Prvi pogled na podatke

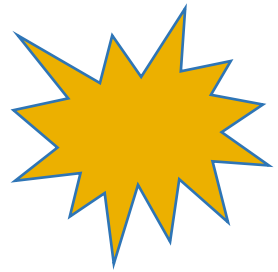
Uvod u znanost o podacima

6. predavanje

doc. dr. sc. Ana Sović Kržić

2021./2022.

Sadržaj

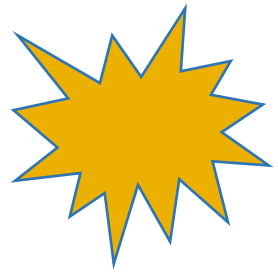


- Deskriptivna statistika
- Inferencijalna statistika

Temeljna pravila vjerojatnosti

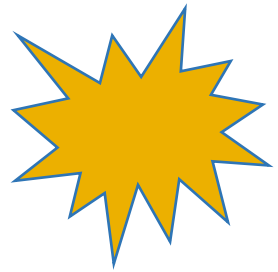
1. Ako je **potpuno sigurno** da će se nešto dogoditi, onda je vjerojatnost tog događaja $p = 1$
2. Ako je **potpuno sigurno** da se nešto **NEĆE** dogoditi, onda je vjerojatnost tog događaja $p = 0$
3. Vjerojatnost da će se između N **događaja** koji su jednako vjerojatni i međusobno nezavisni, **dogoditi jedan određeni** među njima je $p = 1 / N$
4. Vjerojatnost da će se dogoditi **bilo koji od nekoliko mogućih** nezavisnih događaja zbroj je vjerojatnosti svakoga pojedinačnog događaja: $p = p_1 + p_2 + \ldots$
5. Vjerojatnost da će se **zajedno dogoditi dva ili više** nezavisnih događaja produkt je vjerojatnosti svakoga od tih događaja: $p = p_1 \cdot p_2 \cdot \ldots$

Deskriptivna statistika



- Karakteristike konkretnog uzorka

Središnje vrijednosti



- Aritmetička sredina M
- Zajednička aritmetička sredina M_{zaj}
- Centralna vrijednost (medijan) C
- Dominantna vrijednost D
- Geometrijska sredina G
- Harmonijska sredina H

Aritmetička sredina M

$$M = \frac{zbroj\ svih\ rezulata}{broj\ rezultata} = \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$$

- pokazatelj **prave vrijednosti mjerenja**
- Uvjeti korištenja:
  1. rezultati moraju biti **prave mjerene vrijednosti**, dobivene barem na intervalnoj ljestvici
  2. svi rezultati moraju biti dobiveni u **jednakim uvjetima mjerenja**
  3. potreban je dovoljan broj rezultata, **najmanje 30**
  4. distribucija rezultata mora biti **normalna** (što ujedno znači i simetrična)

- **Nesistematski varijabilni faktori**
  - čimbenici koji djeluju prilikom mjerenja, a koje ne možemo kontrolirati ili ih ne znamo
  - vjerujemo da djeluju po zakonu slučaja
  - distribucija im je normalna
- **Sistematski varijabilni faktori**
  - distribucija rezultata nije normalna – na rezultate su djelovali neki sistematski čimbenici
  - **aritmetička sredina neće dati pravu vrijednost mjerenja**
- **Normalna distribucija**
  - **NE jamči da smo izmjerili pravu vrijednost** koju smo željeli mjeriti
  - pokazuje da uz predmet mjerenja i nesistematske varijabilne faktore najvjerojatnije NISU djelovali sistematski faktori

pojedini rezultat nekog mjerenja Y $(M_Y = 0)$
prava vrijednost mjerenja μ
nesistematski varijabilni faktori ϵ $(M_\epsilon = 0)$

$$Y = \mu + \epsilon$$
$$M_Y = \mu + M_\epsilon = \mu$$

Zajednička aritmetička sredina

Ako su aritmetičke sredine izračunate **iz jednakog broja rezultata** (n_M broj aritmetičkih sredina):

$$N_1 = N_2 = N_3 = \cdots = N_{n_m}$$

$$M_{zaj} = \frac{1}{n_M}(M_1 + M_2 + \cdots + M_n)$$

Ako aritmetičke sredine **nisu izračunate iz jednakog broja rezultata**:

$$M_{zaj} = \frac{M_1N_1 + M_2N_2 + \cdots + M_nN_n}{N_1 + N_2 + \cdots + N_n}$$

Primjer – zajednička aritmetička sredina

- Neko mjerenje je ponovljeno 6 puta na različitim skupinama ispitanika

1. mjerenje	2. mjerenje	3. mjerenje	4. mjerenje	5. mjerenje	6. mjerenje
$M_1 = 18.5$ $N_1 = 5$	$M_2 = 22.0$ $N_2 = 17$	$M_3 = 23.9$ $N_3 = 40$	$M_4 = 23.8$ $N_4 = 48$	$M_5 = 22.8$ $N_5 = 19$	$M_6 = 22.6$ $N_6 = 25$

$$M_{zaj} = \frac{M_1N_1 + M_2N_2 + \cdots + M_6N_6}{N_1 + N_2 + \cdots + N_6} = 23.1$$

- Ista vrijednost bi se dobila da je svaki od rezultata uzet pojedinačno
- Kada bismo zanemarili broj uzoraka dobili bismo krivu vrijednost:

$$M_{zaj} = \frac{1}{n_M}(M_1 + M_2 + \cdots + M_6) = 22.3$$

Centralna vrijednost (medijan) C

- vrijednost koja se u nizu rezultata poredanih po veličini nalazi **točno u sredini**
- ako je broj rezultata paran, onda se centralna vrijednost računa kao zbroj dva srednja rezultata i podijeli s 2
- Koristi se ako u nizu rezultata imamo neku **ekstremno veliku ili malu vrijednost** ili je **distribucija rezultata asimetrična**

Dominantna vrijednost (mod) D

- vrijednost koja je u nizu mjerenja **najčešće postignuta**
- na nju ne utječe ni broj ni vrijednost rezultata, već samo frekvencija pojedinih rezultata

Geometrijska sredina G

$$G = \sqrt[N]{X_1 X_2 \ldots X_N}$$

- mjera **prosječne brzine neke promjene**
- npr. koliko puta prosječno je narasla populacija svake godine u nekom mjestu
- ne može se računati ako je neki broj negativan ili nula

Harmonijska sredina H

- **prosjek nekih odnosa**
- npr. prosječni kilometri na sat, prosječan broj slova napisanih u minuti

$$H = \frac{N}{\sum \frac{1}{X}}$$

Primjer

Niz vrijednosti: 1, 2, 4, 4, 4, 5, 6

M=3.71

C=4

D=4

G=3.25

H=2.68

Niz vrijednosti: 1, 2, 4, 4, 4, 5, 60

M=11.43

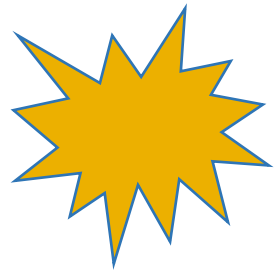
C=4

D=4

G=4.52

H=2.84

Mjere varijabilnosti



- Raspon rezultata R
- Srednje odstupanje (raspršenje) SO
- Varijanca *var*
- Standardna devijacija SD
- Pokuinterkvartalno raspršenje Q
- Koeficijent varijabilnosti V

Raspon rezultata R

- **razlika najvećeg i najmanjeg rezultata**
- outlier značajno povećava raspon, a da se grupacija rezultata oko aritmetičke sredine nije promijenila
- obično je veći što je veći broj mjerenja neke pojave – ako se u obzir uzme samo nekoliko rezultata, smanjena je vjerojatnost da će među njima biti upravo najveći i najmanji rezultat
- često se prikazuje min i max vrijednost umjesto raspona

Srednje odstupanje (raspršenje) SO

- **prosječna veličina odstupanja pojedinačnih rezultata**
- može se računati uz aritmetičku sredinu, centralnu ili dominantnu vrijednost
- grubi pokazatelj razlikovanja rezultata od neke sredine

$$SO = \frac{\sum |X - M|}{N}$$

Varijanca *var*

- prosječni zbroj kvadriranih odstupanja
- Varijanca uzorka

$$var = \frac{\sum(X - M)^2}{N - 1}$$

- Varijanca populacije

$$var = \frac{\sum(X - M)^2}{N}$$

Standardna devijacija SD

- koliko vrijedi dobivena aritmetička sredina – je li ona dobar ili loš reprezentant rezultata
- u jedinicama u kojima su i mjerenja
- smije se računati samo uz aritmetičku sredinu
- Standardna devijacija uzorka

$$SD = \sqrt{var} = \sqrt{\frac{\sum(X - M)^2}{N - 1}}$$

- Standardna devijacija populacije

$$SD = \sqrt{var} = \sqrt{\frac{\sum(X - M)^2}{N}}$$

- kontrola računanja: raspon / standardna devijacija je gotovo uvijek između 2 i 6.5

Poluinterkvartilno raspršenje Q

- niz dobivenih rezultata se poreda po veličini, od manjih prema većima – niz ima 4 kvartila – u svakom se kvartilu nalazi 25% rezultata
- **granične vrijednosti kvartila**: Q_1, Q_2 ($Q_2 = C$ centralna vrijednost dijeli niz na dva dijela), Q_3, Q_4 (gornja granica)
- redno mjesto graničnih vrijednosti: $R_{Q_1} = \frac{N}{4} + 0.5, R_{Q_3} = \frac{N}{4} \cdot 3 + 0.5$
- poluinterkvartilno raspršenje Q – **polovina razlike između graničnih vrijednosti trećeg i prvog kvartila**

$$Q = \frac{Q_3 - Q_1}{2}$$

Koeficijent varijabilnosti (varijacije) V

- koliki postotak vrijednosti aritmetičke sredine iznosi vrijednost standardne devijacije

$$V = \frac{SD}{M} \cdot 100$$

- da bi se mogle **uspoređivati varijabilnosti različitih pojava i svojstava** (npr. što je povoljnije $M_1 = 100, SD_1 = 10$ ili $M_2 = 8, SD_2 = 2$ – prvi slučaj je povoljniji)
- koristi se kada želimo znati:
  - u kojem svojstvu neka grupa varira više, a u kojem manje
  - koja od grupa varira više, a koja manje u istom svojstvu

Primjer – težina i visina desetogodišnjaka

Dječaci, visina	Dječaci, težina	Djevojčice, visina	Djevojčice, težina
$N_1 = 612$ $M_1 = 134.4$ cm $SD_1 = 6.06$ cm $V = \frac{6.06}{134.4} \cdot 100$ $= 4.51\%$	$N_2 = 612$ $M_2 = 29.2$ kg $SD_2 = 3.89$ kg $V = \frac{3.89}{29.2} \cdot 100$ $= 13.32\%$	$N_1 = 684$ $M_1 = 134.9$ cm $SD_1 = 6.43$ cm $V = \frac{6.43}{134.9} \cdot 100$ $= 4.77\%$	$N_2 = 684$ $M_2 = 29.7$ kg $SD_2 = 4.78$ kg $V = \frac{4.78}{29.7} \cdot 100$ $= 16.09\%$

$$V = \frac{SD}{M} \cdot 100$$

Variraju li dječaci više u visini ili težini? U težini.

Variraju li u visini više djevojčice ili dječaci? Djevojčice.

Variraju li u težini više djevojčice ili dječaci? Djevojčice.

