Osnove verjetnosti in statistika

Gašper Fijavž

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

Ljubljana, 7. maj 2010

Statistika

1. Opisna statistika.
2. Vzorčenje.
3. Ocenjevanje parametrov.
4. Testiranje hipotez.
5. Linearna regresija.
6. Morda še kaj.

Vzorčenje

Populacija . . . velika količina podatkov . . . N

Vzorec . . . majhna podmnožica podatkov . . . n

Slučajni vzorec . . . vsak vzorec iste moči ima isto verjetnost, da bo izbran

Slučajni vzorec s ponavljanjem . . . dovolimo, da se podatki v vzorcu ponavljajo

Cilj statističnih metod: na podlagi analize vzorca želimo sklepati na celotno populacijo.

Zgledi

- Volitve.
- Kontrola izdelkov.
- Študenti.
- Psevdonaključna števila.

Ponavljanje, da ali ne

N . . . velikost populacije
n . . . velikost vzorca

Trditev

Če je N velik v primerjavi z n, potem je število vzorcev brez ponavljanja približno enako kot število vzorcev s ponavljanjem.

Ponavljanje, načeloma da, v praksi ne.

Zgled: Ocene šestih študentov so $10, 7, 7, 8, 6, 10$. Vzorčimo brez ponavljanja, vzorci so velikosti 2. Koliko je vzorcev? Kaj lahko poveš o povprečni oceni vzorca?

Kvantili, mediana, kvartili, percentili

Naj X določa slučajno spremenljivko - vrednost parametra na populaciji.
Izberimo $q \in (0,1)$. Vrednosti a pravimo *q-kvantil* za X, če je

$$P(X \leq a) \geq q \text{ in } P(a \leq X) \geq 1 - q.$$

Mediana je 0,5-kvantil.
0,25, 0,5 in 0,75-kvantilom pravimo *kvartili*.
0,01,. . . , 0,99-kvantili so *percentili*.

Vzorčno povprečje

Naj bo Y spremenljivka na populaciji. Izberemo vzorec velikosti n, pridobimo vrednosti $Y_1, Y_2, \ldots, Y_n$.

Vzorčno povprečje je $\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n)$.

Izrek

Matematično upanje vzorčnega povprečja je enako povprečni vrednosti (matematičnemu upanju) na celotni populaciji:
$E(\overline{Y}) = E(Y)$.

Vzorčno povprečje

Izrek

Naj bo Y spremenljivka na populaciji velikosti N, $\mu = E(Y)$, $\sigma^2 = D(Y)$. Vzorčimo brez ponavljanja, vzorci velikosti n. Velja:

(a) $E(\overline{Y}) = E(Y) = \mu$

(b) $D(\overline{Y}) = D(Y) \cdot \frac{N-n}{n(N-1)} \overset{n\to\infty}{\longrightarrow} \frac{1}{n} D(Y) = \frac{1}{n}\sigma^2$
in $\sigma(\overline{Y}) \approx \frac{1}{\sqrt{n}}\sigma$

(c) *Če je Y na populaciji normalno porazdeljena z $N(\mu, \sigma)$, potem je $\overline{Y}$ na vzorcih tudi normalno porazdeljena z $N(\mu, \frac{\sigma}{\sqrt{n}})$.*

(d) *Četudi Y na populaciji ni normalno porazdeljena, je $\overline{Y}$ na populaciji približno normalno porazdeljena z $N(\mu, \frac{\sigma}{\sqrt{n}})$.*

Vzorčno povprečje

Naloga: Populacija šteje 3000 študentov (moških). Predpostavimo, da so njihove višine porazdeljene s popvrečjem 180 cm in standardno deviacijo 5 cm.

Naključno izberemo vzorec 25 študentov in izmerimo povprečno višino vzorca.

- Določi matematično upanje za povprečno višino vzorca.
- Določi standardno deviacijo za povprečno višino vzorca.
- Oceni verjetnost, da bo imel naključno izbrani vzorec povprečje višin
  - med 178 in 182 cm.
  - več kot 183 cm.
  - manj kot 175 cm.

