Zetetika

~ 5 ~

Staša Vujičić Stanković

Moć statistike

Malo ljudi je razume, ali svi je uvažavaju.

Rekli su ... (o statistici)

- Bendžamin Dizraeli (1804 - 1881), britanski državnik

  Postoje tri vrste laži: obične laži, važne laži i statistika.

- Herbert Džordž Vels (1866 - 1946), engleski naučnofantastični književnik

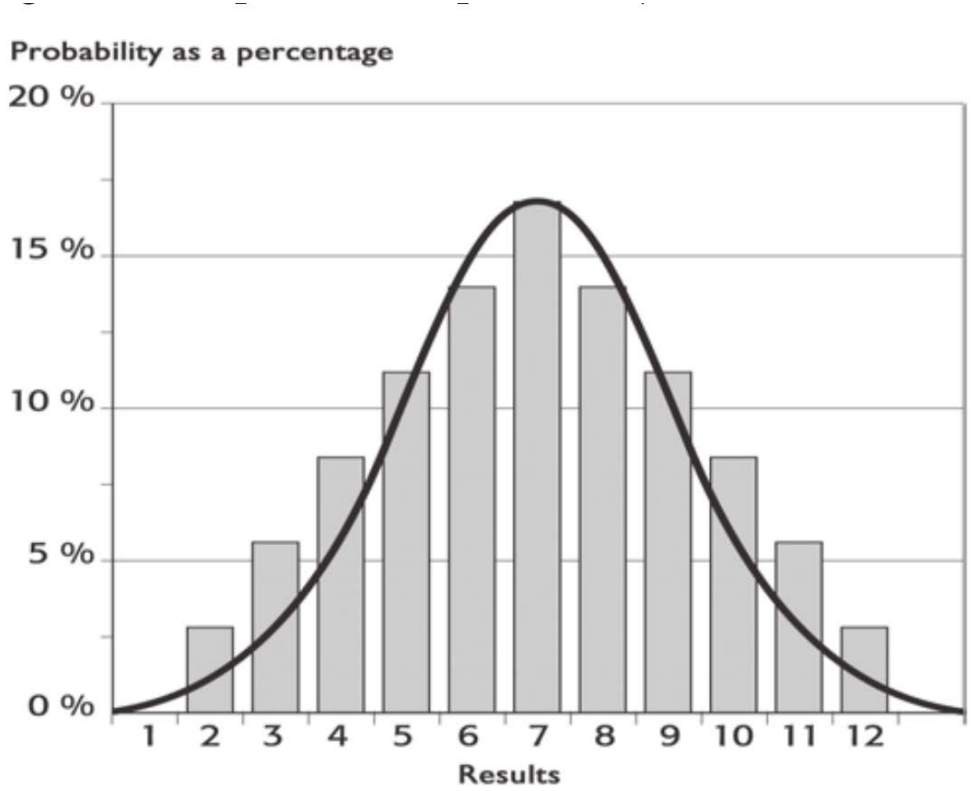
  Kao što je sada potrebno da znamo da pišemo i čitamo,
  tako će u 20. veku biti potrebno da znamo statistiku.

Statistika

- Statistika je funkcija uzorka $x_1, x_2 \ldots x_n$ čiji analitički izraz ne zavisi od nepoznatih parametara raspodele obeležja populacije iz koje je uzorak uzet.

Gausova kriva - primer

- Zbir koji dobijamo bacanjem dve kockice.
- Normalna raspodela ili Gausova kriva



Mere deskriptivne statistike

- Najčešće mere koje se koriste u deskriptivnoj statistici:
  - Aritmetička sredina
  - Medijana
  - Modus
  - Standardna devijacija

