SUFINANCIRANO SREDSTVIMA EUROPSKE UNIJE
EUROPSKI POLJOPRIVREDNI FOND ZA RURALNI RAZVOJ –
EUROPA ULAŽE U RURALNA PODRUČJA

Program
RURALNOG RAZVOJA

EQUINOX
CONSULTING

Osnovne statistike i ekonometrija

Statistika

„Statistika je grana matematike koja se bavi **prikupljanjem, organizacijom, analizom, interpretacijom** i **prezentacijom podataka.**”

Uloga/značaj statistike

- Statistika nam omogućava **izvesti smisleni zaključak iz podataka**
- Ona je temelj svih metoda analize podataka
- Da bismo izveli smisleni zaključak iz podataka i analizirali podatke, moramo razumjeti temeljnu **logiku statističke analize** i moramo biti upoznati s različitim **statističkim metodama** koje se koriste u analizi podataka

SUFINANCIRANO SREDSTVIMA EUROPSKE UNIJE
EUROPSKI POLJOPRIVREDNI FOND ZA RURALNI RAZVOJ –
EUROPA ULAŽE U RURALNA PODRUČJA

Dva glavna područja statistike

Deskriptivna statistika

- Organizirati, sažeti, opisati i prezentirati podatke

Inferencijalna statistika

- Generalizirati iz uzorka u populaciju

Deskriptivna i inferencijalna statistika

	Deskriptivna statistika	Inferencijalna statistika
Ciljevi	Describe and summarize existing data	• Make inferences from (sample) data to the population of interest • Detect and measure relationships and causality
Tools	• Summary measures • Grafički prikaz	• Estimation (point, interval) • Testiranje hipoteze • Detecting relationships (correlation, regression)

1. Deskriptivna statistika

Deskriptivna statistika

Ciljevi

- Opisati i sažeti postojeće podatke

Alati

- Zbirne mjere
- Grafički prikaz

Vrste podataka

- Primarni i sekundarni podaci
- Opservacijski i eksperimentalni podaci
- Popisni i uzorkovani podaci
- Transverzalni podaci, podaci o vremenskom nizu i panel podaci

Vrste podataka – Popisni/Uzorkovani

Popisni podaci: Imamo podatke od svih članova populacije

Uzorkovani podaci: Imamo podatke iz podskupine populacije

	Popisni podaci	Uzorkovani podaci
Parametar interesa	Izravno promatranje	Može se samo procijeniti
Točnost rezultata	Visoka	Ovisi o veličini uzorka i reprezentativnosti podataka
Trošak	Visok	Nizak, ovisi o veličini uzorka

Vrste podataka – Primarni/Sekundarni

Primarni izvori podataka:

- Vlastito prikupljanje podataka

Sekundarni izvori podataka:

- Nacionalne baze podataka
- EU baze podataka
- OECD baze podataka
- Podaci iz javne uprave itd.

Vrste podataka- Opservacijski/Eksperimentalni

Eksperimentalni podaci:

- *Eksperiment* obuhvaća jedan ili više tretmana na sudionicima.
- „Tretman" je pod kontrolom istraživača
  - Istraživač odlučuje tko će biti u tretiranim i ne-tretiranim (kontrolnim) skupinama
  - Također odlučuje o količini tretmana (ako je primjenjivo)
- Dobar eksperiment osmišljen je na način koji nam omogućuje jasne usporedbe između tretiranih i kontrolnih skupina, tako da može pomoći u uspostavljanju uzročno-posljedičnih veza
- Eksperimenti nisu uvijek mogući zbog financijskih i etičkih ograničenja
  - Primjer: Studije o učinku na pušenje uvijek su opservacijske, jer ne možemo nasumično narediti ljudima da puše radi istraživanja

Opservacijski podaci :

- Podaci o pojedincima se prikupljaju na način koji ne utječe na njih (npr. ankete)
- Budući da ne možemo kontrolirati varijablu tretmana, možemo imati samo posredne dokaze za uzročnost
  - Primjer: Možemo naći snažnu povezanost između pušenja i raka pluća, ali ne možemo biti sigurni da ne postoji skriveni zbunjujući faktor koji čini ljude pušačima i čini ih bolesnima
- Važno je da je razlika između tretirane i kontrolne skupine samo tretman (dok su inače slične), a mi moramo kontrolirati sve moguće zbunjujuće faktore u opservacijskoj studiji

Vrste podataka – Transverzalni/Vremenski niz/Panel

Transverzalni podaci

- U jednom vremenskom trenutku
- Više promatranih entiteta

Podaci o vremenskom nizu

- Višestruke vremenske točke
- Jedan promatrani entitet

Panel podaci

- Višestruke vremenske točke
- Više promatranih entiteta

Zbirni Transverzalni podaci

Modeli podataka

(a) Spatial panel data models

i	*t*	*y*	*x*
1	1	$y_{1,1}$	$x_{1,1}$
1	2	$y_{1,2}$	$x_{1,2}$
1	3	$y_{1,3}$	$x_{1,3}$
2	1	$y_{2,1}$	$x_{2,1}$
2	2	$y_{2,2}$	$x_{2,2}$
2	3	$y_{2,3}$	$x_{2,3}$
3	1	$y_{3,1}$	$x_{3,1}$
3	2	$y_{3,2}$	$x_{3,2}$
3	3	$y_{3,3}$	$x_{3,3}$

(b) Cross-sectional data models

i	*t*	*y*	*x*
1	1	$y_{1,1}$	$x_{1,1}$
2	1	$y_{2,1}$	$x_{2,1}$
3	1	$y_{3,1}$	$x_{3,1}$

(c) Time series data models

i	*t*	*y*	*x*
1	1	$y_{1,1}$	$x_{1,1}$
1	2	$y_{1,2}$	$x_{1,2}$
1	3	$y_{1,3}$	$x_{1,3}$

Slučajne varijable

Varijable: Logički skup atributa

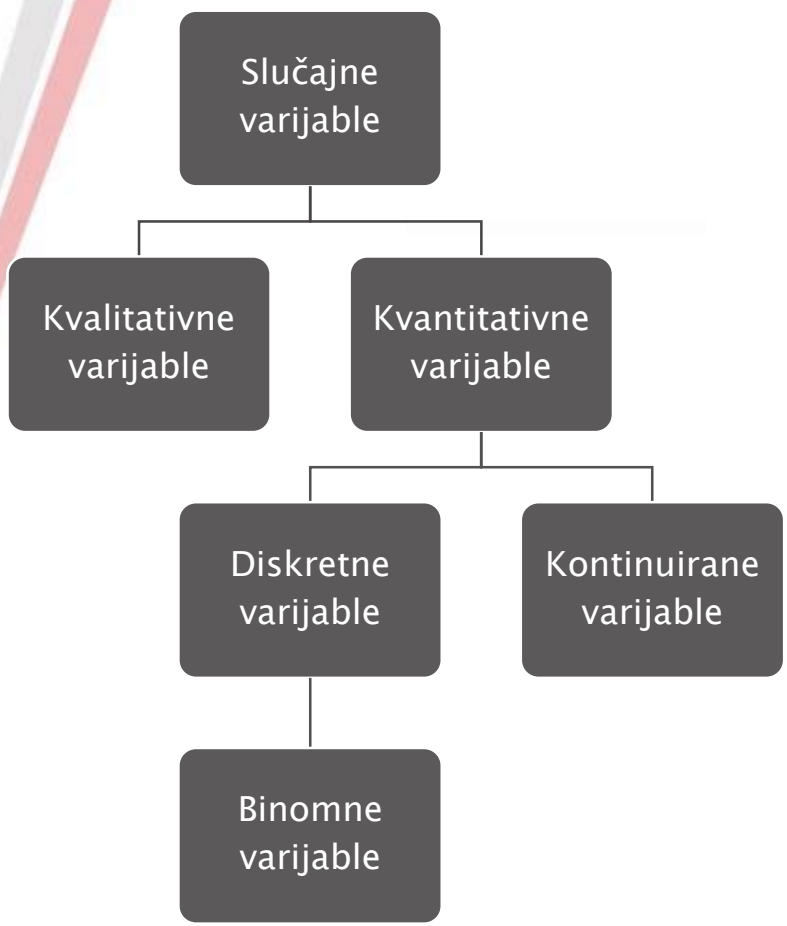
Slučajne varijable: karakteristika, mjerenje ili brojanje koje se nasumično mijenja prema određenom skupu ili uzorku

- Slučajne: ne znamo točan ishod, ali znamo da se izvjesni ishodi događaju s izvjesnom vjerojatnošću
- Notacija: *X, Y, Z,* itd.

- Vrste varijabli:
  - Kvalitativne/kvantitativne
  - **Kategoričke/numeričke**

  - **Metode koje imamo na raspolaganju ovise o vrsti varijable!**

Vrste varijabli



Usporedba kategoričkih in numeričkih varijabli

Kategorička varijabla

Osobe	Teško štedi novac
Anna	Jako
Bella	Nimalo
Cecil	Jako
Dan	Jako
Evelyn	Donekle

Numerička varijabla

Osobe	Mjesečno ušteđeni dolari
Anna	0
Bella	30
Cecil	10
Dan	5
Evelyn	20
Ukupno	**65**

Numeričke varijable

- **Diskretne slučajne varijable**: mogući ishodi mogu se navesti pomoću cijelih brojeva
- **Kontinuirane slučajne varijable**: mogući ishodi mogu se opisati samo pomoću intervala realnih brojeva (npr. svi realni brojevi od nule do beskonačnosti)

Primjer slučajnih varijabli

Diskretne slučajne varijable:

- *Konačne*: možemo navesti sve moguće ishode
  - Primjer: Broj političara koji su glasovali za zakon
- *Prebrojive beskonačne*: možemo navesti moguće ishode, ali oni nemaju specifičan kraj
  - Primjer: Broj nesreća na raskrižju tijekom 10 godina

Kontinuirane slučajne varijable:

- *Neprebrojivo beskonačne*: ima previše mogućih ishoda, obično predstavljaju mjerenja
  - Primjer: Težina novorođenčeta, razina smoga u Zagrebu na određeni dan

Razine mjerenja (Stevens)

Nominalna skala

Ordinalna skala

Intervalna skala

Skala omjera

Metode koje imamo na raspolaganju ovise o vrsti varijable!

