Matematika iza anketa - primjer izbora

Tvrtko Tadić

28. srpnja 2010.

Mediji su puni anketa u kojima se postavljaju razna pitanja. Cilj svih tih anketa je pokušati dati procjenu nekog podatka.

Primjerice: *Koliki postotak stanovništva puši? Koliki postotak gledateljstva je gledao točno određenu nogometnu utakmicu? Koliki postotak stanovništva ima završen fakultet?*

Na neka od njih **točne odgovore** nikada nećemo dobiti. Primjerice do podataka o točnom udjelu pušača u stanovništvu ili postotku ljudi koji su gledali određenu utakmicu, ne možemo doći, jer to ne možemo doznati na drugi način nego ispitivanjem cijele populacije. (Što se za ove podatke ne radi.) S druge strane podatak o postotku stanovništva s završenim fakultetom dobijemo svakih 10 godina popisom stanovništva.

U medijima najčešće ankete su one o popularnosti političkih stranaka i političara. Posebno interesantne su ankete baš na dan samih izbora (tzv. *izlazne ankete*). Nakon zatvaranja birališta rezultati izbora biti će poznati za nekoliko sati, ali javnost voli da joj se da naznaka/procjena kako bi konačni rezultat mogao izgledati.

U tu svrhu se obično ispituje manji dio birača izašlih na izbore i na temelju njega se daje procjena kako bi mogao rezultat izgledati. Praksa je pokazala da ovaj način predviđanja, uglavnom, daje ***približno točne rezultate***. Zato su izlazne ankete najbolji pokazatelj kako ima smisla provoditi istraživanja na **manjem dijelu populacije** i donositi procjene kako bi to moglo izgledati na **nivou cijele populacije**.

Zašto je tome tako? Koje je matematičko opravdanje ovog postupka? Tim pitanjima pozabavit ćemo se u ovom članku.

Odgovore na ova pitanja dati ćemo upravo kroz primjer izbora i izlaznih anketa. U ovom članku pretpostavljamo sljedeće o izborima i anketi koju proučavamo

- na izborima sudjeluju dva kandidata (kandidati A i B);
- svaki birač koji je izašao na izbore glasao je za jednog od njih;

- anketirani birači u anketi daju točne odgovore o tome za koga su glasali.

Uz neke izmjene model lako može funkcionirati i u drugim uvjetima.

1 Primjer izbora i simulacija ankete

Za simulaciju ankete uzet ćemo rezultate drugog kruga predsjedničkih izbora u Hrvatskoj održanog 2010. Rezultati glasova birača izašlih na izbore u Republici Hrvatskoj[1] dani su u tablici.

KANDIDAT	BROJ GLASOVA
Ivo Josipović	1330339
Milan Bandić	778915

Kako bi lakše baratali podacima zapisat ćemo podatke u vektor `izbori` tako da svaki glas za Ivu Josipovića zabilježimo brojem 1, a svaki glas za Milana Bandića zabilježimo brojem 0. (Glasove ovako kodiramo radi jednostavnosti i iz praktičnih razloga koji će se kasnije pokazati.) Simulacije ankete ćemo provesti u statističkom programu **R**.

```
> izbori=rep(c(0,1),c(778915,1330339))
```

U vektoru `izbori` na prvih 778915 mjesta nalazi se brojka 0, a na preostalih 1330339 brojka 1. Kako bi vidjeli koliko je glasova dobio Ivo Josipović u postotcima dovoljno je izračunati aritmetičku sredinu vektora `izbori`.

```
> mean(izbori)
[1] 0.6307154
```

Dakle, Ivo Josipović na području RH je dobio 63.07% (važećih) glasova birača.

Napravimo simulaciju ankete. Na slučajan način odaberimo 2000 različitih osoba i pitajmo ih za koga su glasale. To ćemo ovdje napraviti tako da odaberemo 2000 različitih indeksa vektora `izbori` i vrijednosti na tim mjestima složimo u vektor `anketa`.

```
> anketa=sample(izbori,2000)
> mean(anketa)
[1] 0.627
```

[1] Birači koji glasuju izvan Hrvatske glasuju širom svijeta pa je anketu van RH iz praktičnih razloga nemoguće provesti. (Simuliramo anketu koja se provodi po Hrvatskoj.)

