Grafičke i numeričke metode opisivanja podataka

Rosa Karlić

08.12.2023.

Zašto statistika?

Statistika - grana znanosti koja proučava na koji način je najbolje prikupljati, analizirati i donositi zaključke iz podataka

Statistička analiza - analiza karakteristika (varijabli) ispitanika od interesa

Cilj je opisati, istražiti, donijeti zaključke, predvidjeti i analizirati uzročno-posljedične veze

Zašto statistika?

Pomaže nam da istražimo različite varijable i njihove međusobne odnose i da utvrdimo da li su naša opažanja rezultat slučajnosti

Izvori varijacije: greške u mjerenju (tehnička varijabilnost), varijacije u populaciji (biološka varijabilnost)

Populacije i uzorci

- Populacija (od interesa) : cijela skupina ispitanika o kojima želimo nešto zaključiti
- Uzorak – podskup populacije
- **Parametar** – karakteristika populacije (npr. srednja vrijednost populacije, μ)
- **Statistika** – bilo koja funkcija ispitanika u nasumičnom uzorku (npr. srednja vrijednost uzorka, $\overline{x}$)
- Poželjno je odabrati **nasumičan uzorak** iz populacije kako u analizu ne bismo uvodili **pristranost**

Podaci

VARIJABLE

ISPITANICI

ID	STAROST	SPOL	TRETMAN	GEN1	GEN2	GEN3	STATUS
87	53	F	trt2	17.41	28.23	4.17	zdrav
119	53	M	ctrl	19.84	52.56	47.24	zdrav
67	52	M	trt2	19.19	27.49	12.07	zdrav
62	54	M	trt2	22.77	28.45	9.62	bolestan
131	55	F	ctrl	24.17	49.91	49.55	bolestan
50	54	F	trt1	17.15	15.32	10.67	zdrav
106	54	M	ctrl	17.92	44.95	51.39	zdrav
127	58	F	ctrl	20.06	53.19	49.71	bolestan
30	54	M	trt1	19.97	16.18	13.78	zdrav
72	54	F	trt2	25.44	27.58	11.81	zdrav

Tipovi varijabli

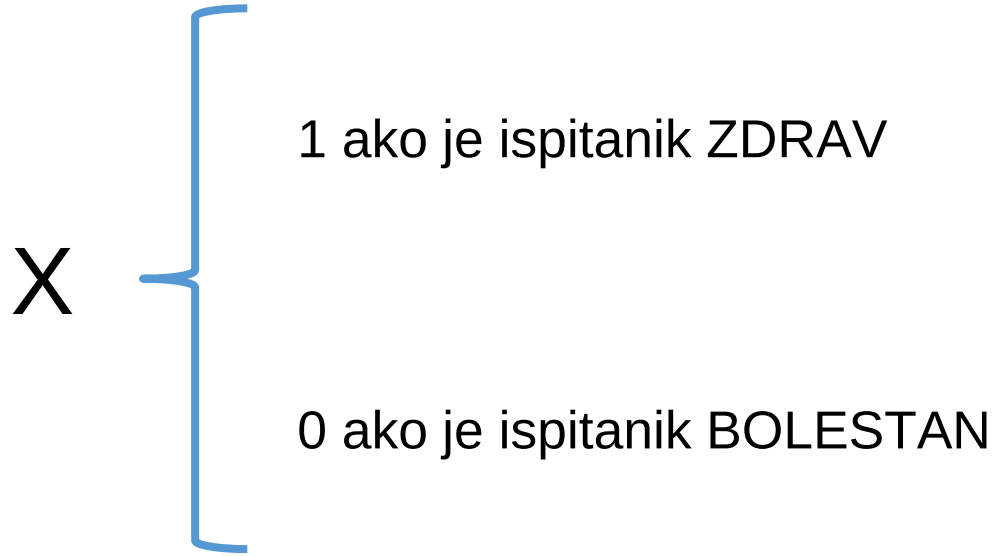
- Numeričke (kvantitativne) - diskretne i kontinuirane
- Kategoričke (kvalitativne) - nominalne i ordinalne

- Različite varijable mogu biti međusobno ovisne ili neovisne

ID	STAROST	SPOL	TRETMAN	GEN1	GEN2	GEN3	STATUS
87	53	F	trt2	17.41	28.23	4.17	zdrav
119	53	M	ctrl	19.84	52.56	47.24	zdrav
67	52	M	trt2	19.19	27.49	12.07	zdrav
62	54	M	trt2	22.77	28.45	9.62	bolestan
131	55	F	ctrl	24.17	49.91	49.55	bolestan
50	54	F	trt1	17.15	15.32	10.67	zdrav
106	54	M	ctrl	17.92	44.95	51.39	zdrav
127	58	F	ctrl	20.06	53.19	49.71	bolestan
30	54	M	trt1	19.97	16.18	13.78	zdrav
72	54	F	trt2	25.44	27.58	11.81	zdrav

Slučajna varijabla

- Numerička funkcija koja svakom ishodu eksperimenta pridružuje realan broj
- Npr.



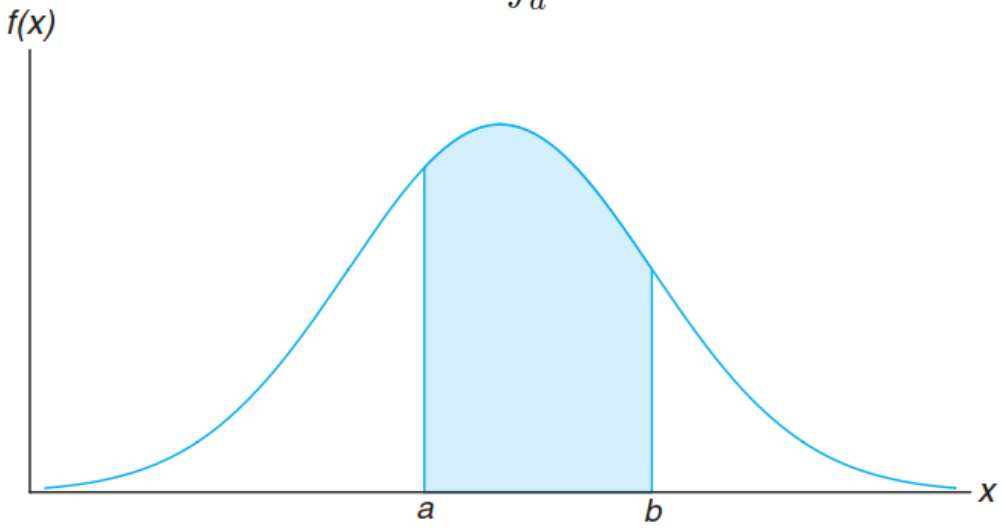
Slučajna varijabla

- Slučajne varijable mogu biti:
  - Diskretne
    - Starost pacijenata u godinama X = {0, 1, 2, 3, …, N}
    - Broj mutacija u DNA lancu X = {0, 1, 2, 3, …, N}
  - Kontinuirane
    - Visina pacijenta X = (0, M)
    - Tjelesna temperatura pacijenta X = (M, N)
- Opisujemo ih distribucijama vjerojatnosti – vjerojatnost da će slučajna varijabla poprimiti određenu vrijednost ili se nalaziti u određenom intervalu

Kontinuirane slučajne varijable

- Mogu poprimiti beskonačno mnogo vrijednosti
- Područje vrijednosti – interval na brojevnom pravcu ili cijeli brojevni pravac
- Određujemo vjerojatnost da će se vrijednost kontinuirane slučajne varijable nalaziti unutar nekog intervala (vjerojatnost da će poprimiti točno određenu vrijednost je 0)
- Opisuju se funkcijama gustoće vjerojatnosti