Deskriptivne metode

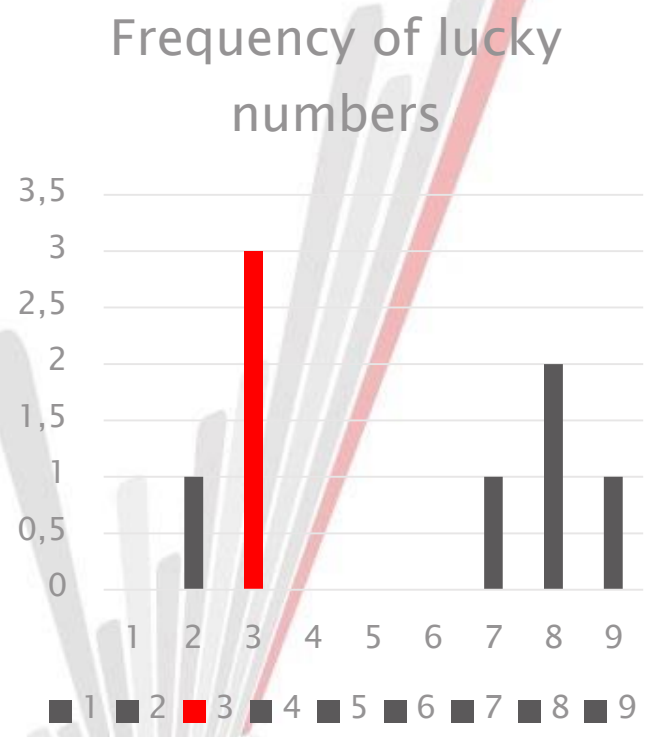
	Kategoričke varijable	Numeričke varijable
Mjere centralne tendencije	Mod	• Srednja vrijednost • Postoci (medijan)
Mjere disperzije	(Relativne) frekvencije	• Minimalno, maksimalno • (Interkvartilni) raspon • Standardna devijacija
Grafički prikaz	Strukturni krug, grafikon stupaca	• Histogram • Kutijasti dijagram

Kategoričke varijable

- Mod
- Frekvencija, relativna frekvencija, tablice frekvencija (unakrsne tablice)
- Strukturni krugovi
- Stupčasti grafikoni

Deskriptivne mjere – Mod

- Vrijednost koja se najčešće pojavljuje



Ime	Sretan broj
Anna	8
Bob	2
Celia	3
Dennis	3
Eve	7
Fiona	9
George	3
Horatius	8

Deskriptivne mjere – Frekvencija

Frekvencija ishoda: koliko se puta ishod pojavio u eksperimentu ili studiji

Relativna frekvencija ishoda: postotak puta kada se ishod pojavio u eksperimentu ili studiji

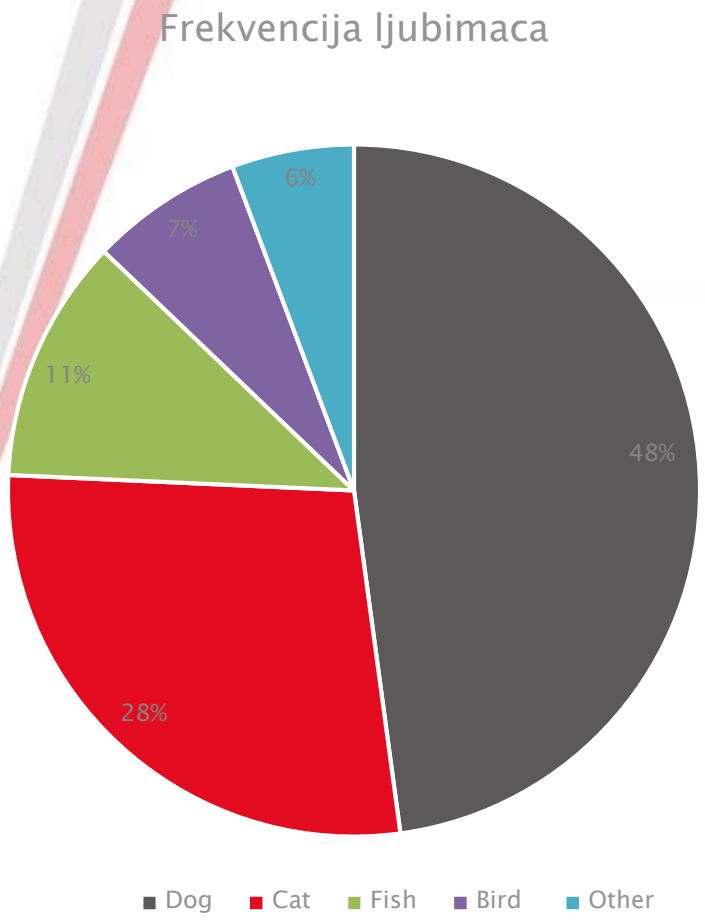
Vrsta ljubimca	Frekvencija	Relativna frekvencija
Pas	67	48%
Mačka	39	28%
Ptica	10	7%
Riba	16	11%
Ostali	8	5%
Ukupno	**140**	**100%**

Deskriptivne mjere – Frekvencija

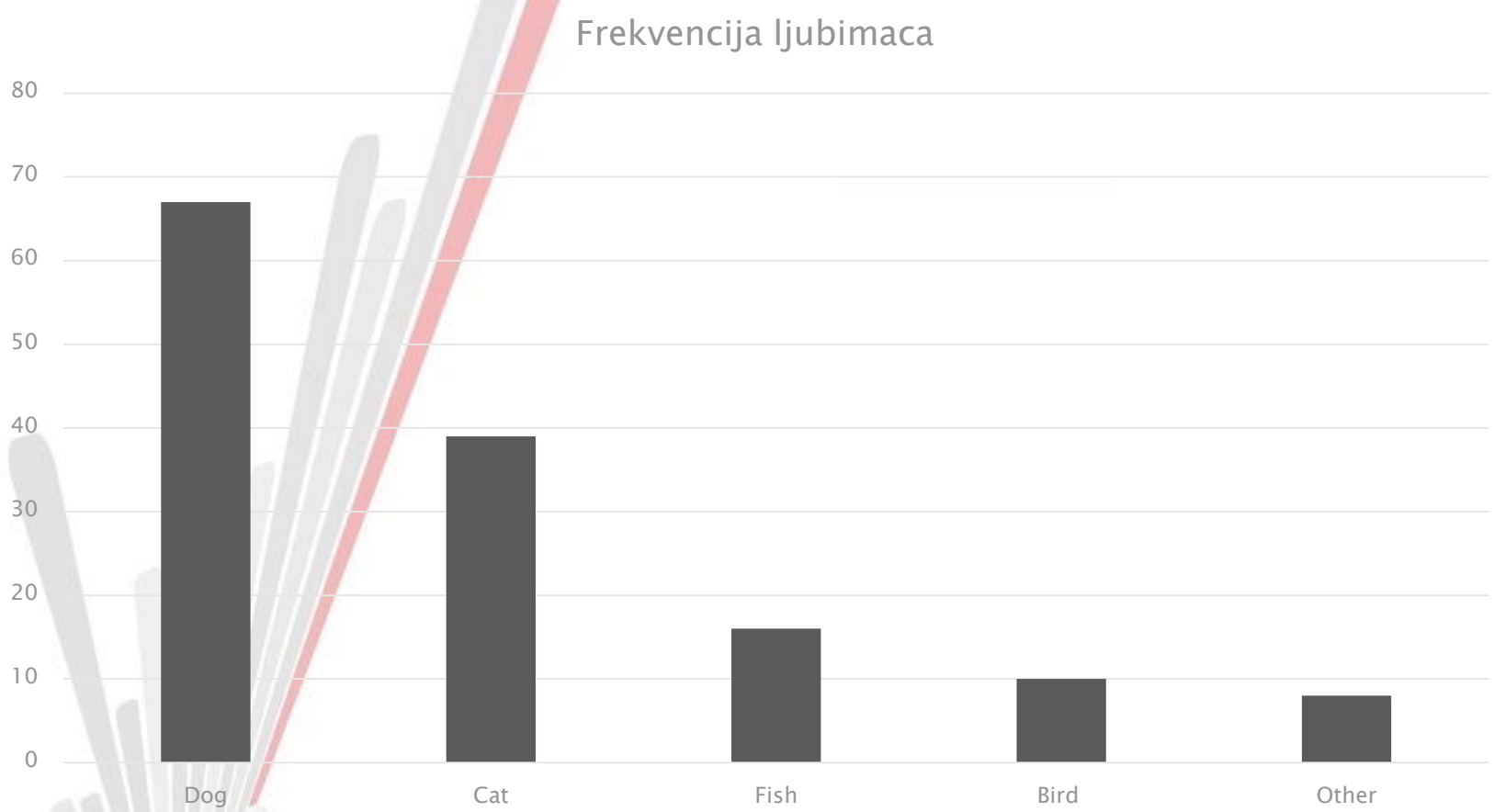
Tablica s dvosmjernom frekvencijom (unakrsna tablica)

Vrsta ljubimca	Spol		Ukupno
	Muški	Ženski	
Pas	38	29	67
Mačka	14	25	39
Ptica	4	6	10
Riba	12	4	16
Ostali	6	2	8
Ukupno	74	66	140

Grafički prikaz – Stukturni krugovi



Grafički prikaz – Stupčasti grafikoni



Numeričke varijable

- Mjere centralne tendencije
  - Srednje
  - Medijan (postoci)
- Mjere disperzije
  - Minimalno, maksimalno
  - Raspon, interkvartilni raspon
  - Standardna devijacija
- Grafički prikaz
  - Kutijasti dijagram
  - Histogram

Deskriptivne metode – srednja vrijednost

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

- Srednja vrijednost s Kobe: 7,029,082 U
- Srednja vrijednost bez Kobe: 5,695,307

Osjetljivost na netipične vrijed

Table 5-2	Salaries for L.A. Lakers NBA Players (2009–2010)
Player	***Salary ($)***
Kobe Bryant	23,034,375
Pau Gasol	16,452,000
Andrew Bynum	12,526,998
Lamar Odom	7,500,000
Ron Artest	5,854,000
Adam Morrison	5,257,229
Derek Fisher	5,048,000
Sasha Vujacic	5,000,000
Luke Walton	4,840,000

Player	***Salary ($)***
Shannon Brown	2,000,000
Jordan Farmar	1,947,240
Didier Ilunga-Mbenga	959,111
Josh Powell	959,111
Total	**91,378,064**

Deskriptivne mjere – Percentili

Percentil: mjera koja pokazuje vrijednost ispod koje se spada određeni postotak opažanja u skupini opažanja

25. percentil (Q1): mjera koja pokazuje vrijednost ispod koje spada 25% opažanja

50. percentil – **Medijan** (Q2): mjera koja pokazuje vrijednost ispod koje spada 50% opažanja

75. percentil (Q3): mjera koja pokazuje vrijednost ispod koje spada 75%

Deskriptivne mjere – Percentili

Primjer:

Medijan godišnjeg prihoda kućanstava u 28 zemalja EU u 2017. godini iznosio je 16563 EUR.

To znači da je 50% kućanstava imalo godišnji prihod manji od 16 653 EUR (a 50% veći).

