Analiza varijanse (ANOVA)

F-raspodela

Neka s_1^2 i s_2^2 predstavljaju uzoračke varijanse za dve populacije. Ako su obe populacije normalno raspoređene, i imaju jednake varijanse, onda se uzoračka raspodela

$$F = \frac{s_1^2}{s_2^2}$$

zove ***F*-raspodela**.

Osobine:

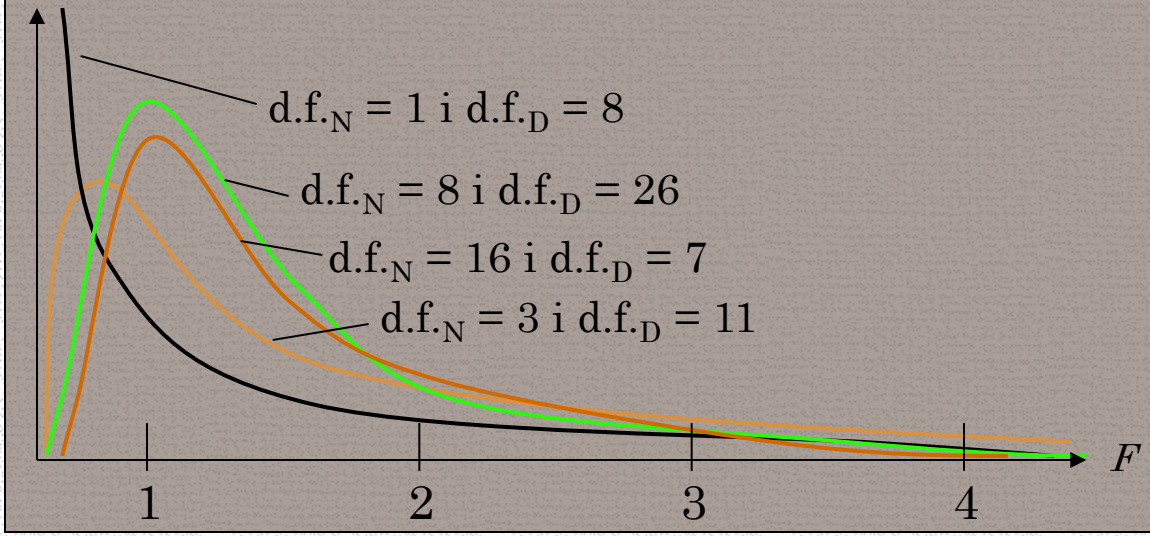
1. *F*-raspodela je familija krivih koje su određene pomoću dva različita stepena slobode:jedan odgovara varijansi u brojiocu $\mathbf{d.f._N}$, a drugi varijansi u imeniocu $\mathbf{d.f._D}$.

F-Raspodela

Osobine:

2. *F*-raspodele su pozitivno zakrivljene
3. Vrednosti *F*-raspodele su uvek nenegativne.
4. Za sve *F*-raspodele, srednja vrednost je približno jednaka 1.

Primeri:



ANOVA

Jednosmerna analiza varijanse je tehnika za upoređivanje srednjih vrednosti kod tri ili više populacija. Obično se koristi skraćenica **ANOVA**.

Tri uslova:

1. Svi uzorci iz približno normalnih populacija.
2. Nezavisni uzorci.
3. Populacije imaju istu varijansu.

Postoje mnoge složenije verzije ANOVA testa (dvosmerna, za zavisne uzorke, itd...)

Jednosmerna ANOVA

Test statistika je količnik varijanse između (between) uzoraka, i varijanse unutar (within) svakog od uzoraka.

1. Varijansa između uzoraka se obeležava sa MS_B i zove **mean square between**.
2. Varijansa unutar uzorka se obeležava sa MS_W i zove **mean square within.**

Što je veći ovaj odnos, veća je verovatnoća da između uzoraka postoji razlika u srednjoj vrednosti, tj. test govori u prilog odbacivanja nulte hipoteze.

Jednosmerna ANOVA

Ukoliko su uslovi zadovoljeni, raspodela je u skladu sa F-raspodelom.

Test statistika je

$$F = \frac{MS_B}{MS_W}.$$

Broj stepeni slobode je

$$\text{d.f.}_N = k - 1$$

i

$$\text{d.f.}_D = N - k$$

Gde je k broj uzoraka (grupa) a N suma svih obima.

