Pričakovana vrednost, varianca

Nataša Kejžar

Povzetek

Spoznali boste kako:

- generiramo vzorce iz končnih in neskončnih populacij
- računamo kvantile v populaciji in na vzorcu
- narišemo gostoto, verjetnostno funkcijo, porazdelitveno funkcijo
- s simulacijami preverjamo pristranskost cenilke: ugotovili smo, da je povprečna ocena natančnejša, če naredimo večje število simulacij

Nekaj osnovnih verjetnostnih izrazov, ki se v statistiki pogosto uporabljajo

Pričakovana vrednost

- izračun po definiciji

$$E(X) = \sum_{i=1}^{\infty} x_i p(x_i)$$
$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- lastnosti

$$E(aX) = a \cdot E(X)$$
$$E(X + Y) = E(X) + E(Y)$$

- Jensenova neenakost: za konkavno funkcijo f velja (za konveksno pa ravno obratno)
- https://www.statlect.com/fundamentals-of-probability/Jensen-inequality

$$f(E(X)) \geq E(f(X))$$

Varianca

- po definiciji

$$var(X) = E\left[(X - E(X))^2\right]$$

- lastnosti

$$
\begin{aligned}
var(X) &= E(X^2) - E(X)^2 \\
var(aX) &= a^2 \cdot var(X) \\
var(X+Y) &= var(X) + var(Y) + 2 \cdot cov(X,Y) \\
var(X) &= E_Y(var(X|Y)) + var_Y(E(X|Y))
\end{aligned}
$$

- izračun variance za diskretne spremenljivke (za izračun $E(X)$ glej zgoraj)

$$var(X) = \sum_{i=1}^{\infty} x_i^2 p(x_i) - E(X)^2$$

Kovarianca

- po definiciji

$$cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

- lastnosti

$$cov(X, Y+Z) = cov(Y+Z, X) = cov(X,Y) + cov(X,Z)$$

Naloge

1. Pomoč za porazdelitve v R: `help(Distributions)`. Za vzorec velikosti 1000 iz porazdelitve $N(120, 30^2)$ narišite:

   a. verjetnostno porazdelitev (`hist`, parameter `probability`, `breaks`, ali s knjižnico `ggplot2`)

      - Kaj je gostota porazdelitve? Kaj dobimo, če narišemo rezultate funkcije `density`?
      - Na histogram dodajte še pričakovano vrednost za to spremenljivko (`abline`, parameter `v`; ali `geom_vline`)
      - Na histogram dodajte še gostoto normalne porazdelitve (uporabite funkciji `dnorm` in `curve`, parameter `add=TRUE`; ali `stat_function`)

   b. porazdelitveno funkcijo (`ecdf`) Ponovite isto za vzorec velikosti 10. Komentirajte opažanja.

2. *Povprečna porodna teža novorojenčka v Sloveniji je 3.300 gramov, večina donošenih novorojenčkov ob rojstvu tehta med 2.500 in 4.100 grami.*

   a. Če velja, da je teža novorojenčkov v populaciji normalno porazdeljena in da sta spodnja in zgornja meja 5. in 95. percentil, z R izračunajte, kolikšen je standardni odklon teže novorojenčkov.
   b. Izračunajte interkvartilni razmik za težo.
   c. V katerem percentilu se nahaja novorojenček, ki je težak 2900 g?
   d. Iz teoretične porazdelitve izberite vzorec 5, 50, 500 enot in na vzorcu izračunajte percentil za novorojenčka s 2900 grami. Primerjajte in komentirajte rezultata.
   e. Za izbran vzorec s 500 enotami izračunajte 90% **interval zaupanja za povprečje**. Primerjajte ga z intervalom, ki so ga poročali na začetku naloge. Komentirajte.

3. Teoretične nalogice:

   a. pokažite s pomočjo definicije za varianco, da drži $var(X) = E(X^2) - E(X)^2$
   b. pokažite s pomočjo vsote varianc, da drži $var(2X) = 4 \cdot var(X)$
   c. pokažite, da varianco končne populacije (velikosti N) lahko zapišemo tudi kot

$$var(X) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2.$$

4. Verjetnostna porazdelitev končne populacije, ki vas zanima, je predstavljena v spodnji tabeli pod nalogo. V njej je 1500 enot.

   a. Izračunajte pričakovano vrednost populacije v R.

      - po definiciji
      - tako da zapišete populacijo v vektor (funkcija `rep`)

   b. Izračunajte varianco populacije v R. Zakaj izračun s funkcijo `var` da drugačen rezultat?

   c. Iz populacije bi radi izbrali vzorec velikosti 15 s ponavljanjem. Zapišite kodo v R. (`sample`)

   d. Iz populacije bi radi izbrali vzorec velikosti 15 brez ponavljanja. Zapišite kodo v R.

