Analiza brojčanih podataka, korelacija i regresija

doc.dr.sc. Vesna Ilakovac

Katedra za biofiziku, medicinsku statistiku i medicinsku informatiku

Medicinski fakultet Osijek

TESTIRANJE RAZLIKA

- **razlike mjerenja neke varijable na dvije ili više skupina ispitanika -> nezavisni uzorci**
- **razlike dva ili više mjerenja neke varijable na istoj skupini ispitanika ->zavisni uzorci**

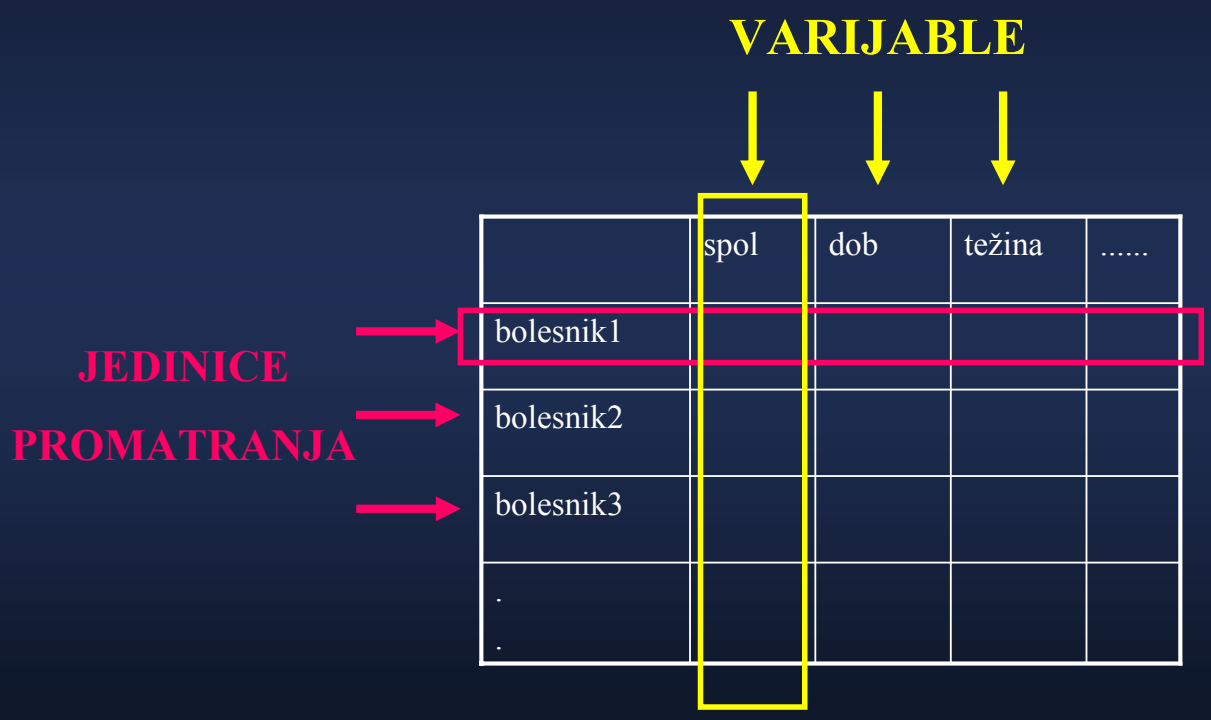
Priprema podataka

1. **jedinica promatranja *(ispitanik, preparat, pokusna životinja, organ ....)***
2. **varijable:**
   - vrsta varijable (numerička, kategorička)
   - ljestvica mjerenja (nominalna, ordinalna, intervalna, omjerna)
   - za numeričke varijable, broj decimalnih mjesta

Upis podataka

- numerički podatci
  - onako kako su izmjereni
- kategorički podatci
  - klasificirati u logičke, isključive skupine (prema problemu)

Unos podataka



Unos podataka o mjerenjima na nezavisnim skupinama

- nezavisne skupine = različiti ispitanici

  (ispitanici koji pripadaju nekoj skupini ne pripadaju niti jednoj od preostalih skupina)

- za unos podataka o nekom mjerenju na nezavisnim skupinama ispitanika UVIJEK imamo 2 varijable (bez obzira koliko je skupina ispitanika):

  1. varijabla koja određuje pripadnost ispitanika pojedinoj skupini

  2. varijabla u koju unosimo vrijednost mjerenja za danog ispitanika

Unos podataka o mjerenjima na nezavisnim skupinama

- npr. mjerenje dobi; skupine po spolu
  - broj mogućih skupina: 2

varijabla koja sadrži vrijednost mjerenja

varijabla koja definira pripadnost skupini

	Dob	Spol	.....
ispitanik1	35	M	
ispitanik2	37	M	
ispitanik3	32	M	
ispitanik4	33	Z	

Unos podataka o mjerenjima na nezavisnim skupinama

- npr. mjerenje visine; skupine po razredu (osnovna škola)
  - broj mogućih skupina: 8

varijabla koja sadrži vrijednost mjerenja

varijabla koja definira pripadnost skupini

	Visina	Razred	.....
ispitanik1	110	2	
ispitanik2	140	2	
ispitanik3	100	1	
ispitanik4	176	7	

Unos podataka o mjerenjima na zavisnim skupinama

- zavisne skupine = ponavljana mjerenja na ISTIM ispitanicima
- SVAKO mjerenje = JEDNA varijabla

koliko mjerenja toliko varijabli

Unos podataka o mjerenjima na zavisnim skupinama

- npr. praćenje dnevnih varijacija sistoličkog tlaka; mjerenja u 6h, 10h, 14h, 18h, 22h

po jedna varijabla za svako mjerenje

	ST6	ST10	ST14	ST18	ST22
ispitanik1	120	135	140	180	160
ispitanik2	115	120	120	125	120
ispitanik3	140	145	150	150	180
ispitanik4	118	110	110	115	120

STUDENTOV T-TEST
(t-test za nezavisne uzorke)

za što se koristi:

- testiranje razlike između dvije nezavisne skupine ispitanika

pod kojim uvjetima:

- varijabla koju testiramo mjerena je najmanje intervalnom skalom
- varijabla koju testiramo ima normalnu razdiobu u promatranim skupinama
- varijance mjerenja varijable koju testiramo u promatranim skupinama su slične (homogenost varijanci)

test statistika:

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{SE(\bar{x}_A - \bar{x}_B)}$$

ima Studentovu (t) razdiobu
s n_A+n_B-2 stupnja slobode

$$SE(\bar{x}_A - \bar{x}_B) = \sqrt{s^2_{zaj}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

standardna pogreška razlike aritmetičkih sredina

$$s^2_{zaj} = \frac{(n_A - 1)s^2_A + (n_B - 1)s^2_B}{(n_A - 1) + (n_B - 1)}$$

zajednička varijanca

test homogenosti varijanci (F-test):

$$F = \frac{s^2_A}{s^2_B}$$

ima F razdiobu
s $n_{A-1,}$ n_B-1 stupnjeva slobode

ako koristimo tablice:

- tablice za F sadrže obično samo desnu stranu distribucije

$\Rightarrow$ u ***brojnik*** **stavljamo** ***veću varijancu***

ZADATAK 1

Ispitivan je utjecaj sniženja tjelesne temperature na protrombinsko vrijeme. Izvršena su mjerenja PV na dvije skupine ispitanika. U jednoj skupini bilo je 16 ispitanika normalne temperature (kontrolna skupina). U drugoj skupini bilo je 14 ispitanika sa sniženom temperaturom (pokusna skupina).

Mjerenjem su dobiveni sljedeći rezultati (u sekundama):

Kontrolna skupina (37^0C)	Pokusna skupina (15^0C)
7	8
8	6
9	8
6	9
8	10
10	12
8	11
12	8
9	7
11	8
10	8
8	9
9	9
7	9
7	
8	

- nezavisne skupine

  2 varijable:

  protrombinsko vrijeme

  skupina

  1 – kontrolna skupina

  2 – pokusna skupina

pvrijeme	skupina
7	1
8	1
9	1
:	:
7	1
7	1
8	1
8	2
6	2
8	2
:	:
9	2
9	2
9	2

Opis varijabli

protrombinsko vrijeme:

– numerička varijabla, omjerna ljestvica

skupina:

– kategorička varijabla, nominalna ljestvica

- **ispitati mjere sredine i raspršenja za PV u svakoj skupini**
- **ispitati normalnost raspodjele PV u svakoj skupini**

Opisna statistika i ispitivanje normalnosti - MedCalc:

