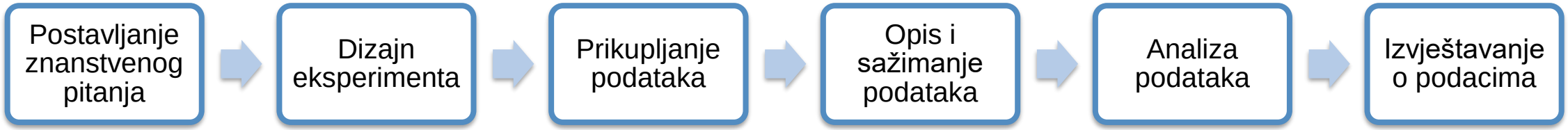
Grafičke i numeričke metode opisivanja podataka

Izv.prof. Rosa Karlić

Predavanje 5, MZIRuB 2024/2025

19.11.2024.

Tijek znanstvenog istraživanja

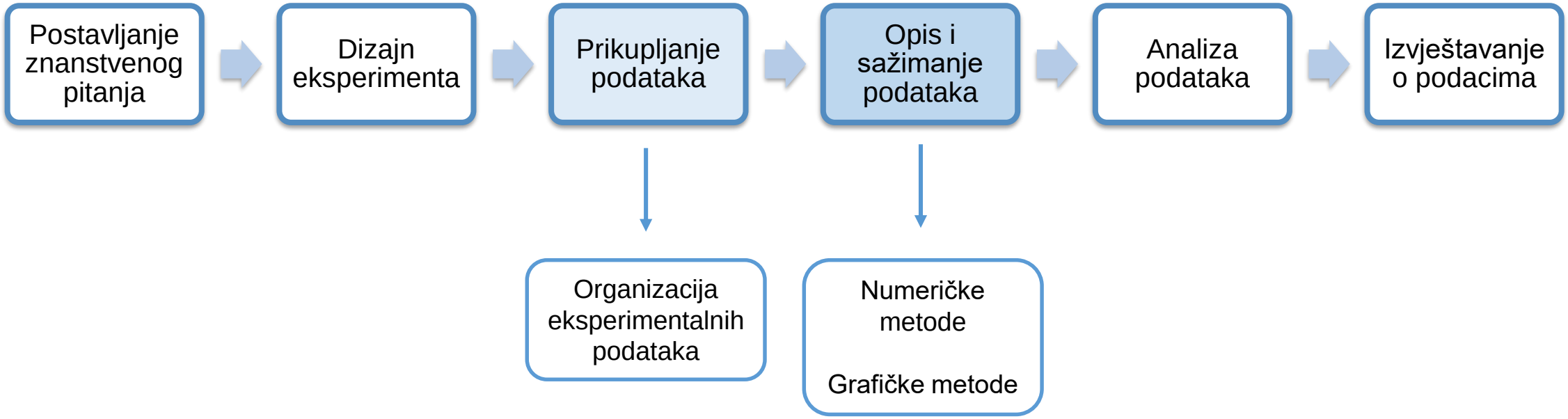


Predavanje 5. 19.11.24. Grafičke i numeričke metode opisivanja podataka
Predavanje 6. 04.12.24. Distribucije vjerojatnosti i intervali pouzdanosti + *Seminar*
Predavanje 7. 11.12.24. Testiranje hipoteza za srednju vrijednost + *Seminar*
Predavanje 8. 18.12.24. **Kolokvij 3** Neparametarski testovi. Kategorički podaci. Greške u testiranju hipoteza.
Predavanje 9. 15.01.25. Korelacija i regresija + *Seminar*
Predavanje 10. 22.01.25. **Kolokvij 4** Planiranje istraživanja. Odgovorno istraživanje i inovacije. Reproducibilnost. + *Seminar*

Kolokviji

- 2 kratka kolokvija (15-20 min)
- 2-3 pitanja (1. kolokvij 6 bodova, 2. kolokvij 4 boda)
- sakupljeno 8 do 10 bodova – oslobađanje od pismenog ispita iz drugog dijela kolegija
- 7 bodova – 2 dodatna boda na pismeni ispit iz drugog dijela kolegija
- 6 bodova – 1 dodatni bod na pismeni ispit iz drugog dijela kolegija

Tijek znanstvenog istraživanja



Podaci

OPAŽANJE (OPSERVACIJA)

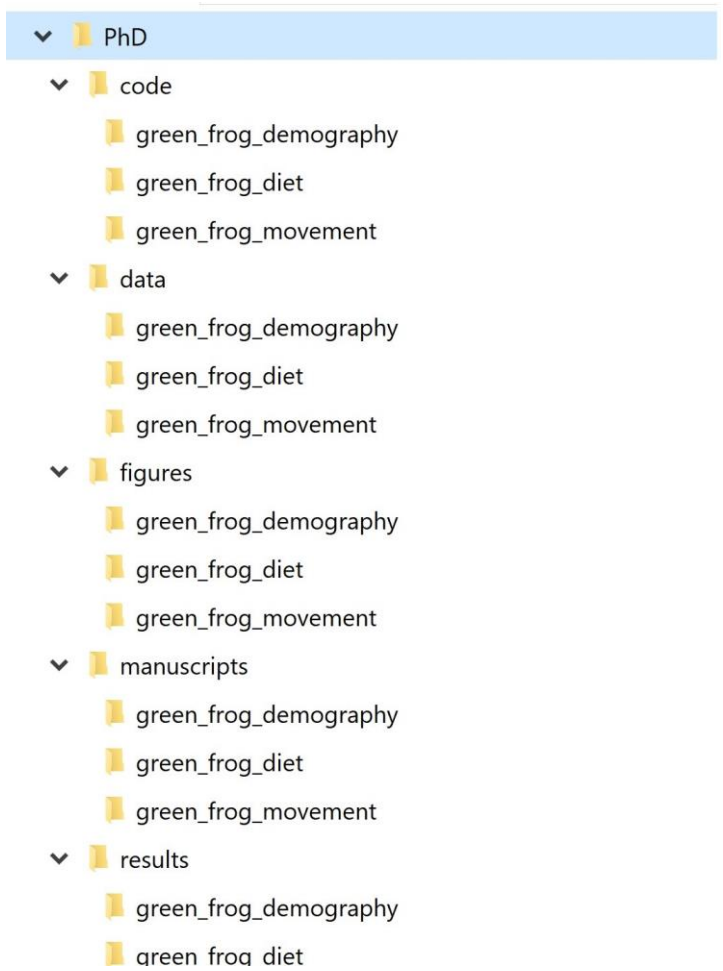
VARIJABLE (KARAKTERISTIKE)

SLUČAJ (*CASE*)

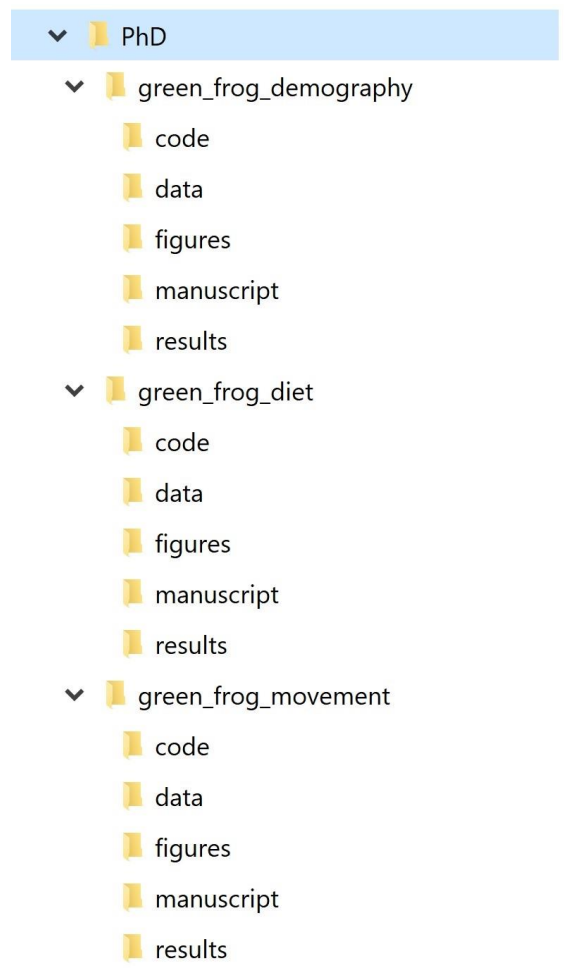
ID	STAROST	SPOL	TRETMAN	GEN1	GEN2	GEN3	STATUS
87	53	F	trt2	17.41	28.23	4.17	zdrav
119	53	M	ctrl	19.84	52.56	47.24	zdrav
67	52	M	trt2	19.19	27.49	12.07	zdrav
62	54	M	trt2	22.77	28.45	9.62	bolestan
131	55	F	ctrl	24.17	49.91	49.55	bolestan
50	54	F	trt1	17.15	15.32	10.67	zdrav
106	54	M	ctrl	17.92	44.95	51.39	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

Organizacija direktorija

- Postoji li preklapanje u podacima između projekata?
- Smanjiti dupliciranje podataka
- Idealno bi projektni direktorij trebao sadržavati sve potrebno za reproduciranje rezultata



Prema aktivnostima



Prema projektima

Dokumentacija

- README – sve informacije potrebne za razumijevanje projekta
- Što su ulazni i izlazni podaci, gdje se nalaze određene datoteke
- Tko su autori pojedinih datoteka, kada i na koji način ste ih dobili
  “File X.csv poslao je Ivan Horvat 19.11.2024. na e-mail adresu rosa@bioinfo.hr”

Sirovi podaci i metapodaci (*Metadata*)

