SIMULACIJA ANKETE

Tvrtko Tadić, Zagreb

Matkači su sigurno zamijetili da mediji često objavljuju razne ankete. To su obično ispitivanja koje provodimo na manjem broju ispitanika (tzv. **uzorku**) kada želimo doznati nešto više o ciljanoj skupini ljudi (tzv. **populaciji**). Vjerojatno su se mnogi zapitali: *Možemo li zbilja na temelju manje skupine ispitanih nešto zaključiti o cijeloj populaciji?*

Jedan od najboljih pokazatelja su tzv. *izlazne ankete* koje se provode na dan izbora. Pokazuje se da one relativno točno procjenjuju konačne rezultate izbora.

U ovom ćemo članku, simulirajući ankete, pokušati ilustrirati smisao njihovog provođenja, te natuknuti koja matematika opravdava[1] zaključke koje donosimo na temelju njih.

Izlazne ankete

Izlazne ankete provode se na sam dan izbora. Anketari na biračkim mjestima ispituju birače za koga su glasali. Nakon zatvaranja birališta Državno izborno povjerenstvo kroz nekoliko će sati objaviti točne rezultate, a cilj naručitelja anketa (obično televizijskih kuća) jest da u što kraćem roku daju procjenu rezultata izbora.

Pokazuje se da te ankete uglavnom daju dobru procjenu stvarnih rezultata. Pretpostavljamo sljedeće:

- na izborima sudjeluju dva kandidata (kandidati *A* i *B*);
- svaki birač koji je izašao na izbore glasao je za jednoga od njih;
- anketirani birači u anketi daju točne odgovore o tome za koga su glasali.

Simulacija ankete

Napravimo simulaciju ankete na 2. krugu izbora za predsjednika Republike Hrvatske, održanom 2010. godine.

Kandidat Ivo Josipović dobio je 1 330 339 glasova, a Milan Bandić 778 915 glasova birača izašlih na izbore u Republici Hrvatskoj[2]. (Simuliramo anketu koja se provodi po Hrvatskoj[3].)

[1] Više o problemu anketa zainteresirani čitatelj može naći u članku *Matematika iza anketa - primjer izbora* u jednom od sljedećih brojeva Poučka.

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

[3] Birači koji glasuju izvan Hrvatske glasuju širom svijeta, pa je anketu izvan RH iz praktičnih razloga nemoguće provesti. Nevažeće glasove zanemarujemo za potrebe primjera.

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.) Simulaciju ankete provest ćemo u statističkom programu R.[4]

```
> 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 dobio Ivo Josipović u postotcima, 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`, a vrijednosti na tim mjestima složimo 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.% 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**.

Simulirajmo više anketa

Dakle, jedna je anketa bila uspješna. No, hoće li baš svaka biti uspješna? Očito je moguće da anketa da krivu procjenu. Recimo, ako od 2000 anketiranih svi budu glasači istog kandidata, dobit ćemo krivu procjenu. Koliko je to vjerojatno? Kako znamo da nismo imali sreće pa nam se baš zalomila ovakva procjena rezultata? Zahvaljujući računalima, ankete možemo ponavljati proizvoljno mnogo puta.

[4] Vodeći besplatni statistički program koji koriste brojne svjetske kompanije (*Google*, *Pfizer*, *Merck*, *Bank of America*, *InterContinental Hotels Group*, *Shell*, ….

Provest ćemo 1000 anketa da vidimo kakvima će se pokazati predviđanja rezultata za kandidata kodiranog brojem 1. Predviđanja rezultata spremit ćemo u vektor `ankete`.

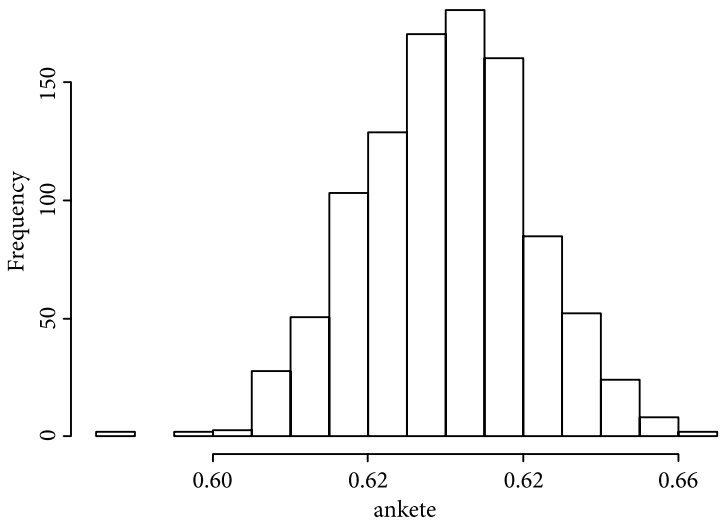
```
> ankete=rep(0,1000)
> for(j in 1:1000) ankete[j]=mean(sample(izbori,2000))
```

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 kandidata kodiranog brojem 1.

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

Dakle, anketa koja je predviđala najveći postotak za kandidata kodiranog 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 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.)

Teorijsko opravdanje

Kao što smo već naveli, 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. Sljedeća tablica ukratko opisuje što koja oznaka znači.

KANDIDAT	BROJ GLASOVA
A	a
B	b
UKUPNO	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.

Zahvaljujući teorijskim rezultatima iz vjerojatnosti, pokazuje se da će u više od 75% slučajeva udio onih koji su se u anketi (u kojoj je anketirano n birača) izjasnili za kandidata A biti u intervalu

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

U slučaju predsjedničkih izbora imamo:

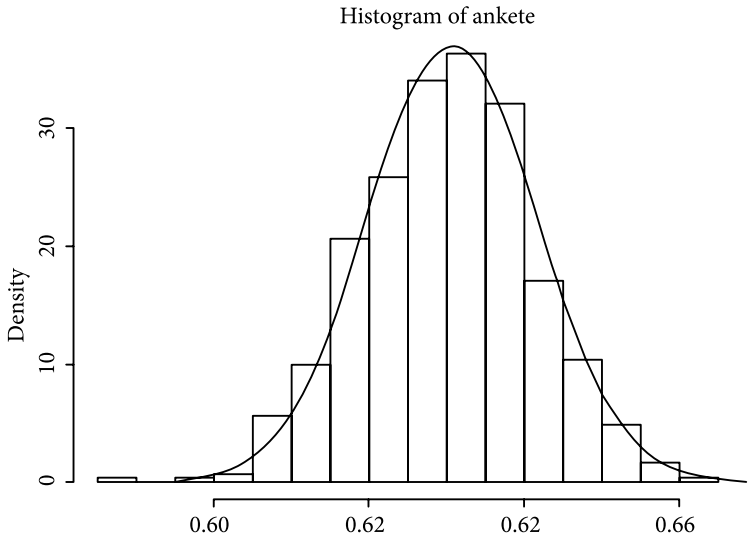
```
> n=2000
> p=mean(izbori)
> d=2*sqrt(p*(1-p)/n)
> c(p-d,p+d)
[1] 0.6091324 0.6522984
```

Dakle, u više od 75% slučajeva u anketama će kandidat kodiran s 1 dobivati od 60% do 65% glasova (nije loše). Provjerimo ovaj rezultat na tisuću simuliranih anketa zapisanih u vektoru `ankete`.

