Grundläggande matematisk statistik

Hypotestest
Del III: ANOVA

Uwe Menzel, 2017
uwe.menzel@matstat.org
www.matstat.org

ANOVA

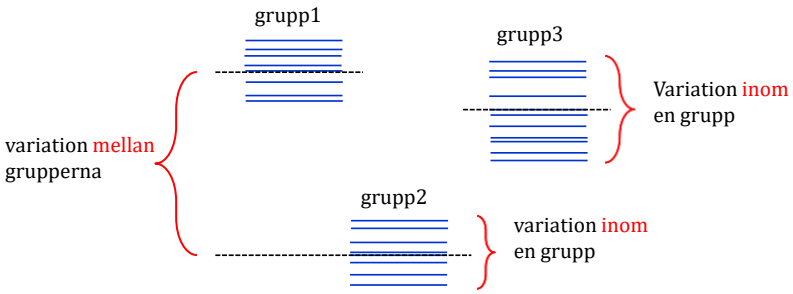
- testar om flera (> 2) normalfördelade populationer (”grupper”) har samma väntevärde (utvidgning av t-testet till fler än 2 grupper)
- därför utnyttjas populationers empiriska varianser (!) (ANOVA = ANalysis Of VAriance)
- Nollhypotes: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (k stickprov)
- alternativ hypotes: minst ett likhetstecken gäller inte, dvs. minst en population har ett avvikande väntevärde
- (ANOVA är alltså ett parametriskt test)
- testet säger ingenting om vilket väntevärde som avviker ifall H_0 förkastas → därför behövs ett så kallad post-hoc test

$$X_{n,i} \sim N(\mu_n, \sigma)$$ fler än 2 populationer

ANOVA

Testet utförs genom att jämföra spridningen mellan grupperna med spridningen inom grupperna.

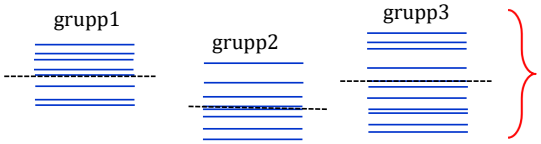
Observationer för tre grupper:



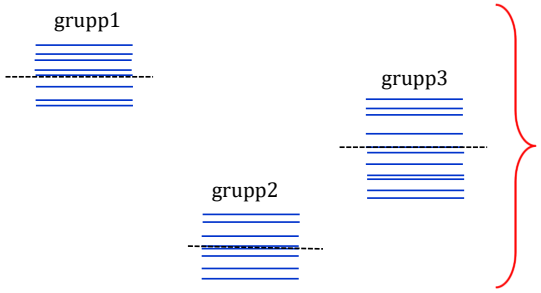
Intuition: grupperna är olika om spridningen mellan grupperna är betydligt större än spridningen inom grupperna.

ANOVA

Intuition: grupperna är olika om spridningen mellan grupperna är betydligt större än spridningen inom grupperna.



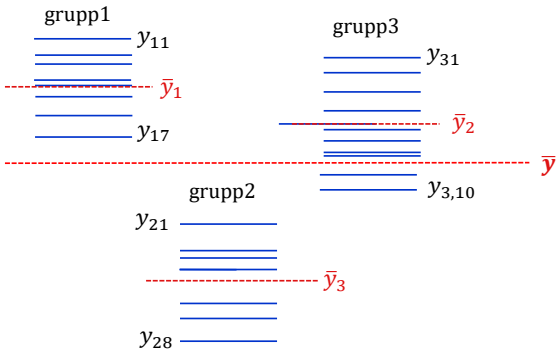
Troligtvis ingen signifikant skillnad mellan väntevärdena (H_0 kan **inte** förkastas), små skillnader kan enbart bero på slumpen.



Populationerna har troligtvis **inte** samma väntevärden (H_0 förkastas inte), skillnaderna kan vara signifikanta.

