Statistika
(za biologe, 1.dio)

Bojan Basrak

2008

1. Statistika

- Uvod (1p)
- Opisne statistike (1.5p)
- Osnovni pojmovi vjerojatnosti (1.5p)
- Modeli razdioba i slučajne varijable (2p)
- Procjena parametara modela (2p)
- Testiranje statističkih hipoteza (3p)
- Jednofaktorska analiza varijance (1p)
- Jednostavna linearna regresija i korelacija (2p)

Ispit

- dva kolokvija: obavezno za potpis: 10% svekupno,

- dva zadatka obradjena programskom jeziku R:

i) normalni uzorak - ar. sredina, median, histogram, intervali pouzdanosti za parametre ili – χ^2–test nezavisnosti ili – zakljucivanje o proporciji u populaciji

ii) regresija ili anova

- popravni (pismeni i usmeni dio ispituje cijelo gradivo), u istom terminu studenti s ispricnicom mogu popraviti jedan od kolokvija.

Uvod

Statistika - skup ideja i metoda koje se bave sljedećim temama:

- prikupljanje podataka,
- prikaz podataka,
- zakljucivanje iz podataka,

Naravno, i njima se možemo baviti amaterski i profesionalno (znanstveno).

Statističari radije koriste "ozbiljniji" jezik, pa govore o sljedeća tri aspekta istraživanja:

- dizajn pokusa, experimental design
- deskriptivna statistika, descriptive statistics
- statističko zaključivanje, statistical inference

Što je statistika?

- Predmet ne bas popularan medju studentima biologije

- Statistika se ne bavi prikupljanjem poznatih činjenica, statistika utvrdjuje što su zapravo činjenice – u znanosti, ekonomiji, politici, na sudu

- Ona je gotovo ekskluzivni način na koji znanstvenici potvrdjuju ili opovrgavaju teorije pokusima.

Statistička analiza pokusa je najčešće i prvi korak prema kreiranju teoretskog modela i novih znanja.

- Statistika je možda primjenjena filozofija u stvari.

- Njena je uloga ključna u modernoj znanosti. Posebno je bliska veza s biologijom.

- Matematičari rado kažu da je statistika i izmjeriva funkcija slučajnog uzorka!?! Oni zaista žele reći da je statistika svaki broj izveden iz uzorka.

- Jezik kojim se služi statistika je jezik vjerojatnosti, a ako povjerujemo kvantnim fizičarima, to je ujedno i jezik kojim govori priroda.

Što je vjerojatnost?

- Uh! To je jedno od najtežih pitanja, osim ako ne pitate matematičare. Naime za njih je to funkcija slučajnih dogadjaja s vrijednostima izmedju 0 i 1, i vrlo zgodna pritom.

Definirati vjerojatnost stvarnih dogadjaja nije jednostavno iako svi imamo neku intuitivnu ideju o tome.
Npr. kolika je vjerojatnost da

- će večeras padati kiša?
- će svi upisani studenti diplomirati?
- će dionica Plive doseći cijenu od 4000 kuna prije kraja godine?
- će u 20 bacanja kocke pasti 20 petica?

Čak i kad prihvatimo matematičku ili intuitivnu definiciju, vjerojatnost ostaje puna naizgled paradoksalnih zaključaka, v. npr. prisoner's dilemma:

Ujutro će odlukom predsjednika biti oslobodjena dvojica od trojice zatvorenika: A, B i C. Sva tri zatvorenika imaju jednaku šansu biti oslobodjena po onome što se zna o dosadašnim predsjednikovim abolicijama.

Zatvorenik A u razgovoru sa stražarem pokuša otkriti svoju sudbinu večer prije. Stražar, iako zna, rezolutno odbija reći hoće li A biti oslobodjen. Tada A zamoli stražara da mu otkrije ime bar jednog od preostale dvojice koji će otići na slobodu. Nakon puno molbi stražar mu otkrije jedno ime. A A? Problijedi, sasvim zbunjen, pomislivši kako su se njegove šanse za aboliciju upravo spustile sa 2/3 na 1/2.

Statistika i vjerojatnost u medijima

Jutarnji list je prije par godina prenio informaciju o američkoj studiji koja pokazuje veliku korelaciju izmedju slušanja seksualno eksplicitnih pjesama i seksualne aktivnosti adolescenata. Naslov je glasio otprilike:

"Slušanje seksualno eksplicitnih pjesama uzrokuje veću seksualnu aktivnost tinejdžera"

Britanski list Big Issue prenio je informaciju o Viagri, ističući u naslov:

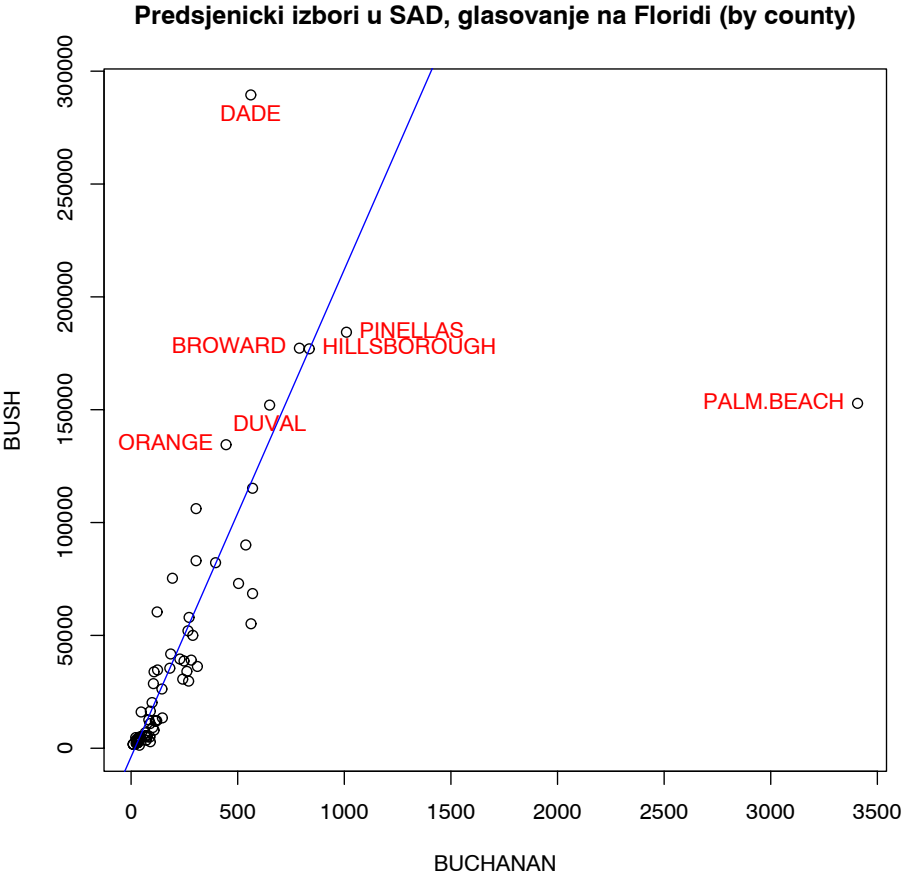
"Viagra special report: Sex drug linked to 31 deaths in one year"

No tekst članka ne daje nikakve informacije o tome zašto bi 31 mrtva osoba medju korisnicima Viagre bila imalo zanimljiv broj. U bilo kojoj populaciji, pa i medju čitateljima novina, za očekivati je da u godini dana netko umre. Treba li zabraniti novine kada taj broj bude veći od npr. 31?

