Statistika II z računalniško analizo podatkov

Preverjanje domnev o aritmetičnih sredinah: analiza variance (enofaktorska (univariatna) analiza variance)

VII Preverjanje domnev o aritmetičnih sredinah

1. Ponovitev iz predmeta Statistika in nadgradnja.

Preverjanje domnev o aritmetični sredini in o razliki dveh aritmetičnih sredin v SPSS-ju: *t* testi

2. Preverjanje domneve o aritmetični sredini
3. Preverjanje domneve o razliki dveh aritmetičnih sredin za neodvisna vzorca
4. Preverjanje domneve o razliki dveh aritmetičnih sredin za odvisna vzorca

Preverjanje domnev o razliki aritmetičnih sredin v treh ali več skupinah v SPSS-ju

5. Analiza variance

VII.5 Analiza variance

- Metoda, ki jo uporabimo, kadar želimo ugotoviti, ali obstajajo razlike v povprečjih med tremi ali več skupinami (preverjanje domneve o treh ali več aritmetičnih sredinah).
- Zakaj ime „analiza variance“ in ne npr. „analiza povprečij“? Zaradi metodologije, ki je uporabljena za določitev, kako različne so si aritmetične sredine, da zaključimo, da so si „statistično značilno“ različne.

- Terminologija
  - spremenljivka, za katero računamo povprečja – odvisna spremenljivka (intervalna ali razmernostna),
  - spremenljivka(e), ki določa(jo) skupine – neodvisna(e) spremenljivka(e) – faktor(ji) (običajno nominalna ali ordinalna)

VII.5 Analiza variance

Več vrst analize variance:

- Če preverjamo domnevo o aritmetičnih sredini ene spremenljivke v več skupinah, ki so določene z eno spremenljivko (enim faktorjem) - enofaktorska analiza variance (angl. one-way analysis of variance – one-way ANOVA) – 1 odvisna, 1 neodvisna spremenljivka.

  Primer: razlika v povprečnem zaupanju v državni zbor med mestnimi, primestnimi in vaškimi prebivalci.
- Če preverjamo domnevo o aritmetičnih sredini ene spremenljivke v več skupinah, ki so določene z dvema spremenljivkama (dvema faktorjema) - dvofaktorska analiza variance (angl. two-way analysis of variance – two-way ANOVA) – 1 odvisna, 2 neodvisni spremenljivki.

  Primer: razlika v povprečnem zaupanju v državni zbor glede na kraj bivanja in spol (med ženskami iz mesta, primestja, vasi in med moškimi iz mesta, primestja, vasi; torej 6 skupin)
- Če preverjamo domnevo o aritmetičnih sredinah večih spremenljivk v večih skupinah – multivariatna analiza variance (angl. multivariate analysis of variance – MANOVA) – več odvisnih, več neodvisnih spr.

  Primer: razlika v povprečnem zaupanju v niz inštitucij (državni zbor, policija, EU parlament itd.) med mestnimi, primestnimi in vaškimi prebivalci ali glede na kraj bivanja in spol .

VII.5 Analiza variance

Kazalo:

5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah
5.2 Enofaktorska analiza variance v primeru enako velikih skupin
5.3 Enofaktorska analiza variance v primeru različno velikih skupin
5.4 Kako izvedemo analizo variance v SPSS-ju?
  5.4.1 Preverjanje predpostavk analize variance
  5.4.2 Izvedba analize variance
  5.4.3 Popravki analize variance v primeru kršitve predpostavk
5.5 Dodatni testi za razlike za pare skupin

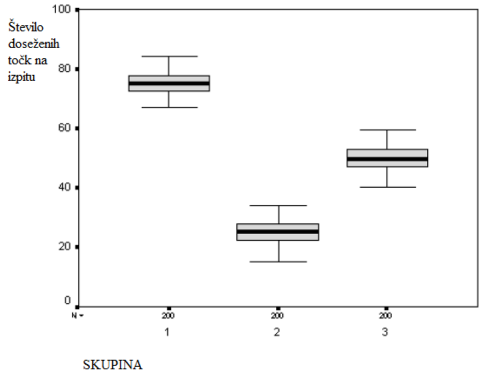
VII.5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah

Primer:

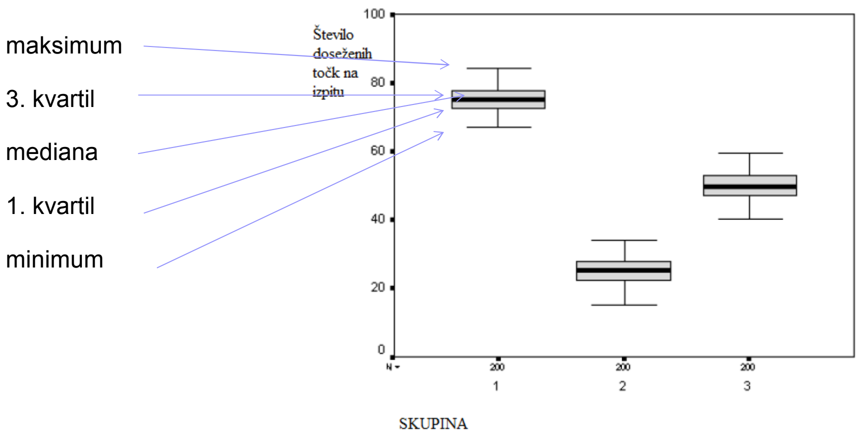
- Enote: osebe.
- Spremenljivka: število doseženih točk na testu (odvisna spremenljivka).
- Podatke imamo za 3 vzorce (3 skupine) po 200 enot (neodvisna spremenljivka je pripadnost vzorcu).
- Vprašanje: Ali vsi trije vzorci prihajajo iz iste populacije (torej ni razlik med tremi vzorci) ali pa prihajajo iz različnih populacij (torej so razlike med tremi vzorci)?
- Reševanje problema: primerjamo razlike v vzorčnih aritmetičnih sredinah 3 vzorcev (razlike med $\overline{X_1}, \overline{X_2}$ in $\overline{X_3}$).

  - ✓ Če so razlike v vzorčnih aritmetičnih sredinah majhne. → Ni razlike med vzorci, vsi vzorci so vzeti iz iste populacije.
  - ✓ Če so razlike v vzorčnih aritmetičnih sredinah velike. → Obstajajo razlike med vzorci, vzorci niso iz iste populacije, pač pa iz različnih populacij.

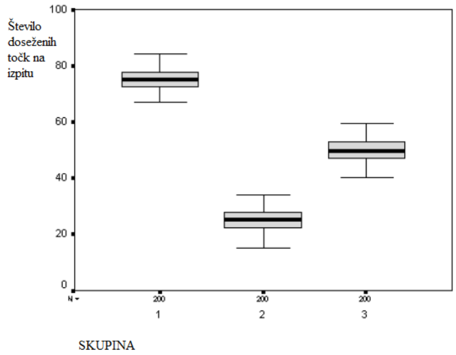
1. možna situacija: Jasno izražene razlike v povprečnem številu doseženih točk: 3 skupine iz treh različni populacij (ločene skupine).



„Škatlasti diagram“ (boxplot): Prikaz porazdelitve spremenljivke znotraj skupine s pomočjo 5 ključnih statistik:

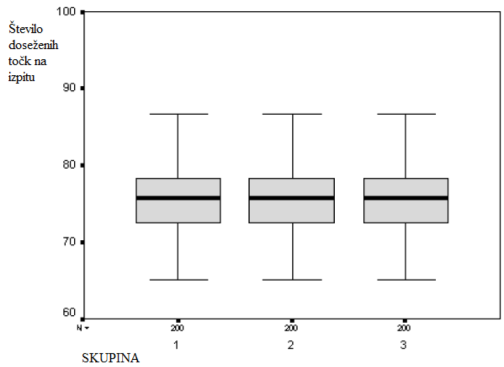


1. možna situacija: Jasno izražene razlike v povprečnem številu doseženih točk: 3 skupine iz treh različni populacij (ločene skupine).



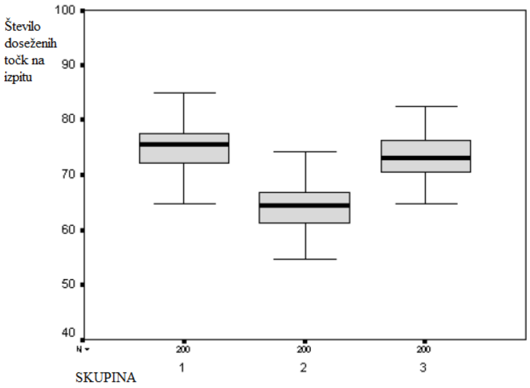
- Ni nobenega prekrivanja v številu doseženih točk med tremi skupinami in mediane so si daleč narazen.
- Primerjava variabilnosti (razlik) med medianami skupin in variabilnosti posameznih vrednosti znotraj vsake skupine: prva je mnogo večja od slednje.

2. možna situacija: Ni razlik v povprečnem številu doseženih točk: 3 skupine iz iste populacije (identične skupine).



- Popolno prekrivanje v število doseženih točk med tremi skupinami, mediane so si popolnoma enake.
- Primerjava variabilnosti (razlik) med medianami skupin in variabilnosti posameznih vrednosti znotraj vsake skupine: variabilnosti (razlik) med medianami skoraj ni in je torej v primerjavi z variabilnostjo posameznih vrednosti znotraj vsake skupine izredno majhna.

