

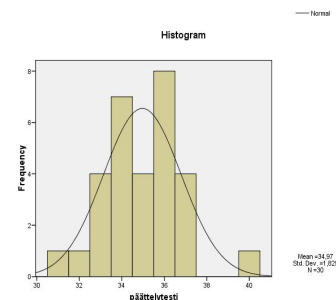
I Tilastollisen aineiston ja analyysin edellytysten tarkistaminen

- Muuttujien jakauman tarkistus
- Muunnokset, uudelleen koodaaminen, summamuuttujien luominen
- Puuttuva tieto ja sen käsittely
- Kuvaileva tilastoanalyysi vs. tilastollinen päättely

1

Normaalijakauma

- Muuttujan jakauman normaalisuus on monien tilastollisten testien oletus
- Normaalijakauma on symmetrinen, sen sijainti ja muoto riippuvat keskiarvosta ja hajonnasta
- Havaintojen jakautumista keskiarvon (mean) ympärille kuvataan keskihajonnalla (standard deviation)
- Normaalijakauman havainnoista 95 % sijoittuu lähemmäs kuin kahden keskihajonnan päähän keskiarvosta.



2

Vinous ja huipukkuus

- Normaalijakauman vinous (skewness) ja huipukkuus (kurtosis) ovat nollia.
- Jakauman vinouden ja huipukkuuden nolasta eroavuutta voi testata jakamalla saatu arvo sitä vastaavalla keskivirheellä (Standard Error). Jos näin saatu luku on $< |2|$, voidaan jakauman vinous ja huipukkuus hyväksyä vielä normaaliseksi ja siten tarkasteltava jakauma on riittävän normaalin tilastollisiin testeihin (vrt. Jos taas on $> |2|$ niin vinous/huipukkuuskerroin eroaa tilastollisesti merkitsevästi nolasta)
- Jakauman vinous ja huipukkuus pulmallisia perinteisissä tilastoanalyysissä.

3

Vinous (vrt. normaalijakauman vinous = 0)

à Skewness = vinouskerroin ilmoittaa, mihin suuntaan jakauma on vino; Jos $Sk > 0$, positiivisesti vino/oikealle vino

=> Jos $Sk < 0$, negatiivisesti vino/vasemmalle vino

• Huipukkuus (vrt. normaalijakauman huipukkuus = 0)

=> Kurtosis = huipukkuuskerroin ilmoittaa, kuinka terävähuippuinen jakauma on

=> Jos $Kur > 0$, jakauma on terävähuippuinen

=> Jos $Kur < 0$, jakauma on huiputon eli litteähuippuinen

oikealle vino

vasemmalle vino

huipukas

huiputon



4

Aineiston tarkistus ja poikkeavat arvot (outlier:it)

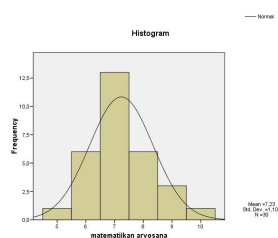
Yksittäiset äärimmäisen suuret tai pienet arvot voivat tuottaa pulmia analyyseissa

- Kannattaa tarkastella myös graafisesti (esim. boxplot)
- Esim. Kuntien asukkaiden keskitulovertailuissa Kauniainen vääristää keskiarvoa, optiomiljönääri vanhempien palkkojen vertailua jne....

5

Muuttujan normaalijakauman testaus

- Kolmogorov-Smirnov -testi
- SPSS: Analyze /Descriptive statistics/Explore – normality plots with tests
à Konservatiivinen testi eli hylkää helposti normaalijakaumaoletuksen... (vrt. Shapiro-Wilk)



6

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
päätelytesti	,147	30	,095	,952	30	,191
verbaalinen testi	,149	30	,088	,962	30	,355
kielten keskiarvo	,075	30	,200*	,974	30	,665
matematiikan arvosana	,250	30	,000	,916	30	,021
opintomenestys	,140	30	,135	,936	30	,069

Aineiston tarkistaminen & poikkeavat havainnot

Tarkistus #1: Descriptives –komento

SPSS/Analyze/Descriptive Statistics/Descriptives

-valinnat mean, std.deviation, minimum ja maximum

- Joskus äärimmäiset arvot johtuvat laite-, mittaus-, koodaus tms. virheistä → Arvojen korjaus oikeaksi, jos se on mahdollista (esim. tarkistus alkuperäisistä lomakkeista)
- Poistetaan datasta, raportoidaan
- Havaintojen käyttäminen sellaisenaan, jos niille on järkevä selitys, mutta huomioidaan niiden vaikutus tuloksiin
- Korvataan keskiarvolla
- kannattaa pohtia löytyykö juuri kyseistä puuttuvaa arvoa kuvaava keskiarvo SPSS/Analyze/Compare Means/Means

7

Mikä on tarpeeksi normaalia

Testien rinnalla jakaumaa kannattaa tarkastella myös graafisesti. Testien oletukseksi riittää, kunhan jakauma on noin 'suurin piirtein' normaalin.