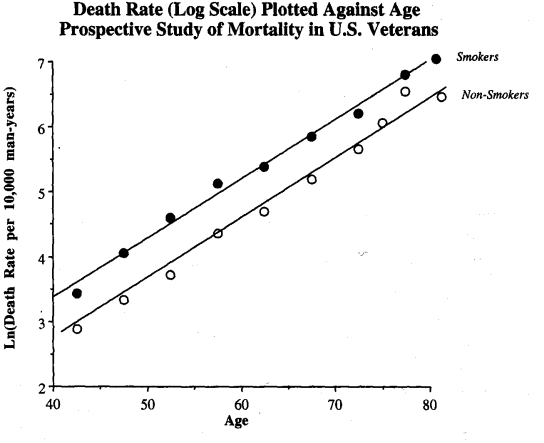
U proljeće 2002, novine su u Britaniji dominirali naslovi o tzv. MMR (measles, mumps, rubella) cjepivu. Uzrok je bio članak u časopisu Lancet u kojem su autori na uzorku od 12 djece utvrdili da sva imaju gastroent. problema i usporen razvoj, te pokazuju znakove autizma. Osmero djece primilo je MMR cjepivo. No 76-92% djece njihove generacije je takodjer primalo cjepivo. Ima li razloga za paniku?

Prema medijima svakako. Istovremeno, premijer Tony Blair odbija reći da li su i njegova djeca cijepljena. No zapravo podaci ne pokazuju ništa spektakularno, a to je potkrijepljeno i naknadnim studijama. Ipak, nekoliko roditeljskih parova je završilo na sudu sporeći se oko toga treba li cijepiti dijete ili ne.

Podaci/Data

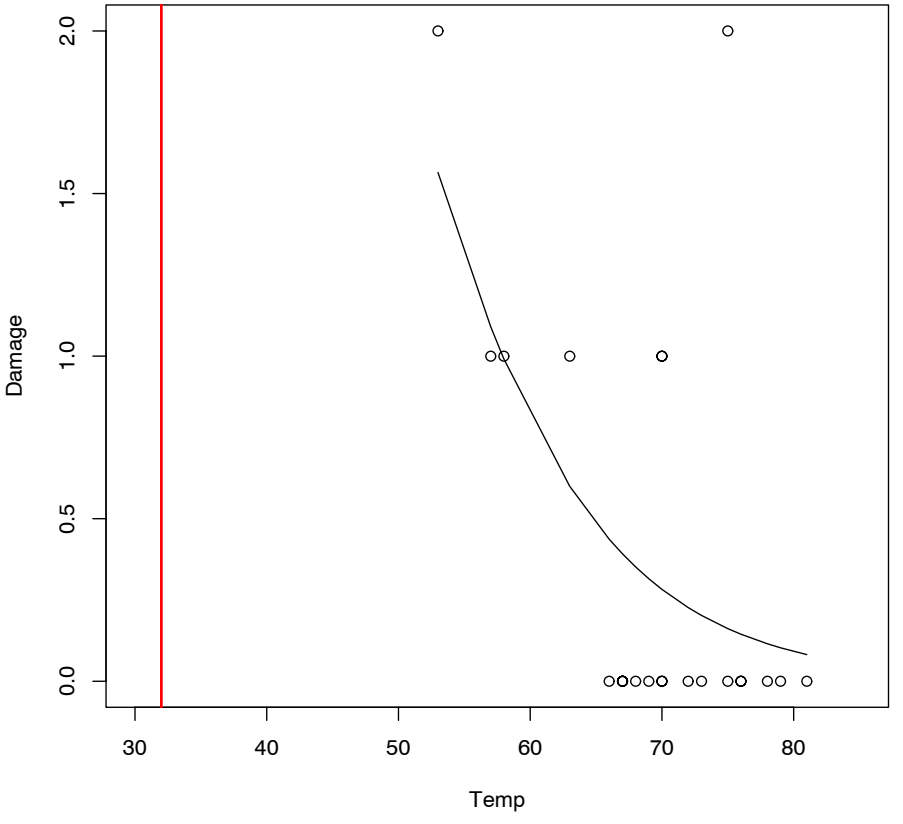


Na predsjedničkim izborima u SAD 2000. bilo je puno kontroverzi o ishodu i prebrojavanju glasova, posebnu pažnju privlačio je ishod u okrugu Palm Beach na Floridi.

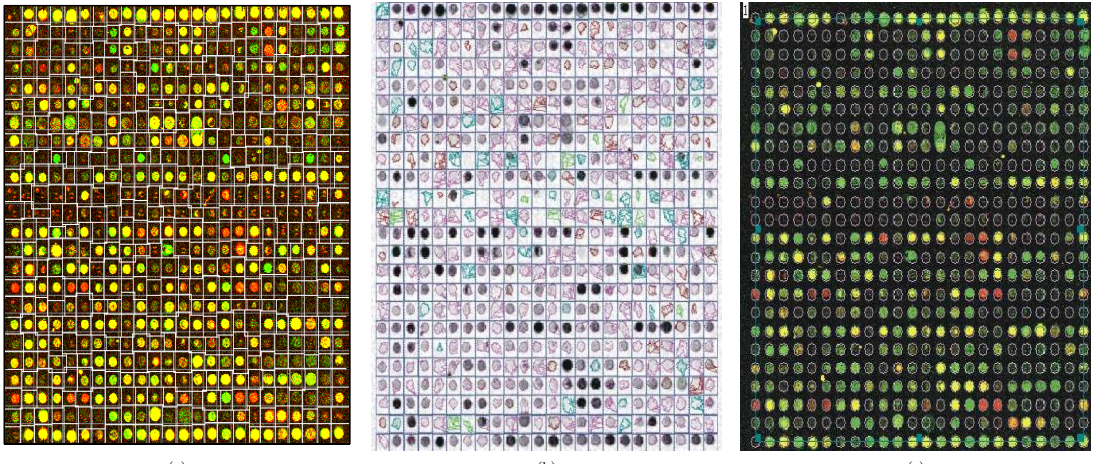


Source: US Surgeon General

Stope smrtnosti pušača i nepušača



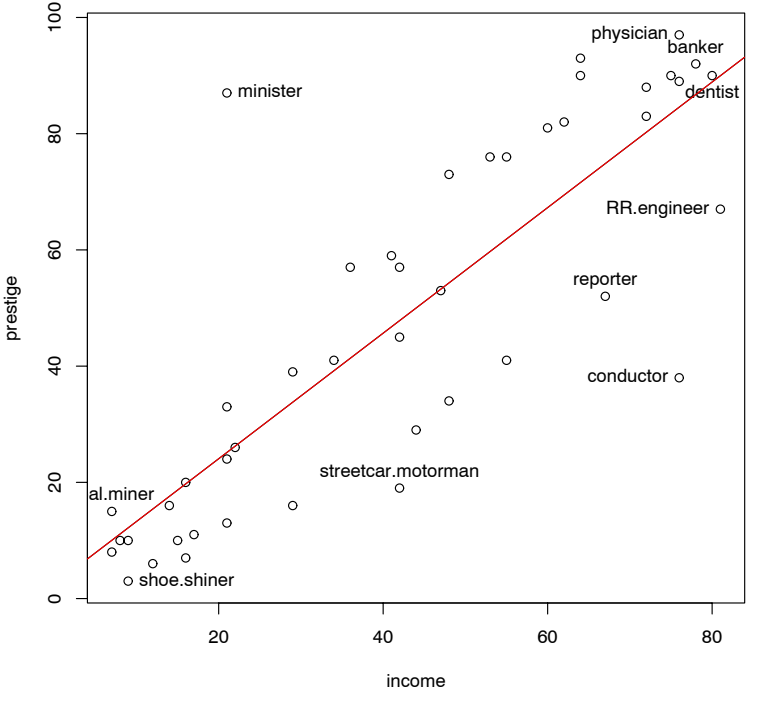
Godine 1986. NASA je lansirala space shuttle Challenger unatoč upozorenjima vlastitih inženjera da podaci sugeriraju nepouzdanost tzv. O-rings na niskim temperaturama



(a) (b) (c)

Figure 2: Comparative gridding results of our method (a) with two widely used microarray image analysis tools: (b) the Spotfinder and (c) the ScanAlyze.