3. možna situacija: Ni jasno, ali so razlike v povprečnem številu doseženih točk: ni jasno, ali gre za 3 skupine iz iste populacije ali iz treh različnih populacij (prekrivajoče se skupine)



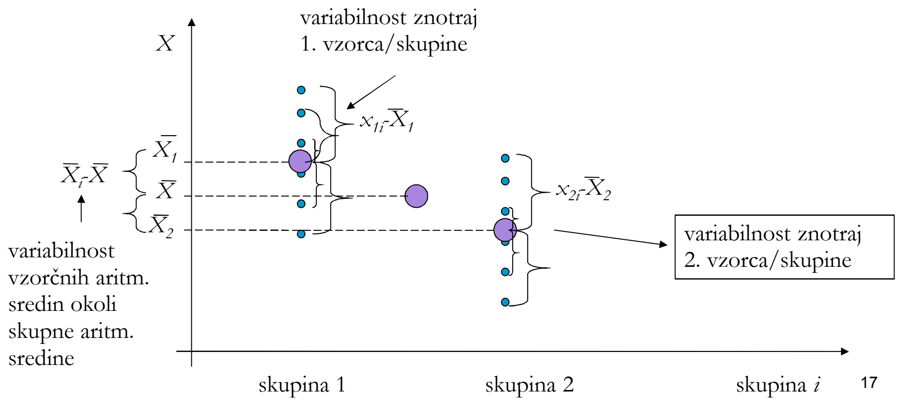
- Vrednosti posameznih skupin se prekrivajo, mediane med skupinami pa so različne. Ni čisto jasno, ali do razlik v srednji vrednosti prihaja zaradi razlik med skupinami ali zaradi slučajne variabilnosti enot znotraj skupine.
- Primerjava variabilnosti (razlik) med medianami skupin in variabilnosti posameznih vrednosti znotraj vsake skupine: ni jasno, ali so razlike med medianami večje od razlik med vrednostmi znotraj vsake skupine. Potreben formalen test analize variance.

VII.5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah

- Metoda, s katero preverjamo domneve o razlikah v aritmetičnih sredinah, se imenuje analiza variance.
- S to metodo primerjamo variabilnost med vzorčnimi aritmetičnimi sredinami skupin z variabilnostjo posameznih enot znotraj vsakega vzorca/skupine – primerjamo variabilnost med vzorci/skupinami z variabilnostjo znotraj vzorcev/skupin.
- Kadar je variabilnost med vzorčnimi aritmetičnimi sredinami bistveno večja od variabilnosti enot znotraj vsakega vzorca, zaključimo, da so populacijske aritmetične sredine različne (oz. da vzorci prihajajo iz različnih populacij).
- Če sta ti dve variabilnosti približno enaki, pa zaključimo, da ni razlik med populacijskimi aritmetičnimi sredinami (oz. da vsi vzorci prihajajo iz iste populacije).

Variabilnost med skupinami, variabilnost znotraj skupin

(angl. between group variance, within group variance)



Variabilnost med skupinami, variabilnost znotraj skupin

(angl. between group variance, within group variance)

variabilnost vseh enot (iz vseh vzorcev) okoli skupne aritmetične sredine	povprečna variabilnost enot v vzorcih okoli vzorčnih aritmetičnih sredin	variabilnost vzorčnih aritmetičnih sredin okoli skupne aritmetične sredine
$\frac{1}{\sum_{i=1}^{a} n_i}\sum_{t=1}^{n_i}(x_{it}-\bar X)^2 =$	$\frac{1}{a}\sum_{i=1}^{a} s_i^2 = \frac{1}{a}(\frac{1}{n_i-1}\sum_{t=1}^{n_i}(x_{it}-\bar X_i)^2))$	$+\ \frac{1}{a-1}\sum_{i=1}^{a}(\bar X_i-\bar X)^2$
s^2	s_p^2	$s_{\bar X}^2$
celotna varianca	variabilnost znotraj skupin (povprečna varianca znotraj vzorcev, angl. pooled variance)	variabilnost med skupinami

nepojasnjena varianca – variabilnost znotraj vzorcev, ki je slučajna

pojasnjena varianca – variabilnost, do katere prihaja, ker gre za vzorce iz različnih populacij

a … število vzorcev/skupin
n_i …število enot v *i*-tem vzorcu
i … indeks za vzorec, *i* = *1*, … *a*
t … indeks za enoto znotraj vzorca, *t* = *1* … n_i

VII.5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah

- Analiza variance za preverjanje domnev o enakosti/razliki populacijskih aritmetičnih sredin torej temelji na primerjavi variabilnost med vzorčnimi aritmetičnimi sredinami z variabilnostjo posameznih enot znotraj vsakega vzorca – primerjamo variabilnost med skupinami z variabilnostjo znotraj skupin.
- Pristop je deduktiven:
- ✓ Če sta dve (ali več) populacijski aritmetični sredini enaki, potem moramo v vzorčnih podatkih zaznati naslednje: variabilnost med skupinami mora biti enaka variabilnosti znotraj skupin. Obe variabilnost namreč izhajata iz istega vira – vse skupine so iz iste populacije. V takem primeru so edini vir razlik med vzorčnimi aritmetičnimi sredinami (torej med skupinami) razlike med posameznimi enotami (ker vzorci vsebujejo različne enote), za katere pa predpostavljamo, da so slučajne. Torej je varianca med vzorci enaka varianci znotraj vzorcev.
- ✓ Če sta dve (ali več) populacijskih aritmetičnih sredin različni, pa moramo v vzorčnih podatkih zaznati naslednje: variabilnost med skupinami mora biti večja od variabilnosti znotraj skupin. Razlike med vzorčnimi aritmetičnimi sredinami (variabilnost med skupinami) so v tem primeru posledica dveh virov: razlik med posameznimi enotami znotraj vzorca in dejanske razlike med populacijskimi aritmetičnimi sredinami. V tem primeru bi bila varianca med vzorci večja od variance znotraj vzorcev.

VII.5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah

Mera, ki primerja variabilnost med vzorčnimi aritmetičnimi sredinami z variabilnostjo posameznih enot znotraj vsakega vzorca, torej variabilnost med skupinami z variabilnostjo znotraj skupin:

$$F = \frac{\text{varianca med vzorci/skupinami}}{\text{varianca znotraj vzorcev/skupin}} = \frac{\text{pojasnjenavarianca}}{\text{nepojasnjenavarianca}}$$

Osnovna ideja:

- Če med populacijskimi aritmetičnimi sredinami ni razlik, je varianca med skupinami enaka varianci znotraj skupin in je zgornje razmerje enako 1 ($F = 1$).
- Če med populacijskimi aritmetičnimi sredinami obstaja razlika, je varianca med skupinami večja od variance znotraj skupin, torej bi bilo zgornje razmerje večje od 1 ($F > 1$).

- Vprašanje: Kdaj je razmerje F zadosti večje od 1, da zaključimo, da je med vzorčnimi aritmetičnimi sredinami razlika, ki izhaja iz dejanskih razlik med populacijskimi aritmetičnimi sredinami?
- Odgovor: Naredimo formalen test – analizo variance.

VII.5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah

Dosedanje metode za preverjanje domnev o dveh aritmetičnih sredinah (predvsem t test za neodvisna vzorca) lahko razumemo kot poseben (poenostavljen primer) analize variance. Rezultati analize variance za 2 skupini bi bili popolnoma enaki analizi *t* testa za 2 neodvisna vzorca.

VII.5 Analiza variance

Kazalo:

5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah
5.2 Enofaktorska analiza variance v primeru enako velikih skupin
5.3 Enofaktorska analiza variance v primeru različno velikih skupin
5.4 Kako izvedemo analizo variance v SPSS-ju?
    5.4.1 Preverjanje predpostavk analize variance
    5.4.2 Izvedba analize variance
    5.4.3 Popravki analize variance v primeru kršitve predpostavk
5.5 Dodatni testi za razlike za pare skupin

VII.5.2 Enofaktorska analiza variance v primeru enako velikih skupin

Metoda za primerjanje razlik med aritmetičnimi sredinami v treh ali večih skupinah, pri čemer so vse skupine enako velike.

5.2.1 Predpostavke analize variance
5.2.2 Postopek analize variance
5.2.3 Primer

VII.5.2.1 Predpostavke analize variance

1. Normalnost (angl. normality): (Odvisna) spremenljivka, za katero računamo povprečje, se znotraj vsake populacije porazdeljuje normalno.
2. Enakost varianc (angl. homogeneity of variance): variabilnost posameznih enot znotraj vsake populacije je enaka ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2 = \sigma^2$).
3. Vsi vzorci so enako veliki ($n_1 = n_2 = ... = n_a = n$).

- Če je zadoščeno tem predpostavkama, lahko uporabimo klasičen F test (analizo variance) za testiranje domnev o razliki aritmetičnih sredin.
- Kršitev predpostavke o normalnosti: Če imamo velik vzorec (npr. večji od 30 v vsaki skupini), lahko uporabimo F test, tudi če predpostavki o normalnosti ni popolnoma zadoščeno. Pomembno je, da je porazdelitev v vseh populacijah približno enaka (npr. povsod enako asimetrična v desno). Pri majhnih vzorcih pa ne moremo uporabiti F testa.
- Kršitev predpostavke o enakosti varianc: Če so vzorci približno enako veliki, je ta predpostavka lahko kršena. V nasprotnem primeru pa uporabimo poseben popravek F testa.
- Kršitev predpostavke o enako velikih vzorcih: Če vzorci niso enako veliki, se uporabi poseben popravek F testa za neenako velike vzorce.