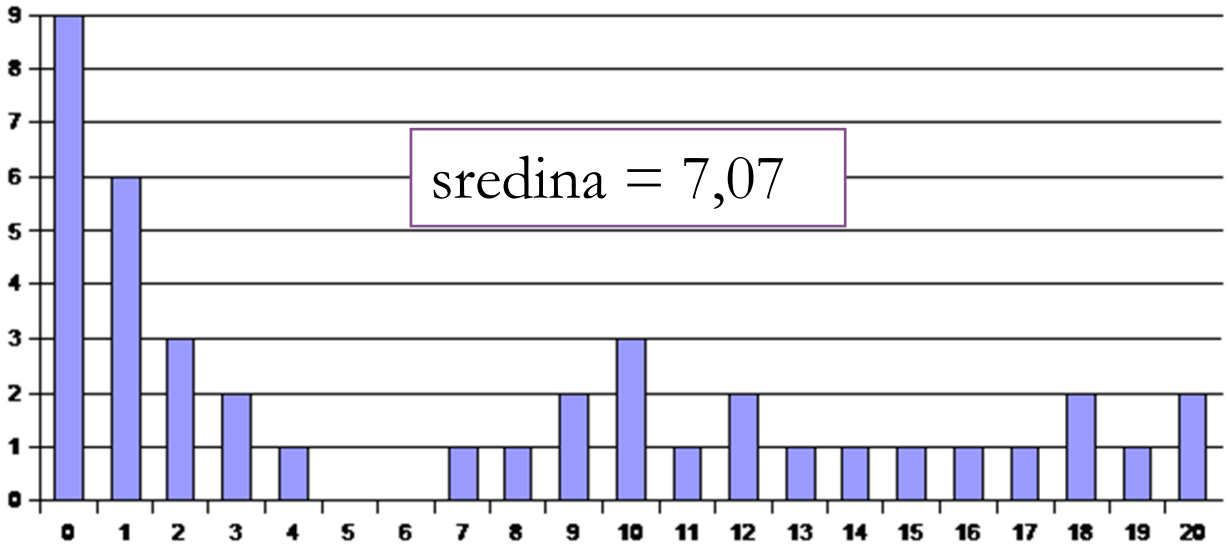
Aritmetička sredina

- Uobičajena (aritmetička) sredina: $\frac{zbir\ elemenata}{broj\ elemenata}$
- Aritmetička sredina se najčešće koristi.
  Uzimaju se svi podaci u razmatranje.
- Ipak, osetljiva je na ekstremne vrednosti.

Aritmetička sredina

- Npr, ako u jednom preduzeću:
  - 47 radnika zarađuje 1.000 € mesečno
  - gazda zarađuje 1.000.000 € mesečno
  - sredina je $\frac{47\times1.000+1.000.000}{48} = 21.182,50$€ mesečno
- Da li je ova sredina reprezentativna za bilo šta?
- Iz nje se ne može izvući podatak o raspodeli vrednosti!

Histogram
grafik raspodele



Poeni na testu iz Programiranja 1

- Ne predstavljamo vrednost, već broj subjekata za tu vrednost
- Ovo omogućava da se relativizuje srednja vrednost
- Iz srednje vrednosti se ne vidi maksimum u raspodeli poena…

Medijana

- Medijana – vrednost veća od jedne polovine predstavljenih vrednosti, a manja od druge polovine

sredina = 7,07
medijana = 4

Medijana

- Medijana se u teoriji verovatnoće i statistici opisuje
  kao broj koji razdvaja gornju polovinu uzorka,
  populacije ili raspodele verovatnoće od donje polovine.
- Medijana se takođe često koristi.
  Ne uzima u obzir sve vrednosti i nije osetljiva na ekstremne vrednosti.
  Ponekad može bolje da opiše podatke od aritmetičke sredine.

Modus

- Modus je vrednost koja se u uzorku ili grupi podataka pojavljuje najčešće.
- Modus se najređe koristi.
  U nekom uzorku može biti više modusa ili ni jedan.
  Ne posmatra sve vrednosti.
- Obično se koristi
  kod nominalnih vrednosti (promenljive opisane imenom) ili
  kod diskretnih promenljivih (mogu imati samo ograničen broj vrednosti).

