aina kuin siihen on mahdollisuus

Parametrinen

Pearsonin tulomomenttikorrelaatio

Riippumattomien otosten t-testi

Yksisuuntainen varianssi analyysi

Toistomittaus

Ei-parametrinen

vs. Spearmanin järjestyskorrelaatio

vs. Mann-Whitney U-testi

vs. Kruskal-Wallis -testi

vs. Wilcoxonin testi



Korrelaatio

- Kertoo kahden muuttujan välisestä lineaarisesta yhteydestä eli kuinka hyvin toisen muuttujan avulla voidaan ennustaa toisen muuttujan vaihtelua

- Ei paljasta käyräviivaista yhteyttä (vrt. ristiintaulukointi)

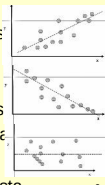
- Ei selitä syy-seuraus-suhdetta.

Päätelmä: Korrelaatiokerroin on lineaarisen riippuvuuden voimakkuutta kuvaava tunnusluku (voi saada arvoja, jotka ovat -1:n ja 1:n välillä)

- Jos kerroin on lähellä arvoa +1, muuttujien välillä on voimakas positiivinen korrelaatio → toisen muuttujan arvojen kasvaessa myös toisen muuttujan arvot kasvavat

- Jos kerroin on lähellä arvoa -1, muuttujien välillä on voimakas negatiivinen korrelaatio → toisen muuttujan arvojen kasvaessa toisen muuttujan arvot pienenevät

- Jos kerroin on lähellä arvoa 0, ei muuttujien välillä ole lineaarista



9



Korrelaatio

- Selitysaste saadaan korottamalla korrelaatiokerroin (.57) potenssiin 2.

- R2 eli korrelaatiokertoimen neliö (jos $r^2 = 0.32$), niin selittävää muuttujaa selittää 32% selitettävän muuttujan varianssista

Spearman

- mittaa kahden muuttujan järjestyslukujen yhteyttä
- rank correlation, r_s

Pearson

- mittaa kahden satunnaismuuttujan x ja y välisen yhteyden voimakkuutta
- satunnaismuuttujien oltava vähintään intervaali-asteikolliset ja normaali-jakautuneet

10



Tilastollisesta päätöksenteon perusteita

Karma-Komulainen s. 65 →

Yleistäminen (kuinka hyvin otos "sample" vastaa perusjoukkoa "population")

Luottamusväli

- Perusjoukon keskiarvoa (tai vastaavaa tunnuslukua) ei voida otoksen perusteella määrittää tarkasti. Otoksen perusteella voidaan vain sanoa, millä todennäköisyydellä arvo sijoittuu tietylle välille. Yleensä lasketaan vähintään 95% todennäköisyys (vrt. 99 ja 99,9%).

- 10 oppilaan perusteella saatu keskiarvo on toki sattumalle alttiimpi (eli suurempi luottamusväli) kuin 100:n tai 1000:n oppilaan vastaava....

- Eli luottamusväli kertoo epävarmuuden määrästä, mitä suuremmat luottamusvälit ovat sitä varmempaa päätely otoksen perusteella on ja päinvastoin....

11



Normaalijakaumaa hyödyntävä päätely

Luottamusväli

Esim. 1. summamuuttujat muistuttavat normaalijakautumaa (muista muuttujan testaus), ja esim. 2. useista kyllä – ei –vaihtoehdoista hyödyntävistä kyselyistä voidaan laskea keskiarvo, joka alkaa muistuttaa normaalijakaumaa... (→ yhteyte ryhmien väliseen vertailuun)

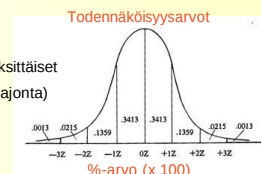
Teoreettisiin todennäköisyyksien arviointiin käytetään Gaussin käyrää, jonka yhteensopivuutta/vastaavuutta verrataan käytännön tutkimustilanteessa saatua otantajakaumaan (joka suhteutetaan aineiston kokoon / hajontaan)

Kaksi sovellustilannetta:

- Otoksen jakauma: Miten otoksen yksittäiset pistemäärät jakautuvat (keskiarvo ja hajonta)

- Otantajakauma: Otoksista tuotettu

keskilukujen jakauma (%)

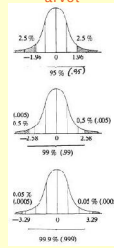


Luottamusvälin laskeminen

Keskivirheen luottamusväli. Tarvitaan tieto:

$$S = \frac{s}{\sqrt{n}}$$

- a) Keskivirhe (standard error) = muuttujan keskihajonta / $\sqrt{n}$ kriittiset z-arvot
- b) 95%:n luottamusväli = keskiarvo $\pm$ 1.96*keskivirhe
- Esim. kun testin $N=250$, $ka=50$, hajonta 8,
- a) $\rightarrow 8 / \sqrt{250}$ / neliöjuuri 250:sta = 0.51
- b) $\rightarrow 50 \pm (1.96 * 0.51 \rightarrow 1) = 50 \pm 1$ (eli 49 – 51)
- (Voitaisiin myös laskea 99%:n tai 99.9%:n luottamusväli!)
- vt. s. 80-81