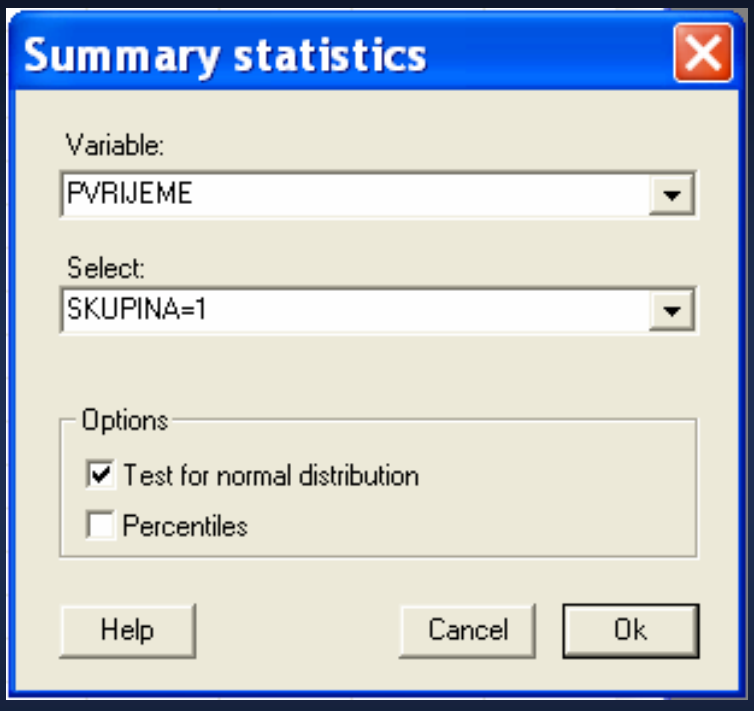
Statistics-> Summary statistics

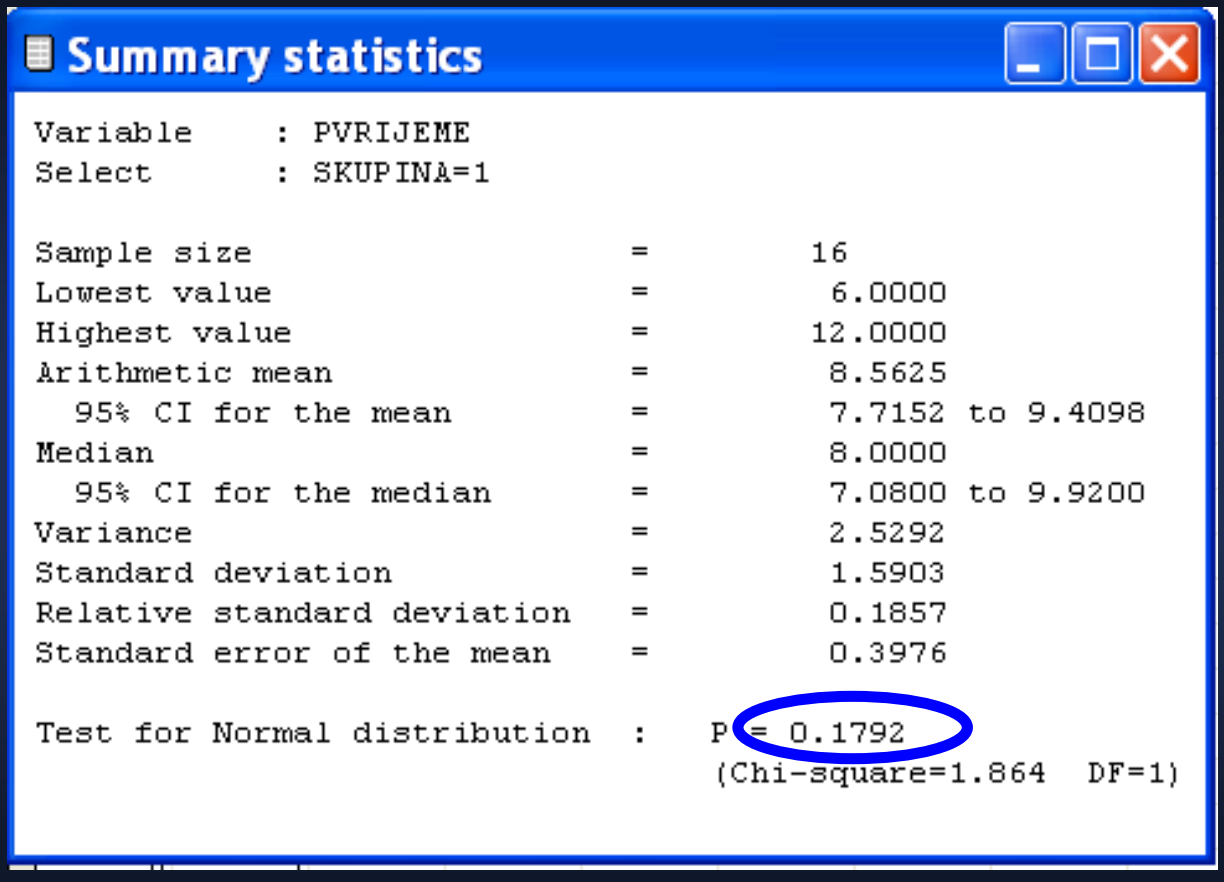
za kontrolnu skupinu:

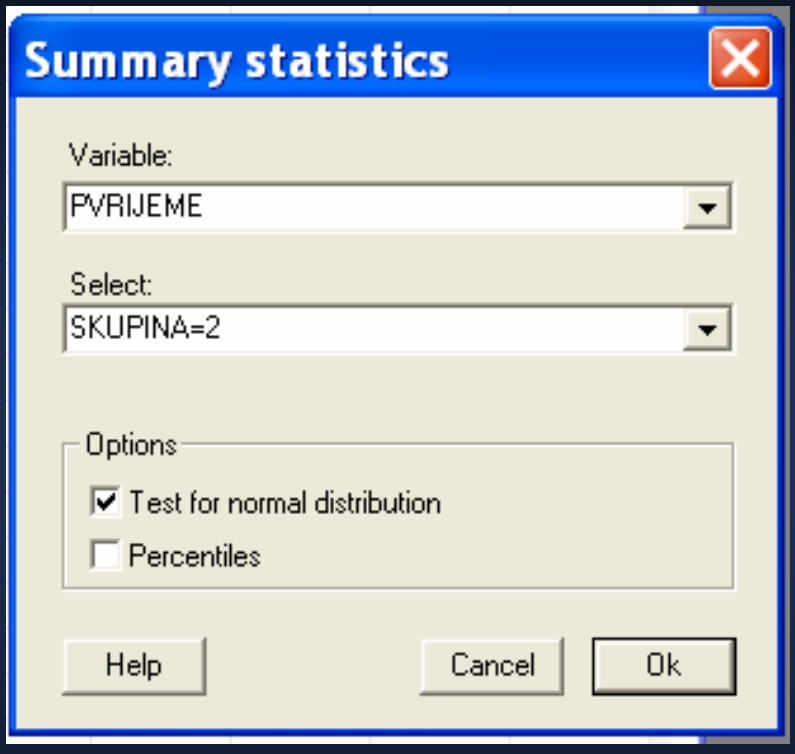
- **Variable -> pvrijeme**
- **Select -> skupina=1**
- **Options -> Test for normal distribution**

za pokusnu skupinu:

- **Variable -> pvrijeme**
- **Select -> skupina=2**
- **Options -> Test for normal distribution**




Summary statistics
Variable : PVRIJEME
Select : SKUPINA=1
Sample size = 16
Lowest value = 6.0000
Highest value = 12.0000
Arithmetic mean = 8.5625
95% CI for the mean = 7.7152 to 9.4098
Median = 8.0000
95% CI for the median = 7.0800 to 9.9200
Variance = 2.5292
Standard deviation = 1.5903
Relative standard deviation = 0.1857
Standard error of the mean = 0.3976
Test for Normal distribution : P = 0.1792
(Chi-square=1.864 DF=1)


Summary statistics
Variable:
PVRIJEME
Select:
SKUPINA=2
Options
Test for normal distribution
Percentiles
Help
Cancel
Ok

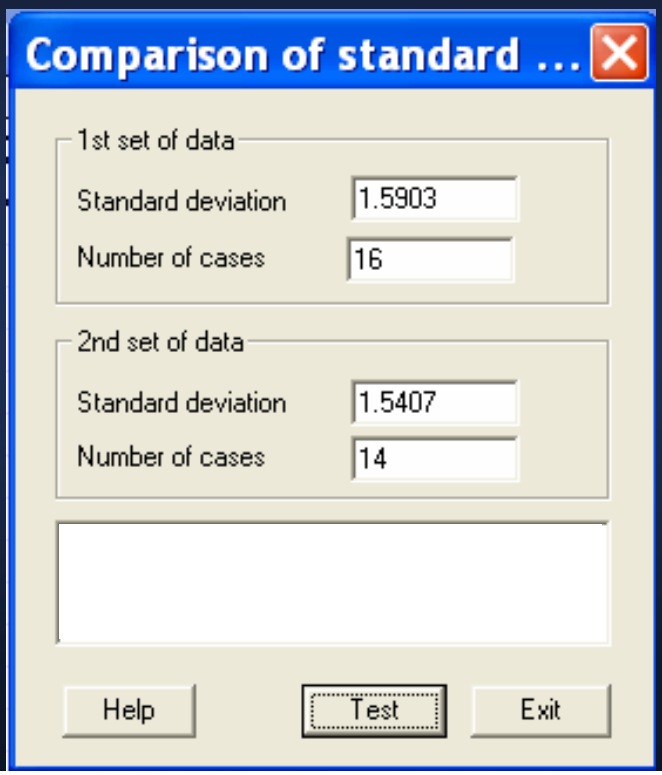
Summary statistics

```
Variable     : PVRIJEME
Select       : SKUPINA=2

Sample size                          =         14
Lowest value                         =          6.0000
Highest value                        =         12.0000
Arithmetic mean                      =          8.7143
  95% CI for the mean                =          7.8249 to 9.6037
Median                               =          8.5000
  95% CI for the median              =          8.0000 to 9.6668
Variance                             =          2.3736
Standard deviation                   =          1.5407
Relative standard deviation          =          0.1768
Standard error of the mean           =          0.4118

Test for Normal distribution  :    P = 0.8357
                                   (Chi-Square=0.047   DF=1)
```

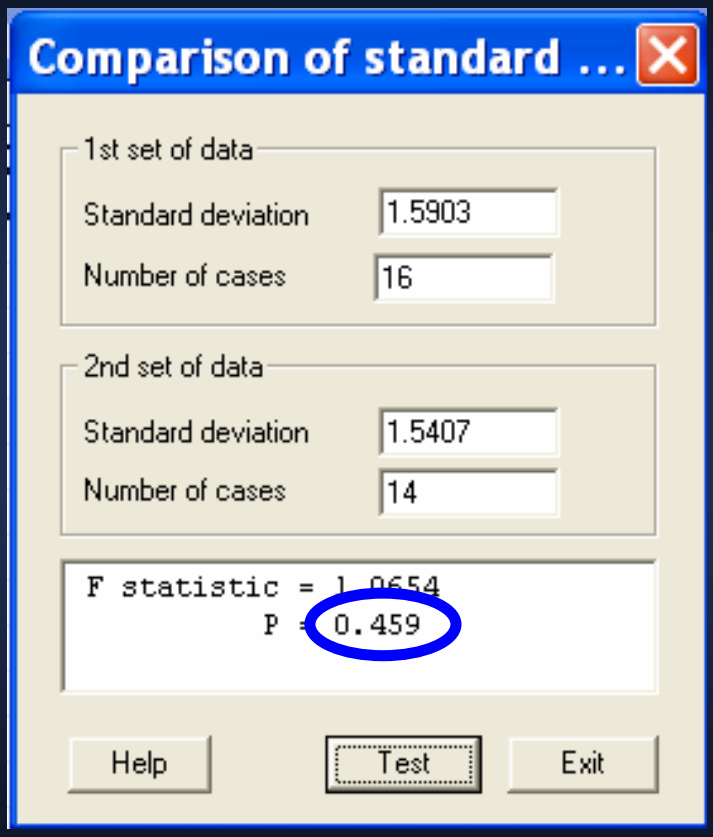

Homogenost varijanci - MedCalc:

