Aritmetička sredina i standardna devijacija

Tvrtko Tadić[1]

Kao što smo vidjeli u prošlom članku ([4]), podatci danas dolaze u ogromnim količinama i znaju biti poprilično nepregledni. Cilj grafičkog prikazivanja podataka je pokazati važne informacije o podatcima. Grafički način prikazivanja podataka može dovesti do određenih problema. Primjerice, **nije uvijek jednostavno usporediti dva grafa ili dijagrama**. Kod brojeva je to mnogo lakše. Zato postoji čitav niz numeričkih vrijednosti koje izračunavamo iz podataka koje nam daju informacije o vrijednostima koje podatci imaju. Jedna od najkorištenijih vrijednosti je **aritmetička sredina**, poznata i kao **prosjek**.

U ovom ćemo članku pokušati objasniti zašto koristimo aritmetičku sredinu, koja su njena dobra svojstva, a koje mane. Puno riječi bit će posvećeno i **standardnoj devijaciji**, tj. *mjeri odstupanja od aritmetičke sredine*. Također, pokušat ćemo na razini srednjoškolske matematike dati uvid u teorijsku pozadinu zašto je aritmetička sredina interesantna.

Definicije

Kada imamo puno podataka, prirodno je razmišljati o nekim **važnim vrijednostima** koje opisuju te podatke. Tako nam, recimo, najveća i najmanja vrijednost govore o *rasponu podataka* koje imamo, no vrlo često ekstremi nam nisu bitni. Važnija nam je *vrijednost kojoj je većina podataka bliska*. Iz brojnih razloga danas se najčešće koristi aritmetička sredina.

Definicija 1. *Za brojevni niz podataka* $x_1,\ldots,x_n$ *njihovom* ***aritmetičkom sredinom*** *zovemo broj*

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Definicija 2. *Za brojevni niz podataka* $x_1,\ldots,x_n$ ***standardnom devijacijom*** *zovemo broj*

$$s = \sqrt{\frac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + \cdots + (\overline{x} - x_n)^2}{n}}.$$

[1]Tvrtko Tadić, Microsoft Corporation, Redmond / University of Washington, Seattle

Aritmetička sredina daje nam broj koji se često u teorijskom i praktičnom pogledu smatra najbližim podatcima. S druge strane, standardna devijacija kaže nam kolika je ta bliskost. Što je s manji, to je x bliži podatcima. Ako je $s = 0$, sve vrijednosti su iste, a x je jednak svim vrijednostima.

Ilustracija pojmova

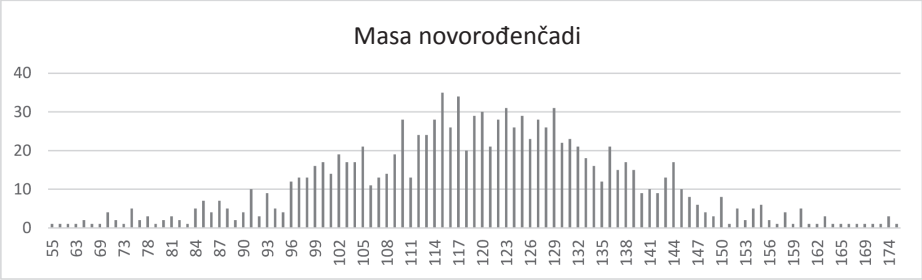
Na primjeru podataka o masi novorođenčadi ilustrirat ćemo kako koristimo aritmetičku sredinu i standardnu devijaciju.

Prikupljeni su podatci o masi novorođenčadi u bolnici *Kaiser Foundation Hospital* u Oaklandu u Kaliforniji (za izvor vidi [2]). Podatci su zapisani u podatkovnu tablicu, uključuju podatke o masi djeteta, podatke o duljini trudnoće, je li dijete prvorođenac ili nije, visinu i masu majke, te je li majka pušač.

Tablica ukupno ima 1174 redaka pa nije moguće sve podatke prikazati u ovome članku. Čitatelji mogu preuzeti ove podatke kako je opisano na kraju članka. Kako bismo dali uvid u to kako podatci izgledaju, prikazat ćemo uzorak od 20 redova u tablici.

Redni broj	Masa bebe (unce)	Trajanje trudnoće	Prvorođenac	Starost majke	Visina majke (inči)	Masa (trudne) majke (funte)	Majka pušač
21	115	274	0	27	67	175	1
27	114	266	0	20	65	175	1
42	87	248	0	37	65	130	1
70	133	284	0	25	66	125	1
81	114	274	0	33	67	148	1
236	125	286	0	21	64	139	0
238	130	285	0	23	63	128	1
268	117	283	0	27	63	108	0
378	102	258	1	22	65	135	0
386	139	279	0	20	64	143	0
390	130	282	0	26	67	147	1
522	103	291	1	26	63	102	0
592	129	280	1	24	65	126	0
695	105	269	0	27	62	100	1
745	96	282	1	30	68	127	1
761	117	255	0	26	61	120	0
975	129	294	1	21	65	132	0
1017	145	316	0	22	67	142	0
1112	71	254	0	19	61	145	1
1125	126	298	0	24	61	112	0

Prvi izazov je da damo grubi opis podataka. Najlakša beba imala je 55 unci, a najteža 176. Najviše beba rodilo se s 115 unci (njih 35). Raspored i zastupljenost vrijednosti dana je na stupčastom dijagramu na Slici 1. (dobiven *Excelom*).



