Statistika

9. predavanje

Barbara Boldin

Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Univerza na Primorskem

Primerjava dveh populacij

Analizirati želimo statistično spremenljivko v dveh populacijah, npr.:

- učinkovitost cepiva v dveh starostnih skupinah,
- količino padavin v dveh regijah,
- delež kadilcev pri moških in ženskah,
- itd.

Izbrano spremenljivko preučujemo z vzorčenjem obeh populacij. Vzorca sta lahko

- **neodvisna**
- **odvisna**

Primerjava dveh populacij: neodvisna vzorca

Naj bo X intervalska slučajna spremenljivka, X_1 naj označuje X v 1. populaciji, X_2 pa X v 2. populaciji. Ločimo več primerov.

♠ Denimo, da je standardni odklon σ znan in enak v obeh populacijah ter

$$X_1 \sim N(\mu_1, \sigma^2), \qquad X_2 \sim N(\mu_2, \sigma^2).$$

Če iz prve populacije izbiramo vzorce velikosti n_1, iz druge populacije pa vzorce velikosti n_2, sta vzorčni porazdelitvi

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right).$$

Če sta vzorca neodvisna, je $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ in torej

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1).$$

Primerjava dveh populacij: neodvisna vzorca

Naj bo X intervalska slučajna spremenljivka, X_1 naj označuje X v 1. populaciji, X_2 pa X v 2. populaciji. Ločimo več primerov.

♠ Denimo, da je standardni odklon σ znan in enak v obeh populacijah ter

$$X_1 \sim N(\mu_1, \sigma^2), \qquad X_2 \sim N(\mu_2, \sigma^2).$$

Če iz prve populacije izbiramo vzorce velikosti n_1, iz druge populacije pa vzorce velikosti n_2, sta vzorčni porazdelitvi

$$\bar{X}_1 \sim N\Big(\mu_1, \frac{\sigma^2}{n_1}\Big), \qquad \bar{X}_2 \sim N\Big(\mu_2, \frac{\sigma^2}{n_2}\Big).$$

Če sta vzorca neodvisna, je $\bar{X}_1 - \bar{X}_2 \sim N\Big(\mu_1 - \mu_2, \sigma^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big)$ in torej

$$\boxed{Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1).}$$

♣ Kadar populacijskih varianc ne poznamo, ju ocenimo iz vzorcev:

◇ Naj bodo

$$x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$$

vrednosti naključnega vzorca velikosti n_1 iz prve populacije in $\bar{x}_1$ povprečje tega vzorca. Tedaj je nepristranska ocena za varianco prve populacije

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1,j} - \bar{x}_1)^2.$$

◇ Če so

$$x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$$

vrednosti naključnega vzorca velikosti n_2 iz druge populacije in $\bar{x}_2$ povprečje tega vzorca, je nepristranska ocena za varianco druge populacije

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2,j} - \bar{x}_2)^2.$$

Kadar se vzorčni varianci ne razlikujeta “preveč”, lahko domnevamo, da sta populacijski varianci enaki in vzorčni varianci zamenjamo s povprečno (oz. skupno) varianco

$$s_{sk}^2 = \frac{s_1^2 + s_2^2}{2}.$$

Tedaj je

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{sk}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Studentova porazdelitev s stopnjami prostosti $df = n_1 + n_2 - 2$.

Kdaj lahko varianci nadomestimo s povprečjem obeh varianc (skupno varianco)? Kadar se oceni varianc ne razlikujeta preveč, konkretno kadar

$$\frac{max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)} < 2,$$

torej, ko je razmerje med večjo in manjšo oceno variance pod 2.

♦ Kadar je

$$\frac{max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)} > 2,$$

potem vzorčni varianci s_1^2 in s_2^2 kažeta na to, da populacijski varianci nista enaki.

Tedaj uporabimo Studentovo porazdelitev

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

ki ima stopnjo prostosti $df = min(n_1 - 1, n_2 - 1)$.

