Medijan i kvantili

Tvrtko Tadić[1]

U prošlom članku (Tadić 2017) upoznali smo se s jednom mjerom za sredinu podataka – artimetičkom sredinom. Imali smo se priliku upoznati s njezinim svojstvima, te pokazati neke probleme koji mogu nastati njezinim korištenjem u interpretaciji podataka. U ovom članku upoznajemo se s jednom novom mjerom za srednju vrijednost podataka – **medijanom**. Vidjet ćemo da se medijan nalazi u širem skupu podatkovnih funkcija koje zovemo **kvantili**. A čemu služe, imat ćemo priliku vidjeti na primjeru podataka o kašnjenju zrakoplova.

Definicije i karakterizacije medijana

Idejno, medijan je broj koji brojevne podatke dijeli na **dva jednakobrojna skupa** – one koji su (strogo) manji i one koji su (strogo) veći od medijana. Svojevrsni *srednji član* niza podataka. Podatci ne moraju uvijek biti različiti. Primjerice, u nizu

$$1, 2, 2, 2, 2,$$

takav broj ne postoji. Zato medijan definiramo nešto *blažom* definicijom.

Definicija 1. Za brojčani niz podataka $x_1, x_2,..., x_n$ medijan M je realan broj sa svojstvom da je:

– bar pola podataka veće ili jednako M;

– bar pola podataka manje ili jednako M.

Primjer 2. Navedimo nekoliko primjera brojčanih skupova i njihovih medijana:

Skup podataka	Medijan	Aritmetička sredina
-2, -1, **0**, 1, 2	0	0
2, 4, 5, **8**, 9, 11, 21	8	8.57
-3, 4	**Bilo koji broj iz intervala** [-3,4]	0.5
1, **2**, **2**, **2**, **2**	2	1.8

[1]Tvrtko Tadić, Microsoft AI & Research, Redmond / University of Washington, Seattle

U ovom primjeru možemo uočiti nekoliko stvari:

– ako je broj podataka neparan, medijan **ima vrijednost u tom skupu**;

– ako je broj podataka paran, medijan **može poprimiti cijeli interval vrijednosti**;

– aritmetička sredina može biti jednaka, veća ili manja od medijana.

Uočimo da su u primjeru 2. članovi niza bili poredani uzlazno i da nam je to omogućilo jednostavno izračunavanje medijana. Zato ćemo uvesti posebnu notaciju kada poredamo članove niza po redu.

Definicija 3. Niz $x_{(1)}, x_{(2)},..., x_{(n)}$ je permutacija niza $x_1, x_2,..., x_n$ takva da je

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

U idućem teoremu opisat ćemo medijan kao vrijednost koja se nalazi u sredini niza podataka nakon što ih poredamo po vrijednosti.

Teorem 4. Ako je

(i) n neparan, medijan je $M = x_{\left(\frac{n+1}{2}\right)}$;

(ii) n paran, svaki $M \in \left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)} \right]$ je medijan.

Dokaz. (i) Očito je da članova niza koji su manji ili jednaki M ima bar $\frac{n+1}{2} \geq \frac{n}{2}$. Isti je slučaj s članovima niza koji su veći ili jednaki od M, pa je prema definiciji 1. M medijan. Tvrdnja (ii) dokazuje se slično.

Računanje medijana

Skupovi podataka dani u primjeru 2. poredani su po veličini, pa nije teško naći medijan. Ako pak brojevi nisu poredani, primjerice:

-1.1, 100, -20, 4.3, 20.9, 3.4, -7.6, 40.3, -80.2, 30,

problem računanja medijana postaje bitno kompliciraniji. A što je više brojeva u nizu, postupak postaje zahtjevniji.

Kod medijana postoji problem koji nismo imali kod aritmetičke sredine. Naime, medijan, osim što nije jednoznačno definiran, **nema jasan postupak kojim ga utvrđujemo**. Za aritmetičku sredinu postupak je izravno dan u definiciji, dok ovdje to nije slučaj. Jedan od načina mogli bismo preuzeti iz Teorema 4:

- Poredaj članove niza $x_1, x_2,..., x_n$.

- Za medijan stavi vrijednost

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{ako je } n \text{ neparan,} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{ako je } n \text{ paran.} \end{cases} \quad (*)$$

Postoje efikasniji načini računanja medijana, koji ne uključuju sortiranje brojeva. Čitatelj može naći druge metode, kao i zanimljivu priču o ovom problemu u knjizi (Knuth 1998., 207-216) koja je posvećena sličnim problemima.

Većina alata za obradu podataka (poput Excela) i programskih jezika ima ugrađenu funkciju za računanje medijana, stoga se time dalje nećemo opterećivati.

Podatci o letovima

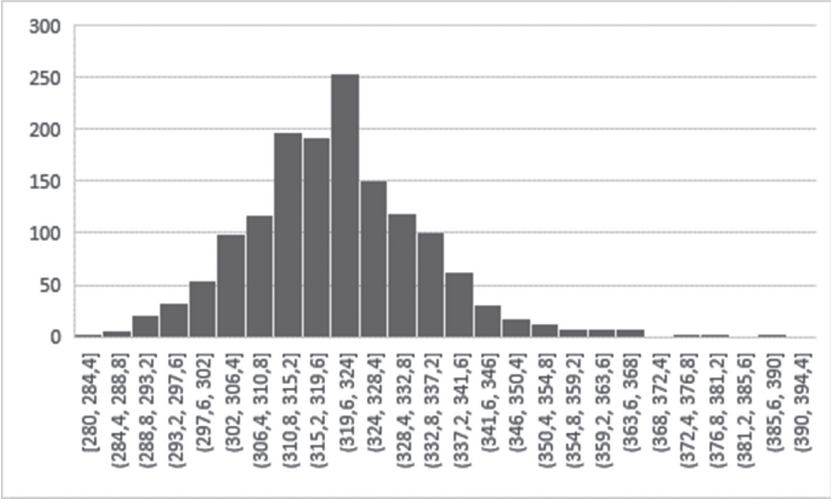
U ovom članku uzet ćemo **podatke o trajanju** i **kašnjenju letova** (vidi (United States Department of Transportation)) koje Ministarstvo prometa SAD-a skuplja i javno su dostupni. Na ovim podatcima proučavat ćemo pojmove o kojima govorimo u ovom članku.