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

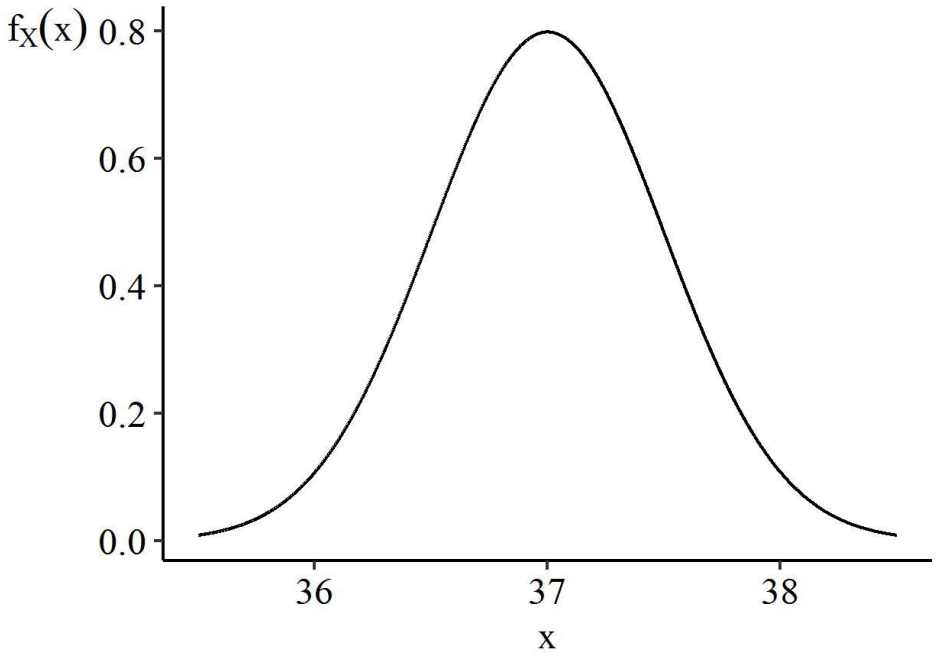


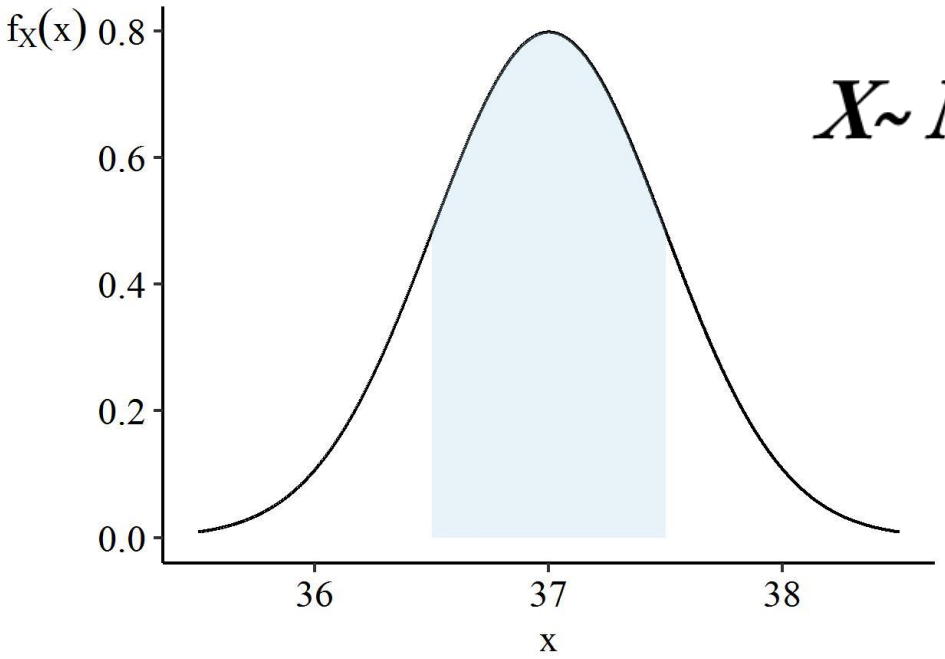
Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2012) *Probability and Statistics for Engineers and Scientists*

Kontinuirane slučajne varijable - primjer

- Tjelesna temperatura ispitanika
- μ (srednja vrijednost) and σ (standardna devijacija) određuju lokaciju i oblik distribucije

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$





$$X \sim N(\mu, \sigma^2)$$

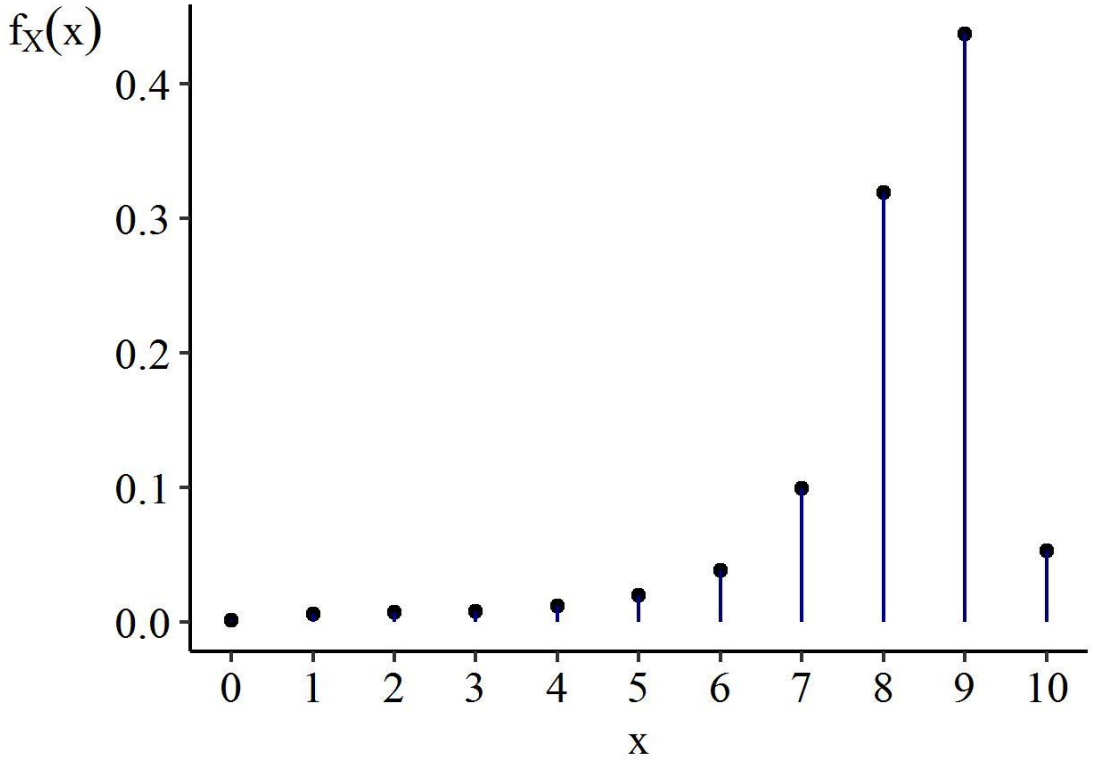
Diskretne slučajne varijable

- Mogu poprimiti prebrojivo mnogo diskretnih vrijednosti
- Svaka vrijednost ima konačnu vjerojatnost
- Opisuju se funkcijama mase vjerojatnosti

$$f_X(x) = P(X = x)$$

Diskretne slučajne varijable - primjer

- Apgar ocjena



$F_x(x)$	x
0.001	0
0.006	1
0.007	2
0.008	3
0.012	4
0.020	5
0.038	6
0.099	7
0.319	8
0.437	9
0.053	10