Asimetrične distribucije

- kada se pri primjeni upitnika i testova dobije **asimetrična distribucija** – radi se najvjerojatnije o nekim pogreškama ili propustima pri provedbi mjerenja (utjecaj nepoželjnih čimbenika na rezultate) – znak istraživaču da provjeri prikupljanje podataka
- npr. ako su različite osobe radile na upitniku u različitim prilikama distrakcije (u tišini ili buci)

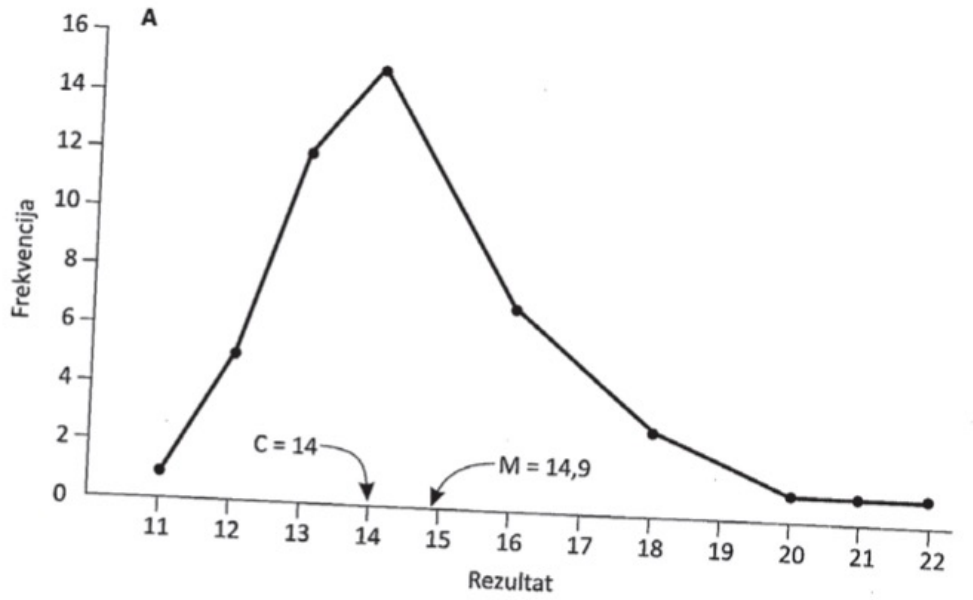
Indeks asimetrije (1)

$$\alpha = \frac{3(M - C)}{SD}$$

- **simetrična distribucija**: M = C, $\alpha = 0$
- **pozitivno asimetrična distribucija**: M>C, $\alpha > 0$
- **negativno asimetrična distribucija**: M<C, $\alpha < 0$
- rijetko se koristi jer nije jasno koliku asimetriju njegova numerička vrijednost pokazuje, pa ne omogućuje usporedbe – za utvrđivanje razlikuje li se dobivena distribucija od normalne se koristi Kolomogorov-Smirnov test

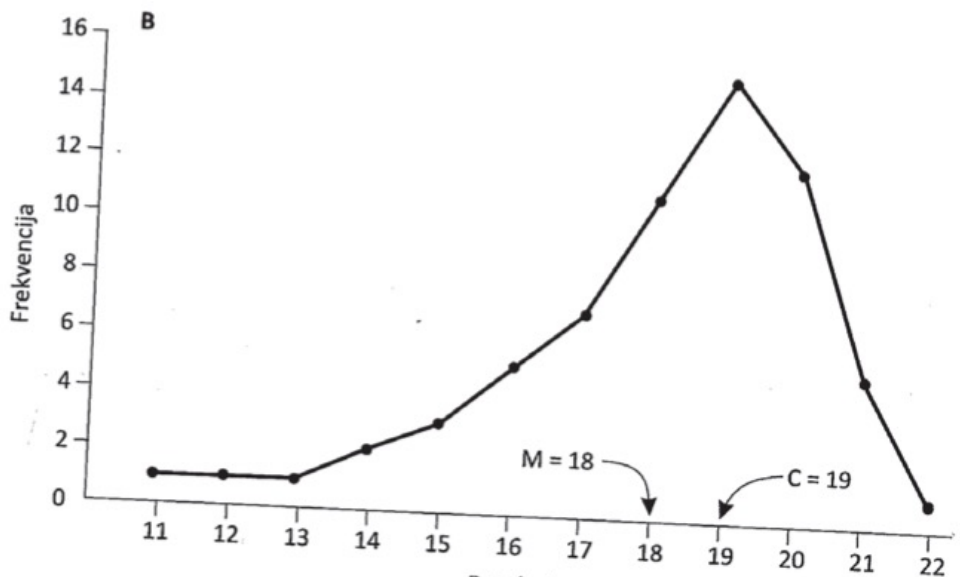
- **Kurtosis** – ispupčenje, zakrivljenost, konveksnost – vrsta odstupanja od normalne distribucije

Indeks asimetrije (2)



pozitivno asimetrična distribucija

$$\alpha = \frac{3(M-C)}{SD} = \frac{3(14.9-14)}{SD} = \frac{2.7}{SD} > 0$$

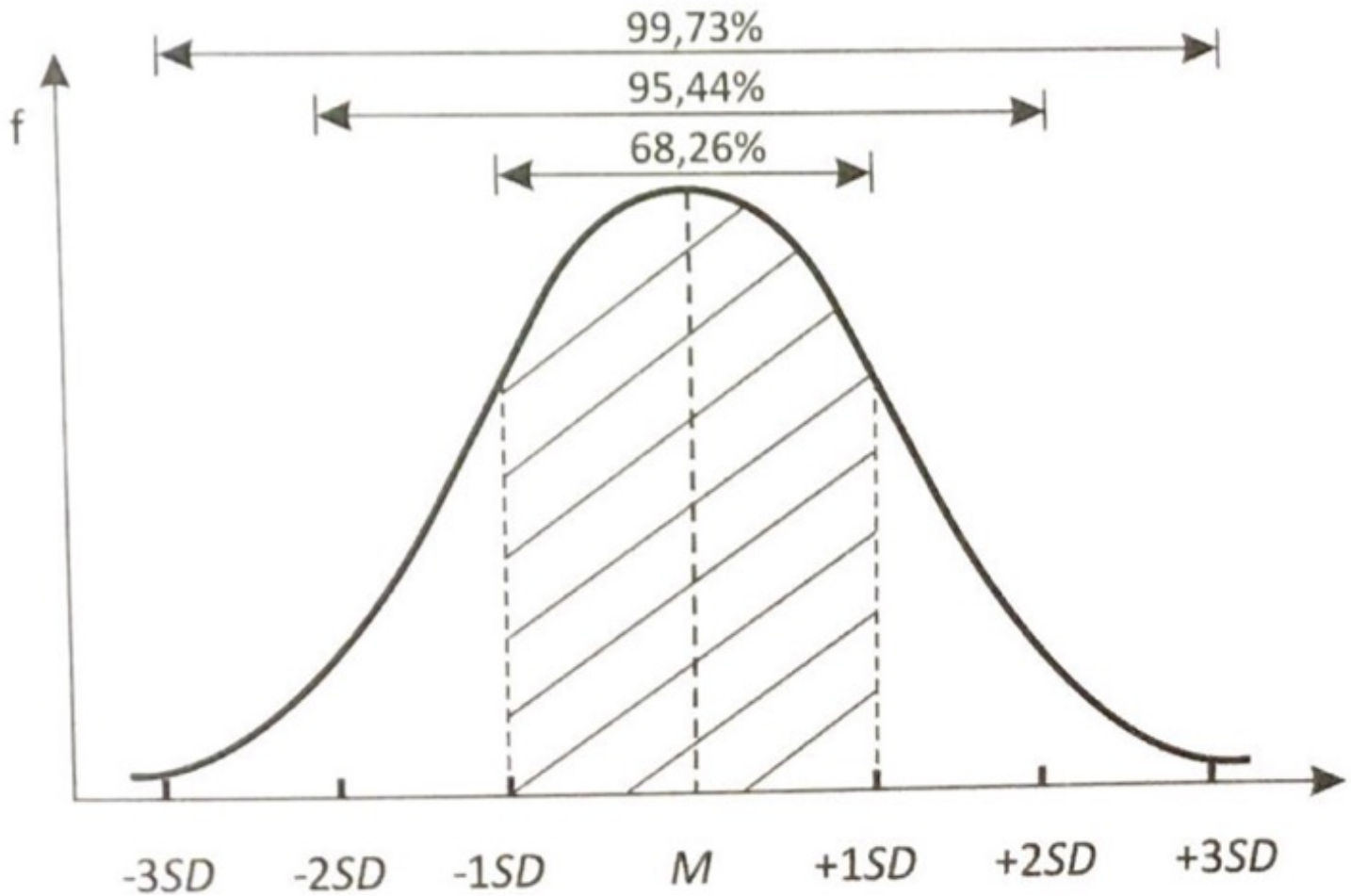


negativno asimetrična distribucija

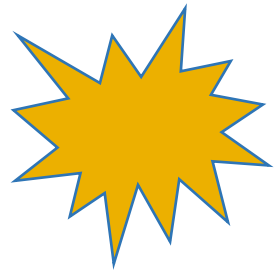
$$\alpha = \frac{3(M-C)}{SD} = \frac{3(18-19)}{SD} = \frac{-3}{SD} < 0$$

Normalna distribucija

- Uvjeti:
  - ako se može pretpostaviti da **postoji prava vrijednost mjerenja** koja je relativno stabilna u vremenu te da pri njenom mjerenju djeluju, osim nje same, samo još **nesistematski varijabilni faktori**
  - da imamo **veliki broj mjerenja**
  - da su sva **mjerenja provedena jednakom metodom** i u što sličnijim vanjskim prilikama (npr. eksperimentalna i kontrolna skupina moraju biti izjednačene u svim ostalim faktorima, osim u onom koji upravo istražujemo)
  - skupina u kojoj obavljamo mjerenja mora biti **homogena po svim drugim svojstvima**, a heterogena po onom svojstvu koji mjerimo


99,73%
95,44%
68,26%
f
-3SD
-2SD
-1SD
M
+1SD
+2SD
+3SD

Položaj rezultata u grupi



- z-vrijednost
- primjeri:
  - postotak rezultata iznad ili ispod neke vrijednosti
  - broj rezultata iznad / ispod / između neke vrijednosti
  - usporedba rezultata različitih ljudi ili različita mjerenja istog čovjeka
  - skupna ili prosječna ocjena iz niza mjerenja s jednakim mjernim jedinicama, ali uz različit varijabilitet rezultata

z-vrijednost

$$z = \frac{X - M}{SD}$$

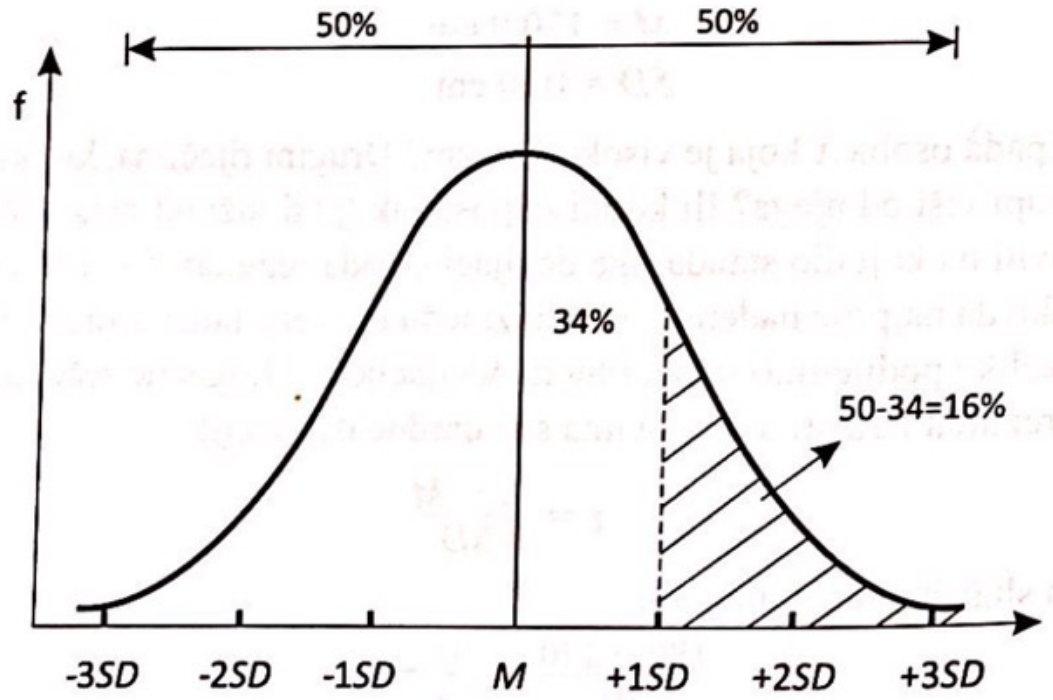
- govori koliko standardnih devijacija SD je promatrani rezultat X udaljen od srednje vrijednosti M
- za lakše određivanje **koliko rezultata je veće ili manje od promatranog rezultata**
- skala z-vrijednosti: M=0, SD kao jedinična vrijednost
- negativni predznak z-vrijednosti = rezultat je ispod aritmetičke sredine