Uzeli smo uzorak od 2000 slučajno odabranih glasača i doznali da među njima, kandidat kodiran s 1 ima 62.7% glasova.

Uočimo da se stvarni postotak dobivenih glasova i postotak dobiven anketom jako malo razlikuju! Relativno *jeftino*, koristeći uzorak manji od jednog promila izašlih birača, dobili smo **približno točnu procjenu konačnog rezultata**.

2 Provedimo više anketa

No dobro, jedna anketa je bila uspješna. Hoće li baš svaka biti uspješna? Očito da je moguće da anketa da krivu procjenu, ali koliko je to vjerojatno? Kako znamo da nismo imali sreće, pa nam se baš zalomila ovako dobra procjena rezultata? Zahvaljujući računalima, ankete možemo ponavljati proizvoljno mnogo puta. Ovdje ćemo provesti 1000 anketa da vidimo kako će se pokazati predviđanja rezultata. Predviđanja rezultata spremit ćemo u vektor `ankete`.

```
> ankete=rep(0,1000)
> for(j in 1:1000) ankete[j]=mean(sample(izbori,2000))
```

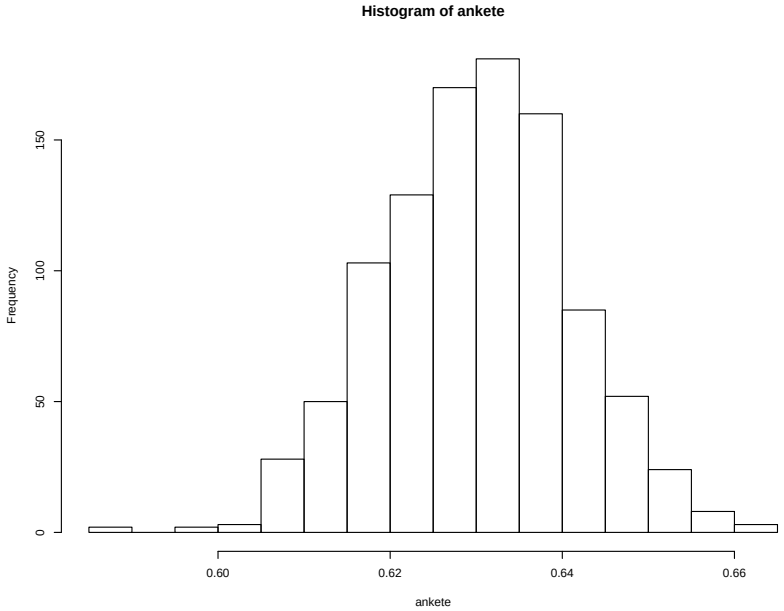
U kodu smo definirali vektor `ankete` u koji spremamo 1000 predviđanja na temelju 1000 anketa. Za svaku anketu ponovo slučajno biramo 2000 ljudi na kojima je provedena.

Pogledajmo u kojem rasponu su naše ankete predviđale postotak pobjednika.

```
> max(ankete)
[1] 0.6645
> min(ankete)
[1] 0.588
```

Dakle, anketa koja je predviđala najveći postotak za kandidata kodiranog s brojem 1 mu je predviđala 66.45% glasova, a anketa koja je predviđala najmanji postotak, mu je predviđala 58.8% glasova. Možemo zaključiti da svih 1000 anketa ne odstupa previše. Pregled kakve postotke te ankete daju možemo vidjeti na histogramu koji je dan na slici 1.

Vidimo da je velika većina anketa predviđa pobjedu kandidata kodiranog sa 1 u rasponu od 60% do 66%, a izvan toga se nalazi *zanemariv* broj anketa. Uočimo da su upravo stupci najbliži stvarnom rezultatu ujedno i najveći! (U većini slučajeva imamo odstupanje od stvarnog rezultata $\pm 3\%$ glasova.)



