Matematika iza anketa - primjer izbora

Tvrtko Tadić[1]

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

Na neka pitanja **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 (a to se za ovakve podatke ne radi). S druge strane, podatak o postotku stanovništva sa završenim fakultetom dobijemo svakih 10 godina popisom stanovništva.

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

U tu svrhu obično se ispituje manji dio birača izašlih na izbore, a na temelju njega daje se procjena kako bi rezultat mogao 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 **razini cijele populacije**.

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

Odgovore na ova pitanja dat 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.

[1] Tvrtko Tadić, PMF - Matematički odjel, Zagreb

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, održnog 2010. Rezultati glasova birača izašlih na izbore u Republici Hrvatskoj[2] dani su u tablici.

KANDIDAT	BROJ GLASOVA
Ivo Josipović	1 330 339
Milan Bandić	778 915

Kako bismo lakše baratali podatcima, 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 provest ćemo u statističkom programu **R**.

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

U vektoru `izbori` na prvih 778 915 mjesta nalazi se brojka 0, a na preostalih 1 330 339 brojka 1. Kako bismo vidjeli koliko je glasova u postotcima dobio Ivo Josipović, dovoljno je izračunati aritmetičku sredinu vektora `izbori`.

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

Dakle, Ivo Josipović na području RH dobio je 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 složmo u vektor `anketa`.

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

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

2. Provedimo više anketa

Jedna je anketa, dakle, bila uspješna. Hoće li baš svaka biti uspješna? Očito je moguće da anketa dade 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`.

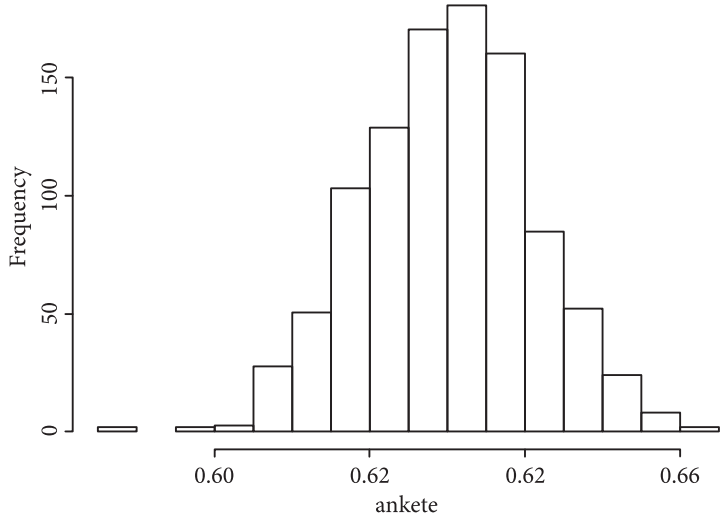
```
> ankete=rep(0,1000)
> for(j in 1:1000) ankete[j]=mean(sample(izbori,2000))
```

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 su rasponu 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 predviđala mu je 66.45% glasova, a anketa koja je predviđala najmanji postotak predviđala mu je 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.



Slika 1. Histogram predviđanja 1000 anketa

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

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 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 provedene ankete 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 se načina može provesti anketa među n birača tako da se njih k izjasni da su glasali za kandidata A? Ovo je također 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 daje nam:

$$\binom{a}{k}\binom{b}{n-k} \qquad (1).$$

5. Vjerojatnosni model

Označimo s 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 u vjerojatnosti zovemo *slučajna varijabla*.) 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 da 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

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

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.[3]

6. Srednje vrijednosti

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

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

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

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

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.

[3] Brojevi a i b su u našem primjeru veći od 700 000, 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).

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

$$\mathrm{Var}X = \mathrm{E}\left[(X - \mathrm{E}X)^2\right] = \sum_{k=0}^{n} (k - \mathrm{E}X)^2 \mathrm{P}(X = k).$$

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

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

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

$$\mathrm{Var}\left[X/n\right] = \mathrm{E}\left[(X/n - p)^2\right] = \frac{1}{n^2}\mathrm{E}\left[(X - \mathrm{E}X)^2\right] = \frac{1}{n} \cdot \frac{a}{a+b} \cdot \frac{b}{a+b} \cdot \frac{a+b-n}{a+b-1}. \qquad (3)$$

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

- uočavamo da će, što više birača anketiramo, tj. što je n veći, 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 ni 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 bi vrijednost 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], str. 144) koja kaže da za slučajnu varijablu Y koja ima varijancu za sve $\varepsilon > 0$ vrijedi

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

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

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

Posljednje nam kaže da je vjerojatnost da X/n bude u intervalu $\langle p - 2\sigma_n, p + 2\sigma_n \rangle$ jednaka barem 75% (ovo je jako gruba ocjena, ali za naše potrebe dovoljna). Dakle, u više od 75% slučajeva procjena postotka glasova odudarat će od stvarnog postotka za $\pm 2\sigma_n$.