Slika 1.

Izračunajmo aritmetičku sredinu i standardnu devijaciju podataka o masi. Ovo ćemo napraviti *Excelom*, odnosno nekim računalnim programom kojim obrađujemo podatke. Dobivamo

aritmetička sredina	standardna devijacija
119.4625213	18.32867144

Promotrimo prvo aritmetičku sredinu. Ako pogledamo grafikon, **velik broj novorođenčadi ima masu blisku broju 119**. Već smo prije uočili da najveći broj djece prilikom rođenja ima masu 115 unci, što kao ni druge česte vrijednosti nije daleko od dobivene vrijednosti aritmetičke sredine. U ovom slučaju **aritmetička je sredina vrijednost na stupčastom dijagramu oko koje su grupirane *najuobičajenije* mase djece**.

Što predstavlja standardna devijacija? Ova vrijednost mjeri u kojem rasponu možemo očekivati da će se kretati vrijednosti mase. Kako bismo ovo ilustrirali, napravimo sljedeću vježbu u *Excelu*.

Utvrdimo koliko masa novorođenčadi i koji je postotak njih na brojevnom pravcu udaljen od aritmetičke sredine do:

a) jedne standardne devijacije, tj. u intervalu $[\overline{x}-s, \overline{x}+s]$;

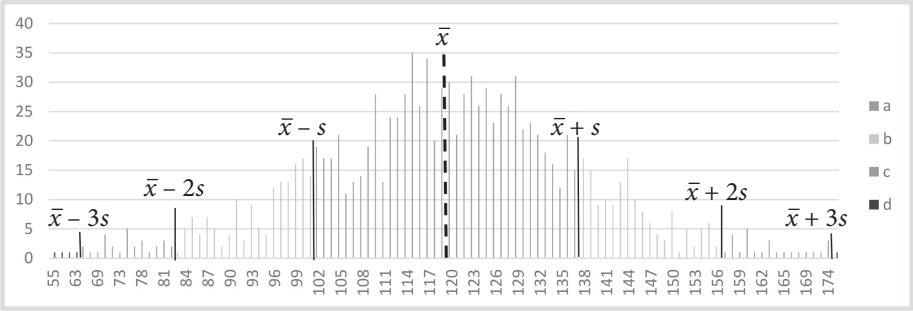
b) dvije standardne devijacije, tj. u intervalu $[\overline{x}-2s, \overline{x}+2s]$;

c) tri standardne devijacije, tj. u intervalu $[\overline{x}-3s, \overline{x}+3s]$.

Rezultati su sljedeći:

Interval	Broj	Postotak
$[\overline{x}-s, \overline{x}+s]$	813	**69.25**
$[\overline{x}-2s, \overline{x}+2s]$	1114	**94.89**
$[\overline{x}-3s, \overline{x}+3s]$	1169	**99.57**

Ovi rezultati postat će jasniji na sljedećem stupčastom dijagramu (Slika 2).



Slika 2.

Pravilo 68-95-99

Ovi podatci izgledaju ***normalno* distribuirani**. Bez ulaženja u preciznu definiciju, navest ćemo neke osobine takvih podataka:

- Podatci su koncentrirani oko aritmetičke sredine.
- Što gledamo vrijednosti udaljenije od aritmetičke sredine, to ćemo manje podataka s tim vrijednostima naći.
- Stupčasti dijagram podataka ima oblik zvonolike krivulje.

Kod tako distribuiranih podataka vrijedi pravilo 68-95-99, poznato i kao pravilo 3-sigma:

- Unutar jedne standardne devijacije, tj. intervala $[\overline{x}-s, \overline{x}+s]$, nalazi se približno 68 % podataka.
- Unutar dvije standardne devijacije, tj. intervala $[\overline{x}-2s, \overline{x}+2s]$, nalazi se približno 95 % podataka.
- Unutar tri standardne devijacije, tj. intervala $[\overline{x}-3s, \overline{x}+3s]$, nalazi se približno 99 % podataka.

Uočimo kako je gornja procjena u slučaju podataka o masi djece vrlo precizna.

Kritike aritmetičke sredine i podatci koji nisu normalno distribuirani

Iako je aritmetička sredina jedna od najkorištenijih vrijednosti koje izračunavamo na temelju podataka, ima brojne mane i mnogi je kritiziraju. Poznata je sljedeća izreka:

Ja jedem meso, ti jedeš zelje, a statistika kaže da jedemo sarmu.

Ova izjava zna biti točna u mnogim slučajevima kad se radi o aritmetičkoj sredini.

Bill Gates u posjetu izbjeglicama

Pogledajmo sljedeći ekstremni primjer:

Zamislimo Billa Gatesa[2] u posjetu prihvatilištu za izbjeglice u kojem živi 1000 ljudi. Prosječno svaka osoba u tom prihvatilištu ima 80 milijuna dolara. Iz ovog podatka može se steći ***krivi dojam da se u prihvatilištu nalazi velik broj iznimno bogatih ljudi***. Točan podatak o financijskom stanju osoba u zgradi nije ni blizu iznosu od 80 milijuna dolara. Izbjeglice imaju iznos od 0 dolara i većina ih nikada neće doći do prosjeka koji je u pitanju. S druge strane, Bill Gates ima 80 milijardi više nego što je prosjek.