Odlučili smo ići na godišnji odmor na Havaje, a prije toga posjetit ćemo Los Angeles. Zanimaju nas letovi između Los Angelesa i Honolulua. Zbog poslovnih razloga letjet ćemo United Airlinesom. Podatke ćemo skinuti sa spomenute internetske stranice. Našli smo 1516 letova ovog prijevoznika na relaciji Los Angeles – Honolulu. Tablica 1. pokazuje početak niza podataka (koji cijeli ne bi stao u ovaj članak).

Šifra prijevoznika	Datum	Broj leta	Broj aviona	Polazište	Odredište	Trajanje leta (min)	Kašnjenje aviona u polasku (min)
UA	1. siječanj 2016.	708	N210UA	LAX	HNL	332	88
UA	1. siječanj 2016.	1431	N57870	LAX	HNL	334	0
UA	1. siječanj 2016.	1158	N57863	LAX	HNL	339	36
UA	1. siječanj 2016.	1170	N77871	LAX	HNL	340	0
UA	1. siječanj 2016.	1232	N78866	LAX	HNL	341	0
UA	1. siječanj 2016.	1224	N77865	LAX	HNL	346	15
UA	2. siječanj 2016.	708	N211UA	LAX	HNL	328	0

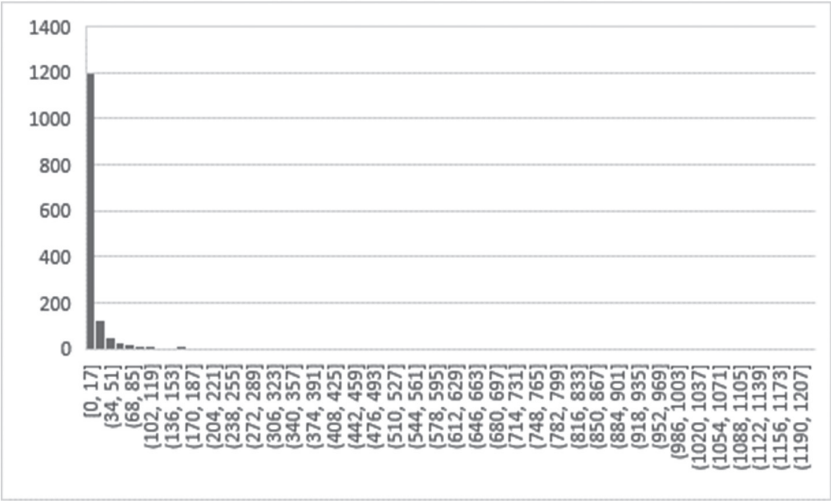
Tablica 1. Popis letova Los Angeles – Honolulu.

Sažetu informaciju o trajanju leta i kašnjenju zrakoplova u dolasku možemo vidjeti na Slici 1. i 2.



Slika 1. Histogram trajanja leta u minutama

U prošlom članku (Tadić 2017) spomenuli smo **podatke koji se ponašaju *normalno***, tj. podatci se grupiraju oko neke vrijednosti i njihov histogram *prati* zvonoliku krivulju. Podatci na slici 1. *ponašaju se normalno*, dok to ne možemo reći za slučaj na slici 2.



Slika 2. Histogram kašnjenja letova (u minutama)

Razlog je što imamo dva niza podataka koja se ponašaju potpuno različito:

- Let u većini slučajeva traje 5.5 sati +/- pola sata. Najkraći let u 2016. godini trajao je oko 4 sata i 40 minuta, a najdulji 6.5 sati.
- Velika većina letova kasni do 17 minuta. No bilo je i letova koji su u polasku kasnili preko 20 sati!

Na primjeru ovih dvaju nizova podataka objasnit ćemo zašto je ponekad medijan bolja mjera za srednju vrijednost od aritmetičke sredine.

Usporedba vrijednosti

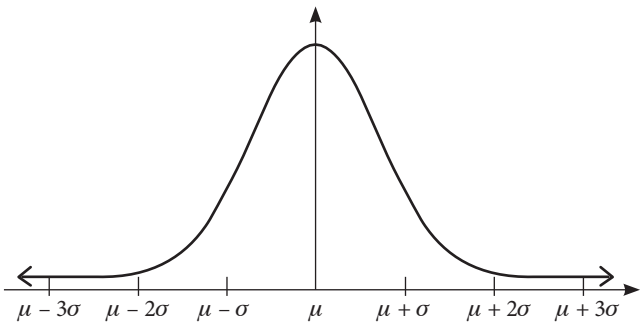
Za nizove podataka o trajanju i kašnjenju letova izračunat ćemo karakteristične vrijednosti nizova – minimum (najmanju vrijednost), aritmetičku sredinu, medijan i maksimum (najveću vrijednost):

Opis vrijednosti	Trajanje leta	Kašnjenje leta
Minimum	280	0
Aritmetička sredina	320.79	6.31
Medijan	320	0
Maksimum	394	1213

Tablica 2. Karakteristične vrijednosti trajnja i kašnjenja letova

Ako pogledamo tablicu, možemo uočiti nekoliko stvari:

- Aritmetička sredina i medijan trajanja letova imaju približno istu vrijednost. To je posljedica činjenice da su podatci normalno distribuirani, tj. histogram prati Gaussovu (zvonoliku) krivulju čija je jednadžba $y = broj\ podataka \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{\sigma^2}}$, gdje su μ, σ parametri. Sa slike 3. jasno je vidljivo da je krivulja simetrična oko vrijednosti parametra μ.