- Uvijek treba **čuvati sirove podatke**
- **Metapodaci – dodatne informacije:**
  - Što su podaci, tko ih je prikupio, kada i gdje
  - Kako pronaći podatke i pristupiti im
  - Upozorenja o poznatim problemima ili nedosljednostima u podacima (npr. stupac koji opisuje kvalitetu)
  - Informacije za provjeru jesu li podaci ispravno uvezeni, npr. broj redaka i stupaca u skupu podataka i ukupni zbroj numeričkih stupaca

Imenovanje datoteka i direktorija

- Informativna imena – računalno čitljiva i razumljiva ljudima
- Idealno samo alfanumerički znakovi, crtica (-) i donja crtica (_)
- Ne sadrže razmake, interpunkcijske znakove i posebne znakove (uključujući dijakritičke)
- Uključiti podatke za sortiranje (datum u YYYYMMDD formatu; broj verzije .v1, .v2; broj verzije 001, 002)
- Konzistentno označavanje metapodataka i riječi

predavanje5_MZIRuB_svi-smjerovi_20241118.ppt

Predavanje5_MZIRuB_SviSmjerovi_20241118.ppt

https://ecorepsci.github.io/reproducible-science/project-organization.html

Imenovanje datoteka i direktorija

Loši primjeri:

data.csv

data_cleaned_March-22-2012.csv

analysis code.R

Green Frogs Manuscript_Final_edits.docx

final.docx

Super Cool Report vFinal.xlsx

Super Cool Report vFinal_1.xlsx

Super Cool Report vFinal_2.xlsx

Super Cool Report vFinal_Final.xlsx

Super Cool Report vFinal_Final-UPDATED.xlsx

Super Cool Report vFinal_Final-UPDATED_NEW.xlsx

Malo bolje, ali ne optimalno:

models.lincRNA.filter_0.2.noOverlapFilter.separate.groups.20240424

models.lincRNA.filter_0.2.OverlapFilter.separate.groups.20240424

models.lincRNA.filter_0.4.noOverlapFilter.separate.groups.20240424

models.lincRNA.filter_0.4.OverlapFilter.separate.groups.20240424

ei-cfDNA.TOO.FC.median.20230823

ei-cfDNA.TOO.FC.median.20230719

ei-cfDNA.TOO.FC.mean.20230823

ei-cfDNA.TOO.FC.mean.20230802

ei-cfDNA.TOO.FC.mean.20230719

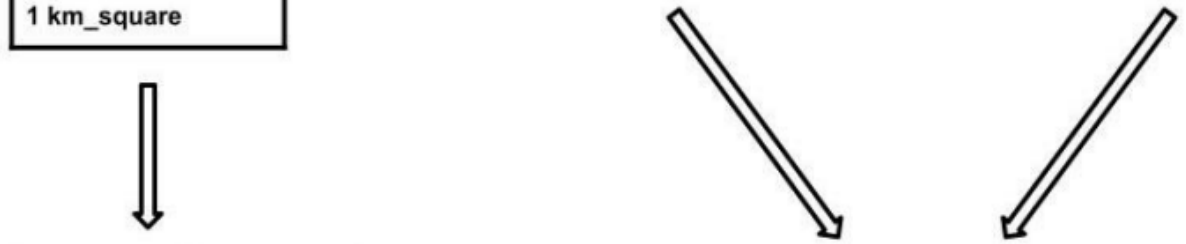
Organizacija podataka unutar tablice

- **Svaki red** treba predstavljati **jedno opažanje** (tj. zapis), a **svaki stupac** treba predstavljati **jednu varijablu** ili vrstu mjerenja
- **Svaka ćelija** treba sadržavati samo **jednu vrijednost** (npr. ne uključivati mjerne jedinice u ćeliju s vrijednostima ili uključivati više mjerenja u jednu ćeliju)
- Za **svaku vrstu informacija** trebao bi postojati **samo jedan stupac** (ovo pravilo najčešće krše strukturirani podaci s unakrsnom tablicom, gdje različiti stupci sadrže mjerenja iste varijable)

Area
10 km_square
100 m_square
1000 m_square
1 km_square

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2



AreaValue	AreaUnits
10	km_square
100	m_square
1000	m_square
1	km_square

Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2

Figure 1. Examples of how to restructure two common issues with tabular data. (a) Each cell should only contain a single value. If more than one value is present then the data should be split into multiple columns. (b) There should be only one column for each type of information. If there are multiple columns then the column header should be stored in one column and the values from each column should be stored in a single column.

White et al. Ideas in Ecology and Evolution 6(2): 1–10, 2013

Koju vrijednost koristiti za podatke koji nedostaju (*missing data*)?

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

White et al. Ideas in Ecology and Evolution 6(2): 1–10, 2013

Organizacija podataka unutar tablice

Canopy measures taken: August 24, 2020
Measurements taken by: Pat Smith

	Plant			Leaf		
Plant	# of leaves	Height (cm)		Location	Length (cm)	Width (cm)
1	14	35		Inner	12	5.5
2	10	29			13.5	8
3	14	30			14.5	7.5
4	12	26			16.5	5.5
5	13	36		Outer	14	8.5
6	15	33			15.5	10
7	15	32			17	9
8	16 (several of the leaves were damaged)	29		Middle	13	7
9	14	37			16.5	6.5
10	15	36			11	9
11	12	29			13	8.5
12	16	28		Inner	15.5	5.5
13	12	39			16	6
14	14	30		Outer	14	7.5
15	16	26			11	7
Average	13.71	31.67		Average	14.20	7.40

DA!

- pravokutni format samo sa stupcima i recima (bez dinamičkog sadržaja - grafikoni, formule i veze između radnih listova) – osigurava kompatibilnost s različitim vrstama programa za analizu

NE!

- Prazni stupci, redovi i ćelije
- Spojene ćelije za grupiranje varijabli
- Bilješke u cijeloj tablici.
- Statistika u zadnjem retku tablice.

https://guides.lib.uoguelph.ca/OrganizingYourResearchData

Organizacija podataka unutar tablice

Canopy measures taken: August 24, 2020
Measurements taken by: Pat Smith

	Plant			Leaf	
Plant	# of leaves	Height (cm)	Location	Length (cm)	Width (cm)
1	14	35	Inner	12	5.5
2	10	29		13.5	8
3	14	30		14.5	7.5
4	12	26			
5	13	36	Outer		
6	15	33			
7	15	32			
8	16 (several of the leaves were damaged)	29	Middle		
9	14	37			
10	15	36			
11	12	29			
12	16	28	Inner		
13	12	39			
	14	30	Outer		
	16	26			
	13.71	31.67	Average		

NE

DA

Plant	PlantNumLeaves	PlantHeight	LeafLocation	LeafLength	LeafWidth	Notes
1	14	35	Inner	12	5.5	
2	10	29	Inner	13.5	8	
3	14	30	Inner	14.5	7.5	
4	12	26	Inner	16.5	5.5	
5	13	36	Outer	14	8.5	
6	15	33	Outer	15.5	10	
7	15	32	Outer	17	9	
8	16	29	Middle	13	7	several of the leaves were damaged
9	14	37	Middle	16.5	6.5	
10	15	36	Middle	11	9	
11	12	29	Middle	13	8.5	
12	16	28	Inner	15.5	5.5	
13	12	39	Inner	16	6	
14	14	30	Outer	14	7.5	
15	16	26	Outer	11	7	

https://guides.lib.uoguelph.ca/OrganizingYourResearchData

Imena varijabli

- Konzistentno koristite mala i velika slova
- Koristite donju crticu (podvlaku), _ kao zamjenu za razmake za odvajanje riječi
- Izbjegavajte interpunkcijske znakove, specijalne znakove i dijakritike
- Ime treba dati informaciju o podacima koje varijabla sadrži
- Ako je riječ skraćena u nazivu varijable, ista se kratica koristi za sva imena

NE

Plant	# of leaves	Height (cm)		Location	Length (cm)	Width (cm)

