CHAPTER 12 — multi-sample inference (ANOVA)

Friday, 2 January 2026, 8:53 PM

* In this chapter we will be comparing multiple groups (all normally distributed) of a population

① Comparing the means of all groups and deciding if at least one of them is different

② Comparing the means of specific groups among all groups

* * * If we have a population of students studying Biostatistics



we then have N groups (sections) in this Population → all having the same n number of students

→ y_{ij} is the jth observation in the ith group

* if you are the y_{ij} observation then you are the jth student in the ith group. if you want to decide whether you did well or not in the exam then you will :

Compare your mark with the mean of marks of your group (you are comparing your mark within your section only)

Compare your mark with the mean of marks of all groups (you are making a comparison between your marks and other ones)

* Note that the variance of all groups is the same

* Note that the groups are normally distributed

y_{ij} is defined by :-

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

- i → the ith group
- j → the jth observation of ith group
- μ → the overall mean of all observations of all groups
- α_i → tells how the mean of group i differs from the overall mean
- e_{ij} → the deviation of the observation j from its group mean; $e_{ij} \sim N(0, \sigma^2)$

→ it is normally distributed $\sim N(\mu + \alpha_i, \sigma^2)$

→ the mean of group i is $\mu + \alpha_i$

if $\alpha_i > 0$ then the mean is larger than the overall mean

if $\alpha_i < 0$ then the mean is smaller than the overall mean

→ So in one-way ANOVA we are studying the differences between α_i's compared to e_{ij}'s

α_i's → differences between groups

e_{ij}'s → differences within the groups

① Comparing the means of all groups and deciding if at least one of them is different

H_0 : the means are all the same vs H_1 : one mean is different from others

→ Some important clarifications :-

* $\bar{y}_i$ → the mean of the ith group
* $\bar{\bar{y}}$ → the mean of all groups
* $\bar{y}_{ij} - \bar{y}_i$ → the deviation of an individual observation from its group mean → this indicates the within-group variability → it is the term e_{ij}
* $\bar{y}_i - \bar{\bar{y}}$ → represents the deviation of a group mean from the overall mean → this indicates the between-group variability → it is the term α_i

What is the difference between k and n ???

if we have the observations :- 3, 2, 5, 4, 3, 6

→ number of observations = k = 6

→ n is the sum of all the observations: 1. 3 4. 4 2. 2 5. 3 3. 5 6. 6

⇒ n = 3+2+5+4+3+6 = 23

→ the equation we will be using

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

total sum of squares = within sum of squares + between sum of squares

total SS = within SS + between SS

short form : $\sum_{i=1}^{k} n_i \bar{y}_i^2 - \frac{\left(\sum_{i=1}^{k} n_i \bar{y}_i\right)^2}{n}$

short form : $\sum_{i=1}^{k} (n_i - 1) S_i^2$

⚠ Important notes :-

* Between SS is large → indicates that the means of groups are different from each others
* within SS is small → indicates that the values in a group are close to their group mean

→ This indicates that the values in each group are close to each other and to their mean, but the means of the groups are different from each other and from the total mean which indicates that the means aren't the same so which indicates rejecting H_0

* Between SS is small → this indicates that the means of the different groups are close to each others and close to the overall mean
* within SS is large → this indicates that the values of one group are different from their group mean

→ This indicates that values in each group are various from the group mean but group means are close to each others which indicates that the means are the same ⇒ which indicates accepting H_0

* Between mean square = Between SS /(k-1) → between MS
* within mean square = within SS /(n-k) → within MS

to test the hypotheses we will be using the F-test of (k-1) and (n-k) degree of freedom :-

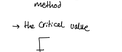
⇒ the test statistic = F_0 = Between MS / within MS

since we are using the sum of squares the F_0 will be ⊕ so we will be working on the right side only

Critical value method

→ the critical value $f_{(k-1),(n-k),1-\alpha}$ → $(k-1)$ degrees of freedom, $(n-k)$ degrees of freedom

