

Intro tilastotieteeseen

Miko Pasanen

`mikos.pasanen@gmail.com`

19.3.2025

Sisältö

- 1 Ennen aineiston keruuta
- 2 Aineiston valmistelu analyysia varten
- 3 Analyysit
- 4 Tulosten tulkinnasta ja estimoinnista
- 5 Esimerkkejä testauksista

Ennen aineiston keruuta

Analyysia ajatellen on hyvä huomioida muutamia asioita jo ennen aineiston keruuta.

- Hypoteesien ja/tai tutkimuskysymyksien asettaminen
→ millaista aineistoa tarvitaan ja millaiset analyysit vastaavat tutkimuskysymyksiin
- Edustavan otoksen kerääminen, jotta tulokset voidaan yleistää kohdepopulaatioon
 - otoksen koko
 - satunnaistaminen
 - koeasetelma
- Kyselylomakkeen toteutus
 - kyselylomakkeen tarkastelu siitä näkökulmasta millaisena aineisto tallentuu, jotta se sopivaa analyysia varten ja välttyään turhalta työltä
→ esim. jos ikää kysytään avoimena kysymyksenä eikä numeerisena, kuinka monella eri formaatilla iän voi kirjoittaa

Aineiston valmistelu

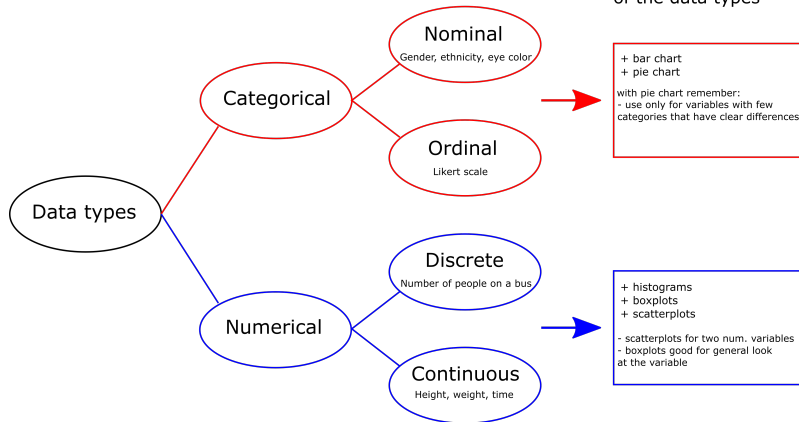
- Aineiston esikäsittely, sisältää esimerkiksi
 - Summamuuttujien muodostamista
 - Numeeristen muuttujien muuttamista kategorisiksi (esim. ikä, työkokemus)
 - Kategoristen muuttujien kategorioiden yhdistely (esim. jos joku ryhmä jää pieneksi)
- Aineiston tarkastelu visualisointien ja kuvailevien tilastollisten suureiden avulla
- Puuttuvien havaintojen tarkastelu ja mahdollinen "paikkaaminen"
 - Onko aineistossa paljon puuttuvia havaintoja?
 - Onko puuttuvissa jonkinlaista systemaattisuutta?
 - Tarvittaessa voidaan tehdä imputointia eli puuttuvien havaintojen paikkaamista jollakin menetelmällä (esim. keskiarvolla).

Muuttujien kuvailu tilastollisilla suureilla

- **Nominaaliset muuttujat** (esim. sukupuoli, ammatti)
 - Frekvenssit, prosenttiosuudet
 - Moodi (=yleisin arvo)
- **Järjestysasteikolliset muuttujat** (esim. Likert-asteikko)
 - Frekvenssit, prosenttiosuudet
 - Moodi
 - Mediaani ja IQR
 - Keskiarvoa ja keskihajontaa on myös mahdollista käyttää, mutta näiden käyttöä voi kritisoida, koska asteikon välien etäisyydet eivät ole niin selvät. Esim. jos arvioimme henkilön ahdistusta yhdellä likert muuttujalla (1-5), onko ero 1 ja 2 välillä sama kuin 2 ja 3? Jos henkilö A vastaa 1 ja henkilö B 5, voimmeko sanoa, että henkilö B on 5 kertaa ahdistuneempi kuin henkilö A?
- **Jatkuvat muuttujat** (esim. ikä, työkokemus)
 - Keskiarvo ja keskihajonta
 - Mediaani ja IQR (erityisesti, jos muuttuja on hyvin vino tai huipukas)
- **Diskreetit muuttujat** (esim. lasten määrä)
 - Keskiarvo ja keskihajonta
 - Mediaani ja IQR (erityisesti, jos muuttuja on hyvin vino tai huipukas)
 - Frekvenssit, prosenttiosuudet
 - Moodi