TOISAALTA: Vaikka "riittävä" normalisuus täyttyisikin tilastollisista oletuksista on hyvä olla tietoinen ja pohdiskella niiden mahdollista vaikutusta tuloksiin...

- ONKO REALISTA OLETTAA, ETTÄ KAIKKI PSYKOLOGISET ILMIÖT KUTEN ESIM. MASENNUS, ONGELMAKÄYTTÄYTYMINEN OLISIVAT KUTTA KUINKIN NORMAALISTI JAKAUTUNEITA?

8

Uudelleen koodaaminen (recode)

SPSS: Transform/recode

- Voidaan tehdä joko alkuperäiseen tai uuteen muuttujaan (mieluummin uuteen)
- Tarvitaan esim. muuttujan "suunnan" vaihtoon
- Tarvitaan myös jatkuvan muuttujan luokittelussa jne...

Esim. ikämuuttuja luokitellaan kolmeen ryhmään

- ikä <35: uusi arvo 1

- ikä 35- 59: uusi arvo 2

- ikä >60: uusi arvo 3

Esim. uuden 'ikämuuttujan' luominen:

IKÄ = VUOSI – SYNTYMÄVUOSI

9

II Tilastollinen päättely

- Perusteet tilastolliselle päättelylle
- Hypoteesien testaus
- Tilastollisen päättelyn ongelmia

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 päättely otoksen perusteella on ja päinvastoin....

11

Luottamusväli

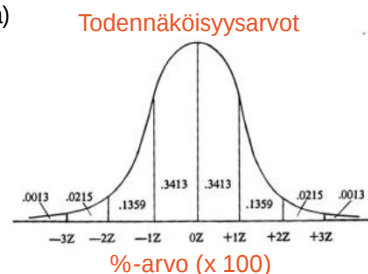
Normaalijakaumaa hyödyntävä päättely

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

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

Kaksi sovellustilannetta:

- Otoksen jakauma: Miten otoksen yksittäiset pistemäärät jakautuvat (keskiarvo ja hajonta)
- Otantajakauma: Otoksista tuotettu keskilukujen jakauma (%)



12

Luottamusvälin laskeminen

Keskiarvon luottamusväli. Tarvitaan tieto:

a) Keskivirhe (standard error) = muuttujan keskihajonta / $\sqrt{n}$

b) 95%:n luottamusväli = keskiarvo $\pm$ 1.96 * keskivirhe

Esim. kun testin $N=250$, $ka=50$, hajonta 8,

a) $\hat{a}$ 8 / neliöjuuri 250:sta = 0.51

b) $\hat{a}$ 50 $\pm$ (1.96 * 0.51 $\hat{a}$ 1) = 50 $\pm$ 1 (eli 49 – 51)

(Voitaisiin myös laskea 99%:n tai 99.9%:n luottamusvälit!)

vrt. s. 80-81

Luottamusvälin laskeminen prosenteista:

esim. natojäsenyyden kannatukselle kun 45% äänestänyt puolesta ja 55% vastaan ($N=1000$)

$\hat{a}$ Keskivirhe = $\sqrt{0.45 * 0.55 / 1000}$ = 1.57

$\hat{a}$ 95%:n luottamusväli kannatukselle: $p \pm$ 1.96 * keskivirhe

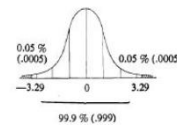
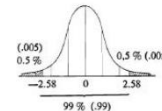
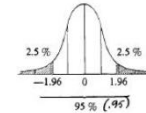
eli 45 $\pm$ (1.96 * 1.57 $\hat{a}$ 3,01), väli on siis 41.9%-48.1%

$\hat{a}$ tarkoittaa samaa kuin **virhemarginaali (vrt. vaaligallupit)**

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

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

kriittiset z-arvot



13

Kandien 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)

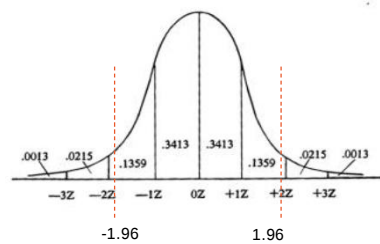
Ajatuksena on, että vaikka molemmat ryhmät on poimittu samasta populaatiosta, sattuman takia tyttöjen ja poikien välillä erot ovat ajoittain **suuria tai pieniä** ja välillä **poikien ka:t olisivat suurempia** ja välillä **tyttöjen ka:t olisivat suurempia**.

$\hat{a}$ Todennäköisin ero suuressa tapausjoukossa sukupuolen välillä olisi nolla eli eroa sukupuolen mukaan ei olisi.