Slika 3. Gaussova krivulja

Krivulja *najbolje prati* histogram ako stavimo da je μ jednak aritmetičkoj sredini. Ako histogram podataka prati krivulju, zbog simetričnosti, očekujemo da broj podataka manjih od μ i većih od μ bude *otprilike* jednak. Za to ni medijan ne može mnogo odstupati od μ, odnosno u ovom slučaju vrijednosti aritmetičke sredine.

- U slučaju kašnjenja letova aritmetička nas sredina može zavarati. Stječe se pogrešan dojam da nećemo zakasniti na let, jer u prosjeku kasni 6 minuta. No od 1506 letova o kojima imamo podatke, njih 1260 (preko 73 %) uopće nije kasnilo u polasku (tj. ti letovi kasnili su 0 minuta). Ako ste zakasnili na let više od 6 minuta, propustili ste ga u 1292 (preko 75 %) slučajeva.

Ekstremni primjer

Kao što smo vidjeli, medijan nekad daje *stvarniju* informaciju o brojčanim podatcima od aritmetičke sredine jer je ili jednak jednoj od tih vrijednosti ili se nalazi između dviju vrijednosti.

Promotrimo ponovo primjer iz prošloga članka u kojemu je Bill Gates posjetio prihvatilište za 1000 izbjeglica (Tadić 2017). Kao što smo naveli, Bill Gates ima 80 i nešto milijardi dolara, dok 1000 izbjeglica nema ništa (0 dolara). Aritmetička sredina bila je približno 80 milijuna, dok je medijan jednak 0. U ovom slučaju medijan izgleda kao *primjerenija* mjera srednje vrijednosti nego aritmetička sredina.

Uočimo da je medijan i *stabilniji* od aritmetičke sredine. Medijan bogatstva ljudi u prihvatilištu ostao je stabilan dolaskom Billa Gatesa, dok se vrijednost aritmetičke sredine dramatično povećala.

Teorijska pozadina medijana

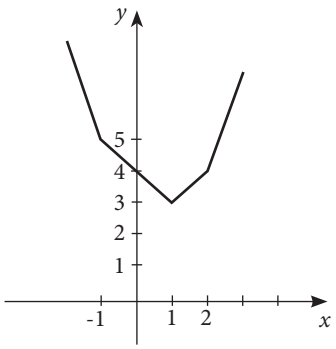
Premda medijan možemo shvatiti kao vrijednost koju dobijemo kada vrijednosti poredamo po veličini i odaberemo srednju, medijan možemo opisati slično kao i aritmetičku sredinu (vidi teorem 3. i posljedice u (Tadić 2017)). Medijan je vrijednost za koju je *zbroj udaljenosti do svih brojeva na brojevnom pravcu* ***minimalan***, tj. vrijedi idući teorem.

Teorem 5. Funkcija

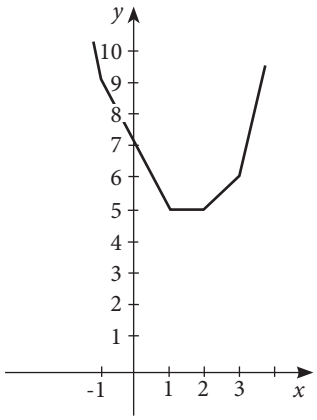
$$f(x) = |x - x_1| + |x - x_2| + \ldots + |x - x_n|$$

postiže minimum za svaku vrijednost x = M za koju je M medijan niza podataka $x_1, x_2, \ldots, x_n$.

Prije samog dokaza na dva grafa funkcije ilustrirat ćemo tvrdnju teorema na slikama 4. i 5.



Slika 4. Graf funkcije f za niz 1, -1, 2. Funkcija postiže minimum u točki x = 1, što je ujedno medijan niza.



Slika 5. Graf funkcije f za niz 1, -1, 2, 3. Funkcija postiže minimum u svakoj točki x iz intervala [1,2], gdje se ujedno nalaze svi medijani niza.

Dokaz. Funkcija f može se zapisati na idući način:

$$f(x)=\left|x-x_{(1)}\right|+\left|x-x_{(2)}\right|+\ldots+\left|x-x_{(n)}\right|.$$

Nadalje, ne moraju svi brojevi $x_{(1)}, x_{(2)},\ldots, x_{(n)}$ biti različite vrijednosti. Neka je $y_1 < y_2 < \ldots < y_r$ skup različitih vrijednosti koje niz $x_{(1)}, x_{(2)},\ldots, x_{(n)}$ poprima, a $n_1,\ldots, n_r$ broj koliko se puta svaka ta vrijednost pojavila u nizu. Funkciju sada možemo zapisati kao

$$f(x)=n_1\left|x-y_1\right|+n_2\left|x-y_2\right|+\ldots+n_r\left|x-y_r\right|.$$

Za $x\in\left[y_j, y_{j+1}\right]$ vrijedi

$$f(x)=\left(n_1+n_2+\ldots+n_j-n_{j+1}-\ldots-n_r\right)x-\left(n_1y_1+\ldots+n_jy_j-n_{j+1}y_{j+1}-n_ry_r\right). \quad (1)$$

Ponovo gledamo dva slučaja ovisno o parnosti broja *n*:

- Medijan niza $x_1, x_2,\ldots, x_n$ je samo jedna vrijednost y_k. Uočimo da je tada prema definiciji medijana $n_1+n_2+\ldots+n_k =$ (broj svih članova niza $\le y_k$) > (broj svih članova niza > y_k) $=n_{k+1}+\ldots+n_r$. Slično, ponovo koristeći definiciju medijana, slijedi $n_1+n_2+\ldots+n_{k-1}<n_k+n_{k+1}+\ldots+n_r$.

  Iz (1) sada lako slijedi da na intervalu $\left(-\infty, y_k\right)$ funkcija *f* pada, dok na intervalu $\left(y_k,\infty\right)$ raste. Stoga je y_k minimum funkcije *f*.