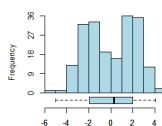
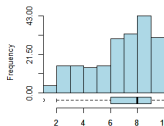
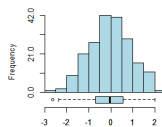
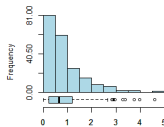
Muuttujien tyypit ja niiden visualisointi

Graphical representations of the data types

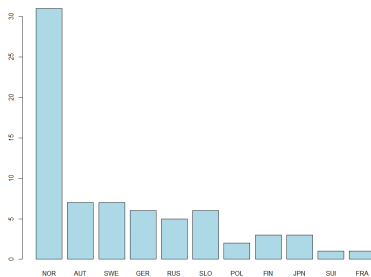


Histograms

- Histograms can be used as an approximation of the distribution of continuous variable.
- The range of the variable is divided to intervals and the frequency of each of those intervals is calculated. Height of the bar represents the frequency.
- Is the distribution symmetric or skewed (left/right)? Is it unimodal/bimodal/multimodal?

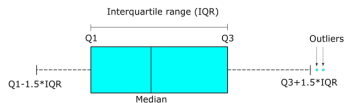


Bar charts



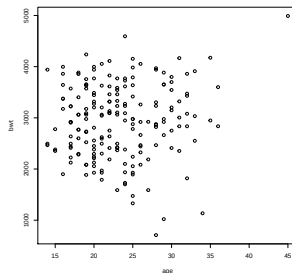
- For categorical/ordinal variables bar chart can be used.
- Bar chart can also be grouped or stacked. E.g. we have likert scale 1-5. For each of these levels we can separate the counts by gender.

Boxplot



- Boxplot is visual method for describing the spread and center of continuous variable.
- "The box" part contains first (Q1), second (median) and third (Q3) quartile of the variable.
- "Whiskers" represent the variability outside first and third quartiles.
- Outliers are the points that lie outside of the whiskers.

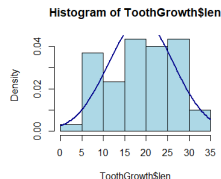
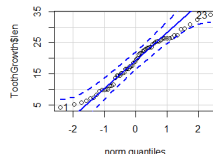
Scatterplot



- Used to visually inspect the relationship between two numerical variables.

Normal Q-Q plot

- The Q-Q plot can be used to visually assess if it is plausible that the data came from some theoretical distribution, e.g. Normal distribution.
- Normal Q-Q plot can be used to assess the normality of sample.
- Points should lie approximately on a 45° straight line.



Data matrix (or data frame)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa

- In your study it is important to think the data in terms of cases and variables.
- In data matrix data is formatted so that
 - each row represents a case and
 - each column represents a variable.

Analyysit

Aineiston ja hypoteesien/tutkimuskysymysten avulla valikoidaan sopivat menetelmät analyysia varten.

- Mikä on hypoteesisi?
- Mikä muuttuja on selitettävä ja mikä/mitkä ovat selittävä/t muuttuja/t?
- Minkä tyyppisiä muuttujat ovat?
- Onko selittävästä muuttujasta toistomittauksia tai toisistaan riippuvaisia populaatioita (esim. lapsi-vanhempi parit)
- Oletusten tarkastelu (normaalisuusoletus, varianssien yhtäsuuruus,...) ja mahdollisesti epäparametristen menetelmien käyttö.

Hypoteesin testaus

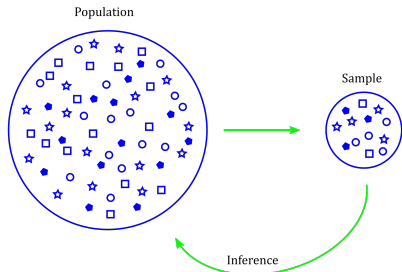
- 1 Tutkimushypoteesi on muotoiltu.
- 2 Määritellään nollahypoteesi (H_0) ja vaihtoehtoinen hypoteesi (H_1).
- 3 Valitaan merkitsevyystaso α . (Tyypillisesti: 0.05 ja 0.01.)
- 4 Kerätään aineistoa.
- 5 Valitaan sopiva testisuure ja lasketaan se otoksesta.
- 6 Joko:
 - (i) Lasketaan kriittinen alue, jolla nollahypoteesi voidaan hylätä. Jos havaittu testisuure osuu tälle alueelle, nollahypoteesi voidaan hylätä.
 - (ii) Lasketaan p-arvo havaitulle testisuurelle ja verrataan sitä asetettuun merkitsevyystasoon. Jos $p\text{-arvo} < \alpha$, nollahypoteesi voidaan hylätä.

