Statistika za jezikoslovno istraživanje

Damir Ćavar

17.3.2010.

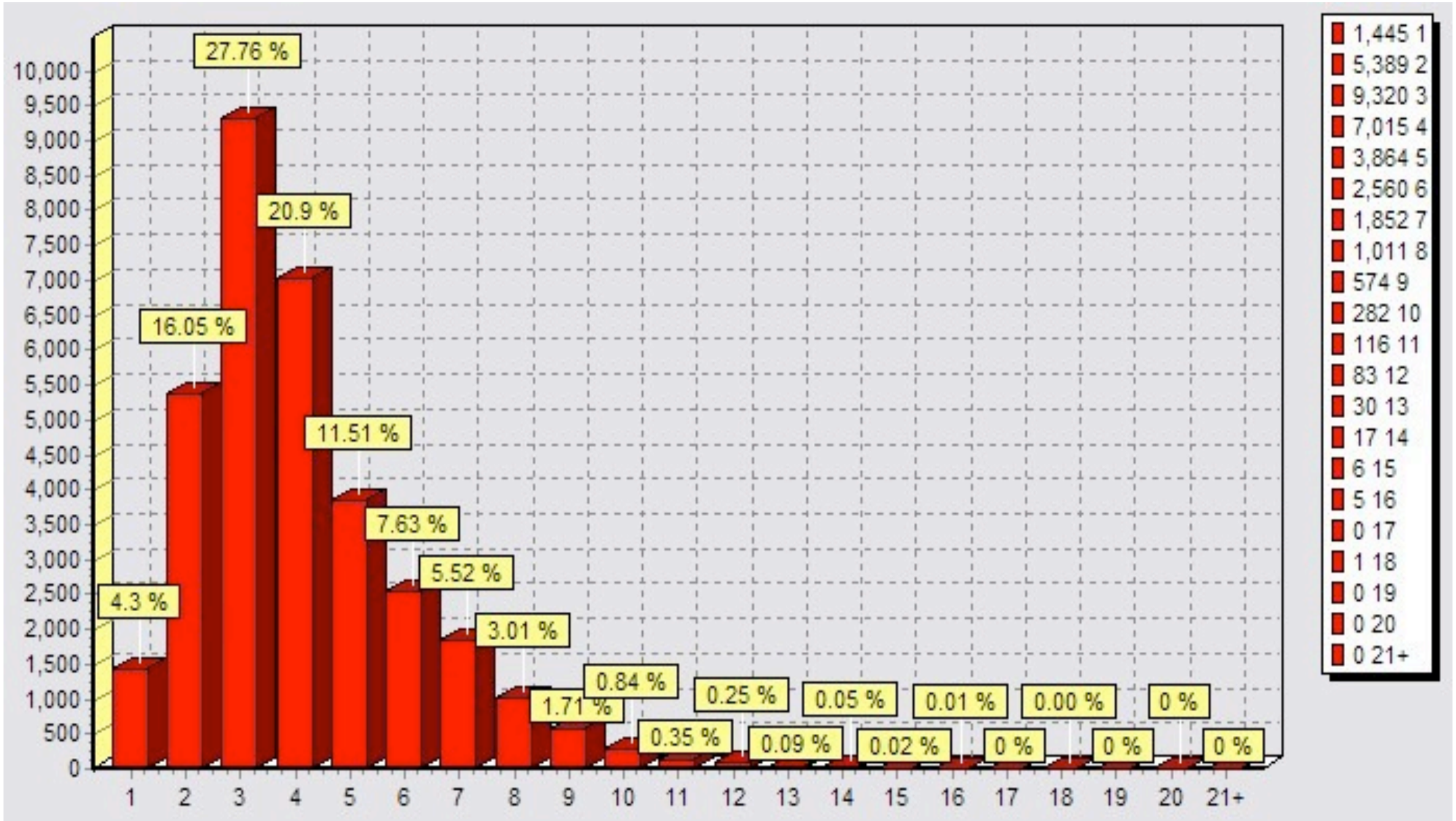
Primjeri u R-u

- dodavanje rezultata
- aritmetička sredina
- srednja vrijednost
- varianca
- učitavanje tabela iz Excela ili OpenOffice Calca
- sortiranje