Deskriptivne mjere – Mjere disperzije

Minimum: najniža vrijednost varijable

Maksimum: najviša vrijednost varijable

Raspon: $Maximum - Minimum$

Interkvartilni raspon: $Q_3 - Q_1$

- Sredina 50% vrijednosti – izostavlja netipične vrijednosti

Deskriptivne mjere – Standardna devijacija

- „Prosječna udaljenost od srednje vrijednosti”
- Ima iste jedinice kao i izvorni podaci

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

Deskriptivne mjere – Standardna devijacija

- Srednja vrijednost s Kobe: 7,029,082
- SD s Kobe: 6.567.000

- Srednja vrijednost bez Kobe: 5,695,3
- SD s Kobe: 4.671.508

Table 5-2 Salaries for L.A. Lakers NBA Players (2009–2010)

Player	*Salary ($)*
Kobe Bryant	23,034,375
Pau Gasol	16,452,000
Andrew Bynum	12,526,998
Lamar Odom	7,500,000
Ron Artest	5,854,000
Adam Morrison	5,257,229
Derek Fisher	5,048,000
Sasha Vujacic	5,000,000
Luke Walton	4,840,000

Player	*Salary ($)*
Shannon Brown	2,000,000
Jordan Farmar	1,947,240
Didier Ilunga-Mbenga	959,111
Josh Powell	959,111
Total	**91,378,064**

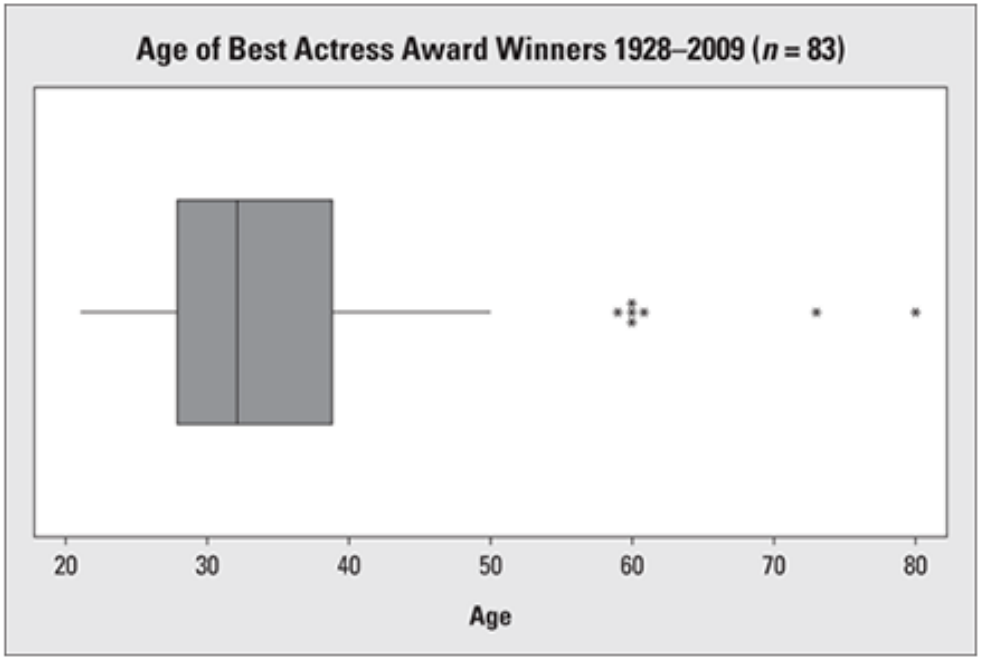
Deskriptivne mjere – Sažetak od pet brojeva

Descriptive Statistics: Time between Eruptions

Variable	Total Count	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Time between	222	71.009	12.799	42.000	60.000	75.000	81.000	95.000	21.000

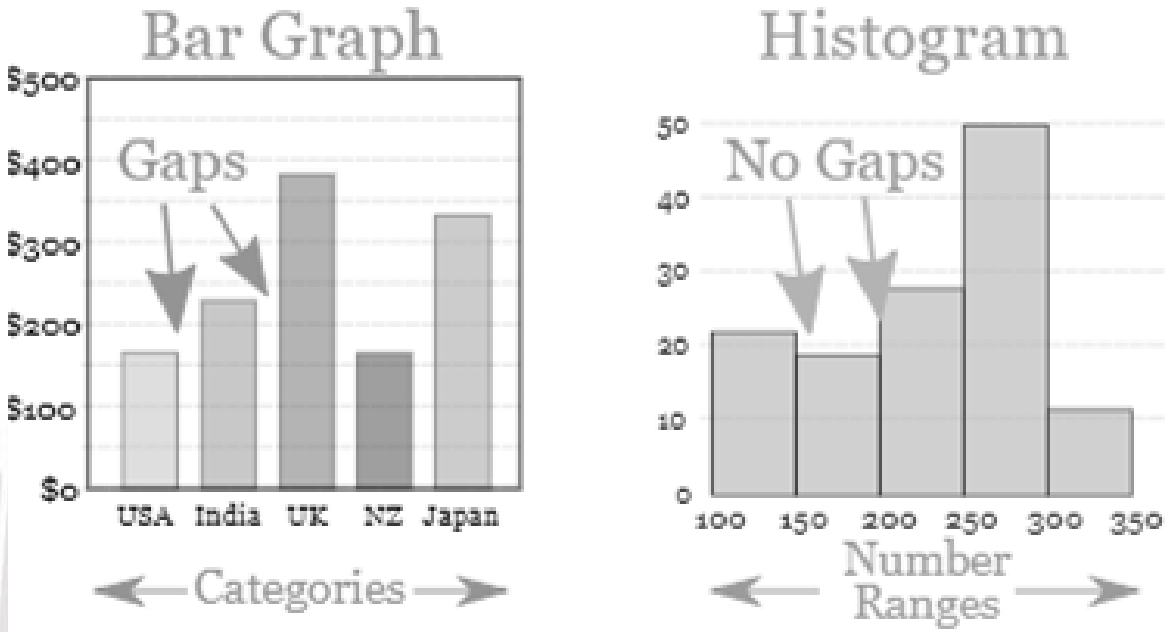
Grafički prikaz – Kutijasti dijagram

Kutijasti dijagram: jednodimenzionalni grafikon numeričkih podataka na temelju sažetka od pet brojeva



Grafički prikaz – Histogram

Histogram točan prikaz raspodjele numeričkih podataka.

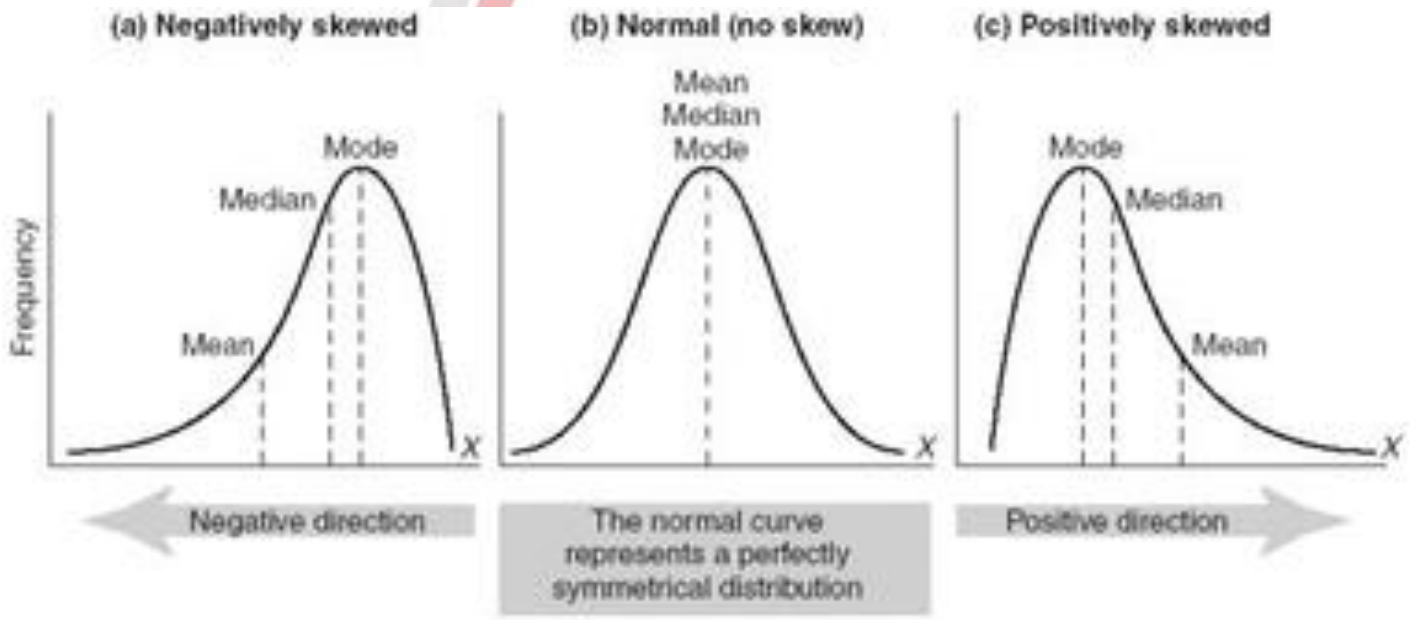


Grafički prikaz – Histogram

Vrste histograma:

- Simetričan: ako ga izrežete na pola, lijeva i desna strana zrcale jedna drugu
- Ukošen nadesno: rep ide nadesno
- Ukošen nalijevo: rep ide nalijevo

Grafički prikaz – Histogram



Sažetak deskriptivne statistike

- Cilj deskriptivne statistike je opisati i sažeti podatke.
- Metode koje imamo na raspolaganju ovise o vrsti varijable.

	Kategoričke varijable	Numeričke varijable
Mjere centralne tendencije	Mod	• Srednja vrijednost • Percentili (medijan)
Mjere disperzije	(Relativne) frekvencije	• Minimalno, maksimalno • (Interkvartilni) raspon • Standardna devijacija
Grafički prikaz	Strukturni krug, grafikon stupaca	• Histogram • Kutijasti dijagram

2. Distribucije vjerojatnosti

Vjerojatnost

Vjerojatnost ishoda: postotak vremena u kojem se očekuje da će se ishod dogoditi, kada se osnovni proces izvodi iznova i iznova, samostalno i pod istim uvjetima

- Šanse su između 0 i 1 (0% i 100%)
- Ako slučajno odaberete iz populacije, svi članovi imaju istu šansu da budu odabrani

Neovisnost: Dva ispitivanja su neovisna ako šanse za prvo ispitivanje ne utječu na šanse za drugo ispitivanje

Primjer:

Ako bacim novčić, imam 50% šanse da dobijem glavu. Naravno, ako bacim novčić, to će biti ili glava ili neće. Ali ako ponovim bacanje 100 puta, učestalost glava će biti blizu 50.

Očekivana vrijednost

Očekivana vrijednost slučajne varijable: dugoročna prosječna vrijednost ponavljanja u istom eksperimentu

- Primjer: Očekivana vrijednost kod bacanja šestostrane kockice je 3.5, jer prosjek svih brojeva koji se pojavljuju iznosi 3.5 kako se broj bacanja približava beskonačnosti

Izračun:

- Diskretna varijabla:
  - $E(X) = \sum X_i P(X_i)$
  - $E(X^2) = \sum X_i^2 P(X_i)$
- Kontinuirana varijabla:
  - $E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$
  - $E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx$

Standardna devijacija:

$$D(X) = \sqrt{E(X^2) - E^2(X)}$$

Distribucije vjerojatnosti

Distribucija vjerojatnosti: Matematička funkcija koja osigurava vjerojatnost pojave različitih mogućih ishoda varijable.

