Statistika

Varstvena biologija & Bioinformatika

1. predavanje

Barbara Boldin

Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Univerza na Primorskem

S čim se ukvarja statistika?

Statistika je veda, ki preučuje množične podatke/pojave in zajema:

- zbiranje podatkov

  kako ob omejitvah (časovnih, finančnih itd.) zbrati podatke, da nam le ti dajo čim bolj zanesljivo informacijo o populaciji oz. pojavu?

  npr.: zanima nas povprečni razpon kril odraslih planinskih orlov v Sloveniji. Ker izvedba meritev za celotno populacijo ni izvedljiva, želimo povprečni razpon kril oceniti na podlagi vzorca. Na kakšen način vzorčiti in kako velik naj bo vzorec?

  Pri tem je pomembna teorija vzorčenja.

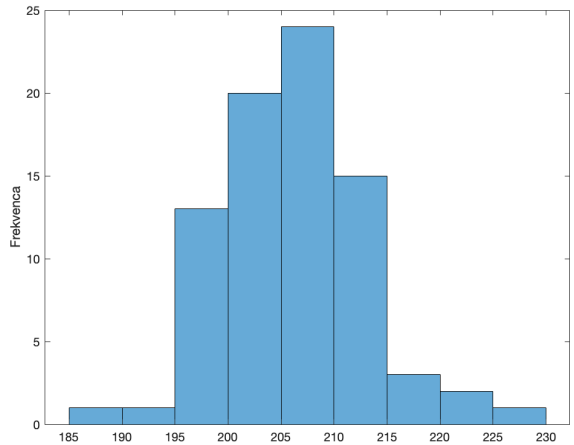
- urejanje in povzemanje podatkov

t.j. kako predstaviti zbrane podatke, da bomo dobili čim bolj pregledno sliko?

Izmerili smo razpon kril 80 naključno izbranim planinskim orlom v Sloveniji in dobili naslednje podatke (v *cm*):

208 216 191 210 207 197 202 207 226 222 197 223 209 205 209 204
204 214 213 214 209 198 209 215 208 211 209 203 207 200 210 198
199 200 187 214 207 200 213 195 204 204 207 207 200 205 204 209
212 212 200 205 198 198 205 214 200 207 204 212 198 205 208 212
214 206 196 201 199 219 201 209 204 210 200 197 196 208 204 204

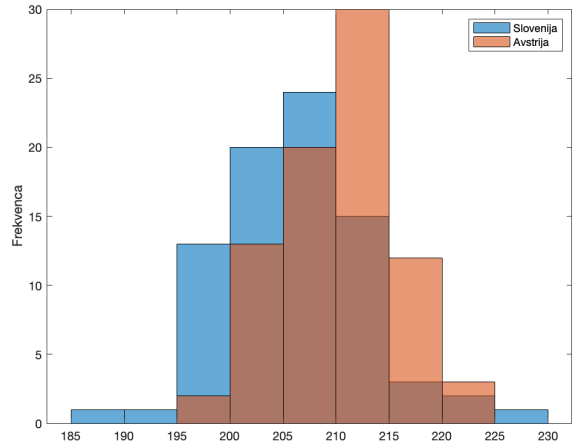
Ti podatki nam ne povedo kaj dosti. Nekaj infomacije nam da *aritmetična sredina* podatkov (205.85 *cm*), informativen pa je tudi *histogram*



Z urejanjem in povzemanjem podatkov se ukvarja opisna statistika.

- vrednotenje podatkov & statistično sklepanje

  Npr.: raziskovalci so izmerili še razpon kril 80 naključno izbranim planinskim orlom v Avstriji. Aritmetična sredina teh meritev je bila 209.9 *cm*, histograma obeh meritev pa sta prikazana na sliki. Ali lahko trdimo, da imajo planinski orli v Avstriji v povprečju večji razpon kril kot orli v Sloveniji?