Modus

$$X = 109, 129, 129, 135, 139, 149, 159, 179$$

$$\mu = \frac{1128}{8} = 141$$

$$\text{Medijana} = \frac{135+139}{2} = 137$$

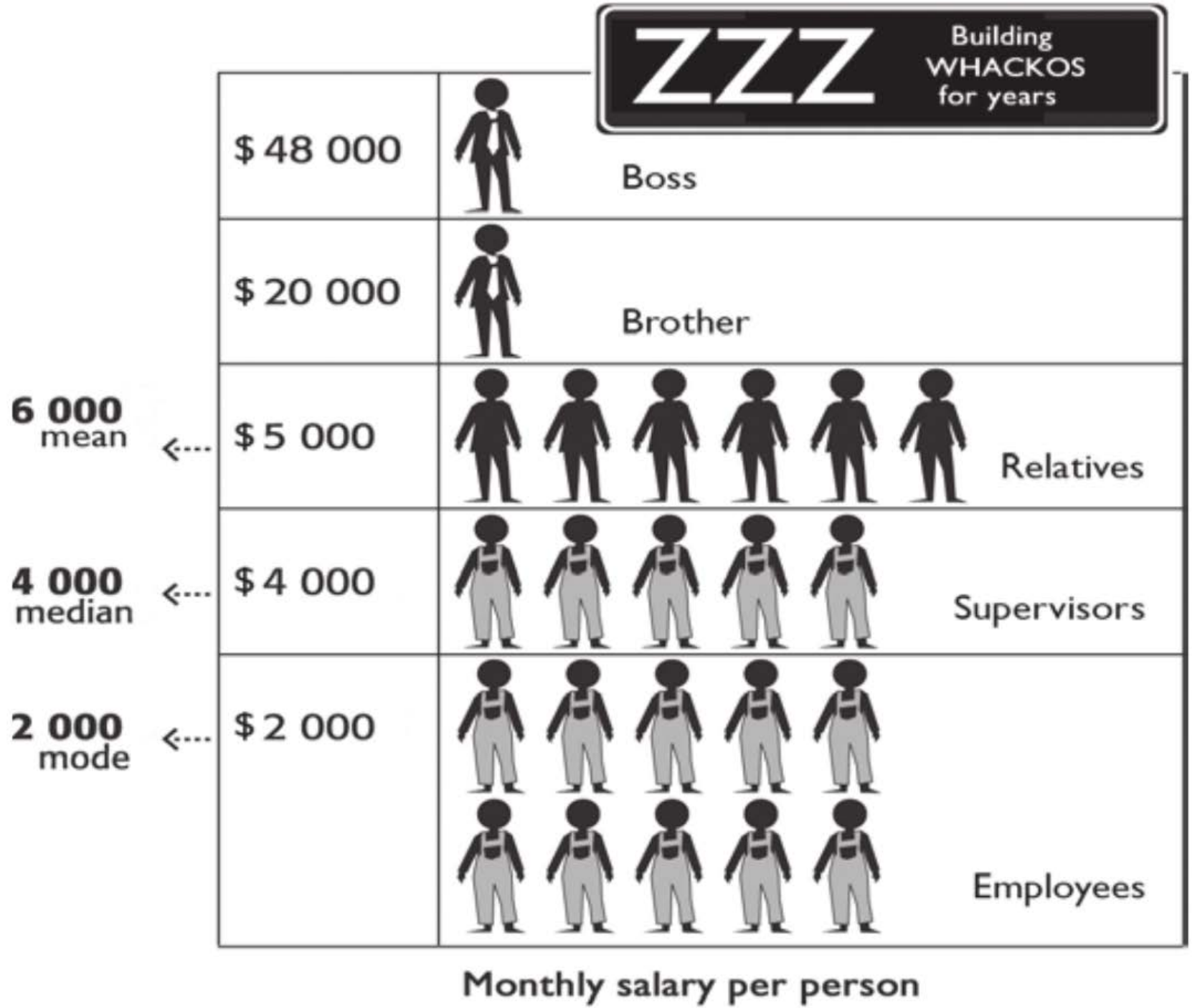
$$\text{Modus} = 129$$

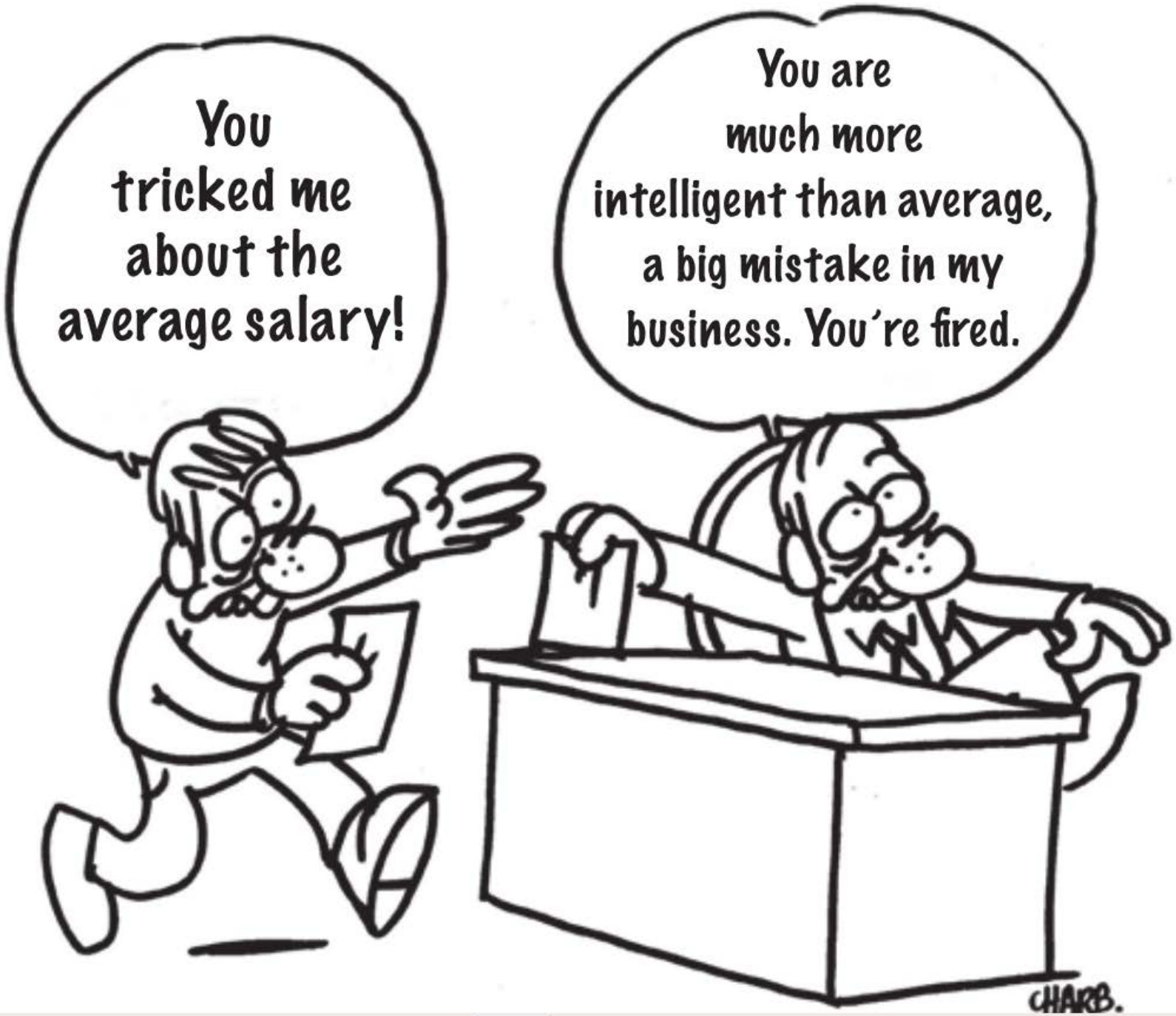
Primer

- Lažno predstavljanje podataka

  Marko se zapošljava u jednoj kompaniji.
  Kompanija ima šefa, njegovog brata i šest rođaka.
  Osoblje čini deset radnika i pet supervizora.
  Šef kaže da je prosečna plata u firmi 6.000$ mesečno.
  Kaže da će Marko u početku primati 1.500$,
  a da će posle probnog perioda plata značajno porasti.

  Posle desetak dana: Marko odlazi kod šefa – lagao si me!