VII.5.2.2 Postopek analize variance

1. Postavimo domnevi:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ ... Ni razlik med populacijskimi aritmetičnimi sredinami.

H_1: $\mu_i \neq \mu_j$, $i \neq j$ .... So razlike med populacijskimi aritmetičnimi sredinami (vsaj dve aritmetični sredini se med seboj razlikujeta oz. vsaj ena aritmetična sredina se razlikuje od ostalih.)

VII.5.2.2 Postopek analize variance

2. Za dane vzorce izračunamo eksperimentalno vrednost testne statistike, ki je razmerje F (angl. F ratio, definiral jo je Sir Ronald Fisher):

$$\boxed{F = \frac{varianca\ med\ skupinami}{varianca\ znotraj\ skupin} = \frac{n \cdot s_{\bar{x}}^2}{s_p^2}}$$

a … število vzorcev/skupin
n … število enot v vzorcu ($n_1 = n_2 = \ldots n_i = n$)
i … indeks za vzorec, $i = 1, \ldots a$
t … indeks za enoto znotraj vzorca, $t = 1 \ldots n$

$s_{\bar{x}}^2 = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{X}_i - \bar{X})^2$ … varianca vzorčnih aritmetičnih sredin okoli skupne aritmetične sredine

$s_p^2 = (\frac{1}{a}\sum_{i=1}^{a} s_i^2 = \frac{1}{a(n\ \ -1)}\sum_{t=1}^{n}(x_{it} - \bar{X}_i)^2$ … povprečna varianca znotraj skupin

$F \approx 1$ ... Ni razlik med populacijskimi aritmetičnimi sredinami (vsi vzorci prihajajo iz enakih populacij oz. iz iste populacije).

$F > 1$ ... Varianca med vzorci je večja, torej so razlike med populacijskimi aritmetičnimi sredinami (vzorci prihajajo iz različnih populacij).

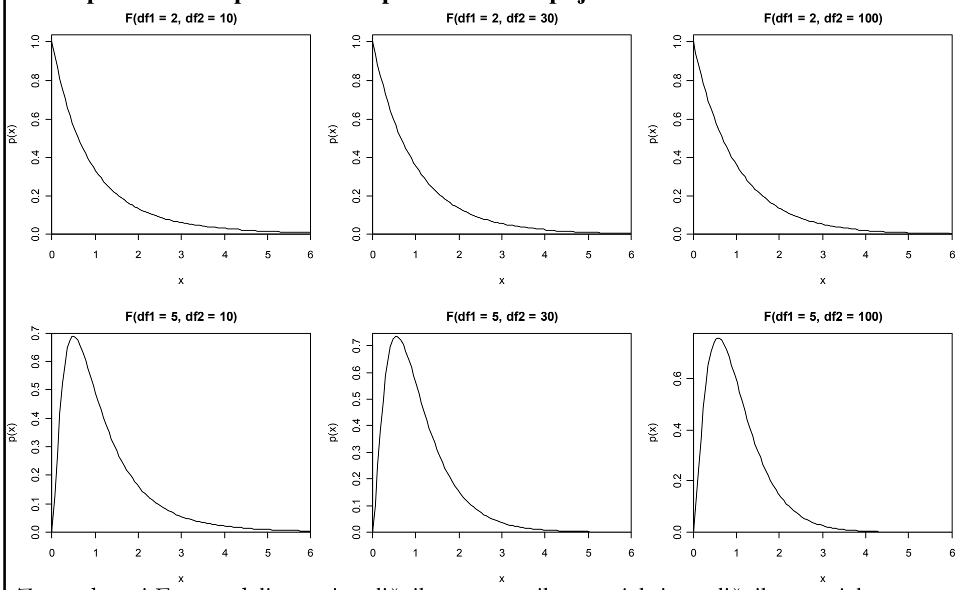
VII.5.2.2 Postopek analize variance

3. Za izračunano vrednost F na danih vzorcih določimo statistično značilnost (p) - kakšna je verjetnost, da dobimo F, ki je večji ali enak izračunanemu F ob predpostavki, da ni razlik med populacijskimi aritmetičnimi sredinami (torej da je pravilna ničelna domneva).

   Testna statistika F se porazdeljuje po F porazdelitvi, ki je asimetrična, njena oblika pa je odvisna od dvoje prostostnih stopenj: $m = (a-1)$ in $a(n-1)$.

a … število skupin/vzorcev
n … velikost vzorcev (vsi enako veliki)

F porazdelitev pri različnih prostostnih stopnjah



Za vrednosti F porazdelitve pri različnih prostostnih stopnjah in različnih stopnjah značilnosti glejte npr. http://www.statsoft.com/textbook/distribution-tables/#f

VII.5.2.2 Postopek analize variance

4. Sklep:

- ✓ Če je $p \leq \alpha$ (npr. $p \leq 0.05$), ničelno domnevo zavrnemo. Vsaj dve populacijski aritmetični sredini se med seboj statistično značilno razlikujeta pri stopnji značilnosti *p*. Vzorci/skupine torej izhajajo iz različnih populacij.

- ✓ Če je $p > \alpha$ (npr. $p > 0.05$), ničelne domneve ne moremo zavrniti. Populacijske aritmetične sredine se med seboj statistično značilno ne razlikujejo pri stopnji značilnosti *p*. Vzorci/skupine torej izhajajo iz iste populacije.

VII.5.2.3 Primer

- **Primer 1:** primerjava produktivnosti treh naprav
- Populacija: število ur delovanja naprave – za 3 naprave
- Enota: 1 ura delovanja
- Spremenljivka: Število proizvedenih izdelkov na uro (odvisna spremenljivka)
- Vzorec: slučajni vzorec $n=5$ ur – za tri naprave (neodvisna spremenljivka je „naprava“)

Naprava 1	Naprava 2	Naprava 3
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49
$\bar{X}_1 = 49$	$\bar{X}_2 = 56$	$\bar{X}_3 = 51$

Vprašanje: Ali so tri naprave različne?

Ali so tri vzorčne aritmetične sredine različne, ker so tri populacijske aritmetične sredine različne (t.j. povprečno dejansko število proizvedenih izdelkov preko celotne življenjske dobe naprave je različno za tri naprave)? Torej, ali so razlike v vzorčnih aritmetičnih sredinah zadosti velike, da jih lahko pripišemo dejanskim razlikam populacijskih aritmetičnih sredin?

Ali pa lahko razlike v vzorčnih aritmetičnih sredinah pripišemo slučajnemu nihanju v številu proizvedenih izdelkov na uro? Torej, da ni razlik med tremi napravami?

VII.5.2.3 Primer

Ali so razlike v vzorčnih aritmetičnih sredinah (skupin) zadosti velike, da sklepamo, da so posledica razlik v populacijskih aritmetičnih sredinah oz. da vzorci prihajajo iz različnih populacij?
ali pa …
Ali so razlike v vzorčnih aritmetičnih sredinah (skupin) zadosti majhne, da sklepamo, da ni razlik v populacijskih aritmetičnih sredinah oz. da vsi vzorci prihajajo iz iste populacije?

(1) Da bi lahko odgovorili na takšno vprašanje, je potrebno določiti, koliko se vzorčne aritmetične sredine med seboj razlikujejo oz. **koliko se vzorčne aritmetične sredine razlikujejo od skupne aritmetične sredine**. Zato izračunamo varianco za vzorčne aritmetične sredine. Ker gre za razlike med vzorčnimi aritmetičnimi sredinami, to variabilnost imenujemo **varianca (variabilnost, razlike) med vzorci/skupinami**.

Varianca med vzorci (between group variance)

	Naprava 1	Naprava 2	Naprava 3	
	47	55	54	
	53	54	50	
	49	58	51	
	50	61	51	
	46	52	49	
$\bar{X}_i$	$\bar{X}_1 = \frac{1}{n}\sum_{t=1}^{n} x_t = 49$	$\bar{X}_2 = \frac{1}{n}\sum_{t=1}^{n} x_t = 56$	$\bar{X}_3 = \frac{1}{n}\sum_{t=1}^{n} x_t = 51$	$\bar{X} = \frac{1}{a}\sum_{i=1}^{a} \bar{X}_i = 52$
$(\bar{X}_i - \bar{X})$	-3	4	-1	$\sum_{i=1}^{a}(\bar{X}_i - \bar{X}) = 0$
$(\bar{X}_i - \bar{X})^2$	9	16	1	$\sum_{i=1}^{a}(\bar{X}_i - \bar{X})^2 = 26$

a … število vzorcev/populacij; n … število enot v 1 vzorcu; *t* … indeks za enoto znotraj posameznega vzorca, *t=1, …, n*; *i* … indeks za vzorec, *i=1,…,a*

$$\boxed{\text{Varianca med vzorci: } s_{\bar{X}}^2 = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{X}_i - \bar{X})^2 = \frac{26}{2} = 13}$$

VII.5.2.3 Primer

Ali so razlike v vzorčnih aritmetičnih sredinah (skupin) zadosti velike, da sklepamo, da so posledica razlik v populacijskih aritmetičnih sredinah oz. da vzorci prihajajo iz različnih populacij?
ali pa …
Ali so razlike v vzorčnih aritmetičnih sredinah (skupin) zadosti majhne, da sklepamo, da ni razlik v populacijskih aritmetičnih sredinah oz. da vsi vzorci prihajajo iz iste populacije?