DNA microarrays predstavljaju jednu od najbitnijih novih tehnologija u molekularnoj biologiji, a rezultat svakog pokusa su deseci tisuća brojeva.



Usporedba statusa i primanja za razna zanimanja, SAD 1950-tih.

Populacija - predmet našeg istraživanja, skup svih potencijalnih (opaženih i neopaženih) podataka na nekoj grupi jedinki.

(Uzoračka) jedinka je jedan izdvojen objekt na kojem potencijalno možemo obaviti mjerenja. Npr. svatko od vas je jedinka u populaciji studenata ovog sveučilišta. Drugi primjeri su: četvrti obavljeni pokus, drvo ispred zgrade fakulteta, jedan pogodak na rukometnoj utakmici, itd.

Cilj statističkog istraživanja je izreći nešto o cijeloj populaciji na osnovu manjeg broja podataka iz populacije, tj. na osnovu uzorka. Pri tome, statističar mora ocjeniti i neizvjesnost kod izricanja takvih tvrdnji.

Uzorak - skup mjerenja (podataka) iz populacije. Tipično svi podaci prikupljeni tijekom istraživanja.

Statističko obilježje ili varijabla je karakteristika jedinki iz populacije, npr. visina u populaciji studenata.

Neizvjesnost je osnovna karakteristika statističkog istraživanja. Naime zaključci se izvode na osnovu dijela populacije, a pojave koje promatramo, baš kao i naša mjerenja neizbježno pokazuju varijabilnost (npr. nivo ekspresije nekog gena u stanicama). Često kažemo da podaci sadrže slučajnu komponentu.

Dakako, statistički će zaključci nužno odražavati tu neizvjesnost.

Primjeri statističkih problema

Primjer

Nakon 100 bacanja novčića, imamo 60 pisama i 40 glava. Sumnjamo da je novčić neispravan. Želimo biti 95 % sigurni u to prije nego zovemo policiju. Kako provjeriti hipotezu o neispravnosti novčića?

Primjer

Na izborima za gradonačelnika pravo glasa ima 80000 osoba. Na uzorku od 1000 ispitanika otkrivamo da sadašnji gradonačelnik uživa potporu 54.4% njih. Želimo procijeniti vjerojatnost da će gradonačelnik zadržati svoj posao.

Primjer

Analiza bioloških nizova je jedan od najvažnijih statističkih problema danas. Posebno puno pažnje istraživači posvećuju:
– forenzičkoj DNA analizi – korištenjem PCR tehnike, uzorak s mjesta zločina se amplificira. Zatim se lociraju tzv short tandem repeats – STR i to više njih (FBI ih koristi 13). Postavlja se pitanje: koliko je vjerojatno da sl. odabrana osoba u referentnoj populaciji ima pronadjeni profil? Koja je to referentna populacija? Kakvu analizu forenzičar treba prijaviti sudu?
– rekonstrukcija zajedničke evolucije DNA nizova (i evt. filogenetskog stabla) korištenjem vjerojatnosnog modela
– lokalno poravnanje DNA (ili proteinskih) nizova. Pitanje je kolika mora biti duljina i kvaliteta poravnanja da bismo posumnjali na zajedničko podrijetla dva gena ili proteina?

Primjer

Kontrolom istovrsnih proizvoda na dva stroja pronadjeni su sljedeći podaci

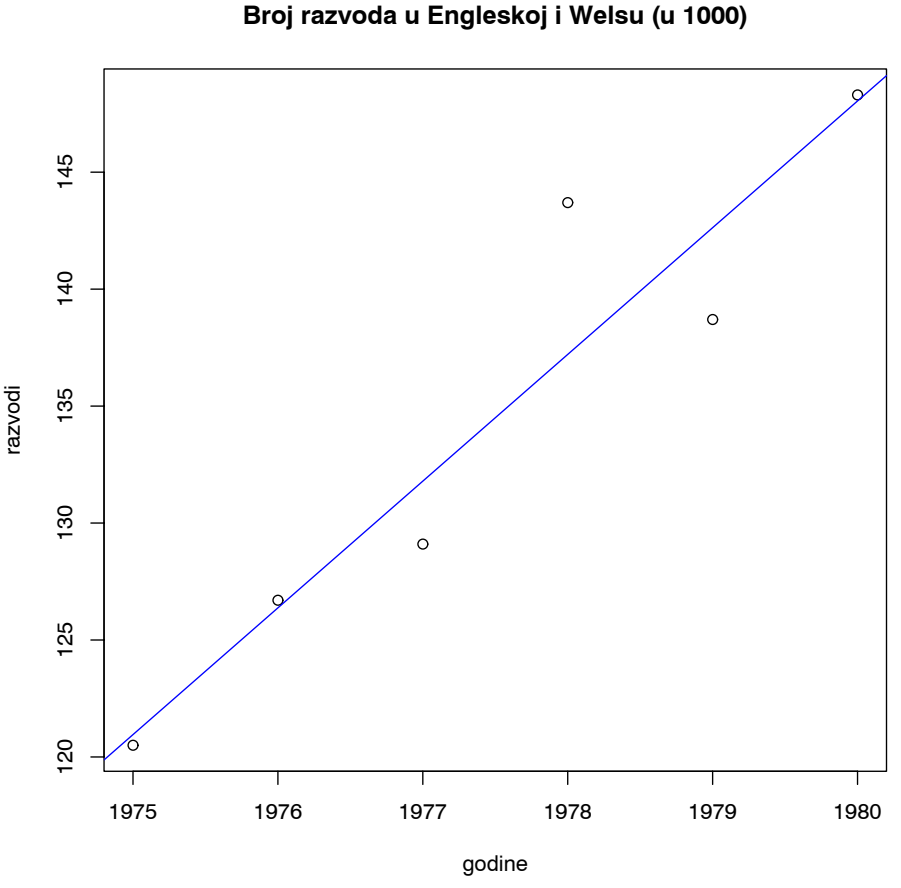
	Stroj A	Stroj B
dobri proizvodi	240	380
loši proizvodi	20	24
ukupno	260	404

Dakle loših prozvoda dobivenih na stroju A je $7.7\% \approx 20/260 \cdot 100\%$, a na stroju B $5.9\% \approx 24/404 \cdot 100\%$. Da li je razlika statistički značajna?