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

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

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 bit će približno $\pm 2.1\%$ (jer gornji broj množimo s 2).

Prebrojimo broj simuliranih anketa gdje 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 su procjene 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 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 čemu je jednak. (Izborno povjerenstvo obično objavi koliko je ljudi izašlo na izbore do nekog trenutka, više puta tijekom 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), pa 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 $\sqrt{\frac{a+b-n}{a+b-1}}$ 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 objave konačnih rezultata, 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

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

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 se k glasača od njih n u anketi 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 gornji interval. Pogledajmo primjer prve simulirane ankete (zapisane u vektoru `anketa`). Anketirali smo $n = 2000$ osoba, a od toga se

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

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

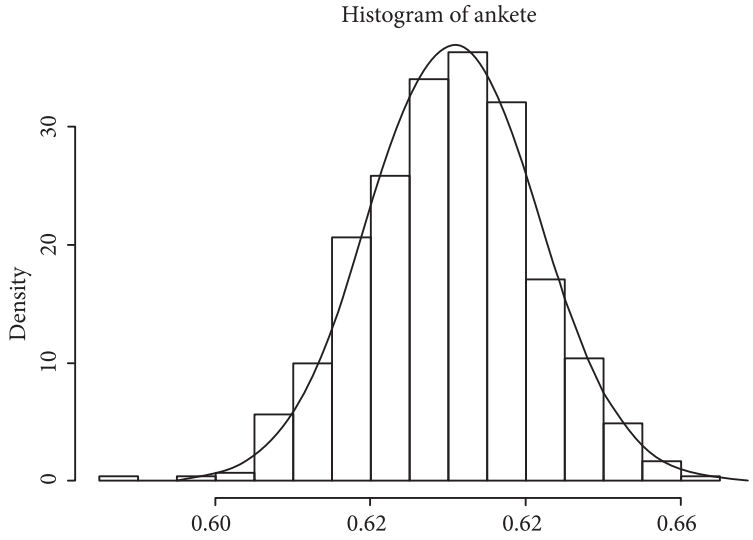
```
> d=2*sqrt((k*(n-k)+n)/n);
> t=k+2;
> c(t-d,t+d)/(n+4);
[1] 0.6051393 0.6483537
```

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 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 podatci grupiraju oko sredine i da formiraju brijeg koji nalikuje normalnoj razdiobi. Ako nacrtamo graf (funkcije gustoće) razdiobe $N\left(p,\sigma_n^2\right)$ i normirani histogram (čiji stupci imaju ukupno površinu 1), vidimo da podatci 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 $N\left(p,\sigma_n^2\right)$. (Sam iskaz, a pogotovo dokaz ove činjenice, nije baš jednostavan, pa zato ne ulazimo dublje u to.) U tom će slučaju[4] p u intervalu (7) biti u približno 95.4% slučajeva. Vidimo da je će taj broj biti daleko veći od 75%, odnosno približno onakav kakav smo dobili u simulacijama.

10. Zaključak

U ovom smo članku razmotrili 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, jest dati ocjenu $p = \frac{a}{a+b}$ dok još ne znamo brojeve a, b i $a + b$.

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

Ako za n slučajno odabranih (anketiranih) birača s X označimo broj onih koji su glasali za kandidata A, onda X/n ima očekivanu vrijednost p pa 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

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

Ovaj smo problem promatrali 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 se terensko uzimanje uzorka mora obaviti na poseban način. Zato se uzorak obično uzima tako 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, Zagreb, 2002.
3. Sarapa N., *Vjerojatnost i statistika 2. dio: Osnove statistike - slučajne varijable*, Školska knjiga, Zagreb, 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