Analiza varijanse - ANOVA

Studentov t-test se primenjuje za testiranje razlike jednog para aritmetičkih sredina.
Ukoliko ispitivanje obuhvata više od dve grupe ispitanika čije aritmetičke sredine je potrebno uporediti primenjuje se analiza varijanse – ANOVA, koja testira odnos između varijanse međugrupnog varijabiliteta i varijanse unutargrupnog varijabiliteta. Dobijeni odnos se obeležava sa F po statističaru Fišeru i nosi naziv F-vrednost.
Izračunava se po obrascu:

$$F=\frac{\text{varijansa medjugrupnog varijabiliteta}}{\text{varijansa unutargrupnog varijabiliteta}}=\frac{SD_m^2}{SD_u^2}$$

Ukoliko je F-vrednost statistički značajna to znači da se ispitivane grupe ne ponašaju kao da pripadaju istom skupu ili populaciji.

Primer 1. Posmatrane su tri bolnice gde je od iste bolesti lečeno po 5 ispitanika, a vreme lečenja u danima je iznosilo:

X1	X2	X3
2	6	10
3	7	11
4	8	12
5	9	13
6	10	14
20	40	60

Da li postoji signifikantna razlika u prosečnom vremenu lečenja između bolnica?

Izračuna se prosečno vreme lečenja po jednom bolesniku u svakoj od ispitivanih bolnica:

$$\overline{X}_1=\frac{\Sigma X_1}{n}=\frac{20}{5}=4 \qquad \overline{X}_2=\frac{\Sigma X_2}{n}=\frac{40}{5}=8 \qquad \overline{X}_3=\frac{\Sigma X_3}{n}=\frac{60}{5}=12$$

Zajednička ili opšta aritmetička sredina iznosi:

$$\overline{X}_{total}=\frac{\Sigma X_1+\Sigma X_2+\Sigma X_3}{n_1+n_2+n_3}=\frac{20+40+60}{5+5+5}=\frac{120}{15}=8$$

Unutrašnji varijabilitet svake grupe je zbir kvadrata odstupanja svake njene pojedinačne vrednosti od aritmetičke sredine:

$(\overline{X}_1 - X_1)^2$	$(\overline{X}_2 - X_2)^2$	$(\overline{X}_3 - X_3)^2$
4 1 0 1 4	4 1 0 1 4	4 1 0 1 4
10	10	10

Ukupni unutrašnji varijabilitet - Vu dobija se kada se saberu varijabiliteti svih grupa:

$$Vu = \Sigma(\overline{X}_1 - X_1)^2 + \Sigma(\overline{X}_2 - X_2)^2 + \Sigma(\overline{X}_3 - X_3)^2 = 10 + 10 + 10 = 30$$

Međugrupni varijabilitet - Vm predstavlja zbir kvadrata odstupanja aritmetičkih sredina grupa od zajedničke aritmetičke sredine pomnoženih brojem članova u grupi. Za naš primer:

$$Vm = (\overline{X}_1 - X_t)^2 \cdot n_1 + (\overline{X}_2 - X_t)^2 \cdot n_2 + (\overline{X}_3 - X_t)^2 \cdot n_3$$

$$Vm = (4-8)^2 \cdot 5 + (8-8)^2 \cdot 5 + (12-8)^2 \cdot 5 = 16 \cdot 5 + 0 \cdot 5 + 16 \cdot 5 = 80 + 80 = 160$$

Za izračunavanje varijansi unutargrupnog i međugrupnog varijabiliteta, potrebni su nam stepeni slobode.

Za međugrupni varijabilitet stepen slobode iznosi k-1, gde je k - broj grupa.
U našem primeru: S.S.m=3-1=2

Za unutargrupni varijabilitet stepen slobode iznosi n-k , gde je n - ukupan broj jedinica posmatranja.