- Medijan niza x_1, x_2,..., x_n svaka je vrijednost iz intervala $[y_k, y_{k+1}]$. Tada je $n_1 + ... + n_k = n_{k+1} + ... + n_r$. Sada iz (1) slijedi da na intervalu $(-\infty, y_k)$ funkcija f pada, na intervalu $[y_k, y_{k+1}]$ je konstantna, a na intervalu (y_{k+1}, ∞) raste. Stoga se minimum postiže na cijelom intervalu $[y_k, y_{k+1}]$.

Ovaj nam teorem kaže da je medijan vrijednost koja je, kad se mjeri na način prikazan u teoremu 5., najbliži broj podatcima. Uočimo kako je za aritmetičku sredinu vrijedilo slično, jedino je način mjerenja bio drukčiji.

Kvantili i 90% pouzdani interval

Podatke o kašnjenju aviona u polasku željeli bismo sažeti da daju stvarnu sliku o tome koliko let može kasniti. Pogledajmo kakav dojam možemo steći iz podataka koje smo do sada izračunali:

- Medijan je jednak 0 i iz njega bi se stekao dojam da zrakoplov na vrijeme polijeće.
- Iz aritmetičke sredine stekao bi se dojam da zrakoplov zakasni u polasku, ali nekoliko minuta.
- Ako damo informaciju o rasponu podataka, tj. intervalu [minimum maksimum], dobit ćemo nejasnu informaciju da let kasni u polasku od 0 minuta do 20 sati.

Ideja je da veličinu koju mjerimo u ovom slučaju procijenimo intervalom oblika [A,B], gdje bismo jedan dio nižih vrijednosti i jedan dio gornjih vrijednosti *zanemarili*. Let koji je kasnio 20 sati dogodio se samo jednom i radi se o *iznimnom slučaju* koji se *gotovo nikada* ne događa. Želimo da interval [A,B] sadrži veliku većinu vrijednosti koje su se dogodile, ali da zanemari ekstremne vrijednosti. U tome će nam pomoći kvantili.

Definicija kvantila i svojstva

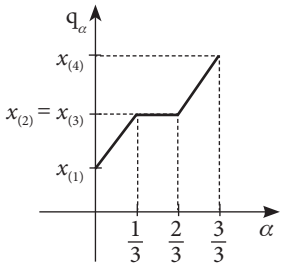
Za $\alpha \in [0,1]$ želimo izabrati broj q_α takav da je

- udio podataka u nizu manjih ili jednakih q_α približno α,
- udio podataka u nizu većih ili jednakih q_α približno $1 - \alpha$.

Postoji više načina kako možemo odrediti q_α. U idućoj definciji dat ćemo jedan način.

Definicija 3. Za brojčani skup podataka x_1, x_2,..., x_n i $\alpha \in [0,1]$. Neka je k takav da je $k < \alpha(n-1)+1 \leq k+1$. Tada α-kvantil definiramo kao

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



Slika 6. Primjer grafa funkcije $f(\alpha) = q_\alpha$.

Napomena. Čitatelj treba biti svjestan nekoliko stvari:

- Za $\alpha = \frac{k-1}{n-1}$ imamo $q_\alpha = x_{(k)}$.
- 0.5-kvantil uvijek je medijan. (Vidi Teorem 4. i relaciju (*).)
- Postoje razne definicije kvantila; ovo je definicija koja se koristiti u Excelu.

Primjer 4. Pogledajmo nekoliko primjera:

Niz podataka	α	α-kvantil
1, 2, 3, 4	2/3	3
-1, 2, 3	**1/2**	**2**
-1, 3, 4, 11	**1/4**	**0.5**

Dodatno vrijedi teorem 5. koji opisuje kvantile slično kao definicija 1. medijane:

Teorem 5. Vrijedi:

- udio podataka u nizu manjih ili jednakih od q_α je barem $\alpha\left(1-\frac{1}{n}\right)$,
- udio podataka u nizu većih ili jednakih od q_α je barem $(1-\alpha)\left(1-\frac{1}{n}\right)$.

Dokaz. U intervalu $(-\infty, q_\alpha]$ sigurno se nalaze vrijednosti $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Iz definicije 3. α-kvantila znamo da je $\alpha(n-1)+1 \le k+1$, iz čega slijedi da je $\alpha\left(1-\frac{1}{n}\right) \le \frac{k}{n}$.

S druge strane, u intervalu $[q_\alpha, +\infty)$ sigurno se nalaze vrijednosti $x_{(k+1)}, x_{(k+2)}, \ldots, x_{(n)}$.

Iz definicije 3. znamo da je $k < \alpha(n-1)+1$, te slijedi $n-1-\alpha(n-1) < n-k$, odnosno $(1-\alpha)\left(1-\frac{1}{n}\right) < \frac{n-k}{n}$.

Napomena. Ukoliko su svi članovi niza podataka jednaki istoj vrijednosti *x*, onda je i $q_\alpha = x$, te se u intervalima $(-\infty, q_\alpha]$ i $[q_\alpha, +\infty)$ nalaze svi podatci.

U praksi je čest slučaj da se vrijednosti podataka neće ponavljati. U tom slučaju dobro je znati iduću posljedicu.

Posljedica 6. Ako niz $x_1, x_2,..., x_n$ sadrži brojeve različitih vrijednosti, vrijedi sljedeće:

- Udio podataka u nizu manjih ili jednakih od q_α je između $\alpha\left(1-\frac{1}{n}\right)$ i $\alpha\left(1-\frac{1}{n}\right)+\frac{1}{n}$.

  Dakle, za ***velike vrijednosti*** broja n, udio je približno α.

- Udio podataka u nizu većih ili jednakih od q_α je barem $(1-\alpha)\left(1-\frac{1}{n}\right)$ i $(1-\alpha)+\frac{\alpha}{n}$.

  Dakle, za ***velike vrijednosti*** broja n, udio je približno $1-\alpha$.

5 %-kvantil i 95 %-kvantil za podatke o letovima

U tablici 3. dani su 5%-kvantil i 95%-kvantil za trajanje leta i kašnjenje leta u polasku.