Nollahypoteesi ja vaihtoehtoinen hypoteesi

- Testaus alkaa olettamuksesta, että nollahypoteesi on tosi. Tavoitteena on saada todisteita nollahypoteesia vastaan.
- Nollahypoteesi esittää "hyväksytyn faktan", esim.
 - ① Ryhmien A ja B välillä ei ole eroa.
 - ② intervention vaikutus vastemuuttujaan on nolla.
- Vaihtoehtoinen hypoteesi on nollahypoteesin vastakohta. Kun nollahypoteesi hylätään, vaihtoehtoinen hypoteesi astuu voimaan.
- Jos nollahypoteesia ei voida hylätä, tämä ei välttämättä tarkoita, että se olisi totta. Esimerkiksi testin voima voi olla liian alhainen nollahypoteesin hylkäämiseksi.

	H_0 tosi	H_0 epätosi
H_0 ei hylätä	OK	Tyypin II virhe (β)
H_0 hylätään	Tyypin I virhe (α)	OK

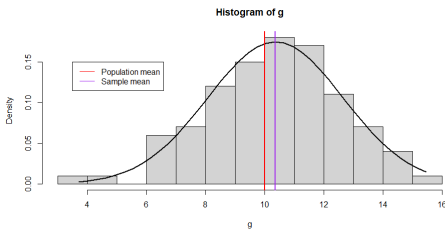
Populaatio ja otos



- **Populaatio** on se joukko, josta tutkimuksessa ollaan kiinnostuneita.
- Hyvin harvoin on mahdollisuus havainnoida kokonaan tämä kohdepopulaatio, joten tästä populaatiosta valitaan osa mukaan tutkimukseen. Tätä kutsutaan **otokseksi**.
- Tämän otoksen tulisi olla **edustava** kohdepopulaation kannalta, jotta otokselle tehtävä tilastollinen päättely voitaisiin yleistää koko populaatioon.

Populaatioparametrien estimointi

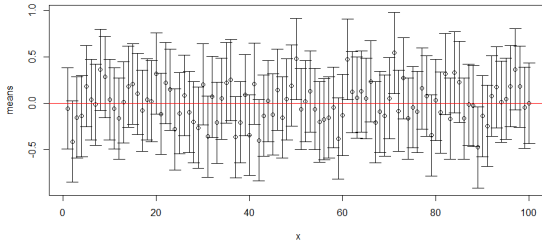
- Koska populaation jakauman parametrit ovat tuntemattomia, otoksesta lasketaan otossuureita estimoimaan todellisia parametreja.
- Otosvirhe** on ero otossuureiden ja populaation parametrien välillä.



- Esimerkiksi normaalijakaumalla on parametrit keskiarvo ja keskihajonta.
- Jos otoksen jakauma on suurinpiirtein normaalijakautunut, voimme olettaa, että myös populaation jakauma voisi olla normaalijakautunut.
- Tällöin otoksen keskiarvoa ja keskihajontaa voidaan käyttää estimoimaan populaation parametreja.

Luottamusvälit

- Koska jokainen mahdollinen populaatiosta otettu otos tuottaa hieman erilaisia otossuureita, estimaateissa on aina otosvirhettä.
- Luottamusvälejä voidaan käyttää estimoimaan estimaattien tarkkuutta.
- Parametrien estimaatit kuten keskiarvo ovat **piste-estimaatteja**.
Luottamusväleillä voidaan arvioida sitä aluetta, joka kattaa populaation parametrin uskottavat arvot otoksen perusteella.
- Tärkeää muistaa: esim. 95% luottamusväli ei tarkoita sitä, että olisi 95% todennäköisyys, että parametrin todellinen arvo olisi tällä välillä! Parempi tapa on ilmaista, että olemme 95% "varmoja" (eng. confident), että todellinen arvo on luottamusvälillä.



Esimerkki 1

- Tutkimuskysymys: Onko rikollisuuden määrä Yhdysvalloissa keskimäärin suurempaa eteläisissä osavaltioissa?
- H_0 : Rikollisuuden määrän odotusarvo on sama, $\mu_s = \mu_{ns}$.
- H_1 : Rikollisuuden määrän odotusarvo on suurempi eteläisissä osavaltioissa, $\mu_s > \mu_{ns} = \mu_s - \mu_{ns} > 0$.
- Valitaan testin merkitsevyystasoksi $\alpha = 0.05$.
- Testiksi voidaan valita kahden otoksen t-testi, jos rikollisuuden määrän jakauman normaalisuusoletus on voimassa. Tämä voidaan testata Shapiro-Wilkin testillä.
- Toinen testattava oletus t-testissä on otosten varianssien yhtäsuuruus. Tämä voidaan testata Levenen testillä.