ZZZ
Building WHACKOS for years
$ 48 000
Boss
$ 20 000
Brother
6 000 mean
$ 5 000
Relatives
4 000 median
$ 4 000
Supervisors
2 000 mode
$ 2 000
Employees
Monthly salary per person


You tricked me about the average salary!
You are much more intelligent than average, a big mistake in my business. You're fired.
CHARB.

Na osnovu aritmetičke sredine ne može se uvek izvući podatak o raspodeli vrednosti!

Značajnost (standardna devijacija)

- Stepen značajnosti jednog statističkog rezultata jednaka je verovatnoći da razmak (*devijacija*) između :

  rezultata posmatranja i teorijskog predviđanja

  bude posledica slučaja!

Standardna devijacija

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

x_i – i-ti član uzorka

μ – aritmetička sredina uzorka

N – broj elemenata u uzorku

Standardna devijacija

- *Standardna devijacija* (σ)
  je u statistici apsolutna mera disperzije u osnovnom skupu.
- Ona nam govori koliko elementi nekog skupa odstupaju od aritmetičke sredine skupa.

Standardna devijacija

- Primer

Otišli ste da pecate na jezero, ali Vam je rečeno da su se ranije u jezeru izlivali toksini iz obližnje fabrike. Rekli su Vam da je za čoveka štetno ako u ribi koju pojede ima 7mg toksina. Takođe su Vam rekli da u proseku u ribi ima 4mg toksina.

Da li biste pojeli ribu?

Da li biste pojeli ribu ako znate da je standardna devijacija 1mg?

A u slučaju da je standardna devijacija 4mg?

Značajnost

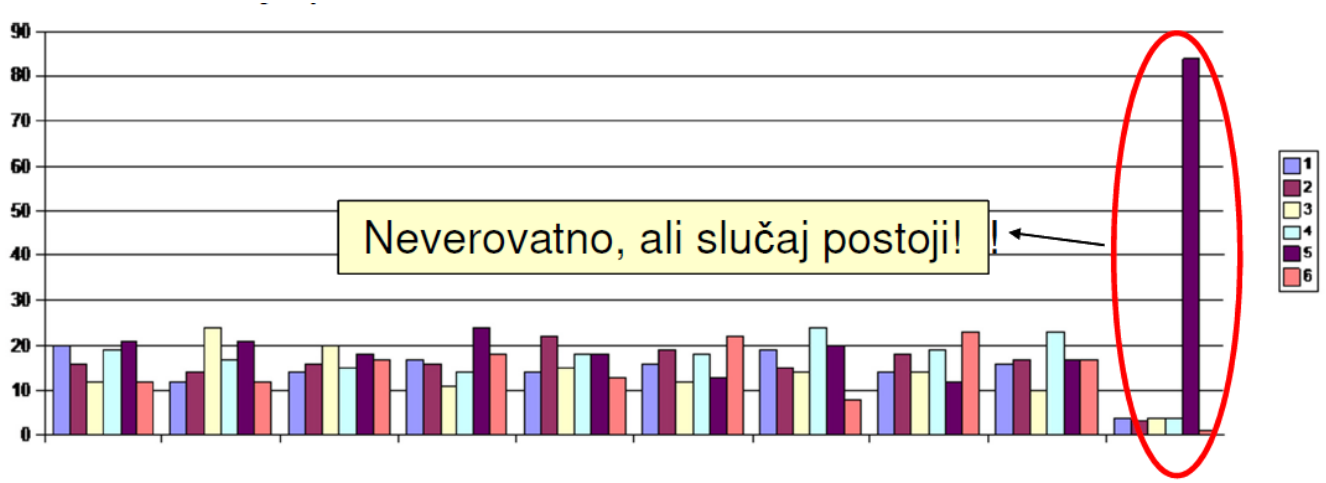
- Primer: testirati da li je kocka dobro balansirana!

- Na 6 bacanja: 3 puta broj 5
  - Uzorak je suviše mali pa rezultat nije značajan
- Na 1 milion bacanja: 900.000 puta broj 5
  - Rezultat je ubedljiv, ali ga je vrlo teško postići
- Na 1.000 bacanja: 248 puta pada broj 5
  - Šta treba zaključiti?