Niz podataka / opis vrijednosti	5 %-kvantili	95 %-kvantil
Trajanje leta	300	345
Kašnjenje leta	0	87.75

Tablica 3. Izračun kvantila za trajanje i kašnjenje leta

Sada je jasno da su za interval [A,B] kojim želimo opisati gdje se nalazi glavnina vrijednosti niza, dobar kandidat

$$A = q_{0.05} \text{ i } B = q_{0.95}.$$

Interval $[q_{0.05}, q_{0.95}]$ ima nekoliko dobrih svojstava:

- Najmanje 90 % članova niza podataka ima vrijednosti u tom intervalu. (Uočimo: ako su sve vrijednosti iste, interval je jedna točka i svi članovi niza imaju vrijednosti u tom intervalu.)
- 5 % najnižih i 5 % najviših vrijednosti su zanemarene. Ovo je dobro jer se ekstremne vrijednosti često postižu u okolnostima koje nisu uobičajene. Primjerice, kašnjenje od 20 sati možda je uzrokovano snježnom olujom koja do tada nikad nije bila zabilježena, ili se let od 280 minuta dogodio jer je taj dan putnike prevozio brži zrakoplov koji nikada ne leti na toj relaciji.

Definicija 7. Interval $[q_{0.05}, q_{0.95}]$ zovemo **90 %-pouzdani interval**.

Koristimo ga kako bismo u praksi priopćili da:

- let između Los Angelesa i Honolulua u (bar) 90 % slučajeva traje od 300 do 345 minuta,
- isti let u polasku u (bar) 90 % slučajeva kasni od 0 do 87 minuta.

Pogled na buduće podatke

Postavlja se pitanje za što možemo iskoristiti ove zaključke. Ovi zaključci korisni su za planiranje leta. Neka od idućih pitanja su:

- Kada okvirno stižemo? Koliko bi let mogao kasniti u polasku?
- Koliko hrane da ponesemo u avion? Koliko stranica knjige možemo pročitati ili filmova pogledati u zrakoplovu?
- Koliko ćemo dugo moći spavati u avionu?

Na temelju podataka sličnih ovima razne aplikacije danas daju preporuke o tome kako uspješno / bolje nešto napraviti.

Podatci iz kojih smo dobili medijan i kvantile su iz 2016. godine, a u vrijeme pisanja ovog članka dostupni su nam i podatci iz siječnja 2017. godine. U idućoj tablici pokazat ćemo koliko dobro statistike iz 2016. predviđaju letove u 2017.

Podatci	**90% pouzdani interval iz 2016**	**Postotak podataka iz siječnja 2017. u intervalu**	**Medijan iz 2016.**	**Postotak podatka ne većih od medijana**	**Postotak podataka ne manjih od medijana**
Trajanje leta	[300,345]	77.23	320	27.72	75.25
Kašnjenje leta	[0,87.5]	93.07	0	77.23	100.00

Tablica 4. Usporedba podataka iz 2016. godine s onima iz siječnja 2017.

Uočimo:

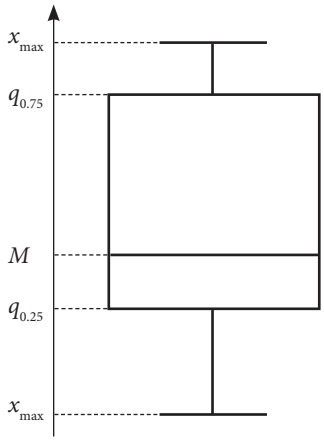
- Kašnjenje leta u siječnju 2017. odvija se prema očekivanjima iz 2016. Glavnina letova ide na vrijeme, medijan je 0 i preko 90 % vrijednosti nalazi se u 90 % pouzdanom intervalu iz 2016.

  Jedno od objašnjenja ovakvog rezultata može biti da kašnjenja tijekom leta ne odstupaju mnogo od očekivanog.

- Trajanje leta ne izgleda loše, ali se čini da let ipak traje nešto dulje nego što je bilo uobičajeno u 2016. godini. Zbog čega je došlo do toga? Zato jer smo podatke za cijelu 2016. uspoređivali s podatcima **samo za siječanj 2017.** Slično bismo dobili kad bismo uspoređivali podatke o temperaturi kroz cijelu godinu s onima za razdoblje od mjesec dana. Zato za bolje predviđanje moramo uzeti u obzir još neke faktore.

Zadatak

Zaključivanje na temelju podataka zahtijeva dosta iskustva. Čitatelji mogu pokušati usporediti podatke o trajanju leta za siječanj 2017. s podatcima za siječanj iz više prethodnih godina (2016., 2015., 2014.,...) i vidjeti u kolikoj se mjeri ti podatci nalaze u pouzdanim intervalima.



Slika 7. Brkata kutija

Prikaz podatka brkatom kutijom

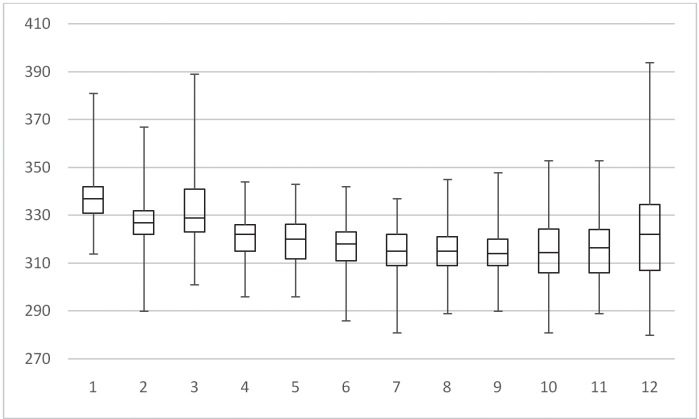
Sažetu informaciju o podatcima, koja uključuje i njihov puni raspon, obično radimo pomoću karakteristične petorke

$$(x_{min}, q_{0.25}, M, q_{0.75}, x_{max}).$$

Kao što vidimo, brojeve smo odabrali tako da vrijednosti između svaka dva uzastopna broja u petorci poprima najmanje četvrtina članova niza. Sada ovu petorku ilustriramo pravokutnikom s brkovima kao na slici 7:

- Gornja i donja stranica pravokutnika nalaze se redom u visini 75 %-kvantila i 25 %-kvantila.
- Pravokutnikom prolazi crta u visini medijana.
- Dvije manje crtice (brkovi) dodani su visini maksimuma i minimuma.
- Polovišta gornje i donje crtice spojena su redom s polovištem gornje i donje stranice pravokutnika.