Deskriptivna statistika

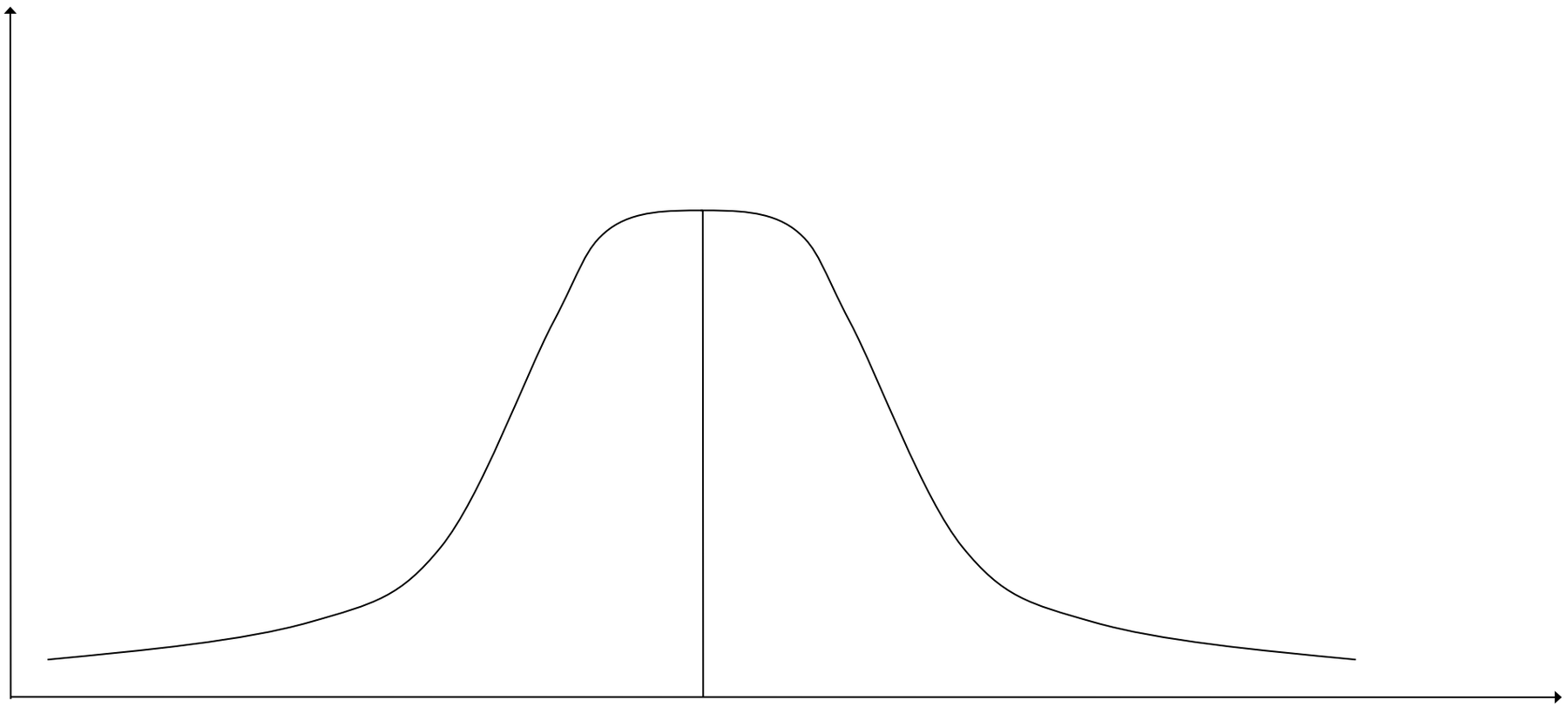
- Primjer:
  - "A House of Pomegranates" od Oscara Wilde
  - Broj pojavnica: 33570
  - Broj različitih riječi: 4066
  - Dužina riječi
  - Frekvencija riječi