Aritmetička sredina je u ovom slučaju vrijednost koja ne daje skoro nikakvu informaciju o podatcima. U navedenom primjeru standardna je devijacija 6 milijardi dolara. Ovo nije osobito korisna informacija. Uočimo da ovi podatci nisu normalno distribuirani, te da pravilo 68-95-99 ne vrijedi. Što ipak možemo reći o podatcima u takvom slučaju?

Čebiševljeva nejednakost

U općenitom slučaju vrijedi tzv. Čebiševljeva nejednakost koju ćemo precizno izreći kasnije, a koja nam jamči

- da se u intervalu $[\overline{x}-2s, \overline{x}+2s]$ nalazi bar 75 % podataka;
- da se u intervalu $[\overline{x}-3s, \overline{x}+3s]$ nalazi bar 88.8 % podataka.

Uočimo kako je ovo zadovoljeno u oba primjera i kod mase novorođenčadi, i kod Bill Gatesa i izbjeglica.

Možemo zaključiti da ako znamo aritmetičku sredinu i standardnu devijaciju ipak imamo neku informaciju kako se podatci ponašaju. Poznavanje aritmetičke sredine neće nam dati mnogo informacija, osim da među podatcima ***postoje*** vrijednosti koje su veće ili jednake i manje ili jednake od te vrijednosti.

Zadatak

Pozivamo čitatelje da provjere izloženo (koristeći *Excel*) na drugim podatcima u danoj tablici poput trajanja trudnoće, visine majke, masa majke...

[2]Prema podatcima iz 2016. godine Bill Gates najbogatiji je čovjek na svijetu, s procijenjenim bogatstvom od 81.7 milijardi dolara.

Teorijska podloga

Ima više razloga zašto baš koristimo aritmetičku sredinu, a jedan od razloga je i to što za kvadratnu funkciju možemo lako naći minimum. Podsjetimo se sljedećeg teorema koji se radi u srednjoj školi:

Teorem 3. Neka je $f(x) = ax^2 + bx + c$ gdje je $a > 0$. Tada funkcija f ima minimum za vrijednost $x = -\frac{b}{2a}$.

Nama je cilj odabrati broj koji je na neki način najbliži podatcima. Udaljenost tog broja od svakog podataka kvadriramo i gledamo prosječnu vrijednost kvadrata tih udaljenosti, tj. tražimo minimum kvadratne funkcije

$$f(x) = \frac{(x-x_1)^2 + \cdots + (x-x_n)^2}{n},$$

Nakon što raspišemo ovu funkciju, dobivamo:

$$f(x) = x^2 - 2\overline{x}x + \frac{x_1^{\,2} + x_2^{\,2} + \cdots + x_n^{\,2}}{n}.$$

Koristeći teorem 3. znamo da se minimum ove funkcije postiže za $x = \overline{x}$ i da je minimum jednak upravo s^2. Korijen uzimamo kako bismo odstupanje mjerili u istim jedinicama kao i podatke.

Teorem 4. (Čebiševljeva nejednakost) Neka je $x_1, x_2, \ldots, x_n$ niz vrijednosti. Za svaku vrijednost $\alpha > 0$ u intervalu $[\overline{x} - \alpha s, \overline{x} + \alpha s]$ nalazi se bar $100\left(1 - \frac{1}{\alpha^2}\right)$ % vrijednosti.

Dokaz. Neka je m broj vrijednosti u intervalu $[\overline{x} - \alpha s, \overline{x} + \alpha s]$ i neka su $j_1, \ldots, j_{n-m}$ indeksi vrijednosti koje nisu u tom intervalu. Sada je

$$s^2 = \frac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + \cdots + (\overline{x} - x_n)^2}{n}$$

$$\geq \frac{(\overline{x} - x_{j_1})^2 + (\overline{x} - x_{j_2})^2 + \cdots + (\overline{x} - x_{j_{n-m}})^2}{n}$$

S obzirom da vrijedi $|x_{jk} - \overline{x}| > \alpha\sigma$ za $k = 1, \ldots, n - m$, slijedi

$$\geq \alpha^2 s^2 \frac{n-m}{n}$$

Dobivamo $\frac{m}{n} \geq 1 - \frac{1}{\alpha^2}$.

Napomena. Uočimo da je za $\alpha \in (0,1)$ tvrdnja teorema točna (jer se u intervalu nalazi nenegativan broj vrijednosti), ali nam ne daje nikakvu ocjenu koliko se brojeva zaista nalazi u intervalu.

Normalnost podataka

Često vrijednosti podataka želimo opisati funkcijom koja modelira njihove vrijednosti. Iz raznih teorijskih razloga i čestog pojavljivanja u praksi **model Gaussove krivulje** ili **normalni model** jedan je od najvažnijih i najkorištenijih modela.