Ovo često koristimo da bismo bolje prikazali usporedbu podataka. Na slici ćemo pomoću brkate kutije pokazati kako opisujemo promjenu trajanja leta kroz mjesece u 2016. godini.



Slika 8. Brkate kutije trajanja leta kroz mjesece u 2016. godini.

Sa slike 8. možemo vidjeti da u 1., 2., 3. i 12. mjesecu letovi mogu biti nešto duži, kao i da je u tim mjesecima medijan trajanja leta nešto veći.

Procjena medijana i kvantila na temelju uzorka

U prethodnim člancima (vidi (Tadić, Podatci i uzorak 2016)) govorili smo o tome kako zbog nedostatka podataka često moramo donositi zaključke na manjem uzorku. Nećemo uvijek biti te sreće da imamo sve podatke dostupne, ali ćemo i iz manjeg slučajnog uzorka moći procijeniti medijan i kvantile. Uzeli smo slučajni uzorak duljine 100.

Vrijednost	Uzorak trajanja leta	Trajanje leta	Uzorak kašnjenja leta	Kašnjenje leta
5%-kvantil	293.9	300	0	0
medijan	320	320	0	0
95%-kvantil	343.05	345	21.15	87.75

Tablica 5. Kvantili i medijan iz uzorka i svih podataka

Uočimo:

- da su medijan i kvantili dobiveni iz uzorka trajanja leta *bliski* onima koje dobijemo iz svih podataka,
- za kašnjenje leta, s druge strane, moglo bi se činiti da imamo veliko odstupanje, no ako pokušamo prebrojiti (pomoću Excela ili nekog programskog jezika) koliko je (svih) članova niza podataka o kašnjenju leta u intervalu između procijenjenih 5 %-kvantila i 95 %-kvantila, na temelju uzorka dobivamo 82.27 %. To je dobro s obzirom da na duljinu uzorka i činjenicu da su podatci s većim vrijednostima *rijetki* i *raspršeni*.

Uzorak	90% pouzdani interval uzorka	Postotak svih podataka u intervalu
Trajanje leta	[293.9, 343.05]	92.36
Kašnjenje leta	[0, 21.15]	82.27

Tablica 6. Postotak podataka u procijenjenom uzoračkom intervalu

Zaključak

U ovom članku imali smo priliku vidjeti da je, uz aritmetičku sredinu, dobro koristiti i medijan kao mjeru za srednju vrijednost podataka. Mogli smo vidjeti da su, u slučaju da se podatci normalno ponašaju, te dvije vrijednosti iste. Također smo vidjeli kakvu nam interpretaciju daje medijan kod podataka koji nisu normalno distribuirani. Medijan također ima zanimljivu teorijsku pozadinu, kao i složeniji način

izračunavanja od aritmetičke sredine. Treba imati na umu da medijan, kao ni aritmetička sredina, **nije savršen**, jer je to tek jedan broj koji pokušava nešto reći o cijelom nizu podataka.

Upoznali smo se s kvantilima koji su nam omogućili da bolje izrazimo u kojem se intervalu kreće većina vrijednosti niza. Na kraju smo se upoznali s prikazom podataka pomoću brkate kutije i karakteristične petorke.

Upoznali smo se s ovim pojmovima i dijelom njihove teorije prilagođene školskom gradivu kroz zanimljive stvarne primjere koristeći računalo. Ovo je jedan od modela kako bi se statistika i obrada podataka mogla približiti učenicima.

Dodatak A. Excel

Svi podatci korišteni u ovom članku, kao i Excel bilježnice, dostupni su na lokaciji https://web.math.pmf.unizg.hr/~tvrtko/metodikaStatistike/clanak3.

Informacije kako djelatnici i polaznici obrazovnog sustava mogu dobiti Excel za svoje potrebe mogu se naći na

https://office365.skole.hr/.

Brkate kutije po mjesecima

Mnoge stvari u Excelu prikazane su u prethodnim člancima. Sada ćemo pokazati kako napraviti dijagram s brkatim kutijama prikazan na slici 8. i pokazati koje funkcije pri tome koristimo. Podatci se nalaze u tablici kao što je prikazano na slici 9.

	A	B	C	D	E	F	G	M
1	Šifra prijevoznika	Datum	Broj leta	Broj aviona	Polazište	Odredište	Trajanje leta (u minutama)	Kašnjenje aviona u polasku (u minutama)
2	UA	1/1/2016	708	N210UA	LAX	HNL	332	88
3	UA	1/1/2016	1431	N57870	LAX	HNL	334	0
4	UA	1/1/2016	1158	N57863	LAX	HNL	339	36
5	UA	1/1/2016	1170	N77871	LAX	HNL	340	0
6	UA	1/1/2016	1232	N78866	LAX	HNL	341	0
7	UA	1/1/2016	1224	N77865	LAX	HNL	346	15

Slika 9. Podatci u Excelu

Izračun potrebnih vrijednosti

Sada ćemo u nekom drugom dijelu lista napraviti tablicu koja će po mjesecima izračunati minimum, 25 %-kvantil, medijan, 75 %-kvantil i maksimum trajanja leta.