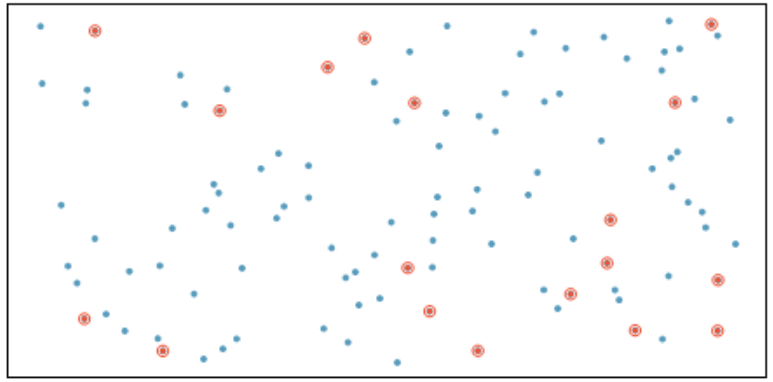
Tipovi pristranosti

- Poželjno je odabrati **nasumičan uzorak** iz populacije kako u analizu ne bismo uvodili **pristranost**

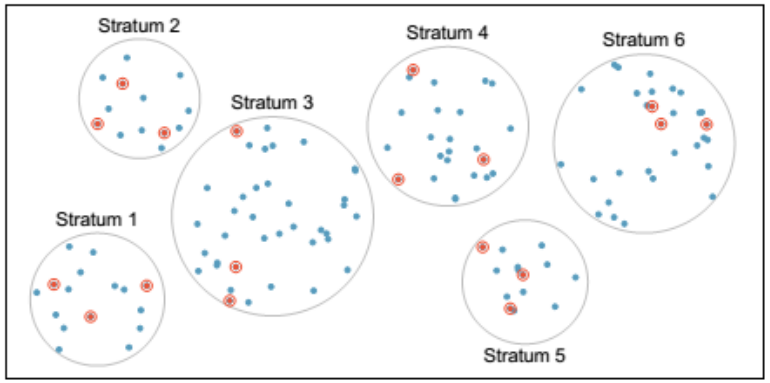
- **Uzorkovanje prema pogodnosti (praktično uzorkovanje) –** veća je vjerojatnost da će biti uključeni pojedinci do kojih se lako može doći
- **Neodziv –** samo (nenasumični) dio nasumično odabranih ljudi odgovori na anketu tako da uzorak više nije reprezentativan za populaciju
- **Dobrovoljni odgovor -** uzorak se sastoji od pojedinaca koji dobrovoljno odgovaraju jer imaju čvrsto mišljenje o problemu

Metode uzorkovanja

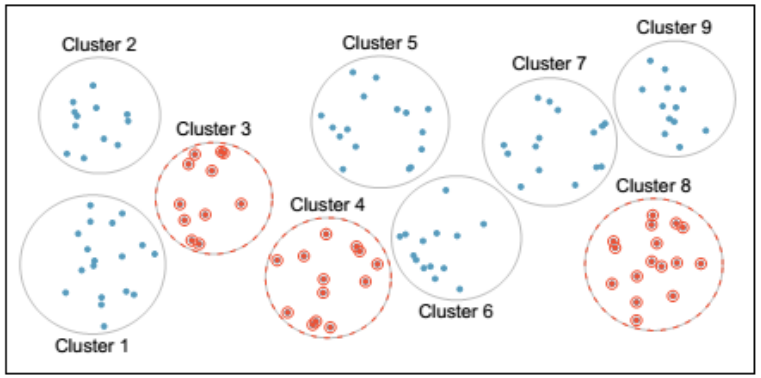
Jednostavno nasumično uzorkovanje



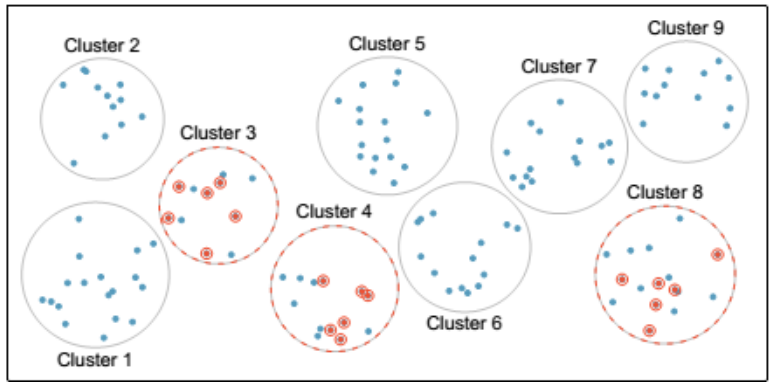
Stratificirano uzorkovanje



Klaster uzorkovanje



višestupanjsko uzorkovanje

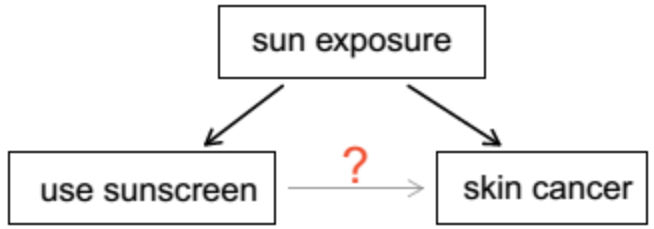


Primjer

- Pretpostavimo da nas zanima procjena stope malarije u gusto naseljenom tropskom dijelu ruralne Indonezije. Saznajemo da u tom dijelu indonezijske džungle ima 30 sela, jedno više-manje slično sljedećem. Cilj nam je testirati 150 osoba na malariju. Koju metodu uzorkovanja treba primijeniti?

Tipovi studija

- **Promatračka (observacijska) studija –** prikupljanje podataka na način koji izravno ne ometa način na koji podaci nastaju
  - Prospektivna i retrospektivna
  - Odredite povezanost između varijabli
- **Eksperiment**
  - Nasumično dodijelite subjekte tretmanima
  - Uspostavite uzročne veze
- **Nezavisne (prediktorske) i zavisne (ishodi varijable**



Korelacija ≠ Uzrok

Primjer

- Smoking Behaviors Among Cancer Survivors
  - Burke L et al., *J Oncol Pract*.

were older than18 years of age, and could read English. We randomly selected a pool of 1,000 patients diagnosed with or treated for cancer between January 1, 2003, and December 31, 2007 at the Mary Babb Randolph Cancer Center at West Virginia University Hospital. These patients were mailed a three-page (22-item) questionnaire with a cover letter explaining the study.

Data obtained from the questionnaire consisted of socio-demographic information, clinical and smoking history, and smoking cessation information. There were no personal identi-

- Weight Loss with a Low-Carbohydrate, Mediterranean, or Low-Fat Diet
  - Kappos L et al., *N Engl J Med*.

METHODS

In this 2-year trial, we randomly assigned 322 moderately obese subjects (mean age, 52 years; mean body-mass index [the weight in kilograms divided by the square of the height in meters], 31; male sex, 86%) to one of three diets: low-fat, restricted-calorie; Mediterranean, restricted-calorie; or low-carbohydrate, non–restricted-calorie.