U našem primeru: S.S.u=15-3=12

Pristupamo izračunavanju varijansi:

Varijansa međugrupnog varijabiliteta:

$$SD_m^2 = \frac{Vm}{S.S.m} = \frac{160}{2} = 80$$

Varijansa unutargrupnog varijabiliteta:

$$SD_u^2 = \frac{Vu}{S.S.u} = \frac{30}{12} = 2,5$$

Kako je već rečeno F-vrednost se izračunava po obrascu:

$$F = \frac{\text{varijansa medjugrupnog varijabiliteta}}{\text{varijansa unutargrupnog varijabiliteta}} = \frac{SD_m^2}{SD_u^2} = \frac{80}{2,5} = 32$$

F-vrednost od 32 govori da je varijansa međugrupnog varijabiliteta 32 puta veća od varijanse unutargrupnog varijabiliteta.
Pitanje je da li je ova vrednost statistički značajna. Odgovor na ovo pitanje dobija se na osnovu Snedecorovih tablica za granične F-vrednosti, a prema F distribuciji.
Granična F-vrednost se očitava na preseku stepena slobode S.S.u i S.S.m.
Za stepene slobode 12 i 2, kao u našem primeru, i prag značajnosti od 0,05, granična tablična F-vrednost iznosi 3,88.
Kako je naša dobijena F-vrednost, F=32 veća od granične tablične vrednosti, odbacujemo nultu hipotezu i sa sigurnošću većom od 95% tvrdimo da postoji statistički značajna razlika između međugrupnog i unutargrupnog varijabiliteta, odnosno da se grupe ponašaju kao da ne pripadaju istom osnovnom skupu.

I dalje ostaje otvoreno pitanje:
Između kojih grupa je izražena razlika?

Na ovo pitanje odgovor daje post-hok analiza, odnosno primena nekog od testova za procenu F-vrednosti.

Tukey-ov test

Izračunaju se apsolutne vrednosti razlika između svakog para grupnih aritmetičkih sredina:

$$\left|\overline{X}_1 - \overline{X}_2\right| = |4-8| = 4$$

$$\left|\overline{X}_1 - \overline{X}_3\right| = |4-12| = 8$$

$$\left|\overline{X}_2 - \overline{X}_3\right| = |8-12| = 4$$

Izračuna se vrednost D po formuli:

$$D = Q \cdot \sqrt{\frac{SD_u^2}{n}}$$

gde Q predstavlja vrednost koja se očitava u posebnim tablicama za odgovarajuće stepene slobode i željenu značajnost.

U našem primeru za S.S.u=12, S.S.m=2 i p=0,05 vrednost Q iznosi 3,16

Iz toga: $D = 3{,}16 \cdot \sqrt{\frac{2{,}5}{5}} = 2{,}23$

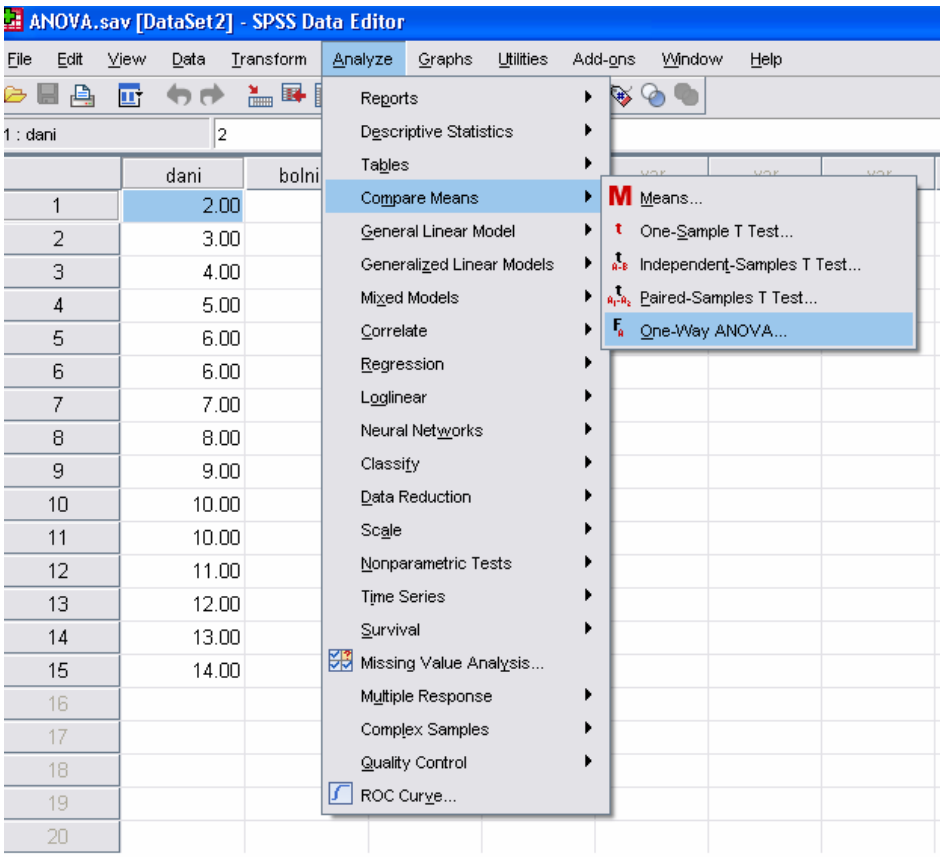
Kako je granična D vrednost 2,23 manja od sve tri vrednosti razlika aritmetičkih sredina između grupa, zaključujemo da postoji statistički značajna razlika između dužine lečenja u sve tri bolnice, odnosno da lečenje u prvoj bolnici traje značajno kraće nego u drugoj i trećoj, a u trećoj traje značajno duže nego u prvoj i drugoj.
Ovo tvrdimo sa sigurnošću većom od 95%.

