Uvod

Multivariatna analiza

Osnovne informacije

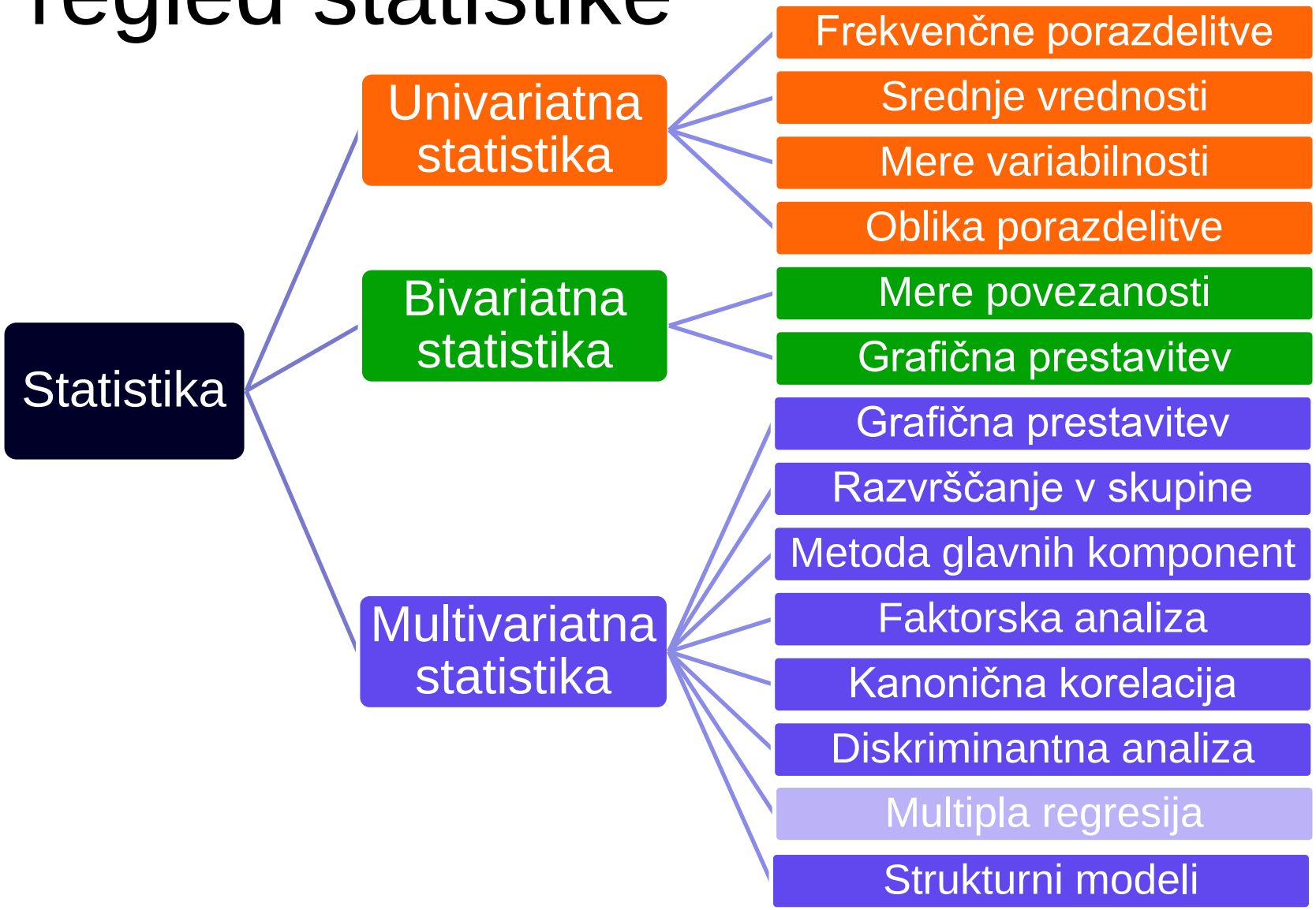
- Nosilec: izr. prof. dr. Aleš Žiberna
- Asistent: doc. dr. Marjan Cugmas

- Obveznosti:
  - Več domačih nalog + predstavitev (50%)
  - Izpit (50%)

Obravnavane teme

- Pregled in ponovitev statistike in umestitev multivariatne analize
- Grafična predstavitev multivariatnih podatkov
- Razvrščanje v skupine
- Metoda glavnih komponent
- Faktorska analiza
- Strukturni modeli
- Kanonična korelacijska analiza in diskriminantna analiza
- Pregled drugih metod multivariatne analize

Pregled statistike



Osnovni pojmi

- **Enota** – posamezni proučevani element.
  - redni študent na Univerzi v Ljubljani v študijskem letu 2010/11
- **Spremenljivka** – lastnost enot; označujemo jih npr. z X, Y, X_1. Vrednost spremenljivke X na i-ti enoti označimo z x_i.
  - spol, ocena na maturi, izobrazba matere, višina zadnjega mesečnega dohodka staršev,
- **Populacija** – množica vseh proučevanih elementov; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko).
  - vsi redni študenti na Univerzi v Ljubljani v študijskem letu 2010/11
- **Vzorec** – podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih cele populacije.
  - slučajni vzorec 300 študentov

Merske lestvice (vrste spremenljivk)

1. **nominalne spremenljivke** oz. **nominalna merska lestvica** (angl. *nominal scale*) – vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol);
2. **ordinalne spremenljivke** oz. **ordinalna merska lestvica** (angl. *ordinal scale*) – vrednosti lahko uredimo od najmanjše do največje (npr. uspeh, strinjanje z vrednostmi “sploh se ne strinjam”, “ne strinjam se”, “niti niti”, “strinjam se”, “zelo se strinjam”);
3. **intervalne spremenljivke** oz. **intervalna merska lestvica** (angl. *interval scale*) – lahko primerjamo razlike med vrednostima dvojic enot; lahko povemo, za koliko vrednosti spremenljivke se neka enota loči od druge (npr. temperatura v °C, koledarsko leto);
4. **razmernostne spremenljivke** oz. **razmernostna merska lestvica** (angl. *ratio scale*) – lahko primerjamo razmerja med vrednostima dvojic enot; lahko povemo, kolikokrat večja/manjša je vrednost neke enote od druge enote (npr. starost, čas).