Primjer

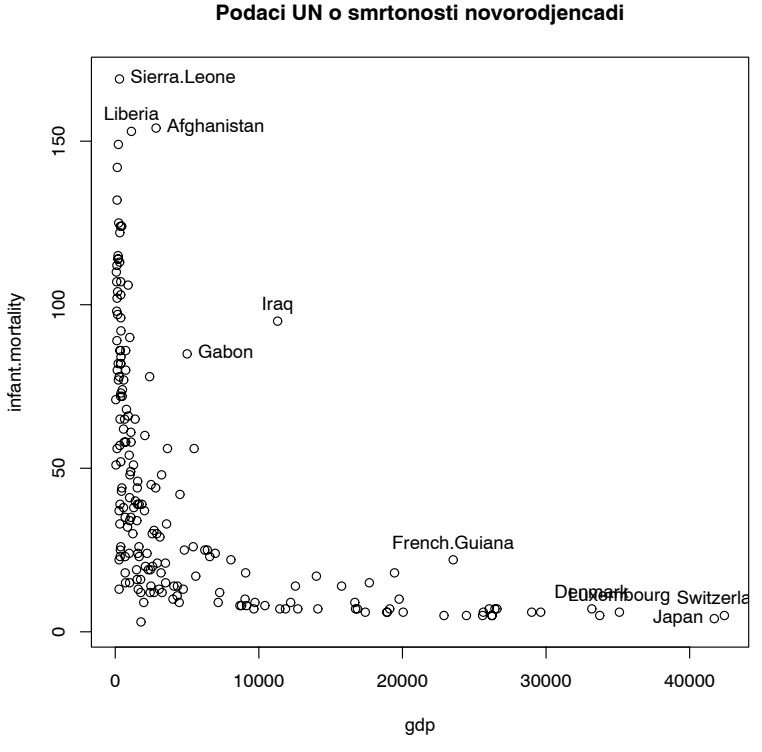
Podaci o broju razvoda u Engleskoj i Welsu u periodu od 1975. do 1980. su u donjoj tablici.

godina	1975	1976	1977	1978	1979	1980
broj razvoda u tisućama	120.5	126.7	129.1	143.7	138.7	148.3



Podaci sugeriraju da postoji rastući trend. Da li je on statistički značajan? Može li se rast modelirati linearnom jednadžbom oblika $y = \alpha + \beta x$?

Primjer



Podaci Ujedinjenih Naroda o BDP i smrtnosti novorodjenčadi. Pitanje je da li su GDP i smrtnost korelirane varijable? Možemo li to testirati?

Primjer

Istraživanje hipertenzije na uzorku od 180 osoba pokazalo je sljedeće rezultate

	nepušač	blagi pušač	teški pušač	ukupno
normalni tlak	48	26	19	93
povišeni tlak	21	36	30	87
ukupno	69	62	49	180

Pitanje je postoji li statistička zavisnost izmedju tlaka i pušenja? S kolikom sigurnošću možemo izreći svoju tvrdnju?

2. Deskriptivna statistika

U gotovo svim granama znanosti i poslovanja danas smo suočeni s iznimnim količinama podataka. Pitanje je kako rezultate mjerenja sažeto predstaviti sebi ili drugima?

Podaci mogu biti

- kvantitativni (numerički), kao u većini naših primjera. Kod numeričkih podataka razlikujemo diskretne i neprekidne.
- kvalitativni (kategorijski), ako se ne mogu predstavljati na nekoj prirodnoj numeričkoj skali.

Nadalje podaci mogu biti jednodimenzionalni ili višedimenzionalni, ovisno o tome da li za svaku jedinku naš uzorak sadrži jedno ili više mjerenja.

Razmislite da li su podaci kategorijski u sljedećim slučajevima

- Odgovorima na 10 pitanja o upotrebi lijeve ili desne ruke kod: pisanja, bacanja, rezanja, itd. studente možemo klasificirati na lijevoruke ili desnoruke.
- Iz podataka o primanjima, sve zaposlene u HR možemo svrstati u grupe visokih, srednjih i niskih primanja.
- Ankete o raspoloženju gradjana HR prema pristupanju EU tipično odvajaju one koju su za, protiv i neodlučni.
- Prema uspjehu na ispitima prve godine studente možemo dijeliti na uspješne, srednje uspješne i neuspješne.
- Školjke prikupljene na terenu možemo dijeliti prema vrsti.

Razmislite da li su podaci kategorijski u sljedećim slučajevima

- Odgovorima na 10 pitanja o upotrebi lijeve ili desne ruke kod: pisanja, bacanja, rezanja, itd. studente možemo klasificirati na lijevoruke ili desnoruke. numdis
- Iz podataka o primanjima, sve zaposlene u HR možemo svrstati u grupe visokih, srednjih i niskih primanja. numdis ili kat
- Ankete o raspoloženju gradjana HR prema pristupanju EU tipično odvajaju one koju su za, protiv i neodlučni. kat
- Prema uspjehu na ispitima prve godine studente možemo dijeliti na uspješne, srednje uspješne i neuspješne. numdis ili kat
- Školjke prikupljene na terenu možemo dijeliti prema vrsti. kat

Mi ćemo se uglavnom baviti jednodimenzionalnim numeričkim mjerenjima, kao što je npr. visina ljudi ili težina ploda. Ako u uzorku imamo n ponovljenih mjerenja, reprezentirat ćemo ih s n brojeva

$$x_1, x_2, \ldots, x_n.$$

Ako se radi npr. o visinama slučajno odabranih muških studenata oni bi mogli izgledati ovako

159	188	175	176	177	168	162	188
183	187	187	162	184	161	180	169
195	171	170	199	181	169	189	191
172	182	183	178	180	165	185	202
183	187	188	182	163	179	178	188

Ako se neki podaci pojavljuju više od jednom tada mjerenja možemo prikazati u tablici koristeći frekvencije njihovog pojavljivanja. Dakle za rezultat a, njegova frekvencija se definira kao

$$f_a = \text{ broj pojavljivanja elementa } a \text{ u nizu } x_1, x_2, \ldots, x_n.$$

Često je interesantno znati postotak pojavljivanja odredjenog elementa u našem uzorku odn. njihove relativne frekvencije koje računamo po formuli

$$r_a = \frac{f_a}{n}$$

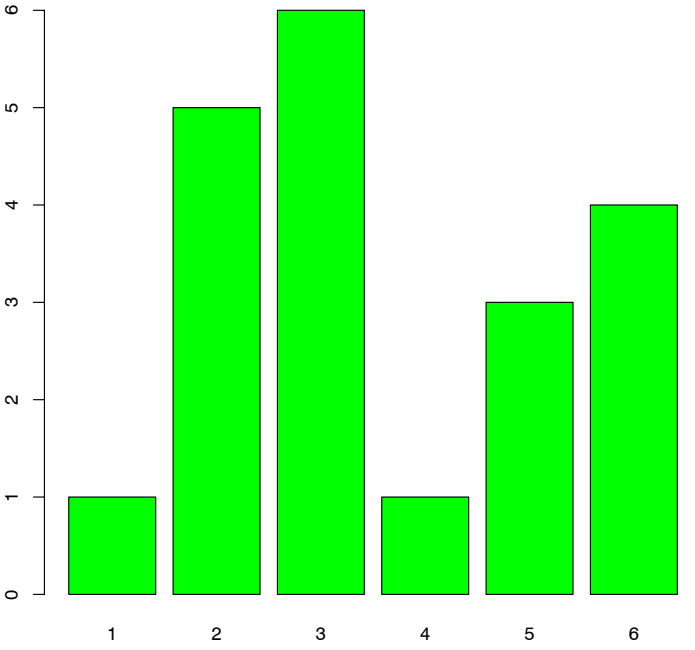
Kažemo da relativne frekvencije opisuju razdiobu (distribuciju) našeg uzorka.