Eksperimenti i observacijske studije

	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

Eksperiment - primjer

- Mjerimo odabrane karakteristike (značajke) eksperimentalnog sustava
- Subjekti (opažanja) – osnovni uzorak (pacijent, miš, stanična linija, ...)
- Neovisne varijable, čimbenici od interesa - atributi (genotip ili spol) ili eksperimentalni uvjeti (vrsta prehrane, temperatura, režim lijekova)
- Razine - različite vrijednosti varijabli (faktora)
- Liječenje - bilo koja kombinacija različitih razina varijabli
- Primjer: studija ekspresije gena u jetri zdravih pojedinaca naspram pacijenata s dijabetesom izloženih različitim tretmanima
  - Zavisna varijabla - mRNA transkripti
  - Ispitanici (opažanja) – pacijenti
  - Neovisne varijable (čimbenici) - status i tretman
  - Razine – zdravi/bolesni i ctrl/trt1/trt2
  - Liječenje – dijabetičar na trt1

Eksperimentalni dizajn

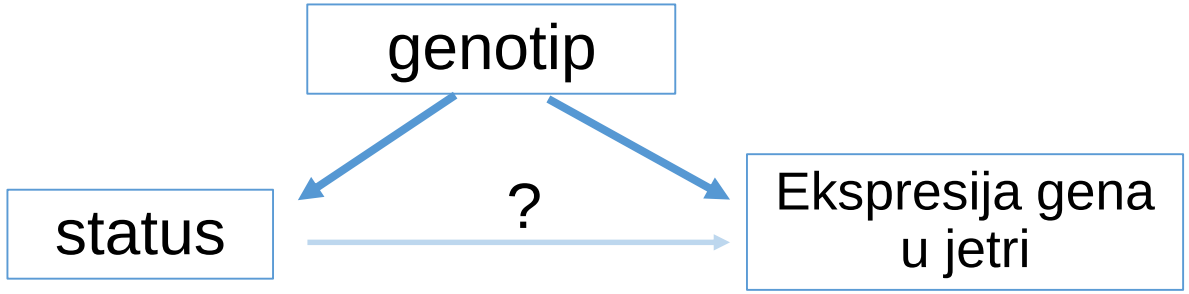
1. Kontrola – usporedite tretman od interesa s kontrolnom skupinom
2. Nasumično – nasumično dodijeli subjekte grupama
3. Ponavljanje – uzmite dovoljno velik uzorak ili ponovite cijelu studiju
4. Blokiraj – blokiraj varijable za koje se ili zna ili se sumnja da utječu na ishod eksperimenta

5. Placebo – lažni tretman
6. Slijepe i dvostruko slijepe studije

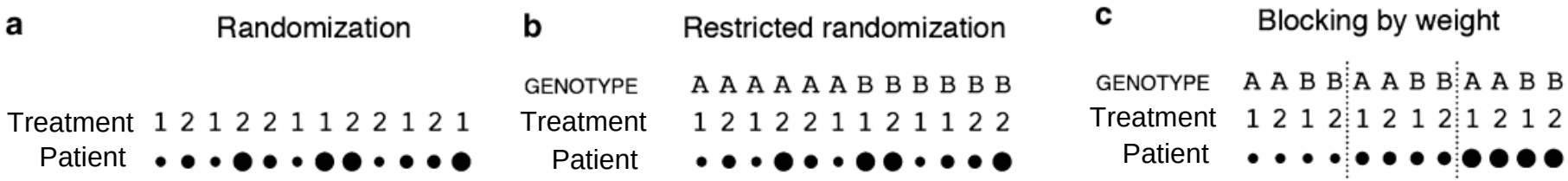
Zbunjujuće varijable

- **Nezavisne (prediktorske) i zavisne (ishodne) varijable**
- **Zbunjujuće varijable – u korelaciji sa zavisnom i nezavisnom varijablom**
- Važno je izbjeći brkanje između interesnih čimbenika (razmislite o dizajnu svog eksperimenta)
- Dobra praksa za izradu studija u kojima se čimbenici od interesa ne miješaju s čimbenicima koji nisu od interesa (na primjer, laboratorijski učinci) – randomizacija!

Korelacija ≠ Uzrok

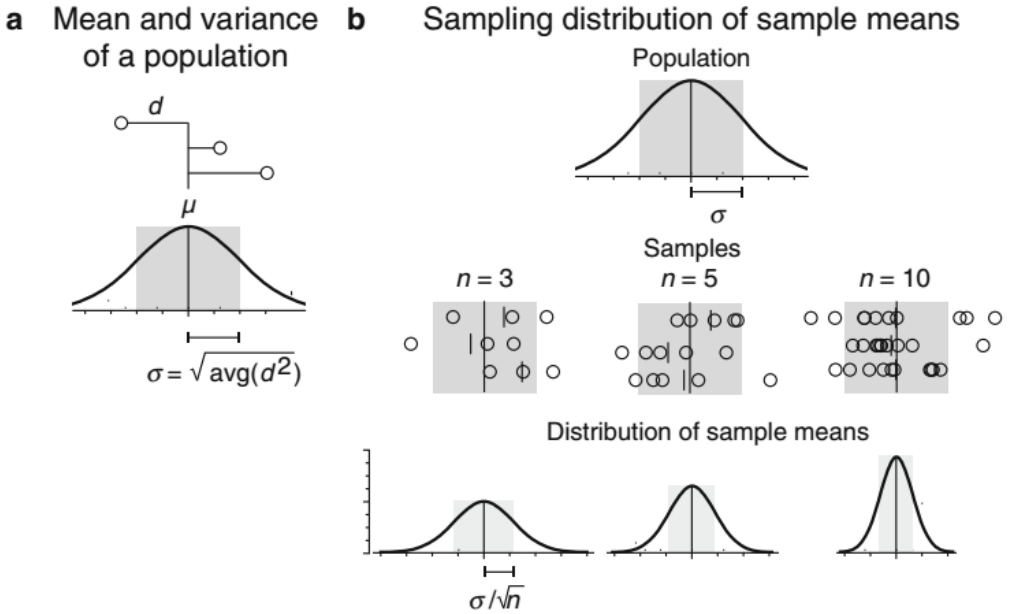


Randomizacija i blokiranje



- Randomizacija – nasumično uzorkujte eksperimentalne jedinice i nasumično ih dodijelite tretmanima
- Blokiranje – pokušavamo isključiti varijaciju u podacima koja nije posljedica tretmana

Distribucija procjena statistike



Honaas *et al.*, 2016, Statistical Genomics: Methods and Protocols

- Distribucija procjena statistike – skup svih mogućih vrijednosti uzorka sažetaka
- Ako ne ponovimo eksperiment, možemo promatrati samo jednu vrijednost iz distribucije uzorka

Replikacija

- Replikacija daje i precizniju procjenu parametara populacije i procjenu varijabilnosti
- Biološki i tehnički izvori varijabilnosti
- Biološki replikati - paralelna mjerenja biološki različitih uzoraka
- Tehničke replike - ponovljena mjerenja istog uzorka
- Naš primjer:
  - Biološki replikati - više pacijenata s istim genotipom na istom tretmanu
  - Tehničke replike - više uzoraka iz jetre istog pacijenta
- Povećanje broja bioloških ponavljanja obično je bolje od povećanja broja tehničkih ponavljanja

Primjer

- Studija ekspresije gena u jetri zdravih pojedinaca naspram pacijenata s dijabetesom izloženih različitim tretmanima

- Što su zavisne i nezavisne varijable u našoj studiji?
- Koja je ovo vrsta studije?
- Može li se ova studija koristiti za zaključivanje uzročnosti?

Deskriptivna i inferencijalna statistika

- **Opisno: s obzirom na uzorak, što možemo reći o uzorku?**
  - opisivanje podataka korištenjem numeričkih sažetaka (kao što su srednje vrijednosti, frekvencije itd.) i grafičkih sažetaka (kao što su histogrami, stupčasti grafikoni itd.)
- **Zaključak: s obzirom na uzorak, što možemo reći o populaciji iz koje je izvučen?**
  - korištenje informacija o uzorku za donošenje zaključaka o većoj skupini predmeta/pojedinaca (populaciji) nego samo o onima u uzorku. Inferencijalna statistika koristi se za izvođenje zaključaka o populaciji iz uzorka.