Distribucije vjerojatnosti

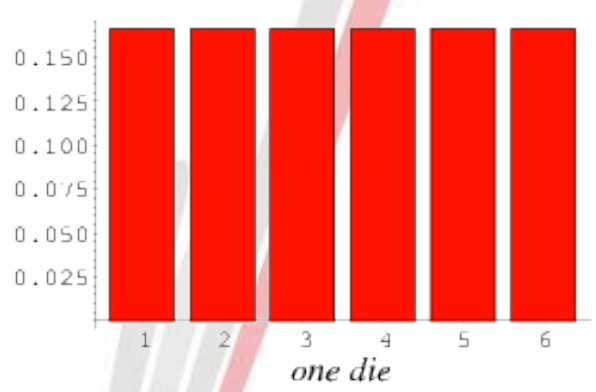
Diskretne slučajne varijable: mogući ishodi mogu se navesti pomoću cijelih brojeva

- Vjerojatnosti svakog ishoda možemo odrediti pomoću teorije vjerojatnosti

Kontinuirane slučajne varijable: mogući ishodi mogu se opisati samo pomoću intervala realnih brojeva (npr. svi realni brojevi od nule do beskonačnosti)

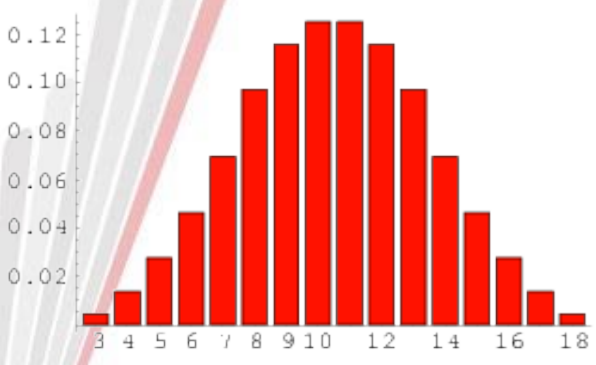
- Ne možemo odrediti vjerojatnost svakog mogućeg ishoda (jer postoji beskonačan ishod)
- Možemo samo odrediti vjerojatnost postojanja u intervalu pomoću integracije

Distribucija – Diskontinuirana varijabla

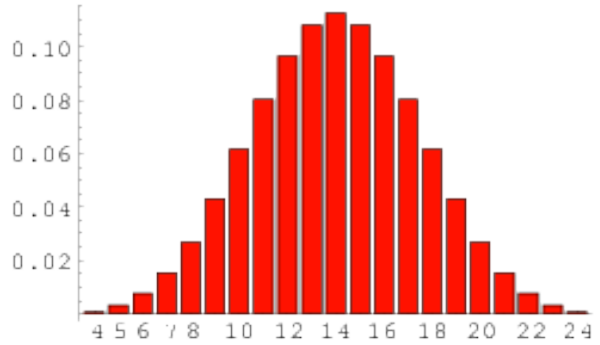


one die

two dice



three dice



four dice

Distribucije vjerojatnosti – diskretne varijable

Diskretna slučajna varijabla: može poprimiti određeni skup mogućih ishoda i svaki od tih rezultata ima određenu vjerojatnost da će se dogoditi

Vjerojatnost pojave određenog ishoda:

$$p(x) \leq 1$$

Distribucija vjerojatnosti: popis ili funkcija koja prikazuje sve moguće vrijednosti diskretne slučajne varijable zajedno s njihovim vjerojatnostima

Distribucije vjerojatnosti – diskretne varijable

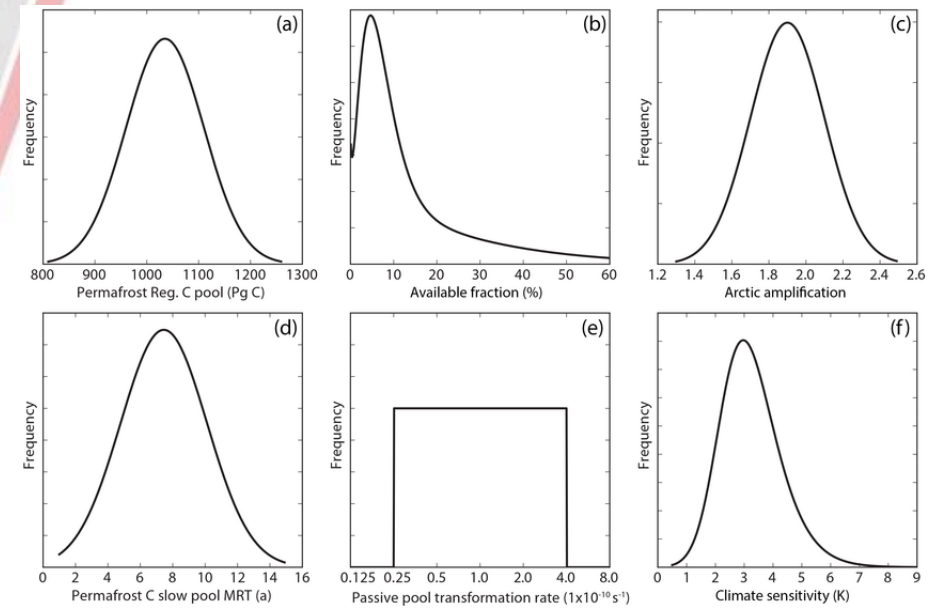
Table 8-1 Probability Distribution for X = Number of Dogs Owned by Apartment Renters

x	$p(x)$
0	0.50
1	0.40
2	0.07
3	0.03

Distribucija – Kontinuirana varijabla

Distribucija kontinuiranih varijabli može imati različite oblike

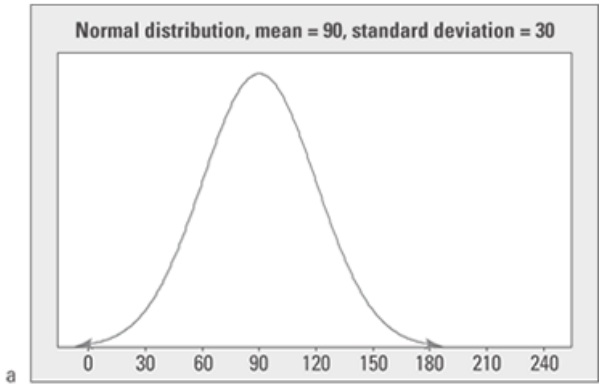
Najpoznatija distribucija je normalna distribucija



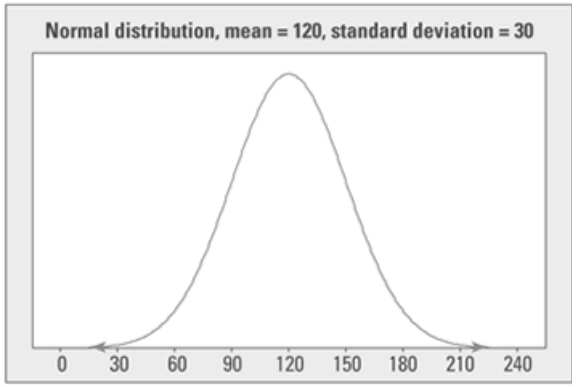
Distribucije vjerojatnosti – Normalna distribucija

Normal distribucija: zvonasta krivulja

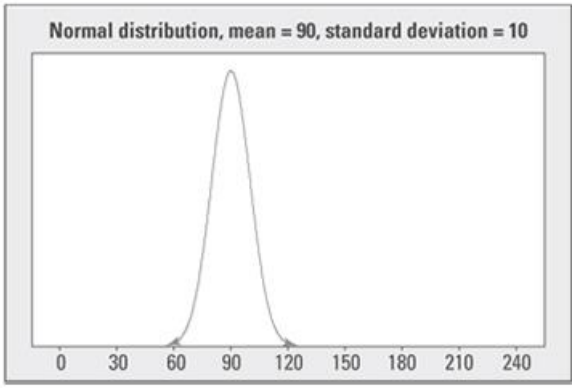
- Sredina: μ
- Standardna devijacija: σ



a



b



c

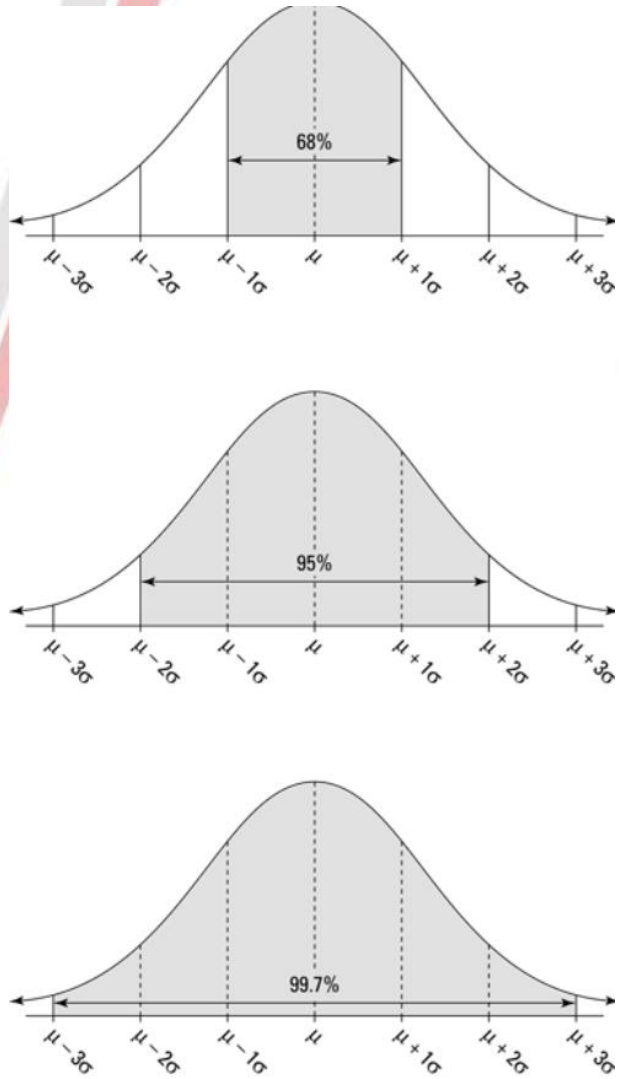
Distribucije vjerojatnosti – Normalna distribucija

Svojstva normalne distribucije:

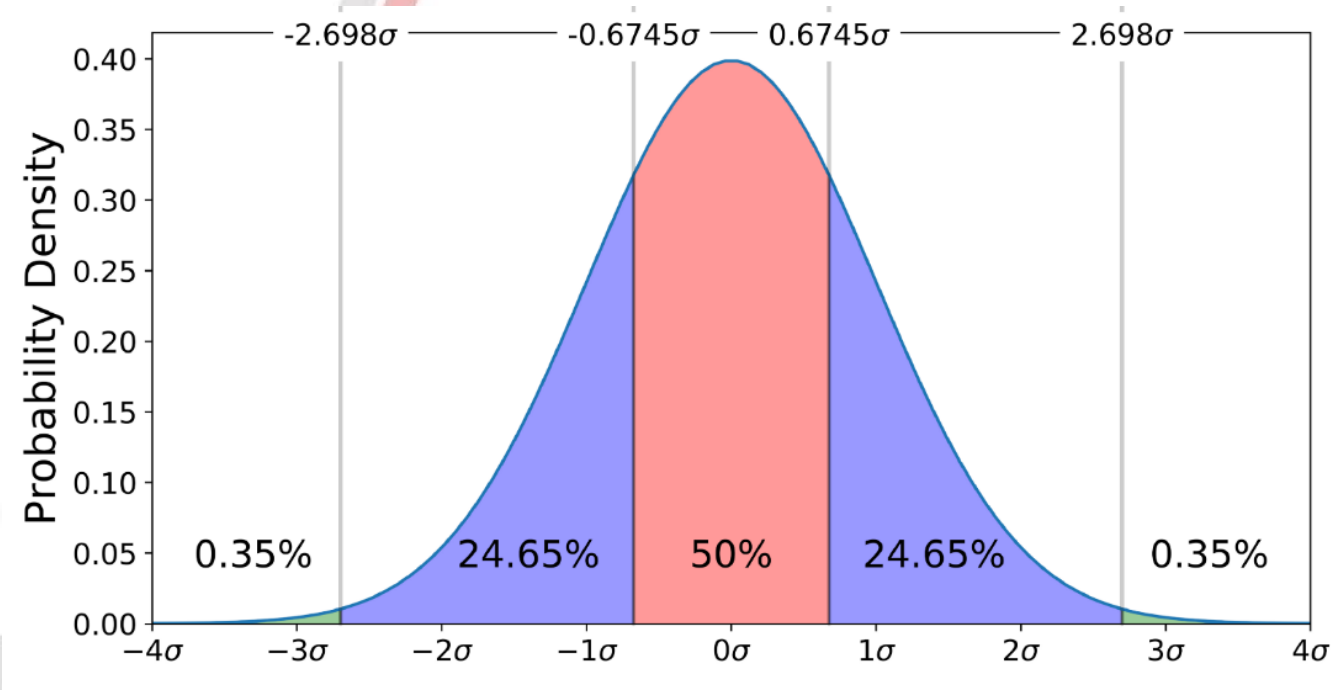
- Simetrični oblik
- Središnja vrijednost i medijan su jednaki i leže u sredini distribucije
- Njena standardna devijacija mjeri udaljenost od središnje vrijednosti do točke infleksije (gdje se krivulja mijenja iz konveksne u konkavnu)