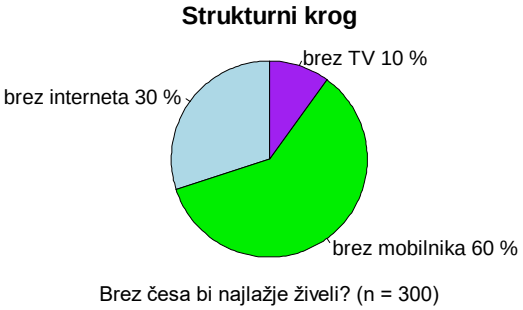
Primeri

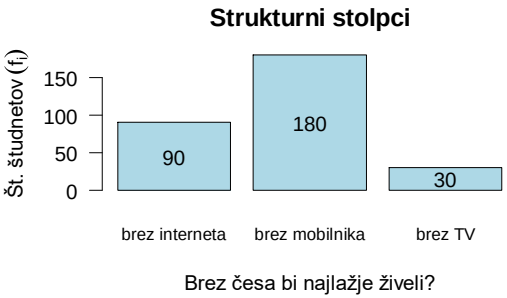
Določite osnovne pojme in merske lestvice spremenljivk za sledeče primere:

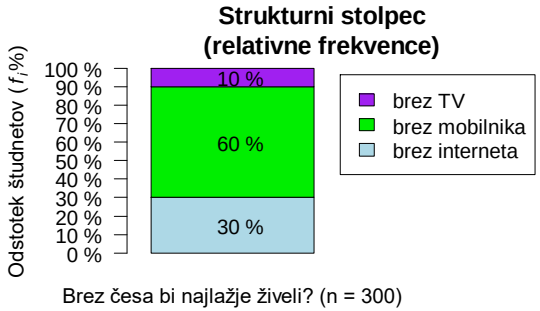
- Zanima nas povezanost med najvišjo dokončano stopnjo izobrazbe in plačo pri zaposlenih v Sloveniji na dan 31. 12. 2016. V ta namen smo pridobili podatke za slučajni vzorec 500 zaposlenih.
- Zanima nas, kaj vpliva na "uspešnost" neke objave na Facebooku izbranega podjetja. Zato smo za vse objave tega podjetja v času od 1. 1. 2016 do 31. 12. 2016 izbrali podatke o številu ogledov, številu "všečkov", številu besed, ali vključuje video ali sliko in času objave (del dneva in tedna).

Grafični prikazi

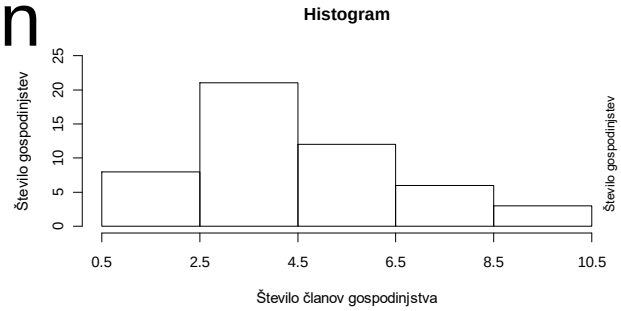
- Nominalne in ordinalne spremenljivke:
  - Strukturni stolpci
  - Strukturni krog

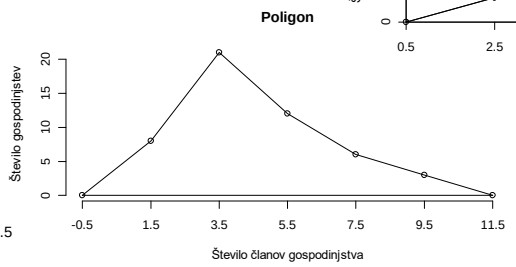


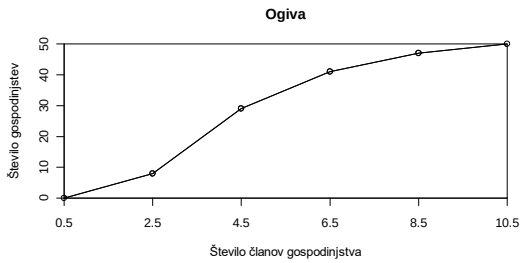




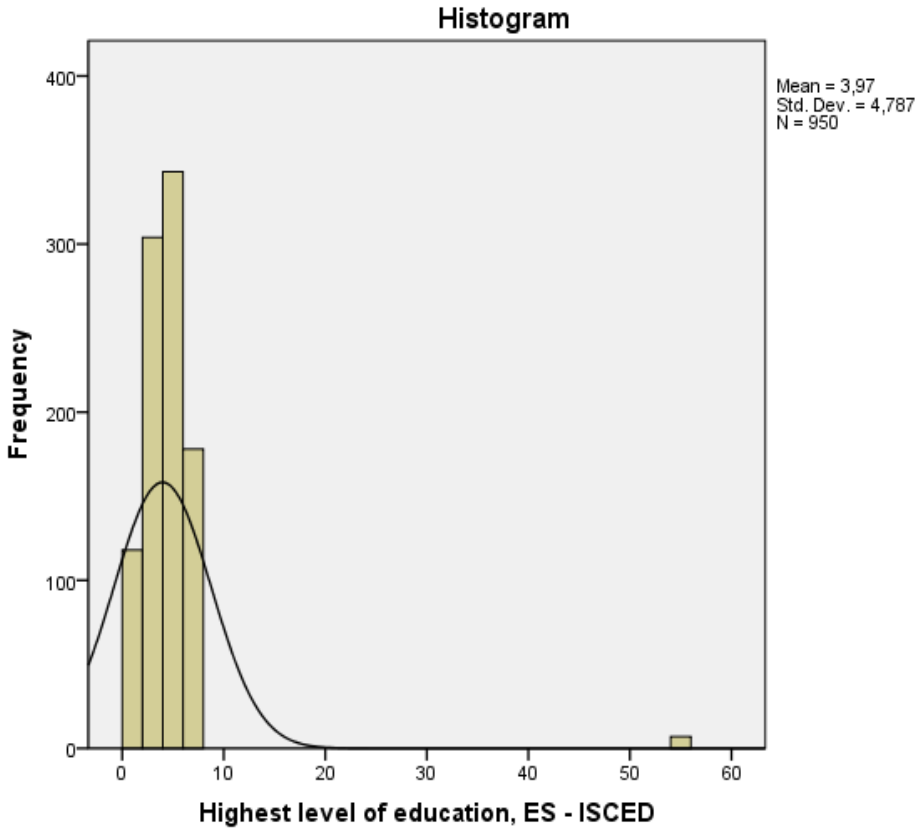
- Intervalne in razmernostne spremenljivke
  - Histogram
  - Poligon
  - Ogiva