Luottamusvälin laskeminen prosenteista:

- esim. natojäsenyyden kannatukselle kun 45% äänestänyt puolesta ja 55% vastaan ($N=1000$)
- $\rightarrow$ Keskivirhe = $\sqrt{pq/n}$ ($45 * 55 / 1000$) = 1.57
- $\rightarrow$ 95%:n luottamusväli kannatukselle: $p \pm 1.96 *$ keskivirhe
- eli $45 \pm (1.96 * 1.57 \rightarrow 3.01)$, väli on siis 41.9%-48.1%
- $\rightarrow$ tarkoittaa samaa kuin virhemarginaali (vt. vaaligallupit)

13

Kahden tunnusluvun erotuksen tilastollinen merkitsevyys (esim. T-testi)

Esimerkkinä tyttöjen (1) ja poikien (2) välisten erojen tarkastelu ja mahdollisuus poimia samankokoisia aineistoja lukematon määrä (esimerkkinä voisi olla myös ryhmä 1 ja ryhmä 2)

Ajatuksena on, että vaikka molemmat ryhmät on poimittu samasta populaatiosta, sattuman takia tyttöjen ja poikien välillä erot ovat ajoittain **suuria tai pieniä** ja välillä **poikien k:t olisivat suurempia** ja välillä **tyttöjen k:t olisivat suurempia**.

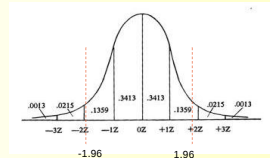
$\rightarrow$ Todennäköisin ero suuressa tapausjoukossa sukupuolen välillä olisi nolla eli eroa sukupuolen mukaan ei olisi.

Useiden mittausten jakauma $\rightarrow$ keskiarvo olisi nolla, ja vasen puolisko kuvaisi esimerkiksi niitä tapauksia, joissa tytöt parempia ja oikea puolisko niitä tapauksia, joissa pojat parempia

$\rightarrow$ hajonta keskiarvon molemmin puolin olisi keskiarvojen erotusten keskivirhe

14

- Saadaan jakauma, joka osoittaa miten samasta populaatiosta poimittujen keskiarvojen erotukset jakautuvat
- Tutkimuksessa saatu ero on yksi näistä moninaisista eromahdollisuuksista
- Jakauman keskellä sijaitsevat todennäköiset, helposti sattumaltakin ilmaantuvat erot
- Laitoja kohti mentäessä (erojen kasvaessa) sattuman osuus pienenee... $\rightarrow$ kun ylitetään 95% riskitaso ryhmät edustavat eri populaatiota $\rightarrow$ erotus on tilastollisesti merkitsevä



15

Tilastollinen testaus

Hypoteesien valinta perustuu edellä esim. edellä esitettyjen erojen testaamiseen

Nollahypoteesi H_0

- Teoriasta johdetun oletuksen vastainen hypoteesi
- Esim. Poikien ja tyttöjen keskiarvon välillä ei ole tilastollisesti merkitsevää eroa

Vastahypoteesi H_1

- Poikien keskiarvo on tyttöjen keskiarvoa pienempi (yksisuuntainen testaus)
- Poikien ja tyttöjen keskiarvot ovat erisuuret (kaksisuuntainen testaus)

16

Tilastollisen päätöksen tekeminen

- Jos p-arvo alle valitun merkitsevyystason nollahypoteesi hylätään
- Tällöin puhutaan tilastollisesti "merkitsevistä" tuloksesta (ei "merkittävistä")
 - Merkittään taulukoihin usein tähdillä:
 - * ($p < 0,05$), ** ($p < 0,01$), *** ($p < 0,001$)

HUOM:

Päätöstä tehtäessä on mahdollisuus kahteen virheeseen:

- Jos nollahypoteesi hylätään, vaikka se on tosi $\rightarrow$ hylkäämisvirhe (eng. type I error)
- Jos nollahypoteesi hyväksytään, vaikka se on epätosi $\rightarrow$ hyväksymisvirhe (eng. type II error)
- Hyväksymis- ja hylkäämisvirheiden todennäköisyydet toisistaan riippuvaisia $\rightarrow$ jos hylkäämisvirheelle asetetaan vaativa raja (esim. 0,1%), kasvaa hyväksymisvirheen todennäköisyys ja päinvastoin

17

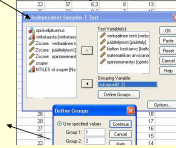
T-testi

- Kun vertaillaan kahta täysin toisistaan riippumatonta osajoukkoa toisiinsa ja halutaan tietää eroavatko ryhmien keskiarvot toisistaan, voidaan tilastollisena menetelmänä käyttää riippumattomien (independent) otosten t -testiä.

- Testin oletuksia ovat muuttujan normaalijakautuneisuus ja riippumattomat otokset.

	t	Sig.	df	Sig. (2-tailed)	Mean Difference	Std. Error
viikkotien leveys	.041	.041	28	.013	2,267	
viikkotien leveys	.041	.041	28	.013	2,267	

raportoidaan



- Testisuureen valintaan vaikuttaa ovatko varianssit (keskihajonnat) molemmissa osajoukoissa yhtäsuuret.

18