SLIKA 1. *Histogram predviđanja 1000 anketa*

3 Pretpostavke problema i oznake

U tekstu pretpostavljamo da imamo **dva kandidata** A i B, koji su redom dobili a i b glasova na izborima. S $N = a + b$ označavamo **ukupan broj izašlih**, a s malo n **broj anketiranih** glasača. Obično ćemo s k označavati broj anketiranih koji su se izjasnili da su glasovali za kandidata A u nekoj točno određenoj anketi. Sljedeća tablica ukratko opisuje što koja oznaka znači.

KANDIDAT	BROJ GLASOVA	BROJ ANKETIRANIH
A	a	k
B	b	$n-k$
UKUPNO	N	n

Ono što mi želimo procijeniti je vrijednost

$$p := \frac{a}{N} = \frac{a}{a+b},$$

tj. udio glasača koji su glasali za kandidata A na temelju ankete provedene na n ljudi.

4 Kombinatorni problem

Na koliko načina možemo od N birača izabrati njih n koje ćemo anketirati (tj. od N-članog skupa biramo n-člani podskup)? To je dobro poznato

$$\binom{N}{n}.$$

Sada će se problem malo zakomplicirati. Pretpostavimo da je od N birača koji su izašli na izbore, njih a glasalo za kandidata A, a $b = N - a$ za kandidata B. Na koliko načina se može provesti anketa među n birača tako da njih k se izjasni da su glasali za kandidata A? Ovo je isto jednostavni kombinatorni problem. Prvo od a glasača koji su glasali za A izabiremo njih k, a $n-k$ biramo među b glasača koji su se izjasnili za kandidata B. Teorem o uzastopnom prebrojavanju nam daje

$$\binom{a}{k}\binom{b}{n-k}. \tag{1}$$

5 Vjerojatnosni model

Označimo sa X broj birača koji je glasao za kandidata A u anketi u kojoj je anketirano slučajno odabranih n ljudi. (Vrijednost od X ovisi o slučajnom odabiru anketiranih. Ovakvu slučajnu veličinu zovemo slučajna varijabla u vjerojatnosti.) Kolika je vjerojatnost da je $X = k$, tj. da se u anketi među n ljudi njih k izjasnilo za kandidata A? (k je neki fiksan prirodan broj.) Iz prethodnih kombinatornih argumenata, budući je odabir anketiranih slučajan, svaki podskup od n ljudi s jednakom vjerojatnošću može biti anketiran, a broj onih podskupova za koje će se k-članova izjasniti za kandidata A dan je s (1). Zato je

$$\mathbb{P}(X = k) = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}, \quad k \in \{0, 1, 2, \ldots, n\}. \tag{2}$$

Kažemo da X ima **hipergeometrijsku razdiobu**. Dobivena formula možda izgleda jednostavna, ali u praksi je ona teško izračunljiva. Razlog leži u činjenici da su ovi binomni koeficijenti iznimno veliki brojevi s kojima se teško operira.[2]

6 Srednje vrijednosti

Koliko u *u prosjeku* očekujemo (kad bi se provodilo više anketa) očekujemo da će se anketiranih izjasniti za kandidata A? Odgovor na pitanje nam daje

[2] a i b su u našem primjeru veći od 700000, a n i k su najviše 2000. Tako će u primjeru koji promatramo $\binom{a+b}{n}$ imati više od 6000 znamenki. (Ovo je jako *gruba* donja procjena.)

matematičko očekivanje, koje je definirano kao

$$\mathbb{E}X = \sum_{k=0}^{n} k\mathbb{P}(X = k).$$

Koristeći svojstva binomnih koeficijenata, (vidi [2, str. 125]) dobiva se da je

$$\mathbb{E}X = n \cdot \frac{a}{a+b} = np.$$

Znači, kada provodimo anketu *u prosjeku* (tj. očekujemo da) će se za prvog kandidata izjasniti isti postotak anketiranih, kao što je postotak glasova koji je kandidat A dobio na izborima. Zbog toga kažemo da je X/n **nepristrani procjenitelj** za vrijednost p. Ovaj rezultat opravdava provođenje anketa na način koji smo opisali na početku.