Preizkus domneve o razliki povprečij dveh populacij

Neodvisna vzorca

Preizkušamo ničelno domnevo

$H_0 : \mu_1 - \mu_2 = \delta$

za nek δ.

Najpogosteje nas zanima enakost obeh povprečij, torej $\delta = 0$

Alternativna domneva je lahko dvostranska ($H_1 : \mu_1 - \mu_2 \neq \delta$) ali ena od enostranskih domnev ($H_1^+ : \mu_1 - \mu_2 > \delta$ ali $H_1^- : \mu_1 - \mu_2 < \delta$)

- kadar $\frac{max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)} < 2$ je testna statistika

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{sk}\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}, \qquad df = n_1 + n_2 - 2$$

- Če je $\frac{max(s_1^2, s_2^2)}{min(s_1^2, s_2^2)} > 2$, je testna statistika

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}, \qquad df = min(n_1 - 1, n_2 - 1).$$

Primer. Učitelja zanima, ali je število ur učenja za kolokvij enako za študentke in študente. Učitelj naključno izbere 7 študentk in 6 študentov in zabeleži, koliko ur so izbrani porabili za priprave:

študentke	15, 18, 3, 10, 20, 13, 12
študenti	14, 15, 12, 5, 10, 4

Pri stopnji značilnosti $\alpha = 0.05$ preverite domnevo, da je povprečno število ur priprav enako pri študentkah in študentih.

Naj bo

$\mu_1 =$ povprečno število ur priprav za študentke,
$\mu_2 =$ povprečno število ur priprav za študente.

Hipotezi sta

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 \neq 0.$

Primer. Učitelja zanima, ali je število ur učenja za kolokvij enako za študentke in študente. Učitelj naključno izbere 7 študentk in 6 študentov in zabeleži, koliko ur so izbrani porabili za priprave:

študentke	15, 18, 3, 10, 20, 13, 12
študenti	14, 15, 12, 5, 10, 4

Pri stopnji značilnosti $\alpha = 0.05$ preverite domnevo, da je povprečno število ur priprav enako pri študentkah in študentih.

Naj bo

$\mu_1 =$ povprečno število ur priprav za študentke,
$\mu_2 =$ povprečno število ur priprav za študente.

Hipotezi sta

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 \neq 0.$

Primer. Učitelja zanima, ali je število ur učenja za kolokvij enako za študentke in študente. Učitelj naključno izbere 7 študentk in 6 študentov in zabeleži, koliko ur so izbrani porabili za priprave:

študentke	15, 18, 3, 10, 20, 13, 12
študenti	14, 15, 12, 5, 10, 4

Pri stopnji značilnosti $\alpha = 0.05$ preverite domnevo, da je povprečno število ur priprav enako pri študentkah in študentih.

Naj bo

$\mu_1 =$ povprečno število ur priprav za študentke,
$\mu_2 =$ povprečno število ur priprav za študente.

Hipotezi sta

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 \neq 0.$

Računamo: $\bar{x}_1 = 13$, $\bar{x}_2 = 10$.

$$s_1^2 = \frac{1}{6}\Big((15-13)^2 + \ldots + (12-13)^2\Big) = 31.3$$
$$s_2^2 = \frac{1}{5}\Big((14-10)^2 + \ldots + (4-10)^2\Big) = 21.2.$$

Ker je $\frac{max(s_1^2,s_2^2)}{min(s_1^2,s_2^2)} = \frac{31.3}{21.2} < 2$ uporabimo $s_{\text{sk}}^2 = \frac{s_1^2+s_2^2}{2} = 26.25$.

Testna statistika je

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\text{sk}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.05$$

Imamo $df = n_1 + n_2 - 2 = 11$. Iz tabele Studentove porazdelitve razberemo kritični vrednosti $\pm t_{0.975}(df) = \pm 2.201$.

Ker $t \in (-2.201, 2.201)$ hipotezo H_0 obrdržimo. Pri stopnji značilnosti $\alpha = 0.05$ trdimo, da je povprečno število ur študija enako za študentke in študente.