* $f_0 >$ critical value ⇒ reject H_0
* $f_0 <$ critical value ⇒ accept H_0



P-value method

$P(F > f_0)$ which can't be found directly from the table



* P-value $> \alpha$ ⇒ accept H_0
* P-value $< \alpha$ ⇒ reject H_0

* in the F - distribution table → d_1 ⇒ k-1, d_2 ⇒ n-k

② Comparing the means of specific groups among all groups

$H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$

* this test shows weather 2 groups have the same mean or not but it does not tell which ones are the ones

if the variance of the population is known → we use the Z test

* the mean of group i is normally distributed as $\bar{y}_i \sim N(\mu + \alpha_i, \sigma^2/n_i)$

→ when we are comparing the means of 2 groups

$\bar{y}_1 \sim N(\mu + \alpha_1, \sigma^2/n_1)$

$\bar{y}_2 \sim N(\mu + \alpha_2, \sigma^2/n_2)$

we will make a normal distribution of the difference between them

$\bar{y}_1 - \bar{y}_2 \sim N(\alpha_1 - \alpha_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$

↓ the standardized normal distribution

$$Z = \frac{\bar{y}_1 - \bar{y}_2 - 0}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

under the condition the population variance is known

⇒ test statistic : $Z = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Critical value method

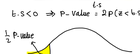
→ critical values $-z_{1-\alpha/2}$, $z_{1-\alpha/2}$



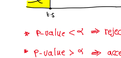
* critical $z_0 <$ -critical ⇒ accept H_0
* $z_0 >$ critical ⇒ reject H_0

P-value method

$z_0 > 0$ ⇒ P-value = $2P(z > z_0)$



$z_0 < 0$ ⇒ P-value = $2P(z < z_0)$



* P-value $< \alpha$ ⇒ reject H_0
* P-value $> \alpha$ ⇒ accept H_0

if the variance of the population isn't known ⇒ (the least difference method (LSD))

→ we use t - test and the variance is estimated using the Pooled variance

⇒ the test statistic = $\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

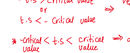
we estimate the variance of the population with the Pooled variance

→ in chapter 8 for 2 samples we used $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

→ in this chapter we use the rule related to (within MS) $S^2 = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{n-k}$ → within MS → within SS / n-k

critical value method

→ critical values $-t_{n-k,1-\alpha/2}$, $t_{n-k,1-\alpha/2}$



* $t_0 >$ critical value or $t_0 <$ -critical value ⇒ reject H_0
* -critical < $t_0 <$ critical value ⇒ accept H_0

P-value method

$t_0 > 0$ ⇒ P-value = $2P(t > t_0)$



$t_0 < 0$ ⇒ P-value = $2P(t < t_0)$



* P-value $> \alpha$ ⇒ accept H_0
* P-value $< \alpha$ ⇒ reject H_0

③ A $100\times(1-\alpha)$ CI for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm t_{n-k,1-\alpha/2}\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Pooled Variance

Biostatistics

→ ch 1 ⇒ introduction

→ ch 2 ⇒ • measures of location • measures of spread

→ ch 3 ⇒ Probabilities

→ ch 4 ⇒ Binomial distribution

→ ch 5 ⇒ Normal distribution

→ ch 6 ⇒ Finding the confidence interval → z - distribution; → t - distribution

→ ch 7 ⇒ hypothesis testing + confidence intervals using the $\frac{z}{t}$ tests to obtain

→ ch 8 ⇒ hypothesis testing for the mean of 2 samples → Critical value method, P-value method → the normal theory method

→ ch 10 ⇒ hypothesis testing for the fraction of 2 samples → normal theory method, ANOVA method (chi-square distribution)

→ ch 11 ⇒ hypothesis testing of the Correlation

→ ch 12 ⇒ ANOVA → Hypothesis testing of the mean related to multiple samples