Važno je znati i koliko će predviđanje dobiveno anketom odstupati od konačnih rezultata. Za tu svrhu koristimo **varijancu**, tj. srednje kvadratno odstupanje od očekivanja:

$$\mathbf{Var}\, X = \mathbb{E}[(X - \mathbb{E}X)^2] = \sum_{k=0}^{n} (k - \mathbb{E}X)^2 \mathbb{P}(X = k).$$

Dobiva se da je (vidi [2, str. 142])

$$\mathbf{Var}\, X = n \cdot \frac{a}{a+b} \cdot \frac{b}{a+b} \cdot \frac{a+b-n}{a+b-1}.$$

Ono što nas zapravo zanima je srednje kvadratno odstupanje vrijednosti X/n (tj. postotka anketiranih koji su se izjasnili za kandidata A) od očekivane vrijednosti. To je

$$\mathbf{Var}[X/n] = \mathbb{E}[(X/n - p)^2] = \frac{1}{n^2}\mathbb{E}[(X - \mathbb{E}X)^2] = \\ = \frac{1}{n} \cdot \frac{a}{a+b} \cdot \frac{b}{a+b} \cdot \frac{a+b-n}{a+b-1}. \quad (3)$$

Uočimo da prethodni rezultat potvrđuje neke intuitivno jasne pretpostavke o provođenju anketa.

- uočavamo da što više birača anketiramo, tj. što je n veći, da će očekivano odstupanje biti manje, tj. procjena vrijednosti p s X/n će biti *pouzdanija*;
- ako je $n = a + b$, tj. ako anketiramo sve birače izašle na izbore odstupanja od stvarnog rezultata neće biti;
- odstupanja od stvarnog rezultata neće biti i ako su svi birači glasali za istog kandidata, odnosno ako je drugi kandidat dobio 0 glasova (onda je $a = 0$ ili $b = 0$).

Više o varijanci i očekivanju čitatelj može naći u knjizi [3, 6. poglavlje].

7 Pouzdani interval

Znamo očekivanu vrijednost i očekivano kvadratno odstupanje od te vrijednosti. Možemo li dati neku procjenu koliku vrijednost bi postotak anketiranih koji su glasali za kandidata A mogao biti u odnosu na postotak birača koji su glasali za njega?

To nam omogućuje **Čebiševljeva nejednakost** (vidi [2, Sarapa, str. 144]) koja kaže da za slučajnu varijablu Y koja ima varijancu, za sve $\varepsilon > 0$ vrijedi

$$\mathbb{P}(|Y - \mathbb{E}Y| \geq \varepsilon) \leq \frac{\mathbf{Var}\, Y}{\varepsilon^2}.$$

Označimo standardnu devijaciju sa $\sigma_n := \sqrt{\mathbf{Var}[X/n]}$. Ako uvrstimo u prethodnu nejednakost $Y = X/n$ i $\varepsilon = 2\sigma_n$, dobivamo da je

$$\mathbb{P}(|X/n - p| \geq 2\sigma_n) \leq \frac{1}{4} \quad \Rightarrow \quad \mathbb{P}(|X/n - p| < 2\sigma_n) \geq \frac{3}{4}.$$

Posljednje nam kaže da je vjerojatnost da X/n bude u intervalu

$$\langle p - 2\sigma_n, p + 2\sigma_n \rangle$$

je barem 75%. (Ovo je jako gruba ocjena, ali za naše potrebe dovoljna.) Dakle, u više od 75% slučajeva procjena postotka glasova će odudarati od stvarnog postotka za $\pm 2\sigma_n$.