X	1	2	3	4	5
p(X)	1/15	1/5	4/15	2/5	1/15

5. Naj bo spremenljivka X porazdeljena po $Bernoulli(\pi = 0.85)$.

   a. Generirajte 100 opazovanj. (uporabite npr. `runif` ali `sample`)
   b. Izračunajte po definiciji $E(X)$ in $var(X)$.
   c. Vemo, da je vsota neodvisnih, enako porazdeljenih Bernoullijevih spremenljivk porazdeljena po

binomski porazdelitvi ($Y = \sum_{i=1}^{n} X_i \sim Bin(n, \pi)$). Pokažite to s simulacijo velikega vzorca.

d. Izpeljite, da velja $var(Y) = n \cdot \pi(1 - \pi)$.

6. Radi bi primerjali povprečni teži dveh velikih skupin morskih prašičkov. Prva skupina so tisti, ki so hišni ljubljenčki v Evropi, druga tisti iz ZDA. V ta namen zberemo enako velika vzorca.

   a. Kateri statistični test bi uporabili? Kaj lahko poveste o kovarianci teh dveh skupin?

   b. Simulirajte populacijsko kovarianco v R (funkcija `cov`) za naslednji spremenljivki (predpostavite lahko, da sta teži v populaciji pribl. normalno porazdeljeni):

      - X_{EU} in X_{ZDA} (koliko je populacijska kovarianca?)
      - X_{EU} in $X_{EU} - X_{ZDA}$ (kaj pričakujete?)

   c. Izračunajte kovarianco teoretično.

7. Imate normalno porazdeljeno spremenljivko ($X \sim N(0, 1)$). Grafično preverite rezultat iz Rice-a (6. poglavje), da je $X^2 \sim \chi^2_{df=1}$.

   a. izberite dovolj velik vzorec iz $N(0, 1)$.
   b. narišite *gostoto* kvadriranih vrednosti
   c. dodajte gostoto porazdelitve χ^2_1

8. Imate dve neodvisni spremenljivki X_1 in X_2, ki sta porazdeljeni po χ^2 porazdelitvi z df_1 in df_2 stopinjami prostosti.

   a. Grafično preverite, da je $X_1 + X_2 = X_3$, kjer je $X_3 \sim \chi^2_{df_1+df_2}$.

   b. Naj bosta sedaj $X_1 \sim \chi^2_{df_1}$ in $X_3 \sim \chi^2_{df_1+df_2}$ neodvisni spremenljivki. Ali mislite, da velja tudi $X_3 - X_1 = X_2$, kjer je $X_2 \sim \chi^2_{df_2}$? Komentirajte zakaj.

      - Narišite porazdelitev spremenljivke $X_3 - X_1$.

   c. Kako mislite, da je porazdeljena razlika dveh neodvisnih standardno normalno porazdeljenih spremenljivk? Zakaj (oblika, variabilnost)? Preverite grafično.

9. Generirajte vrednosti za N = 10 enot veliko populacijo s kodo pod nalogo. Vse naslednje primere v R zakodirajte tako, da boste lahko kadarkoli spremenili vrednosti enot in tudi število enot.

   a. Izračunajte povprečje po formuli $\mu = 1/N \sum_{i=1}^{N} x_i$.
   b. Posebej seštejte pozitivne in negativne odmike od povprečja.
   c. Izračunajte varianco po obeh formulah, ki ste jih omenili na predavanjih.
   d. Izmislite si poljubno vrednost a in izračunajte vsoto kvadriranih odmikov (VKO) od te vrednosti. Za izračun te količine naredite funkcijo.
   e. Za različne vrednosti $a$jev (npr. med $[4, 10]$) izračunajte vsoto kvadriranih odmikov in dvojice narišite na graf. Z rdečo označite vrednost $N\sigma^2$ (pri kateri vrednosti a se pojavi?).

```
set.seed(1)
x = sample(x = 1:10, size=10,replace=TRUE)
```

10. Odgovorite:

    a. S kakšnim namenom za izračun variance kvadriramo odmike od povprečja?
    b. Kaj se dogaja z VKO pri različnih vrednostih a? Pri katerem a je v splošnem VKO najmanjši? Zakaj?
    c. Ponovite prejšnjo vajo z N = 20 enot, ki lahko zavzamejo vrednosti 0 ali 1. Kakšne vrednosti a boste vzeli za risanje grafa in zakaj?

11. Izračunajte delež vrednosti nad 132 za vzorec $n = 1000$, ki ga dobite iz porazdelitve $N(120, 30^2)$. Primerjajte dobljeni delež z verjetnostjo, ki jo izračunamo iz gostote (`pnorm`).

    a. Primerjajte rezultate, ki jih dobite iz vzorca iz 10 enot in vzorca iz 1000 enot. Kje so bistvene razlike in zakaj pride do njih?

12. Zapišite intuitivno cenilko za populacijsko povprečje.
    a. Pokažite teoretično, da je cenilka nepristranska.
    b. Pokažite to s simulacijami na velikosti vzorca 10 iz $N(120, 30^2)$:
        - generirajte 1000 vzorcev (uporabite for zanko ali `replicate`); za vsak vzorec izračunajte oceno za povprečje in te ocene shranite v nov vektor
        - če je cenilka nepristranska, morajo ocene variirati okrog populacijskega povprečja; preverite
    c. Narišite graf
    d. Izračunajte *pričakovano vrednost*