Esimerkki 1: Shapiro-Wilk

Tests of Normality

	Southern state	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Crime rate (number of offences per million population)	No	,090	31	,200*	,972	31	,570
	Yes	,182	16	,164	,960	16	,661

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

- Shapiro-Wilkin testin mukaan nollahypoteesi jää voimaan molemmille jakaumille, eli molemmat voidaan olettaa normaalijakautuneiksi ja jatkaa t-testiin.

Esimerkki 1: t-testi

Independent Samples Test									
Levene's Test for Equality of Variances				t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower Upper
Crime rate (number of offences per million population)	Equal variances assumed	4,139	.048	,359	45	,721	3,22520	8,97957	-14,86059 21,31099
	Equal variances not assumed			,401	40,428	,691	3,22520	8,05195	-13,04304 19,49344

- Levenen testi hylkää nollahypoteesin, jonka mukaan varianssit olisivat yhtä suuret, joten käytetään t-testiä otoksille, joiden varianssit oletetaan eri suuriksi.
- SPSS raportoi ainoastaan kaksisuuntaisen testin p-arvon, josta saadaan yksisuuntainen tässä tapauksessa $\frac{p\text{-arvo}}{2} = \frac{0.691}{2} = 0.3455 > 0.05$. Nollahypoteesi jää siis voimaan.
- Jos vastahypoteesi olisi $H_1 : \mu_s < \mu_{ns} = \mu_s - \mu_{ns} < 0$, niin p-arvo saataisiin laskemalla $1 - \frac{p\text{-arvo}}{2}$.

Esimerkki 2

- Tutkimuskysymys: Interventiotutkimus, jossa 50 tutkittavaa kaikki katsoivat videon tupakoinnista johtuneen syöpäkuoleman vaikutuksesta kuolleen omaisiin. Tutkittavista 25 olivat tupakoitsijoita ja 25 eivät tupakoineet. Kahden viikon jälkeen tutkittavilta kysyttiin tupakoivatko he. Tutkija tahtoo tarkastella oliko videolla vaikutusta tupakointiin.
- H_0 : Mittauskertojen välillä ei ole eroa eli todennäköisyys tupakointiin ensimmäisellä ja toisella mittauskerralla on sama.
- H_1 : Mittauskertojen välillä on eroa.
- Valitaan testin merkitsevyystasoksi $\alpha = 0.05$

Esimerkki 2: Hypoteesin testaus

	After	
	Before	
Non-Smoker	Non-Smoker	20
Smoker	Smoker	9

- Otokset ovat nyt riippuvia, koska samoilta henkilöiltä on kahdesta eri aikapisteestä mittaus.
- Vaste eli tupakointi on kaksiluokkainen muuttuja, joten voidaan käyttää McNemarin testiä.

Esimerkki 2: Hypoteesin testaus

	After	
	Before	
Non-Smoker	Non-Smoker	20
Smoker	Smoker	9

- Otokset ovat nyt riippuvia, koska samoilta henkilöiltä on kahdesta eri aikapisteestä mittaus.
- Vaste eli tupakointi on kaksiluokkainen muuttuja, joten voidaan käyttää McNemarin testiä.

Test Statistics^a

Before & After	
N	50
Exact Sig. (2-tailed)	,027 ^b

a. McNemar Test

b. Binomial distribution used.

Esimerkki 2: Hypoteesin testaus

	After	
Before	Non-Smoker	Smoker
Non-Smoker	20	5
Smoker	16	9

- Otokset ovat nyt riippuvia, koska samoilta henkilöiltä on kahdesta eri aikapisteestä mittaus.
- Vaste eli tupakointi on kaksiluokkainen muuttuja, joten voidaan käyttää McNemarin testiä.

Test Statistics^a

Before & After	
N	50
Exact Sig. (2-tailed)	,027 ^b

a. McNemar Test

b. Binomial distribution used.

- $p\text{-arvo} = 0.027 < 0.05$ eli H_0 voidaan hylätä ja sanoa, että mittausten välillä on eroa tupakoinnin todennäköisyyden suhteen.

Jos tarvitset apua esim.

- otoskoon laskennassa, satunnaistamisessa
- aineiston käsittelyssä
- tilastollisissa analyyseissa
- tilastollisten ohjelmistojen käytössä (SPSS, R, python)
- tilastotieteen ymmärtämiseen

tai johonkin muuhun tilastotieteeseen ja/tai data-analytiikkaan liittyen, laita viestiä **mikos.pasanen@gmail.com** ja katsotaan miten voin olla avuksi!