Tablica za z-vrijednosti

POVRŠINA ISPOD NORMALNE KRIVULJE OD DOBIVENE Z-VRIJEDNOSTI
DO **BLIŽEG** KRAJA KRIVULJE

z	P	z	P	z	P	z	P
0,00	0,5000	0,46	0,3228	0,92	0,1788	1,70	0,0446
0,01	0,4960	0,47	0,3192	0,93	0,1762	1,75	0,0401
0,02	0,4920	0,48	0,3156	0,94	0,1736	1,80	0,0359
0,03	0,4880	0,49	0,3121	0,95	0,1711	1,85	0,0322
0,04	0,4840	0,50	0,3085	0,96	0,1685	1,90	0,0287
0,05	0,4801	0,51	0,3050	0,97	0,1660	1,95	0,0256
0,06	0,4761	0,52	0,3015	0,98	0,1635	2,00	0,0228
0,07	0,4721	0,53	0,2981	0,99	0,1611	2,05	0,0202
0,08	0,4681	0,54	0,2946	1,00	0,1587	2,10	0,0179
0,09	0,4641	0,55	0,2912	1,01	0,1562	2,15	0,0158
0,10	0,4602	0,56	0,2877	1,02	0,1539	2,20	0,0139
0,11	0,4562	0,57	0,2843	1,03	0,1515	2,25	0,0122
0,12	0,4522	0,58	0,2810	1,04	0,1492	2,30	0,0107
0,13	0,4483	0,59	0,2776	1,05	0,1469	2,35	0,0094
0,14	0,4443	0,60	0,2743	1,06	0,1446	2,40	0,0082
0,15	0,4404	0,61	0,2709	1,07	0,1423	2,45	0,0071
0,16	0,4364	0,62	0,2676	1,08	0,1401	2,50	0,0062
0,17	0,4325	0,63	0,2643	1,09	0,1379	2,55	0,0054
0,18	0,4286	0,64	0,2611	1,10	0,1357	2,60	0,0045
0,19	0,4247	0,65	0,2578	1,11	0,1335	2,65	0,0040
0,20	0,4207	0,66	0,2546	1,12	0,1314	2,70	0,0035
0,21	0,4168	0,67	0,2514	1,13	0,1292	2,75	0,0030
0,22	0,4129	0,68	0,2483	1,14	0,1271	2,80	0,0026
0,23	0,4090	0,69	0,2451	1,15	0,1251	2,85	0,0022
0,24	0,4052	0,70	0,2420	1,16	0,1230	2,90	0,0019
0,25	0,4013	0,71	0,2389	1,17	0,1210	2,95	0,0016
0,26	0,3974	0,72	0,2358	1,18	0,1190	3,00	0,0014
0,27	0,3936	0,73	0,2327	1,19	0,1170	3,05	0,0012
0,28	0,3897	0,74	0,2296	1,20	0,1151	3,10	0,0010
0,29	0,3859	0,75	0,2266	1,21	0,1131	3,15	0,0008
0,30	0,3821	0,76	0,2236	1,22	0,1112	3,20	0,0007
0,31	0,3783	0,77	0,2206	1,23	0,1093	3,25	0,0006
0,32	0,3745	0,78	0,2177	1,24	0,1075	3,30	0,0005
0,33	0,3707	0,79	0,2148	1,25	0,1056	3,35	0,0004
0,34	0,3669	0,80	0,2119	1,26	0,1038	3,40	0,00034
0,35	0,3632	0,81	0,2090	1,27	0,1020	3,45	0,00028
0,36	0,3594	0,82	0,2061	1,28	0,1003	3,50	0,00023
0,37	0,3557	0,83	0,2033	1,29	0,0985	3,55	0,00019
0,38	0,3520	0,84	0,2005	1,30	0,0968	3,60	0,00016
0,39	0,3483	0,85	0,1977	1,35	0,0885	3,65	0,00013
0,40	0,3446	0,86	0,1949	1,40	0,0808	3,70	0,00011
0,41	0,3409	0,87	0,1922	1,45	0,0735	3,75	0,00009
0,42	0,3372	0,88	0,1894	1,50	0,0668	3,80	0,00007
0,43	0,3336	0,89	0,1867	1,55	0,0606	3,85	0,00006
0,44	0,3300	0,90	0,1841	1,60	0,0548	3,90	0,000048
0,45	0,3264	0,91	0,1814	1,65	0,0495	4,00	0,000032

Prilagođeno od http://www.adamssixsigma.com/Newsletters/standard_normal_table.htm
i http://www.stat.psu.edu/~babu/418/norm-tables.pdf

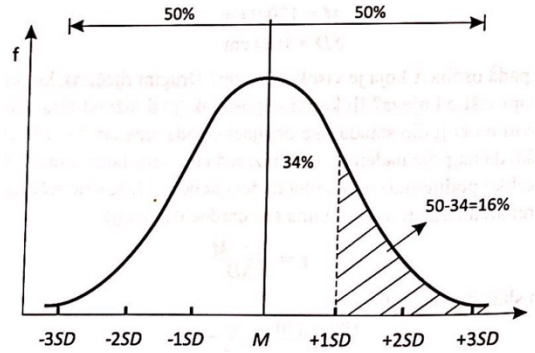


Postotak rezultata iznad ili ispod neke vrijednosti

Problem: Aritmetička sredina visina veće grupe odraslih ljudi je M = 170.0 cm, a standardna devijacija je SD = 10.0 cm. Promatrana osoba X je visoka 180 cm. Koliki postatak ljudi je viši ili niži od ove osobe?

Rješenje:

1. izračunaj z: $z = \frac{X-M}{SD} = \frac{180-170}{10} = 1.0$
2. X pada na +1SD
3. prema grafu normalne distribucije 68,26% rezultata je unutar -1SD i +1SD
   od početka krivulje do +1SD je 50% + 34%=84%
   do kraja krivulje je 100% - 84% = 16%
4. 16% ispitanika je više od 180 cm

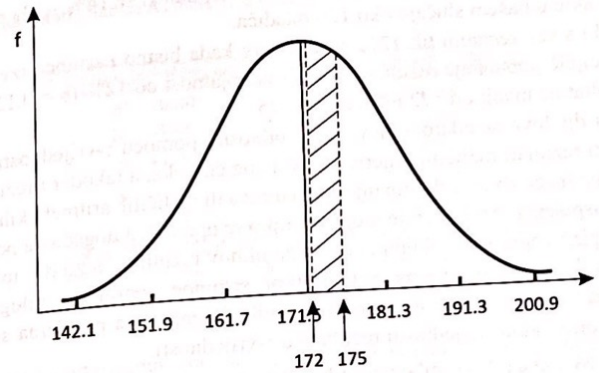


Broj rezultata iznad, ispod ili između neke vrijednosti

Problem: Aritmetička sredina visina grupe od 1000 odraslih ljudi je M = 171.5 cm, a standardna devijacija je SD = 9.8 cm. Koliko ima približno osoba koji su visoki između 172 i 175 cm?

Rješenje:

površina između 172 cm i 175 cm u grafu normalne distribucije



1. izračunaj z: $z_1 = \frac{172-171.5}{9.8} = 0.05, z_2 = \frac{175-171.5}{9.8} = 0.36$
2. granice su 0.05 SD i 0.36 SD
3. **iz tablica z-vrijednosti**:
   $z_1 = 0.05 \rightarrow p_1 = 0.4801 = 48\%$
   $z_2 = 0.36 \rightarrow p_2 = 0.3594 = 36\%$
4. između M i 0.05z se nalazi: 50% - 48.01% = 1.99% = 2% rezultata
   između M i 0.36z se nalazi: 50% - 35.94% = 14.06% = 14% rezultata
5. između 0.05z i 0.36z se nalazi: 14% - 2% = 12% rezultata = 120 osoba

Usporedba rezultata različitih ljudi ili različita mjerenja istog čovjeka

Problem: 100 sportaša se natječe u a) trčanju na 100m, b) skoku u vis, c) skoku u dalj, d) bacanju kugle i e) bacanju diska. Koji sportaš je bolji A ili B u svim disciplinama zajedno?

Rješenje:

M	SD	A	B	z_A	z_B
M_a = 12.8 s	SD_a = 2.0 s	12.2 s	13.0 s	+0.30	-0.10
M_b = 145.0 cm	SD_b = 21.1 cm	140 cm	136.5 cm	-0.24	-0.40
M_c = 485 cm	SD_c = 50 cm	580 cm	490 cm	+1.90	+0.10
M_d = 813 cm	SD_d = 103 cm	804 cm	920 cm	-0.09	+1.04
M_e = 2560 cm	SD_e = 400 cm	2400 cm	2980 cm	-0.40	+1.05
		skupna ocjena		+1.47	+1.69
		prosječna ocjena		+0.29	+0.34

Sportaš B je ukupno bolji od sportaša A (1.69>1.47)

Skupna ili prosječna ocjena iz niza mjerenja

s jednakim mjernim jedinicama, ali uz različit varijabilitet rezultata

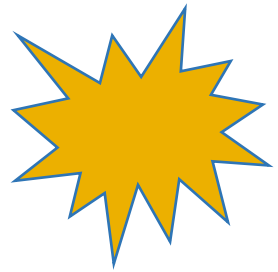
Problem: Skupina ljudi rješava dva psihomotorička testa (T1 i T2). Koji ispitanik je bolji?