- preko usporedbe standardnih devijacija:

Tests-> Comparison of... -> standard deviations (F-test)

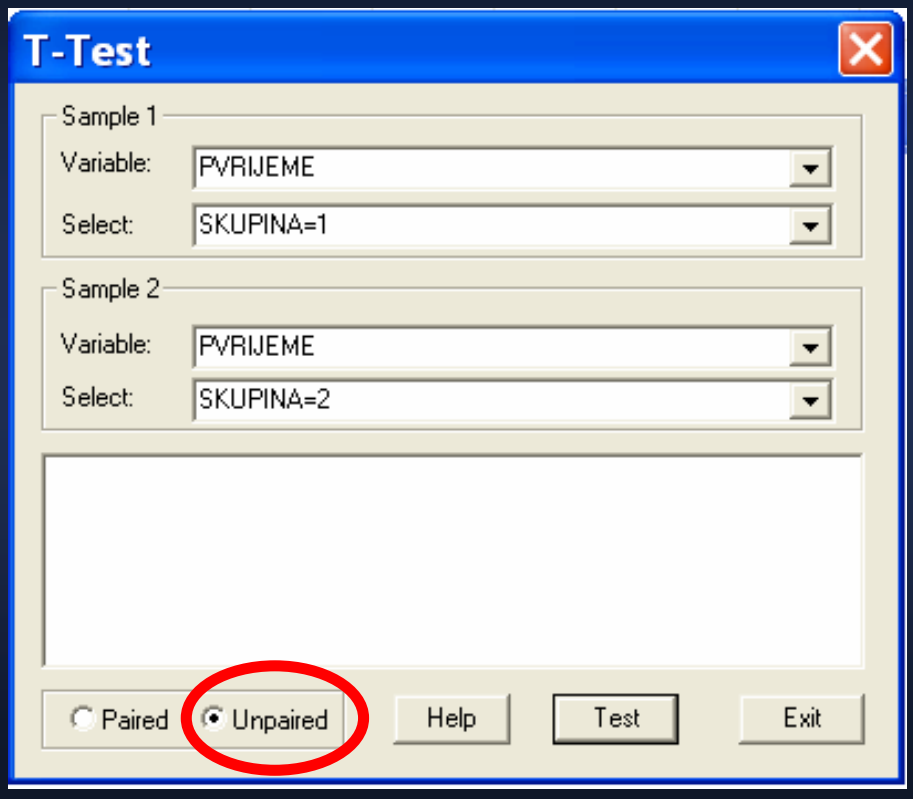


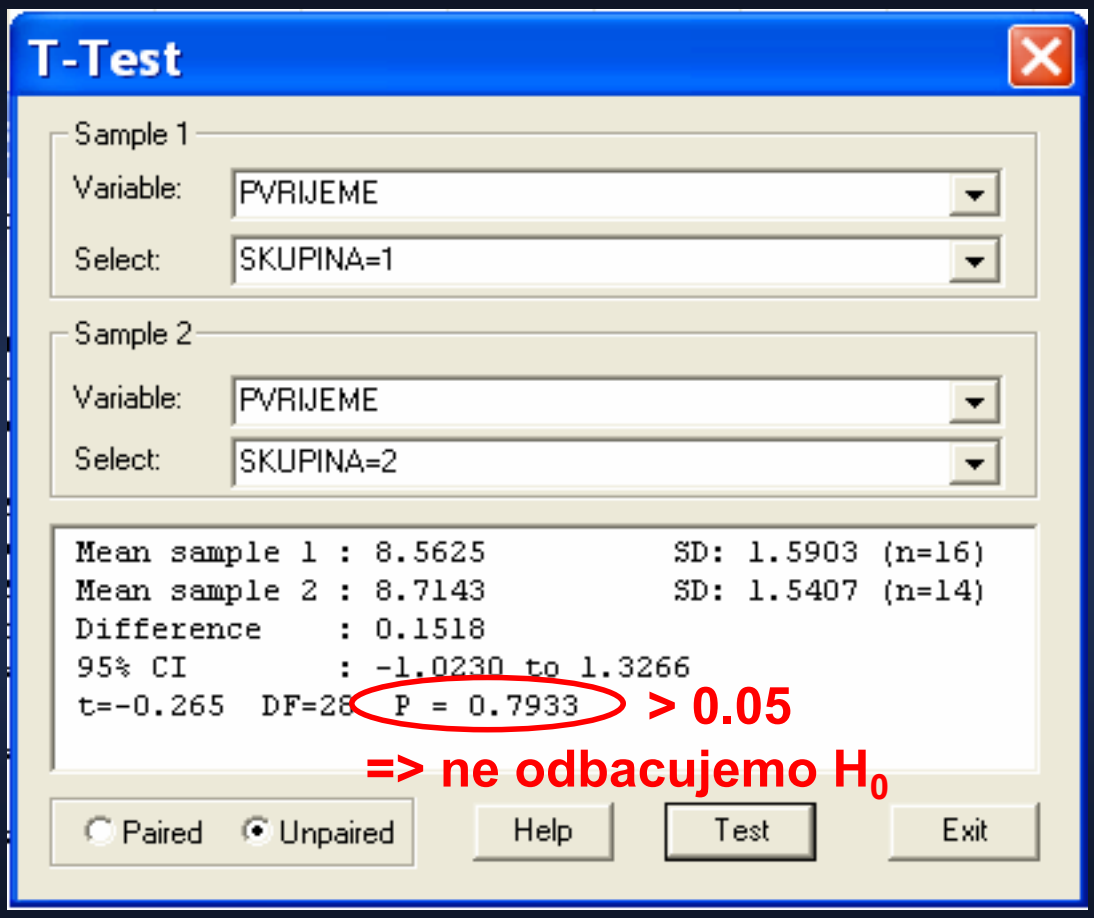
Homogenost varijanci - MedCalc:



Studentov t-test - MedCalc:

Statistics-> T tests



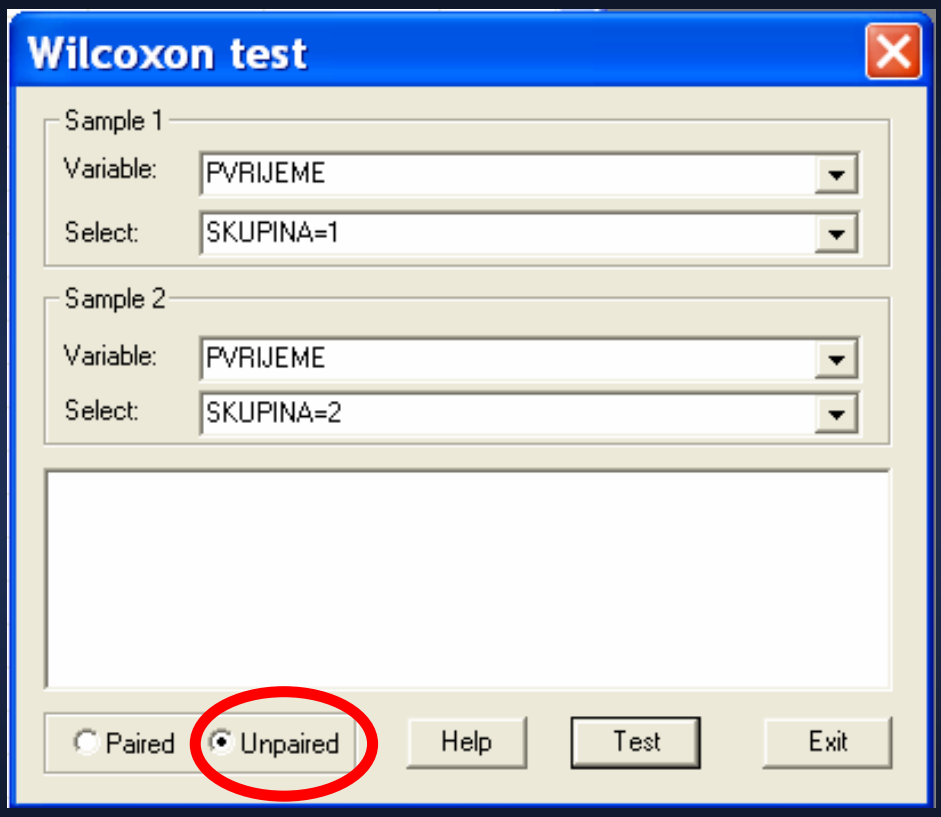


ŠTO AKO NISU ZADOVOLJENI UVJETI ?

Mann-Whitney-Wilcoxon test
(Mann-Whitney U test)

MedCalc:

Statistics-> Wilcoxon tests



T-TEST DIFERENCIJA
(t-test za zavisne uzorke)

za što se koristi:

- testiranje razlike između dvije zavisne skupine ispitanika (ponavljana mjerenja na istim ispitanicima)

pod kojim uvjetima:

- razlike parova vrijednosti mjerene su najmanje intervalnom skalom
- razlike parova vrijednosti imaju normalnu razdiobu

test statistika:

$$t = \frac{\bar{d} - (\mu_A - \mu_B)}{\sqrt{\frac{s_d^2}{n}}}$$

ima Studentovu (t) razdiobu s n-1 stupnjeva slobode

$$\bar{d} = \bar{x}_A - \bar{x}_B$$

srednja razlika

$$s_d^2 = \frac{\sum_i (d_i - \bar{d})^2}{n-1}$$

varijanca razlike

ZADATAK 2

Ispitivan je utjecaj alkohola na vrijeme reakcije vozača. Izvršeno je mjerenje vremena reakcije 14 vozača na standardne zadatke prije i nakon konzumacije određene količine alkohola.

