Ch 8 deals with finding CI and performing hypothesis tests on a single population mean (1-sample) and comparing 2 population means.

Ch 10 is an extension and deals with comparing more than 2 population means. The technique is called Analysis of Variance.

AN O VA

Your coursenote introduces ANOVA (ANalysis Of VAriance) with an example (Ex 10.1 Page 153). We'll work on Ex. 10.1.

The name ANOVA may be a little misleading. We are really comparing population means, but the statistical procedure includes the word variance (this is because population means are compared by dividing the total variation into appropriate pieces).

ANOVA notation

k = the number of populations under investigation.

population	1	2	....	i	....	k
population mean	μ_1	μ_2	....	μ_i	....	μ_k
population variance	σ_1^2	σ_2^2	....	σ_i^2	....	σ_k^2
sample size	n_1	n_2	....	n_i	....	n_k
sample mean	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	....	$\bar{x}_{i\cdot}$	....	$\bar{x}_{k\cdot}$
sample variance	s_1^2	s_2^2	....	s_i^2	....	s_k^2

$n = n_1 + n_2 + ... + n_k$

$=$ total number of observations in the entire dataset

The null and alternative hypothesis are stated in terms of the population means (hence, use Greek symbols)

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \quad \text{(All k population means are equal)}$$

$$H_a : \mu_i \neq \mu_j \text{ some } i \neq j \quad \text{(At least 2 of the k population means differ)}$$

The assumptions for this test procedure are similar to those for a 2-sample t test.

① The k population distributions are normal

② The k population variances are equal (that is, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$)

③ The samples are selected randomly and independently from the respective populations.

To denote observations, $x_{ij} = j^{th}$ measurement taken from the i^{th} population.

i^{th} population j^{th} measurement

Sometimes, we might need a comma $x_{x,j}$ (eg $x_{1,23}$) if there is ambiguity.

<u>Note:</u>

① Mean of the observations in the i^{th} sample is

$$\bar{x}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n_i}(x_{i1} + x_{i2} + ... + x_{in_i})$$

② The mean of <u>all</u> the observations (we call this the grand mean) is

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

③ ANOVA identify

$$SST \quad = \quad SSA \quad + \quad SSE$$

(your notes call this SSt) (your notes call this SSe)

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\cdot\cdot})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}_{\text{between-samples variation}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2}_{\text{within-sample variation}}$$

④ MSA=mean square due to factor=$\frac{SSA}{k-1}$

the characteristic that differentiates the populations from one another, eg gasoline brand if you are studying the effect of 5 different brands of gasoline.

⑤ MSE=mean square error=$\frac{SSE}{n-k}$

⑥ ANOVA test procedure

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_a : \mu_i \neq \mu_j \text{ for some } i \neq j$$

Test statistic is

$$F_{obs} = \frac{MSA}{MSE}$$

this is a new distribution that you have not seen (show class the F table)

Rejection region:

If $F_{obs} \geq F_\alpha$, we reject H_0

numerator df=k-1
denominator df=n-k

⑦ $t_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}$=sum of the observations in the i^{th} sample

$t_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$=sum of all the observations

⑧ shortcut formulas:

faster to compute

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\cdot\cdot})^2 = (\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2) - \frac{t_{\cdot\cdot}^2}{n}$$

$$SSA = \sum_{i=1}^{k} n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\frac{t_{i\cdot}^2}{n_i} - \frac{t_{\cdot\cdot}^2}{n}$$

faster to compute

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2 = SST - SSA$$

Very Important

ANOVA summary table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Factor	k-1	SSA	MSA=$\frac{SSA}{k-1}$	$\frac{MSA}{MSE}$
Error	n-k	SSE	MSE=$\frac{SSE}{n-k}$	
Total	n-1	SST		

Let's work on Ex.10.1 (Pg 153)

Ex.10.1 (a) Are the model's assumptions consistent with the data

- see Pg 155's QQ plots (check normality)

- constant variance, check boxplot (Pg 155)

  (boxes, approximately equal size).

  Some books suggest checking that "largest sample standard deviation is less than twice the smallest sample standard deviation".

- We assume the samples are selected randomly and independently from the respective populations.

ANOVA model

What are the unbiased estimators for the unknown parameters in the model?

$\mu_i, i = 1, 2, ..., k$
and σ^2

① $\bar{x}_{i\cdot}$ is the unbiased estimator for μ_i

② $S_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+...+(n_k-1)s_k^2}{(n_1-1)+(n_2-1)+...+(n_k-1)}$ is an unbiased estimator of σ^2

Note: Notice the numerator

$$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2 = SSE$$

and denominator is really n-k.

It turns that $S_p^2 = MSE$ (this is very useful as you'll see in the next example)

Ex 10.1 (b)

$$\begin{array}{lll} \bar{x}_{1.} = 45.05 & s_1^2 = (7.5)^2 & n_1 = 20 \\ \bar{x}_{2.} = 52.29 & s_2^2 = (8.70)^2 & n_2 = 20 \\ \bar{x}_{3.} = 54.29 & s_3^2 = (6.32)^2 & n_3 = 20 \\ \bar{x}_{4.} = 56.83 & s_4^2 = (10.40)^2 & n_4 = 20 \\ \bar{x}_{5.} = 41.15 & s_5^2 = (8.16)^2 & n_5 = 20 \end{array}$$

$$S_p^2 = \frac{19(7.5)^2 + 19(8.70)^2 + ... + 19(8.16)^2}{100 - 5}$$

$$= 69.3256$$

$$(= MSE)$$

Ex.10.1 (c)

source of variation	df	SS	MS	F
Factor	E 4	D 3461.83	F 865.4575	G 12.48
Error	B 95	C 6585.932	A 69.3256	
Total				

A from S_p^2

B n-k=95

C $\frac{SSE}{95} = 69.3256 \Rightarrow SSE =$

D $SSA = \sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2$. Note that $\bar{x}_{..} = 49.922$.

Hence $SSA = 20(45.05 - 49.922)^2 + 20(52.29 - 49.922)^2 + ... + 20(41.15 - 49.922)^2$

$= 3461.83$

E k-1

F $\frac{SSA}{k-1} = 865.4575$

G $MSA/MSE = \frac{865.4575}{69.3256} = 12.48$

Ex10.1 (c)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_a : \mu_i \neq \mu_j \quad for\ some\ i \neq j, \quad i, j = 1, 2, 3, 4, 5$$

Under H_0, test statistic is

$$F_{obs} = 12.48$$

Look up F table

$$F_{4,95,\alpha=0.05} \approx F_{4,60,\alpha=0.05} = 2.53 \xrightarrow{\text{using}} (\alpha = 0.05)$$

numerator df use smaller df
denominator df

Since $F_{obs} = 12.48 > F_{4,95,0.05}$, we reject H_0 and conclude there are statistically significant differences among the drying methods.