Disperzija vzorca

Vzorec velikosti n z vrednostmi $X_1, \ldots, X_n$ in vzorčnim povprečjem $\overline{X}$. *Vzorčno disperzijo* s^2 in *popravljeno vzorčno disperzijo* $\hat{s}^2$ definiramo kot

$$s^2 = \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n}$$

$$\hat{s}^2 = \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$$

Izrek

$$E(s^2) = \frac{N}{N-1} \cdot \frac{n-1}{n} \cdot \sigma^2 \qquad in \qquad E(\hat{s}^2) = \frac{N}{N-1} \cdot \sigma^2$$

Zgled

Naloga: Vzorec ocen študentov vsebuje ocene $7, 9, 10, 10, 6, 10, 8, 8, 7, 7, 7, 7$. Izračunaj vzorčno povprečje, vzorčno disperzijo in popravljeno vzorčno disperzijo.

Porazdelitev vzorčnega povprečja

Naj bo μ matematično upanje na populaciji. Potem je:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \qquad\qquad T = \frac{\overline{X} - \mu}{\hat{s}/\sqrt{n}}$$

Če σ ne poznamo, jo nadomestimo z $\hat{s}$. Če je vzorec dovolj velik ($n \geq 30$), je tudi $T \dot{\sim} N(0,1)$.

Če je $n \leq 30$, potem je potrebno uporabiti Studentovo porazdelitev.

Zgled

Merimo življensko dobo žarnic; $N = 1.000.000$ in $n = 100$. Na vzorcu izmerimo

500	1000	1500	2000	2500	3000
10	5	5	65	10	5

.

Določi $\overline{X}$, s^2, $\hat{s}^2$. Določi T in oceni interval za μ.

Ocenjevanje parametrov

Naj bo λ parameter, odvisen od slučajne spremenljivke X.
I je *interval zaupanja* za vrednost parametra λ pri stopnji zaupanja γ, če velja naslednje:

- Verjetnost, da parameter λ pripada intervalu I je (ne glede na porazdelitev X) vsaj γ.
- Interval I je najmanjši možen.

Ocenjevanje μ

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

Torej je $\overline{X}$ na vzorcih je porazdeljen približno kot $N(\mu, \frac{\sigma}{\sqrt{n}})$.
Recept:

1. Izberi stopnjo zaupanja γ (tipično $0,9$, $0,95$, $0,99$ ali $0,999$).
2. Določi c za katerega velja $\Phi(c) = \gamma/2$.
3. Izračunaj $\overline{X}$ in $k = \dfrac{c \cdot \sigma}{\sqrt{n}}$.
4. Interval zaupanja je $I = [\overline{X} - k, \overline{X} + k]$.

Ocenjevanje μ

Kaj če σ ne poznamo?
Kaj če je vzorec majhen ($n \leq 30$)?

Porazdelitev χ^2

Naj bodo slučajne spremenljivke $X_1, X_2, \ldots, X_n$ *porazdeljene standardno normalno* in *neodvisne*.

$$\chi^2(n) := {X_1}^2 + {X_2}^2 + \ldots + {X_n}^2$$

Slučajna spremenljivka $\chi^2(n)$ *hi-kvadrat* z *n-prostostnimi stopnjami* je vsota n kvadratov neodvisnih standardno normalno porazdeljenih slučajnih spremenljivk.

$$P(\chi^2 \leq a) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^a x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx$$

če je $a \geq 0$.

Porazdelitev χ^2

Za izračun verjetnosti slučajne spremenljivke χ^2 uporabimo tabele. Fiksirajmo število prostostnih stopenj.

χ^2_p je vrednost, pri kateri je $P(\chi^2 \leq \chi^2_p) = p$.

Interval zaupanja za standardno deviacijo

Izrek

$$\frac{n \cdot s^2}{\sigma^2} = \frac{(n-1) \cdot \hat{s}^2}{\sigma^2} \sim \chi^2(n-1)$$

Radi bi določili interval $[c_1, c_2]$ za katerega velja, da je

$$P(c_1 \leq \sigma \leq c_2) \geq \gamma$$

kjer je γ *izbrana* stopnja zaupanja.

Interval zaupanja za standardno deviacijo

Iščemo interval, za katerega je

- $P(\chi^2 \leq \chi^2_{p_1})$ za $p_1 = \frac{1+\gamma}{2}$
- $P(\chi^2 \leq \chi^2_{p_2})$ za $p_2 = \frac{1-\gamma}{2}$

Potem je

$$P(\chi^2_{p_2} \leq \chi^2 \leq \chi^2_{p_1}) = \frac{1+\gamma}{2} - \frac{1-\gamma}{2} = \gamma$$

Žarnice, znova

$N = 1.000.000, n = 100, \hat{s} = 591,92, \hat{s}^2 = 350379$

Naloga: Določi interval zaupanja za σ pri stopnjah zaupanja $\gamma = 0,99$, $0,95$ in $0,90$.