S takimi vprašanji se ukvarja inferenčna statistika. Matematično ozadje inferenčne statistike je teorija verjetnosti.

Osnovni statistični pojmi

- Statistična populacija je zbirka enot, ki jih preučujemo.

  Npr.: študenti na UP FAMNIT v 2020/21; nova vozila, prvič registrirana v RS v aprilu 2020; itd.

  Vzorec je podmnožica statistične populacije.

- Statistična spremenljivka je predpis, ki vsaki enoti populacije priredi določeno vrednost.

  Npr.: za stat. populacijo "študenti na UP FAMNIT v 2020/21" so statistične spremenljivke "študijska smer", "letnik študija", "spol" itd.

  Statistične spremenljivke označujemo z veliki črkami (npr. X, Y), njihove vrednosti pa z malimi.

  Npr.: vrednosti spremenljivke X označimo z $x_1, x_2, \ldots$

  Statistične spremenljivke delimo na:

  - opisne (kvalitativne)

    Npr.: spol, smer študija, znamka avtomobila itd.
  - številske (kvantitativne)

    Npr.: nadmorska višina, temperatura, letnica rojstva itd.

- Merska lestvica se nanaša na statistično spremenljivko in pove, kakšno strukturo imajo vrednosti statistične spremenljivke ter kakšne operacije so smiselne.

  Pri opisnih slučajnih spremenljivkah lestvice delimo na

  - ♣ *imenske* (oz. *nominalne*)

    spremljamo le vrednosti same (računske operacije niso definirane), npr.: barva oči, pasma in krvna skupina.
  - ♠ *urejenostne* (oz. *ordinalne*),

    lahko povemo, katera vrednost je manjša oz. večja, npr.: stopnja izobrazbe, ocena počutja.

  Številske delimo na

  - ♣ *intervalske*

    smiselno je računanje razlik med dvema vrednostima, medtem ko seštevanje, množenje ali deljenje samih vrednosti ni smiselno (npr.: letnica rojstva, nadmorska višina itd.)
  - ♠ *razmernostne*

    vrednosti smiselno seštevamo, odštevamo in delimo (npr.: moč motorja, dohodek)

Strukture

Naj bo statistična populacija razdeljena v K skupin glede na izbrano statistično spremenljivko.
Naj f_i označuje število enot v i-ti skupini ($i = 1, \ldots, K$). Število f_i imenujemo frekvenca i-te skupine.

Delež enot v i-ti skupini je

$$f_i^\circ = \frac{f_i}{\sum_{i=1}^{K} f_i}.$$

Vrednosti f_i° zavzamejo vrednosti med 0 in 1, njihova vsota pa je enaka 1, t.j. $\sum_{i=1}^{K} f_i^\circ = 1$.

Deleže lahko izrazimo tudi v odstotkih kot

$$f_i\% = 100 \cdot \frac{f_i}{\sum_{i=1}^{K} f_i}.$$

Števila $f_i\%$ zavzamejo vrednosti med 0 in 100, njihova vsota pa je enaka 100, t.j. $\sum_{i=1}^{K} f_i\% = 100$.
Vrednosti f_i° in $f_i\%$ imenujemo relativne frekvence i-te skupine.

Podatke lahko grafično predstavimo v obliki strukturnega stolpca ali strukturnega kroga.

Primer. Podatki o številu novih doktorjev znanosti v letu 2012 po področjih in po spolu. (vir podatkov: www.stat.si)

Področje	Moški	Ženske	Skupaj
Naravoslovne vede (NV)	119	112	231
Tehniške vede (TV)	166	36	202
Medicinske vede (MV)	76	97	173
Kmetijske vede (KV)	12	18	30
Družbene vede (DV)	141	146	287
Humanistične vede (HV)	44	60	104
Skupaj	558	469	1027

- Kakšna je struktura novih doktorjev po spolu?
- Kakšna je struktura novih doktorjev po področjih?
- Kakšna je struktura po spolu znotraj vsakega področja?
- Kakšna je struktura po področjih pri ženskah, kakšna pri moških?