(1) Da bi lahko odgovorili na takšno vprašanje, je potrebno določiti, koliko se vzorčne aritmetične sredine med seboj razlikujejo oz. **koliko se vzorčne aritmetične sredine razlikujejo od skupne aritmetične sredine**. Zato izračunamo varianco za vzorčne aritmetične sredine. Ker gre za razlike med vzorčnimi aritmetičnimi sredinami, to variabilnost imenujemo **varianca (variabilnost, razlike) med vzorci/skupinami**.

(2) Dodatno pa je potrebno tudi določiti, kakšna je variabilnost znotraj posameznih vzorcev/skupin, torej **koliko se vrednosti posameznih enot znotraj vzorca razlikujejo od vzorčne aritmetične sredine**. Takšno variabilnost imenujemo **varianca (variabilnost, razlike) znotraj vzorcev/skupin**.

Varianca znotraj vzorcev (within group variance)

Zakaj je pomembna tudi variabilnost znotraj vzorcev?

Primer 1

Naprava 1	Naprava 2	Naprava 3
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49
$\overline{X}_1 = \frac{1}{n}\sum_{t=1}^{n} x_t = 49$	$\overline{X}_2 = \frac{1}{n}\sum_{t=1}^{n} x_t = 56$	$\overline{X}_3 = \frac{1}{n}\sum_{t=1}^{n} x_t = 51$
$s_1^2 = 7.5$	$s_2^2 = 12.5$	$s_3^2 = 3.5$
$s_1 = 2.7$	$s_2 = 3.5$	$s_3 = 1.9$

Primer 2

Naprava 1	Naprava 2	Naprava 3
50	48	57
42	57	59
53	65	48
45	59	46
55	51	45
$\overline{X}_1 = \frac{1}{n}\sum_{t=1}^{n} x_t = 49$	$\overline{X}_2 = \frac{1}{n}\sum_{t=1}^{n} x_t = 56$	$\overline{X}_3 = \frac{1}{n}\sum_{t=1}^{n} x_t = 51$
$s_1^2 = 29.5$	$s_2^2 = 45$	$s_3^2 = 42.5$
$s_1 = 5.4$	$s_2 = 6.7$	$s_3 = 6.5$

Vzorčne aritmetične sredine in varianca med vzorci je enaka v Primeru 1 in Primeru 2, vendar so nihanja v produktivnosti od ure do ure bistveno večja (večja je torej varianca posameznih vrednosti znotraj vzorca) v Primeru 2.

Primer 1: Zaradi majhne variance znotraj vzorcev lahko sklepamo, da so razlike med aritmetičnimi sredinami posledica razlik med tremi populacijami.

Primer 2: Zaradi velike variance znotraj vzorcev ne moremo zaključiti, da je razlika med vzorci posledica razlik med populacijami, saj bi takšne razlike med vzorci lahko dobili tudi, če bi bili vsi vzorci vzeti iz iste populacije, kjer so nihanja med urnimi vrednostmi (torej varianca enot) velika.

Varianca znotraj vzorcev (within group variance)

Naprava 1	Naprava 2	Naprava 3
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49
$\overline{X}_1 = \frac{1}{n}\sum_{t=1}^{n} x_t = 49$	$\overline{X}_2 = \frac{1}{n}\sum_{t=1}^{n} x_t = 56$	$\overline{X}_3 = \frac{1}{n}\sum_{t=1}^{n} x_t = 51$
$s_1^2 = 7.5$	$s_2^2 = 12.5$	$s_3^2 = 3.5$
$s_1 = 2.7$	$s_2 = 3.5$	$s_3 = 1.9$

$$s_p^2 = \frac{1}{3}(7.5 + 12.5 + 3.5) = 7.83$$

Varianca znotraj enega vzorca: $s_i^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \overline{X}_i)^2$

(Povprečna) varianca znotraj vzorcev (angl. pooled variance):

$$s_p^2 = \frac{1}{a}\sum_{i=1}^{a} s_i^2 = \frac{1}{a(n-1)}\sum_{t=1}^{n}(x_{it} - \overline{X}_i)^2$$

ANOVA tabela

ANOVA – Analysis of Variance

Vmesne izračune pri analizi variance se navadno zapiše v obliki t.im. ANOVA tabele: v prvo vrstico zapišemo izračune za varianco med vzorci (števec pri F statistiki, between group variance), v drugo vrstico pa izračune za varianco znotraj vzorcev (within group variance, imenovalec pri F statistiki).

Vir variabilnosti	Vsota kvadratov (Sum of Squares, SS)	Prost. stopnje	Povprečni kvadrat (Mean Square, MS)	Razmerje F
FAKTOR A: Razlike med vzorčnimi aritmetičnimi sredinami $\overline{X}_i$ in skupno aritmetično sredino $\overline{X}$ (variabilnost med vzorci)	$SS_A = n\sum_{i=1}^{a}(\overline{X}_i - \overline{X})^2$	$(a-1)$	$MS_A = \frac{SS_A}{a-1} = ns_{\overline{X}}^2$	$F = \frac{MS_A}{MS_E} = \frac{ns_{\overline{X}}^2}{s_p^2}$
REZIDUAL (NAPAKA) (angl. Error): Razlike med posameznimi vrednostmi x_{it} in vzorčno aritmetično sredino $\overline{X}_i$ (variabilnost znotraj vzorcev)	$SS_E = \sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it} - \overline{X}_i)^2$	$a(n-1)$	$MS_E = \frac{SS_E}{a(n-1)} = s_p^2$	
SKUPAJ Razlike med posameznimi vrednostmi x_{it} in skupno aritmetično sredino $\overline{X}$ (celotna variabilnost)	$SS = \sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it} - \overline{X})^2 = SS_A + SS_E$	$na-1$		

ANOVA tabela ANOVA – Analysis of Variance

Vmesne izračune pri analizi variance se navadno zapiše v obliki t.im. ANOVA tabele: v prvo vrstico zapišemo izračune za varianco med vzorci (števec pri F statistiki, between group variance), v drugo vrstico pa izračune za varianco znotraj vzorcev (within group variance, imenovalec pri F statistiki).

Razlike med skupinami glede na neodvisno spremenljivko (faktor) A – razlike, pojasnjene s pripadnostjo skupini (pojasnjena variabilnost)

Razlike znotraj skupin – slučajne razlike, posledica vzorčenja (nepojasnjena variabilnost)

Vir variabilnosti	Vsota kvadratov (Sum of Squares, SS)	Prost. stopnje	Povprečni kvadrat (Mean Square, MS)	Razmerje F
FAKTOR A: Razlike med vzorčnimi aritmetičnimi sredinami $\overline{X}_i$ in skupno aritmetično sredino $\overline{X}$ (variabilnost med vzorci)	$SS_A = n\sum_{i=1}^{a}(\overline{X}_i - \overline{X})^2$	$(a-1)$	$MS_A = \frac{SS_A}{a-1} = ns_{\overline{X}}^2$	$F = \frac{MS_A}{MS_E} = \frac{ns_{\overline{X}}^2}{s_p^2}$
REZIDUAL (NAPAKA) (angl. Error): Razlike med posameznimi vrednostmi x_{it} in vzorčno aritmetično sredino $\overline{X}_i$ (variabilnost znotraj vzorcev)	$SS_E = \sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it} - \overline{X}_i)^2$	$a(n-1)$	$MS_E = \frac{SS_E}{a(n-1)} = s_p^2$	
SKUPAJ Razlike med posameznimi vrednostmi x_{it} in skupno aritmetično sredino $\overline{X}$ (celotna variabilnost)	$SS = \sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it} - \overline{X})^2 = SS_A + SS_E$	$na-1$		

ANOVA tabela za Primer 1

Vir variabilnosti	SS	Prost. stopnje	MS	Razmerje F	p vrednost (stat. značilnost)
NAPRAVE	130 (5x26)	2 (3-1)	65 (130/2 oz. 5x13)	65/78.3=8.3	p<0.01
REZIDUAL (NAPAKA)	94 (30+50+14)	12 (3(5-1))	7.83 (94/12)		
SKUPAJ	224 (130+94)	14 (5x3-1)			

Verjetnost, da bi dobili vrednost F statistike 8.3 pri pravilni ničelni domnevi (ni razlik), je majhna (p<0.01), zato podvomimo v pravilnost ničelne domneve. Tri naprave so različne.

ANOVA tabela za Primer 1 …. izpis iz SPSS

ANOVA

stizdelk Število proizvedenih izdelkov

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	130,000	2	65,000	8,298	,005
Within Groups	94,000	12	7,833		
Total	224,000	14			

Skupna variabilnost = varianca med vzorci + varianca znotraj vzorcev (between group variance)

Naprava 1	Naprava 2	Naprava 3
47	55	54
53	54	50
49	58	51
50	61	51
46	52	49

$$\overline{X} = \frac{1}{a}\sum_{i=1}^{a}\overline{X}_i = \frac{49+56+51}{3} = 52$$

$oz.$

$$\overline{X} = \frac{1}{na}\sum_{i=1}^{a}\sum_{t=1}^{n} x_{it} = \frac{47+53+...+49}{5*3} = 52$$

$$s^2 = \frac{1}{na-1}\sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it}-\overline{X})^2 = \frac{1}{5*3-1}((47-52)^2+(53-52)^2+...+(49-52)^2) = \frac{224}{14} = 16$$

$$SS = SS_A + SS_B = 130 + 94 = 224$$

$$F = \frac{ns_{\overline{X}}^2}{s_p^2} = \frac{pojasnjena \text{ varianca}}{nepojasnjena \text{ varianca}}$$

VII.5 Analiza variance

Kazalo:

5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah
5.2 Enofaktorska analiza variance v primeru enako velikih skupin
5.3 Enofaktorska analiza variance v primeru različno velikih skupin
5.4 Kako izvedemo analizo variance v SPSS-ju?
    5.4.1 Preverjanje predpostavk analize variance
    5.4.2 Izvedba analize variance
    5.4.3 Popravki analize variance v primeru kršitve predpostavk
5.5 Dodatni testi za razlike za pare skupin

VII.5.3 Enofaktorska analiza variance v primeru različno velikih skupin

Ena od predpostavk običajne analize variance: Vsi vzorci so enako veliki ($n_1 = n_2 = \ldots = n_a = n$).