Ako smo npr. bacali kocku 20 puta i zabilježili sljedeće rezultate: 6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3. Možemo napraviti sljedeću frekvencijsku tablicu

a	f_a	r_a
1	1	1/20
2	5	1/4
3	6	3/10
4	1	1/20
5	3	3/20
6	4	1/5

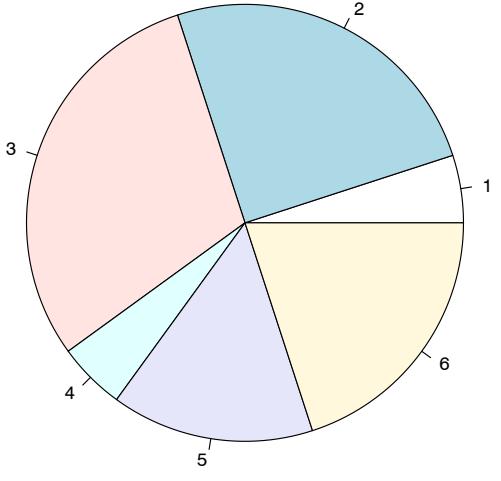
Grafički prikaz podataka

Podatke o bacanju kocke mogli bismo grafički prikazati pomoću stupčastog dijagrama



Rezultati bacanja kocke (koristeći apsolutne frekvencije).

. . . ili preko strukturnog (ili tortnog) dijagrama kakvi se često pojavljuju u tisku



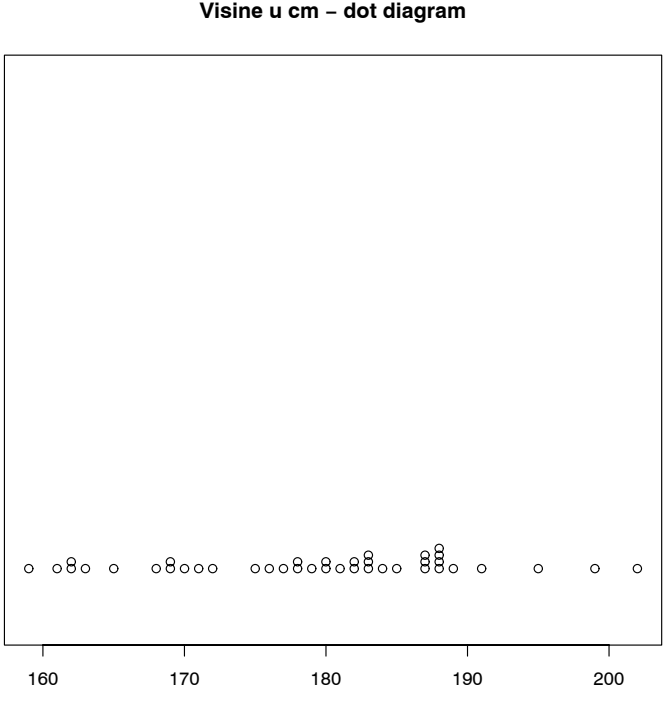
Rezultati bacanja kocke.

Za podatke o bacanju kocke kažemo da su diskretni jer mogu poprimiti samo konačno mnogo unaprijed poznatih vrijednosti, naime: 1, 2 ,3 , 4 , 5 ili 6. U tom slučaju za $x_1, x_2, \ldots, x_n$ kažemo da su mjerenja diskretnog statističkog obilježja X.

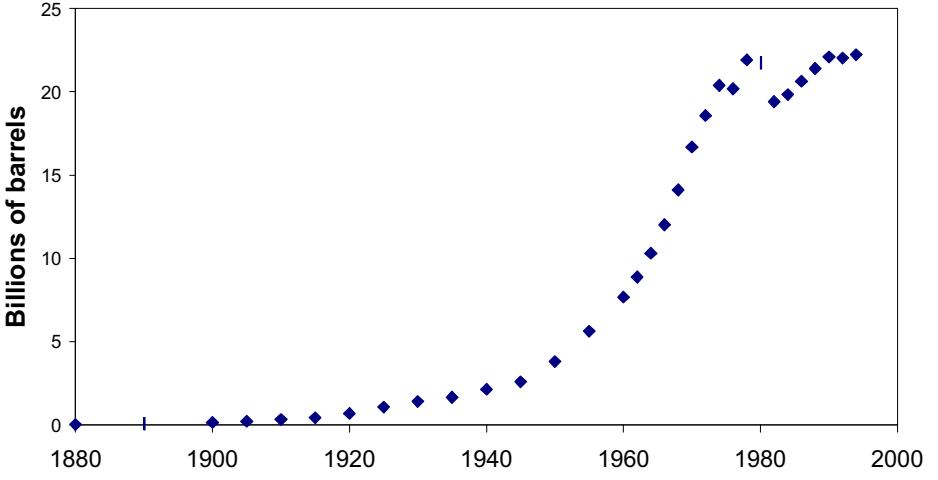
U praksi to nije uvijek slučaj. Posebno kada podatke možemo mjeriti praktično po volji precizno govorimo o neprekidnim numeričkim podacima ili o neprekidnim statističkim obilježjima. Takvima bismo prirodno mogli smatrati podatke o visinama u uzorku ljudi.

Podatke o neprekidnim numeričkim obilježljima prezentiramo najčešće

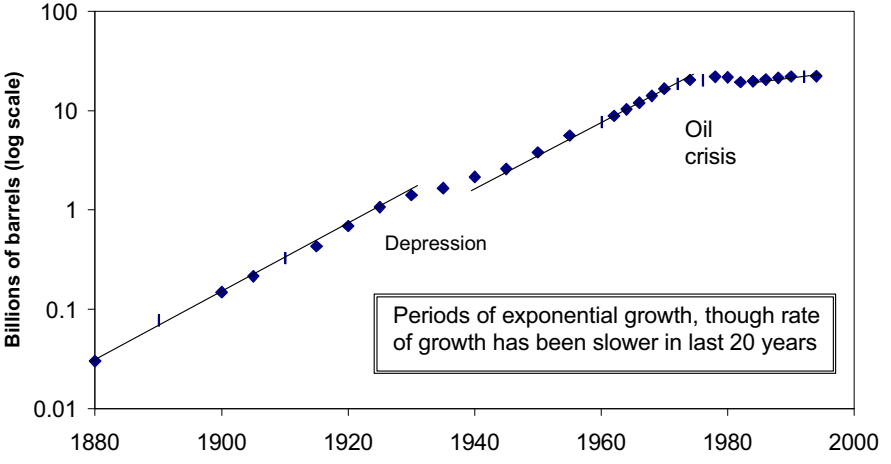
- ▷ točkovnim dijagramom (dot diagram)
- ▷ histogramom



Visine studenata u cm.



Svjetska proizvodnja nafte u barelima



Isti graf na logaritamskoj skali nešto je lakši za interpretiranje

Histogram je zapravo vrlo blizak stupčastom dijagramu, no podaci kod histograma su uvijek numerički, a mi ih grupiramo u razrede. Procedura za izradu histograma može se podijeliti u nekoliko koraka

▷ U uzorku $x_1, x_2, \ldots, x_n$ odredimo $x_{\min}$ i $x_{\max}$, te željeni broj razreda, recimo k.
▷ Iz formule $(x_{\max} - x_{\min})/k$ zaokruživanjem na gore, odredimo širinu razreda.
▷ Odredimo razrede $I_1, \ldots, I_k$, odn. intervale oblika $I_j = [a_{j-1}, a_j)$ za neke brojeve $a_0, \ldots, a_i$, koje zovemo granice razreda.
▷ Za svaki razred izračunamo apsolutnu i relativnu frekvenciju podataka koji u njega ulaze, te dobijamo tablicu oblika