ANOVA: beteckningar

- y_{ij} : observation j för grupp i
  - $i = 1 \ldots k$ (k grupper)
  - $j = 1 \ldots n_i$ (n_i observationer för grupp i)
- $\bar{y}_i$: medelvärde för grupp i
- $\bar{y}$: medelvärde över alla grupper (”grand mean”)



ANOVA: ”Total Sum of Squares”

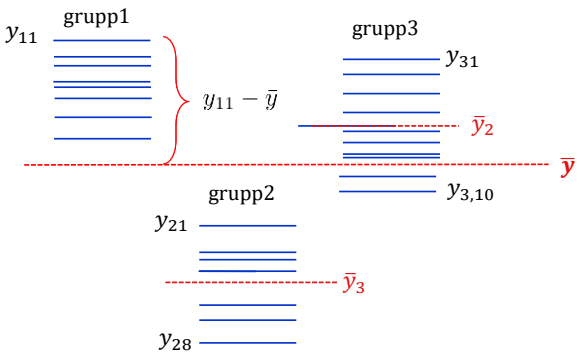
För att kvantifiera variationerna används kvadratsummer (”Sum of Squares”):

$$SS_{tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

y_{ij} : gruppg i ; observation j

$\bar{y}$: ”grand mean”

n_i : antalet observationer för grupp i



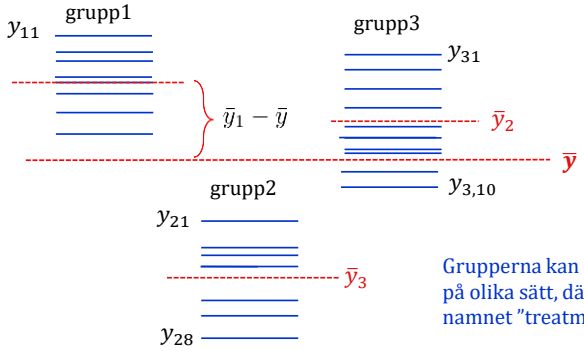
ANOVA: ”Sum of Squares for Treatments”

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2$$

$\bar{y}_i$: medelvärde för grupp i

$\bar{y}$: ”grand mean”

n_i : antalet observationer för grupp i



Grupperna kan ha ”behandlats” på olika sätt, därför också namnet ”treatment”.

www.matstat.org

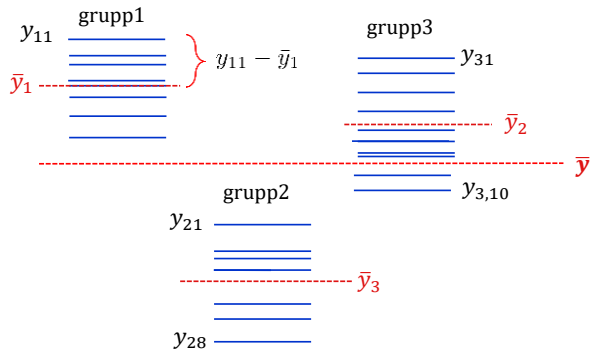
ANOVA: ” Sum of Squares for Error”

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$$

y_{ij} : grupp i ; observation j

$\bar{y}_i$: medelvärde för grupp i

n_i : antalet observationer för grupp i



www.matstat.org

ANOVA

Man kan visa att följande sammanhang gäller:

$$SS_{tot} = SS_{tr} + SS_{err}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

SS_{tot} : "Total Sum of Squares"

SS_{tr} : "Sum of Squares for Treatments"

SS_{err} : "Sum of Squares for Error"

y_{ij} : grupp i ; observation j

$\bar{y}$: "grand mean"

$\bar{y}_i$: medelvärde för grupp i

n_i : antalet observationer för grupp i

$\sum_{i=1}^{k} n_i = n$ totala antalet observationer

ANOVA: Testvariabel

$$SS_{tr} = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 \qquad \frac{1}{\sigma^2}\cdot SS_{tr} \sim \chi^2(k-1)$$

chi-kvadrat-fördelning med $k-1$ frihetsgrader

$$SS_{err} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \qquad \frac{1}{\sigma^2}\cdot SS_{err} \sim \chi^2(n-k)$$

chi-kvadrat-fördelning med $n-k$ frihetsgrader

Allmänt gäller: $\dfrac{\frac{\chi^2(n)}{n}}{\frac{\chi^2(m)}{m}} \sim F(n,m)$

F-fördelning med n frihetsgrader i täljaren och m frihetsgrader i nämnaren

Om nollhypotesen är sann gäller därför:

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} \sim F(k-1, n-k)$$

F-fördelning med n frihetsgrader i täljaren och m frihetsgrader i nämnaren

(Om nollhypotesen är sann kommer $\bar{y}_i$:orna från samma fördelning)

ANOVA: Testvariabel

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} \sim F\,(k-1, n-k)$$

F-fördelning med n frihetsgrader i täljaren och m frihetsgrader i nämnaren

$$SS_{tr} = \sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{y})^2 \qquad SS_{err} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- F-fördelningen antar bara positiva värden (kvot av kvadratsummor)
- testvariabeln F blir stor om avvikelserna av gruppmedelvärdena från det globala medelvärdet (=täljare) blir stor jämfört med avvikelserna inom grupperna (=nämnare).
- med andra ord: testvariabeln F blir desto större ju mer någon grupps medelvärde avviker från de andra (SS_{tr} blir större)**.**
- För stora värden av en observation för F borde därför nollhypotesen förkastas.
- Som förut för Z-testet och t-testet (föreläsning **F12**) förkastas därför nollhypotesen på signifikansnivån α om en observation för F överskrider respektive α-kvantil för F.

ANOVA: Kritiskt område

För en på förhand fastlagd signifikansnivå α kan det kritiska värdet ω_α för testvariabeln F beräknas genom att kräva att sannolikheten för felet typ I ska vara lika med α:

$$P\left(\frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} > \omega_\alpha \mid H_0\ sann\right) = \alpha$$

ekvation för att beräkna ω_α för förbestämd α (= sannolikhet för fel typ I)

H_0 förkastas

fel typ I

Under H_0 är termen på vänster sida i parentesen F-fördelad, vi kan alltså skriva (därmed är betingningen "H_0 *sann*" inräknad):

$$P\,(F > \omega_\alpha) = \alpha \quad \text{med} \quad F \sim F\,(k-1, n-k)$$

Allmänt gäller: $P\,(F > F_\alpha\,(k-1, n-k)) = \alpha$

$F_\alpha(k-1, n-k)$: kvantil för F-fördelning med $k-1$ resp. $n-k$ frihetsgrader

Genom att jämföra de sista båda utrycken får vi $\omega_\alpha = F_\alpha(k-1, n-k)$. Nollhypotesen förkastas alltså om en observation för F $(= f_{obs})$ blir större än $F_\alpha(k-1, n-k)$.

ANOVA: Kritiskt område

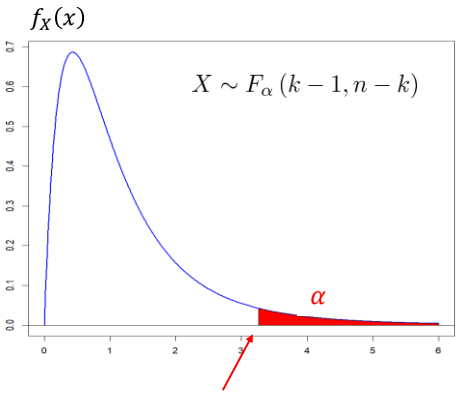
H_0 förkastas om

$$f_{obs} > F_\alpha(k-1, n-k)$$

Kritiskt område, signifikansnivå α :

$$\Omega_\alpha = \{f_{obs} > F_\alpha\,(k-1, n-k)\}$$

Om en observation för F hamnar i det kritiska området (röd) förkastas nollhypotesen.



$F_\alpha(k-1, n-k)$:
kvantil för F-fördelning med $k-1$ resp. $n-k$ frihetsgrader

Antaganden:

- $X_i \sim N$ (eller ungefär så)
- $\sigma_i = \sigma$ (eller ungefär så)
- oberoende stickprovsvärden