Kadar pa vzorci niso enako veliki, uporabimo ustrezen popravek izračunov tako, da upoštevamo različne velikosti vzorcev $n_1, n_2, \ldots, n_a$. Npr. skupna vzorčna aritmetična sredina ni enostavno povprečje treh vzorčnih aritmetičnih sredin, pač pa njihovo uteženo povprečje (k skupnemu povprečju več prispevajo tista vzorčna povprečja, ki pripadajo večjih vzorcem).

VII 5.3.1 Anova tabela v primeru različno velikih vzorcev

Vir variabilnosti	Vsota kvadratov (Sum of Squares, SS)	Prost. stopnje	Povprečni kvadrat (Mean Square, MS)	Razmerje F
FAKTOR A: Razlike med vzorčnimi aritmetičnimi sredinami $\overline{X}_i$ in skupno aritmetično sredino $\overline{X}$ (variabilnost med vzorci)	$SS_A = \sum_{i=1}^{a} n_i(\overline{X}_i - \overline{X})^2$	$(a-1)$	$MS_A = \frac{SS_A}{a-1}$	
REZIDUAL (NAPAKA) (angl. Error): Razlike med posameznimi vrednostmi x_{it} in vzorčno aritmetično sredino $\overline{X}_i$ (variabilnost znotraj vzorcev)	$SS_A = = \sum_{i=1}^{a} n_i(\overline{X}_i - \overline{X})^2 = 2 = \sum_{i=1}^{a}\sum_{t=1}^{n_i}(\overline{X}_i - \overline{X})^2$	$\sum_{i=1}^{a}(n_i - 1)$	$MS_E = \frac{SS_E}{\sum_{i=1}^{a}(n_i - 1)}$	$F = \frac{MS_A}{MS_E}$
SKUPAJ Razlike med posameznimi vrednostmi x_{it} in skupno aritmetično sredino $\overline{X}$ (celotna variabilnost)	$SS = \sum_{i=1}^{a}\sum_{t=1}^{n}(x_{it} - \overline{X})^2 = SS_A + SS_E$	$\sum_{i=1}^{a} n_i - 1$		

kjer je skupna vzorčna aritmetična sredina: $\overline{X} = \frac{\sum_{i=1}^{a}\sum_{t=1}^{n_i} x_{it}}{\sum_{i=1}^{a} n_i} = \frac{\sum_{i=1}^{a} n_i \overline{X}_i}{\sum_{i=1}^{a} n_i}$

VII 5.3.2 Primer

- **Primer 2:** Primerjava povprečna bruto plače glede na 3 stopnje izobrazbe
- Odvisna spremenljivka: Bruto plača
- Enota: zaposlena oseba
- Vzorec: slučajni vzorec 15 zaposlenih oseb

	1 (ne)dokončana osnovna šola	2 2-4 letna srednja šola ali višja šola	3 visoka šola ali več
	100	180	450
	170	87	500
	220	170	300
	169	400	190
		160	350
		170	
n_i	4	6	5
$\overline{x}_i$	164,75	194,5	358

Vprašanje: Ali je povprečna plača odvisna od stopnje izobrazbe?

Ali so razlike v vzorčnih povprečnih plačah zadosti velike, da jih lahko pripišemo dejanskim razlikam v povprečnih plačah na treh populacijah glede na stopnjo izobrazbe?

Ali pa lahko razlike v vzorčnih povprečnih plačah pripišemo slučajnemu nihanju v bruto plačah med posamezniki (so torej zgolj slučajne, posledica vzorčenja) in torej ni razlik v povprečnih plačah na treh populacijah glede na stopnjo izobrazbe? izobrazbe?

Varianca med vzorci

	1 (ne)dokončana osnovna šola	2 2-4 letna srednja šola ali višja šola	3 visoka šola ali več	
	100	180	450	
	170	87	500	
	220	170	300	
	169	400	190	
		160	350	
		170		
n_i	4	6	5	$n=\sum_{i=1}^{3} n_i$
$\bar{x}_i$	$\bar{x}_1=\frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}=164{,}75$	$\bar{x}_2=\frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}=194{,}50$	$\bar{x}_3=\frac{1}{n_3}\sum_{j=1}^{n_3} x_{3j}=358{,}00$	$\bar{x}=\frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{n_i} x_{ij}=241{,}07$
$(\bar{x}_i-\bar{x})^2$	5824,23	2168,45	13673,4	
$n_i(\bar{x}_i-\bar{x})^2$	23296,93	13010,73	68367,02	$\sum_{i=1}^{3} n_i(\bar{x}_i-\bar{x})^2 = 104674{,}68$

a je število vzorcev/populacij, i označuje vzorec, n_i je število enot v vzorcu i, j je indeks za enote znotraj vzorcev

Varianca med vzorci: $MS_A=\frac{1}{a-1}\sum_{i=1}^{a} n_i(\bar{x}_i-\bar{x})^2=\frac{104674{,}68}{2}=52337{,}34$

Varianca znotraj vzorcev

	1 (ne)dokončana osnovna šola	2 2-4 letna srednja šola ali višja šola	3 visoka šola ali več
	100	180	450
	170	87	500
	220	170	300
	169	400	190
		160	350
		170	
n_i	4	6	5
$\bar{x}_i$	164,75	194,50	358,00
$\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$	7290,75	56387,50	60280,00

Vsota kvadratov znotraj vzorev: $SS_E=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2=123958{,}25$

Povprečna (pooled) varianca znotraj vzorcev : $s_p^2=\frac{SS_E}{\sum_{i=1}^{3}(n_i-1)}=\frac{SS_E}{n-a}=\frac{123958{,}25}{15-3}=10329{,}85$

Anova tabela

variabilnost med skupinami variabilnost znotraj skupin

Vir variabilnosti	SS	Stopinje prostosti	MS	F	p (statistična značilnost)
STOPNJA IZOBRAZBE	104674,68	2 (3 - 1)	52337,34 (104674,68 / 2)	5,07 (52337,34 / 10329,85)	0,025
REZIDUAL (NAPAKA)	123958,25	12 (15 - 3)	10329,85 (123958,25 / 12)		
SKUPAJ	228632,93	14 (15 - 1)			

Verjetnost, da bi dobili vrednost F statistike 5,07 pri pravilni ničelni domnevi (ni razlik), je majhna (p = 0,025), zato zavrnemo ničelno domnevo. Tako lahko trdimo, da stopnja izobrazbe vpliva na bruto plačo oziroma da povprečna bruto plača ni enaka v vseh skupinah zaposlenih glede na stopnjo izobrazbe. Vsaj ena skupina se statistično značilno razlikuje od ostalih.

Anova tabela … izpis iz SPSS

ANOVA

G91 Bruto plača v 1000 sit

	Vsota kvadratov (Sum of Squares)	Stopinje prostosti (df)	Povprečni kvadratni odklon (Mean Square)	F	Stopnja značilnosti (Sig.)
Med skupinami (Between Groups)	104674,683	2	52337,342	5,067	,025
Znotraj skupin (Within Groups)	123958,250	12	10329,854		
Skupaj	228632,933	14			

VII.5 Analiza variance

Kazalo:

5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah
5.2 Enofaktorska analiza variance v primeru enako velikih skupin
5.3 Enofaktorska analiza variance v primeru različno velikih skupin
5.4 Kako izvedemo analizo variance v SPSS-ju?
    5.4.1 Preverjanje predpostavk analize variance
    5.4.2 Izvedba analize variance
    5.4.3 Popravki analize variance v primeru kršitve predpostavk
5.5 Dodatni testi za razlike za pare skupin

VII.5.4 Kako izvedemo analizo variance v SPSS-ju?

Analyze – Compare means – One-way ANOVA

Kdaj uporabimo: 1 neodvisna spremenljivka (faktor), ki določa skupine, in 1 odvisna spremenljivka, za katero računamo povprečje.

Enak postopek za enako ali za različno velike vzorce.

Postopek:

1) Izvedemo analizo, s katero preverimo predpostavke ANOVE
   Procedura *Analysis – Descriptive Statistics – Explore*

2) Izvedemo ANOVO, ki vključuje test značilnosti za razlike aritmetičnih sredin
   Procedura *Analysis – Compare Means – One-Way Anova*

VII.5.4 Kako izvedemo analizo variance v SPSS-ju?