Numeričke opisne metode

- Učestalosti (brojevi)
- Relativne frekvencije (postoci, proporcije)
- Kumulativne frekvencije
- Kumulativne relativne frekvencije
- Unakrsne (kontingencijske) tablice

Height	Number of students	Cumulative frequency
160-170	5	5
170-180	10	15
180-190	7	22
190-200	1	23

	Male	Female
Left-handed	2	1
Right-handed	7	8

Numeričke opisne metode

Mjere lokacije

Mod - najčešća observacija u skupu.

Srednja vrijednost (prosjek) - zbroj opažanja podijeljen s brojem tih opažanja.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Medijan skupa opažanja, poredanih od najmanjeg prema najvećem, je vrijednost takva da je najmanje polovica opažanja manja ili jednaka toj vrijednosti.

Kvartili i kvantili - k-ti kvantil skupa vrijednosti ih dijeli tako da k% vrijednosti leži ispod, a (100-k)% vrijednosti leži iznad.

Donji kvartil, medijan, gornji kvartil

Numeričke opisne metode

Mjere rasapa (raspršenosti)

Raspon = Maksimum - Minimum

IQR = 75. kvantil - 25. kvantil, mjeri širenje srednjih 50% podataka

Standardna devijacija je mjera raspršenosti opservacija od srednje vrijednosti.

Standardna devijacija populacije

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

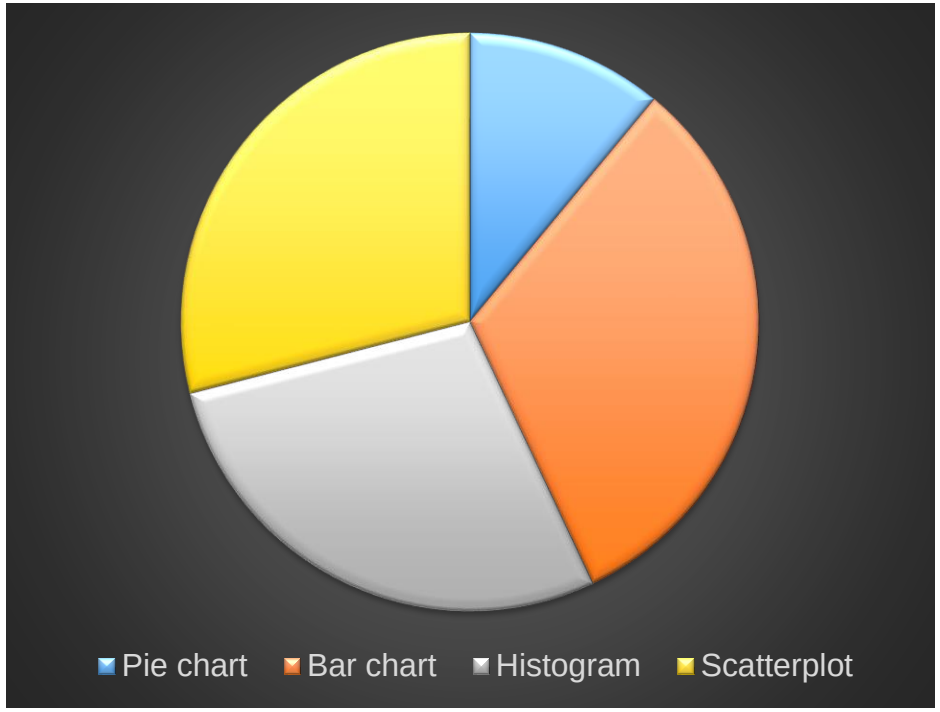
Standardna devijacija uzorka

$$\bar{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Kvadrat standardne devijacije - varijanca

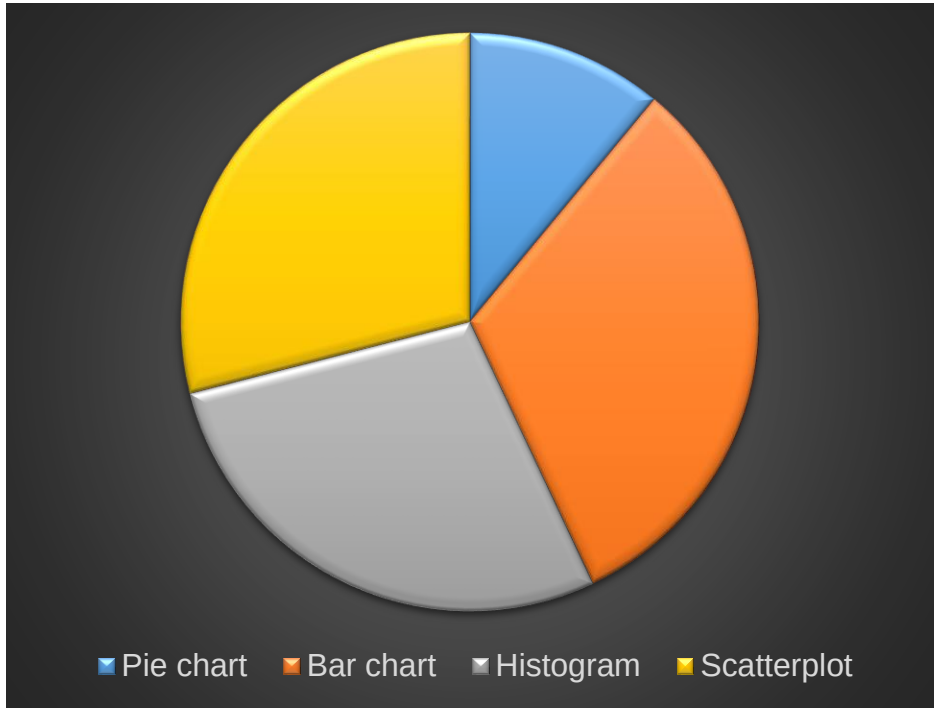
Grafičke metode

Kružni dijagram

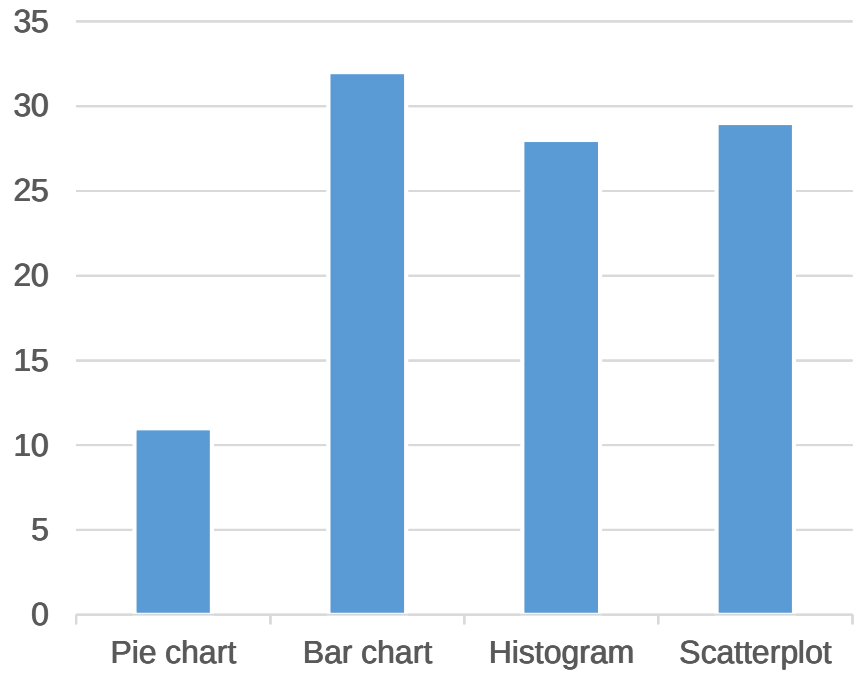


Grafičke metode

Kružni dijagram



Stupčasti dijagram



Distribucija učestalosti

- Distribucija učestalosti - sažeti sažetak podataka
- Podijelite raspon podataka u intervale (intervale klasa, ćelije ili spremnike).
- Ako je moguće, intervali trebaju biti jednake širine
- Broj spremnika ovisi o broju opažanja i količini raspršenosti ili disperzije u podacima
- Odabir broja intervala približno jednak korijenu broja opažanja često dobro funkcionira u praksi.
- Relativne frekvencije nalaze se dijeljenjem promatrane frekvencije u svakom intervalu s ukupnim brojem promatranja.