Rješenje:

	Test T_1	Test T_2	
	min broj bodova = 0	min broj bodova = 0	
	max broj bodova = 15	max broj bodova = 120	
	$M_1 = 7.0$	$M_2 = 60.0$	
	$SD_1 = 1.0$	$SD_2 = 14.0$	
ispitanik A	9 bodova	74 boda	
	$z_1 = +2.0$	$z_2 = +1.0$	$z_A = z_1 + z_2 = +3.0$
ispitanik B	6 bodova	90 bodova	
	$z_1 = -1.0$	$z_2 = +2.14$	$z_B = z_1 + z_2 = +1.14$

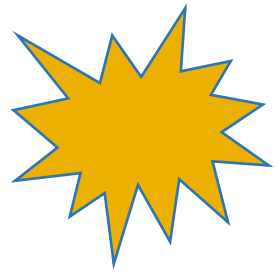
Ispitanik A je bolji od ispitanika B (3.0 > 1.14)

Inferencijalna statistika
Statistika zaključivanja



- iz uzorka nastoji stvoriti zaključak o populaciji

Uzorak i populacija



- Uzorak i populacija
- Frakcija uzorka
- Izbor slučajnog uzorka
- Vrste slučajnih uzoraka
- Vrste neslučajnih uzoraka
- Veličina uzorka
- Teorem centralne granice
- Standardna pogreška aritmetičke sredine
- Granice pouzdanosti

Uzorak i populacija

- **statistička jedinica** (element) – jedinica na kojoj se obavlja mjerenje (npr. osoba, grupa ljudi, razred, proizvod, tvornica…)
- **uzorak** – manji broj definiranih statističkih jedinica ili elemenata koji čine veću cjelinu (populaciju)
  - reprezentira narav populacije u svim važnim značajkama u pogledu mjerenja
- **populacija** = sve statističke jedinice

Vrste uzoraka

- **slučajni uzorak** – ima normalnu razdiobu (teoretski)
- **pristrani uzorak (biased sample)** – kada neki uzorak ima veću šansu da bude izabran
- **neslučajan uzorak** možemo koristiti, ali ne možemo računati pogrešku u odnosu na cijelu populaciju, koristi se zbog praktičnih razloga (npr. studenti, bolesnici, dobrovoljni sudionici istraživanja)

A ovo ćete pročitati zadnje

Ovo ćete pročitati prvo

Tada ćete pročitati ovo

Onda ovo

Pogrešan odabir slučajnog uzorka

- često nesvjesno preferiramo neke brojeve (npr. 3 ili 7) – pa svi brojevi između 1 i 1000 nemaju jednaku šansu
- uzeti imena s popisa koji letimično padnu pod oko – moguć utjecaj duljine ili poznatosti imena

Izbor slučajnog uzorka

- bacanjem **10-terostrane „kockice“** zapisuju se znamenke s „kockice“ (slučajni brojevi dobiveni iz populacije brojeva 0 do 9, vjerojatno pojavljivanja svakog $p = 0.1$)
- **„tablica slučajnih brojeva“** – otvoriti tablicu na bilo kojem mjestu – čitati redom brojeve (po recima, stupcima, dijagonalama) u skupinama znamenaka koliko nam veliki brojevi trebaju – ako je broj preveliki ili ako se ponovno pojavi broj kojeg smo već imali, preskače se
- pomoću **računala** (koristeći hi-kvadrat test se provjeri odstupa li frekvencija svakog pojedinog broja od teoretske frekvencije (p=0.1 za svaku znamenku))

7766	7520	1607	6048	2771	4733	8558	8681	5204	3806
9627	5293	3539	0457	4426	2857	3666	9156	6931	6157
4594	2563	6826	8102	2543	4032	3897	2012	0945	0709
6668	4104	4018	4544	8117	7664	5270	3014	0420	4232
8874	0822	0949	8697	7550	4154	9697	9045	4916	1235
8009	5708	7072	8045	8451	5777	1613	0399	2069	7909
7271	5633	6025	0745	9804	3333	7160	5150	7743	5221
6450	6850	0602	9518	2275	9221	6441	8899	4640	7742
0598	0564	9655	3988	5620	3286	6319	6392	5743	1111
6546	4417	4453	5125	1356	6011	5965	9253	1486	7503
5806	6217	4278	3170	1626	1746	9731	9289	7667	5209
6901	9464	6302	6404	8049	3653	8101	4498	8558	6238
3625	0749	5025	7327	3984	1635	5963	0970	7357	2033
2222	9942	1706	2907	6304	8022	7972	7852	6242	6269
7224	3014	3943	5982	4052	4243	5306	1530	7537	3233
7160	6043	0767	0230	6082	3637	4556	5564	8972	9697
7965	7435	3397	9741	6207	2297	6491	7961	0243	6897
6708	0600	2765	1911	0813	2268	3554	7976	4102	0414
4159	6804	3838	4255	9664	7044	3067	6720	7416	4748
6592	1846	2269	9136	7107	0676	9782	8016	2715	3932
2805	7999	3743	1655	7812	7223	0954	4397	7427	9120
9501	0400	8056	4148	5585	7497	7421	0640	6695	6127
3346	6596	1997	9417	0164	9718	5671	9765	7091	1920
4447	3427	6134	9130	4763	2301	2892	4251	4491	5772
0610	4363	0705	0969	4684	4202	5274	6660	0468	1814
2131	4792	1418	0080	9763	7306	0167	9688	6959	2250
9569	9416	5681	9632	8505	8948	6475	2934	6046	9640
1412	7690	5615	1776	8568	7209	9907	3541	8847	8752
5064	7408	1951	1033	7817	2626	2441	3795	3275	1319
4193	2082	0412	5519	4108	3333	5546	0177	9345	5260
6414	5111	4003	3695	2976	4939	7555	7374	2913	2705
2672	8618	7005	5736	0172	7472	2033	6308	8779	1270
0758	3869	9288	2397	6264	8352	8617	7869	2459	8591
4502	2535	2434	5018	1202	9081	2674	2467	2532	9689
4823	3965	2801	6179	8592	6763	6567	1016	5801	9288
3011	0939	7162	4443	3849	9142	2922	9191	6029	7631
6611	9238	2160	9339	8177	2180	3905	2977	9234	3434
0378	8311	0623	4299	2335	7044	5855	0186	5895	5642
9905	4972	6907	5633	6548	3412	8469	0559	8878	8671
9424	4750	8325	3871	1831	7268	1863	9963	1905	7484
7004	3469	1159	4841	8681	8751	9214	1145	4394	1160
5658	2963	5798	4691	8653	7427	7826	9971	2622	9886
9327	2129	3459	1165	1011	4805	1821	7999	2136	9308
1161	2217	1797	3906	5304	4087	6766	3063	1747	3836
6002	3340	3648	3765	1565	8483	6353	8232	4942	5721
4311	3087	1756	6612	3277	1269	6573	3096	0898	1103
5237	1667	5941	2504	6213	5797	9326	3079	8796	4220
0163	7150	0894	9009	7858	4812	7678	0835	8447	1524
0437	7497	0187	4907	2202	2318	5339	3290	4342	9375
0974	9130	4974	9757	8802	8514	6564	5485	0793	5675

3754	7829	9473	8264	8502	0364	5146	0609	4708	5229
9278	1828	8171	8788	3821	0923	8249	8431	6516	0911
9152	6396	7516	2959	4988	0943	6070	8342	5643	7476
0306	8452	1326	8892	2571	4860	1907	4843	0248	5283
1775	3205	8496	0201	6864	3375	0599	7516	8592	9823
4448	1897	3406	1429	8153	3408	1136	9173	9582	2866
3406	4332	0083	1214	5107	0912	8257	4015	5933	5520
4869	7491	5786	3633	9450	4572	6046	7844	2536	9502
5042	6524	1138	4001	6957	7220	8715	5082	8909	2384
0371	1656	8756	3369	3347	3534	0519	7230	2516	2674
2969	0056	8199	9383	4840	4135	7713	6317	4188	8073
4680	0551	7807	9470	9460	2253	0146	6082	9037	1862
1979	1845	0247	4813	2052	2758	6032	8288	6840	2677
3463	7252	3753	1178	2766	3207	2332	8262	8499	4501
0698	8601	2945	6077	3785	4647	4226	8959	9006	0964
2709	2447	0580	3375	1775	2038	3797	5163	7845	9397
6014	1671	2362	2315	8297	3930	6686	5835	9464	0916
7219	3355	3933	9312	3808	7879	6254	7075	7818	0295
6900	7276	4131	5402	3263	4026	5185	2862	8450	7749
0652	9020	6533	5737	6390	8723	8240	6442	4775	6040
3559	8683	0358	0118	0825	3360	7913	1403	4016	0202
1133	5094	3564	9818	0188	6367	2887	5038	1039	1658
1066	2065	4018	9132	3343	6165	1351	1312	7876	8452
8099	2678	7288	1970	9523	4070	7258	7276	3138	6818
5599	5836	0212	7112	8857	5894	6647	1660	3518	5780
6204	6540	1791	3190	3727	4500	5370	5231	8629	6291
8288	1891	5014	8442	9712	3435	4570	9493	1563	9165
7590	9691	1601	6615	0848	2885	1863	5682	1666	3398
7162	9599	9286	2819	2867	6533	9931	9217	4987	7722
9948	6283	0839	4175	8654	2005	6128	1306	6879	3152
5187	9791	4301	8481	5699	2522	0394	1538	8492	1812
5330	8112	2323	3056	1282	0543	4135	5819	6172	1017
6454	8783	7254	5267	9809	9964	9835	1111	5988	8017
8771	0872	6538	9975	4349	4106	6047	9630	4211	3234
1804	3896	2518	5665	8766	7161	0755	0886	3256	3198
8109	0020	3347	9221	6511	7593	6133	6123	2128	2735
9371	0132	4794	3110	5357	7242	4790	8002	9268	9733
6062	6416	7311	1167	5131	9955	9738	6038	1119	4832
7072	3929	8902	8062	6898	5499	5278	3407	0544	8772
5867	5384	8700	8017	5235	4094	9441	2381	8478	0981
1390	8293	7525	7188	8218	0131	3543	1679	8610	5737
4974	9904	7964	6038	0910	9364	4842	3873	3495	5511
9086	9898	1529	8544	7800	8523	1353	3312	5255	3096
8786	4498	5476	6266	9636	1897	3924	7298	3764	0906
7215	2019	6780	1005	4812	0787	8463	3784	6072	0940
2701	2584	8904	7799	9877	9015	0310	9330	0037	8215
9830	7090	3878	7553	7460	2845	9183	6429	9249	0246
0008	1130	3811	1862	1670	6389	9179	8571	7621	2169
5338	0351	6437	6148	5015	6174	5761	4690	0799	3291
6508	4163	0794	5801	1272	2814	0989	1130	3918	8596

Primjer – slučajni studenti

Problem: Želimo slučajno izabrati 350 studenata od njih 780 upisanih u prvu godinu. Svakog studenta označimo brojem.

Rješenje:

1. Žmireći, tablica se otvori na slučajnu stranicu (npr. bacanjem novčića) i vrhom olovke se slučajno izabere broj
2. Uzimaju se po 3 znamenke, npr. redom:

   7766 7520 1607 6048 2771 4733 8558 8681 5204 3806
3. 827, 855, 886, 815 otpadaju jer su preveliki brojevi
4. Ako se neki broj pojavi ponovno – ne uzimamo ga u obzir
5. Ponavljamo dok ne skupimo 350 studenata

Slučajni broj u računalu

- **Fizikalne metode**
  - Slučajni atomski ili subatomski fizikalni fenomeni (kvantna mehanika)
  - Radioaktivni raspad, termalni šum, šum u Zener diodama, driftanje sata, radio šum…
  - Mogu sadržavati asimetrije ili sustavna odstupanja
- **Računske metode**
  - PRNG (pseudorandom number generator) algoritmi – automatski kreira dugački niz brojeva s dobrim slučajnim svojstvima (ali nakon nekog vremena se ipak sekvence počnu ponavljati ili je memorija preopterećena)
  - Slučajan niz je određen fiksnim brojem „seed”
- **Metode temeljene na ljudima**
  - Skupljanjem različitih ulaza od korisnika i koristeći to kao izvor slučajnosti

Frakcija uzorka

- populacija ima ***N* statističkih jedinica** ili elemenata populacije
- **u uzorak slučajno biramo *n*** statističkih jedinica
- **frakcija uzorka**:

$$f = \frac{n}{N}$$

Primjer: populacija ima 5000 elemenata, u uzorak slučajno biramo njih 150, frakcija uzorka je f = 150/5000 = 0.03 = 3%

Vrste slučajnih uzoraka (1)