Primer. Raziskovalce zanima, ali je povprečen pridelek paradižnika sorte A enak povprečnemu pridelku paradižnika sorte B. Na homogenem zemljišču naključno izberemo 6 parcel s sorto A in 6 parcel s sorto B. Pridelek na izbranih zemljiščih je (v kg):

sorta A :	18, 19, 23, 21, 20, 19
sorta B :	22, 24, 24, 24, 22, 22

Pri stopnji značilnosti $\alpha = 0.05$ preverite domnevo, da je povprečen pridelek sorte A enak povprečnemu pridelku sorte B.

Naj bo

$\mu_A =$ povprečen pridelek sorte A,
$\mu_B =$ povprečen pridelek sorte B.

Hipotezi sta

$H_0 : \mu_A - \mu_B = 0$
$H_1 : \mu_A - \mu_B \neq 0.$

Primer. Raziskovalce zanima, ali je povprečen pridelek paradižnika sorte A enak povprečnemu pridelku paradižnika sorte B. Na homogenem zemljišču naključno izberemo 6 parcel s sorto A in 6 parcel s sorto B. Pridelek na izbranih zemljiščih je (v kg):

sorta A :	18, 19, 23, 21, 20, 19
sorta B :	22, 24, 24, 24, 22, 22

Pri stopnji značilnosti $\alpha = 0.05$ preverite domnevo, da je povprečen pridelek sorte A enak povprečnemu pridelku sorte B.

Naj bo

$\mu_A =$ povprečen pridelek sorte A,
$\mu_B =$ povprečen pridelek sorte B.

Hipotezi sta

$H_0 : \mu_A - \mu_B = 0$
$H_1 : \mu_A - \mu_B \neq 0.$

Računamo: $\bar{x}_A = 20, \bar{x}_B = 23$.

$$s_A^2 = \frac{1}{5}\Big((18-20)^2 + \ldots + (19-20)^2\Big) = 3.2$$
$$s_B^2 = \frac{1}{5}\Big((22-23)^2 + \ldots + (22-23)^2\Big) = 1.2$$

Ker je $\frac{max(s_A^2, s_B^2)}{min(s_A^2, s_B^2)} > 2$ uporabimo

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = -3.5$$

V tem primeru je $df = \min(n_1 - 1, n_2 - 1) = 5$. Iz tabele Studentove porazdelitve razberemo kritični vrednosti $\pm t_{0.975}(df) = \pm 2.571$.

Ker $t \notin (-2.571, 2.572)$ hipotezo H_0 zavrnemo. Rezultati kažejo, da sta povprečna pridelka sort A in B različna.

Primerjava Bernoullijevih verjetnosti

Sedaj želimo primerjati verjetnost nekega dogodka v dveh populacijah (t.j. vrednosti Bernoullijevih verjetnosti), npr.

- verjetnost kalitve dveh sort
- odstotek učinkovitosti zdravila pri moških in ženskah
- itd.

Naj bo X binomska slučajna spremenljivka, X_1 naj označuje porazdelitev X v prvi populaciji, X_2 pa porazdelitev X v drugi populaciji in privzemimo

$$X_1 \sim B(n_1, p_1) \approx N(n_1 p_1, n_1 p_1 (1 - p_1))$$
$$X_2 \sim B(n_2, p_2) \approx N(n_2 p_2, n_2 p_2 (1 - p_2))$$

(X_i ($i = 1, 2$) je torej število dogodkov v populaciji velikosti n_i, kjer je verjetnost dogodka p_i. Nadalje, diskretno binomsko porazdelitev v obeh populacijah aproksimiramo z zvezno normalno).