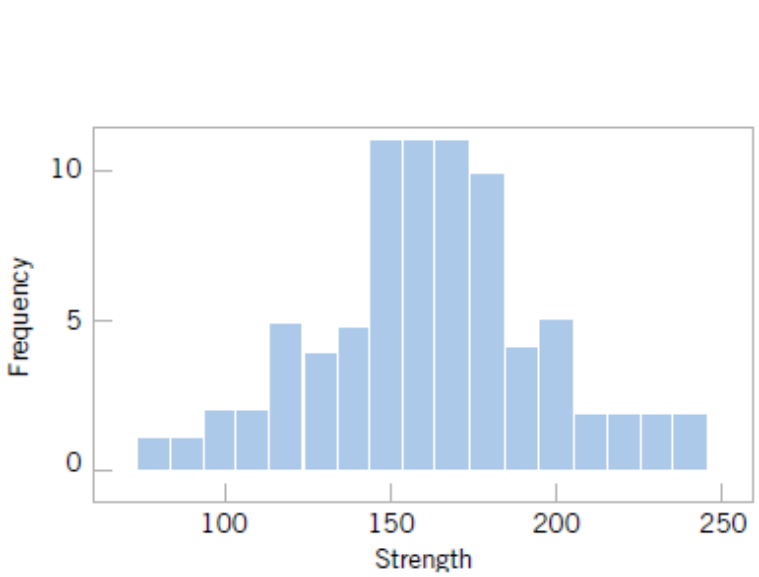
Histogram

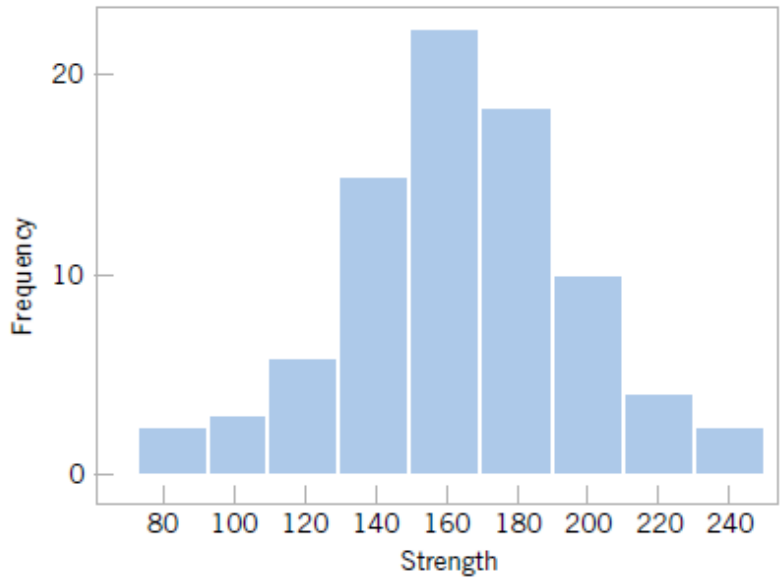
- Vizualni prikaz distribucije učestalosti

  (1) Označite granice intervala na vodoravnoj ljestvici.

  (2) Označite i označite vertikalnu ljestvicu frekvencijama ili relativnim frekvencijama.

  (3) Iznad svakog polja nacrtajte pravokutnik gdje je visina jednaka frekvenciji (ili relativnoj frekvenciji) koja odgovara tom intervalu.





Histogram

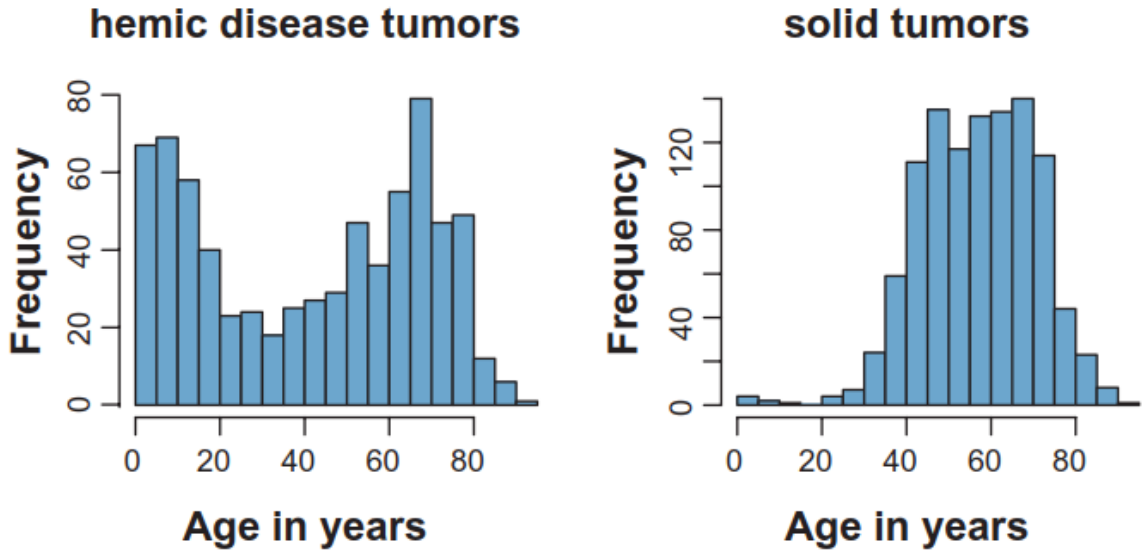


Figure 1. Histogram of the age distribution for the hemic and solid cancer group.

Schmidberger M et al., *Bioinform Biol Insights*. 2011

Kutijasti dijagram (boxplot)

- Boxplot – Grafički prikaz sažetka od pet brojeva.