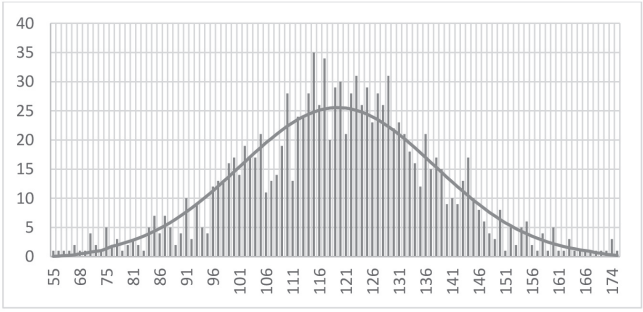
Ideja je da stupčasti dijagram podataka *pokušamo opisati* pomoću funkcije

$$g(x) = \text{broj podataka} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

gdje su $\mu \in \mathbb{R}, \sigma > 0$ **parametri modela** odabrani tako da model *što je bolje moguće opisuje stupčasti dijagram*.

Pokazuje se da funkcija g *najbolje* opisuje normalizirani stupčasti dijagram kad za parametre modela stavimo $\mu = \overline{x}, \sigma = s$.

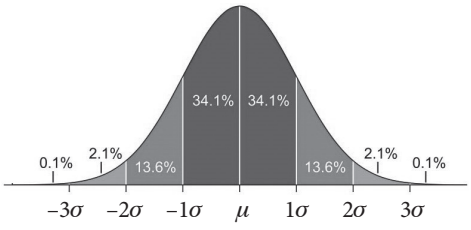
Pokažimo to na primjeru mase novorođenčadi. Nacrtajmo frekvencijski stupčasti dijagram i graf funkcije $g(x)$.



Slika 3. Gaussova krivulja i podaci

Pravilo 68-95-99 vrijedi zbog sljedeće činjenice (za više vidi [3]):

Interval	**Površina ispod krivulje $y = g(x)$, a iznad intervala**
$[\mu - \sigma, \mu + \sigma]$	68.27 % * broj podataka
$[\mu - 2\sigma, \mu + 2\sigma]$	95.45 % * broj podataka
$[\mu - 3\sigma, \mu + 3\sigma]$	99.73 % * broj podataka



Slika 4. Postotak površine ispod Gaussove krivulje

Izazovi procjene aritmetičke sredine i standardne devijacije

Temeljem izloženog znamo da smo, ako imamo aritmetičku sredinu i standardnu devijaciju, dobili ***sažetu informaciju o podatcima*** koje imamo. Postoji jedan problem: aritmetičku sredinu i standardnu devijaciju **nećemo uvijek moći izračunati** jer nam neće biti dostupni svi podatci potrebni za to.

Primjerice:

- Koliko prosječno sati stanovnici Hrvatske provedu spavajući?
- Koliki je postotak prebolio neku bolest?
- Koliko šalica kave dnevno popiju stanovnici Zagreba?
- Kolika je prosječna visina srednjoškolaca u Splitu?

Kod svih ovih primjera uočimo da postoje sljedeći problemi:

- Praktički, **nemoguće je prikupiti sve potrebne podatke** kako bismo točno izračunali prosjek i standardnu devijaciju.
- Za neke **podatke trebamo poseban (skup) postupak da bismo došli do traženog podatka**. Primjerice, kako bismo dobili točnu informaciju o vremenu koje neka osoba provede spavajući, potreban nam je poseban uređaj koji će to izmjeriti.

Međutim, možemo podatke prikupiti na uzorku i procijeniti kolike bi mogle biti vrijednosti prosjeka i standardne devijacije. Kao što smo simulacijama pokazali, uzorak poprima brojne osobine podataka iz kojih je uzet. Sad ćemo pokazati da se može koristiti i za *relativno* preciznu procjenu aritmetičke sredine i standardne devijacije.

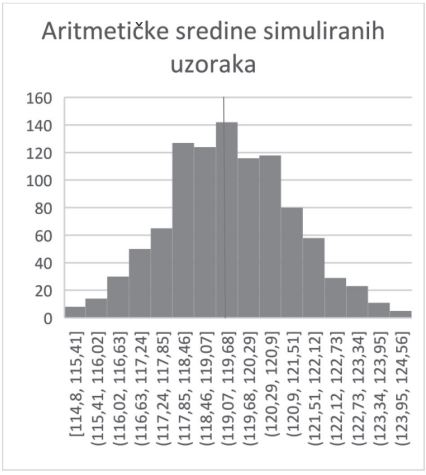
Procjena aritmetičke sredine i standardne devijacije uzorkom

Kod procjene aritmetičke sredine i standardne devijacije svih podataka na manjem uzorku često radimo slučajan odabir uzorka. Tako ćemo, kako bismo simulacijama opravdali ovaj postupak, napraviti sljedeći eksperiment na računalu:

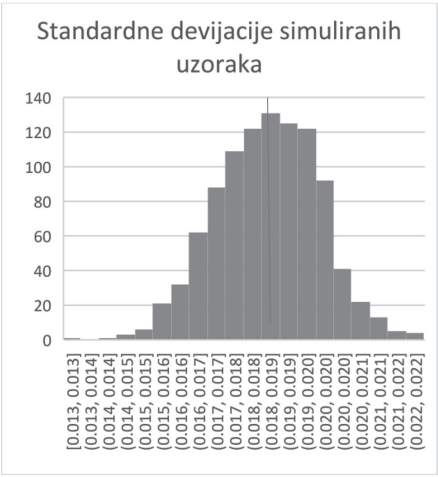
- Slučajno izabrati 100 od 1174 novorođenčadi. Na temelju mase izbrane novorođenčadi izračunat ćemo i zabilježiti aritmetičku sredinu i standardnu devijaciju.
- Prethodni postupak ponovit ćemo 1000 puta te na kraju nacrtati dva histograma; prvi koji pokazuje zabilježene vrijednosti aritmetičke sredine i drugi koji pokazuje zabilježene vrijednosti standardne devijacije.

Uočimo da se zabilježeni podatci (iz uzoraka) grupiraju oko stvarnih vrijednosti aritmetičke sredine i standardne devijacije (svih podataka).

Čak i najveća zabilježena odstupanja još nam uvijek daju dobru predodžbu o svim podatcima. Ovaj fenomen može se precizno matematički objasniti. Naime,



Slika 5.



Slika 6.

histogrami izračunatih vrijednosti na temelju uzoraka ponovo prate krivulje zvonolikog oblika. Pod uvjetom da podatci nisu previše *divlji*, **rijetko će se dogoditi da odstupanje vrijednosti izračunate na uzorku i izračunate na svim podatcima bude preveliko**.

Zaključak

U ovom članku imali smo priliku upoznati se s razlozima zašto koristimo aritmetičku sredinu i standardnu devijaciju kao neke od mjera za izražavanje informacija o numeričkim podatcima. Vidjeli smo i u koje zablude može dovesti kriva upotreba ovih mjera. Također, upoznali smo se izazovima koje predstavlja izračunavanje ovih vrijednosti. Iako postoje određene teorijske činjenice koje se mogu objasniti na nivou srednjoškolske matematike, glavne ideje mogu se prikazati obradom na računalu koristeći stvarne (dovoljno velike) skupove podataka

Dodatak A. Podatci i kod

Podatci i kod korišteni u izradi ovoga članka modu se preuzeti s internetske stranice: https://web.math.pmf.unizg.hr/~tvrtko/metodikaStatistike/clanak2

Dodatak B. Obrada podataka u *Excelu*

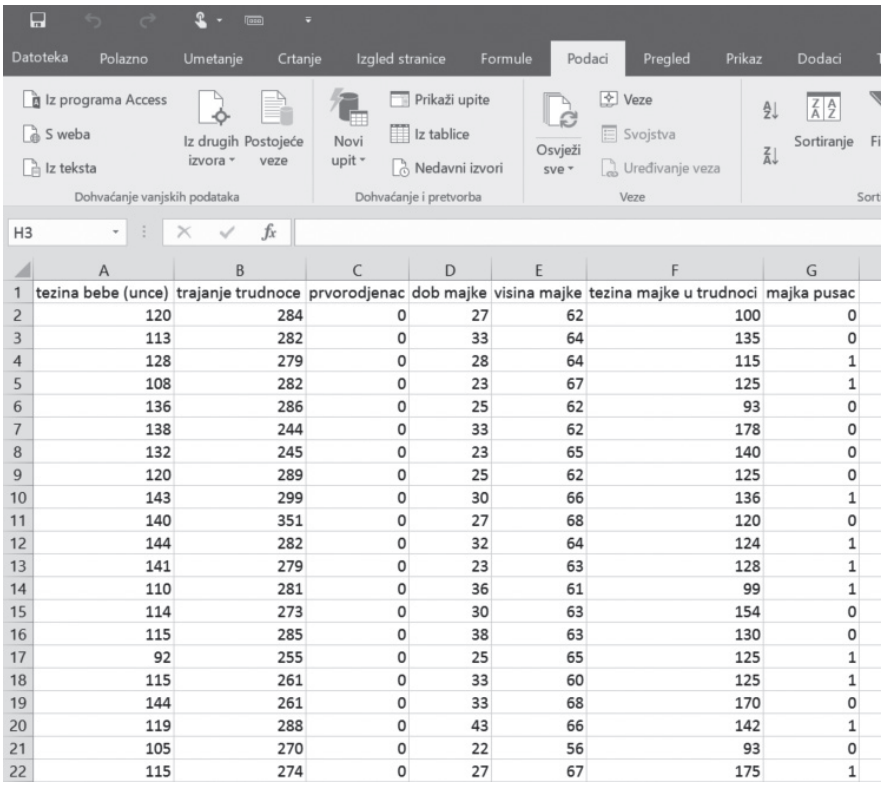
Obradu podataka koje smo predstavili u ovome članku napravili smo u *Excelu* koji je standardni dio ponude *Microsoft Officea*. *Excel*, osim na osobnim računalima, ima verzije za razne uređaje poput mobitela (pametnih telefona) i tableta, kao i online verziju. Microsoft je pomoću tehnologije računalnog oblaka omogućio da se podatci među uređajima **sinkroniziraju i dijele** (pod pretpostavkom da uređaj ima vezu za internet).

