Analiza variance

Analiza variabilnosti (razpršenosti)

Zgled: Imamo vzorčne podatke o življenjski dobi lisic.

Nekateri dejavniki, ki vplivajo na življenjsko dobo: življenjski prostor, prehrana, bolezni, naravni sovražniki...

Ali tudi spol vpliva na življenjsko dobo?

Kako to preverimo?

1. MOŽNOST: Testiramo enakost povprečne življenjske dobe pri samcih in samicah s Studentovim T-testom.

2. MOŽNOST: Uporabimo analizo variabilnosti.

Pri tej analizi lahko hkrati primerjamo povprečja več skupin.

Lisice razdelimo v dve skupini:

- SAMCI
- SAMICE

- SAMCI
  - X_1 meri življenjsko dobo samcev.
  - Predpostavimo: $X_1 \sim N(\mu_1, \sigma)$.
  - Velikost vzorca: n_1.
  - Vzorčni podatki: $x_{11}, x_{12}, \ldots, x_{1n_1}$.
  - Vzorčno povprečje: $\overline{x_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$.
  - Vzorčna disperzija: $S^2_{x_1} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1i} - \overline{x_1})^2$.
- SAMICE
  - X_2 meri življenjsko dobo samic.
  - Predpostavimo: $X_2 \sim N(\mu_2, \sigma)$.
  - Velikost vzorca: n_2.
  - Vzorčni podatki: $x_{21}, x_{22}, \ldots, x_{2n_2}$.
  - Vzorčno povprečje: $\overline{x_2} = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$.
  - Vzorčna disperzija: $S^2_{x_2} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2i} - \overline{x_2})^2$.

DEJSTVA:

- Razlike v življenjski dobi znotraj skupine SAMCI (SAMICE) niso odvisne od spola.

  Na te razlike vplivajo drugi dejavniki.

  Variabilnost znotraj skupine = nepojasnjena variabilnost.
- Na razlike v življenjski dobi med skupinama SAMCI in SAMICE vpliva spol.

  Variabilnost med skupinama = pojasnjena variabilnost.

  Spol vpliva na življenjsko dobo, če je ta variabilnost velika oz. statistično značilna.
- Skupna variabilnost = nepojasnjena var. + pojasnjena var.

Izračun variabilnosti

- Spomnimo se:
  - Vzorčno povprečje pri samcih: $\overline{x_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$.
  - Vzorčno povprečje pri samicah: $\overline{x_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$.

  Označimo:
  - Skupno povprečje: $\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$ (tehtano povprečje).
- Odkloni znotraj skupine in med skupinama.

- Variabilnost znotraj skupin:
  - $\sum_{j=1}^{n_1}(x_{1j} - \overline{x_1})^2 + \sum_{j=1}^{n_2}(x_{2j} - \overline{x_2})^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2$.
- Variabilnost med skupinama:
  - $n_1(\overline{x_1} - \overline{x})^2 + n_2(\overline{x_2} - \overline{x})^2 = \sum_{i=1}^{2} n_i(\overline{x_i} - \overline{x})^2$.
- Skupna variabilnost:
  - $\sum_{j=1}^{n_1}(x_{1j} - \overline{x})^2 + \sum_{j=1}^{n_2}(x_{2j} - \overline{x})^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2$.
- Ker velja:

  Skupna variabilnost = nepojasnjena var. + pojasnjena var.

  Dobimo:
  $\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2 + \sum_{i=1}^{2} n_i(\overline{x_i} - \overline{x})^2$.

Skupna varianca

- Če dobljeno enačbo delimo z $n = n_1 + n_2$, dobimo:
  $\sigma_s^2 = \sigma_z^2 + \sigma_m^2$.
  - s...skupna,
  - z...znotraj skupin,
  - m...med skupinama.

  Vir variance je torej znotraj skupin ali med skupinami.

  Skupna varianca = nepojasnjena var. + pojasnjena var.

Analiza variabilnosti za tri skupine

Zgled: Na neki šoli za pse so uporabljali tri različne metode urjenja:

- 1. skupino psov so urili po I. metodi,
- 2. skupino psov so urili po II. metodi in
- 3. skupino psov so urili po III. metodi.

Izurjenost naključno izbranih psov so na koncu preverili s poligonom in jih ocenili s točkami.