Uzorkovanje

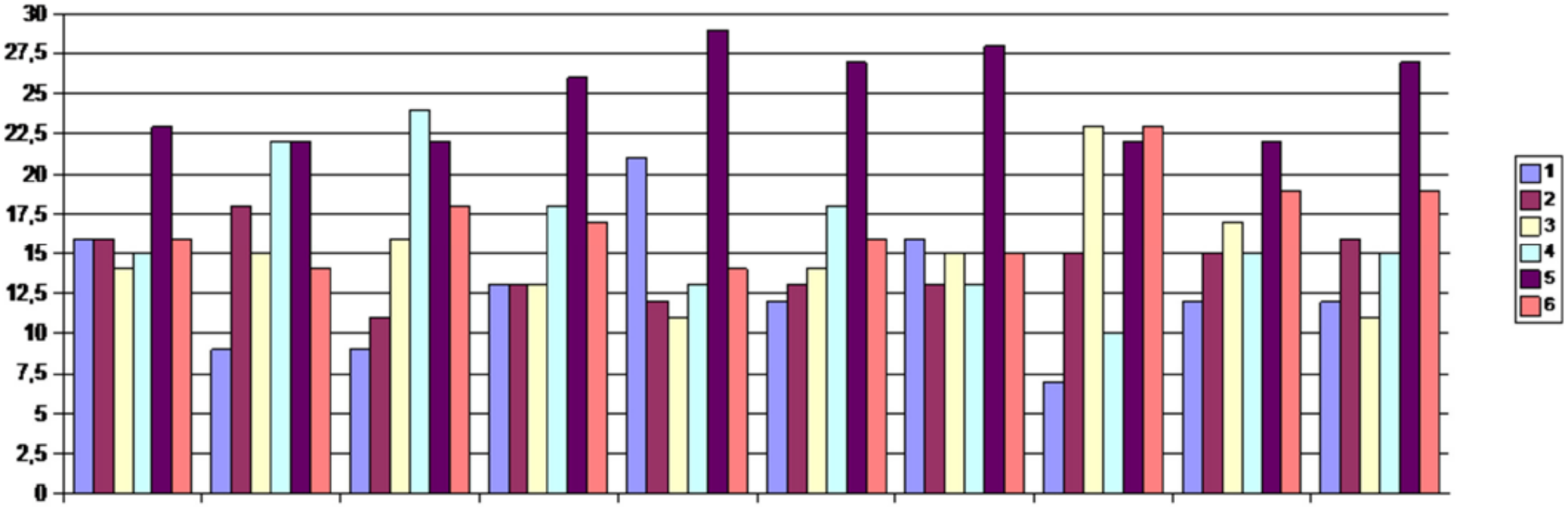
- Posmatramo distribuciju (10 serija od po 100 bacanja)

- Ako je kocka nameštena, 5 će se pojaviti više puta u svakoj seriji bacanja!



Uzorkovanje

- Šta zaključiti o sledećem ogledu (10 serija od po 100 bacanja)?
- Ovoga puta je trebalo imati sreće 10 puta što nije razumna pretpostavka



Značajnost

- Prag značajnosti zavisi od testa:
  - Za kockice, brojimo slučajeve
  - Za anketu, moramo voditi računa o uticaju
    (ne smemo navesti da mi glasamo za X, itd.)
  - Za meru sličnosti između dva pismena zadatka,
    mora se voditi računa o tome da nema baš 50 načina
    da se odgovori na izvesna pitanja…

Istraživanja i uzimanje uzorka

Istraživanja i uzimanje uzorka

- Statistika nam omogućava da zaključujemo o osobinama neke populacije na osnovu malog dela te populacije koji nazivamo uzorak.
- Uglavnom nije moguće ispitati celu populaciju (zbog vremena ili novca koji jedan za to potreban). Zato uzimamo uzorak.
- Kreiranje uzoraka i pravljenje zaključaka na osnovu tih uzoraka je najrasprostranjenija i najvažnija primena statistike.