P39 {=MIN(IF(MONTH(B$2:B$1506)=O39,G$2:G$1506))}

	O	P	Q	R	S	T
38	Mjesec	Minimum	25%-kvanitl	Medijan	75%-kvanitl	Maksimum
39	1	314	331	337	342	381
40	2	290	322	327	332	367
41	3	301	323	329	341	389
42	4	296	315	322	326	344
43	5	296	311.75	320	326.25	343
44	6	286	311	318	323	342
45	7	281	309	315	322	337
46	8	289	309	315	321	345
47	9	290	309	314	320	348
48	10	281	306	314.5	324.25	353
49	11	289	306	316.5	324	353
50	12	280	307	322	334.5	394

Slika 10. Tablica izračunatih vrijednosti

Za izračunavanje koristimo iduće Excelove funkcije:

Naziv	Primjer
MIN (vrijednosti)	MIN (A1:A100) - najmanja vrijednost među prvih 100 u stupcu A
MAX (vrijednosti)	MAX (A1:A100) - najveća vrijednost među prvih 100 u stupcu A
MEDIAN (vrijednosti)	MEDIAN (A1:A100) - medijan prvih 100 vrijednosti u stupcu A
PERCENTILE (vrijednosti, alpha)	PERCENTILE (A1:A100, 0.2)[2] - 0.2-kvantil prvih 100 vrijednosti u stupcu A

Za izračun koristimo dane funkcije u Excelu koristeći datum u stupcu B kako bismo ograničili vrijednosti samo na one koje su se dogodile u željenom mjesecu. Za to koristimo sljedeću sintaksu:

IF(MONTH(B$2:B$1506)=O39,G$2:G$1506)

Ova naredba daje nam vrijednosti iz stupca G, gdje je vrijednost mjeseca izračunatog na temelju vrijednosti u stupcu B istog retka jednaka onoj u polju O39 (u ovom slučaju 1). Na slici 9. može se vidjeti kako je vrijednost u polju P39 izračunata. Kako je IF ugniježđena naredba unutar funkcije, treba istovremeno stisnuti Shift i Ctrl te onda Enter kako bi se pravilno izvrednovao izraz.

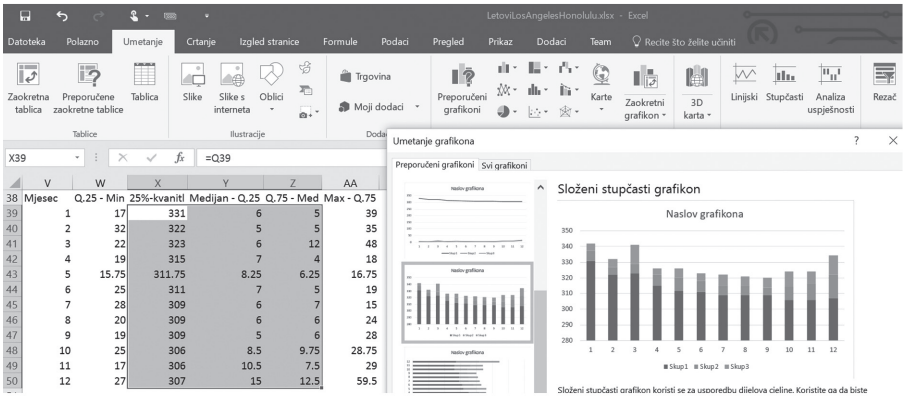
Crtanje brkatih kutija

Kako bismo nacrtali brkatu kutiju, trebamo malo modificirati tablicu sa slike 10. Uz vrijednost 25%-kvantila trebat će nam razlike između:

25 %-kvantila i minimuma; medijana i 25 %-kvantila;
75 %-kvantila i medijana; maksimuma i 75 %-kvantila.

[2]PERCENTILE je stara funkcija u Excelu koja zamijenjena u novim verzijama funkcijom PERCENTILE.INC te je još dodana funkcija PERCENTILE.EXC koja računa kvantile na jedan drugi način. PERCENTILE još radi zbog kompatibilnosti starih verzija Excela, ali se preporuča korištenje novih funkcija.

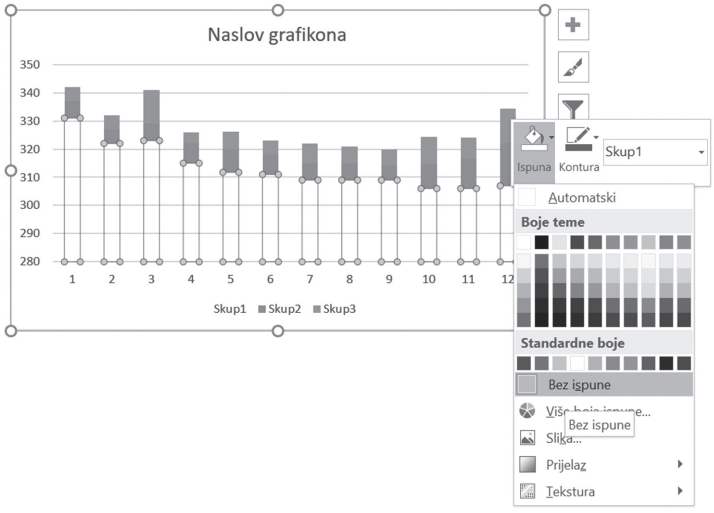
Nakon što napravimo novu tablicu, kao na slici 11., odaberemo 3 unutrašnja polja te u izborniku *Umetanje* odaberemo polje *Preporučeni grafikoni*. U izborniku koji se pojavio pod *Preporučeni grafikoni* odaberemo *Složeni stupčasti grafikon*.