Plant	PlantNumLeaves	PlantHeight	LeafLocation	LeafLength	LeafWidth	Notes

DA

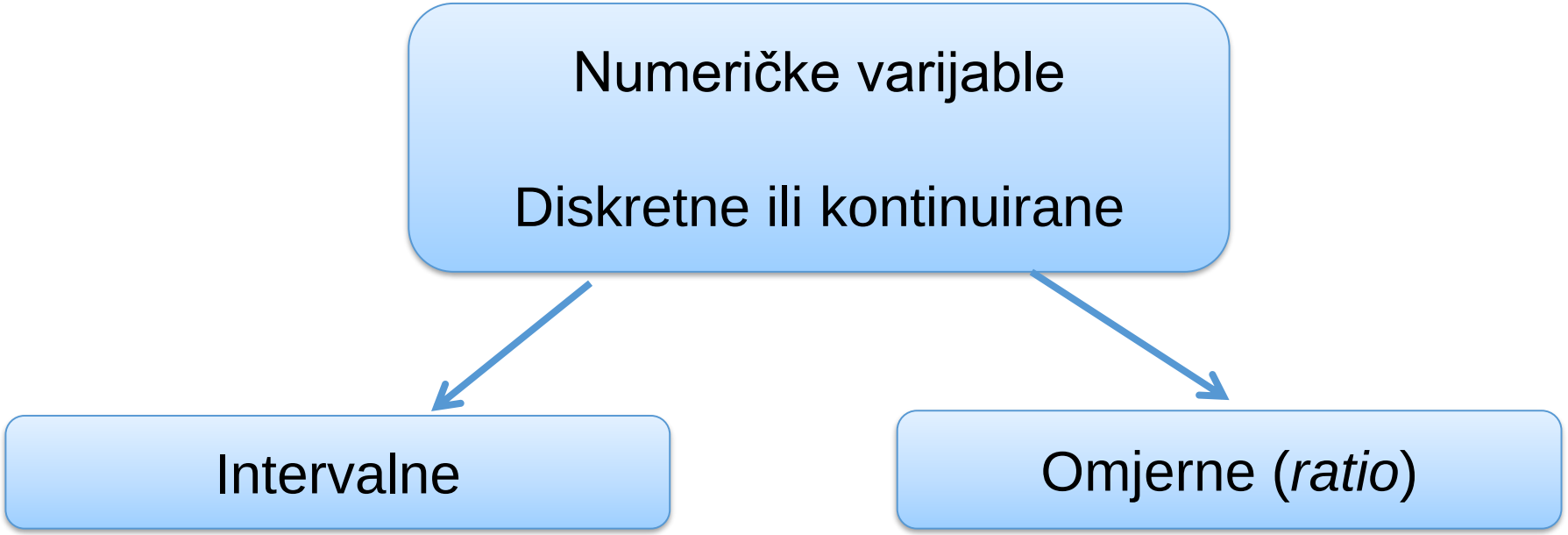
Tipovi varijabli

- 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

Tipovi varijabli



Intervalna varijabla je ona kod koje postoji poredak i razlika između dvije vrijednosti je smislena
Omjer dviju vrijednosti nema smislenu interpretaciju
Npr. Temperatura (F,°C),

Omjerna varijabla ima sva svojstva intervalne varijable i jasnu definiciju 0,0. Kada je varijabla jednaka 0,0, nema te varijable.
Npr. Aktivnost enzima, koncentracija tvari

Tipovi varijabli

- Primjer: broj kopija gena u stanici – diskretna omjerna varijabla
  - Diskretno: varijabla je prebrojiva i poprima cijele brojeve (npr. 0, 1, 2, 3 kopije gena).
  - Omjer: ima pravu nultu točku (stanica može imati nula kopija gena).
  - Omjeri su značajni (npr. stanica s 4 kopije gena ima dvostruko više od stanice s 2 kopije).

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**

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 (tipične vrijednosti)

Mod - najčešća observacija u skupu.

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

- vrijednost takva da je zbroj pozitivnih i negativnih udaljenosti od srednje vrijednosti jednak nuli

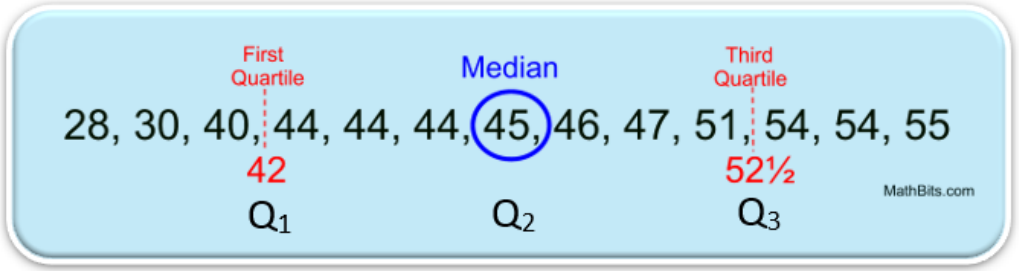
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

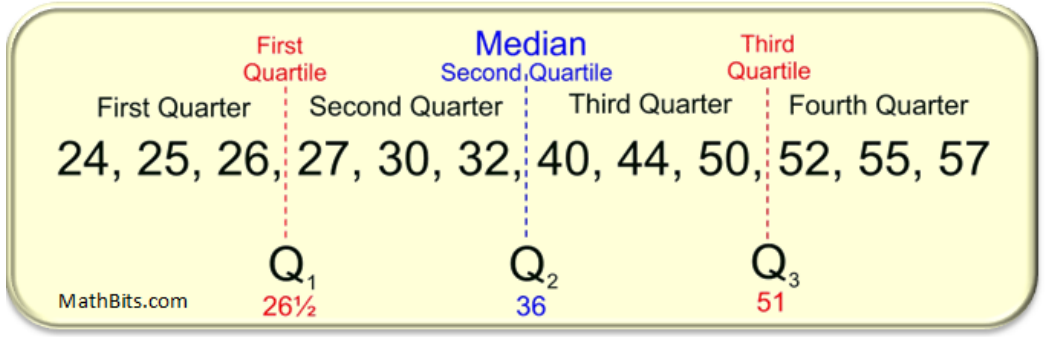
Numeričke opisne metode

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 (Q1), medijan (Q2), gornji kvartil (Q3)





Numeričke opisne metode

Mjere rasapa (raspršenosti)

Raspon = Maksimum - Minimum

IQR = 75. kvantil - 25. kvantil, mjeri raspršenost 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

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

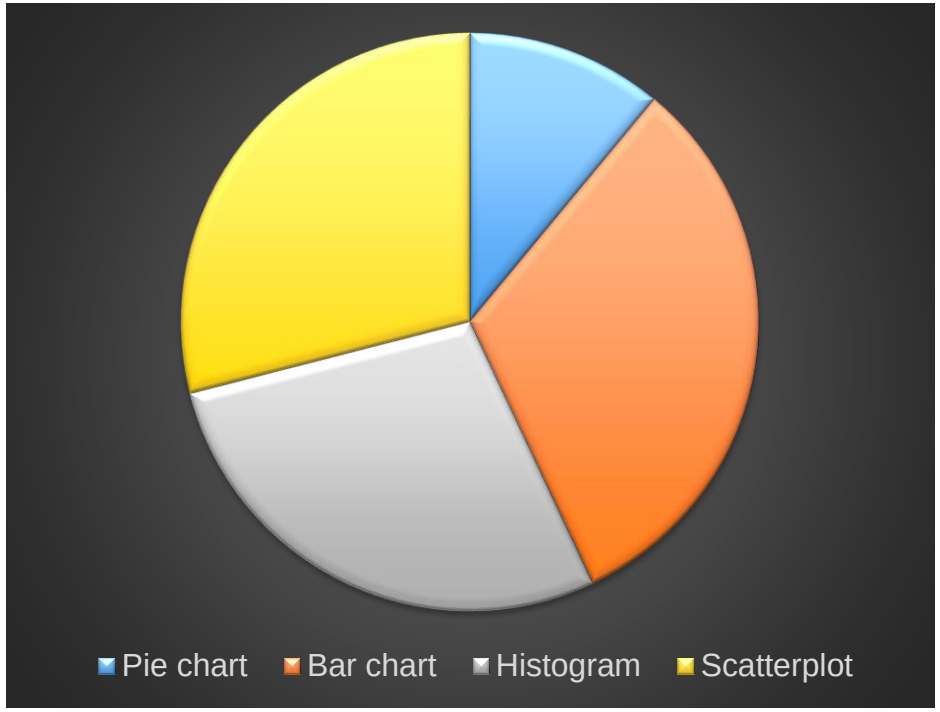
Kvadrat standardne devijacije - varijanca

Numeričke opisne metode prema tipovima varijabli

OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution	Yes	Yes	Yes	Yes
median and percentiles	No	Yes	Yes	Yes
sum or difference	No	No	Yes	Yes
mean, standard deviation, standard error of the mean	No	No	Yes	Yes
ratio, or coefficient of variation	No	No	No	Yes