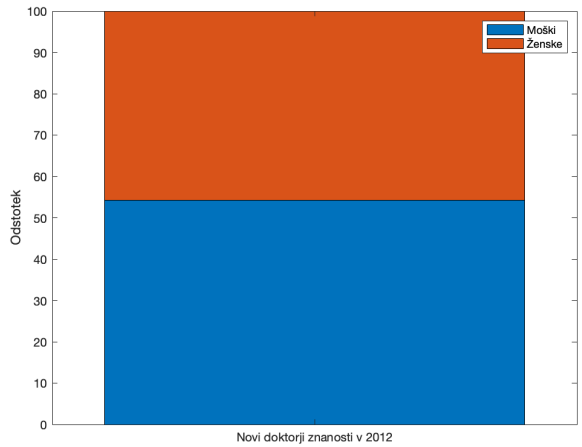
Primer. Podatki o številu novih doktorjev znanosti v letu 2012 po področjih in po spolu. (vir podatkov: www.stat.si)

Področje	Moški	Ženske	Skupaj
Naravoslovne vede (NV)	119	112	231
Tehniške vede (TV)	166	36	202
Medicinske vede (MV)	76	97	173
Kmetijske vede (KV)	12	18	30
Družbene vede (DV)	141	146	287
Humanistične vede (HV)	44	60	104
Skupaj	558	469	1027

- Kakšna je struktura novih doktorjev po spolu?
- Kakšna je struktura novih doktorjev po področjih?
- Kakšna je struktura po spolu znotraj vsakega področja?
- Kakšna je struktura po področjih pri ženskah, kakšna pri moških?

Struktura po spolu:

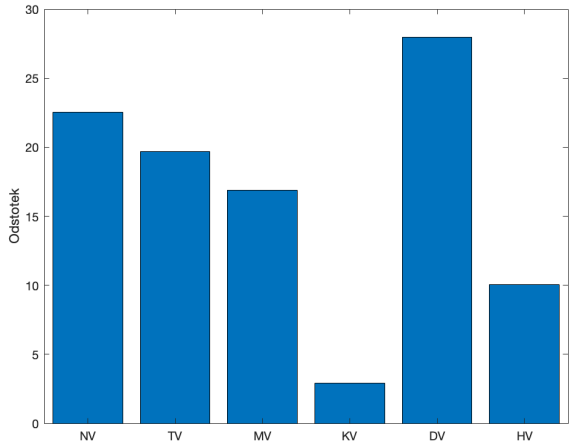
$$f_M = \frac{558}{1027} = 54.3\%, \qquad f_Z = \frac{469}{1027} = 45.7\%.$$



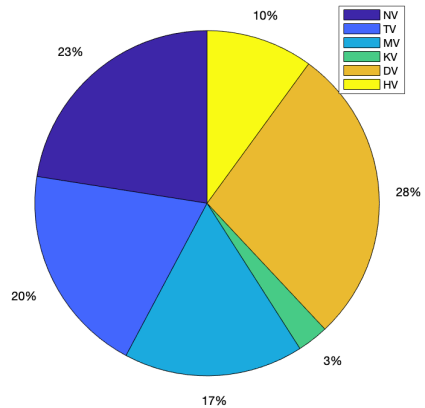
Interpretacija?