ANOVA

Ibland har man inte alla observationer y_{ij} tillgängliga, utan bara stickprovsstorlekar, medelvärden och standardavvikelser för varje grupp. I så fall beräknas kvadratsummorna på följande sätt:

$$\bar{y} = \frac{\sum n_i \cdot \bar{y}_i}{\sum n_i}$$

$$SS_{tr} = \sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{y})^2$$

n_i: antalet observationer för grupp i

$\bar{y}_i$: medelvärde för grupp i

s_i: standardavvikelse för grupp i

$$SS_{err} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2$$

se appendixet för härledning

ANOVA: antaganden

- Alla observationer måste vara oberoende av varandra.
- Populationerna måste (ungefär) vara normalfördelade.
  - Kolmogorow-Smirnow testet, Shapiro-Wilk-testet, eller Anderson-Darling-testet kan användas för att testa om fördelningarna avviker signifikant från normalfördelningen.
- Varianserna av alla grupper måste vara lika. Har man samma antal observationer i varje grupp måste detta bara gälla ungefärligt.
  - Levene testet eller Bartlett's test kan användas för att testa om varianserna skiljer sig signifikant.

Vilket medelvärde avviker ?

- Ett post-hoc test kan svara på denna fråga (efter att H_0 förkastas i ANOVA), t. ex. Tukey's test.
- Tukey's test gör parvisa jämförelser och korrigerar samtidigt för "multiple testing".

ANOVA: exempel, 4 grupper

$\alpha = 0.05$

A	B	C	D
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

$$n_1 = 6\,; \quad n_2 = 7\,; \quad n_3 = 6\,; \quad n_4 = 4$$

$$\bar{y}_1 = 75.67\,; \quad \bar{y}_2 = 78.43\,; \quad \bar{y}_3 = 70.83\,; \quad \bar{y}_4 = 87.75$$

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + n_3 \cdot \bar{y}_3 + n_4 \cdot \bar{y}_4}{n_1 + n_2 + n_3 + n_4} = \frac{1179}{23} = 77.35$$

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = 712.6 \qquad M_{tr} = \frac{SS_{tr}}{k-1} = 237.5$$

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 = 1196.6 \qquad M_{err} = \frac{SS_{err}}{n-k} = 63.0$$

$$f_{obs} = \frac{M_{tr}}{M_{err}} = \frac{237.5}{63.0} = 3.77$$

$$\Omega_\alpha = \{F > F_\alpha\,(k-1, n-k)\} = \{F > F_{0.05}(3, 19)\} = \{F > 3.13\}$$

Nollhypotesen förkastas. Minst ett väntevärde skiljer sig signifikant från de andra.

uwe.menzel@matstat.org

ANOVA: exempel, alternativ beräkning

A	B	C	D
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

⇨

	A	B	C	D
n_i	6	7	6	4
$\bar{y}_i$	75.67	78.43	70.83	87.75
s_i^2	66.67	50.62	91.77	33.58

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + n_3 \cdot \bar{y}_3 + n_4 \cdot \bar{y}_4}{n_1 + n_2 + n_3 + n_4} = \frac{1179}{23} = 77.35$$

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = 712.8 \qquad M_{tr} = \frac{SS_{tr}}{k-1} = \frac{712.8}{3} = 237.5$$

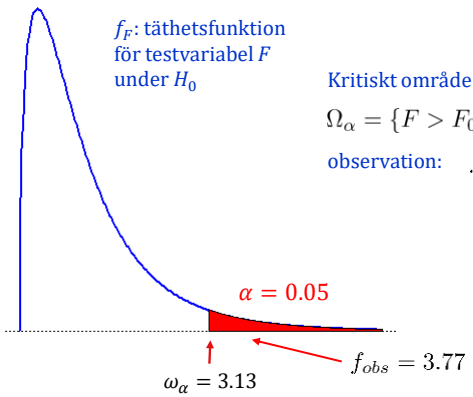
$$SS_{err} = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2 = 1196.66 \qquad M_{err} = \frac{SS_{err}}{n-k} = \frac{1196.66}{19} = 63$$

alternativ formel (se appendix)

$$f_{obs} = \frac{M_{tr}}{M_{err}} = \frac{237.5}{63.0} = 3.77$$

$$\Omega_\alpha = \{F > F_\alpha\,(k-1, n-k)\} = \{F > F_{0.05}(3, 19)\} = \{F > 3.13\}$$

ANOVA: exempel 4 grupper



Observationen för testvariabeln F överskrider det kritiska värdet. Nollhypotesen förkastas därför. Minst ett väntevärde avviker signifikant från de andra. (signifikansnivå $\alpha = 0.05$)