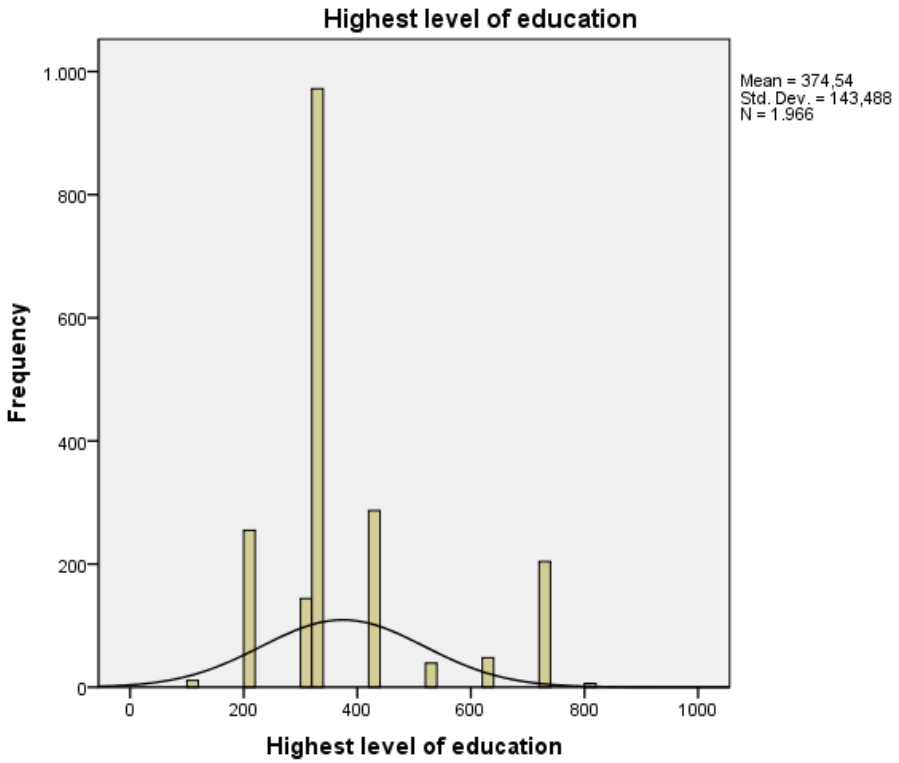




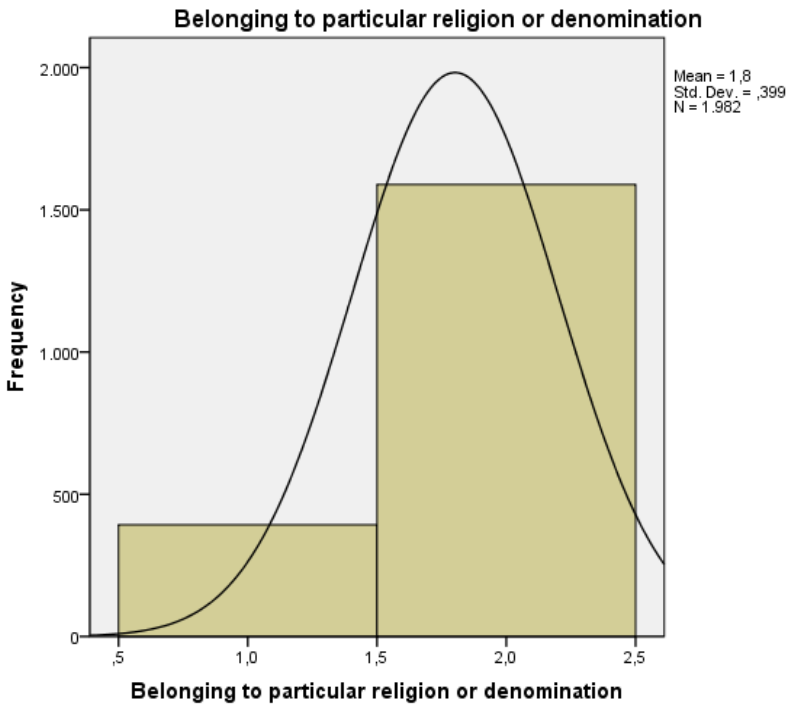


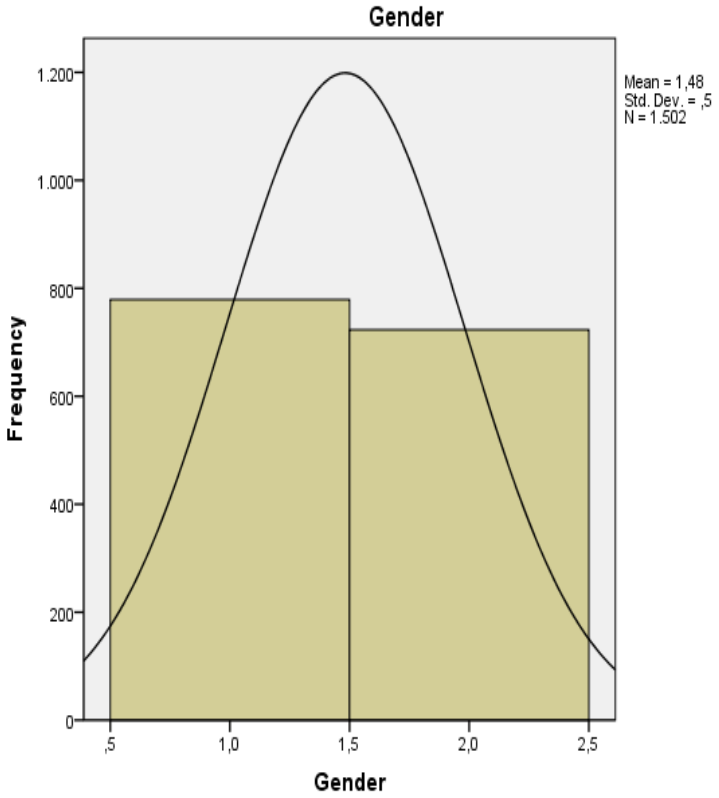
Nekaj primerov – kaj je narobe



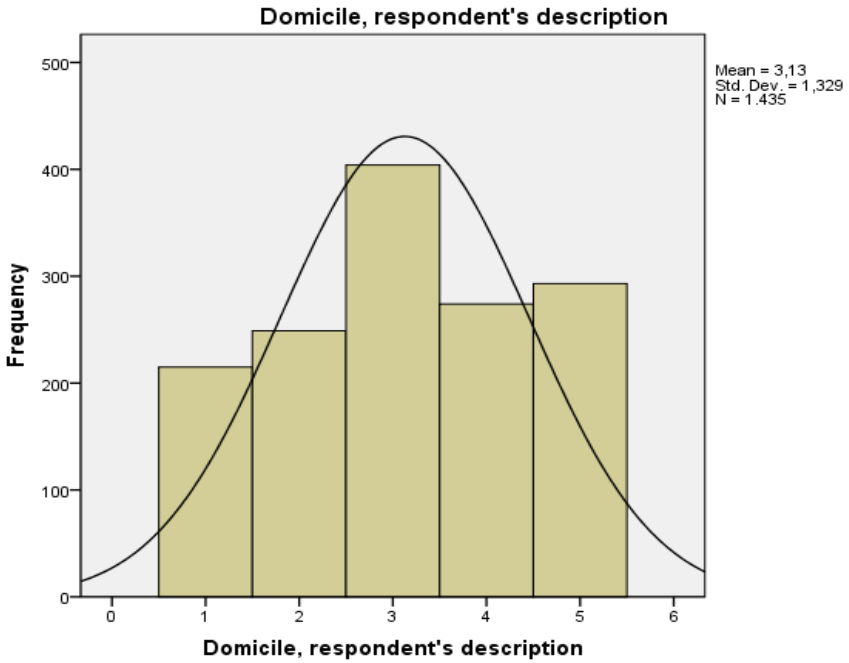


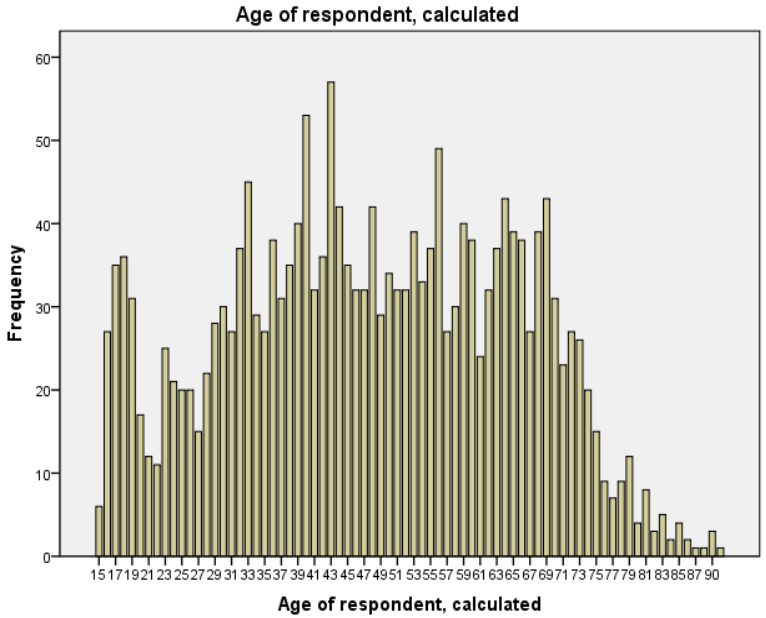
Nekaj primerov – kaj je narobe



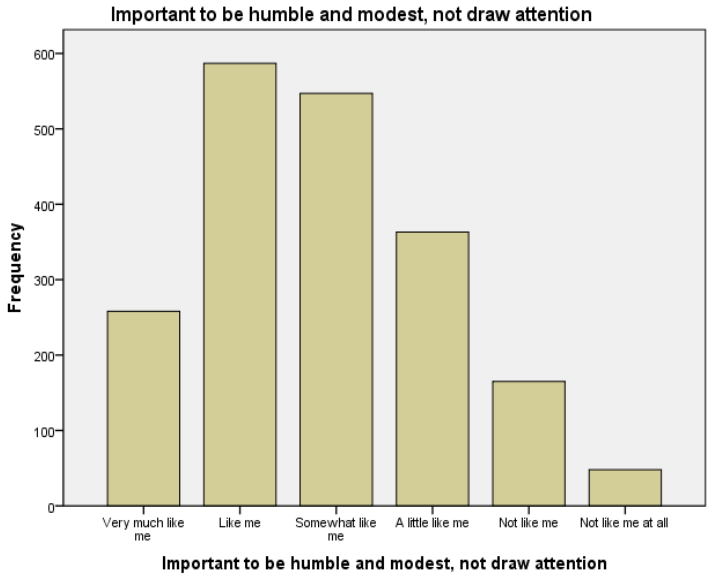


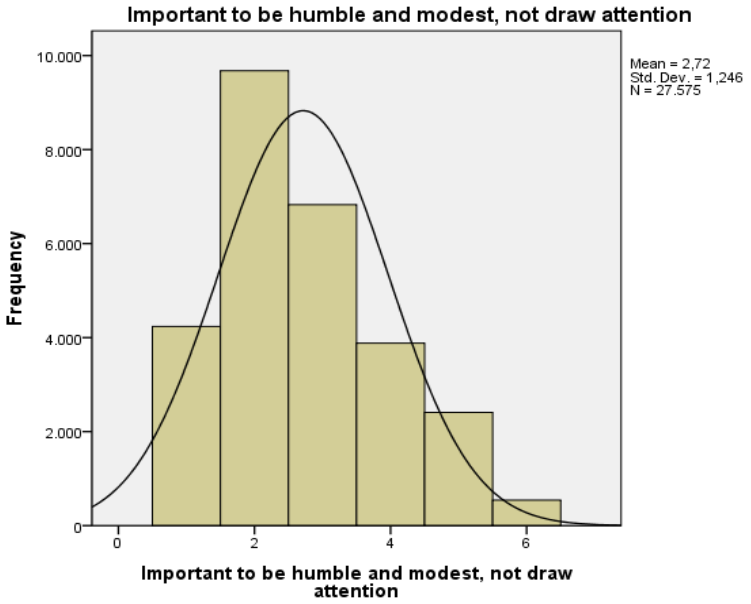
Nekaj primerov – kaj je narobe



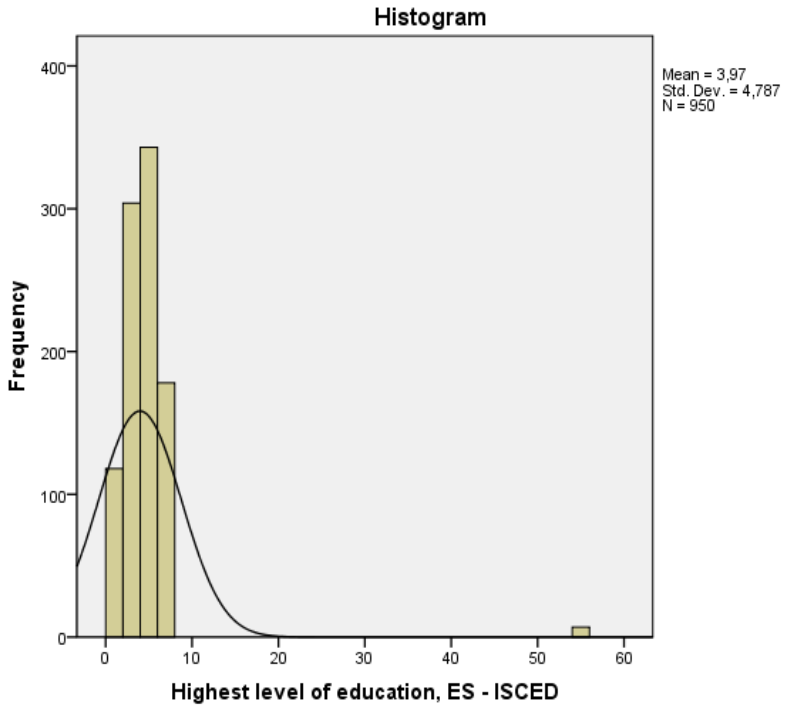