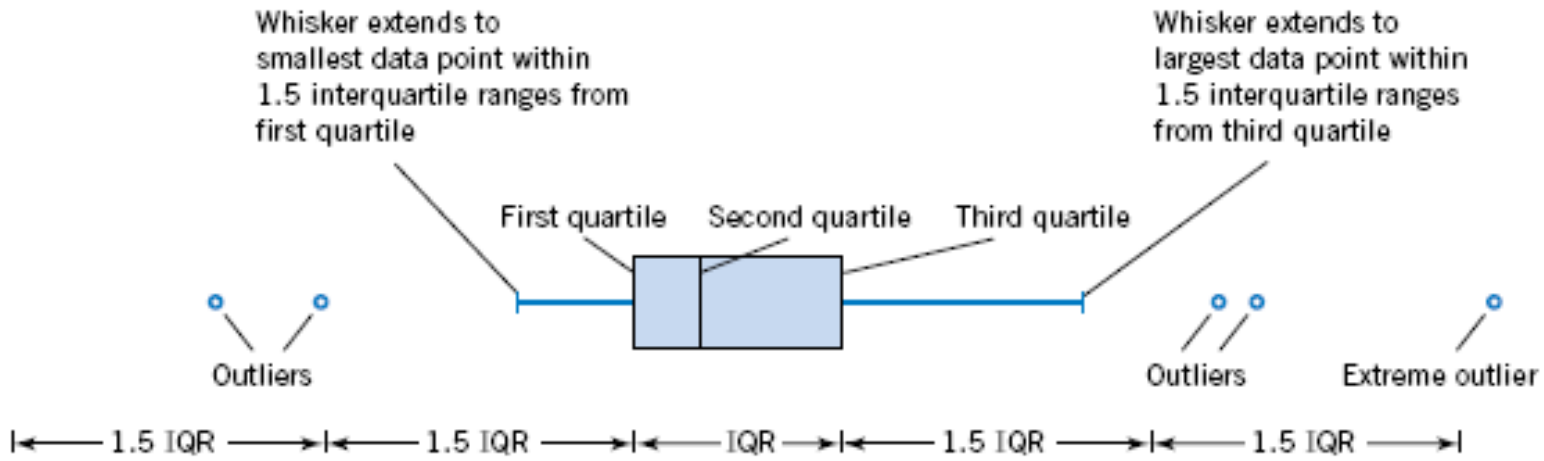
- Konstruirajte kutiju:
  - Nacrtajte i označite ljestvicu koja predstavlja varijablu.
  - Nacrtajte okvir preko ljestvice s lijevim i desnim krajevima na Q1 i Q3.
  - Nacrtajte okomitu liniju kroz okvir na sredini.
  - Nacrtajte lijevi rep (brkove) od okvira do minimuma.
  - Nacrtajte desni rep iz okvira do maksimuma.

Kutijasti dijagram (boxplot)

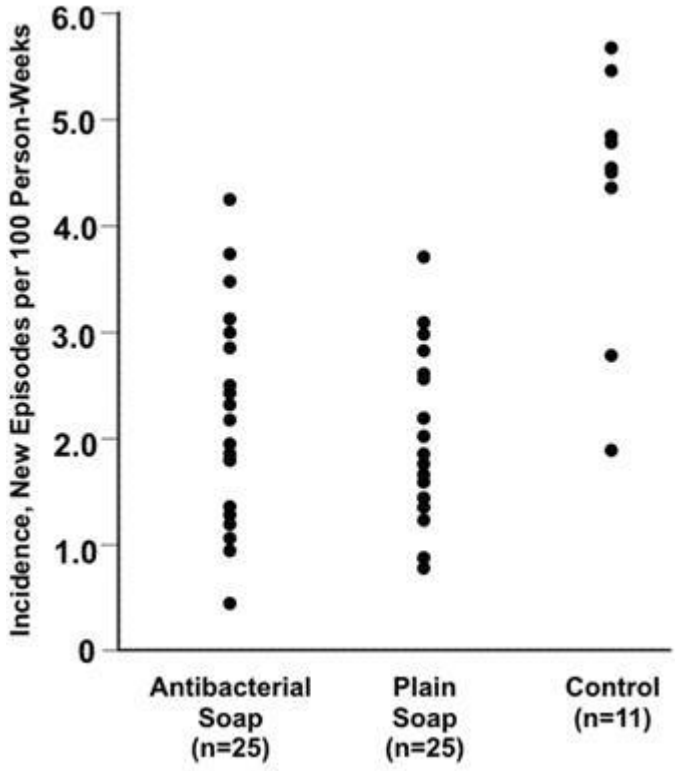
Brkovi mogu predstavljati:

- minimum i maksimum svih podataka
- najniži podatak još uvijek unutar 1,5 IQR od donjeg kvartila, a najviši podatak još uvijek unutar 1,5 IQR od gornjeg kvartila
- jednu standardnu devijaciju iznad i ispod srednje vrijednosti podataka
- 9. percentil i 91. percentil
- 2. percentil i 98. percentil.

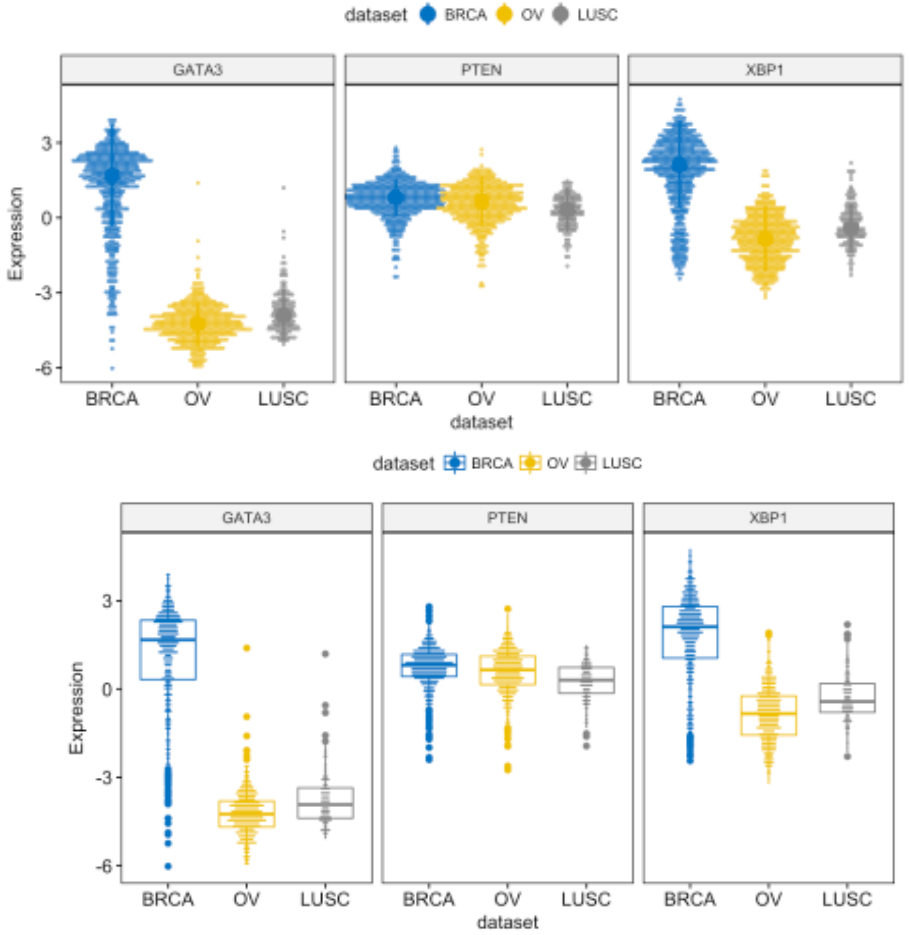
Kutijasti dijagram (boxplot)