PRIMER: Preverjanje domneve o razliki v povprečni bruto plači glede na kraj bivanja

- Datoteka: EES 2010 za Slovenijo
- Enota analize: 1 anketiranec
- Odvisna spremenljivka: mesečna bruto plača v EUR („brutoplaca“)
- Neodvisna spremenljivka (faktor): Kraj „Kraj bivanja“
- Izvedemo analizo, s katero preverimo predpostavke ANOVE

  Procedura *Analysis – Descriptive Statistics – Explore*
- Izvedemo test značilnosti aritmetičnih sredin (ANOVO)

  Procedura *Analysis – Compare Means – One-Way Anova*

VII.5.4.1 Preverjanje predpostavk

Preden izvedemo analizo variance (oz. katerokoli statistično analizo), je smiselno najprej pregledati in preveriti podatke, ki jih bomo uporabili. Več razlogov za to:

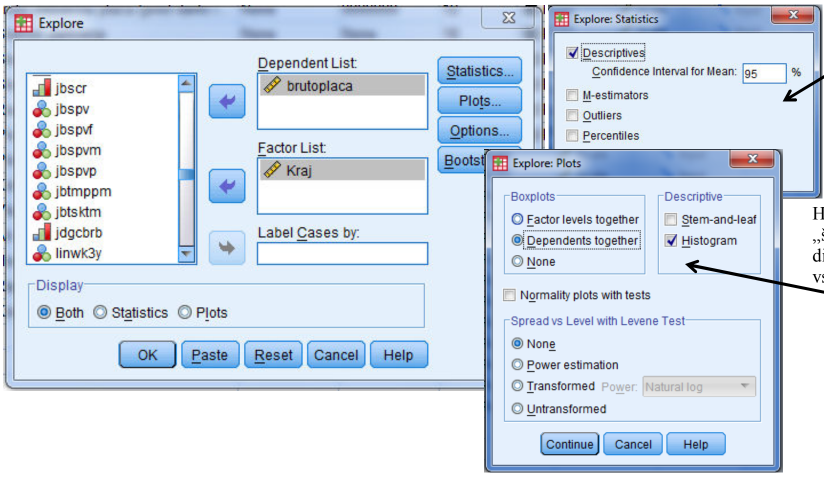
- identificiramo morebitne napake v podatkih,
- identificiramo odstopajoče enote (angl. *outliers*),
- spoznamo se z morebitnimi nepričakovanimi ali zanimivimi vzorci v podatkih,
- preverimo predpostavke načrtovane analize. Za ANOVO to pomeni:
  - preverimo predpostavko o homogenosti (enakosti) varianc,
  - preverimo predpostavko o normalnosti porazdelitve znotraj vsake populacije.

Analiza, ki jo izvedemo z namenom pregleda podatkov, imenujemo „**eksplorativna analiza podatkov**“ (angl. exploratory data analysis).

Procedura v SPSS-ju: *Analyze – Decriptive Statistics – Explore*.

VII.5.4.1 Preverjanje predpostavk

- *Analysis – Descriptive Statistics – Explore*
- Opravimo pregled porazdelitve vrednosti in opisnih statistik spremenljivke „bruto plača“ (odvisna spremenljivka) glede na vrednosti spremenljivke „kraj bivanja“ (neodvisna spremenljivka oz. faktor).



Opisne statistike po skupinah

Histogram in „škatlasti diagram“ za vsako skupino

Podatki o veljavnih in neveljavnih vrednostih (velikosti vzorcev)

Case Processing Summary

	Kraj Kraj bivanja	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
brutoplaca bruto mesečna plača (pred davki in prispevki)	1 veliko mesto (vključno s predmestjem)	116	30,9%	259	69,1%	375	100,0%
	2 malo mesto	70	27,0%	189	73,0%	259	100,0%
	3 podeželje	169	22,1%	597	77,9%	766	100,0%

Analiza je opravljena na 116 enotah iz velikega mesta, 70 enotah iz malega mesta in 165 enotah iz velikega mesta. Ostali anketiranci pa imajo manjkajoče vrednosti, večinoma zato, ker nimajo ali niso navedli bruto plače ali pa jim to vprašanje sploh ni bilo zastavljeno (zastavljeno je bilo le tistim, ki so v zadnjih 7 dneh opravljali delo za plačilo).

Ker so vzorci različno veliki, bo v nadaljevanju opravljena analiza variance z ustreznim popravkom za neenako velike vzorce (SPSS sam to izvede brez kakšnih dodatnih uporabnikovih zahtev).

Opisne statistike za bruto plačo po krajih bivanja

Descriptives

			Kraj Kraj bivanja					
			1 veliko mesto (vključno s predmestjem)		2 malo mesto		3 podeželje	
			Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
brutoplaca bruto mesečna plača (pred davki in prispevki)	Mean		1390,85	68,617	1768,10	152,921	1295,34	53,937
	95% Confidence Interval for Mean	Lower Bound	1254,94		1463,03		1188,86	
		Upper Bound	1526,77		2073,17		1401,82	
	5% Trimmed Mean		1344,28		1613,13		1239,93	
	Median		1200,00		1500,00		1050,00	
	Variance		546155,361		1636931,541		491649,284	
	Std. Deviation		739,023		1279,426		701,177	
	Minimum		60		0		200	
	Maximum		4000		7500		4500	
	Range		3940		7500		4300	
	Interquartile Range		1100		1091		863	
	Skewness		1,034	,225	2,328	,287	1,400	,187
	Kurtosis		1,236	,446	6,670	,566	2,399	,371

Opisne statistike za bruto plačo po krajih bivanja: interpretacija

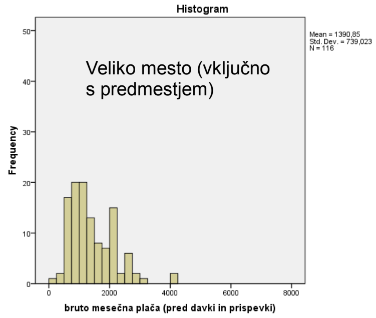
Vzorčna povprečja (*Mean*) se po skupinah razlikujejo: v povprečju imajo najvišjo bruto plačo (1768 EUR) v malih mestih, najnižjo (1295 EUR) pa na podeželju. V velikem mestu je le nekoliko višja kot na podeželju (1391 EUR).
Če iz analize izločimo 5% enot (*5% Trimmed Mean*), ki imajo najbolj ekstremne vrednosti, se razlike v vzorčnih povprečnih bruto plačah med skupinami zmanjšajo, vendar ostaja prepoznaven zgornji vzorec.

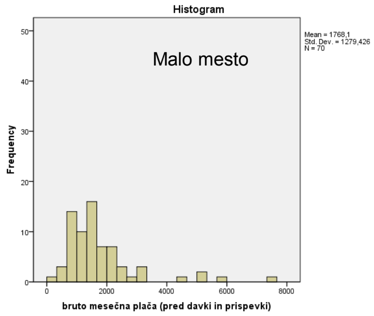
Predpostavka o homogenosti varianc: variance in iz njih izračunani standardni odkloni (*Variance* in *Std. Deviation*) se med skupinami precej razlikujejo – predvsem v malih mestih (st. odklon je 1279 EUR) so razlike med anketiranci mnogo večje kot v drugih dveh skupinah (st. odklon 701 EUR na podeželju in 739 EUR v velikih mestih). Uporabiti bo potrebno popravek F testa za nehomogene variance.

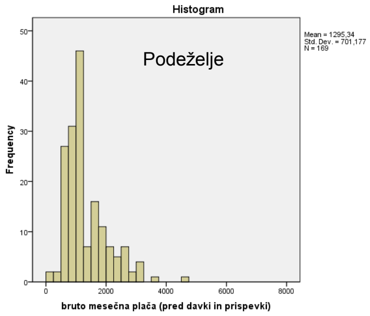
Predpostavka o normalnosti porazdelitve znotraj vsake skupine: Porazdelitev spremenljivke „bruto plača“ ni normalna, pač pa asimetrična v desno in koničasta (koeficient asimetrije (*skewness*) so v vseh treh skupinah pozitivni in veliki (večji od 2-kratne vrednosti svoje standardne napake), prav tako koeficienti sploščenosti/koničavosti (*kurtosis*)). Predpsotavka o normalnosti je torej kršena. Vendar, ker imamo velike vzorce in ker je porazdelitev spremenljive v vseh treh skupinah približno enaka, vseeno lahko opravimo analizo variance.

Grafična predstavitev porazdelitve spremenljivke „bruto plača“ za vsako skupino s histogrami

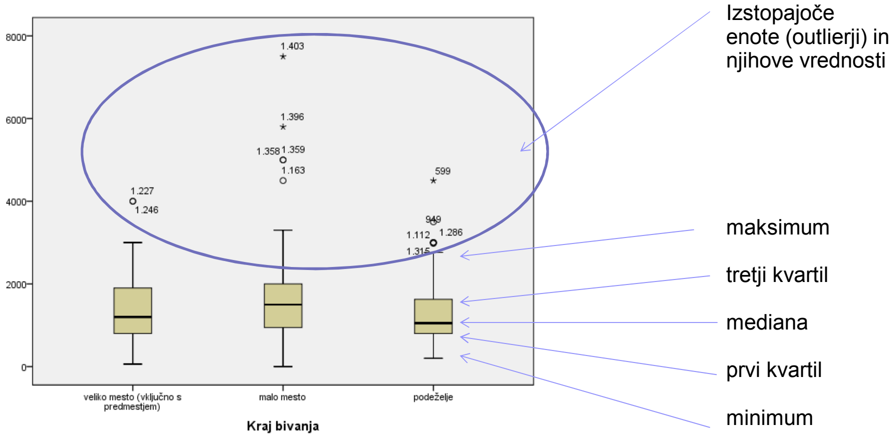
V vseh treh skupinah je porazdelitev asimetrična v desno, variabilnost je največja za anketirance iz malih mest.