Dužina i frekvencija riječi



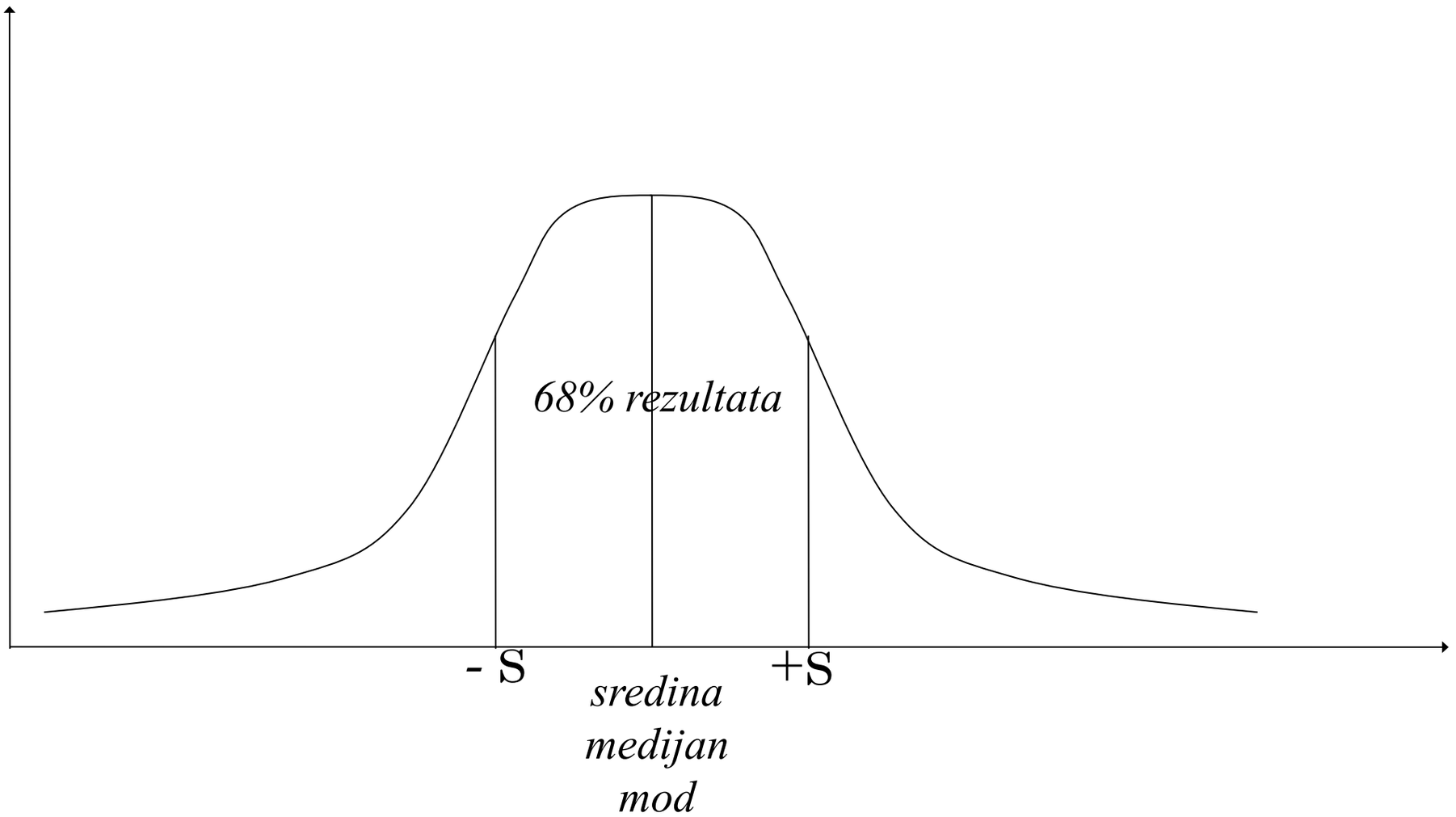
Deskriptivna statistika

- Simetrična distribucija: aritmetična sredina, medijan i mod su isti

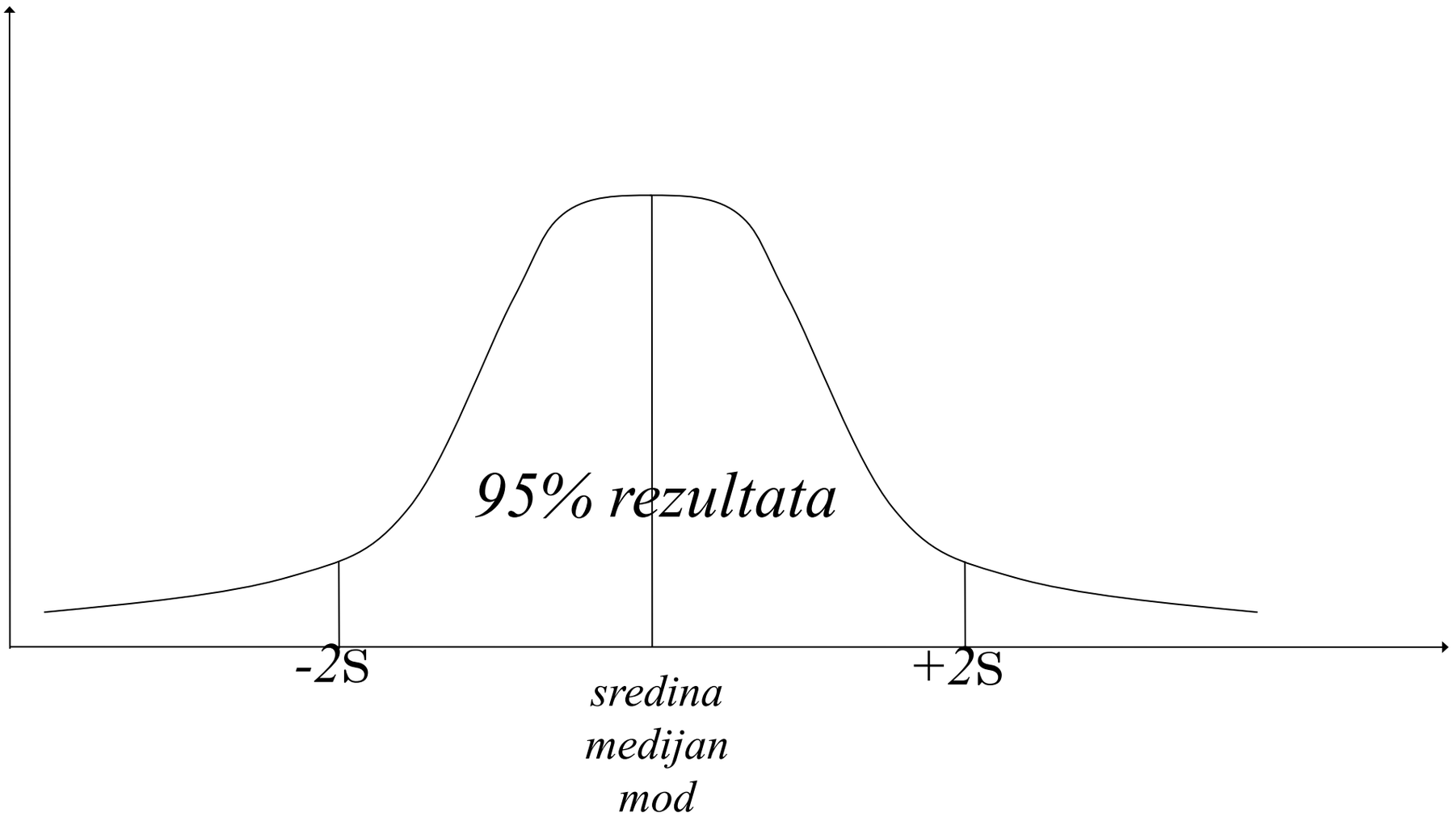


sredina
medijan
mod

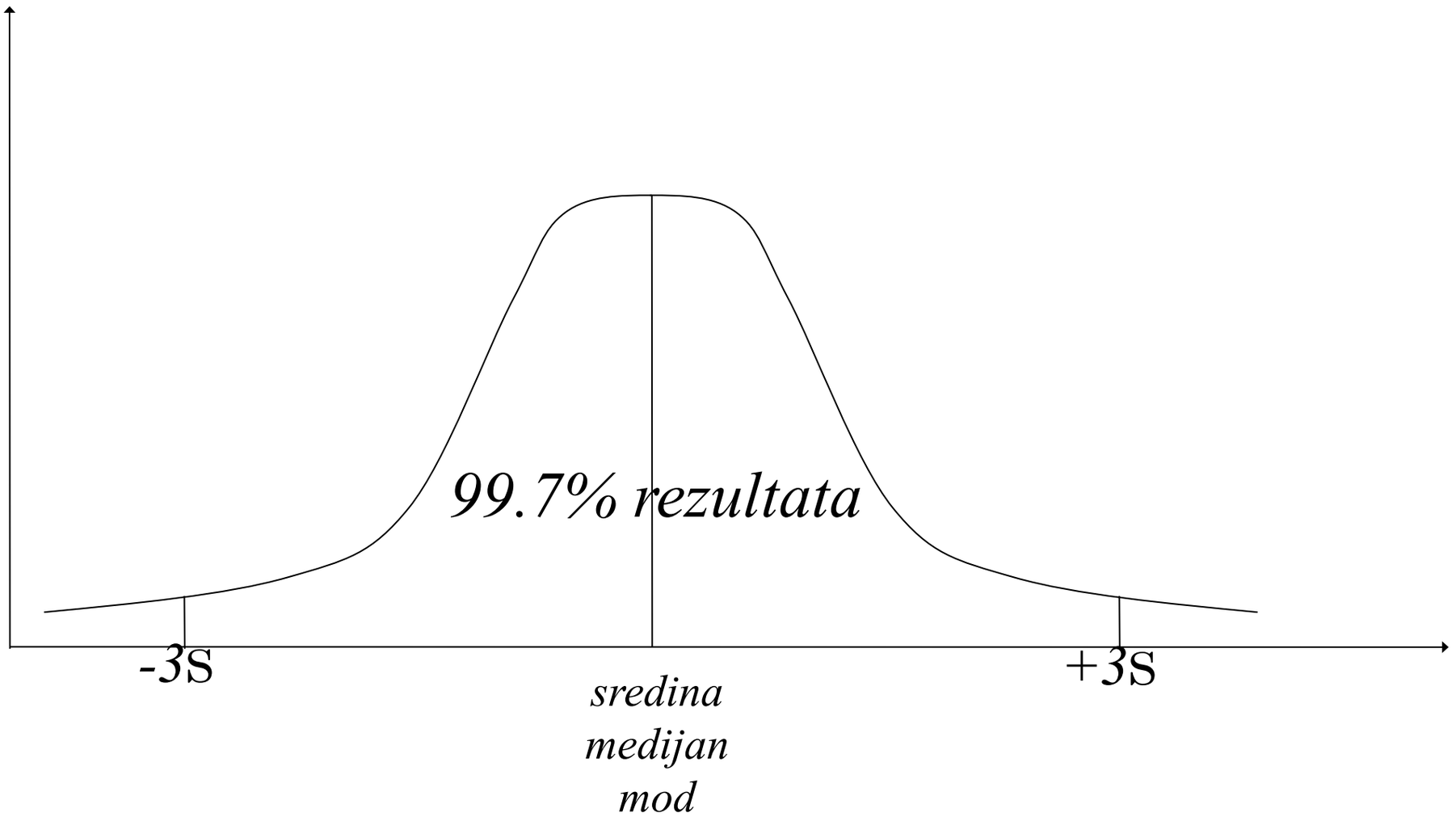
Deskriptivna statistika



Deskriptivna statistika



Deskriptivna statistika



Teorija vjerojatnosti

- Vjerojatnost jednog specifičnog događaja možemo odrediti kao odnos između broja pogodnih ishoda i cjelokupnog broja mogućih ishoda.

$$P(A) = \frac{\text{broj pogodnih ishoda}}{\text{broj svih mogućih ishoda}}$$

- Pristup: frekvencijski

Teorija vjerojatnosti

- Značajne činjenice:

  – Ako dvije varijable nisu povezane:
    - $P(XY) = P(X)\,P(Y)$

  – Ako jesu povezane:
    - $P(XY) \neq P(X)\,P(Y)$

Primjer: Kolokacije