Koliki je σ_n u primjeru iz predsjedničkih izbora? Zapišimo malo drukčije σ_n:

$$\sigma_n = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{a+b-n}{a+b-1}}. \tag{4}$$

Pogledajmo koliki je σ_n u primjeru predsjedničkih izbora. Uvrštavanjem za a i b konkretnih brojeva u gornjoj jednakosti dobivamo

```
> sg_n=sqrt((1-p)*p *(a+b-n)/n/(a+b-1))
> sg_n
[1] 0.01078640
```

Dakle u preko 75% slučajeva odstupanje procjene anketa će biti približno $\pm 2.1\%$ (jer gornji broj množimo sa 2).

Prebrojimo broj simuliranih anketa u kojem je procjena ankete bila unutar intervala $\langle p - 2\sigma_n, p + 2\sigma_n \rangle$. **R** će nam to lako napraviti.

```
> unu=(ankete<p+2*sg_n)&(ankete>p-2*sg_n)
> length(unu[unu])
[1] 955
```

Prvi red bilježi koje procjene su u intervalu, a drugi ih broji. Dakle, u čak 95.5% slučajeva je predviđanje ankete bilo unutar ovog intervala. Razlog zašto je ovaj broj daleko veći od 75% leži u činjenici da se radilo o vrlo gruboj ocjeni pouzdanog intervala, čija je pouzdanost daleko veća. (To je posljedica asimptotske normalnosti hipergeometrijske razdiobe. To ćemo kratko spomenuti na kraju.)

8 Statistički problem

Sada smo razvili vjerojatnosni model. Statistika se s druge strane bavi pitanjem kako na temelju opaženih mjerenja procijeniti parametre nekog vjerojatnosnog modela. U našem slučaju želimo procijeniti p. Stoga se vraćamo se praktičnom problemu procjene rezultata. U samoj izbornoj noći mnogo toga neće biti poznato.

Primjerice, brojeve a i b nećemo uopće znati, eventualno ćemo znati od čega je zbroj $a+b$ veći ili jednak. (Izborno povjerenstvo obično objavi koliko je ljudi izašlo na izbore do nekog trenutka, više puta u toku dana.)

Kako riješiti problem nepoznavanja brojeva a i b? Treba nam malo praktičnog razmišljanja. Obično će biti $a + b \gg n$ (broj anketiranih birača je bitno manji od broja izašlih birača), te će biti

$$\frac{a+b-n}{a+b-1} \approx 1.$$

Nadalje, kako je ovaj broj iz $\langle 0, 1\rangle$, funkcija korijen ($\sqrt{}$) će ga preslikati još bliže broju 1. Tako u primjeru predsjedničkih izbora broj $\sqrt{(a+b-n)/(a+b-1)}$ ima vrijednost

```
> sqrt((a+b-n)/(a+b-1))
[1] 0.999526
```

Iz iznesenog vidimo da drugi dio umnoška u jednakosti (4) možemo zanemariti, tj. smatrati da je jednak 1. Stoga uvodimo

$$\sigma'_n = \sqrt{\frac{p(1-p)}{n}}. \tag{5}$$

Kako je $\sigma_n \leq \sigma'_n$, vrijedi

$$\langle p - 2\sigma_n, p + 2\sigma_n\rangle \subseteq \langle p - 2\sigma'_n, p + 2\sigma'_n\rangle,$$

pa će procjene anketa s još većom vjerojatnošću biti u intervalu $\langle p-2\sigma'_n, p+2\sigma'_n\rangle$.

Kada smo se riješili potrebe da procjenjujemo parametre a i b, preostaje samo procijeniti koliki može biti p s dovoljno velikom vjerojatnošću. Kako

na temelju rezultata ankete procijeniti koliki je p? (Njega također ne znamo, do obajve konačnih rezultat, a cilj ankete je upravo procijeniti njega.) Nejednakost

$$\left|\frac{X}{n} - p\right| \leq 2\sigma_n'$$

vrijedi s vjerojatnošću od bar 75%. Kvadriranjem prethodne nejednakosti i uvrštavanjem jednakosti (5) dobivamo da nejednakost

$$\left(p - \frac{X}{n}\right)^2 \leq 4\frac{p(1-p)}{n},$$

odnosno nakon sređivanja,

$$p^2\left(1 + \frac{4}{n}\right) + p\left(-2\frac{X}{n} - \frac{4}{n}\right) + \left(\frac{X}{n}\right)^2 \leq 0 \tag{6}$$