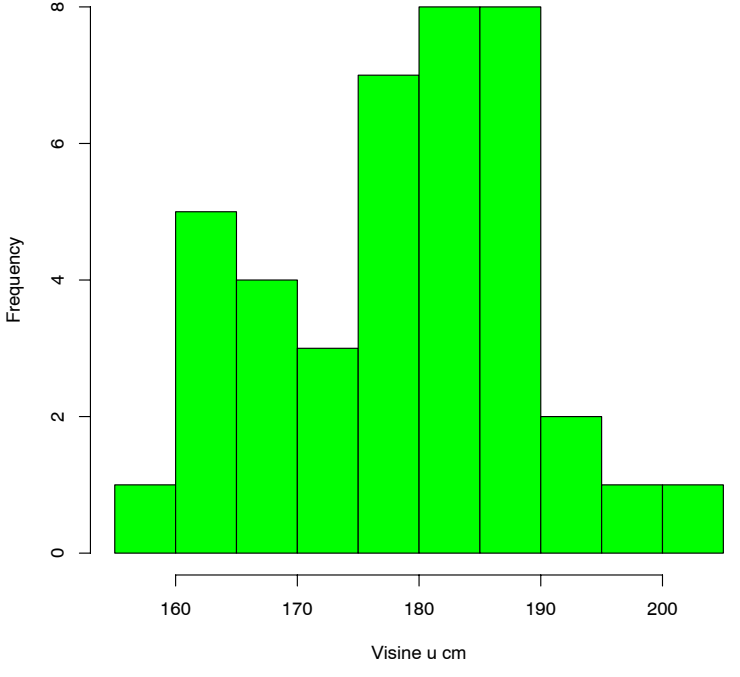
razred	granice	f_a	r_a
I_1	a_0, a_1	f_1	r_1
I_2	a_1, a_2	f_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	a_{k-1}, a_k	f_k	r_k

▷ Konačno nacrtamo graf na kojem iznad pojedinog razreda povučemo liniju na visini koja odgovara relativnoj frekvenciji tog razreda podijeljenoj s duljinom razreda. Alternativno ponekad crtamo histogram apsolutnih frekvencija povlačeći liniju na visini apsolutne frekvencije pojedinog razreda.

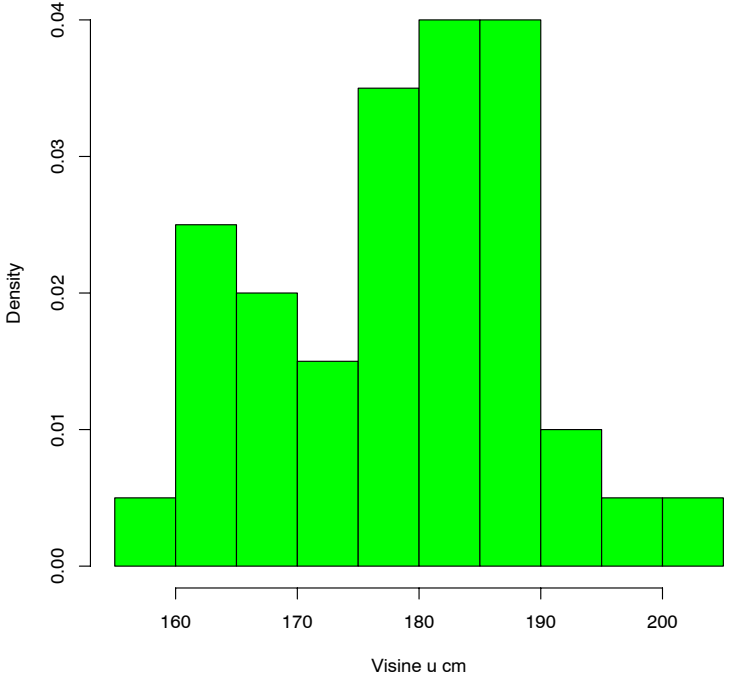
Kod izrade histograma bitno je da svaki podatak ulazi u jedan i točno jedan razred. Konačni izgled dijagrama kod raznih statističkih paketa u praksi ovisi o tretmanu podataka koji leže na granici dva razreda.

Obično su razredi intervali jednake duljine, no to ne mora uvijek biti slučaj. U tom općenitijem slučaju iznad razreda I_j liniju povlačimo na visini

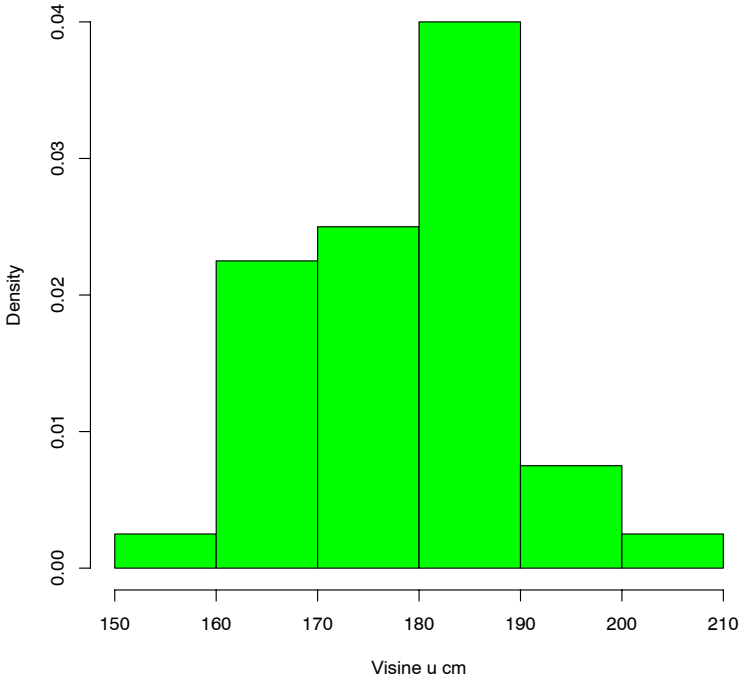
$$\frac{r_j}{a_j - a_{j-1}}.$$



Histogram apsolutnih frekvencija.



Histogram s razredima širine 5



Histogram s razredima širine 10

Ponekad se numerčki podaci prikazuju i pomoću tzv. stem and leaf dijagrama (Tukey, 1977). Npr. podaci o visinama se ovim dijagramom mogu prikazati na sljedeći način

15	9
16	1223
16	5899
17	012
17	567889
18	001223334
18	577788889
19	1
19	59
20	2

Ako ste primjetili mi smo već susreli i dvodimenzionalne numeričke podatke npr. promatrajući status i primanja različitih zanimanja ili podatke UN o različitim zemljama. Simbolički takvi podaci predstavljaju niz uredjenih parova realnih brojeva

$$(x_i, y_i),\ i = 1, \ldots, n.$$

Njih tipiňo ilustriramo koristeći tzv. scatterplot kakav smo već vidjeli u tim primjerima.

Mjere centra ili sredine uzorka

Aritmetička sredina uzorka $x_1, x_2, \ldots, x_n$ je

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Primjer
Za uzorak 6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3, koji smo sakupili bacanjem kocke, aritmetičku sredinu možemo naći na sljedeći način

$$\bar{x} = \frac{1}{20}(1 \cdot 1 + 2 \cdot 5 + 3 \cdot 6 + 4 \cdot 1 + 5 \cdot 3 + 6 \cdot 4) = 3.6.$$

A koliko ste vi očekivali?

Ako uredimo podatke $x_1, x_2, \ldots, x_n$ po veličini dobit ćemo uzlazni niz

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Posebno je $x_{\max} = x_{(n)}$, a $x_{\min} = x_{(1)}$.
Medijan se općenito definira kao onaj broj za koji je 50% uzorka i manje ili jednako od njega, ali i veće ili jednako od njega. No takvih brojeva može biti više zbog toga mi definiramo medijan na sljedeći način:

- ako je n neparan, medijan je

$$m = x_{(\frac{n+1}{2})}$$

- ako je n paran, medijan je

$$m = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Primjer

Ako rezultate bacanja kocke uredimo dobijamo niz:

$$1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6.$$

Kako je $n = 20$ paran broj medijan je sredina izmedju 3 i 3, dakle 3.