1. stratificirani ili slojeviti uzorak

- populacija se dijeli u **„potpopulacije“** ili „slojeve“ tj. **„stratume“ prema nekim karakteristikama** i **iz svake grupe** se uzme **slučajni uzorak**
- stratumi mogu biti prema godinama, spolu, socijalnom sastavu
- niti jedna statistička jedinica se ne smije nalaziti u više stratuma
- prednost: kada je **potrebno znati o svakom stratumu**, a ne samo o populaciji kao cjelini
- **veličina uzorka iz svakog stratuma**:
  - **proporcionalna** veličini grupe u cijeloj populaciji

- među 10000 ljudi ima 60% mladih, 30% srednjih, 10% starih
- uzorak se treba sastojati od 60% mladih, 30% srednjih i 10% starih među 1000 ljudi
- 600 + 300 + 100
- $f_{mladi} = \frac{600}{6000} = 0.1, f_{srednji} = \frac{300}{3000} = 0.1, f_{stari} = \frac{100}{1000} = 0.1, f_{opći} = \frac{1000}{10000} = 0.1$

Vrste slučajnih uzoraka (2)

- **neproporcionalan uzorak**
  - ako je neki **stratum jako malen –** frakcija uzorka se može povećati
  - **odnosi veličina pojedinih stratuma** = isti omjer kao što su produkti standardne devijacije i veličine uzorka u pojedinom stratumu

- $N_{stratumA} = 1000, \quad SD_{stratumA} = 5$
- $N_{stratumB} = 100, \quad SD_{stratumB} = 20$
- odnos umnoška veličine uzorka i SD: $\frac{1000 \cdot 5}{100 \cdot 20} = \frac{5}{2}$
- optimalni odnos veličina 5:2
- za N=70, 50 podataka iz prvog stratuma i 20 podataka iz drugog

Vrste slučajnih uzoraka (3)

- **aritmetička sredina** = ponderirana aritmetička sredina stratuma
  - ako je vrlo velik slučajni uzorak iz populacije - opasnost krivog računanja je manja

- $N_{stratumA} = 1000, \quad M_{stratumA} = 100$
- $N_{stratumB} = 100, \quad M_{stratumB} = 80$
- $M_{zaj} = \frac{M_1 N_1 + M_2 N_2}{N_1 + N_2} = \frac{100 \cdot 1000 + 80 \cdot 100}{1000 + 100} = 98.18$
- pogrešno: $M_{zaj} = \frac{100+80}{2} = 90$

Vrste slučajnih uzoraka (3)

2. klaster slučajni uzorak

- često u ekonomskim, političkim ili tržišnim istraživanjima
  - potrebno je skupiti mišljenja stanovnika grada
  - grad se podijeli na blokove,
  - pa se po slučaju odabere određeni broj tih blokova
  - intervjui sa **svim** stanovnicima odabranih blokova – **jednostupanjski klaster uzorak**
  - **dvostupanjski, trostupanjski, višestupanjski uzorci** – po slučaju izabrati samo neki stanovnici

3. sistematski slučajni uzorak

- prema popisu statističkih jedinica populacije
  - popis ljudi neke tvornice – koji je slučajno napravljen
  - odabere slučajnim izborom jedan,
  - a nakon toga se uzima svaki n-ti uzorak (npr. po abecedi)

Vrste neslučajnih uzoraka (1)

1. prigodni uzorak

- uzorak koji se „**nađe pri ruci**“, jer drugog nemamo: trenutno prisutni bolesnici, slučajni prolaznici na ulici, dobrovoljni sudionici – opasnost da su ekstremno **nereprezentativni**
- uzorci do kojih **najlakše dolazimo** – ako ne postoje suprotni dokazi, možemo ih koristiti kao slučajni uzorak

2. namjerni ili svrhoviti uzorak

- koji se uzima **radi određenog cilja ili svrhe**
- npr. kupci u trgovačkom centru o zadovoljstvu cijenama

3. modalni uzorak

- varijanta namjernog uzorka – **najčešći ili tipični slučajevi** („tipični“ stanovnik grada)

4. uzorak eksperata

- u uzorku su **stručnjaci u nekom području**

Vrste neslučajnih uzoraka (2)

5. kvota uzorak

- **proporcionalni kvota uzorak**
  - uzimaju se sudionici u anketiranju prema unaprijed utvrđenoj kvoti (npr. mišljenja građana o nekom pitanju) – reprezentirati glavne karakteristike populacije tako da se odabere proporcionalni dio svake od karakteristika
  - unaprijed se izabere broj ljudi svakog pojedinog stratuma koji će se intervjuirati – na cesti se slučajno odaberu ti ljudi
- **neproporcionalni kvota uzorak** – ne gleda se na proporcionalnu zastupljenost u svakoj od karakteristika populacije

6. uzorak heterogenosti

- kada se želi **uključiti sva različita mišljenja** ili pogledi pri anketiranju
- kada ne želimo prosjek, nego **utvrditi razlike**

7. uzorak „snježne grude“

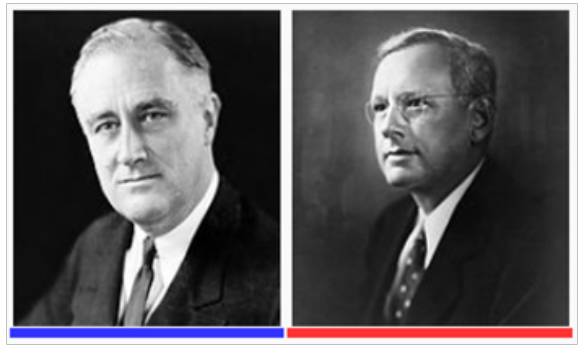
- članovi uzorka se prikupljaju na temelju **preporuke prethodnog člana** koji je bio uključen u uzorak

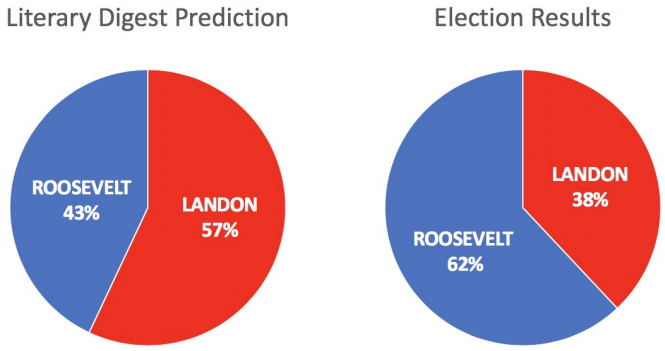
Veličina uzorka

- uzorak mora biti **reprezentativan** – nije toliko bitan broj
- ovisi o **varijabilnosti** pojave koju mjerimo (mala varijabilnost – potrebno malo uzoraka), **preciznosti** kojom želimo izmjeriti pojavu (želimo manju preciznost – manje uzoraka)
- **Weber metoda**: ako možemo grubo predvidjeti u kojem postotku je neko svojstvo zastupljeno u populaciji
  - veličina uzorka = taj postotak pomnožimo postotkom koji nedostaje do 100%
  - npr. 5% populacije posjeduje karakteristiku koju mjerimo – uzorak treba biti veličine 5×95=475
  - za 50% populacije – uzorak treba biti veličine 50x50=2500
  - veliki varijabilitet – veći broj uzoraka

Primjer - izbori

- **Istraživanje**: časopis Literary Digest je proveo istraživanje tko će pobijediti na izborima 1936: Landon ili Roosevelt
- **Anketirano 2.4 milijuna ljudi**
- **Rezultat ankete**: izgubio Roosevelt 43%
- **Izbori**: pobijedio Roosevelt 62%
- **Objašnjenje**: istraživanje je napravljeno telefonski. U to vrijeme samo bogati su mogli imati telefon.
- Pristran uzorak, neovisno o broju anketiranih
- Literary Digest je uskoro propao

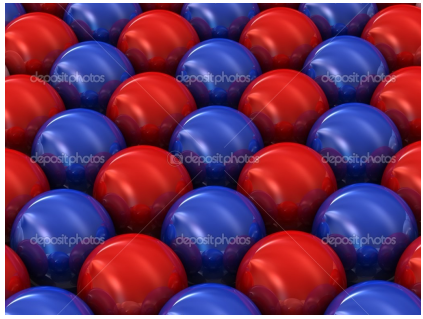




Uzorak i populacija

Populacija	Uzorak
prava aritmetička sredina μ **prava standardna devijacija σ**	**procjena prave aritmetičke sredine M** **procjena prave standardne devijacije SD**
veći varijabilitet rezultata u populaciji → bit će veći varijabilitet uzoraka uzetih iz te populacije	uzorak nije minijaturni duplikat populacije → pri uzimanju slučajnog uzorka dolazi do slučajnih varijacija
najviše aritmetičkih sredina uzoraka M će biti grupirano oko prave aritmetičke sredine populacije μ	veći uzorci N → raspršenje aritmetičkih sredina uzoraka M oko μ je manje (distribucija aritmetičkih sredina se približava normalnoj razdiobi)

Primjer 1 - kuglice

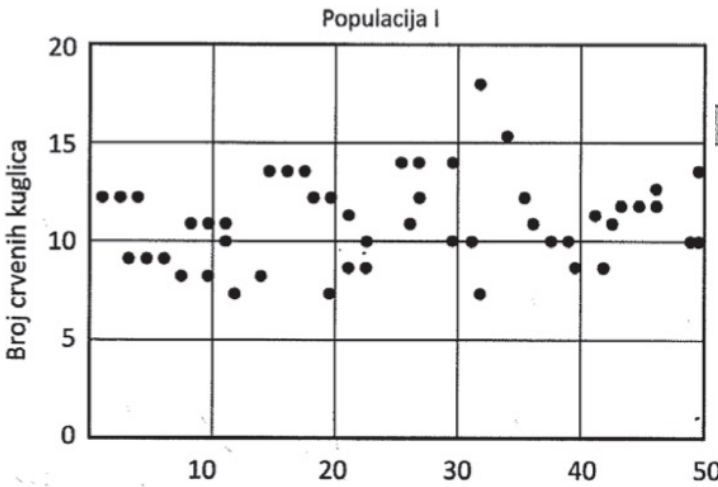


Populacija I: Kutija s nepoznatim brojem crvenih (C) i plavih (P) kuglica. Vadimo N=20 kuglica i brojimo crvene kuglice (uzorak). Vratimo kuglice u kutiju i ponavljamo 50 puta.

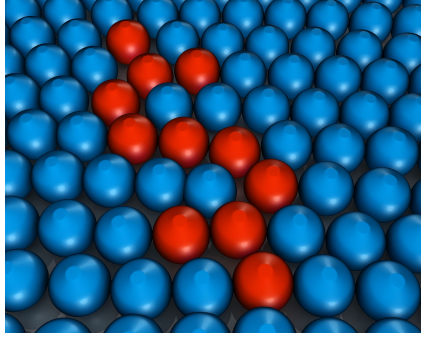
Redni broj uzorka	Broj crvenih kuglica
1	12
2	12
3	9
4	12
5	9
6	9
7	11
8	8
9	11
10	8
11	11
12	10
13	7
14	8
15	13
16	13
17	13

Redni broj uzorka	Broj crvenih kuglica
18	12
19	8
20	12
21	9
22	11
23	9
24	10
25	14
26	11
27	12
28	14
29	10
30	14
31	10
32	18
33	8
34	15

Redni broj uzorka	Broj crvenih kuglica
35	12
36	11
37	10
38	10
39	9
40	10
41	11
42	9
43	11
44	12
45	12
46	12
47	13
48	10
49	10
50	13
Ukupno	548



Broj crvenih kuglica u uzorku	f
manje od 7	0
7	1
8	5
9	7
10	9
11	8
12	10
13	5
14	3
15	1
16	0
17	0
18	1
više od 18	0
Ukupno	50



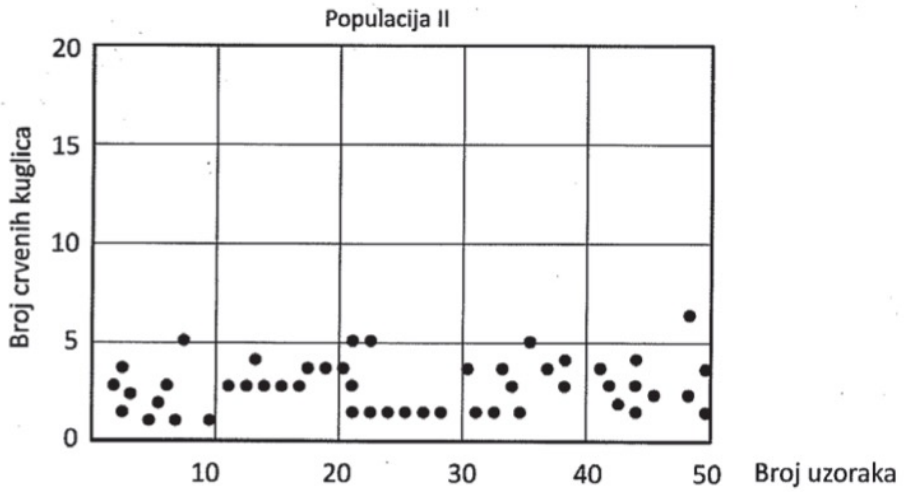
Populacija II: Promijenimo broj crvenih (C) i plavih (P) kuglica i ponovimo pokus: 50 uzoraka, veličine N=20.