Predpostavimo:
- X_i meri rezultate skupine i $(i = 1, 2, 3)$.
- $X_i \sim N(\mu_i, \sigma)$, $i = 1, 2, 3$.

Katera metoda je bila najbolj učinkovita?

1.MOŽNOST: Uporabimo Studentov T-test o enakosti povprečij neodvisnih vzorcev za vse tri možne pare dveh skupin.

Testiramo:

- $H_0(\mu_1 = \mu_2) : H_1(\mu_1 \neq \mu_2)$,
- $H_0(\mu_1 = \mu_3) : H_1(\mu_1 \neq \mu_3)$ in
- $H_0(\mu_2 = \mu_3) : H_1(\mu_2 \neq \mu_3)$.

2.MOŽNOST: Analiza variabilnosti.

Testiramo:

- $H_0(\mu_1 = \mu_2 = \mu_3) : H_1(\text{vsaj eno povp. je različno od ostalih})$.

DEJSTVA:

- Variabilnost rezultatov znotraj posamezne skupine ni odvisna od metode urjenja (nepojasnjena variabilnost).

  Na te razlike vplivajo drugi dejavniki: starost, gibalne sposobnosti...
- Na variabilnost rezultatov med skupinami vpliva metoda urjenja (pojasnjena variabilnost).
- Metode urjenja so različno učinkovite, če je ta variabilnost velika oz. statistično značilna.

ANOVA – Analiza variance

- ANOVA – analysis of variance.
- Uporabimo jo za testiranje enakosti povprečij pri $k \geq 4$ neodvisnih skupinah (lahko tudi za $k = 2$ in $k = 3$.)
- BISTVO: Skupno variabilnost razdelimo na dve komponenti (pojasnjeno in nepojasnjeno).

Oznake in predpostavke:

- Imamo k skupin.
- Imamo k neodvisnih vzorcev velikosti $n_1, n_2, \ldots, n_k$ $(n_1 + n_2 + \cdots + n_k = n)$.
- X_i meri količino, ki nas zanima na vzorcu i $(i = 1, 2, \ldots, k)$.
- $X_1 \sim N(\mu_1, \sigma), X_2 \sim N(\mu_2, \sigma), \ldots, X_k \sim N(\mu_k, \sigma)$.
- Vzorčni podatki i-te skupine: $x_{i1}, x_{i2}, \ldots, x_{in_i}$.
- Izberemo stopnjo značilnosti α.

- Na stopnji značilnosti α testiramo ničelno hipotezo

$$H_0(\mu_1 = \mu_2 = \ldots = \mu_k)$$

proti alternativi

$$H_1(\text{vsaj eno povprečje je različno od ostalih}).$$

- Naj bodo

$$\overline{x_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad \overline{x_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}, \ldots, \overline{x_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}$$

vzorčna povprečja skupin in

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + \cdots + n_k\overline{x_k}}{n}$$

skupno povprečje (tehtano povprečje).

- Velja:

$$\sum_{i,j}(x_{ij}-\overline{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2 + \sum_{i=1}^{k} n_i(\overline{x_i}-\overline{x})^2.$$

Skupna variabilnost = nepojasnjena var. + pojasnjena var.

$VK_s = VK_z + VK_m$

  - VK...vsota kvadratov,
  - s...skupna,
  - z...znotraj skupin,
  - m...med skupinami.

ANOVA – testna statistika

- Izkaže se, da je statistika (testna statistika metode ANOVA)

$F = \frac{PK_m}{PK_z} \sim F(k-1, n-k)$, kjer je

  - $PK_m = \frac{VK_m}{k-1}$ povprečje kvadratov med skupinami,
  - $PK_z = \frac{VK_z}{n-k}$ povprečje kvadratov znotraj skupin,
  - $F(k-1, n-k)$ Fisherjeva porazdelitev s $k-1$ in $n-k$ prostostnimi stopnjami.