Što mislite koja je bolja mjera za centar uzorka, medijan ili aritmetička sredina?

Pitanje je vrlo važno. Ono je jedno od važnijih u trenutnoj predizbornoj kampanji u SAD. Republikanci ukazuju na gospodarski rast od 4.2%, koji dakle kaže i da se bogatstvo prosječnog gradjanina uvećalo za isti postotak. Demokrati upozoravaju da je medijan zarade gradjana u isto vrijeme pao, te da se povećalo siromaštvo.

Kao mod uzorka $x_1, x_2, \ldots, x_n$ definiramo vijednost koja se pojavljuje s najvećom frekvencijom. Za naš uzorak iz prethodnih primjera i mod bi dakle bio 3.

Ako se u uzorku dvije odn. više vrijednosti pojavljuju s najvećom frekvencijom govori se da je uzorak bimodalan, odn. višemodalan.

Primjetite da u uzorku koji nije dovoljno diskretiziran mod može biti čak i najmanja ili najveća vrijednost u uzorku, dakle sasvim na njegovom rubu, a ne u centru.

Mjere raspršenosti ili disperzije uzorka

Raspon uzorka je udaljenost najmanje i najveće vrijednosti, dakle

$$d = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Uočite: uzorci $1, 2, 2, 10$ i $1, 4, 7, 10$ imaju isti raspon, premda očito nisu jednako disperzirani. Zato ćemo tražiti bolju mjeru za disperziju.

Bilo koji broj s izmedju 1 i n, možemo jednoznačno prikazati u obliku

$$s = k + r$$

tako da je $k = 1, 2, \ldots, n$ njegov cijeli, a $0 \leq r < 1$ njegov razlomljeni dio. Ako želimo oznaku $x_{(s)}$ koristiti i za brojeve $1 \leq s \leq n$ koji nisu nužno cijeli, definiramo uz gornji prikaz

$$x_{(s)} = (1 - r)x_{(k)} + rx_{(k+1)}\,.$$

Uz ove oznake je medijan i za parne i za neparne n

$$m = x_{(\frac{1}{2}(n+1))}.$$

Gornji i donji kvartil definiramo kao veličine

$$Q_3 = x_{(\frac{3}{4}(n+1))}, \quad Q_1 = x_{(\frac{1}{4}(n+1))}.$$

Udaljenost izmedju ova dva broja zovemo interkvartilni raspon

$$d_Q = Q_3 - Q_1.$$

Brojeve

$$(x_{\min}, Q_1, m, Q_3, x_{\max})$$

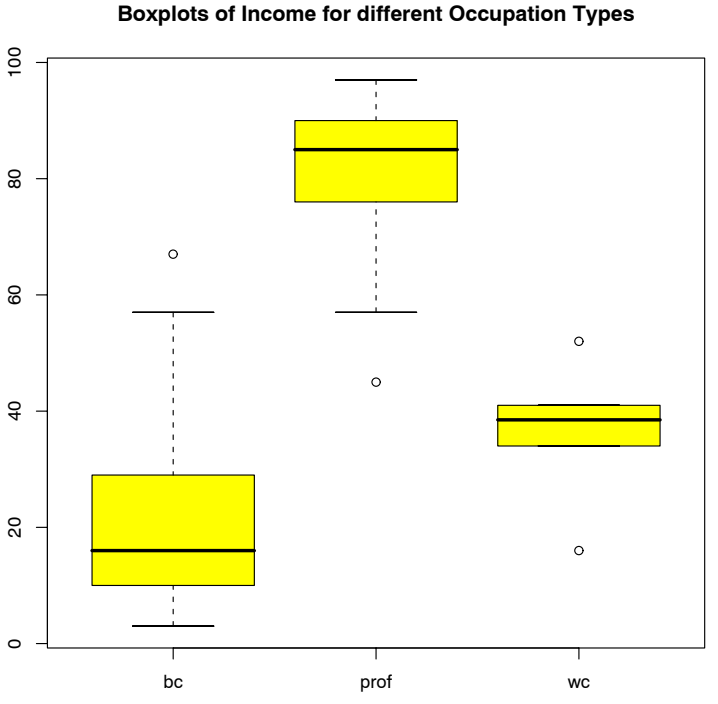
zovemo karakteristična petorka uzorka.

Primjetite da interkvartilni raspon ovisi o mjernim jedinicama u kojima su izraženi naši podaci, da bismo to izbjegli ponekad za nenegativne podatke koristimo koeficijent kvartilne varijacije

$$v_Q = \frac{d_Q}{Q_3 + Q_1}.$$

Postoji elegantan grafički prikaz uzorka preko njegove karakteristične petorke koji se zove box and whiskers dijagram (dakle dijagram kutije s brkovima ili pravokutni dijagram).

On nastaje tako što na brojevnom pravcu označimo medijan m i oko njega izgradimo "kutiju" od Q_1 do Q_3. Nakon toga na jednu i drugu stranu medijana pronadjemo podatak koji se nalazi najdalje od njega, ali ne dalje od jedan i pol d_Q. Povučemo linije od tih podataka do kutije. Sve podatke koji se nalaze izvan tog intervala oko medijana posebno označimo i smatramo ekstremnim.



Box and whiskers dijagram primanja za tri grupe zanimanja.

Za bilo koji $0 < \alpha < 1$ možemo definirati α-kvantil uzorka $x_1, x_2, \ldots, x_n$, kao

$$q_\alpha = x_{(\alpha(n+1))}\,.$$

gdje pretpostavljamo da je $s = \alpha(n+1)$ realan broj izmedju 1 i $n+1$. Kao i prije da bismo našli $x_{(s)}$, moramo rastaviti s na cijeli i razlomljeni dio. Uočite

$$Q_1 = q_{0.25}\,,\ Q_3 = q_{0.75}.$$

Specijalno, ako je $\alpha = k/100$ za neki $k = 1, \ldots, 99$, kažemo da je q_α k%-tni percentil uzorka. Npr. medijan je 50%-tni, a donji kvartil je 25%-tni percentil uzorka.

Varijanca uzorka $x_1, x_2, \ldots, x_n$ definirana je izrazom

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

Ova često korištena mjera disperzije uvijek je nenegativna, baš kao i prethodne dvije. Poprima vrijednost 0, ako i samo ako je i raspon 0, tj. ako su sve vrijednosti x_i jednake.

Primjetite da varijanca mjeri prosječno kvadratno odstupanje podataka od aritmetičke sredine, te je dakle izražena u drugačijim mjernim jedinicama od samih podataka (u kvadratu originalnih jedinica). To možemo ispraviti koristeći njen drugi korijen tj. standardnu devijaciju. Standardna devijacija uzorka definirana je na sljedeći način

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

Kako je varijanca, kao i standardna devijacija osjetljiva na mjernu jedinicu u kojoj izražavamo naže podatke, ponekad se raspršenost za nenegativne podatke mjeri i koeficijentom varijacije uzorka

$$v = \frac{s}{\bar{x}}.$$

Pitanje je kako naći aritm. sredinu, varijancu odn. stand. devijaciju za grupirane podatke? Tj. podatke za koje znamo samo u koji interval pripadaju.