Porazdelitev deležev je tedaj

$$\frac{X_1}{n_1} \approx N\Big(p_1, \frac{p_1(1-p_1)}{n_1}\Big), \qquad \frac{X_2}{n_2} \approx N\Big(p_2, \frac{p_2(1-p_2)}{n_2}\Big).$$

Če sta X_1 in X_2 neodvisna, je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big).$$

Izračun se poenostavi, če sta p_1 in p_2 enaka, označimo ju s p. Tedaj je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(0, p(1-p)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big).$$

Porazdelitev deležev je tedaj

$$\frac{X_1}{n_1} \approx N\Big(p_1, \frac{p_1(1-p_1)}{n_1}\Big), \qquad \frac{X_2}{n_2} \approx N\Big(p_2, \frac{p_2(1-p_2)}{n_2}\Big).$$

Če sta X_1 in X_2 neodvisna, je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big).$$

Izračun se poenostavi, če sta p_1 in p_2 enaka, označimo ju s p. Tedaj je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(0, p(1-p)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\Big).$$

Porazdelitev deležev je tedaj

$$\frac{X_1}{n_1} \approx N\Big(p_1, \frac{p_1(1-p_1)}{n_1}\Big), \qquad \frac{X_2}{n_2} \approx N\Big(p_2, \frac{p_2(1-p_2)}{n_2}\Big).$$

Če sta X_1 in X_2 neodvisna, je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big).$$

Izračun se poenostavi, če sta p_1 in p_2 enaka, označimo ju s p. Tedaj je

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \approx N\Big(0, p(1-p)\big(\frac{1}{n_1} + \frac{1}{n_2}\big)\Big).$$

Kako izračunamo ocene za p_1 in p_2 iz vzorca?

- iz vzorca 1. populacije izračunamo oceno za p_1, označimo jo s $\hat{p}_1$

$$\hat{p}_1 = \frac{x_1}{n_1},$$

kjer je n_1 velikost vzorca, x_1 pa število enot v vzorcu pri katerih se je dogodek zgodil

- iz vzorca 2. populacije izračunamo oceno za p_2, označimo jo s $\hat{p}_2$

$$\hat{p}_2 = \frac{x_2}{n_2},$$

kjer je n_2 velikost vzorca, x_2 pa število enot v vzorcu pri katerih se je dogodek zgodil

- ocena za p iz obeh vzorcev je

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Recimo, da želimo preveriti domnevo, da je verjetnost nekega dogodka v 1. populaciji enaka kot verjetnost istega dogodka v 2. populaciji, torej ali je p_1 enak p_2.

Postavimo ničelno domnevo
$H_0 : p_1 = p_2$

in alternativno domnevo, ki je lahko dvostranska ($H_1 : p_1 \neq p_2$) ali ena od enostranskih domnev ($H_1^+ : p_1 > p_2$ ali $H_1^- p_1 < p_2$)

Testna statistika je sedaj

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Kadar preizkušamo domeno, da je razlika verjetnosti nek neničelen p_0, torej
$H_0 : p_1 - p_2 = p_0$
potem je testna statistika

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Primer. Raziskovalci primerjajo učinkovitost dveh zdravil proti visokemu krvnemu tlaku. Imajo 240 pacientov z visokim krvnim tlakom. S slučajno izbiro bolnike razdelijo v dve enako veliki skupini, prvi skupini dajo zdravilo A, drugi pa zdravilo B. Stanje se je izboljšalo 85 pacientom, ki so prejeli zdravilo A in 95 bolnikom, ki so prejeli zdravilo B. Pri stopnji značilnosti $\alpha = 0.05$ preverite učinkovitost obeh zdravil.

Naj bo

$p_A =$ verjetnost, da je zdravilo A učinkovito
$p_B =$ verjetnost, da je zdravilo B učinkovito.

Hipotezi sta

$H_0 : p_A - p_B = 0$
$H_1 : p_A - p_B \neq 0.$

Primer. Raziskovalci primerjajo učinkovitost dveh zdravil proti visokemu krvnemu tlaku. Imajo 240 pacientov z visokim krvnim tlakom. S slučajno izbiro bolnike razdelijo v dve enako veliki skupini, prvi skupini dajo zdravilo A, drugi pa zdravilo B. Stanje se je izboljšalo 85 pacientom, ki so prejeli zdravilo A in 95 bolnikom, ki so prejeli zdravilo B. Pri stopnji značilnosti $\alpha = 0.05$ preverite učinkovitost obeh zdravil.