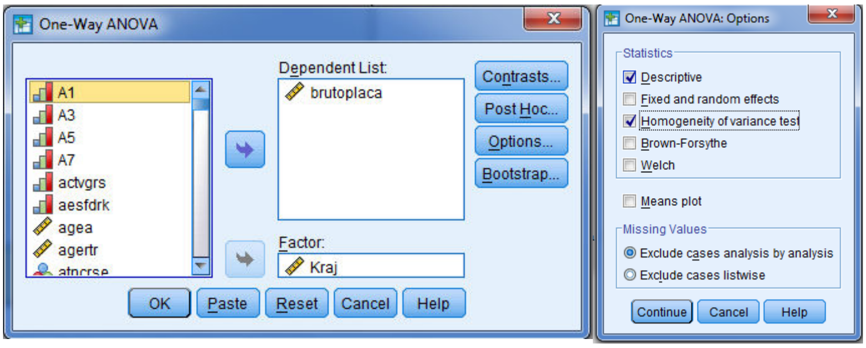
Grafična predstavitev porazdelitve spremenljivke „bruto plača“ za vsako skupino s škatlastimi diagrami



Obstajajo neke razlike med medianami (v mestu je mediana višja), vendar se porazdelitve v treh skupinah precej prekrivajo. Variabilnost med skupinami (med medianami oz. med aritmetičnimi sredinami) zgleda majhna v primerjavi z variabilnostjo znotraj skupin.

VII.5.4.2 Izvedba analize variance (ANOVA): test značilnosti za razlike aritmetičnih sredin

- *Analysis – Compare Means – One-Way ANOVA*. Zahtevamo izvedbo testa značilnosti za razliko v aritmetičnih sredinah spremenljivke „bruto plača“ (odvisna spremenljivka) po skupinah glede na spremenljivko „kraj bivanja“ (neodvisna spremenljivka oz. faktor).
- Dodatne zahteve - gumb *Options*: po želji zahtevamo še enkrat izpis opisnih statistik (*Descriptives*), nujno pa zahtevamo test homogenosti varianc (*Homogeneity of variance test*).



Rezultati ANOVE: Opisne statistike

Descriptives

brutoplaca bruto mesečna plača (pred davki in prispevki)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 veliko mesto (vključno s predmestjem)	116	1390,85	739,023	68,617	1254,94	1526,77	60	4000
2 malo mesto	70	1768,10	1279,426	152,921	1463,03	2073,17	0	7500
3 podeželje	169	1295,34	701,177	53,937	1188,86	1401,82	200	4500
Total	355	1419,77	872,622	46,314	1328,68	1510,85	0	7500

- Interpretacija ista kot zgoraj.

Rezultati ANOVE: Preverjanje predpostavk o homogenosti (enakosti) varianc

Test of Homogeneity of Variances

brutoplaca bruto mesečna plača (pred davki in prispevki)

Levene Statistic	df1	df2	Sig.
6,742	2	352	,001

PROBLEM NEENAKIH VARIANC:

- Ena izmed predpostavk ANOVE je tudi enakost varianc med skupinami:

  Enakost varianc (angl. homogeneity of variance): variabilnost posameznih enot znotraj vsake populacije je enaka ($\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2 = \sigma^2$).

- Populacijskih varianc seveda ne poznamo, iz izpisa vzorčnih varianc (prejšnja tabela) pa lahko vidimo, ali je variabilnost znotraj skupin (torej vzorčna varianca) približno enaka po skupinah.
- Natančneje to določimo z Levene-vim testom; test, s katerim preverjamo predpostavko o enakosti populacijskih varianc (poznamo ga že iz procedure za *t* test za neodvisna vzorca).

Preverjanje predpostavk o homogenosti (enakosti) varianc: Levenev test

- Domnevi:

  H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_a^2$ Ni razlik med populacijskimi variancami.

  H_1: $\sigma_i^2 \neq \sigma_2^2$, $i \neq j$ So razlike med populacijskimi variancami (vsaj dve populacijski varianci se med seboj razlikujeta oz. vsaj ena populacijska varianca se razlikuje od ostalih.)

- Levenov test je pravzaprav ANOVA na novi spremenljivki *Z*, ki za vsako enoto meri absolutni odklon (originalne) spremenljivke od povprečja skupine, ki jih enota pripada. → $z_{it} = |x_{it} - \bar{X}_i|$ (*z* vrednost za enoto *t* iz skupine *j*).
- Če so povprečja te spremenljivke *Z* enaka v vseh skupinah, pomeni, da so odkloni vrednosti od povprečja (torej variabilnost) enaki.
- Uporabi se torej *F* test kot za ANOVO, pri čemer preverjamo domnevo o enakosti povprečnih absolutnih odklonov od povprečja po skupinah.
- Levene-v test je splošno uporaben, ker zanj ni potrebna predpostavka o normalnosti porazdelitve spremenljivke po skupinah.

Rezultati ANOVE: Preverjanje predpostavk o homogenosti (enakosti) varianc

Test of Homogeneity of Variances

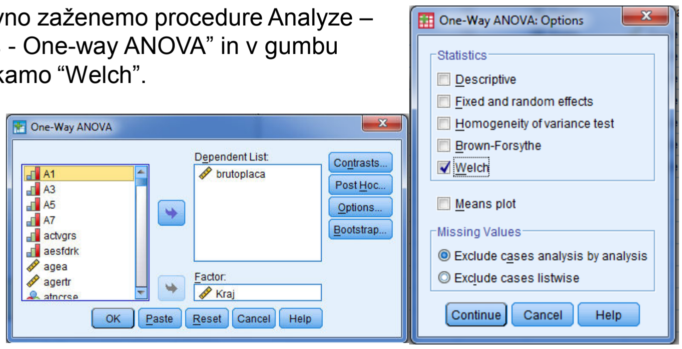
brutoplaca bruto mesečna plača (pred davki in prispevki)

Levene Statistic	df1	df2	Sig.
6,742	2	352	,001

- Domnevi:
  H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ Variance v vseh treh skupinah so enake.
  H_1: $\sigma_i^2 \neq \sigma_2^2,\ i \neq j$ Vsaj v eni skupini se varianca razlikuje.
- Testna statistika (Leveneva statistika) ima vrednost *6,742*, porazdeljuje se po *F* porazdelitvi s prostostnimi stopnjami *2* in *352*, njena statistična značilnost pa je $p = 0.001$.
- Domnevo o enakosti varianc moramo zavrniti ob stopnji značilnost $p = 0.001$.
- Ker so populacijske variance različne, moramo uporabiti popravek običajnega F testa za razliko povprečij.

VII.5.4.3 Popravki analize variance v primeru kršitve predpostavk

- Če z Levene-vin testom ugotovimo, da je predpostavka o enakosti varianc kršena, moramo (podobno kot pri *t* testu za neodvisna vzorca) uporabiti popravek običajnega F testa.
- Uporabi se Welch-ova *F* statistika, ki upošteva velikosti varianc in skupin tako pri izračunu *F* statistike kot tudi pri izračunu prostostnih stopenj. (Test je soroden (oba je predlagal tudi isti avtor) različici *t*-testa za neodvisne vzorce za neenake variance.)
- V SPSS-u ponovno zaženemo procedure Analyze – Compare Means - One-way ANOVA” in v gumbu “Options” odkljukamo “Welch”.



Rezultati ANOVE: Rezultati običajne ANOVA tabele

ANOVA

brutoplaca bruto mesečna plača (pred davki in prispevki)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1,121E7	2	5603536,238	7,635	,001
Within Groups	2,584E8	352	733958,019		
Total	2,696E8	354			

- Običajna ANOVA tabela: iz nje lahko razberemo, kakšna je variabilnost med skupinami in kakšna je variabilnost znotraj skupin.
- Običajna F statistika v tem primeru ni primerna, ker je kršena predpostavka o enakosti varianc. Uporabimo popravek – Welch F statistiko.

Rezultati ANOVE: Rezultati popravka F testa – Welch F statistika

Robust Tests of Equality of Means

brutoplaca bruto mesečna plača (pred davki in prispevki)

	Statistic[a]	df1	df2	Sig.
Welch	4,359	2	155,047	,014

a. Asymptotically F distributed.

- Domnevi:

  H_0: $\mu_1^2 = \mu_2^2 = \mu_3^2$ Ni razlik med populacijskimi aritmetičnimi sredinami.

  H_1: $\mu_i^2 \neq \mu_2^2,\ i \neq j$ So razlike med populacijskimi aritmetičnimi sredinami (vsaj dve populacijski aritmetični sredini se med seboj razlikujeta oz. vsaj ena populacijska aritmetična sredina se razlikuje od ostalih.)
- Welch F ima vrednost *4,359*, porazdeljuje pa se po F porazdelitvi (s popravljenimi prostostnimi stopnjami, upoštevajoč neenakost skupin in varianc). Njena statistična značilnost je $p = 0{,}014$.
- Ker je statistična značilnost majhna, zavrnemo domnevo o enakosti aritmetičnih sredin. Vsaj v eni skupini se populacijska aritmetična sredina statistično značilno (pri stopnji značilnosti p = 0.014) razlikuje od ostalih dveh.
- Kateri aritmetična sredina je razlika od ostalih?