vrijedi s vjerojatnošću od bar 75%. Zadnje je po p kvadratna nejednadžba, čijim rješavanjem dobivamo da je

$$p \in \left[\frac{\frac{X}{n} + \frac{2}{n} - 2\sqrt{\frac{(X/n)(1-X/n)+1/n}{n}}}{1+\frac{4}{n}}, \frac{\frac{X}{n} + \frac{2}{n} + 2\sqrt{\frac{(X/n)(1-X/n)+1/n}{n}}}{1+\frac{4}{n}}\right]$$

u bar 75% slučajeva. Dakle, kada je k glasača od njih n u anketi se izjasnilo za kandidata A onda se zamjenom X s k u gornjem intervalu dobiva interval

$$\left[\frac{\frac{k}{n} + \frac{2}{n} - 2\sqrt{\frac{(k/n)(1-k/n)+1/n}{n}}}{1+\frac{4}{n}}, \frac{\frac{k}{n} + \frac{2}{n} + 2\sqrt{\frac{(k/n)(1-k/n)+1/n}{n}}}{1+\frac{4}{n}}\right]$$

koji zovemo *procjena* bar 75% pouzdanog intervala za p. Ovo možemo zapisati malo preglednije pa dobijemo da je procjena pouzdanog intervala

$$\left[\frac{k+2-2\sqrt{\frac{k(n-k)+n}{n}}}{n+4}, \frac{k+2+2\sqrt{\frac{k(n-k)+n}{n}}}{n+4}\right]. \tag{7}$$

U **R**-u lako računamo možemo gornji interval. Pogledajmo primjer prve simulirane ankete (zapisane u vektoru `anketa`). Anketirali smo $n = 2000$ osoba, a od toga je

```
> k=sum(anketa)
> k
[1] 1254
```

se izjasnilo za 1. kandidata. (Dakle, $k = 1254$.) Dobivamo da je procjena pouzdanog intervala za p

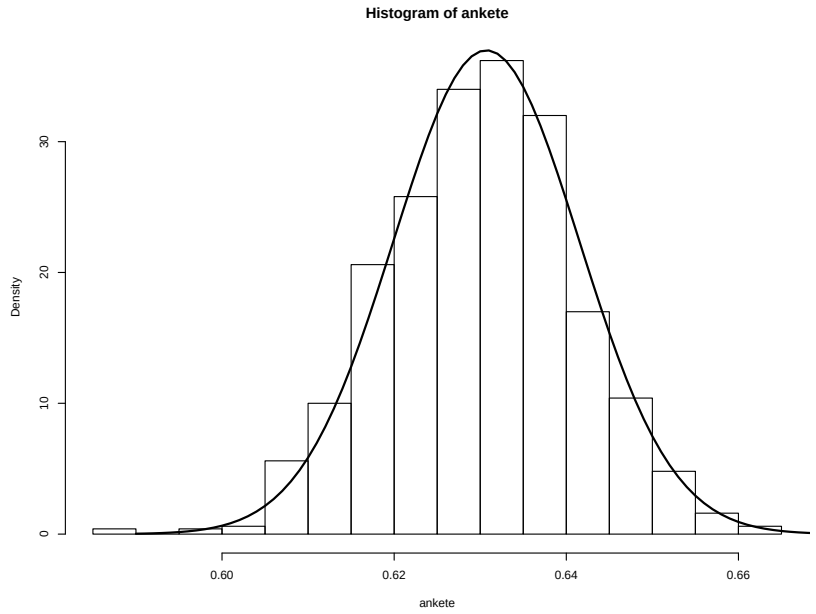
```
> d=2*sqrt((k*(n-k)+n)/n);
> t=k+2;
> c(t-d,t+d)/(n+4);
[1] 0.6051393 0.6483537
```

Kako je $p \approx 0.63$, vidimo da je (u ovoj anketi) dobro procijenjen p (jer pripada tom intervalu).