ANOVA med R

```
data(InsectSprays)
levels(InsectSprays$spray)
summary(InsectSprays$count)
boxplot(count ~ spray, data = InsectSprays, col="green")

# 1. funktion "oneway.test"
oneway.test(count ~ spray, data = InsectSprays)

# Olika varianser i grupperna?
bartlett.test(count ~ spray, data = InsectSprays)  #  olika – problem!

# Icke-parametriskt test:
kruskal.test(count ~ spray, data = InsectSprays)

# 2. alternativ funktion för ANOVA: aov
aov.out = aov(count ~ spray, data = InsectSprays)
summary(aov.out)

TukeyHSD(aov.out)  # post-hoc test
plot(TukeyHSD(aov.out))  # parvisa differenser – signifikant skillnad om KI:et inte omfattar nollan.
```


Appendix

Hypotestest
Del 3: ANOVA

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Alternativ formel för SS_{err}

$$SS_{err} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$ dra isär summan över i

$$SS_{err} = \underbrace{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2}_{(n_1-1)\cdot s_1^2} + \underbrace{\sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}_{(n_2-1)\cdot s_2^2} + \ldots + \underbrace{\sum_{j=1}^{n_k}(y_{kj} - \bar{y}_k)^2}_{(n_k-1)\cdot s_k^2}$$

ty $s_i^2 = \frac{1}{n_i - 1}\cdot\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)$ varians för grupp i. Det summeras över alla observationer för gruppen (över index j)

t. ex. $s_1^2 = \frac{1}{n_1 - 1}\cdot\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)$

y_{ij} : grupp i ; observation j
$\bar{y}_i$: medelvärde för grupp i
n_i: antalet observationer i grupp i

Om man skriver det sista uttrycket för SS_{err} som en summa igen följer:

$$SS_{err} = \sum_{i=1}^{k}(n_i - 1)\cdot s_i^2$$

F-test, fördelning för testvariabeln

$$\frac{SS_{err}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$ dra isär summan över i

$$\frac{SS_{err}}{\sigma^2} = \underbrace{\frac{1}{\sigma^2}\cdot\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)}_{\sim\chi^2(n_1-1)} + \underbrace{\frac{1}{\sigma^2}\cdot\sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)}_{\sim\chi^2(n_2-1)} + \ldots + \underbrace{\frac{1}{\sigma^2}\cdot\sum_{j=1}^{n_k}(y_{kj} - \bar{y}_k)}_{\sim\chi^2(n_k-1)}$$

$\Rightarrow$ $\frac{SS_{err}}{\sigma^2} \sim \chi^2(n-k)$ för att $\sum_{i=1}^{k}(n_i - 1) = n - k$

med $\sum_{i=1}^{k} n_i = n$

$Y_{n,i} \sim N(\mu, \sigma)$ under H_0

y_{ij} : grupp i ; observation j
$\bar{y}_i$: medelvärde för grupp i
n_i: antalet observationer i grupp i

F-test, fördelning för testvariabeln

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

$$\frac{SS_{tot}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \sim \chi^2(n-1) \quad \text{med} \quad \sum_{i=1}^{k} n_i = n$$

$$\underbrace{\frac{SS_{tot}}{\sigma^2}}_{\sim \chi^2(n-1)} = \frac{SS_{tr}}{\sigma^2} + \underbrace{\frac{SS_{err}}{\sigma^2}}_{\sim \chi^2(n-k)}$$

$$\Rightarrow \quad \frac{SS_{tr}}{\sigma^2} \sim \chi^(k-1)$$

därför att: adderas χ^2-fördelade slumpvariabler, så adderas deras frihetsgrader (χ^2 är ”reproduktiv”)

F-Test, fördelning för testvariabeln

$$\frac{SS_{tr}}{\sigma^2} \sim \chi^(k-1)$$

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} = \frac{\dfrac{SS_{tr}}{\sigma^2 \cdot (k-1)}}{\dfrac{SS_{err}}{\sigma^2 \cdot (n-k)}} \sim \frac{\dfrac{\chi^2(k-1)}{(k-1)}}{\dfrac{\chi^2(n-k)}{(n-k)}} \sim F\,(k-1, n-k)$$

$$\frac{SS_{err}}{\sigma^2} \sim \chi^2(n-k)$$