- Empirijsko pravilo:
  - 68% svih vrijednosti leži unutar 1 standardne devijacije od srednje vrijednosti
  - 95% svih vrijednosti leži unutar 2 standardne devijacije od srednje vrijednosti
  - 99,7% svih vrijednosti leži unutar 3 standardne devijacije od srednje vrijednosti

Empirijsko pravilo:



Empirijsko pravilo:

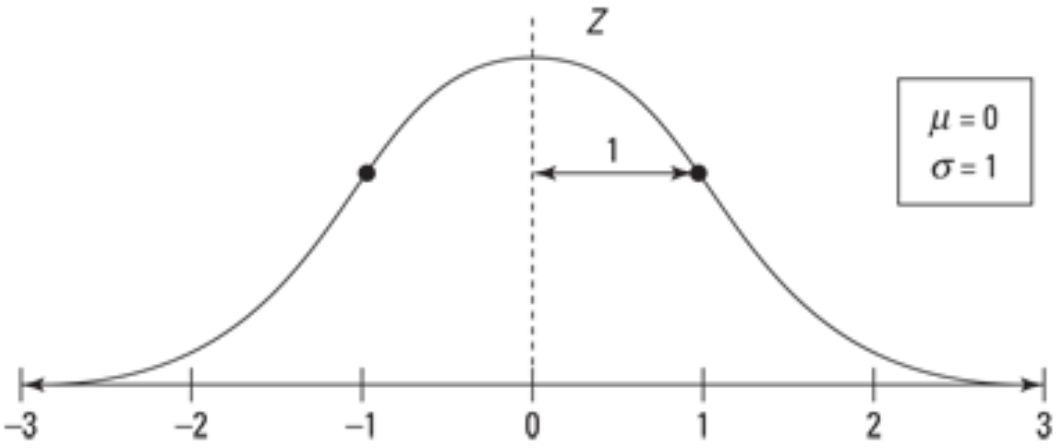


Izračunavanje vjerojatnosti za kontinuirane varijable

- Vjerojatnosti za bilo kakvu kontinuiranu distribuciju nalaze se pronalaženjem područja ispod krivulje (s integracijom).
- Normalne distribucije se razlikuju (različite središnje vrijednosti i standardne devijacije), tako da bi dobivanje vjerojatnosti zahtijevalo mnogo računanja (integraciju)
- Standardna normalna distribucija: možemo normirati bilo koju normalnu distribuciju u z-distribuciju, čije su vjerojatnosti već izračunate

Distribucije vjerojatnosti – Standardna normalna (Z-) distribucija

Z-distribucija: normalna distribucija sa srednjom vrijednošću 0 i standardnom devijacijom 1



Distribucije vjerojatnosti – Standardna normalna (Z-) distribucija

Z-vrijednost: broj standardne devijacije gdje određena vrijednost leži iznad ili ispod srednje vrijednosti

$$z = \frac{x - \mu}{\sigma}$$

Primjer: Bob ima 80 bodova na oba ispita iz matematike (koji ima srednju vrijednost 70 i standardnu devijaciju 10) i engleskog (koji ima srednju vrijednost 85 i standardnu devijaciju 5).

Standardizacija Bobovog rezultata iz matematike: $z = \frac{80-70}{10} = 1$

Njegova matematička ocjena je 1 standardna devijacija iznad prosjeka razreda.

Standardizacija Bobovog rezultata iz engleskog: $z = \frac{80-85}{5} = -1$

Njegova ocjena iz engleskog je 1 standardna devijacija ispod prosjeka.

Pronalaženje vjerojatnosti za normalnu distribuciju – koraci

1. Prevedite problem u jedno od sljedećeg:
   1. p(X < a)
   2. p(X > b)
   3. p(a < X < b).
2. Standardizirajte a (i/ili b) u z-vrijednost pomoću z-formule:

$$z = \frac{x - \mu}{\sigma}$$

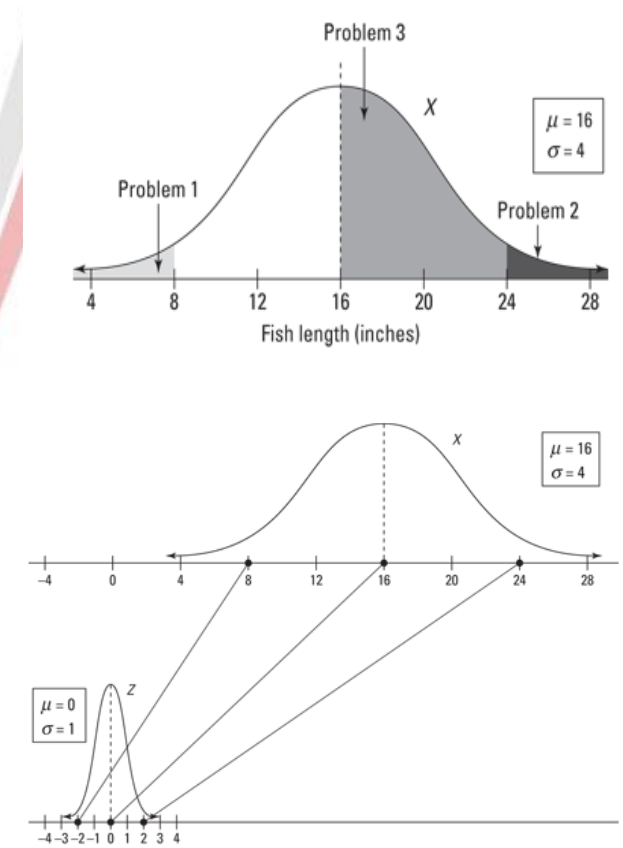
3. Potražite z-ocjenu u Z-tablici (ili izračunajte pomoću softvera)

4a. Ako trebate „manje od“ vjerojatnosti – p (X < a) – gotovi ste

4b Ako želite vjerojatnost „veće od“ – to jest, p(X > b) – uzmite jedan minus rezultat iz 4. koraka.

4c Ako vam je potrebna vjerojatnost „između dvije vrijednosti“ – to jest, p(a < X < b) – učinite korake 1-4 za *b* (veću od dvije vrijednosti) i opet za *a* (manju od dvije vrijednosti), te potom oduzmite rezultate.

Pronalaženje vjerojatnosti za normalnu distribuciju – Ilustracija



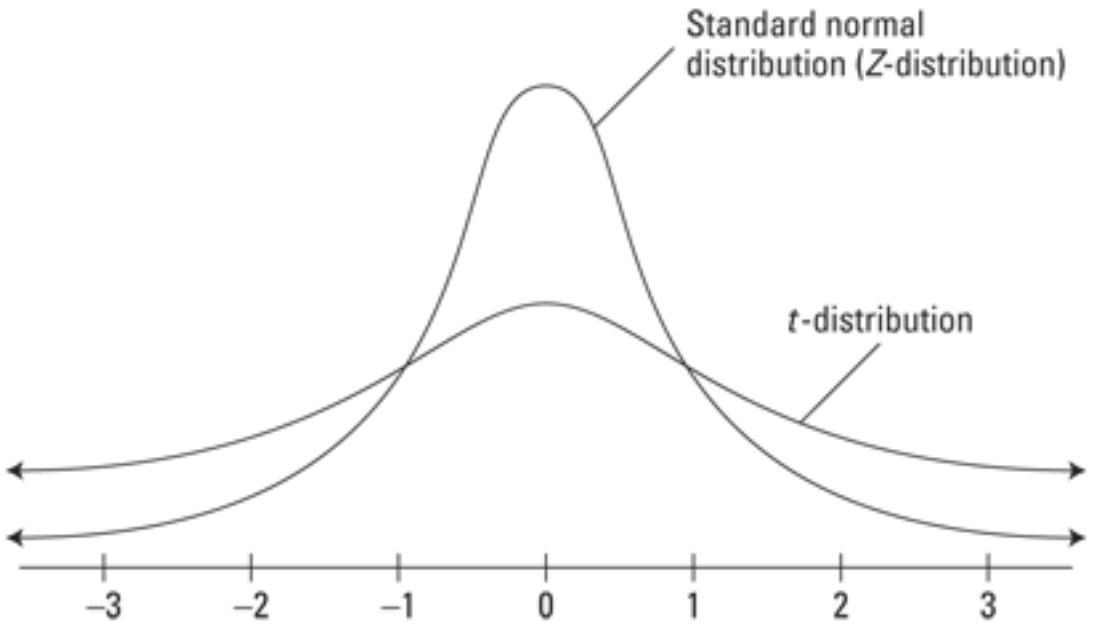
Distribucije vjerojatnosti – Standardna normalna distribucija

Za korištenje Z-distribucije trebamo:

- Relativno veliku veličinu uzorka (otprilike n>30)
- Standardnu devijaciju populacije – obično nije poznata

Distribucije vjerojatnosti – T-distribucije

T-distribucija: normalna distribucija sa srednjom vrijednošću 0 ali ravnomjernijom od standardne normalne distribucije



Distribucije vjerojatnosti – T-distribucije

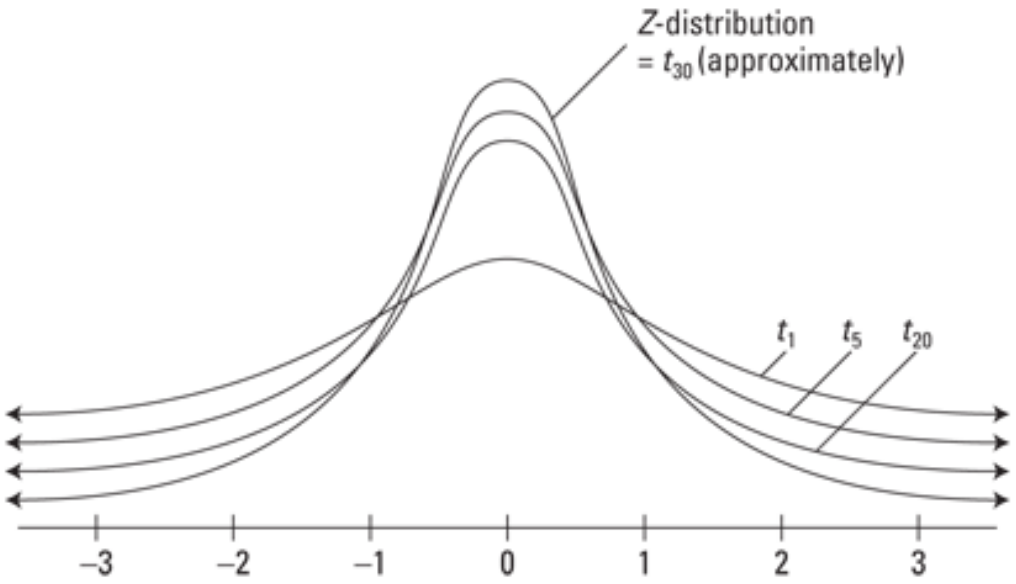
Višestruke t-distribucije, ovisno o veličini uzorka

Stupnjevi slobode: $n - 1$,

gdje je n: veličina uzorka

- T-distribucije temeljene na manjim veličinama uzorka imaju veće standardne devijacije
- Što je veća veličina uzorka, to više t-distribucija izgleda kao standardna normalna distribucija (slično od oko n = 30)
- T-distribucije koristimo ako je veličina uzorka mala ili ako je standardna devijacija populacije nepoznata (što je obično slučaj)

Distribucije vjerojatnosti – t-distribucija



Distribucije vjerojatnosti – Sažetak

Distribucija vjerojatnosti: Matematička funkcija koja osigurava vjerojatnost pojave različitih mogućih ishoda varijable.