Struktura po področjih:

$$f_{NV} = 22.49\% \qquad f_{TV} = 19.67\% \qquad f_{MV} = 16.85\%$$
$$f_{KV} = 2.92\% \qquad f_{DV} = 27.95\% \qquad f_{HV} = 10.13\%$$



Strukturni stolpci

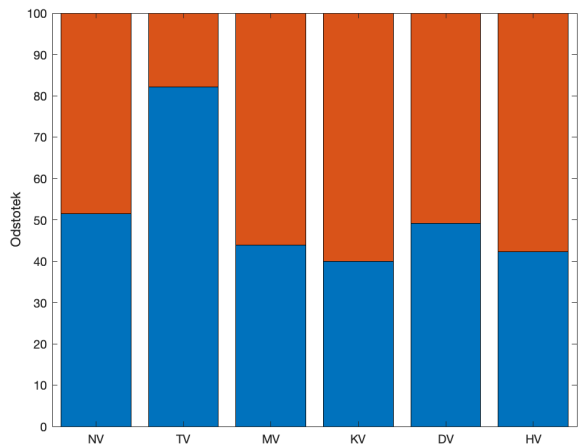


Strukturni krog

Interpretacija?

Struktura po spolu znotraj vsake znanstvene vede:

$$f_M^{NV} = 51.52\% \qquad f_{\check{Z}}^{NV} = 48.48\%$$
$$f_M^{TV} = 82.18\% \qquad f_{\check{Z}}^{TV} = 17.82\%$$
$$f_M^{MV} = 43.93\% \qquad f_{\check{Z}}^{MV} = 56.07\%$$
$$f_M^{KV} = 40.00\% \qquad f_{\check{Z}}^{KV} = 60.00\%$$
$$f_M^{DV} = 49.13\% \qquad f_{\check{Z}}^{DV} = 50.87\%$$
$$f_M^{HV} = 42.31\% \qquad f_{\check{Z}}^{HV} = 57.69\%$$



Interpretacija?

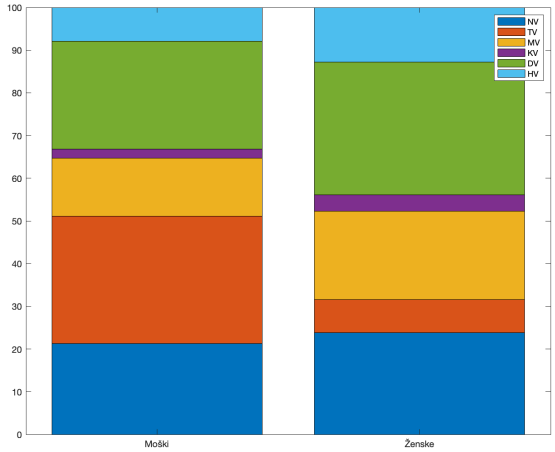
Struktura po področjih pri vsakem od obeh spolov.

Moški:

$$f^M_{NV} = 21.32\% \qquad f^M_{TV} = 29.75\% \qquad f^M_{MV} = 13.62\%$$
$$f^M_{KV} = 2.15\% \qquad f^M_{DV} = 25.27\% \qquad f^M_{HV} = 7.89\%$$

Ženske:

$$f^{\check{Z}}_{NV} = 23.88\% \qquad f^{\check{Z}}_{TV} = 7.68\% \qquad f^{\check{Z}}_{MV} = 20.68\%$$
$$f^{\check{Z}}_{KV} = 3.83\% \qquad f^{\check{Z}}_{DV} = 31.13\% \qquad f^{\check{Z}}_{HV} = 12.79\%$$



Interpretacija?

Koeficienti

Koeficient je razmerje dveh podatkov.

Npr.:

- gostota prebivalstva (število prebivalcev na km^2)
- stopnja rodnosti (število živorojenih na 1000 prebivalcev)
- število bolnikov na zdravnika
- hitrost
- poraba goriva
- itd.

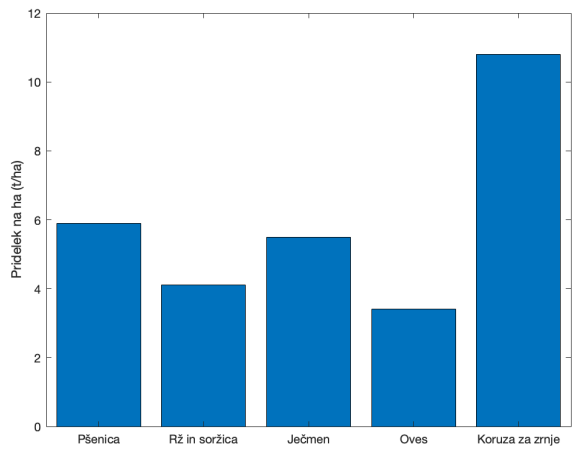
Posebne vrste koeficientov so t.i. stopnje, ki jih pogosto srečamo v medicini in demografiji.