Žarnice, znova

99 prostostnih stopenj:

1. χ^2 lahko izračunamo:
   $\chi^2_{0,005} = 66,5$, $\chi^2_{0,025} = 73,4$, $\chi^2_{0,05} = 77,0$,
   $\chi^2_{0,95} = 123,2$, $\chi^2_{0,975} = 128,4$, $\chi^2_{0,995} = 138,9$

2. χ^2 lahko ocenimo z linearno interpolacijo:
   $\chi^2_{0,005} = 66,5$, $\chi^2_{0,025} = 73,7$, $\chi^2_{0,05} = 77,0$,
   $\chi^2_{0,95} = 122,9$, $\chi^2_{0,975} = 128,8$, $\chi^2_{0,995} = 138,8$

3. χ^2 lahko ocenimo s pomočjo centralnega limitnega izreka:
   $\chi^2_{0,005} = 62,8$, $\chi^2_{0,025} = 71,4$, $\chi^2_{0,05} = 75,9$,
   $\chi^2_{0,95} = 122,1$, $\chi^2_{0,975} = 126,6$, $\chi^2_{0,995} = 135,2$

Žarnice, primerjava

Intervali zaupanja za σ, glede na različne metode izračuna $\chi^2(99)$.

stopnja zaupanja γ	1.	2.	3.
0.99	[500,722]	[500,722]	[506,743]
0.95	[520,678]	[519,687]	[523,697]
0.90	[531,671]	[531,671]	[533,676]

Statistične hipoteze

Statistična hipoteza je hipoteza o porazdelitvi slučajne spremenljivke.

Statistični test je bodisi

- *parametričen* . . . znan je tip porazdelitve, hipoteza govori o parametru.
- *neparametričen* . . . porazdelitev je neznana, hipoteza govori o vrsti porazdelitve. Parametrični testi so *enostranski* in *dvostranski*.

Statistične hipoteze

H_0 *ničelna hipoteza* o porazdelitvi slučajne spremenljivke.
. . . Slučajna spremenljivka je porazdeljena normalno.
. . . Slučajna spremenljivka ima matematično upanje enako 2.
H_{alt} *alternativna hipoteza* o porazdelitvi slučajne spremenljivke.
. . . Slučajna spremenljivka ni porazdeljena normalno.
. . . Slučajna spremenljivka ima matematično upanje različno od 2.

S statističnim testom testiramo ničelno hipotezo H_0 *proti* alternativni hipotezi H_{alt}.

Napake, ki jih lahko naredimo

Napaka 1. vrste ničelna hipoteza je pravilna, s testom jo zavrnemo.
Verjetnost, da naredimo napako 1. vrste je merljiva in jo lahko poljubno zmanjšamo.

Napaka 2. vrste ničelna hipoteza je napačna, s testom jo sprejmemo.
Verjetnosti, da naredimo napako 2. vrste ni mogoče oceniti. Zato napak 2. vrste ne delamo.
To pomeni, da ničelnih hipotez nikoli ne sprejmemo.

Stopnja značilnosti testa je verjetnost, da zavrnemo pravilno hipotezo (naredimo napako 1. vrste).

Stopnja značilnosti α je tipično $0,1$, $0,05$ ali $0,01$.

Parametrični test — dvostranski

Naloga: Imamo vzorec $n = 3000$ otrok, od katerih je 1578 dečkov.

Pri stopnji značilnosti $\alpha = 0,01$ testiraj ničelno hipotezo

H_0 *verjetnost rojstva dečka je enaka* $\frac{1}{2}$.

proti alternativni hipotezi

H_{alt} *verjetnost rojstva dečka je različna od* $\frac{1}{2}$.

Parametrični test — enostranski

Naloga: Imamo vzorec $n = 3000$ otrok, od katerih je 1578 dečkov.
Pri stopnji značilnosti $\alpha = 0,01$ testiraj ničelno hipotezo

H_0 *verjetnost rojstva dečka je večja ali enaka* $\frac{1}{2}$.

proti alternativni hipotezi

H_{alt} *verjetnost rojstva dečka je manjša od* $\frac{1}{2}$.