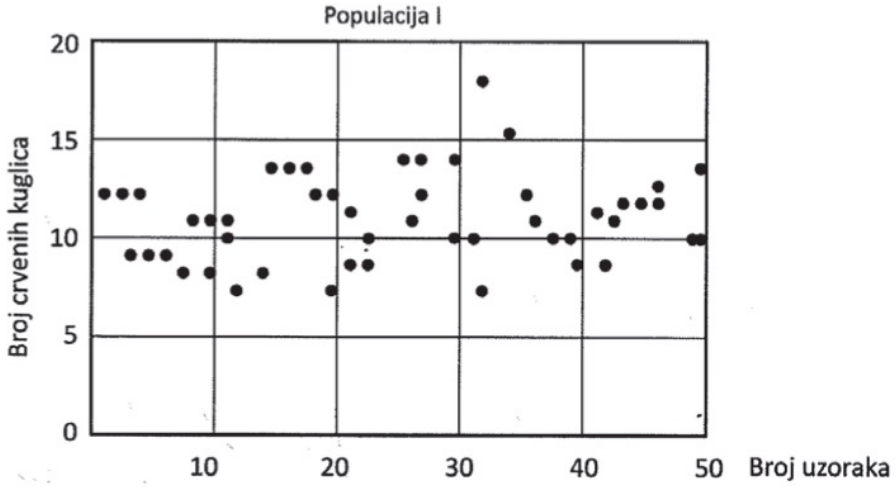
Redni broj uzorka	Broj crvenih kuglica
1	3
2	4
3	1
4	2
5	1
6	2
7	3
8	1
9	5
10	1
11	3
12	3
13	4
14	3
15	3
16	3
17	3

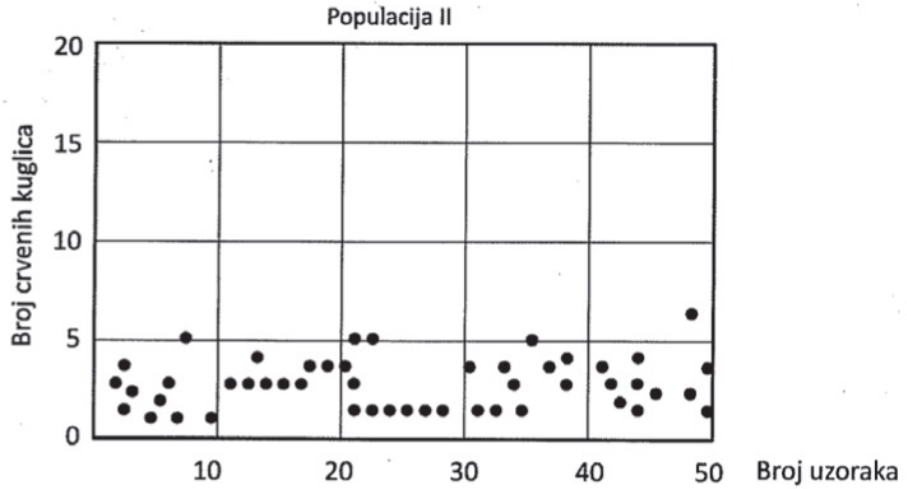
Redni broj uzorka	Broj crvenih kuglica
18	4
19	4
20	4
21	5
22	3
23	5
24	2
25	2
26	2
27	2
28	2
29	2
30	4
31	2
32	2
33	4
34	3

Redni broj uzorka	Broj crvenih kuglica
35	2
36	5
37	4
38	3
39	4
40	4
41	3
42	2
43	3
44	4
45	2
46	3
47	7
48	3
49	2
50	4
Ukupno	152

Broj crvenih kuglica u uzorku	f
0	0
1	4
2	14
3	15
4	12
5	4
6	0
7	1
više od 7	0
Ukupno	50







Uzorci u populaciji I se grupiraju oko **VEĆE vrijednosti** nego u populaciji II.

Gušće grupiranje uzoraka u populaciji II (**manje raspršenje**).

↓

sve kuglice crvene → nema varijabiliteta ni u populaciji, ni u uzorcima

više prevladava jedna boja kuglica → varijabilitet populacije i uzoraka manji

50% C i 50% P → najveći varijabilitet

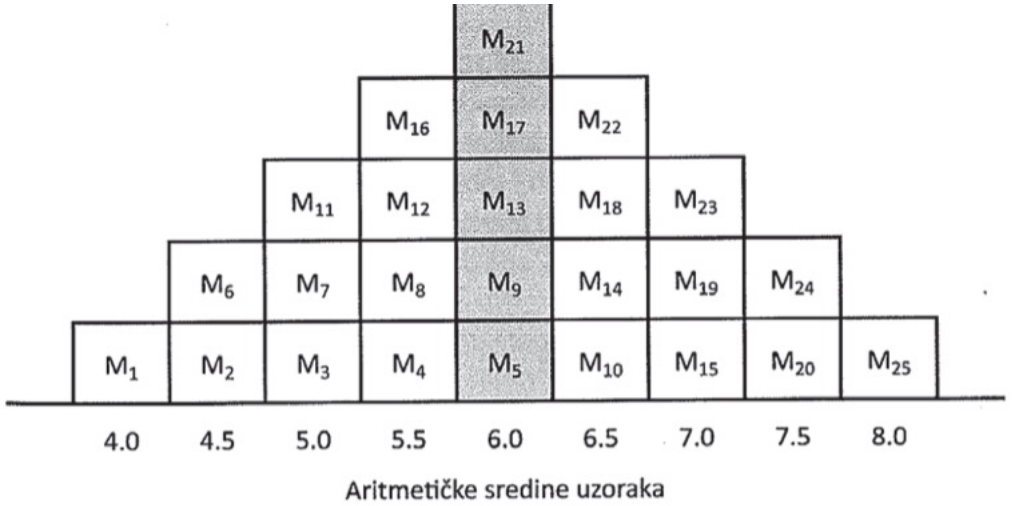
↓

Populacija I: odnos C i P približno 50% : 50%

Populacija II: puno manje C od P

Primjer 2 – pet rezultata

- Populacija od 5 rezultata: 4, 5, 6, 7 i 8
- Aritmetička sredina: $\mu = \frac{30}{5} = 6$
- Svi mogući uzorci veličine N=2 (uz povrat)



Svi mogući uzorci veličine $N = 2$	Aritmetičke sredine uzoraka	Simbol za aritmetičku sredinu
4, 4	4.0	M_1
4, 5	4.5	M_2
4, 6	5.0	M_3
4, 7	5.5	M_4
4, 8	6.0	M_5
5, 4	4.5	M_6
5, 5	5.0	M_7
5, 6	5.5	M_8
5, 7	6.0	M_9
5, 8	6.5	M_{10}
6, 4	5.0	M_{11}
6, 5	5.5	M_{12}
6, 6	6.0	M_{13}
6, 7	6.5	M_{14}
6, 8	7.0	M_{15}
7, 4	5.5	M_{16}
7, 5	6.0	M_{17}
7, 6	6.5	M_{18}
7, 7	7.0	M_{19}
7, 8	7.5	M_{20}
8, 4	6.0	M_{21}
8, 5	6.5	M_{22}
8, 6	7.0	M_{23}
8, 7	7.5	M_{24}
8, 8	8.0	M_{25}

Teorem centralne granice

- za određenu populaciju s aritmetičkom sredinom μ i varijancom σ^2, **distribucija aritmetičkih sredina uzoraka** imat će:
  - **aritmetičku sredinu** $M_M = \mu$,
  - **varijancu** $SD^2 = \frac{\sigma^2}{N}$
  - **standardnu devijaciju** $SD_M = \frac{\sigma}{\sqrt{N}}$.
- distribucija aritmetičkih sredina uzoraka približava se normalnoj razdiobi kako N uzorka raste

- Pogreška koja se veže uz procjenu prave aritmetičke sredine na temelju aritmetičkih sredina uzorka:
  - veća ako je pojava koju mjerimo više varijabilna (ne možemo utjecati)
  - veća ako je uzorak manji
- **Pogrešku možemo smanjiti – povećanjem uzorka**

- **Standardna pogreška aritmetičke sredine** (procjena pogreške aritmetičke sredine)

$$SD_M = \frac{SD}{\sqrt{N}}$$

  - mjera variranja aritmetičkih sredina uzoraka M oko prave aritmetičke sredine populacije μ
  - SD je mjera variranja individualnih rezultata oko njihove aritmetičke sredine

- **Prava standardna pogreška** $\sigma_M = \frac{\sigma}{\sqrt{N}}$

Primjer – standardna pogreška aritmetičke sredine

- Ispitivanje je obavljeno na N=100 ispitanika. Aritmetička sredina M=90 i standardna devijacija SD=10.
- Standardna pogreška aritmetičke sredine (procjena pogreške aritmetičke sredine)

$$SD_M = \frac{10}{\sqrt{100}} = 1$$

- 68% je vjerojatno da aritmetička sredina M ne odstupa od prave aritmetičke sredine više od ±1 (1 puta SD_M)
- 95% je vjerojatno da ne odstupa više od ±2 (2 puta SD_M)
- 99.7% je vjerojatno da je prava aritmetička sredina unutar intervala ±3, tj. između 87 i 93

Granice pouzdanosti

- **Granice pouzdanosti (granice sigurnosti)** – interval dobiven iz standardne pogreške aritmetičke sredine

- 68%-tne granice sigurnosti su M±1 SD_M (između 89 i 91),
  95%-tne granice su M±2 SD_M (između 88 i 92),
  99.7%-tne granice su M±3 SD_M (između 87 i 93) → prava aritmetička sredina populacije je sa 99.7%-tnom sigurnosti između 87 i 93.