```
> log=(ankete<=p+d)&(ankete>=p-d)
> length(ankete[log])
[1] 955
```

Prva naredba ispituje koje su ankete predviđale rezultate unutar intervala, a druga naredba ispituje koliko je takvih anketa bilo.

Vidimo da je od 1000 anketa njih 955 predviđalo rezultate unutar navedenoga intervala. Uočimo kako je njih 95.5% pogodilo unutar ovog intervala, što je puno više od 75%. Razlog tome je što se zbog velikih brojeva koji se pojavljuju ovi podatci počinju ponašati *normalno*, tj. histogrami prate tzv. **Gaussovu krivulju** (krivulju normalne razdiobe). Ako sve stupce histograma proporcionalno smanjimo tako da im je ukupna površina jednaka 1 i nacrtamo Gaussovu krivulju s pripadnim parametrima, vidjet ćemo da se podatci zaista ponašaju kao da prate ovu krivulju.



Slika 2. Normirani histogram predviđanja anketa i graf normalne razdiobe

Zbog te (približne) normalnosti podataka, teorija nam kaže da bi ankete trebale navedeni interval pogađati u približno 95.4% slučajeva. To u ovom slučaju potvrđuju i simulacije.

Statistički problem

Pod pretpostavkom da znamo p, znamo i interval u kojemu će ankete s velikom vjerojatnošću predviđati udio glasova kandidata A.

No, kod ankete je glavni problem što mi **ne znamo** parametar p (kao ni a, ni b). ***Mi taj parametar na temelju rezultata ankete želimo procijeniti!***

Upravo se na temelju prethodno navedenog intervala dolazi do procjene pouzdanog intervala za parametar p. Ako se među n anketiranih s k izjasnilo da su glasali za kandidata A, u približno 95.4% slučajeva p će se nalaziti u intervalu

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

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

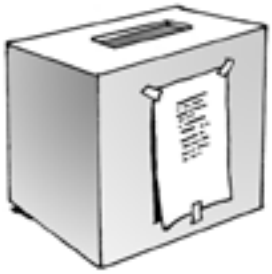
```
> k=sum(anketa)
> k
[1] 1254
```

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).

Završni osvrt



Kod provođenja anketa najvažnije je dobro uzeti uzorak. Obično se, kako bi se simulirala slučajnost, anketari trude ispitati što različitije tipove osoba, s ciljem da sve skupine društva budu zastupljene. Nama je računalo riješilo taj problem.

Ovdje izneseni podatci o anketama mogu se, naravno, primijeniti i na druge tipove anketa u kojima nas zanima koliko je nešto zastupljeno u nekoj populaciji.

Simulacije su izvrstan pokazatelj kojim smjerom trebamo istraživati i pokušati teorijski opravdati ono što tvrdimo (slično kao crteži u geometriji). U mnogim praktičnim pitanjima teorijski modeli mogu biti prekomplicirani da bi se došlo do korisnih rezultata, stoga smo u tim slučajevima primorani pouzdati se u informacije koje dobijemo simulacijama tih modela.

Ovdje se i uz pomoć teorijskih rezultata možemo uvjeriti da ima smisla provoditi ankete (na opisani način) i na temelju njih donositi zaključke o cijeloj populaciji.