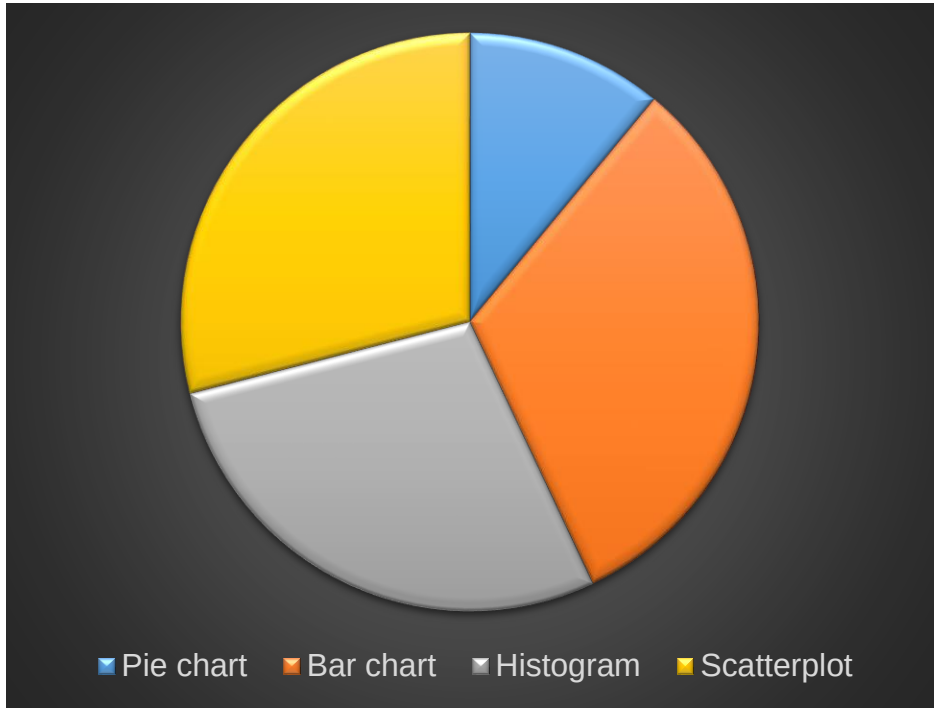
Grafičke metode

Kružni dijagram

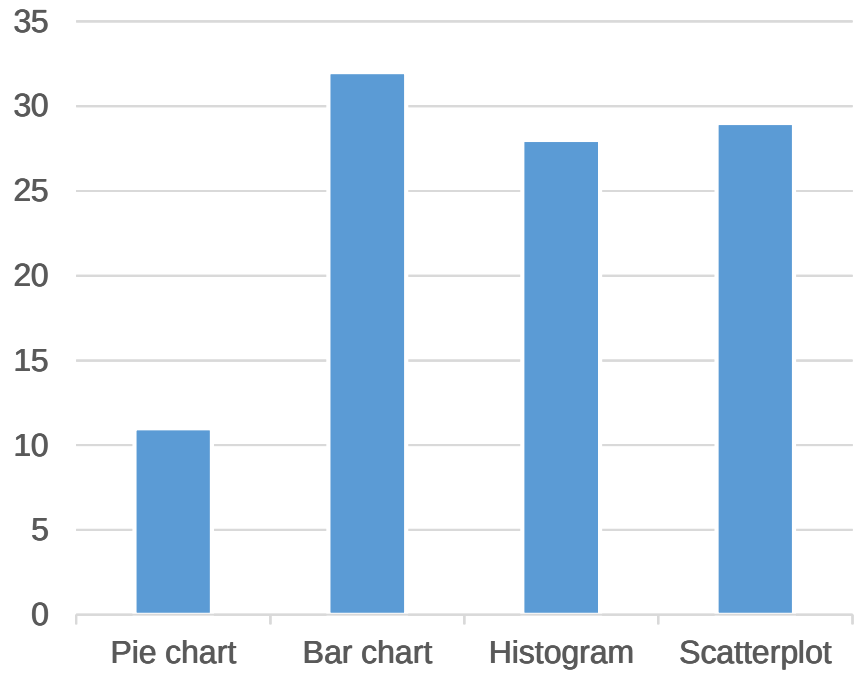


Grafičke metode

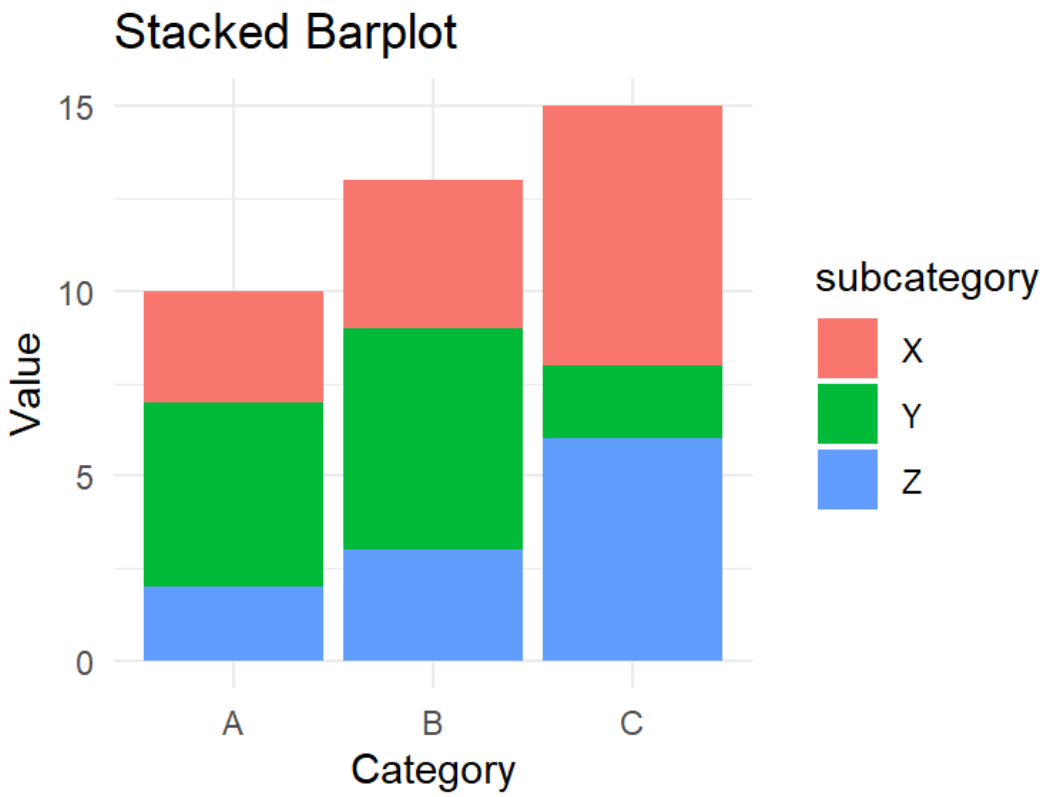
Kružni dijagram

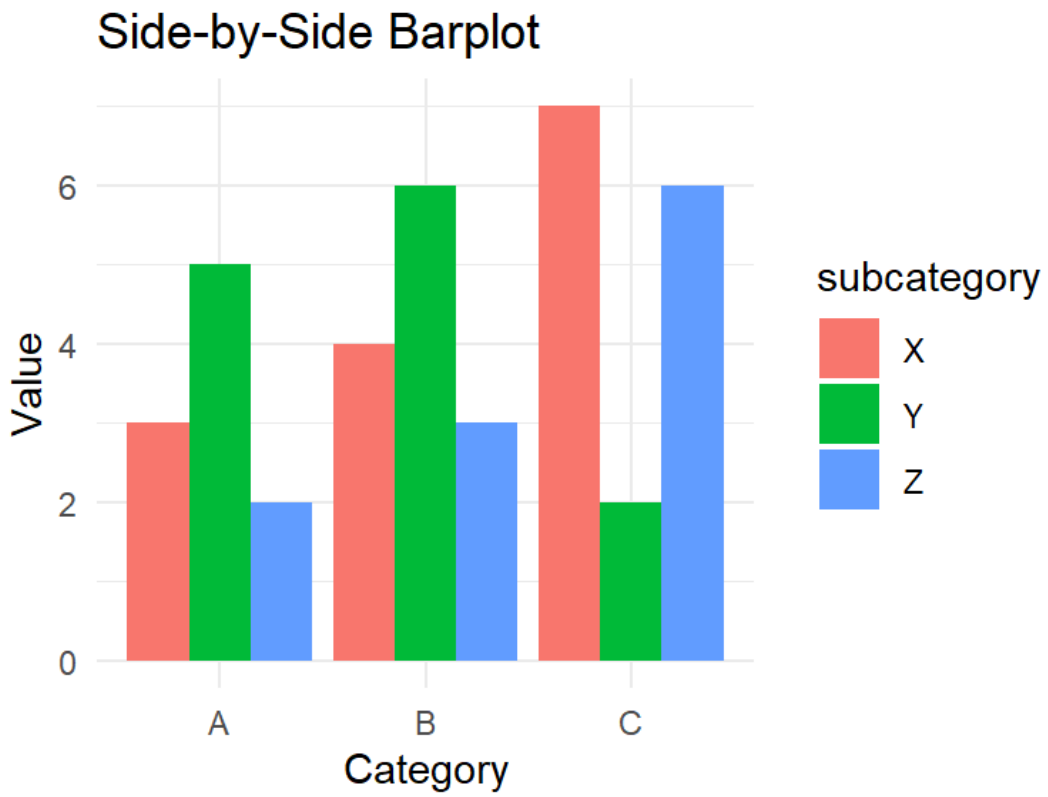


Stupčasti dijagram

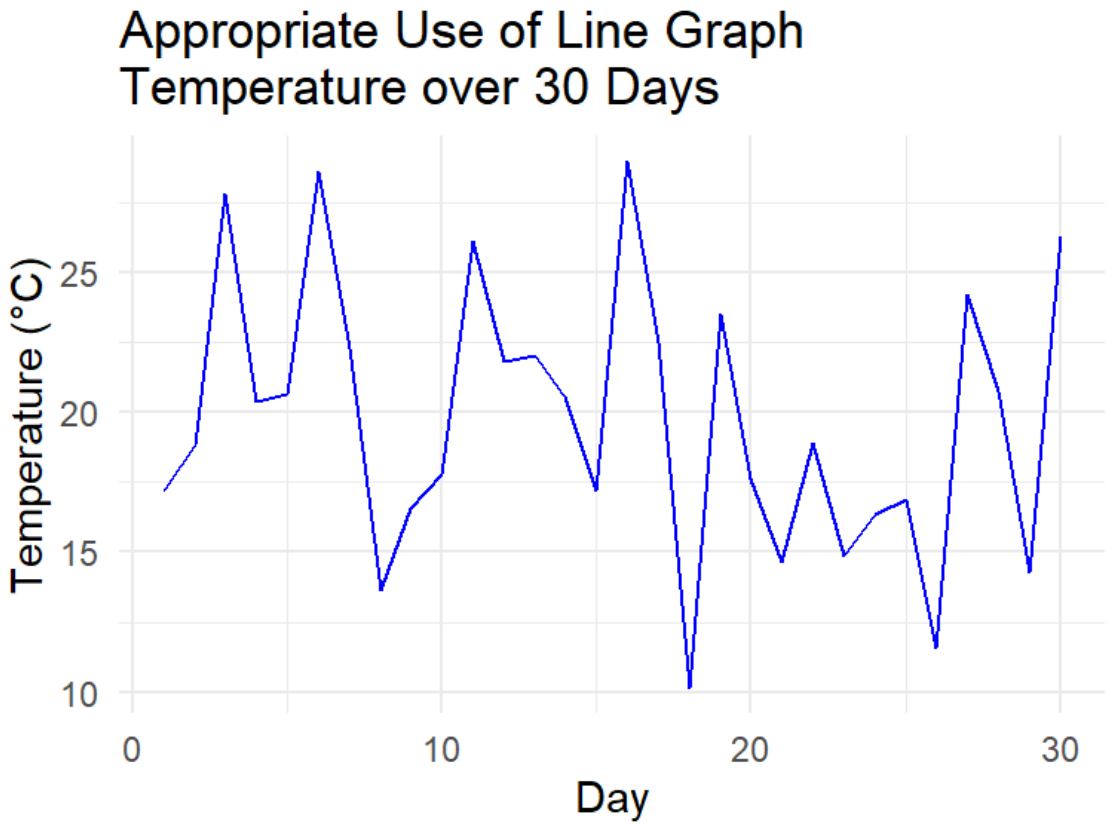


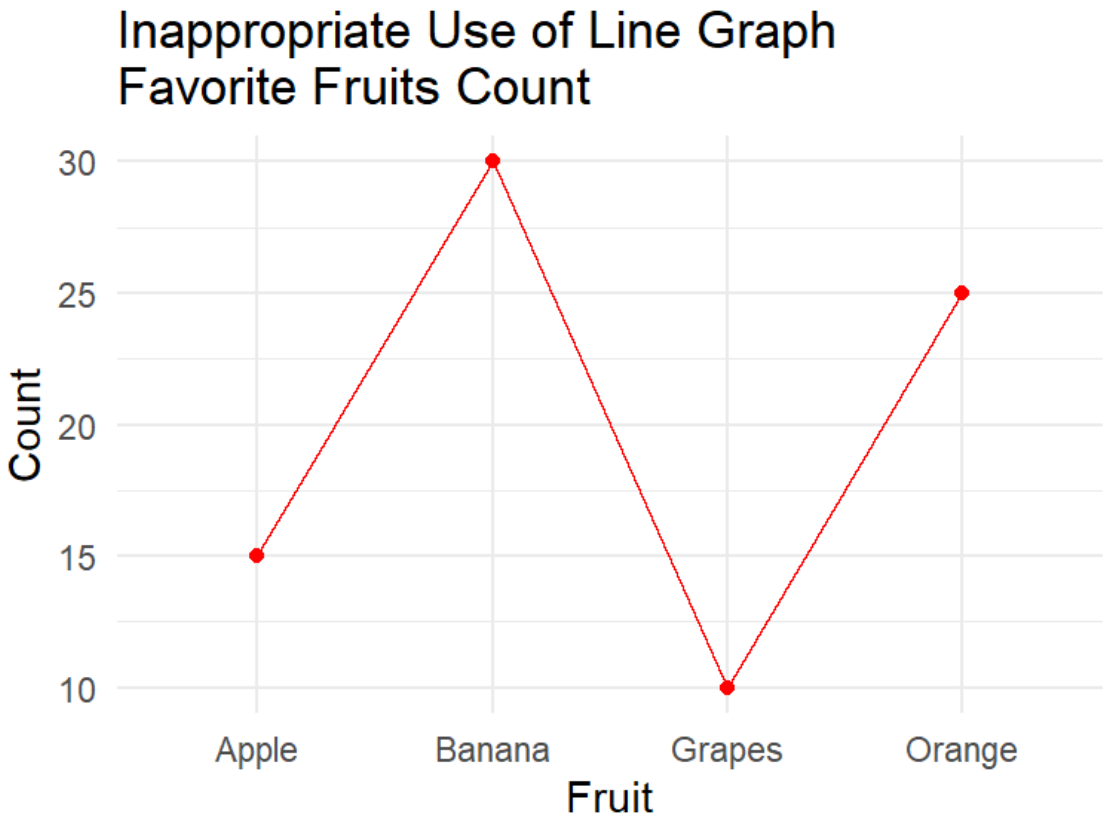
Stupčasti dijagram – prikaz više varijabli





Linijski graf





