Lecture 31 Analysis of Variance (ANOVA)

BIO210 Biostatistics

Xi Chen

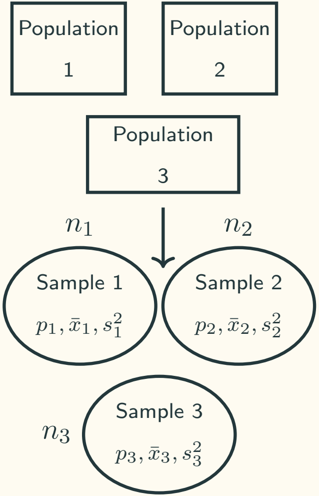
Fall, 2024

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Compare More Than Two Means

More than two-samples



Intuitive way: compare all possible pairs using two-sample independent t test:

Samples 1 vs 2: $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$

Samples 1 vs 3: $H_0 : \mu_1 = \mu_3$; $H_1 : \mu_1 \neq \mu_3$

Samples 2 vs 3: $H_0 : \mu_2 = \mu_3$; $H_1 : \mu_2 \neq \mu_3$

Good enough?

Compare More Than Two Means

What if we have 15 samples from 15 different populations ?

- **Intuitive way:** compare all possible pairs using two-sample independent t test:
  - Number of comparisons: $\binom{15}{2} = \frac{15\times 14}{2} = 105$
  - Significance level: $\alpha = 0.05$
  - When we set $\alpha = 0.05$, we want to tolerate a 5% of chance of making a type I error. That is, the number of tests that have made a type I error: ≈ 5
- Assume that the means are all the same, what is the probability of making a type I error in at least one test ?

$$\begin{aligned}
&\mathbb{P}\left(\text{reject } H_0 \text{ in at least one test} \,\middle|\, H_0 \text{ is true}\right) \\
=&1 - \mathbb{P}\left(\text{not rejecting } H_0 \text{ in all tests} \,\middle|\, H_0 \text{ is true}\right) \\
=&1 - 0.95^{105} \\
=&0.995
\end{aligned}$$

Source of Variation - Total

Sample 1	Sample 2	Sample 3
3	5	5
2	3	6
1	4	7
$\bar{x}_1 = 2$	$\bar{x}_2 = 4$	$\bar{x}_3 = 6$

sum of squares (SS): add up the squared distance between an observation and the mean:

$$\sum(\boldsymbol{X} - \bar{\boldsymbol{X}})^2$$

SST: total sum of squares

The grand mean: $\bar{\bar{x}} = \dfrac{3+2+1+5+3+4+5+6+7}{9} = 4$

SST $= (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 30$

What is the df ? $df_T = 9 - 1 = 8$

Source of Variation - Within Groups

Sample 1	Sample 2	Sample 3
3	5	5
2	3	6
1	4	7
$\bar{x}_1 = 2$	$\bar{x}_2 = 4$	$\bar{x}_3 = 6$

sum of squares (SS): add up the squared distance between an observation and the mean:

$$\sum (\boldsymbol{X} - \bar{\boldsymbol{X}})^2$$

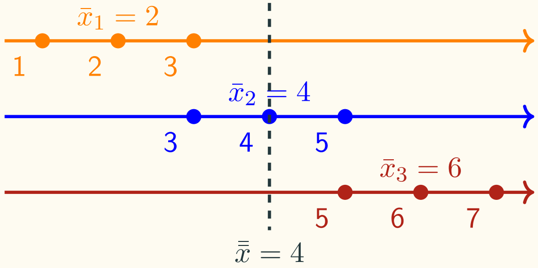
SSW: sum of squares within

$$\begin{aligned} \text{SSW} &= (3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 \\ &= df_1 \cdot s_1^2 + df_2 \cdot s_2^2 + df_3 \cdot s_3^2 \\ &= 6 \end{aligned}$$