U SPSS-u zadatak se radi na sledeći način:

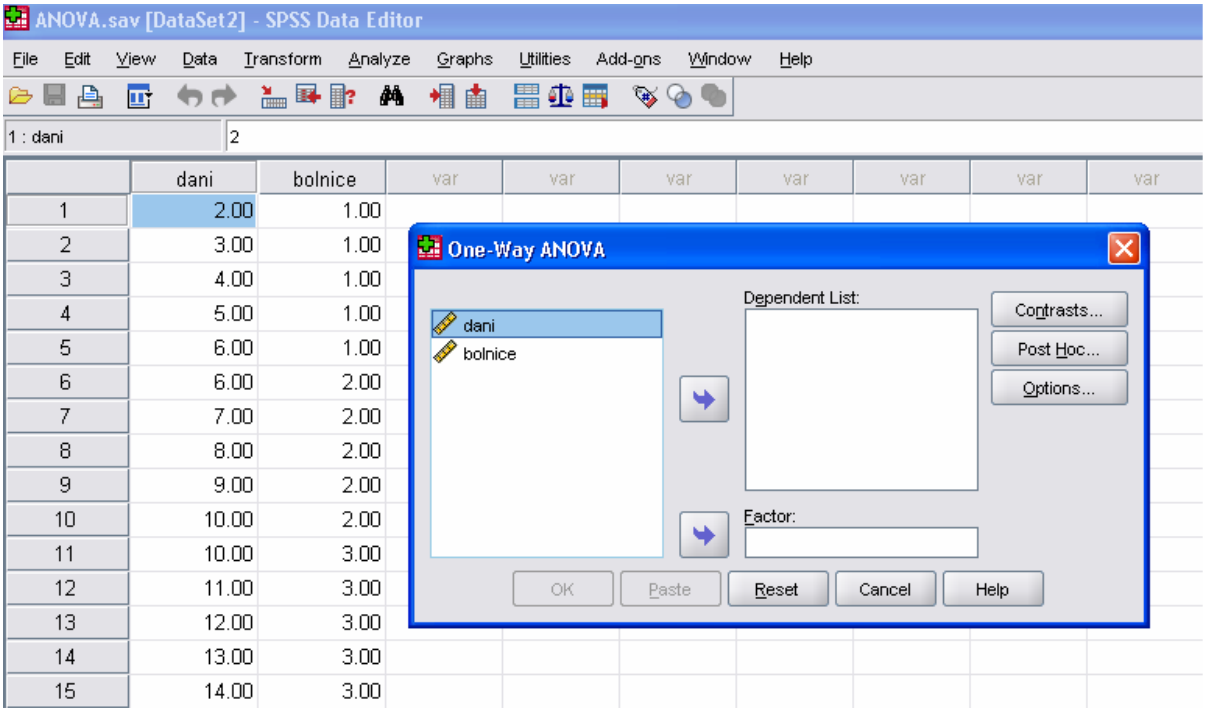
Podatke u SPSS obrazac unosimo tako što ćemo prvu varijablu nazvati dani. U ovu kolonu unećemo broj dana odnosno dužinu lečenja. Druga varijabla koju unosimo je bolnica: sa 1- označićemo prvu bolnicu, sa 2 – drugu bolnicu i sa 3 – treću bolnicu.