Ako znamo distribuciju varijable, možemo izračunati vjerojatnost nastanka određenih ishoda.

Diskretne varijable:

- Vjerojatnosti svakog ishoda možemo izračunati pomoću teorije vjerojatnosti

Kontinuirane varijable:

- Pomoću integracije, možemo samo izračunati vjerojatnost da smo u intervalu

Distribucije vjerojatnosti – Sažetak

Kontinuirane varijable:

- Pomoću integracije, možemo samo izračunati vjerojatnost da smo u intervalu
- Računanje vjerojatnosti za različite distribucije zahtijeva mnogo računanja
- Ako imamo normalno distribuiranu varijablu, možemo to učiniti lakše pomoću:
  - Z-distribucije, ako je veličina uzorka dovoljno velika i poznata je standardna devijacija populacije
  - T-distribucije, ako je veličina uzorka mala i/ili je standardna devijacija populacije nepoznata

3. Inferencijalna statistika

Inferencijalna statistika

Ciljevi:

- Izvođenje zaključke iz (uzorka) podataka za populaciju od interesa
- Otkrivanje i mjerenje odnosa i uzročnosti

Metode:

- Procjena parametara
- Testiranje hipoteze
- Otkrivanje odnosa (korelacija, regresija)

Parametar – Statistika – Procjenitelj

- Parametar:
  - Jedinstvena mjera nekog atributa **populacije** (npr. aritmetička sredina populacije varijable)
- Statistika:
  - Jedinstvena mjera nekog atributa**uzorka** (npr. sredina uzorka varijable)
- Procjenitelj:
  - Statistika koja se koristi za procjenu (populacijskog) parametra

Uzorak

- **Statistički uzorak:** Podskupina populacije
- Iz uzorka želimo izvesti zaključak o populaciji

- **Pogreška uzorkovanja:**
  - razlika između statističkog uzorka korištenog za procjenu parametra populacije i stvarne ali nepoznate vrijednosti parametra populacije
  - Ove varijacije mogućih uzorkovanih vrijednosti statistike se teoretski mogu izraziti kao pogreške uzorkovanja, iako je u praksi točna pogreška uzorkovanja u pravilu nepoznata

Primjer pogreške uzorkovanja

Primjer:

Želimo znati prosječnu visinu hrvatskog naroda. Odabiremo uzorak od 100.

Ako odaberemo 100 najnižih osoba, tada će prosječna visina u uzorku (statistički uzorak) biti mnogo manja od populacijskog parametra.

Vrste uzoraka

Slučajni uzorci (uzorci vjerojatnosti):

1. Svaki član populacije ima šanse da bude izabran
2. Za svakog člana populacije možemo utvrditi šansu da bude odabran – to možemo učiniti samo ako postoji neki slučajni element

- Možemo procijeniti pogrešku uzorkovanja

Ne-slučajni uzorci (nevjerojatni uzorci):

- Dva uvjeta slučajnih uzoraka nisu ispunjena
- Ne možemo procijeniti pogrešku uzorkovanja

Ne-slučajni uzorci

- Vrste ne-slučajnih uzoraka
- Što učiniti ako imamo ne-slučajni uzorak

U nastavku ćemo uvijek pretpostaviti slučajni uzorak!

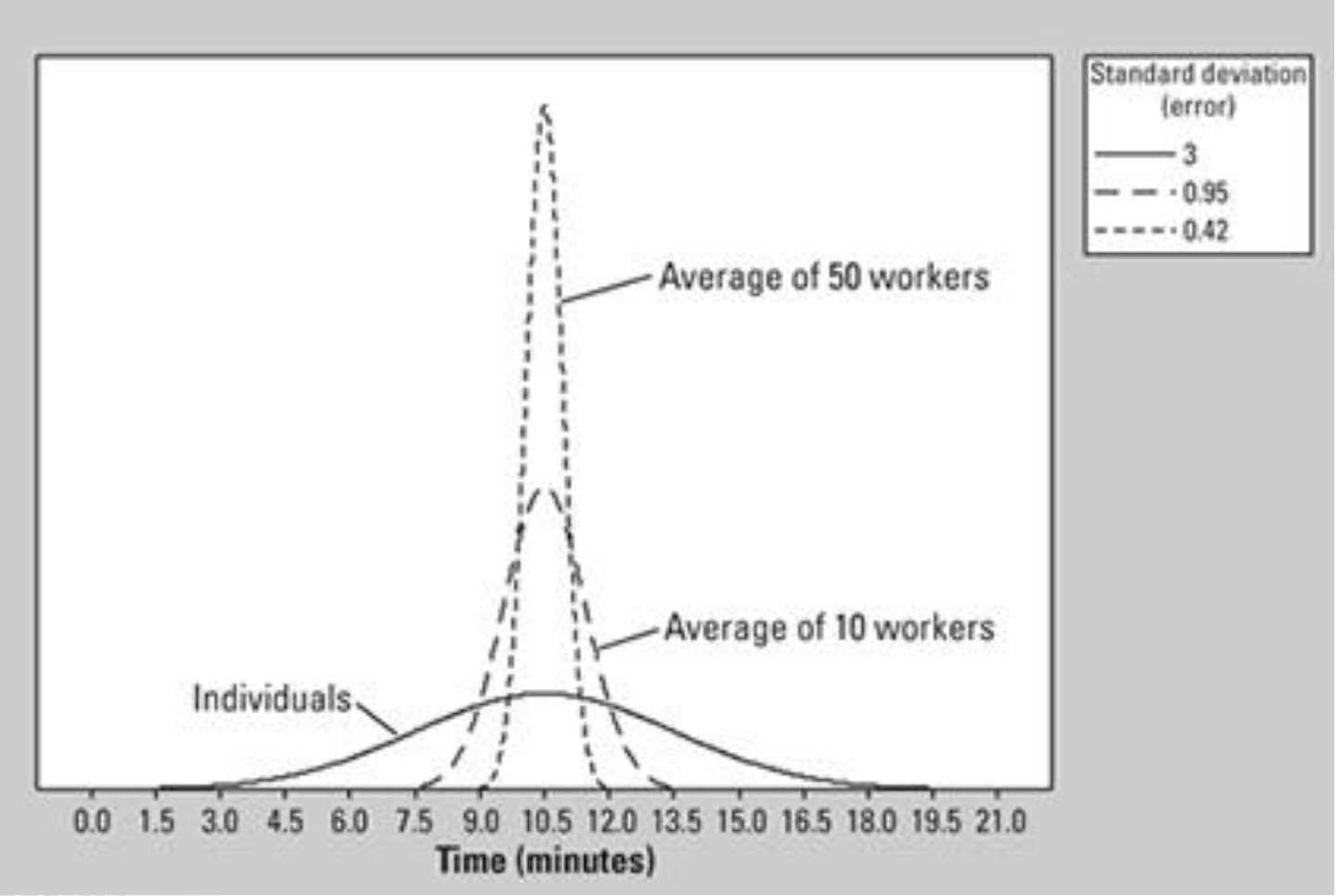
Uzorak

Standardna pogreška: mjeri varijabilnost u statistici uzorka

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

- SE se smanjuje kako *n* raste
- SE se povećava kako SD raste

Uzorak


Standard deviation (error)
3
0.95
0.42
Average of 50 workers
Average of 10 workers
Individuals
0.0 1.5 3.0 4.5 6.0 7.5 9.0 10.5 12.0 13.5 15.0 16.5 18.0 19.5 21.0
Time (minutes)

Distribucija uzorkovanja

Centralni granični teorem: ako imamo slučajni uzorak i veličina uzorka je dovoljno velika, distribucija uzorkovanja (parametra) je blizu normalnog

Empirijsko pravilo: ako imamo normalne distribucije,

- 68% svih vrijednosti leži unutar 1 standardne devijacije od srednje vrijednosti
- 95% svih vrijednosti leži unutar 2 standardne devijacije od srednje vrijednosti
- 99,7% svih vrijednosti leži unutar 3 standardne devijacije od srednje vrijednosti

Distribucija uzorkovanja

Budući da je standardna devijacija raspodjele uzorka nepoznata, možemo koristiti t-distribucije za izračunavanje:

- Margina pogreške: mjera koliko se blizu očekuje statistika uzorka od parametra populacije
- Vjerojatnost odabira uzorka s danom središnjom vrijednosti, ako pretpostavimo populacijsku sredinu

Procjena parametara

Cilj: procijeniti nepoznati parametar populacije

Procjena točke: procjena vrijednosti parametra

Procjena intervala: procijeniti dvije vrijednosti između kojih leži parametar populacije

Procjena parametara

- Ako imamo uzorak, možemo procijeniti populacijski parametar (procjena točke), na temelju statistike uzorka
- Ali zato što nismo uzorkovali cijelu populaciju, statistički uzorak će se razlikovati od parametra populacije

Sama procjena točke nije dovoljna, ali možemo dati interval u kojem se najvjerojatnije nalazi

Koraci procjene parametara

1. Odaberite razinu pouzdanosti i veličinu uzorka
2. Odaberite slučajni uzorak iz populacije
3. Prikupljanje podataka
4. Sažmite podatke u statistiku (npr. srednja vrijednost, udio itd.)
5. Izračunajte granicu pogreške
6. Uzmite statistiku $\pm$ marginu pogreške da biste dobili procjenu parametra

Razina pouzdanosti

Razina pouzdanosti: kvantificira razinu pouzdanosti da se parametar nalazi u intervalu.

- Razina pouzdanosti od 95% znači da uzimamo u obzir 95% svih mogućih uzoraka prilikom predviđanja
  - Mogli smo slučajno odabrati uzorak koji se nalazi izvan našeg intervala povjerenja!

Interval pouzdanosti za populacijsku sredinu

Ako je standardna devijacija populacije poznata:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Ako je standardna devijacija populacije nepoznata i/ili je veličina uzorka mala:

$$x \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

Veličina uzorka

Određivanje veličine uzorka za interval pouzdanosti za populacijsku sredinu:

$$n \geq \left(\frac{z^* \sigma}{margin\ of\ error} \right)^2$$

Primjeri procjene intervala

Primjer 1:

Procjenjujemo prosječnu visinu muškaraca odabirom slučajnog uzorka od 100 i dobivamo sljedeću procjenu s razinom pouzdanosti od 95%:

$$\bar{x} = 175 \pm 6{,}2\ cm$$

Srednja vrijednost populacije vjerojatno će biti između 168,8 i 181,2 cm, ali imamo 5% šanse da interval pouzdanosti ne uključuje srednju vrijednost populacije.