- Riječi u kontekstu
  - distribucija
  - idijomi
  - kolokacije
    - statističke osobine
    - funkcionalne riječi

- Kako testirati da li je niz riječi kolokacija?
  - Statistički, signifikantnost

Testiranje hipoteza

- Osnovni statistički rezultati
- Strategija:
  - Formuliramo hipotezu što znanstveno očekujemo:

Istraživačka hipoteza ili alternativna hipoteza (Ha)

- Što očekujemo i testiramo

Testiranje hipoteza

- Strategija:
  - Formuliramo hipotezu koja tvrdi suprotno od naše znanstvene hipoteze: Nulta hipoteza (H0)
    - Testiramo samo nultu hipotezu
  - Ako možemo odbiti ili falsificirati nultu hipotezu, imamo podršku za istraživačku hipotezu.
  - Hipotezu obično ne možemo “dokazati”, samo možemo naći potporu za neku hipotezu.

Testiranje hipoteza

- Alternativna hipoteza:
  - Na Sveučilištu u Zadru ocjene studenata lingvistike iz područja statistike razlikuju se od ocjena studenata psihologije.
    - Ha: $\mu 1 \neq \mu 2$
    - Ha: $\mu 1 - \mu 2 \neq 0$

Testiranje hipoteza

- Nulta hipoteza:
  - Na Sveučilištu u Zadru ocjene studenata lingvistike iz područja statistike ne razlikuju se od ocjena studenata psihologije.
    - H0: μ1 = μ2
    - H0: μ1 - μ2 = 0

Testiranje hipoteza

- Još specifičnija Alternativna hipoteza:
  - Na Sveučilištu u Zadru studenti lingvistike imaju bolje ocjene iz statistike od studenata psihologije.
    - Ha: μ1 > μ2
    - Ha: μ1 - μ2 > 0

Testiranje hipoteza

- Još specifičnija Nulta hipoteza
  - Na Sveučilištu u Zadru studenti lingvistike imaju lošije ocjene iz statistike od studenata psihologije.
    - H0: $\mu_1 \leq \mu_2$
    - H0: $\mu_1 - \mu_2 \leq 0$