Mjerenjem su dobiveni slijedeći rezultati:

	prije	nakon
1	0.68	0.73
2	0.64	0.66
3	0.68	0.66
4	0.82	0.92
5	0.58	0.68
6	0.80	0.87
7	0.72	0.77
8	0.65	0.70
9	0.84	0.88
10	0.73	0.79
11	0.63	0.68
12	0.72	0.68
13	0.68	0.75
14	0.69	0.78

- zavisne skupine

  2 varijable:

  prije

  nakon

prije	nakon
0.68	0.73
0.64	0.66
0.68	0.66
0.82	0.92
0.58	0.68
0.80	0.87
0.72	0.77
0.65	0.70
0.84	0.88
0.73	0.79
0.63	0.68
0.72	0.68
0.68	0.75
0.69	0.78

Opis varijabli

prije, nakon:

– numeričke, omjerna ljestvica

za obje varijable:

- **ispitati mjere sredine i raspršenja**

kreirati novu varijablu prije-nakon:

- **ispitati normalnost**

Opisna statistika i ispitivanje normalnosti - MedCalc:

Statistics-> Summary statistics

prije:

Variable -> prije

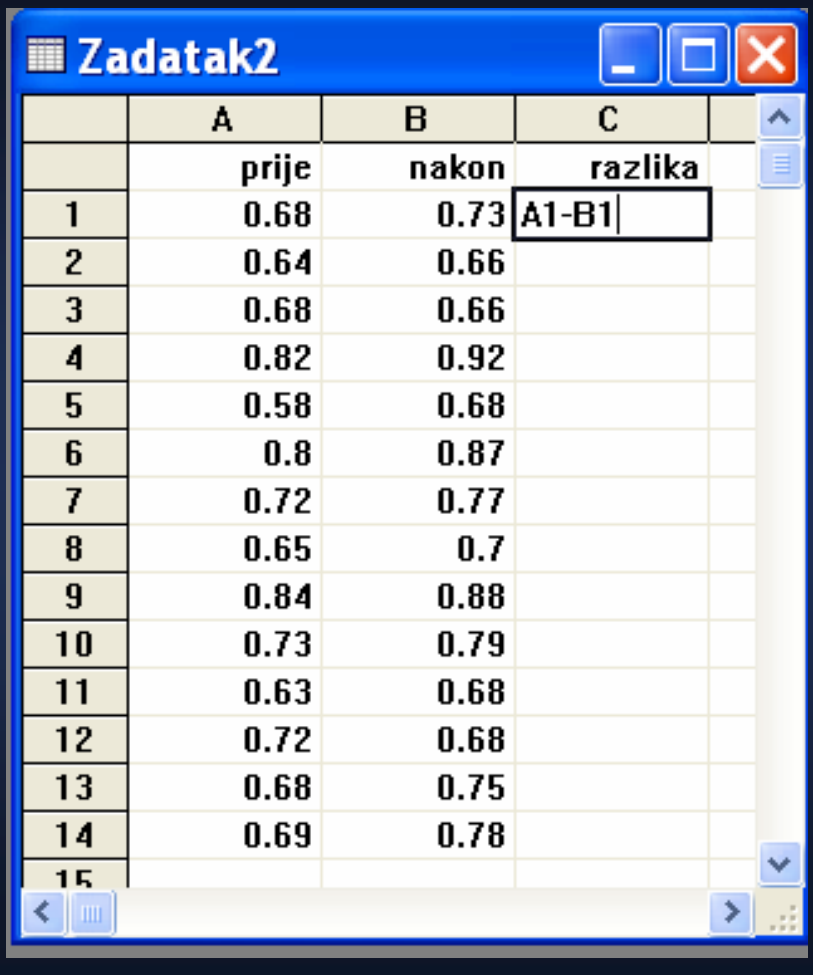
poslije:

Variable -> poslije

razlika:

Variable-> razlika

Options -> Test for normal distribution



Zadatak2

	A	B	C
	prije	nakon	razlika
1	0.68	0.73	-0.05
2	0.64	0.66	-0.02
3	0.68	0.66	0.02
4	0.82	0.92	-0.1
5	0.58	0.68	-0.1
6	0.8	0.87	-0.07
7	0.72	0.77	-0.05
8	0.65	0.7	-0.05
9	0.84	0.88	-0.04
10	0.73	0.79	-0.06
11	0.63	0.68	-0.05
12	0.72	0.68	0.04
13	0.68	0.75	-0.07
14	0.69	0.78	-0.09

Summary statistics

Variable : PRIJE

Sample size	=	14
Lowest value	=	0.5800
Highest value	=	0.8400
Arithmetic mean	=	0.7043
95% CI for the mean	=	0.6613 to 0.7473
Median	=	0.6850
95% CI for the median	=	0.6433 to 0.7767
Variance	=	0.0055
Standard deviation	=	0.0745
Relative standard deviation	=	0.1058
Standard error of the mean	=	0.0199

Summary statistics

Variable : NAKON

Sample size	=	14
Lowest value	=	0.6600
Highest value	=	0.9200
Arithmetic mean	=	0.7536
95% CI for the mean	=	0.7038 to 0.8034
Median	=	0.7400
95% CI for the median	=	0.6800 to 0.8433
Variance	=	0.0074
Standard deviation	=	0.0863
Relative standard deviation	=	0.1145
Standard error of the mean	=	0.0231

Summary statistics

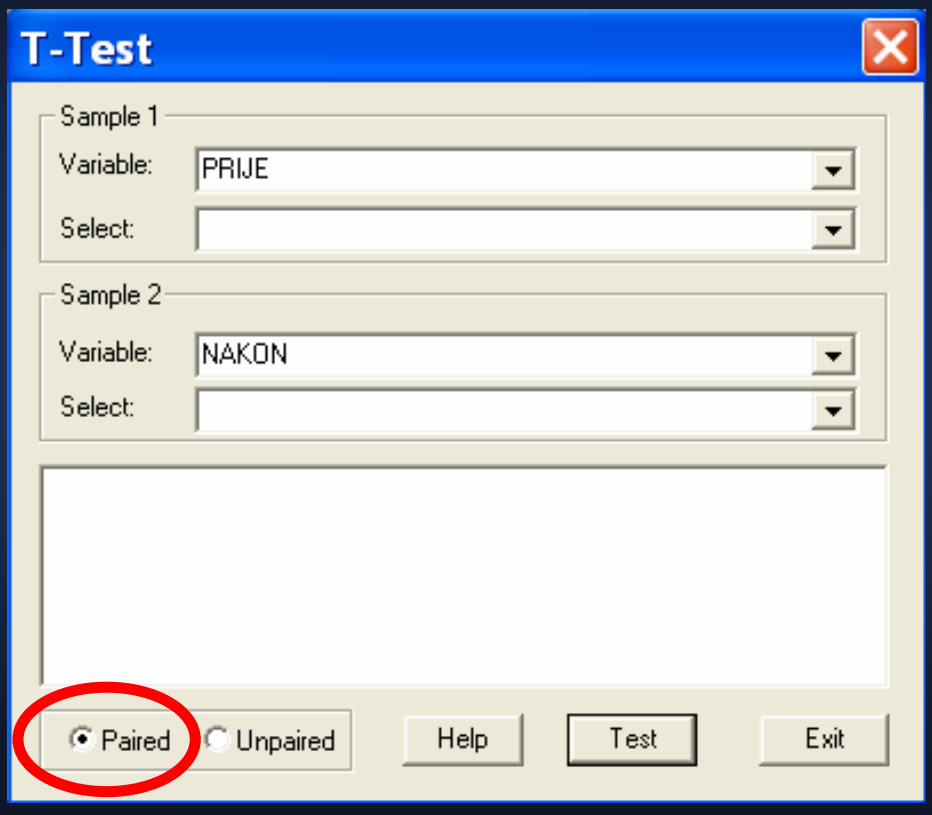
Variable : RAZLIKA

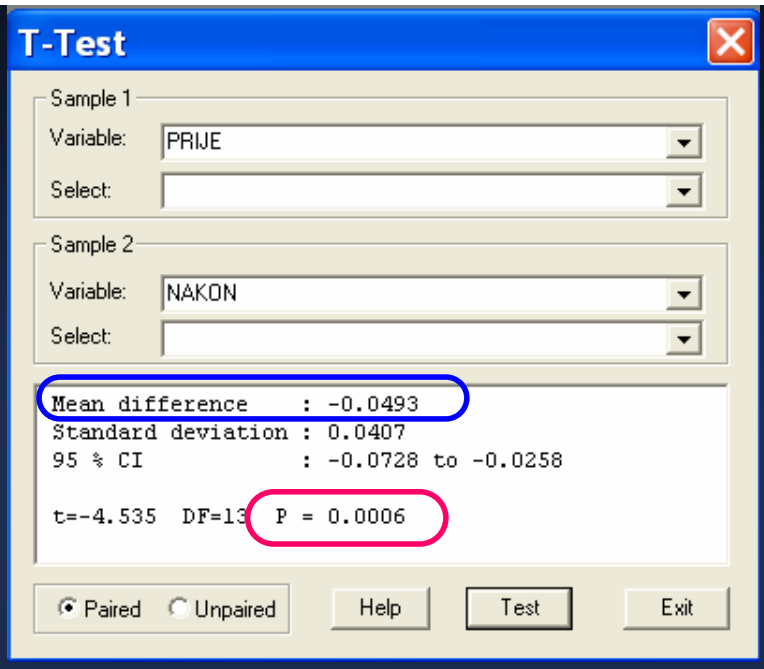
Sample size	=	14
Lowest value	=	-0.1000
Highest value	=	0.0400
Arithmetic mean	=	-0.0493
95% CI for the mean	=	-0.0728 to -0.0258
Median	=	-0.0500
95% CI for the median	=	-0.0833 to -0.0267
Variance	=	0.0017
Standard deviation	=	0.0407
Relative standard deviation	=	-0.8250
Standard error of the mean	=	0.0109

Test for Normal distribution : P = 0.4908
(Chi-square=0.478 DF=1)

T-test diferencija - MedCalc:

Statistics-> T tests





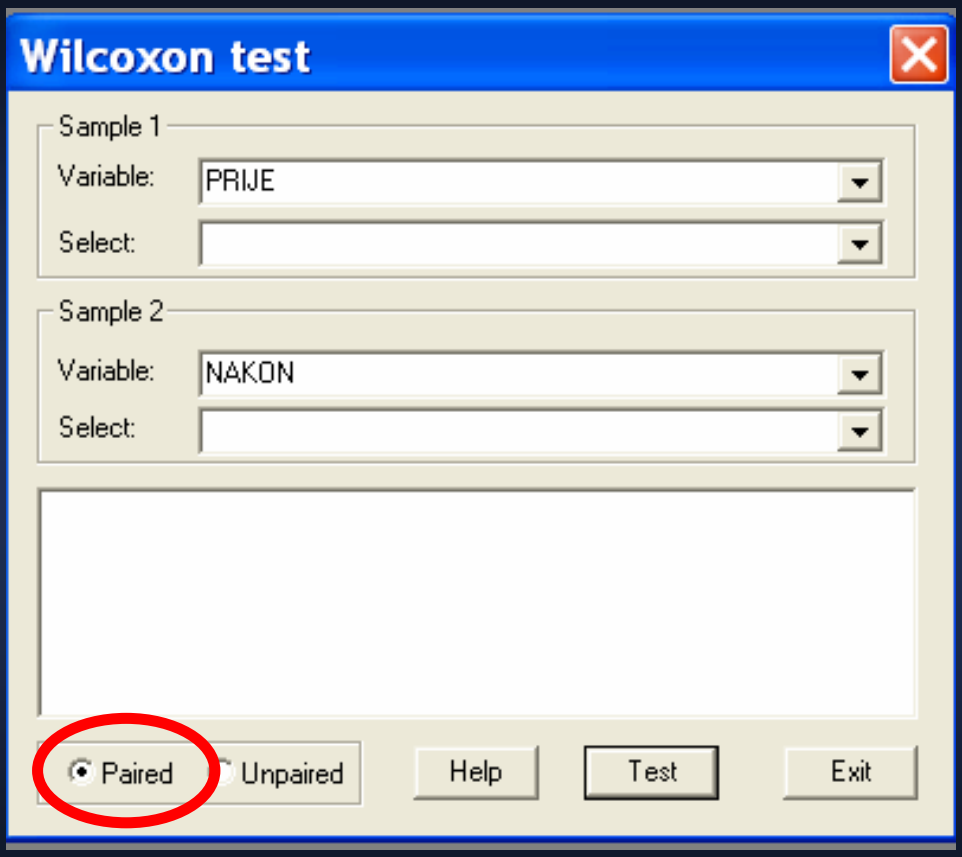
razlika < 0 i p < 0.05 => vrijeme reakcije vozača nakon konzumacije te količine alkohola značajno je dulje nego prije konzumacije te količine alkohola

ŠTO AKO NISU ZADOVOLJENI UVJETI ?

Wilcoxonov test

MedCalc:

Statistics-> Wilcoxon tests



JEDNOSMJERNA ANALIZA VARIJANCE (One-way ANOVA)

za što se koristi:

- testiranje razlike između tri i više skupina

faktor

- kategorička varijabla prema kojoj su definirane skupine

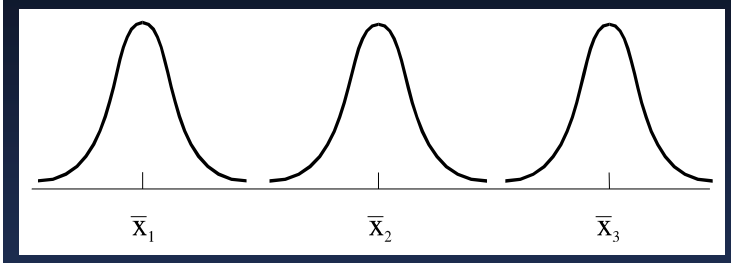
jednosmjerna analiza varijance

- analiza varijance s jednim faktorom

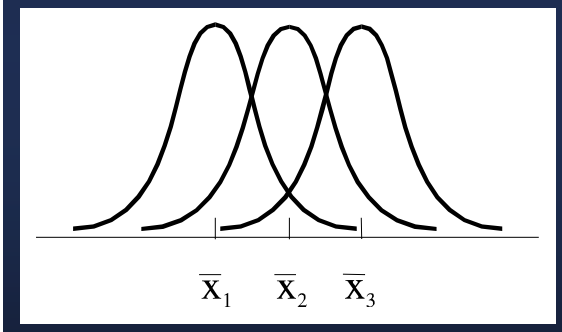
- postupak u kojem donosimo dvije procjene varijance za promatrane grupe (model):
  - procjenu koja odražava ***varijabilitet između grupa***
  - procjenu koja odražava ***varijabilitet unutar grupa***

OSNOVNA IDEJA:

- utvrditi je li varijabilitet ***između grupa*** veći od varijabiliteta ***unutar grupa***



VARIJABILITET IZMEĐU GRUPA VEĆI JE OD VARIJABILITETA UNUTAR GRUPA



VARIJABILITET UNUTAR GRUPA VEĆI JE OD VARIJABILITETA IZMEĐU GRUPA

pretpostavke:

- varijabla koju testiramo mjerena je najmanje intervalnom skalom
- varijabla koju testiramo ima normalnu razdiobu u promatranim skupinama
- varijance mjerenja varijable koju testiramo u promatranim skupinama su slične (**homogenost varijanci**)

test statistika:

$$F = \frac{MS_{tretman}}{MS_{pogreška}}$$

procjena koja odražava varijabilitet IZMEĐU grupa

procjena koja odražava varijabilitet UNUTAR grupa

- F ima F razdiobu s k-1, N-k stupnjeva slobode

ZADATAK 3

Bolesnici s uznapredovalim stadijem raka želuca, bronhija, kolona i dojke tretirani su novim lijekom. Svrha istraživanja je utvrditi je li preživljenje bolesnika povezano sa zahvaćenim organom. Vrijeme preživljenja (u mjesecima) dano je u tablici:

Želudac	Bronhiji	Kolon	Dojke
11	9	16	35
8	21	19	45
9	4	14	40
7	21	8	34
20	16	13	46
11	13	23	35
33	8	23	52
12	8	21	28
10	12	20	42
19	29	19	49
12	12	31	38
18	13	28	32
20	6	19	43
18	15	13	28
17	12	10	
10	8	4	
	16	17	

- nezavisne skupine

  4 skupine , ali **2 varijable**:

  vrijeme

  organ

  1 - želudac

  2 - bronhiji

  3 - kolon

  4 - dojke

organ	vrijeme
1	11
:	:
1	10
2	9
:	:
2	16
3	16
:	:
3	17
4	35
:	:
4	28

Opisna statistika i ispitivanje normalnosti - MedCalc:

Statistics-> Summary statistics

za želudac:
- Variable -> vrijeme
- Select -> organ=1
- Options -> Test for normal distribution

za bronhije:
- Variable -> vrijeme
- Select -> organ=2
- Options -> Test for normal distribution

za kolon:
- Variable -> vrijeme
- Select -> organ=3
- Options -> Test for normal distribution

za dojke:
- Variable -> vrijeme
- Select -> organ=4
- Options -> Test for normal distribution

Summary statistics

```
Variable     : VRIJEME
Select       : ORGAN=1

Sample size                    =        16
Lowest value                   =         7.0000
Highest value                  =        33.0000
Arithmetic mean                =        14.6875
  95% CI for the mean          =        11.1551 to 18.2199
Median                         =        12.0000
  95% CI for the median        =        10.0000 to 18.9200
Variance                       =        43.9625
Standard deviation             =         6.6304
Relative standard deviation    =         0.4514
Standard error of the mean     =         1.6576

Test for Normal distribution   :   P = 0.199
                                   (Chi-Square=1.646   DF=1)
```

Summary statistics

```
Variable     : VRIJEME
Select       : ORGAN=2

Sample size                    =        17
Lowest value                   =         4.0000
Highest value                  =        29.0000
Arithmetic mean                =        13.1176
  95% CI for the mean          =         9.8918 to 16.3435
Median                         =        12.0000
  95% CI for the median        =         8.0000 to 16.0000
Variance                       =        39.3603
Standard deviation             =         6.2738
Relative standard deviation    =         0.4783
Standard error of the mean     =         1.5216

Test for Normal distribution   :   P = 0.2308
                                   (Chi-Square=1.467   DF=1)
```

Summary statistics

```
Variable     : VRIJEME
Select       : ORGAN=3

Sample size                    =        17
Lowest value                   =         4.0000
Highest value                  =        31.0000
Arithmetic mean                =        17.5294
  95% CI for the mean          =        13.9805 to 21.0783
Median                         =        19.0000
  95% CI for the median        =        13.0000 to 22.0813
Variance                       =        47.6397
Standard deviation             =         6.9022
Relative standard deviation    =         0.3937
Standard error of the mean     =         1.6740

Test for Normal distribution   :   P = 0.6711
                                   (Chi-Square=0.185   DF=1)
```

Summary statistics

```
Variable     : VRIJEME
Select       : ORGAN=4

Sample size                    =        14
Lowest value                   =        28.0000
Highest value                  =        52.0000
Arithmetic mean                =        39.0714
  95% CI for the mean          =        34.7487 to 43.3942
Median                         =        39.0000
  95% CI for the median        =        32.6664 to 45.6668
Variance                       =        56.0714
Standard deviation             =         7.4881
Relative standard deviation    =         0.1917
Standard error of the mean     =         2.0013

Test for Normal distribution   :   P = 0.6836
                                   (Chi-Square=0.169   DF=1)
```

Test homogenosti varijanci - MedCalc:

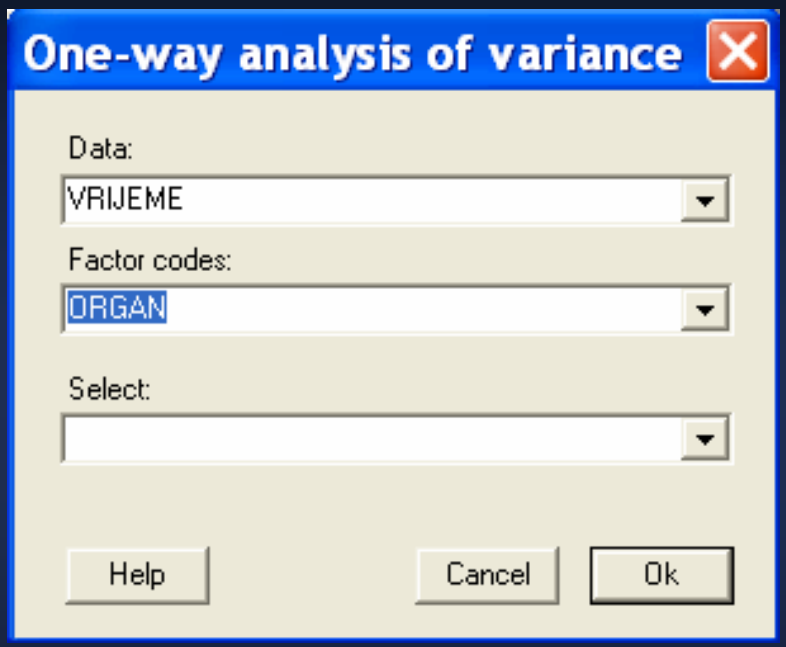
- ver. 4.1 NEMA!!!!