Primjeri procjene intervala

Primjer 2:

Predstoje izbori u Zemlji Čudesa. Procjenjujemo koliko će ljudi glasati za stranku A. Procjenjujemo da je udio ljudi koji podupiru stranku A jednak $48 \pm 3\%$.

To znači da je udio osoba koje podupiru stranku A između 45 i 51%.

Testiranje hipoteze

Logika testiranja hipoteza – primjer:

- Iskazujemo tvrdnju o parametru populacije
  - Recimo, da je prosječna visina $\mu = 165\ cm$
- Znamo da prema teoremu o središnjoj granici, ako imamo dovoljno veliku veličinu uzorka, distribucija uzorkovanja bit će blizu normalnoj
- Za normalnu distribuciju možemo izračunati vjerojatnost postojanja uzorka s prosječnom središnjom vrijednošću $\bar{x}$, s obzirom na parametar populacije $\mu = 165$

Testiranje hipoteze

- Postavimo tvrdnju o vrijednosti parametra populacije
- Tvrdnju možemo testirati s obzirom na naš uzorak, koliko je vjerojatno da je naša tvrdnja o parametru populacije istinita
  - Hipotezu možemo odbacit ukoliko je malo vjerojatno da smo naš uzorak izvukli iz populacije s utvrđenim parametrom
    - Odlučujemo koliko sigurni želimo biti – stupanj važnosti

Testiranje hipoteze

- Nulta hipoteza:
  - Jednadžba koja se odnosi na parametar
    - Npr. Tvrdim da je prosječna visina 165 cm
      - H_0: $\mu = 165$
  - Testiramo navedenu tvrdnju
- Alternativna hipoteza:
  - Parametar populacije nije jednak vrijednosti koju zahtijeva nulta hipoteza
    - H_0: $\mu \neq 165$
  - Vrijednost parametra populacije je viša…
    - H_0: $\mu > 165$
  - Vrijednost parametra populacije je niža
    - H_0: $\mu < 165$

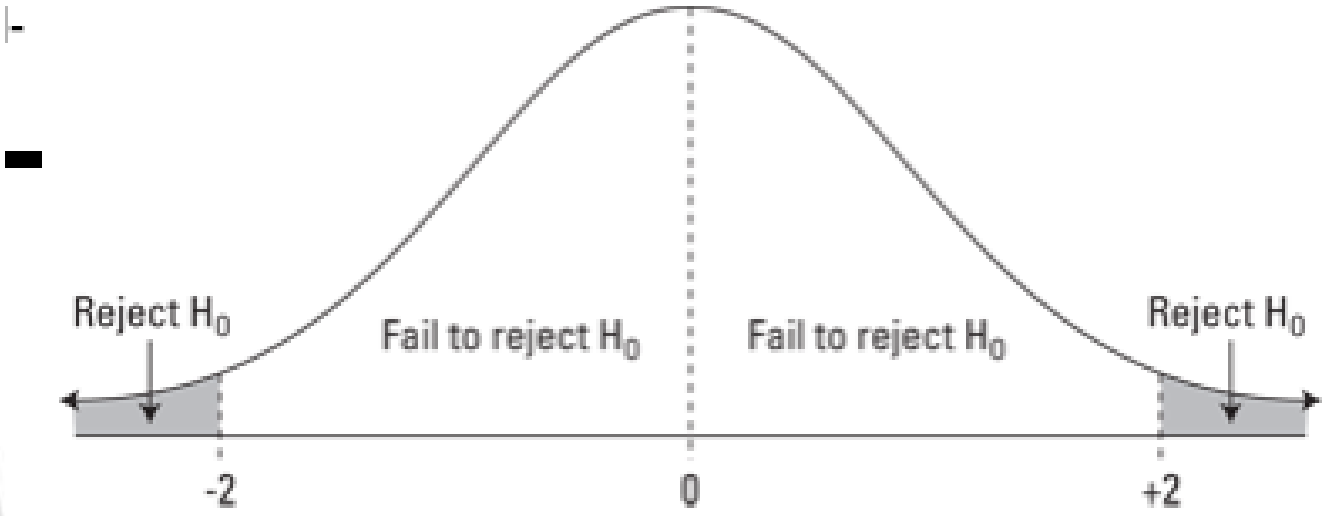
Testiranje hipoteze

- Koraci testiranja hipoteze:
  - 1. Definirati nultu i alternativnu hipotezu
  - 2. Odrediti statistiku testa
  - 3. Odrediti stupanj važnosti i kritičnu regiju
  - 4. Izračunati p-vrijednost
  - 5. Odlučiti između odbacivanja i nemogućnosti odbacivanja hipoteze
    - Ako p-vrijednost pripada u kritičnu regiju, možemo odbaciti hipotezu

Testiranje hipoteze

- Zapamtite:
  - Određujemo stupanj važnosti i kritični raspon
- Vrste pogrešaka:
  - Pogreška tipa 1:
    - Odbacujemo nultu hipotezu, ali je ona istinita – imamo istu šansu za to kao naš stupanj važnosti
  - Pogreška tipa 2:
    - Ne možemo odbaciti nultu hipotezu iako je neistinita
      - Ne znamo da li je slučajnost
      - Ovu pogrešku možemo smanjiti povećanjem veličine uzorka

Testiranje hipoteze



Testiranje hipoteze – primjer

```
. ttest female == .508

One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
female	2322	.5697674	.0102769	.4952152	.5496145	.5899203

```
    mean = mean(female)                                     t =   6.0103
Ho: mean = .508                         degrees of freedom =     2321

   Ha: mean < .508               Ha: mean != .508                 Ha: mean > .508
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Ispitivanje odnosa

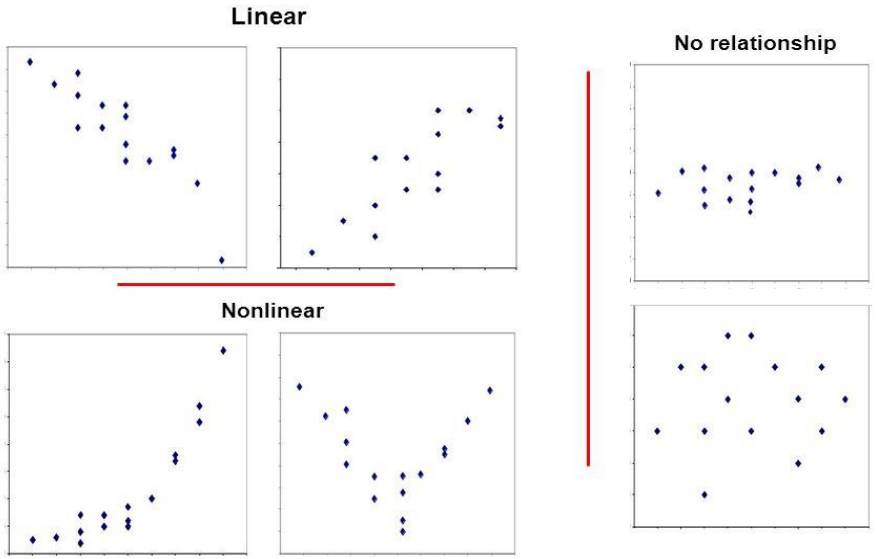
- Odnos između varijabli:
  - Uzročnost naspram paralelnih promjena
- Vrsta odnosa
  - **Linearni**
  - Kvadratni
  - Itd…

Linearni odnos

- Korelacija
- Regresija

Linearni odnos

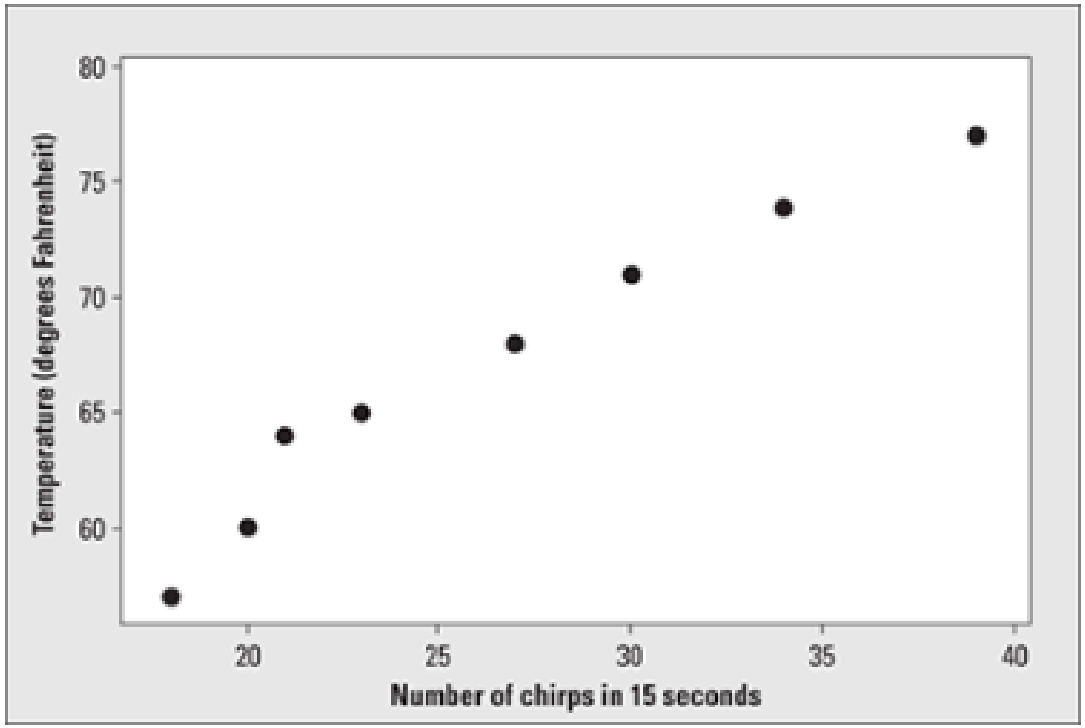
- Opći oblik: $y = mx + b$
  - y: zavisna varijabla
  - X: nezavisna varijabla
  - M: nagib
  - B: sjecište