Primer. Podatki o pridelovalnih površinah in pridelku za izbrana žita v Republiki Sloveniji v letu 2020 (vir: www.stat.si).
Smiselni koeficient v tem primeru je pridelek na hektar.

Vrsta	Površina (*ha*)	Pridelek (*t*)	Pridelek na *ha* (*t**ha*)
Pšenica	26761	156771	
Rž in soržica	827	3412	
Ječmen	22212	122214	
Oves	809	2718	
Koruza za zrnje	39836	429925	

Primer. Podatki o pridelovalnih površinah in pridelku za izbrana žita v Republiki Sloveniji v letu 2020 (vir: www.stat.si).
Smiselni koeficient v tem primeru je pridelek na hektar.

Vrsta	Površina (*ha*)	Pridelek (*t*)	Pridelek na *ha* (*t**ha*)
Pšenica	26761	156771	5.9
Rž in soržica	827	3412	4.1
Ječmen	22212	122214	5.5
Oves	809	2718	3.4
Koruza za zrnje	39836	429925	10.8



Indeksi

Indeksi s stalno osnovo

Naj bo X številska slučajna spremenljivka in naj x_i označuje vrednost spremenljivke X v i-ti skupini ($i = 1, \ldots, K$).

Zaporedje

$$x_1, x_2, \ldots, x_K$$

imenujemo statistična vrsta.

Izberimo en podatek v statistični vrsti in ga označimo z x_0.
Število x_0 imenujemo (stalna) *osnova*. Indeksi z osnovo x_0 so

$$I_{i\backslash 0} = 100 \cdot \frac{x_i}{x_0}, \quad i = 1, \ldots, K.$$

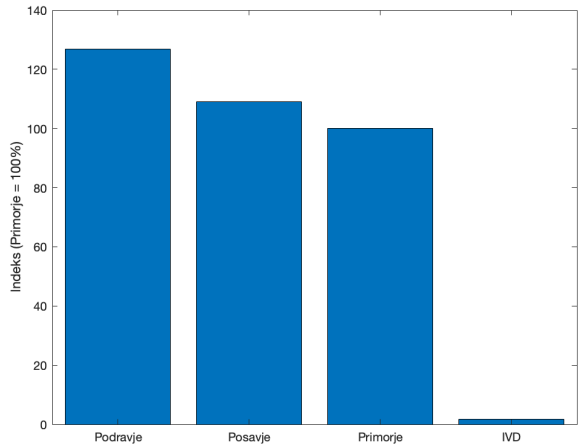
Za izračun indeksov mora biti merska lestvica spremenljivke razmernostna, vrednosti pa pozitivne.

Primer. Podatki o številu nasadov v vinogradih v različnih regijah v Sloveniji v letu 2020 (vir: www.stat.si).
Izračunajmo indekse z osnovo “Primorje”.

Vinorodna dežela	Število nasadov	Indeksi (osnova = “Primorje”)
Podravje	17769	
Posavje	15292	
Primorje	14017	
Izven vinorodnih dežel (IVD)	246	

Primer. Podatki o številu nasadov v vinogradih v različnih regijah v Sloveniji v letu 2020 (vir: www.stat.si).
Izračunajmo indekse z osnovo “Primorje”.