VII.5 Analiza variance

Kazalo:

5.1 Logika preverjanja domnev o razlikah v aritmetičnih sredinah
5.2 Enofaktorska analiza variance v primeru enako velikih skupin
5.3 Enofaktorska analiza variance v primeru različno velikih skupin
5.4 Kako izvedemo analizo variance v SPSS-ju?
  5.4.1 Preverjanje predpostavk analize variance
  5.4.2 Izvedba analize variance
  5.4.3 Popravki analize variance v primeru kršitve predpostavk
5.5 Dodatni testi za razlike za pare skupin

VII.5.5 Dodatni testi za razlike za pare skupin

- Na prejšnjem primeru smo ugotovili, da vsa povprečja niso enaka.
- Pojavi se nam vprašanje, katera so različna.
- V ta namen lahko uporabimo dopolnilne teste „Post hoc tests". (Opravimo jih šele po tem, ko F test pokaže, da obstajajo neke razlike med populacijskimi aritmetičnimi sredinami.)
- Ti testi so običajno narejeni na osnovi nam znanega *t* testa za neodvisni skupini. Pri izračunu statistične značilnosti pa upoštevajo, da je verjetnost α, da naredimo napako prve vrste (verjetnost, da se zmotimo, ko zavrnemo ničelne domnevo) večja, ker istočasno opravimo več testov (en test za vsak možni par skupin).
- Post-hoc testov je več; ni enega testa, ki bi bil primeren v vseh oz. večini situacij. (Dober test bi kontroliral verjetnost napake prve vrste, bi bil zmožen zaznati prave populacijske razlike in bil robusten na kršitve predpostavk o homogenosti varianc in normalnosti porazdelitve.)

VII.5.5 Dodatni testi za razlike za pare skupin

- Teste lahko razdelimo v dve skupini glede na to, ali predpostavljamo enake variance med skupinami ali ne (enakost varianc smo predhodno že preverili z Levene-vim testom o homogenesti varianc). Med njimi ločimo še na tiste, ki so primerni, kadar imamo enake ali kadar imamo različne velike skupine/vzorce.
- Znotraj vsake skupine predstavljamo izbor nekaterih testov, urejenih od najbolj liberalnih (tisti, ki prej (pri manjših vzorcih) zaznajo populacijske razlike, torej imajo večjo statistično moč, imajo pa večjo verjetnost napake 1. vrste, torej da zavrnemo pravilno domnevo) do najbolj konzervativnih (tistih, ki kasneje (šele pri večjih vzorcih) zaznajo populacijske razlike, imajo torej manjšo statistično moč, vendar pa imajo manjšo napako 1. vrste).

VII.5.5 Dodatni testi za razlike za pare skupin

- Enake variance in enako velike skupine (izbor):
  - LSD – običajen *t* test za vsak par skupin, nobenega popravka zaradi večjega števila izvedenih testov. Ni najbolj primeren. Najbolj liberalen test.
  - Duncan – opravi več zaporednih testov. Najprej uredi vzorčne aritmetične sredine od najnižje do najvišje, nato pa testira razliko za dve najbolj ekstremni. Če nista statistično značilni, je testiranja konec. Sicer se testiranje nadaljuje z naslednjima dvema najbolj ekstremnima aritmetičnima sredina. Na ta način dobi skupine statistično neznačilnih aritmetičnih sredin.
  - Bonferroni – uporabi *t* test, vendar upošteva število parov skupin. Statistična značilnost je popravljena, upoštevajoč število parov skupin.
  - Tukey – izboljšani LSD. Najbolj konzervativen test.
  - Dunnett – uporabljamo, kadar vse primerjamo z eno skupino.
- Enake variance in enako velike skupine (izbor):
  - Hochberg's GT2 ali Gabriel – v obeh primerih za velikost skupin vzame harmonično sredino.

VII.5.5 Dodatni testi za razlike za pare skupin

- Različne variance in različno velike skupine (izbor):
  - Tamhane's T2 – najbolj liberalen.
  - Dunnett's T3
  - Games-Howell - najbolj konzervativen.

VII.5.5 Dodatni testi za razlike za pare skupin

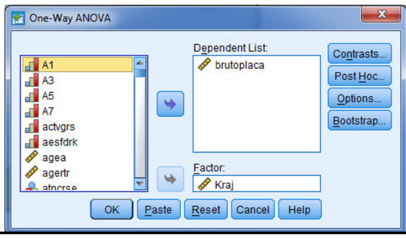
Kako izbrati primeren post-hoc test?

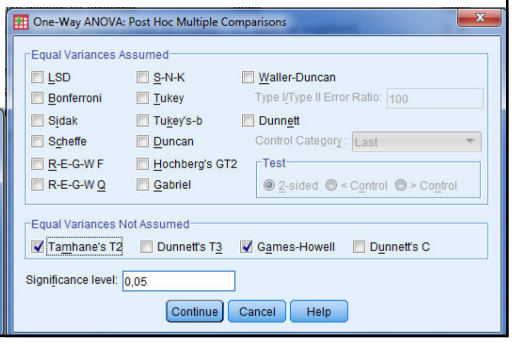
1) Odvisno od tega, ali so variance po skupinah različne ali enake.
2) Odvisno od tega, ali imamo enako ali različno velike skupine.
3) Naredimo več testov (npr. en bolj liberalen test in en bolj konzervativen test). Tisti pari skupin, ki se kažejo kot statistično značilni v obeh testih, so zagotovo statistično značilni (rečemo, da je rešitev „stabilna“).

Izvedba post-hoc testov v SPSS-ju

- Po tem, ko smo z ANOVA proceduro opravili Levene'v test o homogenosti varianc (vemo, ali so variance po skupinah enake ali različne) in ko je *F* test (običajen ali popravljeni) pokazal, da se v vsaj eni skupini povprečje razlikuje od ostalih, opravimo še post-hoc teste za posamezne pare skupin. za enakost povpr
- V SPSS-u ponovno zaženemo procedure *Analyze – Compare Means - One-way ANOVA*" in v gumbu "Options" odkljukamo "Post Hoc". Med možnostmi izberemo ustrezen test (ali več testov).
- V našem primeru (nadaljevanje prejšnjega primera) bomo izvedli en liberalen (Temhane's T2) in en konzervativen test (Games-Howell)

  za primer neenakih varianc.





Rezultati ANOVE: Rezultati post-hoc testov.

razlika vzorčnih povprečij dveh skupin

standardna napaka za razliko vzorčnih povprečij dveh skupin

statistična značilnost

Multiple Comparisons

brutoplaca bruto mesečna plača (pred davki in prispevki)
Tamhane

(I) Kraj Kraj bivanja	(J) Kraj Kraj bivanja	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1 veliko mesto (vključno s predmestjem)	2 malo mesto	-377,247	167,610	,078	-784,47	29,97
	3 podeželje	95,516	87,278	,619	-114,35	305,39
2 malo mesto	1 veliko mesto (vključno s predmestjem)	377,247	167,610	,078	-29,97	784,47
	3 podeželje	472,763*	162,154	,013	77,96	867,57
3 podeželje	1 veliko mesto (vključno s predmestjem)	-95,516	87,278	,619	-305,39	114,35
	2 malo mesto	-472,763*	162,154	,013	-867,57	-77,96

*. The mean difference is significant at the 0.05 level.

Rezultati so izpisani za vsak par skupin posebej, pri čemer se vsi pari ponovijo (1-2 in 2-1). Zadostuje seveda ena interpretacija (ali 1-2 ali 2-1).

- Statistično značilna je razlika v povprečni bruto plači med prebivalci iz podeželja in malega mesta (pri *p = 0.013*) in sicer prebivalci malega mesta imajo v povprečju za 472 EUR večjo plačo (vzorčna razlika podeželje – malo mesto je -473 EUR).
- Mejno (pri *p = 0.078*, torej stopnji značilnosti manjši od 10%) je statistično značilna tudi razlika med velikim in malim mestom (v malem mestu imajo v povprečju za 377 EUR večjo plačo).
- Povprečna bruto plača se torej v malem mestu razlikuje od povprečne bruto plače na podeželju ali v velikem mestu, medtem ko med podeželjem in velikim mestom ni statistično značilnih razlik.

Rezultati ANOVE: Rezultati post-hoc testov.

razlika vzorčnih povprečij dveh skupin

standardna napaka za razliko vzorčnih povprečij dveh skupin

statistična značilnost

Multiple Comparisons

brutoplaca bruto mesečna plača (pred davki in prispevki)
Games-Howell

(I) Kraj Kraj bivanja	(J) Kraj Kraj bivanja	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 veliko mesto (vključno s predmestjem)	2 malo mesto	-377,247	167,610	,068	-776,18	21,69
	3 podeželje	95,516	87,278	,518	-110,32	301,35
2 malo mesto	1 veliko mesto (vključno s predmestjem)	377,247	167,610	,068	-21,69	776,18
	3 podeželje	472,763*	162,154	,012	86,09	859,44
3 podeželje	1 veliko mesto (vključno s predmestjem)	-95,516	87,278	,518	-301,35	110,32
	2 malo mesto	-472,763*	162,154	,012	-859,44	-86,09

*. The mean difference is significant at the 0.05 level.

Rezultati so enaki kot v prejšnjem primeru, le da so statistične značilnosti nekoliko manjše (bolj konzervativen test, manjša je verjetnost napake 1. vrste).

- Statistično značilna je razlika v povprečni bruto plači med prebivalci iz podeželja in malega mesta (pri $p = 0.012$).
- Mejno (pri $p = 0.068$, torej stopnji značilnosti manjši od 10%) je statistično značilna tudi razlika med velikim in malim mestom.
- Povprečna bruto plača se torej v malem mestu razlikuje od povprečne bruto plače na podeželju ali v velikem mestu, medtem ko med podeželjem in velikim mestom ni statistično značilnih razlik. Rezultat je stabilen, saj sta ga pokazala oba testa.