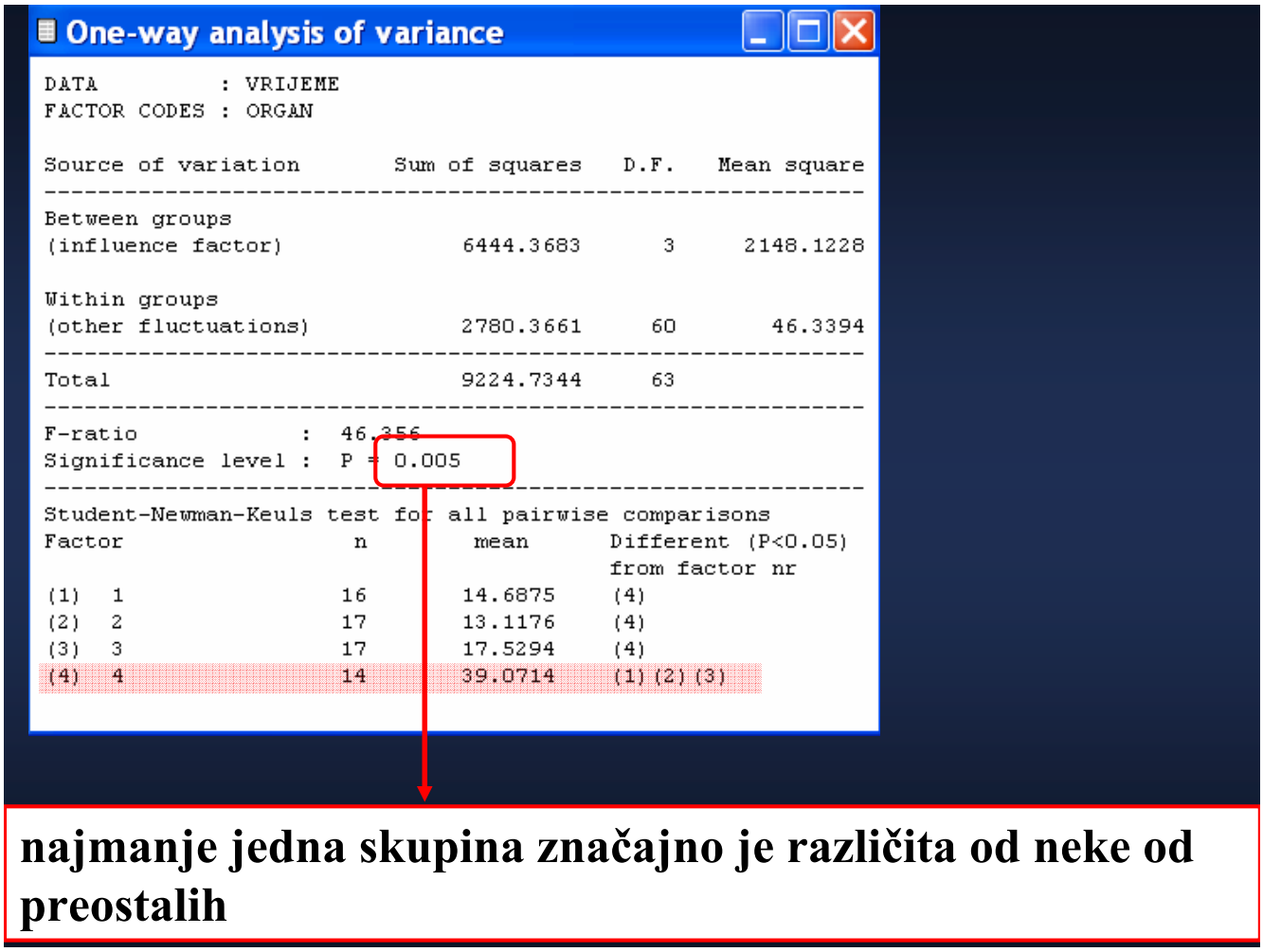
Test homogenosti varijanci - SPSS:

Test of Homogeneity of Variances

vrijeme

Levene Statistic	df1	df2	Sig.
.410	3	60	.746





najmanje jedna skupina značajno je različita od neke od preostalih

KORELACIJA

KORELACIJA

- veza među obilježjima (varijablama)
- obilježja koja "variraju zajedno"

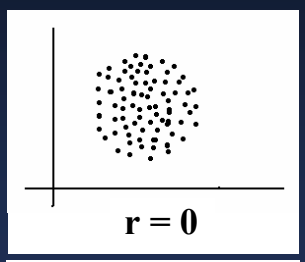
KOEFICIJENT KORELACIJE

- mjera stupnja povezanosti

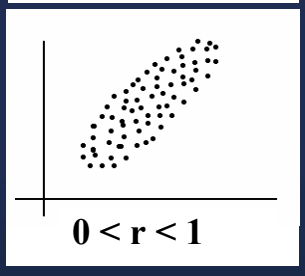
PEARSONOV KOEFICIJENT KORELACIJE r

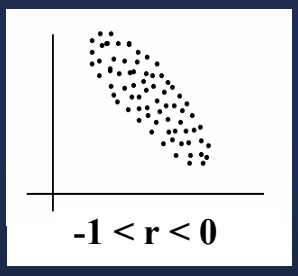
- mjera stupnja ***linearne*** povezanosti dviju kvantitativnih varijabli

$$-1 \leq r \leq 1$$

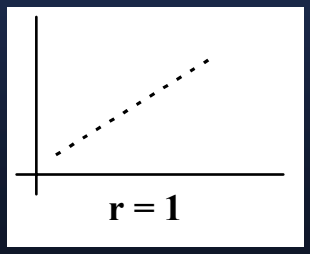


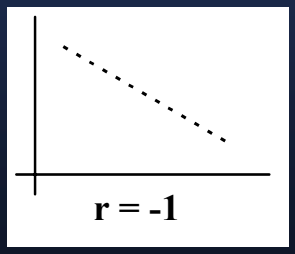
nema povezanosti





stohastička povezanost





funkcionalna povezanost

x, y ....nizovi vrijednosti varijabli čiju povezanost ocjenjujemo

POSTUPAK ZA OCJENU KORELACIJE

a) crtanje korelacionog dijagrama

b) ocjena postojanja povezanosti

c) u slučaju da postoji **linearna** povezanost, računamo koeficijent korelacije r

$$r = \frac{\sum_{i=1}^{N} z_{xi} z_{yi}}{N-1}$$

z_{xi}, z_{yi} .... standardizirane vrijednosti pojedinačnih vrijednosti varijabli **x i y, tj.**

$$z_{xi} = \frac{x_i - \bar{x}}{s_x} \qquad z_{yi} = \frac{y_i - \bar{y}}{s_y}$$

skraćeni postupak računanja r:

$$r = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right]\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right]}}$$

ZNAČAJNOST KOEFICIJENTA KORELACIJE

- testiramo je li **r** značajno različit od 0
- test statistika

$$t = r\frac{\sqrt{N-2}}{\sqrt{1-r^2}}$$

slijedi t razdiobu uz $df = N - 2$

ZADATAK 4

Izmjerena je visina u centimetrima i vitalni kapacitet pluća (VC) u litrama 33 studentice prve godine. Dobiveni su sljedeći rezultati:

Rbr.	Visina	VC	Rbr.	Visina	VC	Rbr.	Visina	VC
1.	180.6	4.74	12.	155.0	2.20	23.	174.2	4.27
2.	168.0	3.63	13.	171.0	3.38	24.	167.0	3.45
3.	163.0	3.40	14.	171.5	3.82	25.	162.0	2.88
4.	171.0	3.75	15.	167.6	3.26	26.	172.0	4.13
5.	177.0	4.23	16.	160.2	2.63	27.	161.0	2.90
6.	169.4	3.20	17.	166.6	3.06	28.	155.0	2.65
7.	161.0	2.90	18.	167.0	3.52	29.	162.0	3.12
8.	170.0	3.88	19.	163.0	2.82	30.	174.0	4.02
9.	158.0	2.40	20.	172.0	3.41	31.	161.0	2.80
10.	161.0	2.60	21.	158.0	2.81	32.	166.0	3.46
11.	163.0	2.72	22.	165.0	3.07	33.	166.0	3.26