Tako primjerice nastavnik može podijeliti podatke za obradu sa svojim učenicima. Učenik može pripremiti podatke na svom osobnom računalu i kasnije ih dorađivati na mobitelu ili tabletu.

Ovo je sve moguće putem *Office 365* pretplate (koja pretplatnicima omogućuje da skinu najnoviju verziju svih programa u sklopu *Microsoft Officea*). **Polaznici i djelatnici obrazovnog sustava u Hrvatskoj mogu dobiti *Office 365* pretplatu besplatno**. Za više detalja vidi stranicu https://office365.skole.hr/.

Unos podataka

Podatke možemo unijeti tako da otvorimo novu radnu knjigu, kliknemo na izbornik *Podatci* i odaberemo *Dohvaćanje vanjskih podataka – Iz teksta* te odaberemo datoteku s podatcima u ovom slučaju `novorodjencadPodatci.txt`. U *Čarobnjaku* za uvoz teksta moramo odabrati da naša datoteka ima zaglavlje (nazive stupaca) te da su graničnici podataka zarezi.

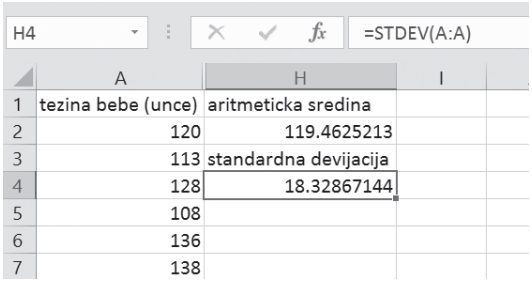


	A	B	C	D	E	F	G
1	tezina bebe (unce)	trajanje trudnoce	prvorodjenac	dob majke	visina majke	tezina majke u trudnoci	majka pusac
2	120	284	0	27	62	100	0
3	113	282	0	33	64	135	0
4	128	279	0	28	64	115	1
5	108	282	0	23	67	125	1
6	136	286	0	25	62	93	0
7	138	244	0	33	62	178	0
8	132	245	0	23	65	140	0
9	120	289	0	25	62	125	0
10	143	299	0	30	66	136	1
11	140	351	0	27	68	120	0
12	144	282	0	32	64	124	1
13	141	279	0	23	63	128	1
14	110	281	0	36	61	99	1
15	114	273	0	30	63	154	0
16	115	285	0	38	63	130	0
17	92	255	0	25	65	125	1
18	115	261	0	33	60	125	1
19	144	261	0	33	68	170	0
20	119	288	0	43	66	142	1
21	105	270	0	22	56	93	0
22	115	274	0	27	67	175	1

Slika 7. Podaci uvezeni u Excel

Nakon toga imamo sve podatke u radnom listu i po želji ih možemo dodatno urediti. Mi smo ih uvezli tako da se masa bebe nalazi u stupcu A. Ostale ćemo stupce sakriti jer nam neće trebati.

Računanje aritmetičke sredine i standardne devijacije



Slika 8. Račun u Excel-u

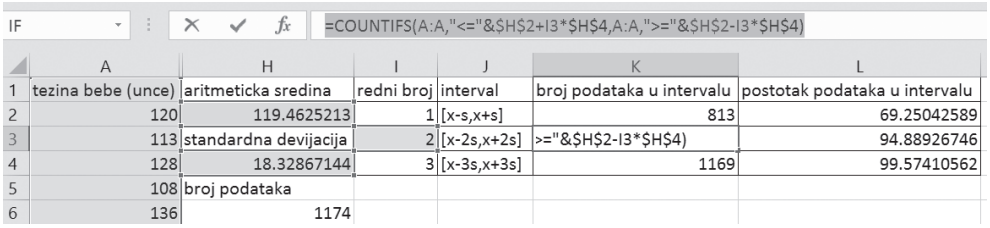
Za izračunavanje aritmetičke sredine i standardne devijacije koristimo funkcije AVERAGE (*polja s podatcima*) i STDEV (*polja s podatcima*). Kako su nam svi podatci u stupcu A, to radimo tako da u polju gdje želimo da nam se pokaže izračun upišemo AVREAGE(A:A) i STDEV(A:A) (polja s nenumeričkim podatcima bit će zanemarena). Vidi Sliku 8.

Provjera pravila 68-95-99

Sada ćemo iskoristiti *Excel* kako bismo provjerili pravilo 68-95-99 na podatcima o masi novorođenčadi.

Prvo ćemo izračunati koliko podataka u stupcu A imamo tako da upišemo COUNT(A:A).

U sljedećem koraku prebrojat ćemo sva polja u stupcu A čije su vrijednosti unutar intervala $[\overline{x}-is, \overline{x}+is]$ za i = 1, 2, 3.



Slika 9. Provjera pravila 68-95-99 u Excelu

To radimo tako da u polje upišemo

COUNTIFS(A:A,"<="&H2+I3*H4,A:A,">="&H2-I3*H4).

H2 sadrži podatak o aritmetičkoj sredini, H4 o standardnoj devijaciji, dok stupac I sadrži vrijednost od i (vidi Sliku 9.). Naredba provjerava sva polja u stupcu A čija vrijednost ima svojstva da je

- manja ili jednaka od $\overline{x}+is$;
- veća ili jednaka od $\overline{x}-is$.