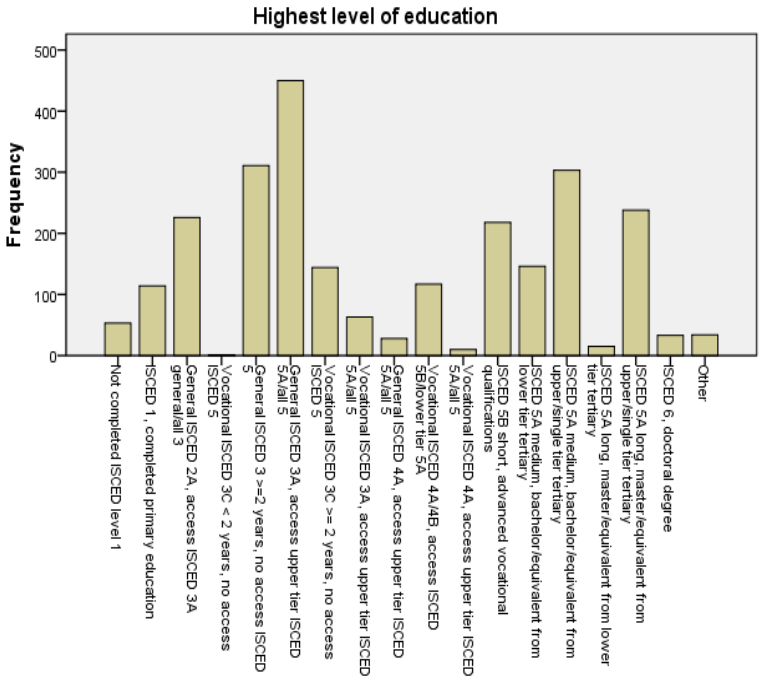
Nekaj primerov – kaj je "prav"



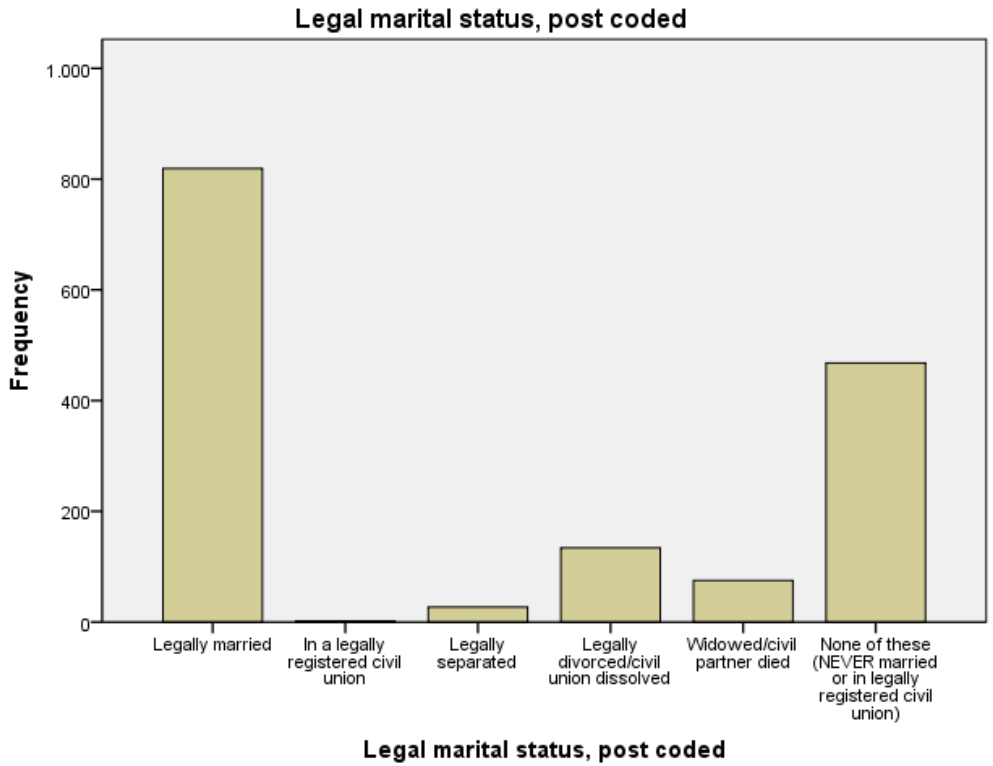


Nekaj primerov – kako naprej





Nekaj primerov – kako naprej



Nekaj primerov – kaj je narobe

Tabela 1: Statistics

	N					
	Valid	Missing	Mean	Std. Deviation	Skewness	Kurtosis
agea Age of respondent, calculated	1981	28	47,76	17,146	-,031	-,881
edulvlb Highest level of education	1966	43	374,54	143,488	1,429	1,246
rlgblg Belonging to particular religion or denomination	1982	27	1,80	,399	-1,515	,294
ipmodst Important to be humble and modest, not draw attention	1968	41	2,86	1,243	,415	-,414
imptrad Important to follow traditions and customs	1977	32	2,65	1,222	,655	-,011
ipstrgv Important that government is strong and ensures safety	1965	44	2,24	1,218	,882	,222
domicil Domicile, respondent's description	1940	69	2,73	1,124	-,520	-1,101