- Kritično območje enostranskega testa:
  - Izberemo tak f_α, da velja $P(F(k-1, n-k) \geq f_\alpha) = \alpha$. Uporabimo tabelo D.
  - $K_\alpha = [f_\alpha, \infty)$.
  - Izračunamo vrednost testne statistike F na vzorcu: F_e.
  - Če je $F_e \geq f_\alpha$, hipotezo H_0 zavrnemo in potrdimo H_1.
  - Če je $F_e < f_\alpha$, o hipotezi H_0 ne odločimo.
- Naj bo $p = P(F \geq F_e)$ (signifikanca testa).
  Če je $p \leq \alpha$, H_0 zavrnemo.
- Tabela metode ANOVA

VV vir variance	VK vsota kvadratov	PS prostostne stopnje	$PK = \frac{VK}{PS}$ povprečje kvadratov	F
med skupinami	VK_m	$k-1$	$PK_m = \frac{VK_m}{k-1}$	
znotraj skupin	VK_z	$n-k$	$PK_z = \frac{VK_z}{n-k}$	
skupna	VK_s	$n-1$	$S^2 = \frac{VK_s}{n-1}$	$F = \frac{PK_m}{PK_z}$

Zgled: Primerjali bomo 5 različnih pooperativnih postopkov za okrevanje po operaciji (5 skupin). Zanima nas, ali kateri od postopkov statistično značilno vpliva na čas okrevanja (v dnevih).

- Za vsako skupino izberemo vzorec pacientov.
- X_i meri čas okrevanja v skupini i ($i = 1, 2, \ldots, 5$).
- $X_i \sim N(\mu_i, \sigma)$.
- Izberemo $\alpha = 0,05$.
- Na stopnji značilnosti $\alpha = 0,05$ testiramo ničelno hipotezo

$$H_0(\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$$

proti alternativi

$$H_1(\text{vsaj eno povprečje je različno od ostalih}).$$

Vzorčni podatki:

Skupina	Velikost vzorca	Vzorčni podatki	Vzorčno povprečje
1	$n_1 = 8$	$3, 5, 5, 2, 4, 5, 3, 3$	$\overline{x_1} = 3,75$
2	$n_2 = 6$	$4, 7, 6, 6, 9, 7$	$\overline{x_2} = 6,50$
3	$n_3 = 8$	$2, 3, 4, 3, 3, 5, 5, 4$	$\overline{x_3} = 3,63$
4	$n_4 = 6$	$6, 3, 5, 5, 2, 4$	$\overline{x_4} = 4,17$
5	$n_5 = 7$	$8, 3, 4, 5, 6, 6, 5$	$\overline{x_5} = 5,29$
Skupaj	$n = 35$		$\overline{x} = 4,57$

- Tabela metode ANOVA

VV	*VK*	*PS*	*PK*	F_e
med	$39,435$	4	$9,859$	
znotraj	$57,137$	30	$1,905$	
skupna	$96,571$	34		$5,176$

  - $f_\alpha = 2,69$.
  - Ker je $F_e > f_\alpha$, H_0 zavrnemo in potrdimo H_1.
  - Ker je $p = 0,003 \leq 0,05 = \alpha$, H_0 zavrnemo in potrdimo H_1.

Opombe k metodi ANOVA

1. Ena od predpostavk metode je, da so standardni odkloni spremenljivk enaki. Zato moramo pred uporabo metode ANOVA opraviti test homogenosti varianc.

   Testirati moramo ničelno hipotezo

   $$H_0(\sigma_1 = \sigma_2 = \ldots = \sigma_k)$$

   proti alternativi

   $$H_1(\text{vsaj en odklon se razlikuje od ostalih}).$$

   SPSS to opravi s pomočjo Leveneovega testa – testna statistika F je porazdeljena po $F(k-1, n-k)$.

- Primer:

  $F_e = 0,242$, $f_\alpha = 2,69$

  Ker je $F_e < f_\alpha$, o H_0 ne odločimo (je ne zavrnemo).

  Ker je $p = 0,912 > 0,05 = \alpha$, o H_0 ne odločimo.

Opombe k metodi ANOVA

- Če se standardni odkloni razlikujejo, ANOVA ni utemeljena. V tem primeru uporabimo Welchov test ali Brown-Forsythov test.

2. Če predpostavke o normalnosti porazdelitev niso izpolnjene, uporabimo neparametrični Kruskal-Wallisov test.
3. ANOVA pri dveh skupinah je ekvivalentna Studentovemu T-testu o enakosti povprečij.