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 intervala (*bins*) 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.

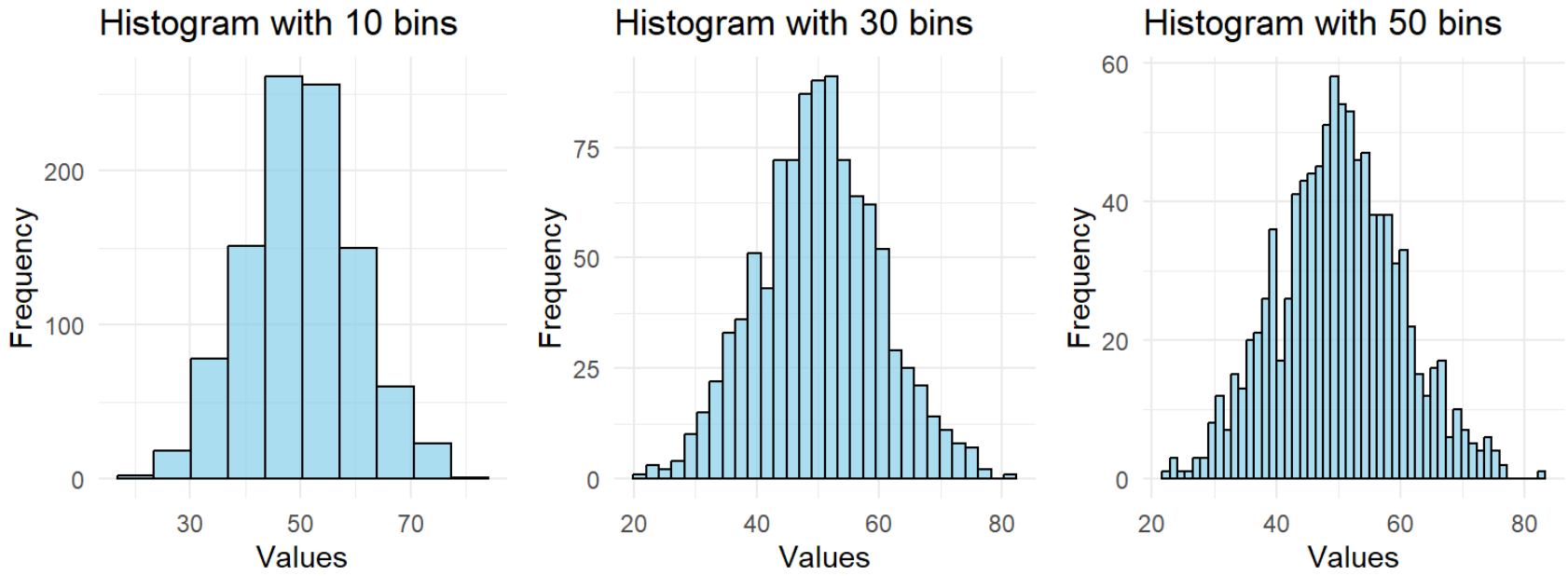
Histogram

- Vizualni prikaz distribucije učestalosti

(1) Označite granice intervala na vodoravnoj ljestvici.

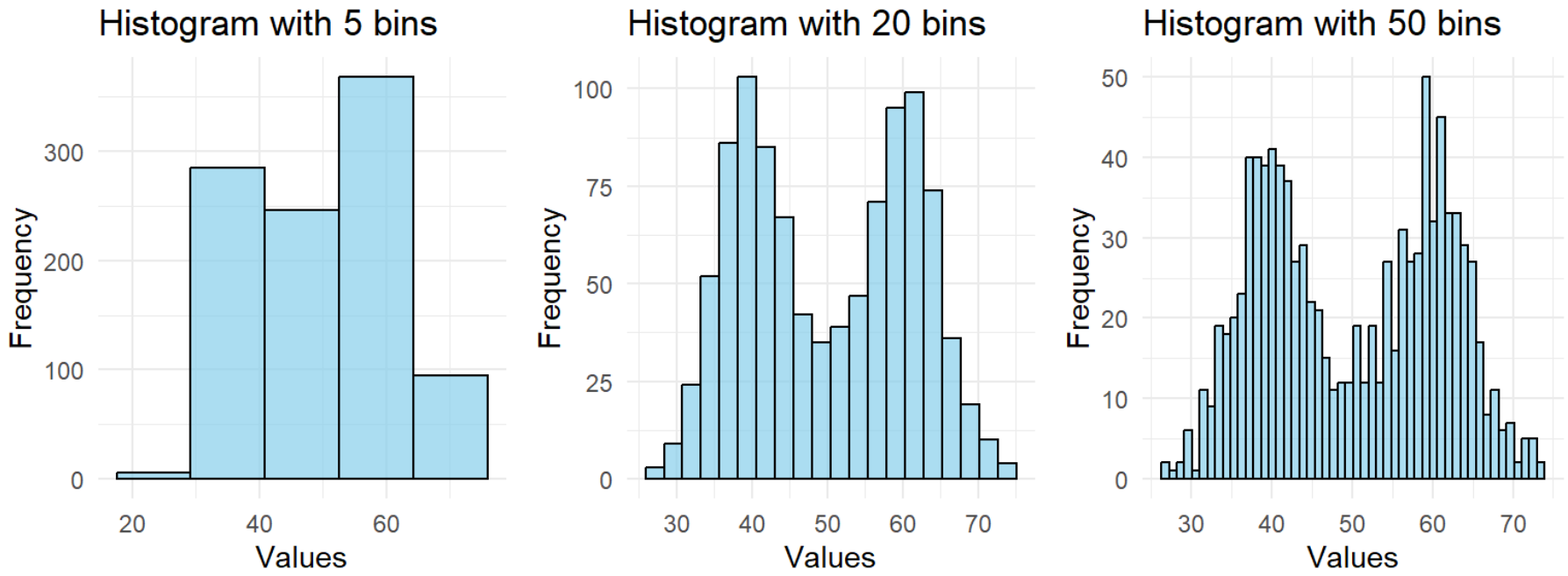
(2) Označite i označite vertikalnu ljestvicu frekvencijama ili relativnim frekvencijama.