Provjerom na računalu pokazuje se da će p biti u 95.5% procjena pouzdanih intervala od anketa simuliranih i zapisanih u vektoru `ankete`. Tako da vidimo da će procjena p intervalom (7), biti točna u više od 75% slučajeva.

9 Napomena o asimptotskoj normalnosti

Na slici 1. vidimo da se podaci grupiraju oko sredine i da formiraju brijeg koji nalikuje normalnoj razdiobi. Ako nacrtamo graf (funkcije gustoće) razdiobe $N(p, \sigma_n^2)$ i normirani histogram (čiji stupci imaju ukupno površinu 1) vidimo da podaci zaista približno prate ovu distribuciju. (Vidi sliku 2.)



SLIKA 2. *Normirani histogram predviđanja anketa i graf normalne razdiobe*

Uz određene uvjete za velike brojeve X/n ima približnu distribuciju $N(p, \sigma_n^2)$. (Sam iskaz, a pogotovo dokaz ove činjenice nije baš jednostavan, te zato ne ulazimo u dublje u to.) U tom će slučaju[3] p u intervalu (7)

[3] Naime ako je $X \sim N(\mu, \sigma^2)$, onda je $\mathbb{P}(X \in [\mu - 2\sigma, \mu + 2\sigma]) \approx 0.9545$ (vidi [3, str. 56]). Zato nejednakost (6) vrijedi s vjerojatnošću od približno 95.4%.

biti u približno 95.4% slučajeva. Pa vidimo da je će taj broj biti daleko veći od 75%, odnosno približno onakav kakav smo dobili u simulacijama.

Zaključak

U ovom članku smo prošli način procjene postotka (udjela) glasova koje je dobio pojedini kandidat. Na izbore je izašlo $a + b$ glasača, od kojih je a glasovalo za kandidata A, a b za kandidata B. Cilj ankete u kojoj je anketirano n glasača od kojih se k izjasnilo za kandidata A, dati ocjenu $p = \frac{a}{a+b}$, dok još ne znamo brojeve a, b i $a + b$.

Ako se za n slučajno odabranih (anketiranih) birača s X označimo broj nih koji su glasali za kandidata A, onda X/n ima očekivanu vrijednost p i da će X/n pripadati intervalu

$$\left[p - 2\sqrt{\frac{p(1-p)}{n}}, p + 2\sqrt{\frac{p(1-p)}{n}}\right],$$

s vjerojatnošću od bar 75%.

Zbog prethodno navedenog, kada je $X = k$ s velikom pouzdanošću možemo tvrditi da će se p nalaziti u intervalu

$$\left[\frac{k + 2 - 2\sqrt{\frac{k(n-k)+n}{n}}}{n+4}, \frac{k + 2 + 2\sqrt{\frac{k(n-k)+n}{n}}}{n+4}\right].$$

Mi smo promatrali ovaj problem samo za pitanje izbora, ali rezultati dobiveni ovdje mogu se primijeniti na razne druge probleme u kojima ispitujemo zastupljenost nečega u populaciji na temelju uzorka.

Poseban problem kod provođenja anketa je kako uzeti dobar uzorak. Nama je uzorak napravilo računalo, dok terensko uzimanje uzorka mora se na poseban način. Zato se uzorak obično uzima da budu anketirani što različitiji ispitanici.

Nadam se da se čitatelj imao priliku uvjeriti da provođenje anketa ima smisla, te da ih opravdava ne baš tako jednostavna matematika.

Literatura

[1] Pauše Ž., *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[2] Sarapa N., *Teorija vjerojatnosti*, Školska knjiga, 2002.

[3] Sarapa N., *Vjerojatnost i statistika 2. dio: Osnove statistike - slučajne varijable*, Školska knjiga, 1996.

[4] Venables W.N., Smith, D.M., R Development Core Team, *An Introduction to R : Notes on R: A Programming Enviornment for Data Analysis and Graphics*, 2008.,
`http://cran.r-project.org/doc/manuals/R-intro.pdf`

[5] *Državno izborno povjerenstvo RH*, `http://www.izbori.hr`