Test statistika kod jednosmerne ANOVA-e

Računanje test statistike:

1. Određujemo AS i varijansu za svaki uzorak.

$$\bar{x} = \frac{\sum x}{n} \qquad s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

2. Određujemo AS za sve uzorke zajedno (zbir obima je N)

$$\bar{\bar{x}} = \frac{\sum x}{N}$$

3. Nalazimo sumu kvadrata između uzoraka (between)

$$SS_B = \sum n_i (\bar{x}_i - \bar{\bar{x}})^2$$

4. Nalazimo sumu kvadrata unutar uzoraka (within)

$$SS_W = \sum (n_i - 1) s_i^2$$

Test statistika kod ANOVA-e

Nastavak računanja:

5. Određujemo varijansu između uzoraka (between)

$$MS_B = \frac{SS_B}{k-1} = \frac{SS_B}{\text{d.f.}_\text{N}}$$

6. Određujemo varijansu unutar uzoraka (within)

$$MS_W = \frac{SS_W}{N-k} = \frac{SS_W}{\text{d.f.}_\text{D}}$$

7. Njihov količnik je test statistika sa F-raspodelom

$$F = \frac{MS_B}{MS_W}$$

ANOVA sumirana tabela

- Rezultati ANOVA testa se obično prikazuju ovakvom sumiranom tabelom.

Variation	**Sum of squares**	**Degrees of freedom**	**Mean squares**	***F***
Between	SS_B	d.f._N	$MS_B = \frac{SS_B}{\text{d.f.}_\text{N}}$	$MS_B \div MS_W$
Within	SS_W	d.f._D	$MS_W = \frac{SS_W}{\text{d.f.}_\text{D}}$	

- Zaključivanje na osnovu p-vrednosti: ako je p manje od α, postoji razlika između grupa.
- U tom slučaju, post-hoc testovi za ispitivanje pojedinih parova.

Primer ANOVA-e

Primer:
Tabela prikazuje godišnje zarade slučajno odabranih stanovnika 4 US grada. Sa pragom $\alpha = 0.05$, proveriti da li između ovih gradova postoji razlika u prosečnim platama?

Pittsburgh	Dallas	Chicago	Minneapolis
27,800	30,000	32,000	30,000
28,000	33,900	35,800	40,000
25,500	29,750	28,000	35,000
29,150	25,000	38,900	33,000
30,295	34,055	27,245	29,805

Primer ANOVA-e

Formulacija hipoteza:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Bar jedna srednja vrednost se razlikuje od ostalih. (tvrdnja)

Budući da imamo k = 4 uzorka, $\text{d.f.}_N = k - 1 = 4 - 1 = 3$.

Suma obima uzoraka je
$N = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 5 + 5 = 20$.

$\text{d.f.}_D = N - k = 20 - 4 = 16$

Za zadato $\alpha = 0.05$, $\text{d.f.}_N = 3$, i $\text{d.f.}_D = 16$,
Kritična vrednost F-raspodele je $F_0 = 3.24$.

Primer ANOVA-e

Test statistika:

Da bi izračunali vrednost test statistike, prvo računamo sledeće:

$$\bar{\bar{x}} = \frac{\sum x}{N} = \frac{140745 + 152705 + 161945 + 167805}{20} = 31160$$

$$MS_B = \frac{SS_B}{\text{d.f.}_\text{N}} = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$= \frac{5(28149 - 31160)^2 + 5(30541 - 31160)^2}{4-1} +$$

$$\frac{5(32389 - 31160)^2 + 5(33561 - 31160)^2}{4-1}$$

$$\approx 27874206.67$$

Primer ANOVA-e

Nastavak računanja:

$$MS_W = \frac{SS_W}{\text{d.f.}_\text{D}} = \frac{\sum(n_i - 1)s_i^2}{N - k}$$

$$\approx \frac{(5-1)(3192128.94) + (5-1)(13813030.08)}{20-4} +$$

$$\frac{(5-1)(24975855.83) + (5-1)(17658605.02)}{20-4}$$

$$= 14909904.97$$

$$F = \frac{MS_B}{MS_W} = \frac{27874206.67}{14909904.34} \approx 1.870$$

Test statistika: 1.870 < Kritična vrednost: 3.24.

Ne odbacujemo H_0.

Zaključak: Sa pragom značajnosti 5% ne možemo potvrditi hipotezu da se prosečne plate razlikuju u posmatrana 4 grada.