Da bi se aktivirala ANOVA U meniju izaberete:

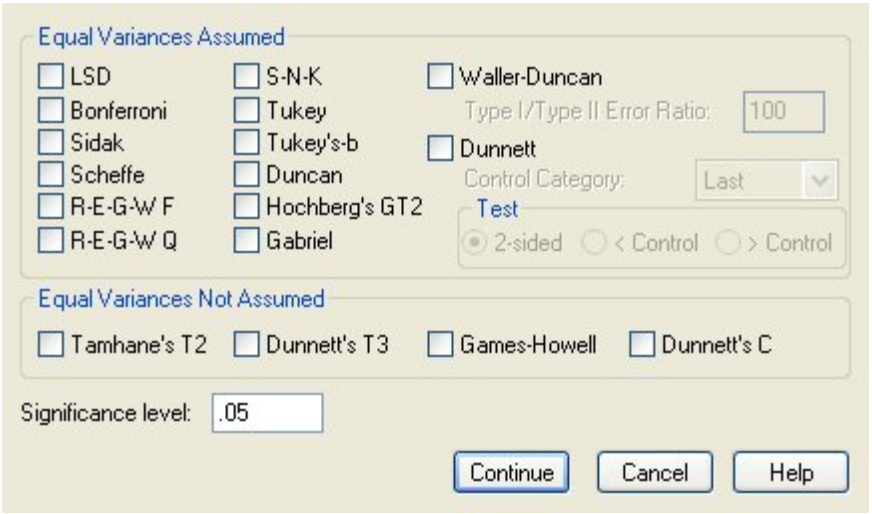
Analyze > Compare Means > One-Way ANOVA...



Nakon toga pojavljuje se sledeći prozor:



U **Dependent list** ubacujete zavisnu varijablu koju ispitujete, to je u ovom primeru varijabla dani, a u **Factor ubacujete** varijablu sa šiframa grupa koje ispitujete, a to su u ovom slučaju bolnice.
Nakon unosa varijabli u predviđena polja, klikne se na dugme Post Hoc i otvara se sledeći prozor:



Kada dobijete rezultate ANOVA analize, radi se post hock analiza da se vidi između koje tačno dve grupe je razlika značajna.
Ako je kod ANOVA-e p bilo <0,05, čekira se neki od testova u Equal variances not assumed
Ako je kod ANOVA-e p bilo >0,05, čekira se neki od testova u Equal variances assumed.

Dobijaju se sledeće tabele sa rezultatima:

ANOVA

dani

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	160.000	2	80.000	32.000	.000
Within Groups	30.000	12	2.500		
Total	190.000	14			

Za Post Hoc analizu dobija se naredna tabela:

Post Hoc

Multiple Comparisons

Dependent Variable:dani

	(I) bolnice	(J) bolnice	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	-4.00000*	1.00000	.005	-6.6679	-1.3321
		3	-8.00000*	1.00000	.000	-10.6679	-5.3321
	2	1	4.00000*	1.00000	.005	1.3321	6.6679
		3	-4.00000*	1.00000	.005	-6.6679	-1.3321
	3	1	8.00000*	1.00000	.000	5.3321	10.6679
		2	4.00000*	1.00000	.005	1.3321	6.6679
Dunnett T3	1	2	-4.00000*	1.00000	.011	-6.9580	-1.0420
		3	-8.00000*	1.00000	.000	-10.9580	-5.0420
	2	1	4.00000*	1.00000	.011	1.0420	6.9580
		3	-4.00000*	1.00000	.011	-6.9580	-1.0420
	3	1	8.00000*	1.00000	.000	5.0420	10.9580
		2	4.00000*	1.00000	.011	1.0420	6.9580
Dunnett t (2-sided)[a]	1	3	-8.00000*	1.00000	.000	-10.5024	-5.4976
	2	3	-4.00000*	1.00000	.003	-6.5024	-1.4976

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

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Post Hoc analiza je pokazala da postoji statistički značajna razlika u dužini lečenja između sve tri bolnice.