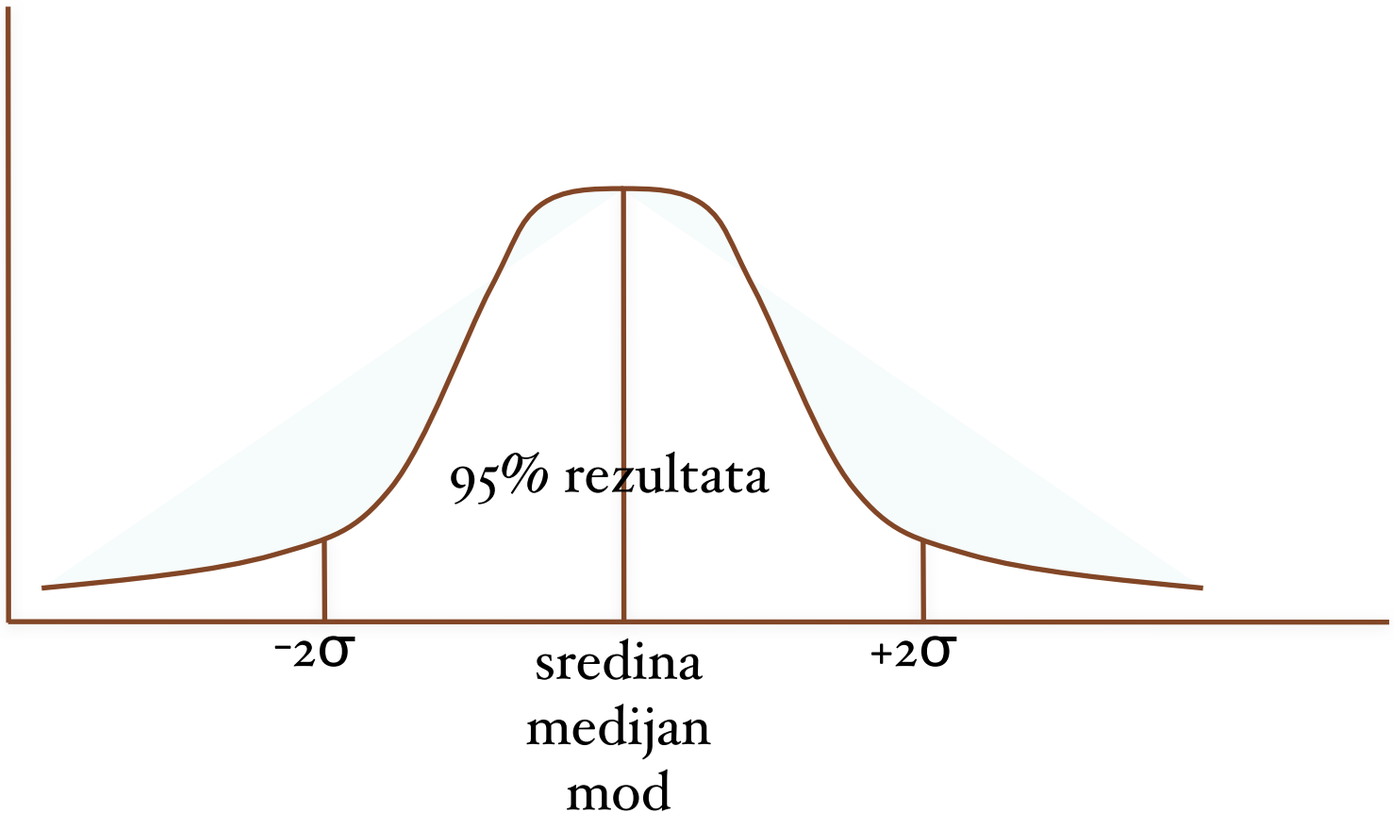
Testiranje hipoteza

- Ako imamo distribuciju nekog poznatog područja
  - npr. Normalnu distribuciju
- Izračunamo vjerojatnost da su se rezultati dobili slučajno.
- Ako je ta vjerojatnost niska, alternativna hipoteza (da nisu rezultat slučajnosti) ima potporu.

