Lecture 31 Analysis of Variance (ANOVA)

BIO210 Biostatistics

Xi Chen

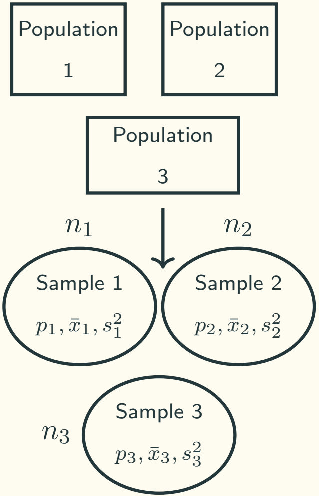
Spring, 2022

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: $\binom{15}{2} = \dfrac{15 \times 14}{2} = 105$
  - 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 intended number of tests of making 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 ?

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

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:

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

SS_T: total sum of squares

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

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

What is d.f. ? $df_T = 9 - 1 = 8$

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:

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

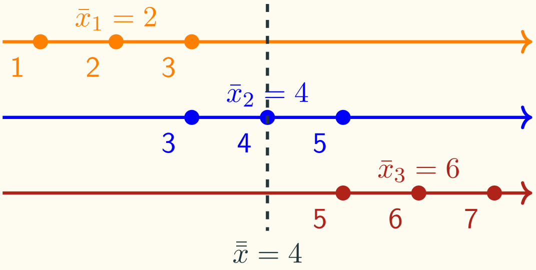
SS_W: sum of squares within

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

What is d.f. ? $df_W = (3-1) + (3-1) + (3-1) = 6$

Source of Variation - Between Groups

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



SS_B: sum of squares between

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

What is d.f. ? $df_B = 3 - 1 = 2$

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

Source of Variation	**SS (sum of squares)**	**d.f.**	**MS (mean square)**
Between	24	2	12
Within	6	6	1
Total	30	8	

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

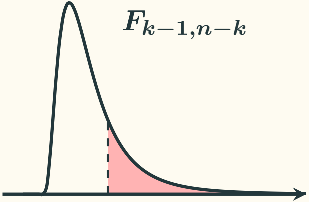
Source of Variation	SS	d.f.	MS
Between	$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	$s_B^2 = \dfrac{SS_B}{k-1}$
Within	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{k} df_i s_i^2$	$n-k$	$s_W^2 = \dfrac{SS_W}{n-k}$
Total	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2 = SS_B + SS_W$	$n-1$	

F-test

$$F = \frac{s_B^2}{s_W^2} \cdot \frac{\sigma_W^2}{\sigma_B^2}$$

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: & \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k \end{cases} \Leftrightarrow \begin{cases} H_0: & \text{Main var. is from } SS_W \\ H_1: & \text{Main var. is from } SS_B \end{cases} \Leftrightarrow \begin{cases} H_0: & \dfrac{\sigma_W^2}{\sigma_B^2} \geqslant 1 \\ H_1: & \dfrac{\sigma_W^2}{\sigma_B^2} < 1 \end{cases}$$

$$p\text{-value: } P(\text{data} \mid H_0 \text{ is true}) = P\left(F = \frac{s_B^2}{s_W^2} \cdot \frac{\sigma_W^2}{\sigma_B^2}\right) = P\left(F_{k-1,n-k} \geqslant \frac{s_B^2}{s_W^2}\right)$$



Summary of an ANOVA result

Source of Variation	SS	d.f.	MS	F	p-value
Between	$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	$s_B^2 = \dfrac{SS_B}{k-1}$	$\dfrac{s_B^2}{s_W^2}$	$P\left(F \geqslant \dfrac{s_B^2}{s_W^2}\right)$
Within	$SS_W = \sum_{i=1}^{k} df_i s_i^2$	$n-k$	$s_W^2 = \dfrac{SS_W}{n-k}$		
Total	$SS_T = SS_B + SS_W$	$n-1$			