Konačno, kako bismo izračunali postotak, podijelimo broj podataka u intervalu s ukupnim brojem podataka u stupcu A i pomnožimo sa 100. Primjerice, polje L4 na Slici 9. izračunali smo kao K4/H6*100.

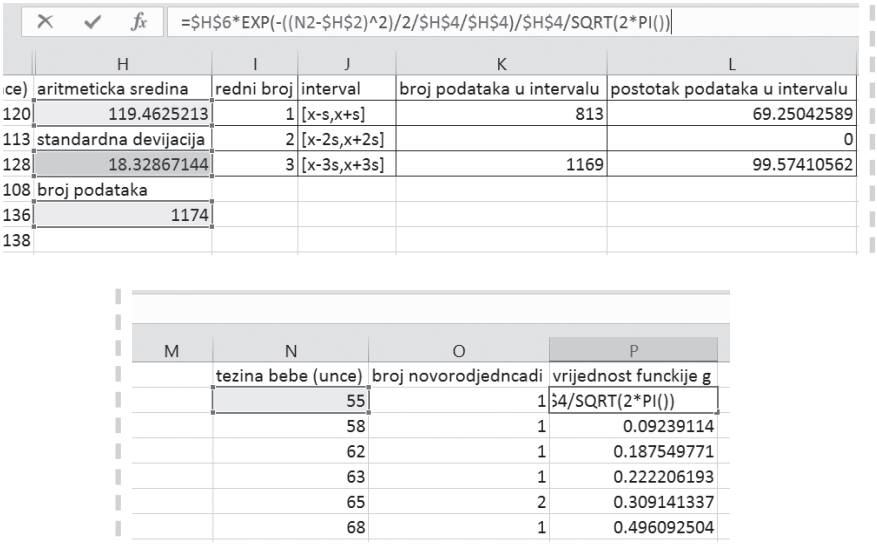
Crtanje stupčastog dijagrama i Gaussove krivulje

Nacrtajmo stupčasti dijagram s frekvencijama i pripadajućom Gaussovom krivuljom. To ćemo napraviti u nekoliko koraka:

- Kopirajmo podatke iz stupca A u neki slobodni stupac, primjerice, N. Uklonimo duplikate iz stupca N odabirom svih podataka u tom stupcu i koristeći izbornik *Podatci->Ukloni duplikate*. Nakon toga poredajmo podatke uzlazno (*Podatci->Sortiranje*).
- U stupcu O izračunat ćemo koliko novorođenčadi iz stupca A ima mase navedene u stupcu N. Primjerice, da bismo utvrdili koliko polja u stupcu A ima vrijednost koja se nalazi u polju N2, koristimo naredbu COUNTIF(A:A,N2). *Excel* zna sam automatski kreirati naredbe za ostala polja iz stupca N.
- U stupcu P izračunat ćemo vrijednost funkcije

$$g(x) = \text{broj podataka} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

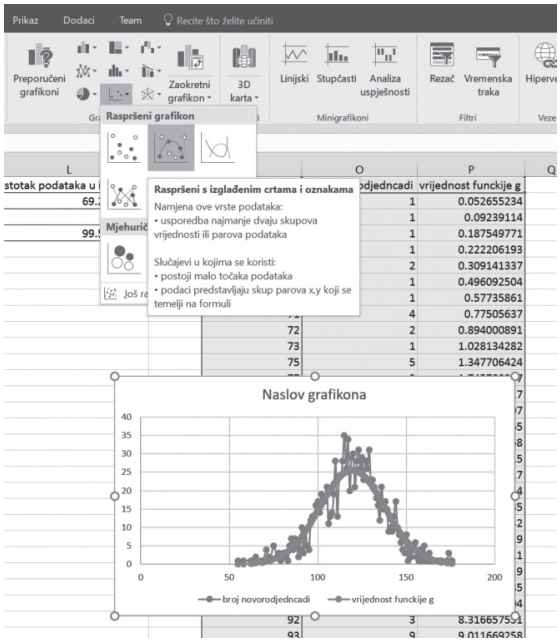
gdje je x vrijednost iz stupca N, parametar μ jednak vrijednosti izračunate aritmetičke sredine i parametar σ jednak vrijednosti izračunate standardne devijacije. Ovo radimo tako da upišemo u polje P2 naredbu $H$6*EXP(-((N2-$H$2)^2)/2/$H$4/$H$4)/$H$4/SQRT(2*PI()) i kopiramo u ostala polja.