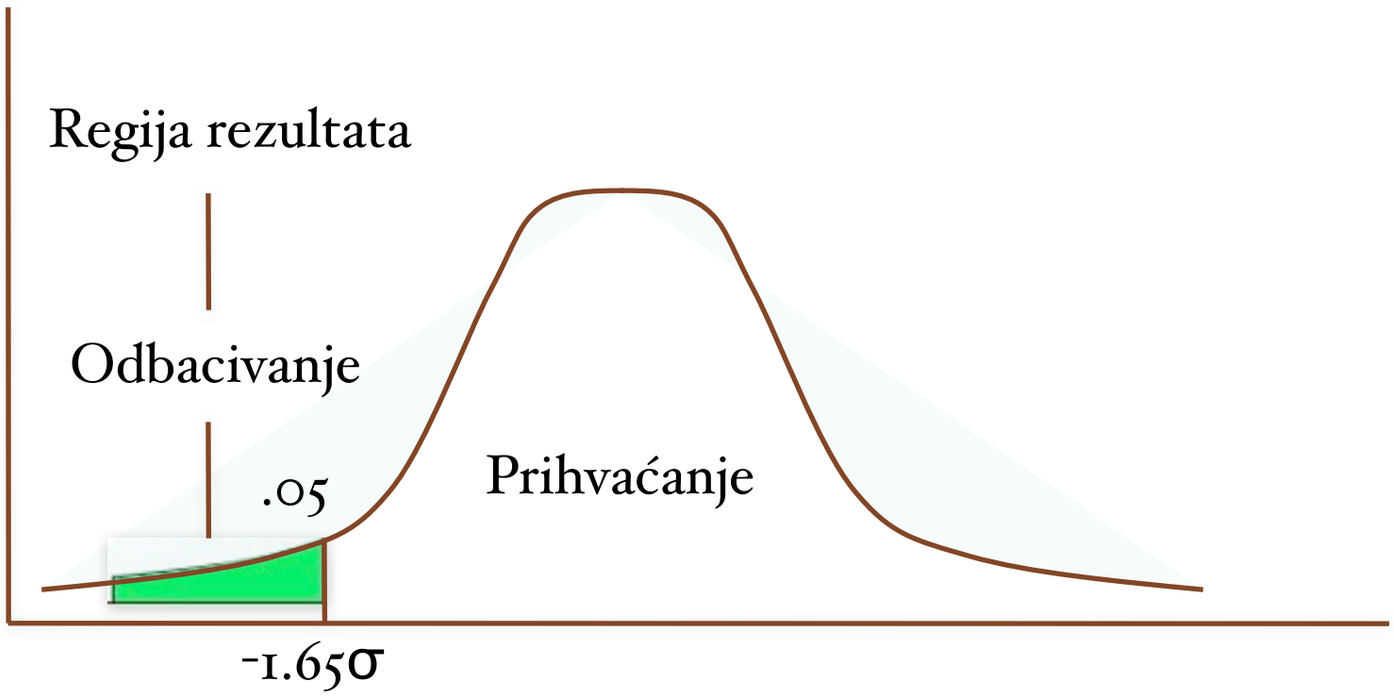
Testiranje hipoteza

- Dvije mogućnosti:
  - Odbijamo Nultu hipotezu
  - Prihvaćamo Nultu hipotezu

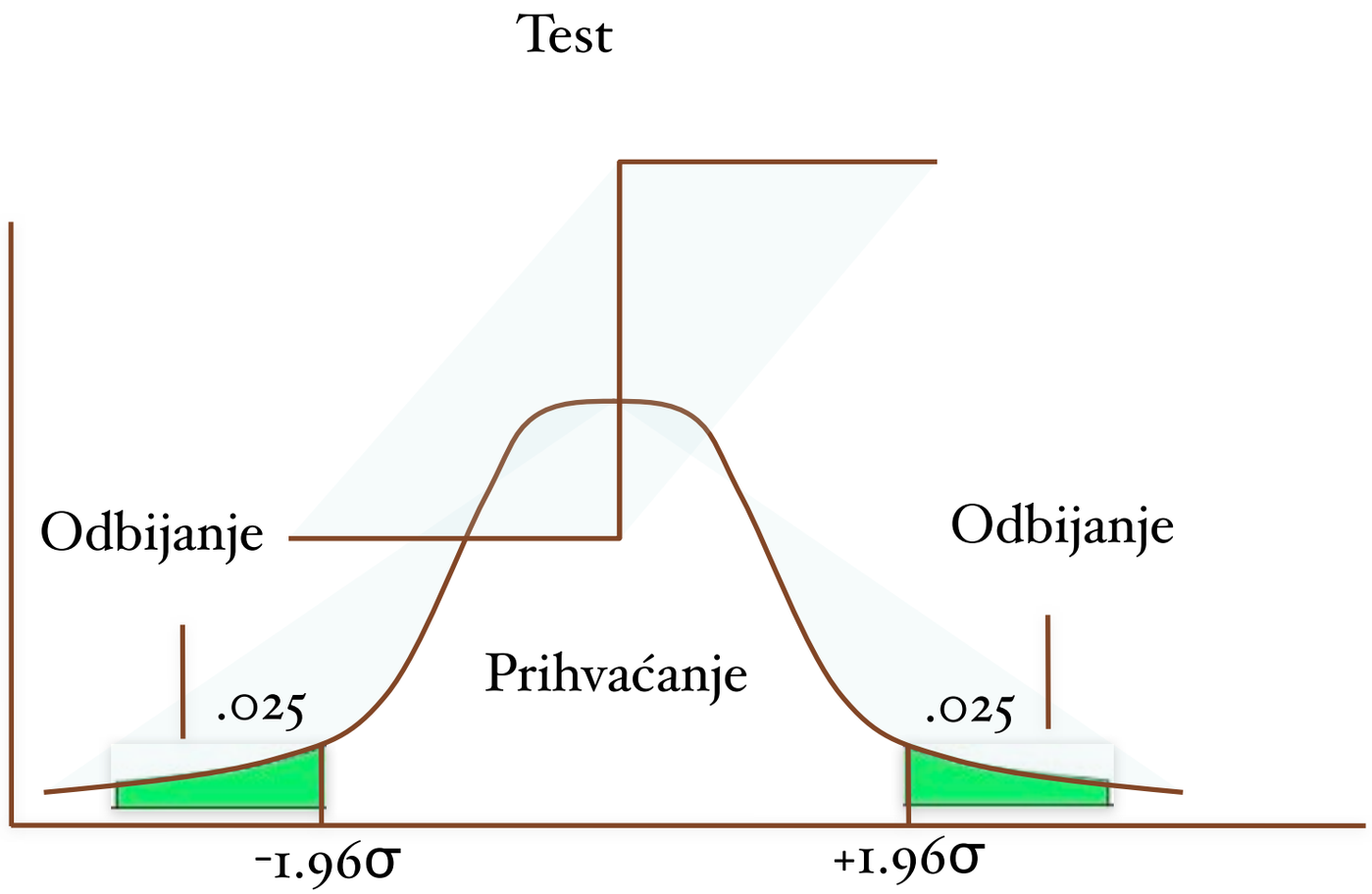
Testiranje hipoteze



Testiranje hipoteze



Testiranje hipoteze



chi^2 (χ2) test

- Favorizirane aktivnosti iz jedne populacije od 125 studenata:

	nogomet	**ples**	**računala**	**ukupno**
muški	30	29	16	75
ženski	12	33	5	50
ukupno	42	62	21	125

chi^2 (χ2) test

- Je li izbor favorizirane aktivnosti slučajan i neovisan o spolu ispitanika?
  - –Ako te dvije varijable spol i aktivnost ne ovise jedna od druge, možemo predvidjeti koliko ispitanika možemo očekivati u svakoj kombinaciji.
  - –Ako se očekivanje značajno razlikuje od rezultata, te dvije varijable vjerojatno ovise jedna od druge.

chi^2 (χ2) test

- Alternativna hipoteza:
  - U ovom eksperimentu te varijable ovise jedna od druge.

- Nulta hipoteza:
  - Te varijable ne ovise jedna od druge.

chi^2 (χ2) test

- Vjerojatnost da je slučajno izabrani ispitanik:
  - muško: 75/125 = .6
  - žensko: 50/125 = .4

- Vjerojatnost da slučajno izabrani ispitanik preferira:
  - nogomet: 42/125 = .336
  - ples: 62/125 = .496
  - računala: 21/125 = .168

chi^2 (χ2) test

- Ne ovisne varijable: umnožak pojedinačne vjerojatnosti
  - P(XY) = P(X) P(Y)

chi^2 (χ2) test

- Očekivana vjerojatnost da je jedan slučajno izabrani ispitanik muško i preferira nogomet:
  - P(muško & nogomet): .6 x .336 = .202
  - Očekivani broj: .202 x 125 = 25.2

chi^2 (χ2) test

- Ili:
  - umnožak sumacije reda sa sumacijom razreda, i djeljenjem kroz ukupni broj ispitanika:
- (75 x 42) / 125 = 25.2

	nogomet	ples	računalo	ukupno
muški	30 (25.2)	29 (37.2)	16 (12.6)	75
ženski	12 (16.8)	33 (24.8)	5 (8.4)	50
ukupno	42	62	21	125

chi² (χ2) test

- Formula:

$$\chi^2 = \sum \frac{(\text{rezultat} - \text{očekivano})^2}{\text{očekivano}}$$

$$\chi^2 = \frac{(30-25.2)^2}{25.2} + \frac{(29-37.2)^2}{37.2} + \frac{(16-12.6)^2}{12.6} + \frac{(12-16.8)^2}{16.8} + \frac{(33-24.8)^2}{24.8} + \frac{(5-8.4)^2}{8.4} = 9.097$$

chi^2 (χ2) test

- Što veći χ2 to vjerojatnije su varijable ovisne.
- Velike razlike se povećaju uz efekt kvadrata.

chi^2 (χ2) test

- Naći rezultat u tabeli:
  - Degree-of-freedom:
    - df = (broj redova - 1) x (broj razreda – 1)
    - Naš primjer: (2 – 1) x (3 – 1) = 2
  - U tabeli: 9.097 (< .025; > .01)

chi^2 (χ2) test

- Primjer: 9.097 (< .025; > .01)
  - Signifikantnost (na razini: .05, .01)!
  - Odbijamo Nultu hipotesu (ne ovisnost varijabli)

chi^2 (χ2) test

- Collocations
  - new, companies

	r1=new	r1¬new	ukupno
r2=companies	8	4667	4675
r2¬companies	15820	14287181	14303001
ukupno	15828	14291848	14307676

chi^2 (χ2) test

- Kolokacije
  - new, companies

	r1=new	r1¬new	ukupno
r2=companies	8 (5)	4667 (4669)	4675
r2¬companies	15820 (15822)	14287181 (14287178)	14303001
ukupno	15828	14291848	14307676

chi^2 (χ2) test

- Kolokacije
  - ban, derenčin = 267771.9929697935

	r1=ban	r1¬ban	ukupno
r2=derenčin	31 (...)	69 (...)	100
r2¬derenčin	3019 (...)	84972930 (...)	84975949
ukupno	3050	84972999	84976049

Domaći

- R: posložite podatke sample.dat po dužini riječi i izgenerirajte grafiku po frekvenciji kao barplot

- Nađite tako jedan par pojavnica u korpusu: Korpus hrvatskog jezika (riznica.ihjj.hr) i izračunajte Chi2 vrijednost i vjerojatnost da se radi o kolokaciji, ili uradite to za dvije prethodne tabele.