Primjer (vježbe)

raspon	f_a	r_a
1.0-1.2	1	1/14
1.3-1.4	3	3/14
1.5-1.6	6	6/14
1.7-1.8	1	1/14
1.9-2.0	3	3/14

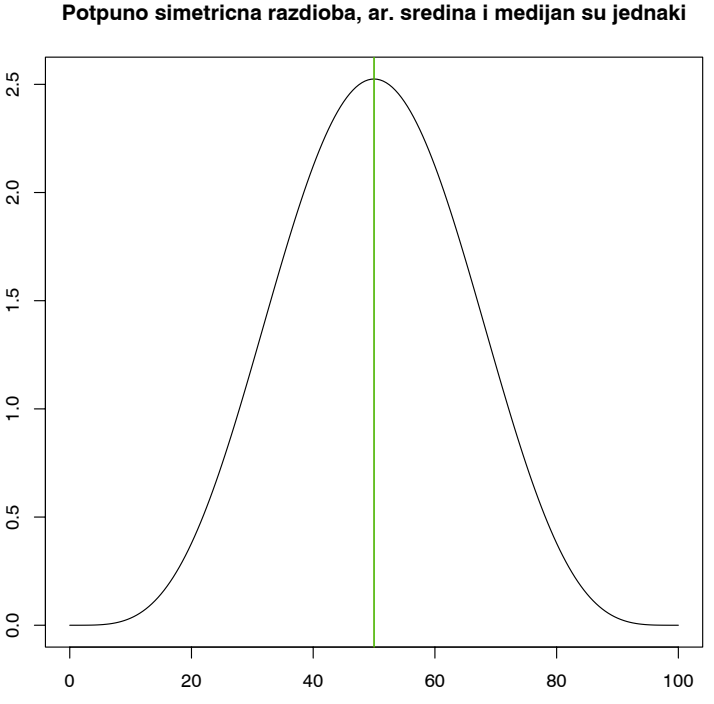
Oblik razdiobe

Kao što ćemo kasnije vidjeti razdiobu za neprekidna numerička obilježja X možemo teoretski modelirati grafom krivulje koja oblikom odgovara histogramu.

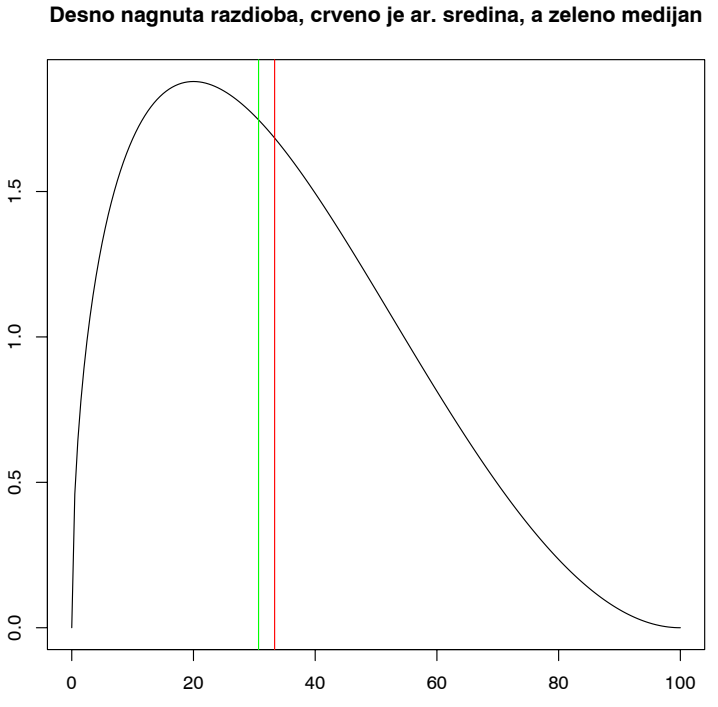
Ovisno o obliku grafa ove krivulje (ili tzv. funkcije gustoće) razdioba može biti

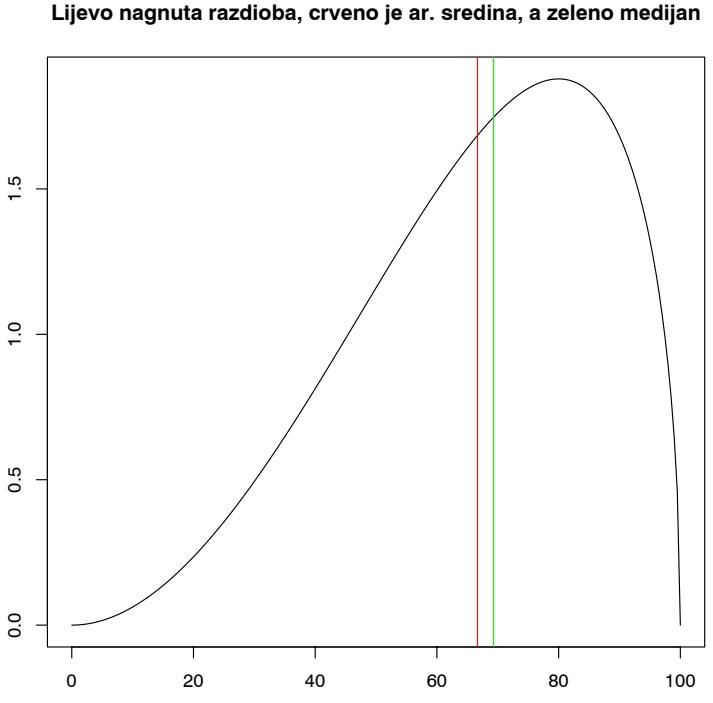
- simetrična oko svog očekivanja.
- desno nagnuta
- lijevo nagnuta

Kažemo da je razdioba desno odn. lijevo nagnuta, skewed, ako je njen desni odn. lijevi rep teži od onog drugog.



Simetrična razdioba.


Desno nagnuta razdioba, crveno je ar. sredina, a zeleno medijan
0.0
0.5
1.0
1.5
0
20
40
60
80
100



Lijevo nagnuta razdioba.

Pitanje je možemo li lijevo odn. desno nagnute razdiobe definirati rigoroznije tj. ne oslanjajući se samo na promatranje histograma čiji oblik ipak ovisi i o odabranim razredima.

Definiramo za prirodni broj k prvo k-ti centralni moment uzorka kao

$$\mu_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^k.$$

Očito je $\mu_1 = 0$ i $\mu_2 = s^2$ (provjerite).

Broj

$$\alpha_3 = \frac{\mu_3}{s^3}$$

zovemo koeficijent asimetrije uzorka (skewness). Ako je $\alpha_3 > 0$ radioba je lijevo nagnuta (ili pozitivno asimetrična), a za $\alpha_3 < 0$ ona je desno nagnuta (ili negativno asimetrična). Primjetite za uzorak simetričan oko $\bar{x}$ $\alpha_3 = 0$.

Već smo rekli da se statistikama nazivaju i sve numeričke vrijednosti izvedene iz uzorka. Tako bismo mogli reći da su

$$\bar{x},\ s,\ s^2,\ m,\ Q_1,\ Q_3,\ d_Q,\ q_\alpha,\ \mu_k,\ \alpha_3, \text{itd.}$$

primjeri različitih statistika.

Ne/Statističnost

Za svaku od tvrdnji na skali od 0 do 10, odredite koliko se slažete s njom. Pa zbrojite da dobijete svoj score.

- Ispit za kolegij Statistika mi neće predstavljati problem po svemu sudeći.
- Moja očekivanja o statistici su dosadašnjim tijekom kolegija uglavnom ispunjena.
- Prije početka kolegija statistika mi se činila kao možda koristan i zanimljiv predmet.
- Matematički predmeti su mi išli tokom školovanja.
- Kad vidim statističke podatke u novinskom tekstu ili udžbeniku, uglavnom ih pročitam vrlo pažljivo.