Linearni odnos

- Otkrivanje linearnog odnosa
  - Dijagram rasipanja
  - Korelacija

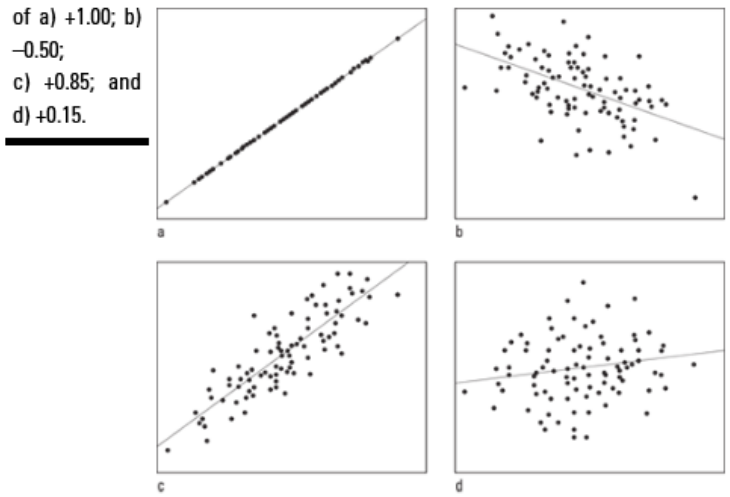
Dijagram rasipanja



Korelacija

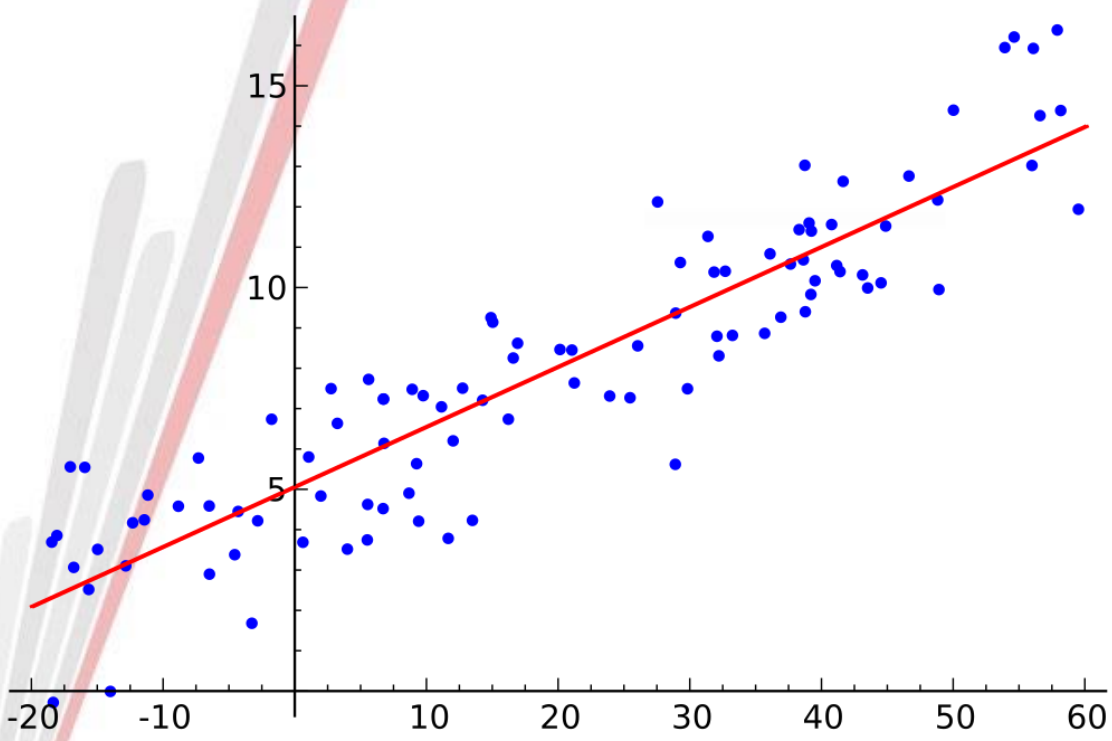
- Korelacija
  - Između −1 i +1
    - −1: savršeni negativni linearni odnos
    - +1: savršeni pozitivni linearni odnos
    - 0: no linear relationship

$$r = \frac{1}{n-1}\left(\frac{\sum_x \sum_y (x-\bar{x})(y-\bar{y})}{s_x s_y}\right)$$



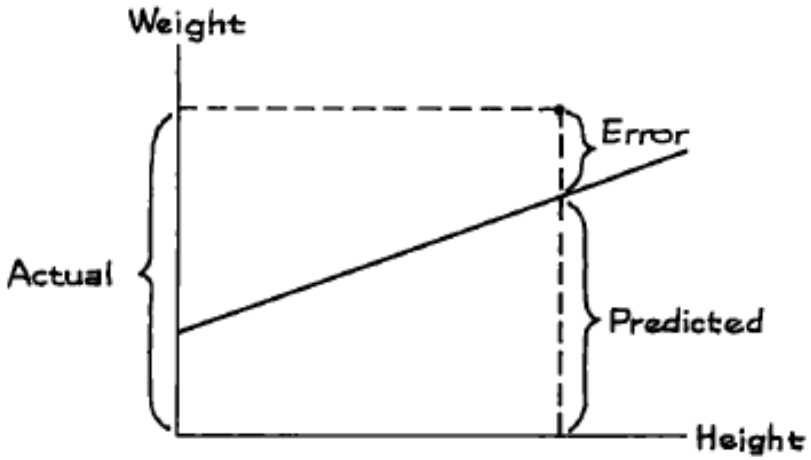
Regresija

- Ako imamo linearne odnose između varijabli (na temelju dijagrama rasipanja i korelacije)
- Tijekom regresijske analize stavljamo liniju na podatke uzorka
  - Obični najmanji kvadrati (OLS): smanjuje prosječnu kvadratnu udaljenost podatkovnih točaka od linije
- Opći oblik linearne funkcije:
  - $y = \beta_0 + \beta_1 x \ldots + u$
    - y: Zavisna varijabla
    - x_n: nezavisne varijable
    - β_0: sjecište
    - β_n: koeficijenti nezavisnih varijabli
    - *u: pogreška*


15
10
5
-20
-10
10
20
30
40
50
60

Regresija

- Regresija se koristi za predviđanje y iz x
- Stvarna opažanja razlikuju se od vrijednosti procijenjene regresijskom linijom
  - $error(residual) = actual\ value\ - predicted\ value$



Regresija

Rezidualna odstupanja:

- ako je odnos linearan i model dobro pristaje, prosječna rezidualna odstupanja su 0
- Ako rezidualna odstupanja pokazuju jak uzorak, uklapanje linearne linije možda neće biti dostatno

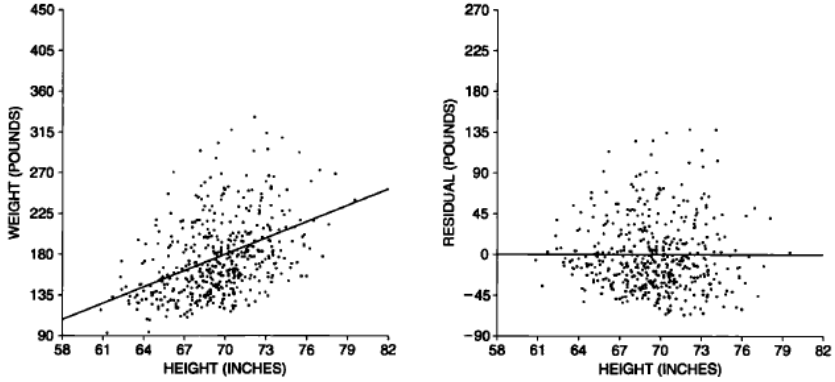


Figure 7. A residual plot with a strong pattern. It may have been a mistake to fit the regression line.

Regresija – primjer

```
. regress mpg weight
```

Source	SS	df	MS	Number of obs	=	74
				F(1, 72)	=	134.62
Model	1591.9902	1	1591.9902	Prob > F	=	0.0000
Residual	851.469256	72	11.8259619	R-squared	=	0.6515
				Adj R-squared	=	0.6467
Total	2443.45946	73	33.4720474	Root MSE	=	3.4389

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

Kauzalnost

Korelacija: ne podrazumijeva kauzalnost, već samo paralelnu promjenu varijabli

Regresija: može značiti kauzalnost, ako su ispunjeni određeni uvjeti:

- Možemo kontrolirati sve zbunjujuće varijable koje bi mogle utjecati na zavisne i nezavisne varijable
  - To se uglavnom događa u eksperimentalnim situacijama

Valjanost

Valjanost: Valjanost je stupanj do kojeg je koncept, zaključak ili mjerenje dobro utemeljen i koji vjerojatno točno odgovara stvarnom svijetu.

Vanjska valjanost: odnosi se na stupanj do kojeg možemo generalizirati naše nalaze – uska povezanost s metodom uzorkovanja

Unutarnja valjanost: odnosi se na mjeru u kojoj dokazi potkrepljuju tvrdnju o uzroku i posljedici – usko povezana s rigoroznošću analize

Statistički proces

1. **Planiranje istraživanja**

- Definiranje interesne populacije
- Definiranje istraživačkih pitanja i hipoteza
- Planiranje prikupljanje podataka (veličina uzorka, metoda odabira uzorka)

2. Prikupljanje podataka
3. Organizacija podataka
4. Analiza podataka
5. **Izvođenje zaključaka i donošenje zaključaka o populaciji**
6. Prezentacija nalaza

1. Planiranje istraživanja

Populacija:

- Precizna definicija populacije
  - Moramo biti sposobni odlučiti za svaki entitet da li je dio populacije ili ne
    - Primjer: Hrvatski narod – hrvatski građani/ljudi koji žive u Hrvatskoj/hrvatski državljani koji žive u Hrvatskoj

1. Planiranje istraživanja

Istraživačka pitanja:

- Precizno postavljena istraživačka pitanja
  - Teritorijalna dimenzija
  - Dimenzija vremena
  - Varijable interesa

Hipoteze:

- Hipoteze koje je moguće testirati izražene matematički

1. Planiranje istraživanja

Odabir uzorka:

- Logika statističke interferencije temelji se na pretpostavci reprezentativnog uzorka dovoljne veličine
  - Ako nemamo slučajni uzorak, ne možemo procijeniti standardnu pogrešku
  - Ako nemamo slučajni uzorak, moramo uzeti u obzir moguće pristranosti koje proizlaze iz odabira uzorka

1. Planiranje istraživanja

Vrste uzorka:

- Ne-slučajni uzorci: koriste se za dobivanje preliminarnih informacija o fenomenu ili kada slučajno uzorkovanje nije moguće
  - Proizvoljno uzorkovanje
  - Praktično uzorkovanje
  - Uzastopno uzorkovanje
  - Uzorkovanje metodom ‘snježne grude’
  - Kritičko/svrsishodno uzorkovanje
- Slučajni-uzorci: koriste se za izvođenje zaključaka o populaciji iz uzorka
  - Jednaka vjerojatnost odabira (EPS)
  - Jednostavno slučajno uzorkovanje
  - Sustavno uzorkovanje
  - Stratificirano uzorkovanje itd.

1. Planiranje istraživanja

Pristranost uzorkovanja:uvijek moramo računati na moguće pristranosti!

- Populacija uzorkovanja:
  - Nije uvijek moguće odabrati uzorak iz cijele populacije (npr. nemamo popis svih građana s kontaktnim informacijama)
- Metoda uzorkovanja može uzrokovati pristranosti (ne-slučajni odabir)
- Tehnički detalji prikupljanja podataka mogu uzrokovati pristranosti (npr. anketa putem telefonskih poziva, vodeća pitanja u anketama itd.)

4. Analiza podataka

Analitičke metode ovise o:

- Karakteristikama istraživanog fenomena
- Vrsti ispitivanih varijabli
- Istraživačkih pitanja i hipoteze

5. Zaključci

- Unutarnja valjanost
- Vanjska valjanost

Vježba i diskusija

SUFINANCIRANO SREDSTVIMA EUROPSKE UNIJE
EUROPSKI POLJOPRIVREDNI FOND ZA RURALNI RAZVOJ –
EUROPA ULAŽE U RURALNA PODRUČJA

Program
RURALNOG RAZVOJA

EQUINOX
CONSULTING

EQUINOX
CONSULTING