Ocijenite postoji li povezanost visine i vitalnog kapaciteta pluća

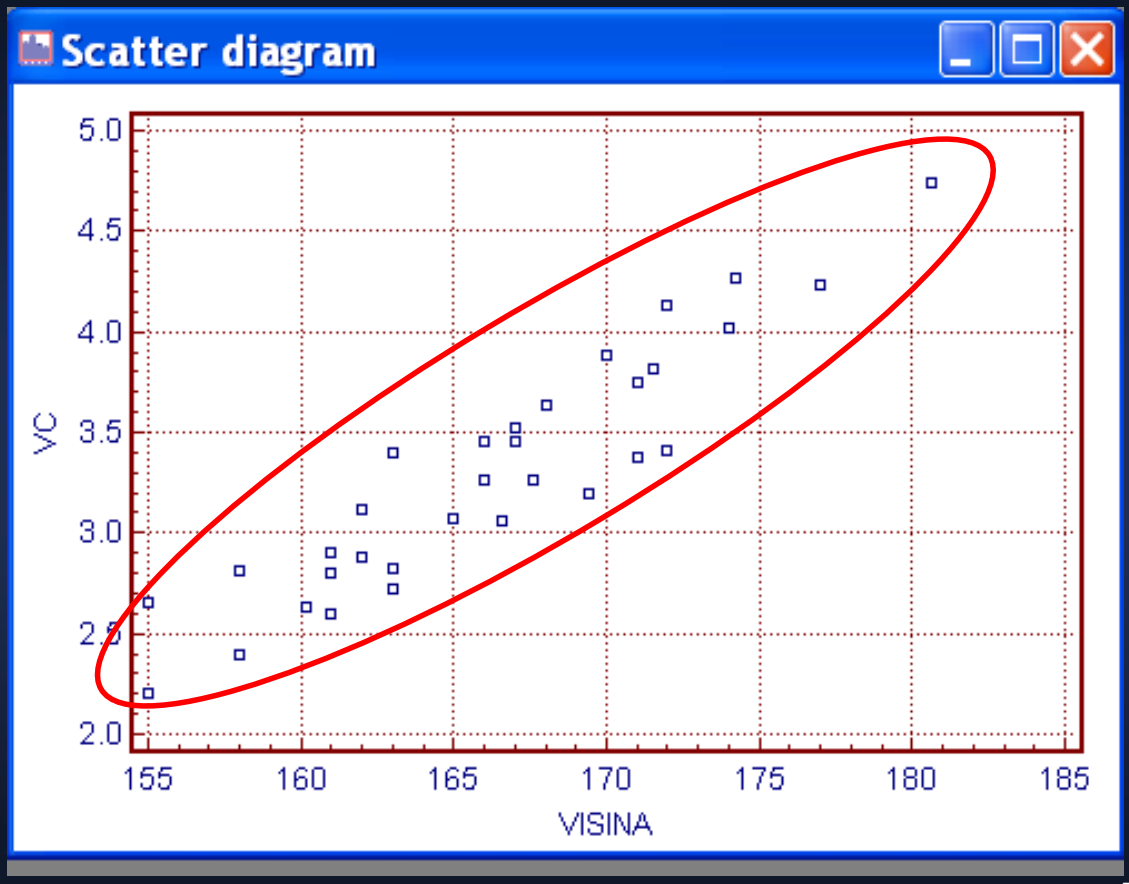
Crtanje korelacionog dijagrama (točkasti "scatter" grafikon)

MedCalc:

Statistics -> Correlation -> Scatter diagram

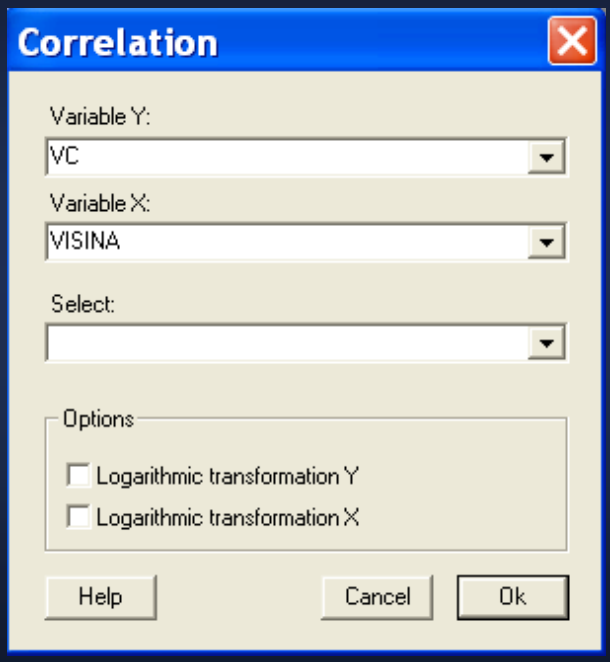
visina -> X os

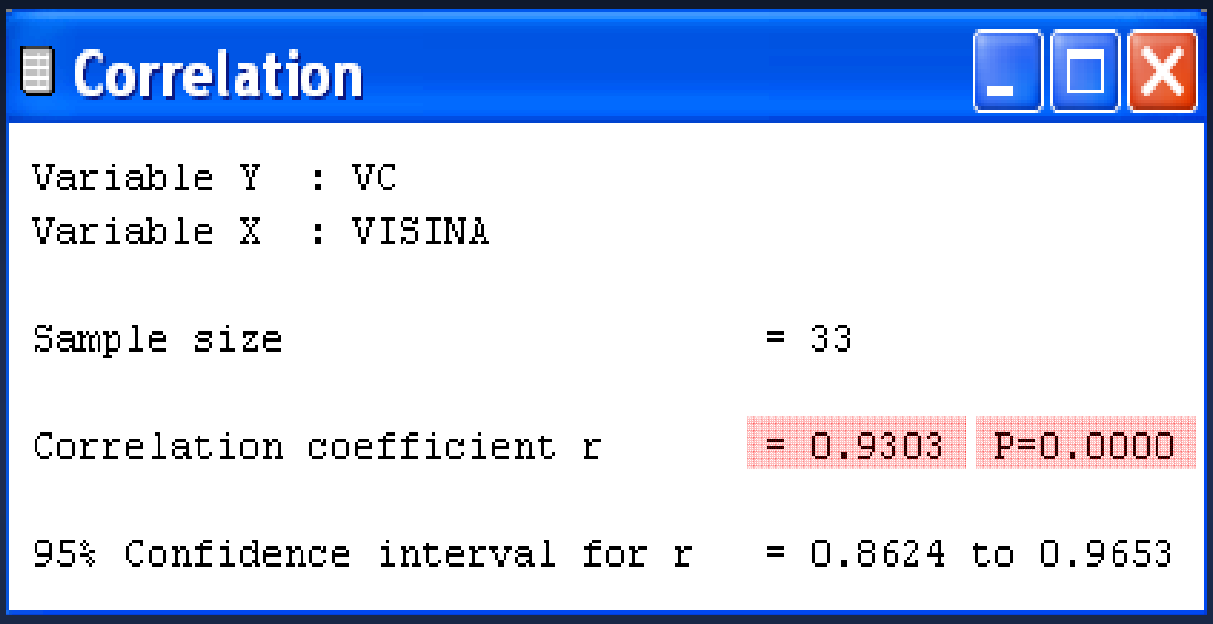
VC -> Y os



Izračun koeficijenta korelacije - MedCalc:

Statistics -> Correlation -> Correlation coefficient





Interpretacija koeficijenta korelacije

statistička značajnost

- ocjenjuje je li **r** značajno različit od 0
- ovisi o veličini uzorka - za velike uzorke, mali r će biti značajan

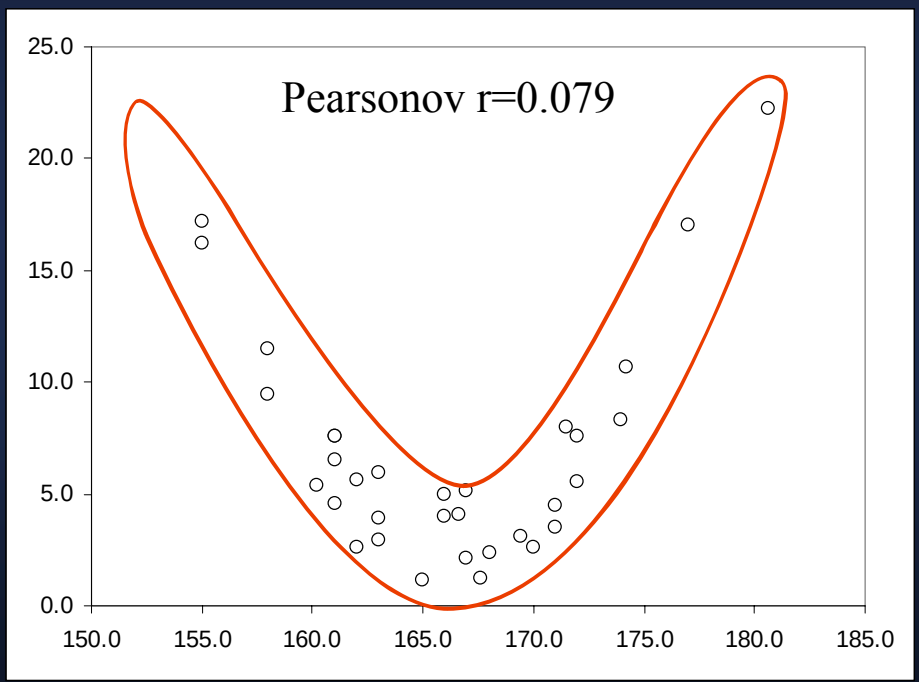
praktična značajnost

- ocjenjuje se pomoću ***koeficijenta determinacije*** r^2
- koliki udio varijabilnosti je "zajednički"

Interpretacija koeficijenta korelacije

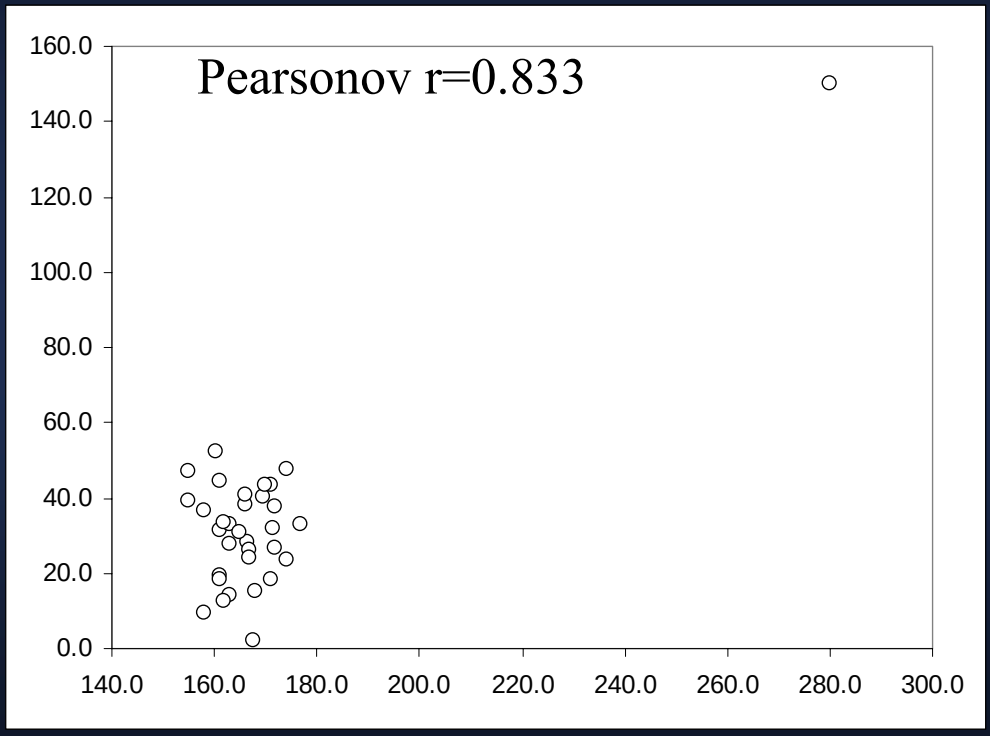
N	Najmanji značajni r (p<0.05)	r^2
10	0.632	0.399
20	0.444	0.197
30	0.361	0.130
40	0.312	0.097
50	0.279	0.078
100	0.197	0.039
200	0.139	0.019
300	0.113	0.013
500	0.088	0.008