Vinorodna dežela	Število nasadov	Indeksi (osnova = “Primorje”)
Podravje	17769	126.77
Posavje	15292	109.10
Primorje	14017	100
Izven vinorodnih dežel (IVD)	246	1.76



Stolpični diagram

Indeksi

Verižni indeksi

Kadar številsko spremenljivko X spremljamo v času, potem statistično vrsto $x_1, x_2, \ldots, x_T$ imenujemo časovna vrsta, število T je dolžina statistične vrste.
Za časovne vrste lahko poleg indeksov s stalno osnovo obravnavamo tudi indekse s premično osnovo, oziroma verižne indekse. Za osnovo izberemo predhodni podatek v časovni vrsti

$$I_{t\backslash t-1} = 100 \cdot \frac{x_t}{x_{t-1}}, \quad t = 2, \ldots, T$$

Ker prvi podatek v časovni vrsti nima predhodnika, prvi verižni indeks ne obstaja. Verižni indeksi so smiselni le, če je časovna vrsta ekvidistantna.

Definirajmo še stopnjo rasti

$$S_{t\backslash t-1} = I_{t\backslash t-1} - 100\%.$$

Stopnja rasti v odstotkih izraža prirast oziroma upad v primerjavi s prejšnjim časovnim obdobjem. Lahko jo zapišemo tudi kot

$$S_{t\backslash t-1} = 100 \cdot \frac{x_t - x_{t-1}}{x_{t-1}}.$$

Primer. Podatki o pridelavi čebule in paradižnika (v tonah) v Republiki Sloveniji v letih 2010-2020. (vir: www.stat.si)

Izračunajmo in interpretirajmo indekse s stalno osnovo “2010” in verižne indekse za čebulo (izračune za paradižnik naredite za vajo!).

Leto	Čebula	Paradižnik
2010	4667	3766
2011	6333	5512
2012	5869	7313
2013	6074	6892
2014	7563	6607
2015	7286	8704
2016	10885	8652
2017	9709	8396
2018	9612	8392
2019	10101	9013
2020	12042	10415

Primer. Podatki o pridelavi čebule in paradižnika (v tonah) v Republiki Sloveniji v letih 2010-2020. (vir: www.stat.si)

Izračunajmo in interpretirajmo indekse s stalno osnovo "2010" in verižne indekse za čebulo (izračune za paradižnik naredite za vajo!).

Leto	Čebula	Indeksi (osnova = "2010")	Verižni indeksi	Stopnja rasti
2010	4667	100		
2011	6333	135.70		
2012	5869	125.76		
2013	6074	130.15		
2014	7563	162.05		
2015	7286	156.12		
2016	10885	233.23		
2017	9709	208.04		
2018	9612	205.96		
2019	10101	216.43		
2020	12042	258.02		

Primer. Podatki o pridelavi čebule in paradižnika (v tonah) v Republiki Sloveniji v letih 2010-2020. (vir: www.stat.si)

Izračunajmo in interpretirajmo indekse s stalno osnovo "2010" in verižne indekse za čebulo (izračune za paradižnik naredite za vajo!).

Leto	Čebula	Indeksi (osnova = "2010")	Verižni indeksi	Stopnja rasti
2010	4667	100	-	-
2011	6333	135.70	135.70	35.70
2012	5869	125.76	92.67	-7.33
2013	6074	130.15	103.49	3.49
2014	7563	162.05	124.51	24.51
2015	7286	156.12	96.34	-3.66
2016	10885	233.23	149.40	49.40
2017	9709	208.04	89.20	-10.8
2018	9612	205.96	99.00	-1.00
2019	10101	216.43	105.09	5.09
2020	12042	258.02	119.22	19.22

Grafično nam pridelek paradižnika in čebule v Sloveniji v letih 2010-2020 prikaže linijski grafikon

Slika: (a) Pridelava čebule (modra) in paradižnika (rdeča) v RS v letih 2010-2020. (b) Indeksi s stalno osnovo 2010 (modra) in premično osnovo (t.j. verižni indeksi) za pridelek paradižnika v RS v letih 2010-2020.

Interpretacija?