	V	W	X	Y	Z	AA
38	Mjesec	Q.25 - Min	25%-kvantil	Medijan - Q.25	Q.75 - Med	Max - Q.75
39	1	17	331	6	5	39
40	2	32	322	5	5	35
41	3	22	323	6	12	48
42	4	19	315	7	4	18
43	5	15.75	311.75	8.25	6.25	16.75
44	6	25	311	7	5	19
45	7	28	309	6	7	15
46	8	20	309	6	6	24
47	9	19	309	5	6	28
48	10	25	306	8.5	9.75	28.75
49	11	17	306	10.5	7.5	29
50	12	27	307	15	12.5	59.5

Slika 11. Cratnje brkate kutije prvi korak

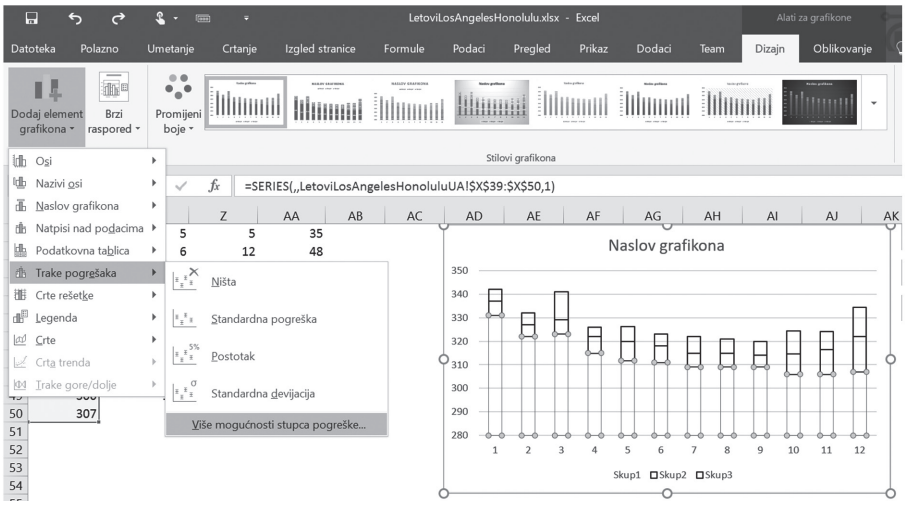
U idućem koraku riješit ćemo se donje boje tako da kao na slici 12. kliknemo na taj dio stupca i odaberemo polje *Ispuna,* a zatim u izborniku odaberemo bez *Bez ispune*. Na taj način dobili smo dva dijela kutije. Kutiju možemo obojiti ponuđenom bojom ili možemo ostaviti praznom i dodati okvir u nekoj boji.



Slika 12. Dobivanje kutije

U daljnjem koraku kutijama ćemo dodati trake pogreške, tj. brkove. Ovo radimo u dva koraka:

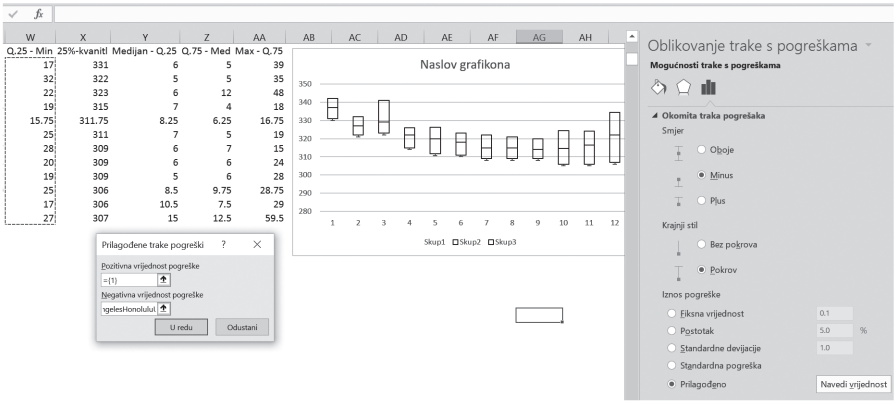
– za gornji dio kutije
– za donji dio stupca koji se više ne vidi



Slika 13. Crtanje brkova

Pod *Alati za grafikone* > *Dizajn* > *Dodaj element grafikona* > *Trake pogrešaka* odaberimo *Više mogućnosti stupca pogreške* kao na slici 13. Sada u izborniku *Oblikovanje trake s pogreškama* odaberemo:

– *Smjer*: *Minus* (za gornji dio kutije smjer plus)
– *Krajnji stil*: *Pokrov*
– Iznos pogreške: Prilagođeno i odaberemo stupac (W na slici 14) koji sadrži podatke o razlici 25 %-kvantila i minimuma (za gornji dio kutije stupac AA na slici 14. koji sadrži podatke o razlici maksimuma i 75 %-kvantila).



Slika 14. Odabir visine brkova

Dodatno, odabirom na vrijednosti grafikona možemo smanjiti raspon osi ordinate.

Dodatak B. Python

Za razliku od prethodnih članaka, u ovom članku nije korišten programski jezik Python. Ova tematika zadire duboko u područje računarstva koje se bavi *sortiranjem i traženjem podataka*, što je vrlo detaljno obrađeno u knjizi (Knuth 1998.). Neke od tih tema su standardni dio naprednijeg predmeta iz informatike (vidi recimo (Budin, i dr. 2014.)). Dio tih tema zahtijeva uvođenje posebnih struktura podatka.

Bilo bi svakako zgodno kad bi se učenici koji imaju priliku obrađivati takve sadržaje u nastavi informatike okušali sa stvarnim podatcima i isprobali neke od primjera navedenih u ovom članku na većoj količini podataka.

Literatura:

1. Budin, Leo, Predrag Brođanac, Zlatka Markučić, Smiljana Perić. *Napredno rješavanje problema programiranjem u Pythonu.* Zagreb: Element, 2014.
2. Knuth, Donald E. *The Art of Computer Programing : Volume 3 / Sorting and Searching.* Addison – Wesley, 1998.
3. Tadić, Tvrtko. Artimetička sredina i standardna devijacija. *Poučak 69*, 2017: 10-28. Podatci i uzorak. *Poučak 67*, 2016: 16-26.
4. United States Department of Transportation. Airline On-Time Statistics . *Bureau of Transportation statistics.* n.d. https://www.transtats.bts.gov/ONTIME/Airborne.aspx.
5. Varošanec, Sanja. Peteljka-list dijagaram. Medijan i mod. *Matematika i škola 71*, 2013.: 10-13. Karakteristična petorka i brkata kutija. *Matematika i škola 72*, 2013.: 56-59.
6. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference.* New York: Springer, 2005.