Useiden mittausten jakauma $\hat{a}$ keskiarvo olisi nolla, ja vasen puolisko kuvaisi esimerkiksi niitä tapauksia, joissa tytöt parempia ja oikea puolisko niitä tapauksia joissa pojat parempia $\hat{a}$ hajonta keskiarvon molemmin puolin olisi keskiarvojen erotusten keskivirhe

14

- Saadaan jakauma, joka osoittaa miten samasta populaatiosta poimittujen keskiarvoparien erotukset jakautuvat
 - Tutkimuksessa saatu ero on yksi näistä moninaisista eromahdollisuuksista
 - Jakauman keskellä sijaitsevat todennäköiset, helposti sattumaltakin ilmaantuvat erot
 - Laitoja kohti mentäessä (erojen) ylitetään 95% riskitaso ryhmät eivät ole enää tilastollisesti merkitseviä



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
”merkitsevää” tuloksesta (ei
”merkittävää”)
- Merkitään 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 à
hylkäämisvirhe (eng. type I error)
- Jos nollahypoteesi hyväksytään, vaikka se on epätosi à
hyväksymisvirhe (eng. type II error)
- Hyväksymis- ja hylkäämisvirheiden todennäköisyydet toisistaan
riippuvaisia à jos hylkäämisvirheelle asetetaan vaativa raja (esim.
0,01%), kasvaa hyväksymisvirheen todennäköisyys ja päinvastoin

Standardointi tekee eri asteikollisista muuttujista vertailukelpoisia

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

à arvot verrattavissa minkä muun normeeratun muuttujan
kanssa

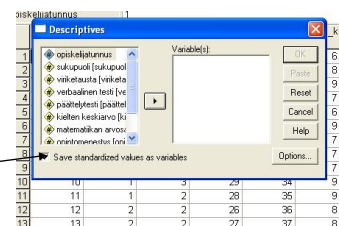
à puhutaan Z-pisteistä

à voidaan muodostaa esim.
summamuuttujia

à miten z-pisteet tehdään

Analyze à Descriptive statistics

à Descriptives



Mitta-asteikon vaikutus analyysimenetelmän ja merkitsevyystestin valintaan

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

19

Tilastollisen menetelmän valinta

		Selitettävä -, riippuva -, y-muuttuja	
		<i>Luokittelu- tai järjestysasteikko</i>	<i>Välimatka tai suhdeasteikko</i>
Selittävä -, riippumaton -, X -muuttuja	<i>Luokittelu- tai järjestysasteikko</i>	Ristiintaulukointi, Loglineaariset mallit	T-testi, [*] Manova, Anova
	<i>Välimatka tai suhdeasteikko</i>	Logistinen ja multinomiaalinen regressioanalyysi	Regressioanalyysi, polku- ja rakenneyhtälömallit

20

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

Ristiintaulukointi eli kontingenssitaulukointi

Käyttö:

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

21

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
koulutustoive

KOE = 2 tai
useampi virhe

KONTROLLI = 0-1
virhettä

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

22

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 lukumääristä, $df=(r-1)(s-1)$
- Testin havaittu merkitsevyydestä 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 $< 0,05$

23

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 kokonaisia nollarivejä tai -sarakkeita

24

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 normaaliuus

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**Ei-parametrinen**

Pearsonin tulomomenttikorrelaatio

vs. Spearmanin järjestyskorrelaatio

Riippumattomien otosten *t*-testi

vs. Mann-Whitney U-testi

Yksisuuntainen varianssi analyysi

vs. Kruskal-Wallis –testi

Toistomittaus

vs. Wilcoxonin testi

SUMMAMUUTTUIJEN LUOMINEN

- SPSS: Transform / Compute:

$$\text{SUMMA} = (\text{KYS1} + \text{KYS2} + \text{KYS3} + \text{KYS4} + \text{KYS5})/5$$

$$\text{SUMMA} = \text{MEAN}(\text{KYS1}, \text{KYS2}, \text{KYS3}, \text{KYS4}, \text{KYS5})$$

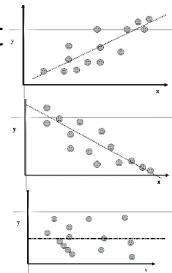
- Suoran summamuuttujan muodostaminen edellyttää skaalan riittävää reliabiliteettia (esim. Cronbachin alfa > .70)
- Muista: Kysymysten (item) kääntäminen, niin että kaikki summamuuttujan osakysymykset ovat samansuuntaisia siten, että mitä suurempi arvo sitä enemmän se mittaa mitattavaa muuttujaa

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 riippuvuutta



27

Korrelaatio

- Selitysaste saadaan korottamalla korrelaatiokerroin (.57) potenssiin 2.
- R^2 eli korrelaatiokertoimen neliö (jos $r^2 = 0.32$), niin selittävämuuttuja 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 intervalli-asteikolliset ja normaali-jakautuneet

28

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.

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
verbaalinen testi	Equal variances assumed	,041	,841	2,856	28	,013	2,267
	Equal variances not assumed			2,856	27,981	,013	2,267