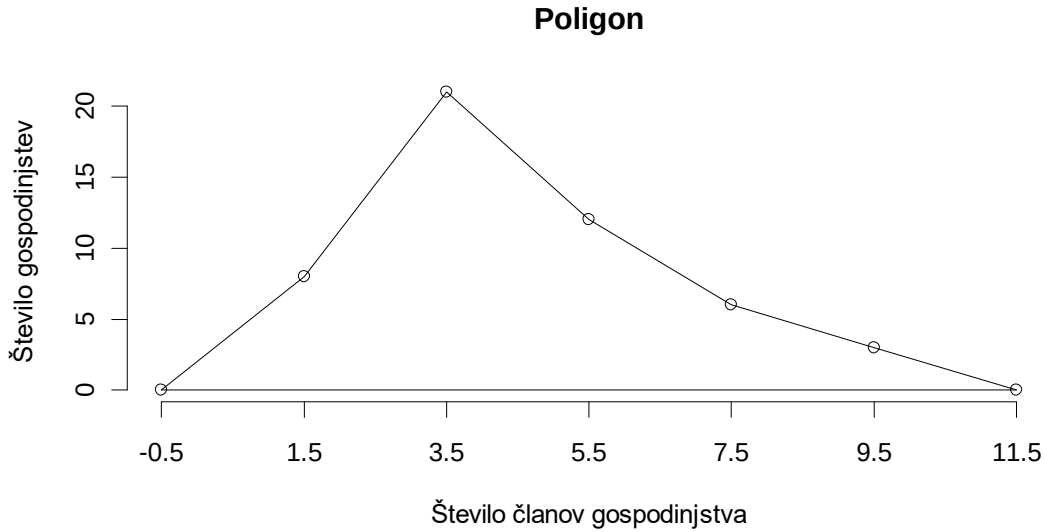
Standardizacija

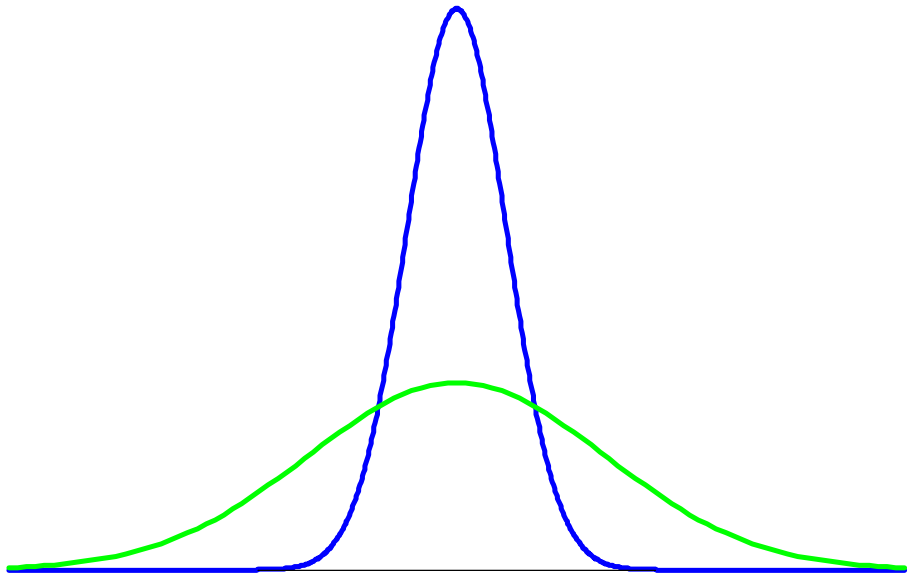
$$z_i = \frac{x_i - \mu_X}{\sigma_X}$$

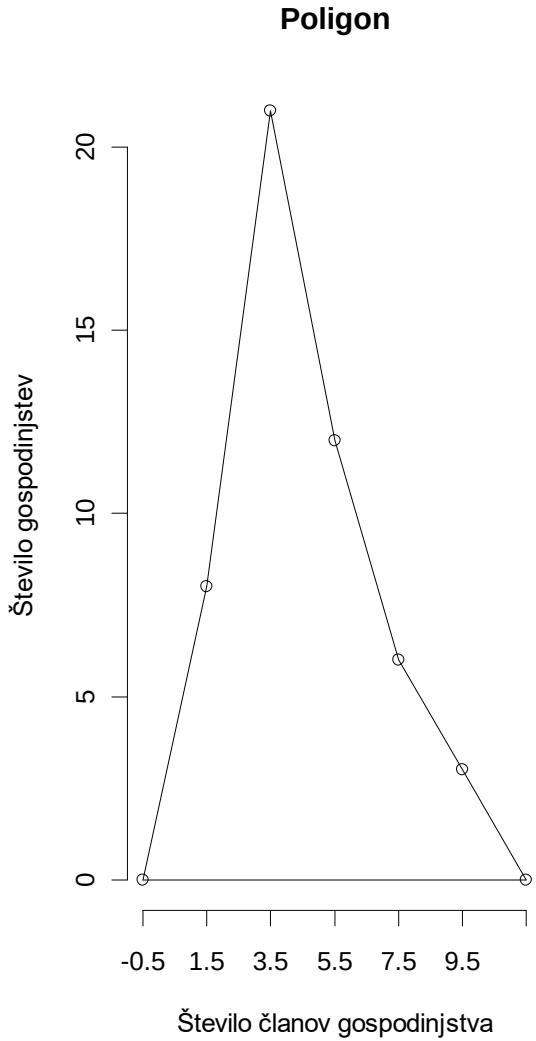
Standardizirana vrednost z_i za vrednost x_i predstavlja relativni odklon od aritmetične sredine.

Vrednosti različnih spremenljivk v splošnem niso primerljive. Če pa spremenljivke standardiziramo, lahko primerjamo njihove standardizirane vrednosti.