Validnost uzorka

- Da bi zaključak o nekoj populaciji bio validan, uzeti uzorak mora da bude predstavnik te populacije.
- Da bi zadovoljio ovaj uslov uzorak mora da bude veliki i nepristrasan.

Validnost uzorka

- Primer

Literary Digest je od 1920. sprovodio istraživanje o pobedniku Američkih izbora. Iako je nekoliko godina bio uspešan, 1936. rezultat istraživanja je bio pogrešan.

Časopis je svojim pretplatnicima slao anonimne upitnike i poslao bi oko 10 miliona upitnika, a dobio bi oko 2 miliona odgovora.

Interesantno, istovremeno je druga agencija sprovela istraživanje nad manjim skupom ljudi (4.500) i oni su imali ispravnu predikciju.

Zašto je tako veliki uzorak imao loš rezultat?

Validnost uzorka - Primer

- Pristrasnost: pretplatnici časopisa su uglavnom bogatiji, konzervativni ljudi koji će verovatno radije glasati za Republikance.

Veličina uzorka

- Šta je dobra veličina uzorka?
- Zavisi od mnogo faktora:
  - veličina posmatrane populacije
  - ekonomska isplativost
  - nivo preciznosti koji je potrebno postići
  - pitanja koja se istražuju
  - …

Veličina uzorka

- Većina istraživanja o mišljenju o nekoj temi ima između 1.000 i 2.000 ispitanika.
- Preciznost dobijena uzimanjem većeg skupa uglavnom nije vredna troška.

Biranje uzorka

- Najbolji način je nasumično odabrati jedinke u uzorku.
- Ipak u praksi ovo je teško realizovati.

Biranje uzorka – nekoliko metoda

- Stratifikacija

  populacija se podeli na grupe i iz svake grupe se bira nasumično

- Klasterovanje

  populacija se podeli na grupe i nasumično se izabere nekoliko grupa

Biranje uzorka - Primer 1

Radio stanica se bavi istraživanjem o legalizacije marihuane.
Nakon javljanja slušalaca zaključak je da je 78% ispitanika podržalo predlog i da vlada odmah treba da donese zakon.

- Ispitanici koji su se javili su samo oni koji slušaju tu radio stanicu (koja možda podržava takav stav).
- Uz to, javili su se samo oni ispitanici kojima je ova tema jako važna.

Biranje uzorka - Primer 2

Najčešće se prilikom istraživanja mišljenja o nekoj temi biraju nasumični brojevi telefona. Korisnici se pozivaju i odgovaraju na pitanja.

- Da li ipak i ovde postoji pristrasnost?
- Najsiromašniji ljudi, beskućnici i slične grupe nemaju telefon.

Biranje uzorka

- U dobrim istraživanjima će pisati kolika je margina greške.
  Ova greška se upravo može desiti zbog problema sa uzorkom.

- Margina greške je statističko izražavanje količine slučajne greške prilikom uzimanja uzorka.
  Obično se definiše i kao radijus intervala pouzdanosti za određenu statistiku istraživanja.
  To je cena koju plaćamo jer ne možemo anketirati celu populaciju.

Biranje uzorka - Primer 3

U januaru su rezultati ankete za nekog predsedničkog kandidata pokazivali da ima 54% podrške. Margina greške je 5%.

U junu, rezultati ankete pokazuju da je podrška 56% i novinari zaključuju da se podrška povećala.

- Da li je zaključak ispravan?
- Prvi rezultat sugeriše da je podrška između 49% i 58%, a drugi da je podrška između 51% i 61%.

Formulisanje pitanja u anketi

- Pitanja u anketi ne smeju da budu pristrasna ili višeznačna.

  Svako ko odgovara na anketu mora na isti način da razume pitanja
  i da na njih iskreno odgovori.

Formulisanje pitanja u anketi

- Ipak, ove uslove u vezi pitanja nije lako postići.
- Primer sa „*brojem partnera*“.
- Primer: „*Da li čitate Politiku?*“
  - Da li svaki dan ili ponekad ili jednom nedeljno? Da li ceo list ili samo delove?
- Primer: „*Da li pijete puno alkohola?*“
  - Šta znači puno? Za različite ljude to može biti različita vrednost.