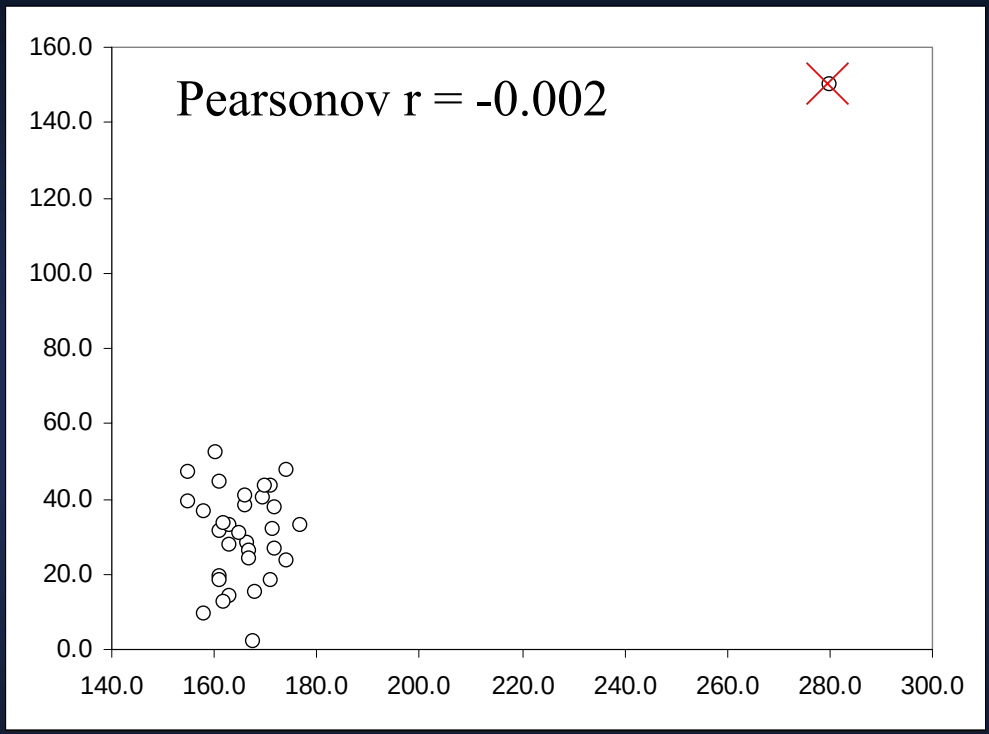
VAŽNO:
Pearsonov koeficijent korelacije daje stupanj LINEARNE povezanosti dviju varijabli!



VAŽNO:
Korelacija daje povezanost, a ne UZROČNOST !

VAŽNO:
Na koeficijent korelacije jako utječu ekstremne vrijednosti!





SPEARMANOV KOEFICIJENT KORELACIJE ρ

- neparametrijski koeficijent korelacije

KADA?

- Dvije ordinalne varijable
- Jedna ili obje numeričke varijable nisu normalno distribuirane
- Prisustvo ekstremnih vrijednosti

LINEARNA REGRESIJA

- ako parovi varijabli pokazuju prisustvo korelacije, funkcionalnu vezu prikazuje JEDNADŽBA REGRESIJE

REGRESIJA - prognoza iz jedne varijable u drugu

linearni slučaj - povezanost varijabli je linearna

- ***jednadžba regresije* je *jednadžba pravca* oko kojeg se grupiraju parovi varijabli u korelacionom dijagramu**

$$y = a + bx$$

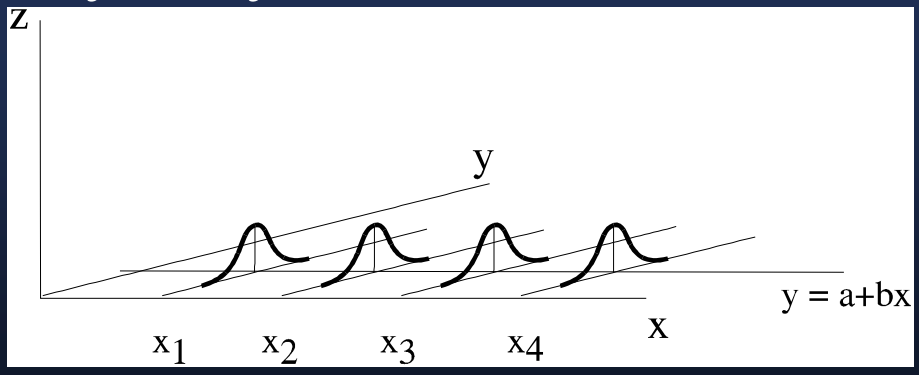
OPĆI OBLIK JEDNADŽBE LINEARNE REGRESIJE

x ... nezavisna varijabla (prediktorska)

y ... zavisna varijabla (kriterijska)

b ... koeficijent smjera

–u realnoj situaciji:



- jednadžba regresijskog pravca dobiva se METODOM NAJMANJIH KVADRATA

y'_i ... vrijednost na regresijskom pravcu koja odgovara x_i

$$\sum_i y_i - y'_i = 0$$

$$\sum_i (y_i - y'_i)^2 = \min$$

iz normalnih jednadžbi

$$\sum_{i=1}^{N} y_i = Na + b\sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i y_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2$$

$$b = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}$$

KOEFICIJENT REGRESIJE

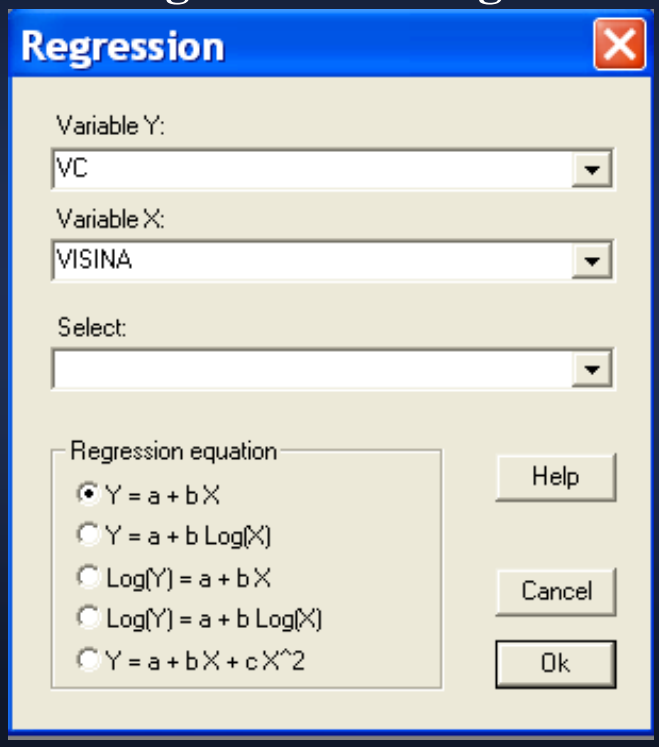
a ... odsječak na ordinati

$$a = \bar{y} - b\bar{x}$$

- pravac regresije izražava "prosječni odnos" ("prosječnu vezu") varijabli x i y

Linearna regresija - MedCalc:

Statistics -> Regression -> Regression



Regression

```
Dependent Y   : VC
Independent X : VISINA

Sample size                          =        33
Coefficient of determination         =         0.8655
Residual standard deviation          =         0.2206

-- REGRESSION EQUATION ----------------------------------------
Y = -11.5374  +  0.0893  X

Parameter     Coefficient      Std.Error       T-value         P
Intercept      -11.53739        1.05028      -10.9851      0.0000
Slope            0.08927        0.00632       14.1213      0.0000

-- ANALYSIS OF VARIANCE ---------------------------------------
Source              DF        Sum of Squares          Mean Square
Regression           1                9.7037                9.7037
Residual            31                1.5085                0.0487

F-Ratio = 199.4107        P = 0.000
```

86% varijabilnosti vitalnog kapaciteta pluća može se objasniti visinom

Sample size	=	33
Coefficient of determination	=	0.8655
Residual standard deviation	=	0.2206

standardna devijacija reziduala (standardna pogreška procjene)


```
-- REGRESSION EQUATION -----------------------------------
Y = -11.5374  +  0.0893  X

Parameter    Coefficient     Std.Error      T-value      P
Intercept      -11.53739       1.05028     -10.9851    0.0000
Slope            0.08927       0.00632      14.1213    0.0000
```

β_0 β_1

```
-- REGRESSION EQUATION ----------------------------------
Y = -11.5374  +  0.0893  X

Parameter     Coefficient     Std.Error      T-value      P
Intercept      -11.53739       1.05028      -10.9851    0.0000
Slope            0.08927       0.00632       14.1213    0.0000
```

vitalni kapacitet pluća=$\beta_0+\beta_1$ * Visina = –11.537+0.089 * Visina

VAŽNO:
Predviđanja se smiju raditi samo za vrijednosti iz postojećeg raspona varijabli!

npr. za visinu 175,

vitalni kapacitet pluća= -11.537+0.089 x 175 = 4.04

razlika SS_T-SS_R; (SS_M); predstavlja poboljšanje u predviđanju zbog korištenja regresijskog modela

suma kvadrata odstupanja od vrijednosti predviđene regresijskim pravcem (SS_R)

```
-- ANALYSIS OF VARIANCE ---------------------------------
Source            DF        Sum of Squares        Mean Square
Regression         1                9.7037              9.7037
Residual          31                1.5085              0.0487

F-Ratio = 199.4107        P = 0.000
```

SS_T - suma kvadrata odstupanja od aritmetičke sredine

regresijski model značajno bolje predviđa zavisnu varijablu od predviđanja aritmetičkom sredinom