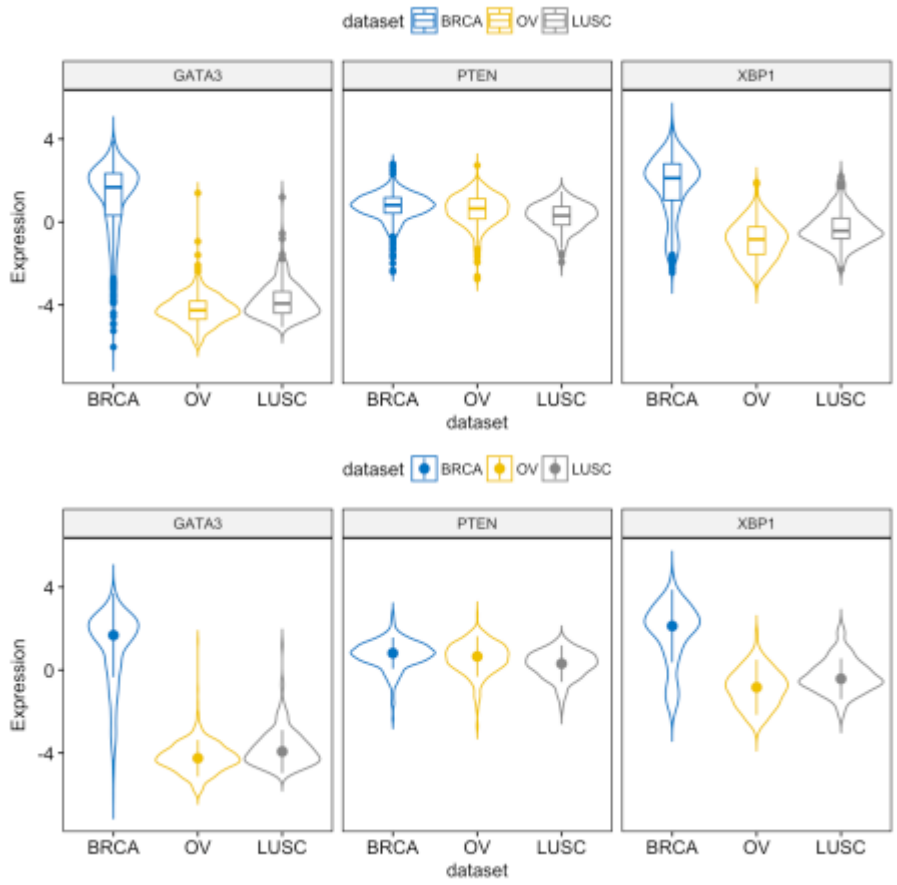
Dot plots



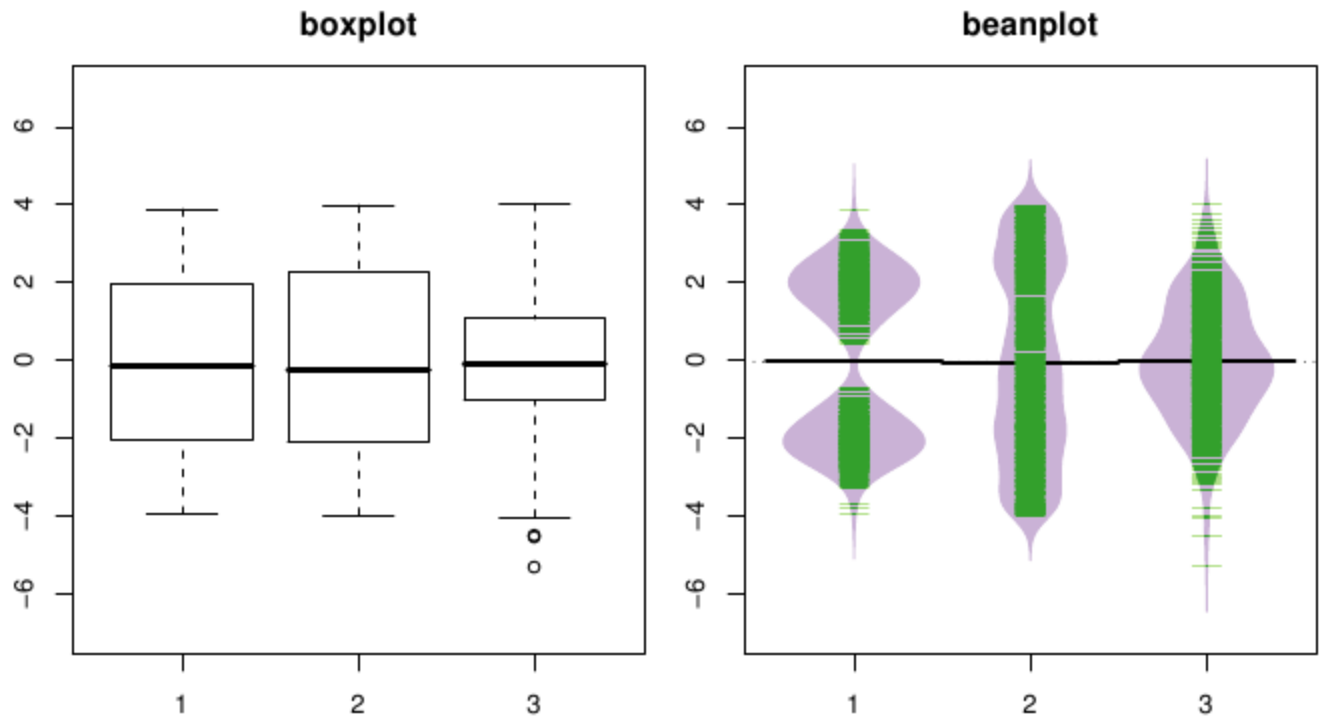
Luby et al., 2004, JAMA



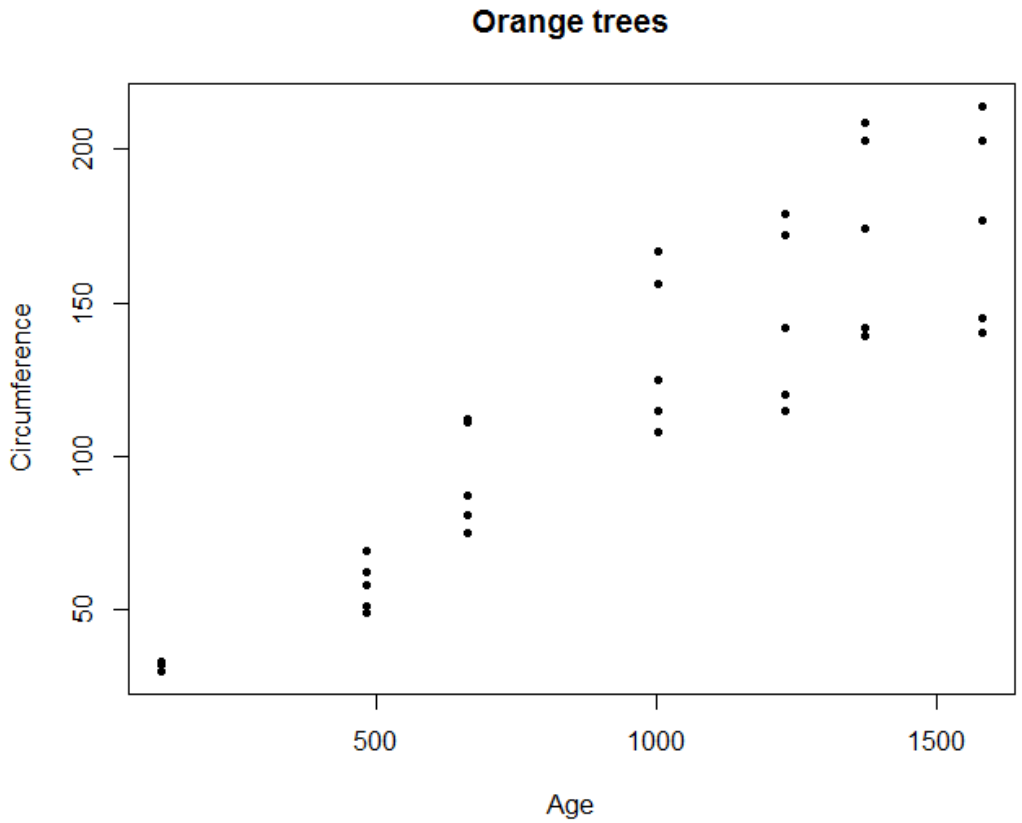
Violin plot


dataset BRCA OV LUSC
GATA3
PTEN
XBP1
Expression
4
0
-4
BRCA
OV
LUSC
dataset
dataset BRCA OV LUSC
GATA3
PTEN
XBP1
Expression
4
0
-4
BRCA
OV
LUSC
dataset

Beanplot



Dijagram raspršenosti *(scatterplot)*



Ponekad je korisno transformirati podatke