Formulisanje pitanja u anketi

- Dobre ankete prvo „testiraju/probaju“ pitanja na manjem uzorku!

Korelacija

Korelacija

- Korelacija je međuodnos ili međusobna povezanost između različitih pojava predstavljenih vrednostima dve promenljive.
  Pri tome povezanost znači da je vrednost jedne promenljive moguće s određenom verovatnoćom predvideti na osnovu saznanja o vrednosti druge.

  Korelacija predstavlja i obrazac variranja promenljivih u zavisnosti od načina na koji su povezane.

Korelacija

- Korelacija je često pogrešno shvaćena.
- Kada kažemo da su dve promenljive **A** i **B** u korelaciji, to ne znači da među njima postoji uzročno posledična veza!
- Naći uzrok nekog ponašanja jako je teško i predmet je naučnih istraživanja.

Korelacija

- A i B su u korelaciji može da znači:
  - A je uzrok pojave B
  - B je uzrok pojave A
  - A i B su slučajno povezani bez bilo kakve uzročne veze između njih
  - A i B su oboje zavisni od nekog trećeg faktora C

Korelacija - primeri

- Đaci koji puše cigare – njihove ocene su lošije
- Cena kafe u Oregonu – količina kiše koja pada u Oregonu
- Broj dimnjaka u kući – broj dece u toj kući

Korelacija - primeri

Monasi u Kini: Pomračenje meseca izaziva nebeski pas koji pojede mesec. Zato moraju da udare u veliki gong da bi psa oterali.

- Upravo pogrešno razumevanje korelacije, uzroka i posledice vodi mnogim sujeverjima.

Regresija ka sredini i sujeverje

Regresija ka sredini

- Regresija ka sredini
  se javlja kada kod dve vrednosti koje nisu u savršenoj korelaciji,
  a ekstremne vrednosti jedne promenljive su u korelaciji sa neekstremnim vrednostima druge promenljive (često u korelaciji sa srednjom vrednosti druge promenljive).

- Pojam je uveo engleski naučnik Frensis Galton
  (statističar, sociolog, psiholog, antoropolog, …).

Regresija ka sredini
primer sa sinovima i očevima

Frensis Galton je ispitivao odnos visine između očeva i sinova:

- Otkrio je očigledno: visoki očevi imaju visoke sinove, a niski očevi imaju niske sinove.
- Međutim, otkrio je još nešto iznenađujuće…
- Veoma visoki očevi imaju sinove koji su niži od njih.
- Veoma niski očevi imaju sinove koji su viši od njih.

Šta ovo znači?

Regresija ka sredini
primer sa sinovima i očevima

Ovo je primer **nesavršene korelacije** dve promenljive.

Mnogo toga utiče na visinu dece.

- Majčina visina.
- Način odgajanja, ishrane, mesto odgajanja, bavljenje različitim aktivnostima itd.
- Puno faktora mora da se poklopi da bi osoba bila ekstremno visoka ili niska.

Regresija ka sredini
primer sa sinovima i očevima

- Verovatnoća da se svi uslovi poklope je jako mala.
- To objašnjava zašto je slučaj ekstremno visokog oca u korelaciji sa neekstremnim vrednostima visine sina.

Tj. ova pojava je predvidiva i naziva se regresija ka sredini.

Regresija ka sredini
primer sujeverja

- Ranije pomenuti primer vrhunskih sportista koji se boje da se pojave na naslovnoj strani časopisa kako ne bi pokvarili svoje rezultate.

Baksuz, sujeverje ili prosto regresija ka sredini?

Literatura

- Normand Baillargeon, *A Short Course in Intellectual Self-Defense*, UQAM, Seven Stories, 2008.
- John Allen Paulos, *A mathematician reads the newspaper*, New York: Anchor Books, 1995

Pogledajte i…

- Uzorkovanje za statističke analize
- Stratifikacija – jedan od načina za uzimanje uzorka za statističku analizu
- Klasterovanje – jedan od načina za uzimanje uzorka za statističku analizu
- Regresija ka sredini
- Logička zabluda

Hvala

Staša Vujičić Stanković

stasa@math.rs

www.matf.bg.ac.rs/~stasa