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



Histogram

- Ponekad pravi oblik distribucije učestalosti nije vidljiv zbog broja intervala u histogramu



Histogram

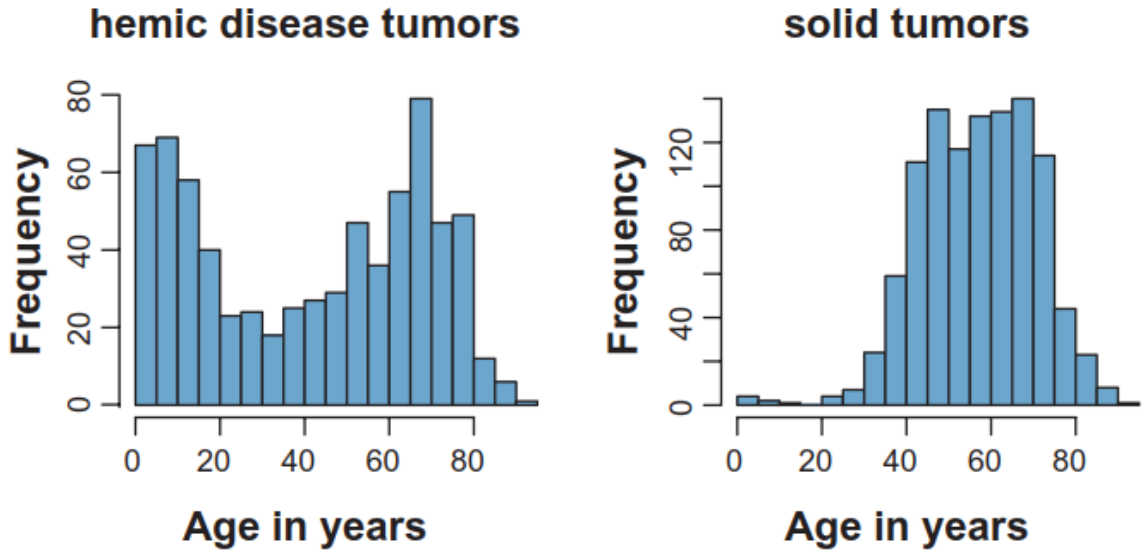


Figure 1. Histogram of the age distribution for the hemic and solid cancer group.

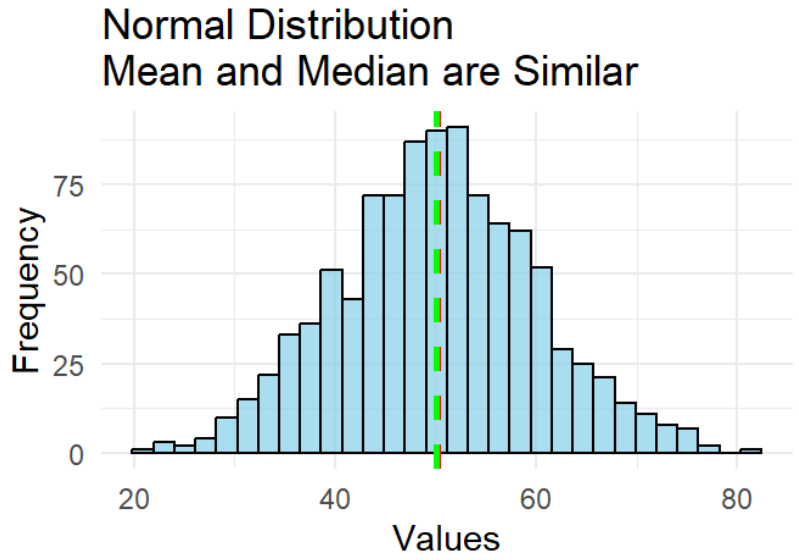
Schmidberger M et al., *Bioinform Biol Insights*. 2011

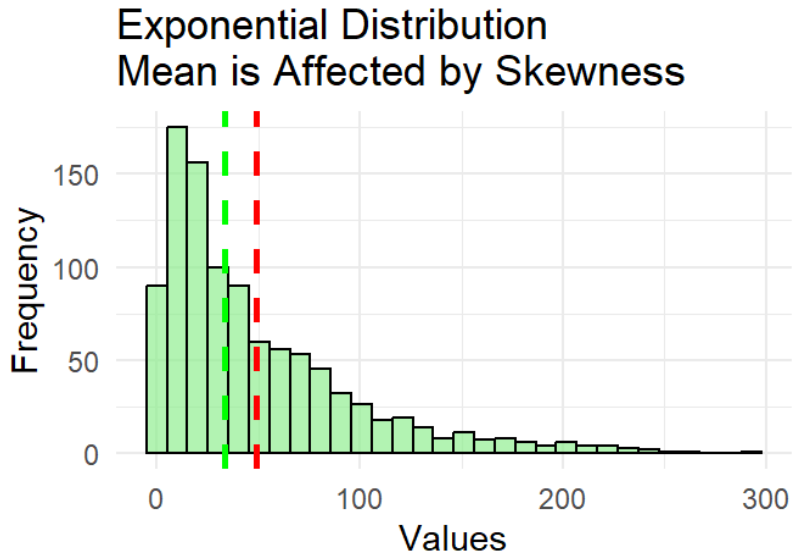
Različiti oblici distribucija učestalosti

- Oblik empirijske distribucije učestalosti karakterizira se pomoću nekoliko mjera:
  - **Mjere lokacije** – npr. srednja vrijednost ili medijan
  - **Mjere raspršenja** – npr. varijanca ili interkvartilni raspon
  - **Mjere simetrije (*skewness*)** - kvantificira asimetriju distribucije:
    - Pozitivna: rep na desnoj strani je duži; distribucija je desno zakošena.
    - Negativna: rep na lijevoj strani je duži; distribucija je lijevo zakošena.
    - Simetrija: asimetrija ≈ 0 za simetrične distribucije.
  - **Mjere zaobljenosti (*kurtosis*) – opisuje oštrinu vrha distribucije u odnosu na rep distribucije**
    - Leptokurtična - Oštriji vrh i teški repovi (npr. t-distribucija).
    - Mesokurtična (Kurtosis = 3): normalna distribucija.
    - Platikurtična ravniji vrh i manji repovi.

Različiti oblici distribucija učestalosti

- **Srednje vrijednosti** najbolje je koristiti za približno simetrične podatke: na srednju vrijednost utječu stršeće vrijednost i asimetrija.
- **Medijan** je bolje koristiti za podatke koji su asimetrično distribuirani ili sadrže odstupanja: na medijan ne utječu odstupanja i asimetrija.





Kutijasti dijagram (boxplot)

- Boxplot – Grafički prikaz sažetka od pet brojeva.

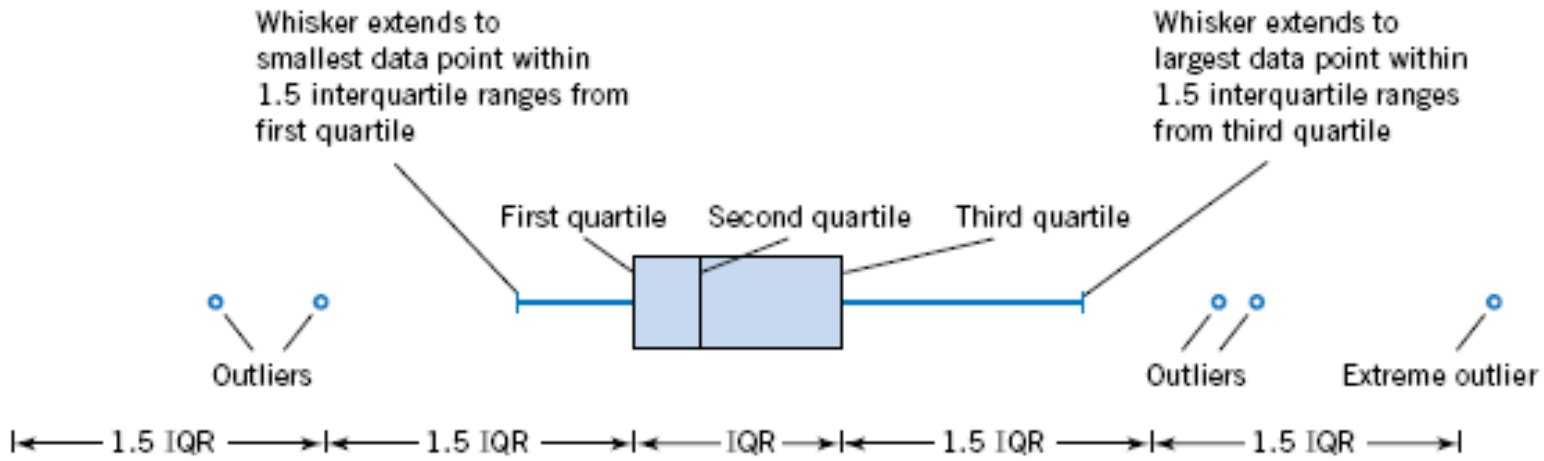
- Konstruirajte kutiju:
  - Nacrtajte i označite ljestvicu koja predstavlja varijablu.
  - Nacrtajte okvir preko ljestvice s lijevim i desnim krajevima na Q1 i Q3.
  - Nacrtajte okomitu liniju kroz okvir na sredini.
  - Nacrtajte lijevi rep (brkove) od okvira do minimuma.
  - Nacrtajte desni rep iz okvira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

Kutijasti dijagram (boxplot)