What is the df ? $df_W = (3-1) + (3-1) + (3-1) = 6$

Source of Variation - Between Groups

Sample 1	Sample 2	Sample 3
3	5	5
2	3	6
1	4	7
$\bar{x}_1 = 2$	$\bar{x}_2 = 4$	$\bar{x}_3 = 6$



SSB: sum of squares between

$$
\begin{aligned}
\text{SSB} &= (2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2 \\
&= n_1 \cdot (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 \cdot (\bar{x}_2 - \bar{\bar{x}})^2 + n_3 \cdot (\bar{x}_3 - \bar{\bar{x}})^2 \\
&= 24
\end{aligned}
$$

What is the df ? $\quad df_B = 3 - 1 = 2$

Summary of The Source of Variation

Sample 1	Sample 2	Sample 3
3	5	5
2	3	6
1	4	7
$\bar{x}_1 = 2$	$\bar{x}_2 = 4$	$\bar{x}_3 = 6$

Variance-like

Source of Variation	SS (sum of squares)	*df*	MS (mean square)
Between	24	2	12
Within	6	6	1
Total	30	8	[redacted]

Multiple Samples From Multiple Populations

Population 1	Sample 1 $(n_1, \bar{x}_1, s_1^2)$
Population 2	Sample 2 $(n_2, \bar{x}_2, s_2^2)$
Population 3	Sample 3 $(n_3, \bar{x}_3, s_3^2)$
$\vdots$	$\vdots$
Population k	Sample k $(n_k, \bar{x}_k, s_k^2)$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$

The ANOVA Table

Source of Variation	SS	*df*	MS
Between	$\mathrm{SSB} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$	$k - 1$	$\mathrm{MSB} = \frac{\mathrm{SSB}}{k-1}$
Within	$\mathrm{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{k} df_i s_i^2$	$n - k$	$\mathrm{MSW} = \frac{\mathrm{SSW}}{n-k}$
Total	$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2 = \mathrm{SSB} + \mathrm{SSW}$	$n - 1$	

The F-test

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: & \text{not all equal} \end{cases} \Leftrightarrow \begin{cases} H_0: & \text{The main variation is from SSW} \\ H_1: & \text{The main variation is from SSB} \end{cases}$$

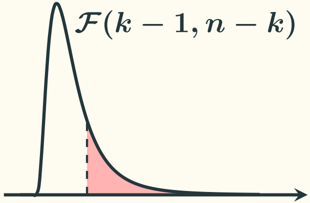
Under the null hypothesis:

$$\frac{\mathrm{SSB}}{\sigma^2} \sim \boldsymbol{\chi}^{\mathbf{2}}(k-1) \text{ and } \frac{\mathrm{SSW}}{\sigma^2} \sim \boldsymbol{\chi}^{\mathbf{2}}(n-k) \text{ , where } \sigma^2 \text{ is the common variance}$$

The test statistic:

$$F = \frac{\frac{\mathrm{SSB}}{(k-1)\sigma^2}}{\frac{\mathrm{SSW}}{(n-k)\sigma^2}} = \frac{\mathrm{MSB}}{\mathrm{MSW}} \sim \boldsymbol{\mathcal{F}}(k-1, n-k)$$

p-value: $\mathbb{P}\left(\text{data} \mid H_0 \text{ is true}\right) = \mathbb{P}\left(F_{k-1,n-k} \geqslant \frac{\mathrm{MSB}}{\mathrm{MSW}}\right)$



Summary of an ANOVA result

Source of Variation	SS	*df*	MS	F	*p*-value
Between	$\text{SSB} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	$\text{MSB} = \frac{\text{SSB}}{k-1}$	$\frac{\text{MSB}}{\text{MSW}}$	$\mathbb{P}\left(F \geqslant \frac{\text{MSB}}{\text{MSW}}\right)$
Within	$\text{SSW} = \sum_{i=1}^{k} df_i s_i^2$	$n-k$	$\text{MSW} = \frac{\text{SSW}}{n-k}$		
Total	$\text{SST} = \text{SSB} + \text{SSW}$	$n-1$			