Neparametrični test χ^2

Naj bo X_{ref} znana slučajna spremenljivka. Poznamo njeno verjetnostno shemo.

$$X_{ref} \sim \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

Naj bo X *opazovana* slučajna spremenljivka.
Z neparametričnim testom χ^2 testiramo ničelno hipotezo

H_0 *slučajna spremenljivka* X *je porazdeljena enako kot slučajna spremenljivka* X_{ref}.

proti alternativni hipotezi

H_{alt} *slučajna spremenljivka* X *ni porazdeljena enako kot slučajna spremenljivka* X_{ref}.

Neparametrični test χ^2

Izberemo X-vzorec velikosti n in sestavimo tabelico :

Dogodek	$X = x_1$	$X = x_2$	$\ldots$	$X = x_k$
Izmerjena frekvenca	X_1	X_2	$\ldots$	X_k
Pričakovana frekvenca	$n \cdot p_1$	$n \cdot p_2$	$\ldots$	$n \cdot p_k$

Izračunamo statistiko

$$\chi^2 = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} + \ldots + \frac{(X_k - np_k)^2}{np_k} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

pri čemer je $X_1 + X_2 + \ldots + X_k = n$.

Neparametrični test χ^2

Izrek

Pri veljavni ničelni hipotezi, dovolj velikem vzorcu ($n \geq 30$) *in če je* $n \cdot p_i \geq 5$ *za vsak* i, *je statistika* χ^2 *porazdeljena približno kot hi-kvadrat s* $(k-1)$ *prostostnimi stopnjami* $\chi^2(k-1)$.

Izberemo stopnjo značilnosti α .

Če je $\chi^2 > \chi^2_{1-\alpha}(k-1)$, potem hipotezo **zavrnemo**.
Če je $\chi^2 \leq \chi^2_{1-\alpha}(k-1)$, potem hipoteze **ne zavrnemo**.

Neparametrični test χ^2

Naloga: Kocko vržemo 600 krat. Padlo je 105 enic, 99 dvojk, 102 trojki, 104 štirice, 93 petic in 97 šestic. Pri stopnji zaupanja $\alpha = 0,05$ testiraj ničelno hipotezo *"Kocka je poštena."*

Test χ^2 za testiranje neodvisnosti

V prodajalni bureka vzorčimo kupce in spremljamo njihova naročila. Dobili smo naslednjo kontingenčno tabelo.

starost/vrsta	mesni	sirov	pizza
mladi	25	17	18
mlajši	42	25	23
starejši	33	25	22
stari	12	10	8

Naloga: Testiraj ničelno hipotezo *"Izbira vrste bureka je neodvisna od starosti kupca."*

Test χ^2 za testiranje neodvisnosti

Kot v prejšnjem primeru izračunamo količino

$$\chi^2 = \sum_{\text{po vseh celicah}} \frac{(\text{izmerjena vrednost} - \text{pričakovana vrednost})^2}{\text{pričakovana vrednost}}$$

Izrek

Pri veljavni ničelni hipotezi , dovolj velikem vzorcu $(n \geq 30)$ *in če je vsaka pričakovana vrednost* ≥ 5, *je statistika* χ^2 *porazdeljena približno kot hi-kvadrat s* $(r-1)(s-1)$ *prostostnimi stopnjami.*

Test χ^2 za testiranje neodvisnosti

Naj bosta X in Y slučajni spremenljivki.

Test χ^2 uporabljamo za testiranje ničelne hipoteze

H_0 *X in Y sta neodvisni.*

proti alternativni hipotezi

H_{alt} *X in Y sta odvisni.*

Test z znaki

Naj bosta X in Y slučajni spremenljivki.
Test z znaki uporabljamo za testiranje ničelne hipoteze
H_0 $P(X > Y) = 0.5$.
proti alternativni hipotezi
H_{alt} $P(X > Y) \neq 0.5$.

Potrebujemo isto število meritev $X_1, X_2, \ldots, X_n$ in $Y_1, Y_2, \ldots, Y_n$.

Mann-Whitneyev test

Naj bosta X in Y slučajni spremenljivki.
Mann-Whitneyev test uporabljamo za testiranje ničelne hipoteze
H_0 *X in Y sta enako porazdeljeni.*
proti alternativni hipotezi
H_{alt} *X in Y nista enako porazdeljeni.*