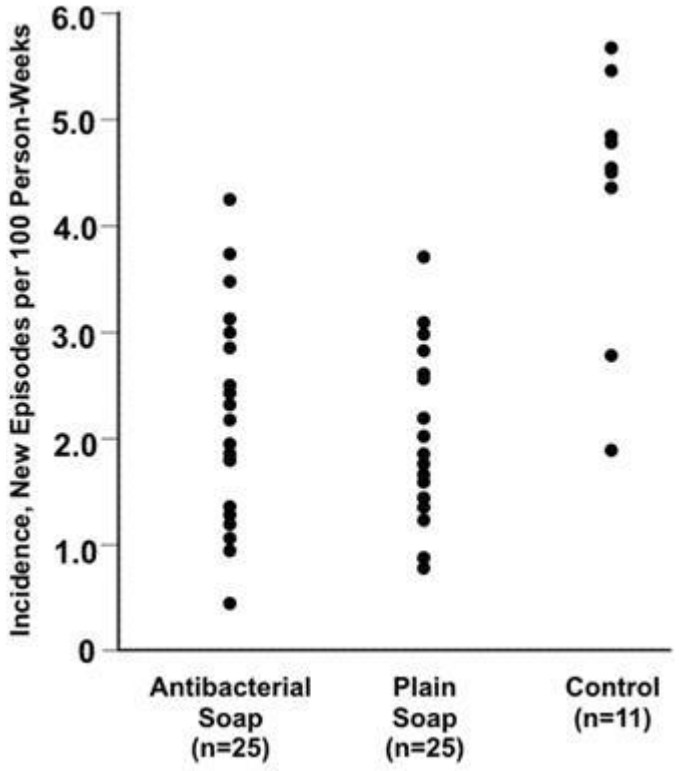
Brkovi mogu predstavljati:

- 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.

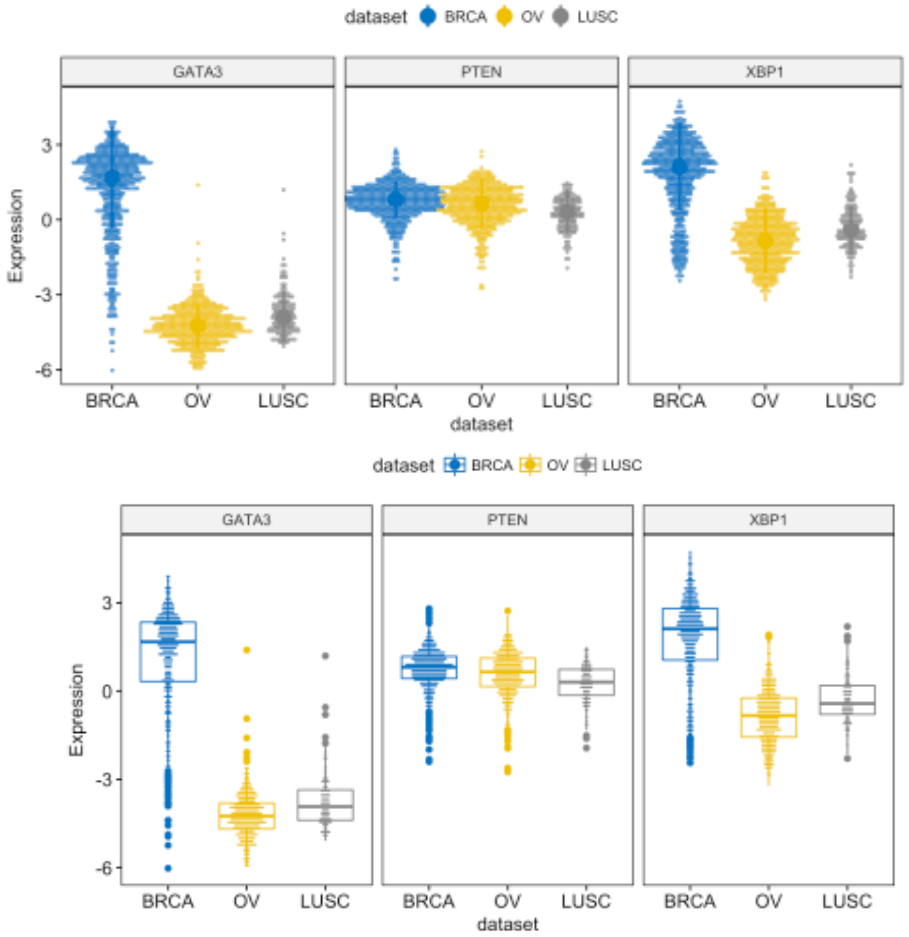
Kutijasti dijagram (boxplot)