Naj bo

$p_A =$ verjetnost, da je zdravilo A učinkovito
$p_B =$ verjetnost, da je zdravilo B učinkovito.

Hipotezi sta

$H_0 : p_A - p_B = 0$
$H_1 : p_A - p_B \neq 0.$

Računamo

$$\hat{p}_A = \frac{85}{120} = 0.71$$
$$\hat{p}_B = \frac{95}{120} = 0.79$$
$$\hat{p} = \frac{180}{240} = 0.75.$$

Testna statistika je

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_A} + \frac{1}{n_B})}} = -1.43.$$

Iz tabele za $N(0,1)$ razberemo kritični vrednosti $\pm z_0.975 = \pm 1.96$. Ker $z \in (-1.96, 1.96)$, hipotezo H_0 obdržimo.

Primerjava dveh populacij: dva odvisna vzorca

Pri neodvisnih vzorcih informacija prvega vzorca ni povezana z drugim vzorcem. Sedaj obravnavamo primer odvisnih vzorcev, ko je vsaka enota iz prve populacije v paru z enoto iz druge populacije.

Npr.:

- primerjamo znanje angleškega jezika študentov pred in po tečaju
- primerjamo težo pred in po dieti
- itd.

Privzemimo

$$X_1 \sim N(\mu_1, \sigma_1^2) \qquad X_2 \sim N(\mu_2, \sigma_2^2)$$

Na vsakem paru izračunamo razliko $D = X_1 - X_2$. Ker sta spremenljivki odvisni, variance ne moremo določiti samo iz σ_1 in σ_2,

$$D \sim N(\mu_1 - \mu_2, \sigma_D^2)$$

Če pišemo $\mu_D = \mu_1 - \mu_2$ je porazdelitev vzorčnih razlik $\bar{D} = N(\mu_D, \frac{\sigma_D^2}{n})$, kjer je n velikost vzorca. Torej je

$$Z = \frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}} \sim N(0, 1)$$

Kadar σ_D ne poznamo, jo nadomestimo z oceno s_D in je porazdelitev

$$T = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

Studentova porazdelitev s stopnjo prostosti $df = n - 1$.

Primer. Učenci so se med zimskimi počitnicami udeležili plavalnega tečaja. Trenerji so njihove plavalne sposobnosti preverili pred začetkom in po koncu tečaja, in sicer so merili dolžimo (v m), ki jo vsak plavalec preplava v dvanajstih minutah. Trenerji trdijo, da so plavalci po tečaju izboljšali svoj dosežek vsaj za 25m. Da bi preverili domnevo naključno izberejo pet plavalcev. Dobijo naslednje rezultate

Pred tečajem (X_{pred})	Po tečaju (X_{po})	Razlika ($D = X_{\text{po}} - X_{\text{pred}}$)
375	400	25
500	550	50
375	390	15
400	425	25
450	465	15

Kaj lahko rečete o hipotezi trenerjev pri stopnji značilnosti $\alpha = 0.05$?

Naj bo

μ_{po} = povprečna preplavana dolžina po tečaju
μ_{pred} = povprečna preplavana dolžina pred tečajem
$\mu_D = \mu_{\text{po}} - \mu_{\text{pred}}$.

Domnevi sta
$H_0 : \mu_D = 25m$
$H_1 : \mu_D > 25m$

Računamo: $\bar{d} = 26$

$$s_D^2 = 205$$

Testna statistika:

$$t = \frac{\bar{d} - \mu_d}{\sqrt{\frac{s_D^2}{n}}} = 0.1562.$$

Imamo $df = n - 1 = 4$. Iz tabele Studentove porazdelitve razberemo prag zavrnitve za enostransko alternativno domnevo, t.j. 2.132.

Ker $t < 2.132$ hipotezo H_0 obdržimo. Rezultati torej ne kažejo, da bi se rezultati po tečaju izboljšali za vsaj 25m.