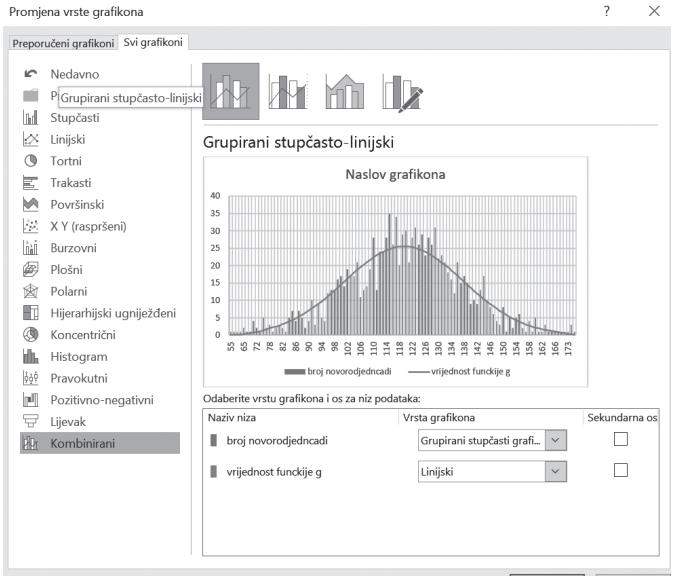
Slika 10. Priprema za crtanje Gaussove krivulje i stupčastog dijagrama

- Sada imamo sve spremno da nacrtamo stupčasti dijagram frekvencija i Gaussovu krivulju.

Odaberimo stupce N, O i P te odaberimo *Umetanje->Raspršeni grafikon* i odaberimo raspršeni grafikon s uglađenim crtama i oznakama (vidi Sliku 11.).



Slika 11. Stvaranje grafikona



Slika 12. Odabir tipa grafikona

Preostaje nam da podatke o broju novorođenčadi predstavimo u obliku stupčastog dijagrama. Kliknemo (desnim gubom miša) na liniju koja predstavlja te podatke te odaberemo *Promjena vrste grafikona za niz...* (krivulja na Slici 11.).

U prozoru koji nam se otvorio odaberemo Grupirani stupčasto-linijski dijagram (vidi Sliku 12.) i dobili smo traženo. Sada graf možemo dodatno urediti.

Čitatelji mogu skinuti *Excel* radnu knjigu s lokacije navedene u dodatku A.

Dodatak C. Računanje aritmetičke sredine i standardne devijacije na reprezentativnom uzorku u *Pythonu*

Kako bismo simulirali uzorke, koristimo programski jezik *Python* koji se uči u srednjim školama (vidi primjerice [1]). Kod izvršava sljedeći postupak:

učitaj podatke

ispisi aritmetičku sredinu i standardnu devijaciju

za $k = 1$ **do** *broj Ekperimenata*:

- uzmi reprezentativni uzorak;
- izračunaj aritmetičku sredinu i standardnu devijaciju na uzorku;
- zabilježi podatke u odgovarajuće liste (datoteke);

ispiši raspon aritmetičkih sredina dobivenih na uzorcima;

ispiši raspon standardnih devijacija dobivenih na uzorcima;

```
#unos paketa za generianje slucajnih brojeva
import random;
#unos paketa za racunaje statistickih funckija
import statistics;
duljinaUzorka = 100;
brojEksperimenata = 1000;

#ucitavanje podataka
with open('masaNovorodjencadi.txt') as podaciIzvor:
  podaci = [int(x) for x in podaciIzvor.readlines()]

print("Aritmeticka sredina je");
print(statistics.mean(podaci));

print("Standardna devijacija je");
print(statistics.stdev(podaci));

duljinaPodataka = len(podaci);

#incijalizacija liste
aritmetickeSredineUzoraka = [0] * brojEksperimenata;
```

```
standardneDevijacijeUzoraka = [0] * brojEksperimenata;

#datoteke u koje cemo spremiti podatke
aritmetickeSredineUzorakaDatoteka = open("aritmeticke-
SredineUzoraka.txt","w");
standardneDevijacijeUzorakaDatoteka = open("standardne-
DevijacijeUzoraka.txt","w");

for k in range(0,brojEksperimenata):
#reprezenativni uzorak
  uzorak = random.sample(podaci, duljinaUzorka);
  aritmetickeSredineUzoraka[k] = statistics.mean(uzorak);
  aritmetickeSredineUzorakaDatoteka.write(str(aritme-
  tickeSredineUzoraka[k]) + "\n");
  standardneDevijacijeUzoraka[k] = statistics.stdev(uzo-
  rak);
  standardneDevijacijeUzorakaDatoteka.write(str(stan-
  dardneDevijacijeUzoraka[k]) + "\n");

print("Aritmeticka sredina uzoraka se krece u intervalu");
print([ min(aritmetickeSredineUzoraka) , max(aritmetic-
keSredineUzoraka) ]);

print("Standardna devijacija uzoraka se krece u intervalu");
print([ min(standardneDevijacijeUzoraka), max(standar-
dneDevijacijeUzoraka) ]);
```

Bibliografija

1. L. Budin, P. Brođanac, Z. Markučič, S. Perić. *Rješavanje problema programiranjem u Pythonu: za 2. i 3. razred gimnazije.* Zagreb: Element, 2012.
2. Nolan D., Speed T. *Stat Labs: Mathematical Statistics through Applications.* New York: Springer-Verlag, 2000.
3. Sarapa, N. *Vjerojatnost i statistika 2. dio: Osnove statistike – slučajne varijable.* Zagreb: Školska knjiga, 1996.
4. Tadić, Tvrtko: „Podatci i uzorak." *Poučak 67*, 2016: 16-26.
5. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference.* New York: Springer, 2005.