Mere asimetrije in sploščenosti







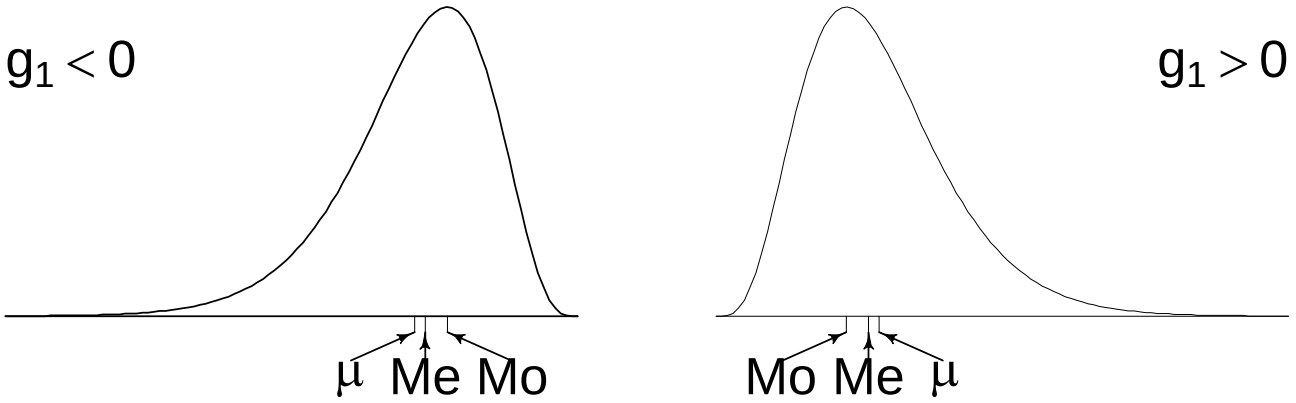
Koeficient asimetrije

$$g_1 = \frac{m_3}{\sqrt{m_2}^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}\right)^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left(\sqrt{\sigma^2}\right)^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\sigma^3}$$

$g_1 > 0$ ... asimetrija v desno

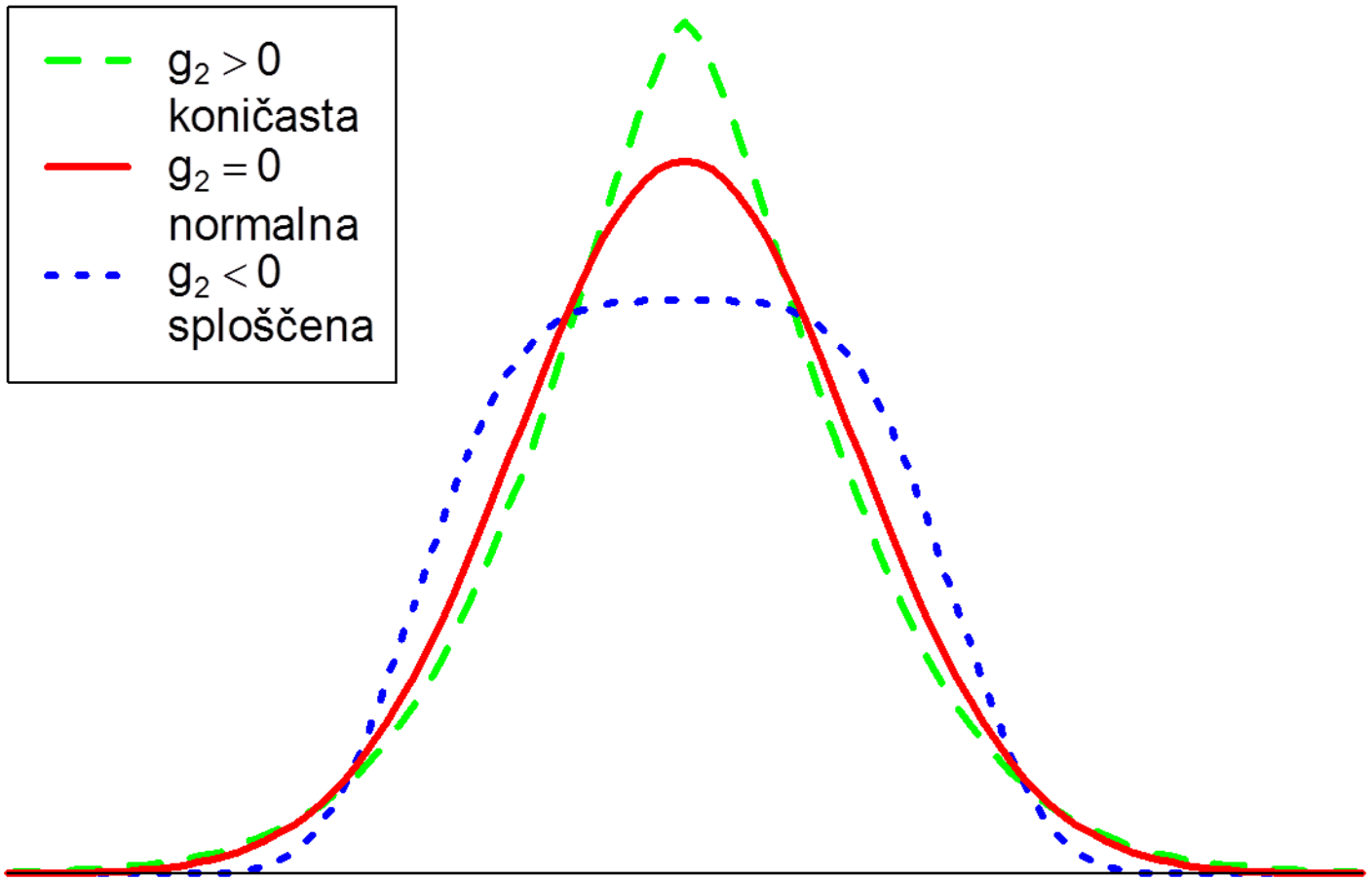
$g_1 = 0$ ... simetrična

$g_1 < 0$ ... asimetrična v levo



Koeficient sploščenosti

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2)^2} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{(\sigma^2)^2} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} - 3$$



Statistično sklepanje

Sklepanje o populaciji na osnovi vzorcev, ki so bili slučajno izbrani iz te (neznane) populacije ➔ **statistično sklepanje** (angl. *statistical inference*), s katerim se ukvarja področje **inferenčne statistike**.

Statistično sklepanje uporablja verjetnostno teorijo za to, da pove, koliko lahko zaupamo rezultatom, pridobljenim na verjetnostnem vzorcu.

Statistično sklepanje = sklepanje iz vzorca na populacijo:

1. ocena populacijskih parametrov na osnovi vzorčnih statistik – točkovne ocene, intervali zaupanja;
2. testiranje domnev o populaciji.

Intervali zaupanja

interval zaupanja – nakazuje na točnost ocene ter vsebuje informacijo o zanesljivosti te ocene (z določeno verjetnostjo “zaupamo”, “smo gotovi” v pravilnost ocene).

$P(a < \gamma < b) = 1 - \alpha$

Interpretiramo:

S tveganjem α (oz. z gotovostjo 1 - α) lahko trdimo, da se populacijski parameter γ nahaja v tem intervalu. **Pozor:** Verjetnost se nanaša na postopek izračuna (vključno z vzorčenjem) in ne na sam interval oz. vrednosti.