Primjer – četiri rezultata

- Populacija od 4 rezultata: 2, 10, 4 i 8
- Aritmetička sredina: $\mu = \frac{24}{4} = 6$
- Standardna devijacija: $\sigma = \sqrt{\frac{\sum(X-M)^2}{N}} = \sqrt{10}$
  (N u nazivniku jer se radi o populaciji)
- Svi mogući uzorci veličine N=2 (uz povrat)

Originalna populacija			Uzorci veličine $N = 2$	Populacija aritmetičkih sredina			Populacija varijanci
1	2	3	4	5	6	7	8
X	$X - \mu$	$(X - \mu)^2$		M	$M - \mu_M$	$(M - \mu_M)^2$	SD^2
2	-4	16	2, 2	2	-4	16	0
10	4	16	2, 10	6	0	0	32
4	-2	4	2, 4	3	-3	9	2
8	2	4	2, 8	5	-1	1	18
$\sum = 24$		$\sum = 40$	10, 2	6	0	0	32
$\mu = 24/4 = 6$			10, 10	10	4	16	0
$\sigma^2 = 40/4 = 10$			10, 4	7	1	1	18
			10, 8	9	3	9	2
			4, 2	3	-3	9	2
			4, 10	7	1	1	18
			4, 4	4	-2	4	0
			4, 8	6	0	0	8
			8, 2	5	-1	1	18
			8, 10	9	3	9	2
			8, 4	6	0	0	8
			8, 8	8	2	4	0
				$\sum = 96$	$\sum = 0$	$\sum = 80$	$\sum = 160$

$$\mu = \frac{96}{16} = 6$$

$$\mu_{SD2} = 160/16 = 10$$

$$\sigma^2_M = 80/16 = 5$$

Zaključak

1. **Aritmetička sredina svih mogućih aritmetičkih sredina uzoraka** iste veličine jednaka je **pravoj aritmetičkoj sredini**, tj. aritmetičkoj sredini populacije. (stupac 5)
2. **Varijanca populacije aritmetičkih sredina uzoraka** jednaka je **varijanci originalne populacije, podijeljenoj veličinom uzorka**. (stupci 6 i 7)

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

3. **Varijance uzoraka** čine takvu raspodjelu oko prave varijance da im **aritmetička sredina** odgovara **pravoj varijanci**

$$\mu_{SD^2} = \sigma^2$$

(zato u nazivniku računanja varijanci uzoraka treba biti N-1, a populacije N → da se ovi rezultati poklapaju)

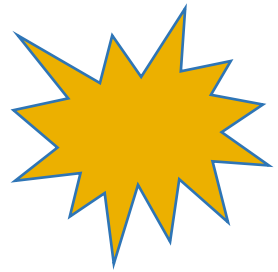
Varijanca uzorka

$$SD^2 = \frac{\sum(X - M)^2}{N - 1}$$

Varijanca populacije

$$\sigma^2 = \frac{\sum(X - M)^2}{N}$$

Razlika između aritmetičkih sredina velikih nezavisnih uzoraka



- Standardna pogreška razlike
- t-test
- Razina značajnosti
- Nul-hipoteza

Primjer 1: rezultati testa iz fizike (1)

N_1 = 900 studenata računarstva M_1 = 120 bodova SD_1 = 10 bodova	N_2 = 865 studenata elektrotehnike M_2 = 123 boda SD_2 = 11 bodova

- Razlika $M_2 – M_1$ = 3 boda
- Je li razlika statistički značajna?
- Promatramo 2 uzorka (a ne cijele populacije studenata elektrotehnike i računarstva)

Standardna pogreška razlike

Standardna pogreška razlike između dviju aritmetičkih sredina kod velikih uzoraka (oko 100 i više uzoraka, jednake ili slične veličine uzoraka):

$$SD_{M_1-M_2} = \sqrt{SD_{M_1}^2 + SD_{M_2}^2} = \sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}}$$

Primjer 1: rezultati testa iz fizike (2)

N_1 = 900 studenata računarstva
M_1 = 120 bodova
SD_1 = 10 bodova

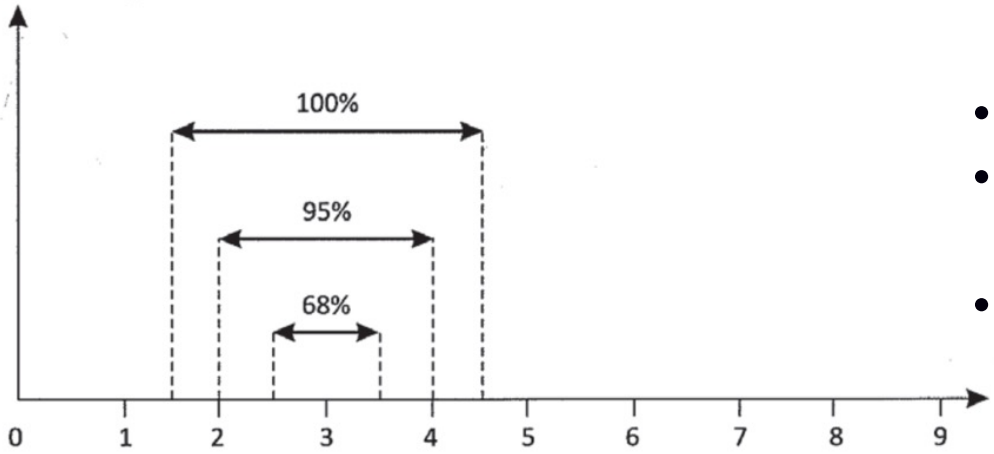
$$SD_{M1} = \frac{SD_1}{\sqrt{N_1}} = \frac{10}{\sqrt{900}} = 0.33$$

N_2 = 865 studenata elektrotehnike
M_2 = 123 boda
SD_2 = 11 bodova

$$SD_{M2} = \frac{SD_2}{\sqrt{N_2}} = \frac{11}{\sqrt{865}} = 0.37$$

$M_2 - M_1$ = 3 boda

$$SD_{M_1-M_2} = \sqrt{0.33^2 + 0.37^2} = 0.4958 \approx 0.5$$



- prava razlika je > 0 → razlika je statistički značajna
- praktički se ne može dogoditi da je razlika među pravim aritmetičkim sredinama = 0
- uzorci studenata E i R nisu iz iste populacije → E bolji na tom testu

Primjer 2: rezultati testa iz matematike (1)

N_1 = 36 studenata računarstva
M_1 = 120 bodova
SD_1 = 20 bodova

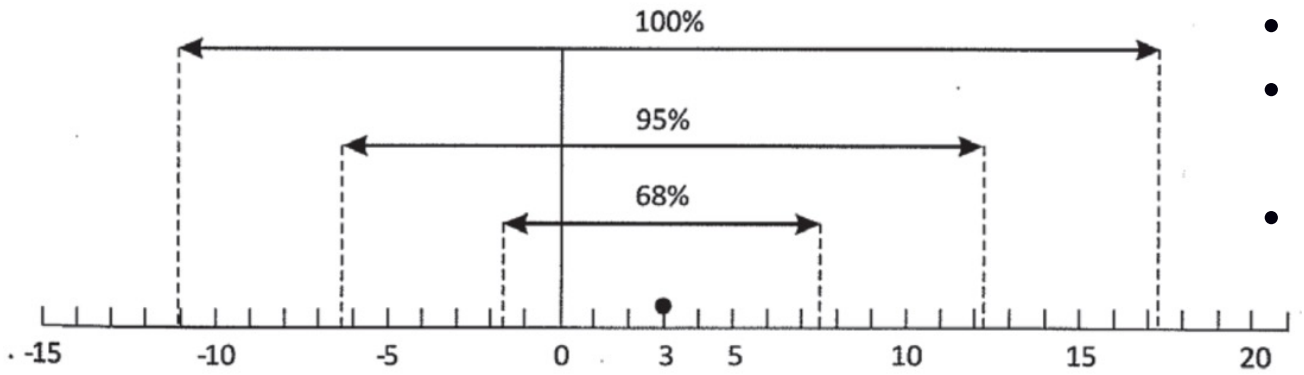
$$SD_{M1} = \frac{SD_1}{\sqrt{N_1}} = \frac{20}{\sqrt{36}} = 3.33$$

N_2 = 36 studenata elektrotehnike
M_2 = 123 boda
SD_2 = 20 bodova

$$SD_{M2} = \frac{SD_2}{\sqrt{N_2}} = \frac{20}{\sqrt{36}} = 3.33$$

Standardna pogreška razlike:

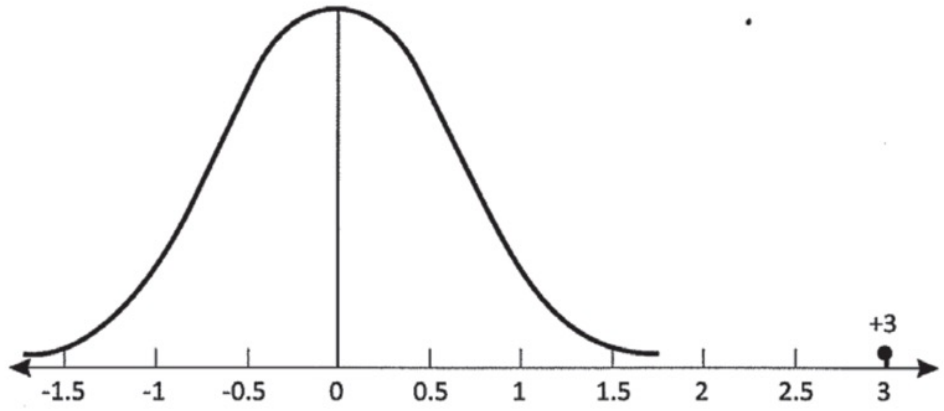
$$SD_{M_1-M_2} = \sqrt{3.33^2 + 3.33^2} = 4.71$$



- već 68%-tne granice pouzdanosti zahvaćaju 0
- možda među populacijama nema razlike ili je možda čak i prva skupina bolja od druge
- ovu razliku ne možemo smatrati statistički značajnom

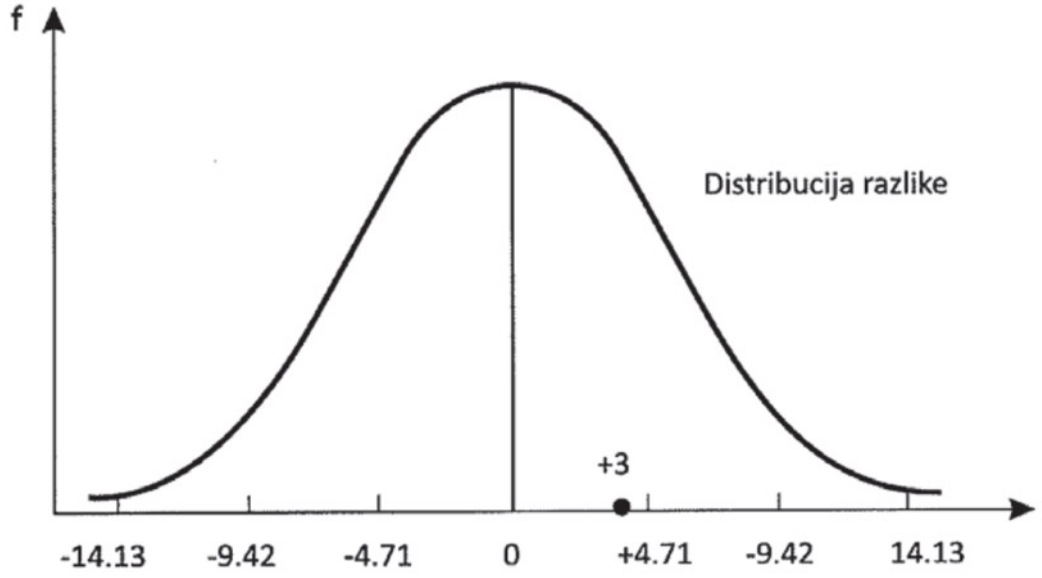
Primjer 1: rezultati testa iz fizike (3)