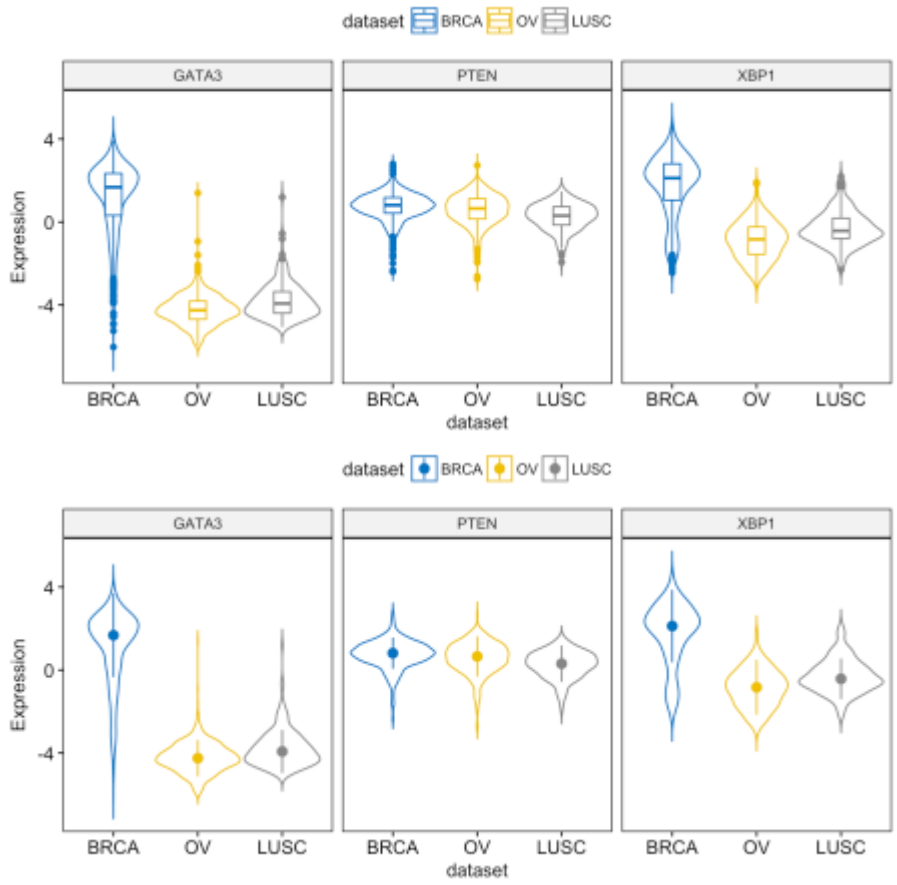
Dot plots



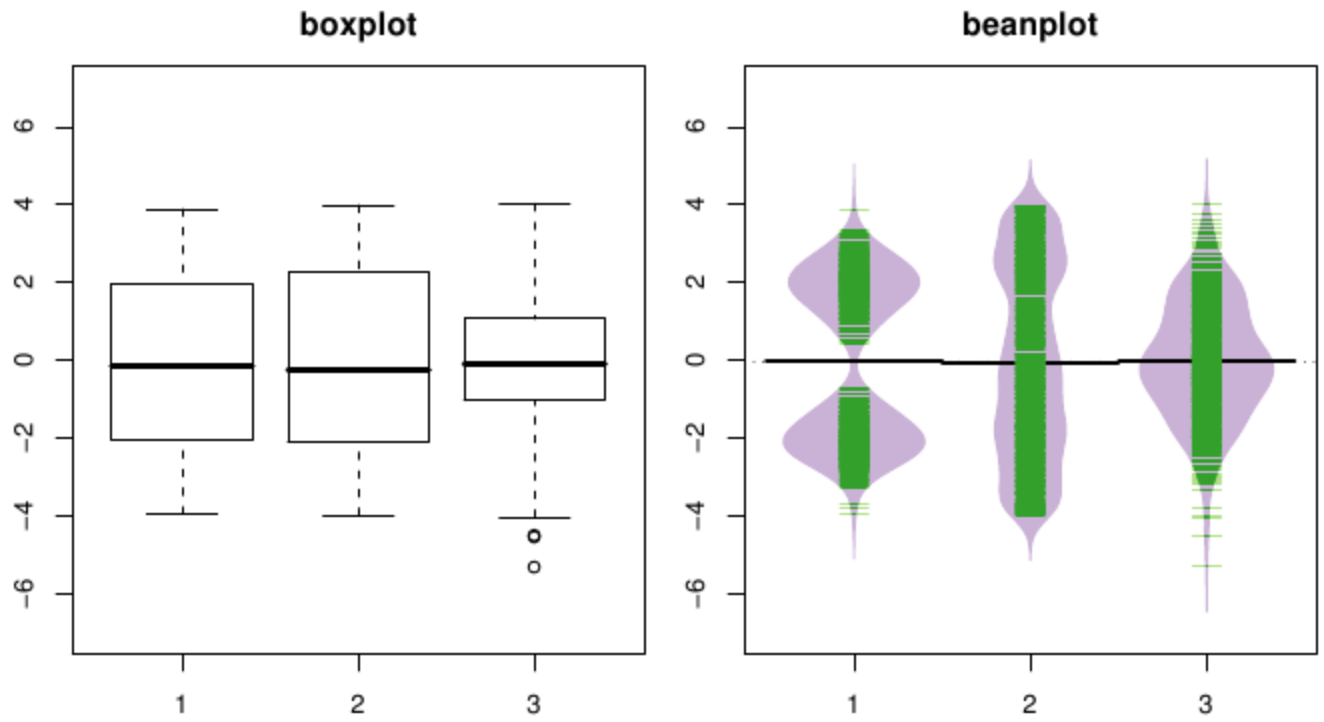
Luby et al., 2004, JAMA



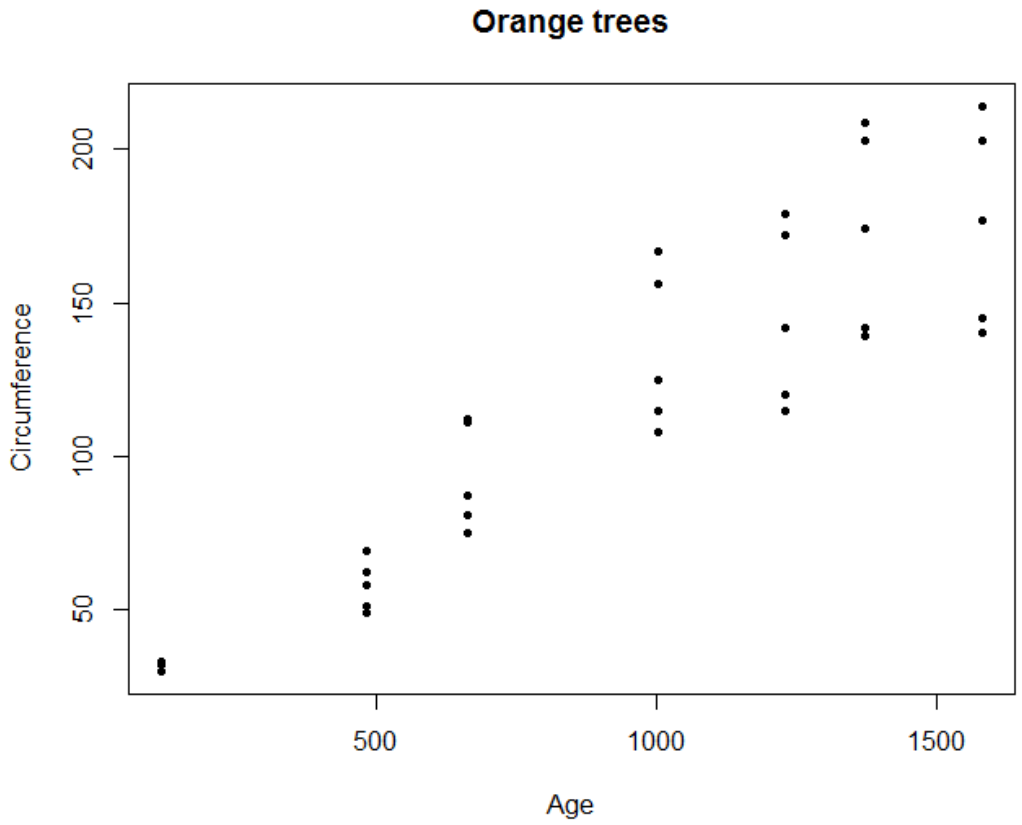
Violin plot


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

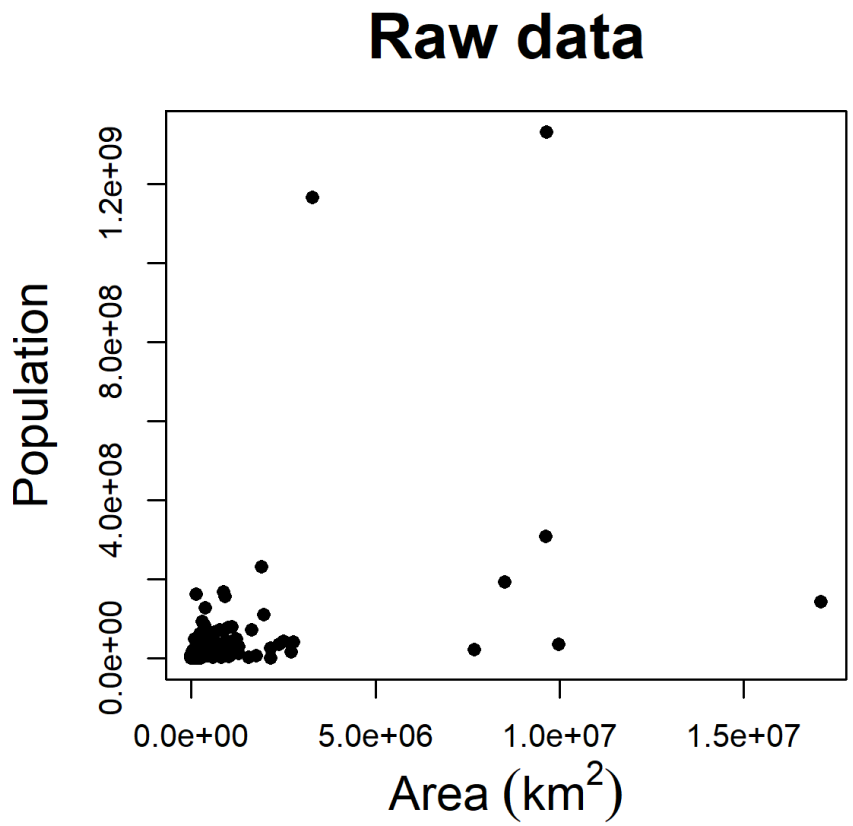
Beanplot

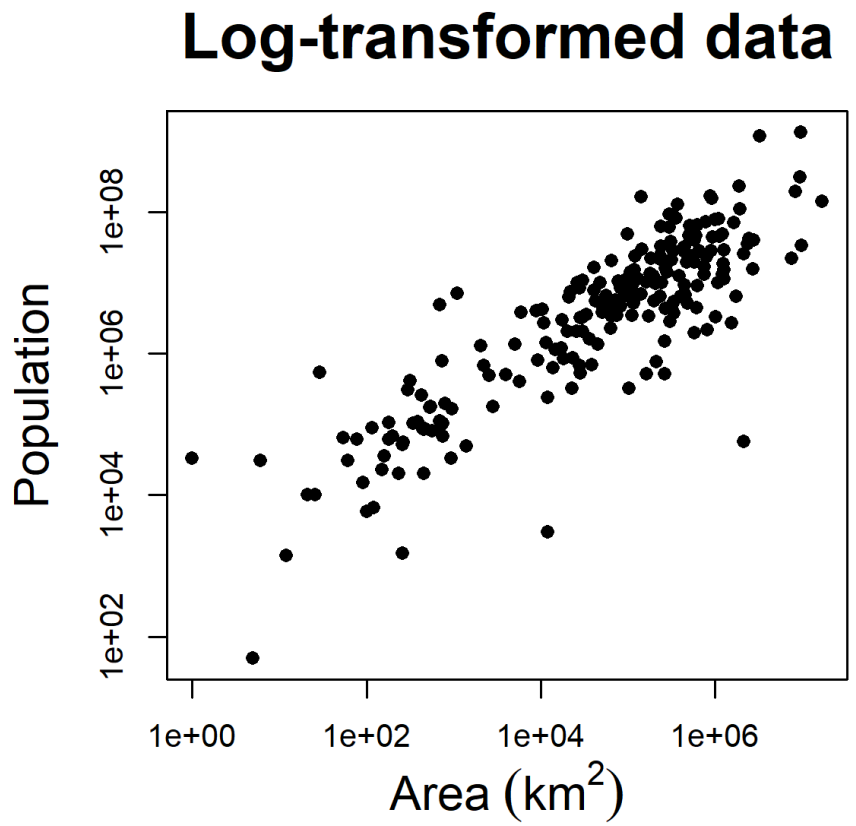


Dijagram raspršenosti *(scatterplot)*



Ponekad je korisno transformirati podatke





Naputci za dobre grafove

- Ne zatrpavajte graf nepotrebnim dodacima (uzorci, 3D efekti, nepotrebne legende) – što jednostavnije to bolje
- Uključite sve potrebne informacije (označite obje osi). Simboli, boje i uzorci trebaju biti definirani u legendi (na slici ili ispod slike). Rezultati na grafu moraju se moći razumjeti bez čitanja glavnog teksta.
- Budite oprezni s korištenjem boja u grafovima (trebaju biti podobne za crno-bijelo printanje i za daltoniste) .
- Nemojte pokušavati prikriti ili krivo interpretirati rezultate.

Literatura

- Dunn, P.K. (2019) *Scientific Research Methods: An introduction to quantitative research in science and health.* Available at: https://srm-course.netlify.com (Accessed: November 19, 2024).

Organizacija datoteka i direktorija:

- Picardi, S. (2024) *Reproducible Data Science*. Available at: https://ecorepsci.github.io/reproducible-science/project-organization.html (Accessed: November 19, 2024).
- White, E. *et al.* (2013) "Nine simple ways to make it easier to (re)use your data," *Ideas in Ecology and Evolution*, 6(2). Available at: https://doi.org/10.4033/iee.2013.6b.6.f.
- Breton, C. and Costanzo, L. (2024) *Organizing Your Research Data*. Available at: https://guides.lib.uoguelph.ca/OrganizingYourResearchData (Accessed: November 19, 2024).