Preverjanje domnev

Uporabimo ga za preverjanje, ali podatki podpirajo našo (ničelno domnevo)

Postopek:

- Postavimo ničelno in alternativno domnevo
- Izberemo sprejemljivo stopnjo tveganja (α)
- Izberemo ustrezno testno statistiko in jo izračunamo
- Izračunamo *p* vrednost (kako verjetno je, da bi tako vrednost testne statistike dobili ob pravilni ničelni hipotezi)
- Če je $p < \alpha$, zavrnemo ničelno hipotezo

Preverjanje domnev o aritmetičnih sredinah

- o vrednosti aritmetične sredine → t-test za en vzorec
- o enakosti aritmetičnih sredin na dveh odvisnih vzorcih → t-test za odvisne vzorce
- o enakosti aritmetičnih sredin na dveh neodvisnih vzorcih → t-test za neodvisne vzorce
- o enakosti aritmetičnih sredin na (več kot dveh) neodvisnih vzorcih → Enofaktorska ANOVA

Bivariatna statistika

Povezanost med:

- Nominalnimi spremenljivkami → χ^2 in kontingenčni koeficienti
- Ordinalni spremenljivki → Spearmanov koeficient korelacije
- Intervalne/razmernostne spremenljivke → Pearsonov koeficient korelacije
- Intervalno in nominalno spremenljivko → Primerjava povprečji po kategorijah

Povezanost med nominalnimi spremenljivkami

- Podatke uredimo v kontingenčno tabelo.
- Za preverjanje domneve o povezanosti uporabimo χ^2 statistiko
- Za izračun moči uporabimo kontingenčne koeficiente (npr. Cramerjev α in Pearsonov (popravljen) koeficient kontingence)

Povezanost med ordinalnimi spremenljivkami

- Za merjenje moči in smeri povezanosti uporabimo Spearmanov koeficient korelacije (rangov) – ρ ali ρ_s (na vzorcu r_s)
- Za preverjanje domneve o povezanosti uporabimo testno statistiko izračunano iz r_s

Povezanost med intervalnimi in razmernostnimi spremenljivkami

- Za merjenje moči in smeri povezanosti uporabimo Pearsonov koeficient (**linearne**) korelacije – r ali ρ
- Za preverjanje domneve o povezanosti uporabimo testno statistiko izračunano iz r.

Povezanost med intervalno in nominalno spremenljivko

- Za merjenje moči lahko primerjamo razlike med povprečji po kategorijah (čeprav to ni ravno neka statistika na intervalu [0,1]). Tako statistiko bi bilo sicer moč izračunati, a ne bomo kompliciral.
- Za preverjanje domneve o povezanosti uporabimo t-test za neodvisne vzorce (če ima nominalna spremenljivka samo 2 kategoriji) ali ANOVA test (več kategorij).

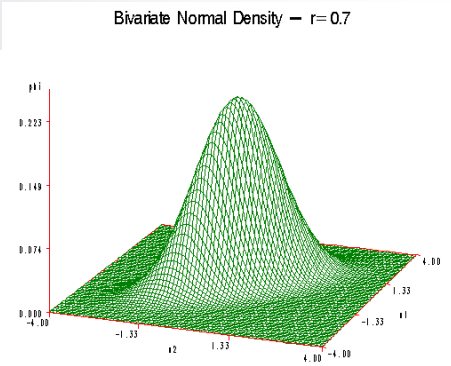
Likartova lestvica

- Likartovo lestvico konstruiramo iz večjega števila spremenljivk, ki nam vse merijo isti koncept.
- Vrednosti vseh spremenljivk seštejemo (ter dobljeno vsoto delimo s številom spremenljivk).
- Vse spremenljivke morajo biti “obrnjene v isto smer”, oz. da vrednosti posamezne spremenljivke pri vsaki spremenljivki pomenijo
- Nova (konstruirana) spremenljivka je intervalnega tipa.
- Npr: X = (X1 + X2 + X3 +X4 + X5)/5

Multivariatna normalna porazdelitev

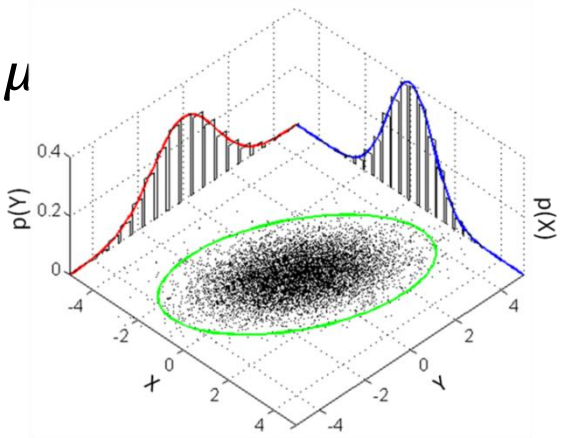
- Večina klasičnih MV metod (regresija, FA, diskriminantna, kanonična kor. analiza, SEM, ...) predpostavlja MV normalno porazdelitev
- Kršenje predpostavke ima za posledico:
  - Napačne parametre modelov
  - Napačno testiranje hipotez / statistično značilnost (p).
- Velikost napake je odvisna od:
  - Vrste metode (različne metode so različno občutljive na kršitve)
  - Velikosti vzorca

Multivariatna normalna porazdelitev

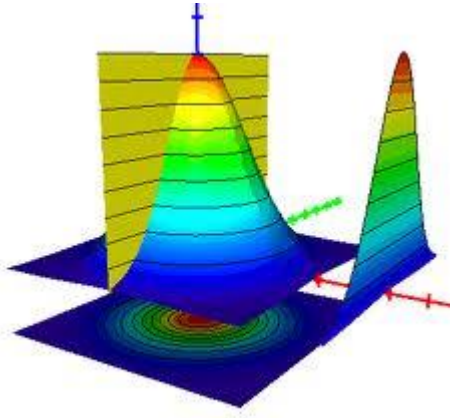


- Gostota:

$$f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu}$$



- kjer je *x* slučajni vektor, *p* število spremenljivk, μ vektor aritmetičnih sredin, Σ pa variančno-kovariančna matrika.
- $x \sim N_p(\mu, \Sigma)$
- Tudi robna in pogojna porazdelitev sta (multivariatno) normalni



Multivariatna normalna porazdelitev – interaktivni prikazi

- http://socr.ucla.edu/htmls/HTML5/BivariateNormal/
- https://demonstrations.wolfram.com/TheBivariateNormalDistribution/
- https://demonstrations.wolfram.com/TheBivariateNormalAndConditionalDistributions/