- standardna devijacija razlika aritmetičkih sredina = 0.5
- kada ne bi bilo razlike među populacijama:
  - iz tih populacija vadimo slučajne uzorke N_1=900 i N_2=865 (mnogo izvlačenja) → raspodjela tih mnogo razlika bi bila **normalna → aritmetička sredina razlika bi bila 0**, **standardna devijacija razlika bi bila 0.5**
  - **slučajno** možemo dobiti razlike kod uzoraka u rasponu **±1.5** (3 · SD=3 · 0.5)
  - **teško** ćemo slučajno dobiti razliku M2 – M1 = **3 boda** (3 se nalazi na 6SD od aritmetičke sredine) → zato **razlika 3 nije slučajna, već statistički značajna**



Primjer 2: rezultati testa iz matematike (2)

- standardna devijacija razlika aritmetičkih sredina = 4.71
- razlika M2 – M1 = **3 boda je unutar distribucije razlika** → moguće da je dobivena slučajno
- slučajno možemo dobiti i razlike koje su veće od 14

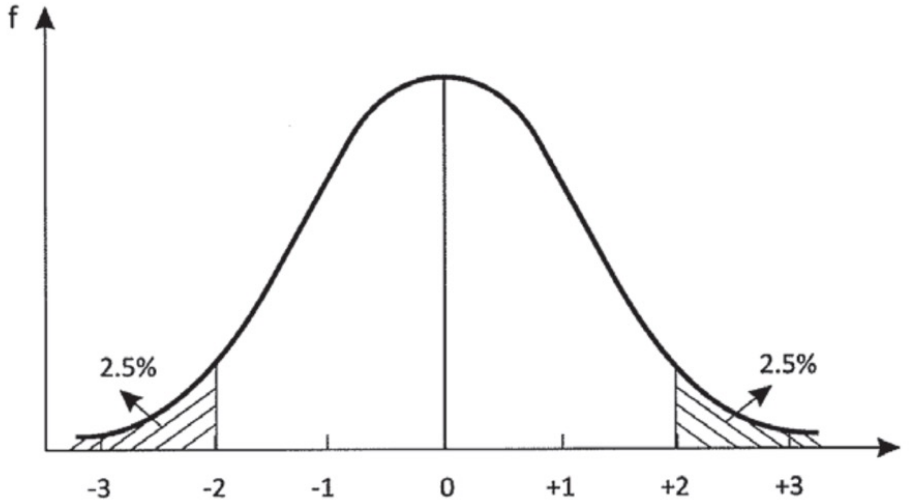


t-test (1)

- **Naivni zaključak**: ako je neka **razlika između dvije aritmetičke sredine → bar 3 puta veća** od svoje vlastite pogreške → onda ju možemo smatrati statistički značajnom → malo je vjerojatno da će se dogoditi slučajno
- **t-odnos (t-vrijednost)** – koliko je puta razlika veća od svoje pogreške

$$t = \frac{razlika}{standardna\ pogreška\ razlike} = \frac{M_1 - M_2}{SD_{M_1 - M_2}}$$

- **t=1** → nađena razlika među aritmetičkim sredinama je na **1SD** raspršenja svih slučajnih razlika koje se mogu dogoditi (pa makar nema razlike među aritmetičkim sredinama populacije)



t-test (2)

- **praktičan zaključak**: t=3 je prestrog kriterij → možemo proglasiti statistički neznačajnim razlike koju su zapravo značajne
- **za t>1.96** → razina značajnosti **manja od 5%** → šansa manja od 5% da smo pogriješili → obilježava se $\boldsymbol{p < 0.05}$
- **za t<1.96** → ako razlika padne u interval $\boldsymbol{0 \pm 1.96\, SD_{M_1 - M_2}}$ → 95% rezultata se nalazi u tom intervalu → ne smatramo ju značajnom → šansa za pogreškom je veća od odabranog kriterija
- iz z-tablica: za z=1.96 → p = 0.0025 = 2.5% za jednu stranu krivulje → oko 5% za obje strane krivulje

t-test (3)

- **t-test** – koliko je puta razlika veća od svoje pogreške

$$t = \frac{razlika}{standardna\ pogreška\ razlike} = \frac{M_1 - M_2}{SD_{M_1 - M_2}}$$

- **t > 1.96 → $p < 0.05$** → uz rizik 5% → **razlika JE statistički značajna** (tj. nije slučajna)
- **t > 2.58 → $p < 0.01$** → uz rizik 1% → **razlika JE statistički značajna** (tj. nije slučajna)

Koju razinu značajnosti uzeti?

- **ovisi o važnosti posljedica ukoliko se pogrešno zaključi – treba biti oprezan**
- utjecaj 2 lijeka: jedan jako opasan i isključen iz upotrebe → želimo da se drugi sigurno razlikuje od prvoga → trebamo veći stupanj sigurnosti: 1% ili 0.1%
- ispitivanje nuspojava kod sredstva za smirenje → bolje priznati da nuspojava postoji, nego ju previdjeti → proglasiti ju značajnom → manji stupanj sigurnosti: 5% ili 10%
- ako smo skloni prihvatiti i problematične dokaze o optuženikovoj krivnji – riskiramo da kaznimo mnogo nedužnih ljudi → ako ne uzimamo sve u obzir, već samo najjače dokaze o nečijoj krivnji – mnogi ljudi koji su krivi će ostati nekažnjeni
- da li nova organizacija rada ima efekta, jer je inače postupak preskup ili se ne isplati → stroži nivo, <1%
- da li nova organizacija rada ima efekta, a postupak nije ni skuplji, ni kompleksniji, ni opasniji → blaži nivo, >5%
- nova tehnika operacije tumora (ako sigurno bolesniku ne može nanijeti štetu, a ima veći postotak ozdravljenja) → možemo pristati na tu metodu čak i uz 20% ili 30%
- kod spašavanja života pristajemo na novu metodu i uz 1% vjerojatnosti da je bolja od stare metode

Nul-hipoteza

1. **nul-hipoteza = tvrdi da NE postoji razlika** – nema razlike među pojavama koje mjerimo

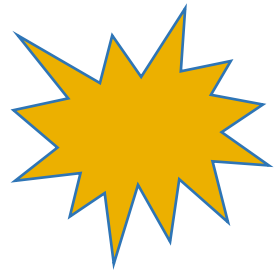
„među njima nema statistički značajne razlike“

2. **nul-hipoteza = svaka hipoteza koju želimo provjeriti** → vjerojatno se može napisati i u obliku prve definicije

„je li neka razlika između dvije aritmetičke sredine statistički značajna u smislu da je statistički značajno veća od npr. 10“ = „nema statistički značajne razlike između naše dobivene razlike i zamišljene razlike od 10“

Odluka	**Stvarno stanje u populaciji**	
	Nema razlike između dvije aritmetičke sredine	Postoji razlika između dvije aritmetičke sredine
Odbacujemo nul-hipotezu	Pogreška tipa I	Nema pogreške ako t-test pokaže da **razlika** među aritmetičkim sredinama **JE statistički značajna** (npr. $t > 1.96$)
Prihvaćamo nul-hipotezu	Nema pogreške ako t-test pokaže da **razlika** među aritmetičkim sredinama **NIJE statistički značajna** (npr. $t < 1.96$)	Pogreška tipa II

Značajnost razlike između jedne aritmetičke sredine i neke unaprijed fiksirane vrijednosti



- fiksirana vrijednost ne mora biti dobivena mjerenjem, nema SD

Primjer 1 – težina djece

- N = 144 djeteta, prosječna težina M=34.0 kg, SD = 4.8 kg
- Razlikuje li ta težina statistički značajno od normalne težine 32 kg?
- **Standardna pogreška aritmetičke sredine**:

$$SD_M = \frac{SD}{\sqrt{N}} = \frac{4.8}{\sqrt{144}} = 0.4$$

- 95% vjerojatnosti da M ne odstupa od prave aritmetičke sredine više ili manje od $1.96 \cdot 0.4 = 0.78$
- prava aritmetička sredina je u intervalu 34 ± 0.78, tj. 33.22 kg i 34.78 kg
- 33.22 kg ne zahvaća 32 kg
- M=34 kg se statistički značajno razlikuje od 32 kg

Primjer 2 - pilići

- Nova hrana za piliće koja je dosta skuplja od standardne hrane se isplati jedino ako bi pilići bili teži barem 500g od standardnih.

kontrolna skupina (pilići hranjeni standardnom hranom)	eksperimentalna skupina (pilići hranjeni novom hranom)
$N_1 = 400$ $M_1 = 2350$ g $SD_1 = 200$ g	$N_1 = 361$ $M_1 = 3040$ g $SD_1 = 220$ g

$$SD_{M_1-M_2} = \sqrt{SD_{M_1}^2 + SD_{M_2}^2} = \sqrt{\frac{SD_1^2}{N_1} + \frac{SD_2^2}{N_2}} = \sqrt{\frac{200^2}{400} + \frac{220^2}{361}} = 15.3$$

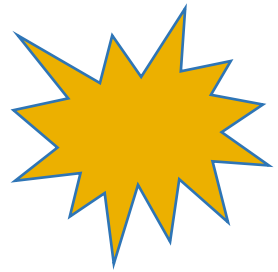
$$M_2 - M_1 = 690\text{g}$$

- Ne zanima nas razlikuje li se 690g statistički značajno od 0, **već od 500 g → t-test**:

$$t = \frac{razlika}{standardna\ pogreška\ razlike} = \frac{(M_2 - M_1) - 500}{15.3} = 12.4$$

t>1.96, čak i t>2.58 → posve smo sigurni da nova hrana uzrokuje dobitak na težini

Razlika između aritmetičkih sredina velikih zavisnih uzoraka



- kada su dvije varijable kojima tražimo razliku **u korelaciji** ($r_{1,2}$ koeficijent korelacije)

$$SD_{M_1-M_2} = \sqrt{SD_{M_1}^2 + SD_{M_2}^2 - 2r_{1,2}SD_{M_1}SD_{M_2}}$$

- korelaciju očekujemo kada **ista skupina ispitanika služi i kao kontrolna skupina** (često se naziva „metoda jedne grupe“)

Primjer - vježbanje

prosječna snaga stiska šake	posebno vježbanje i ponovno mjerenje prosječne snage stiska šake za tjedan dana
$N_1 = 64$ $M_1 = 45.0$ cm $SD_1 = 6.0$ cm $SD_{M_1} = 0.75$	$N_1 = 64$ $M_2 = 46.5$ cm $SD_1 = 5.0$ cm $SD_{M_2} = 0.63$

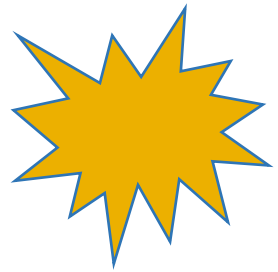
- Nakon vježbanja je veća prosječna snaga stiska: $M_2 - M_1 = 1.5$ cm
- uz $r = 0.6$:

$$SD_{M_1-M_2} = \sqrt{0.75^2 + 0.63^2 - 2 \cdot 0.6 \cdot 0.75 \cdot 0.63} = 0.63$$

$$t = \frac{razlika}{standardna\ pogreška\ razlike} = \frac{1.5}{0.63} = 2.38$$

- za p<0.05 → t>1.96 →razlika je statistički značajna
- (da nije uzet u obzir koeficijent korelacije, $SD_{M_1-M_2} = 0.98, t = 1.53$ → razlika nije statistički značajna)

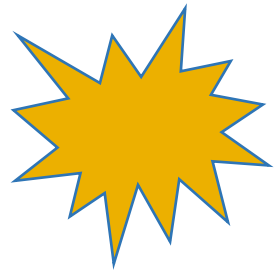
Samostalni rad



Razlika između aritmetičkih sredina malih nezavisnih uzoraka

Razlika između aritmetičkih sredina malih zavisnih uzoraka

Literatura



Boris Petz, Vladimir Kolesarić, Dragutin Ivanec. Petzova statistika. Osnovne statističke metode za nematematičare. Naklada Slap.