3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 37

Last time Fisher LSD Method for Multiple Comparisons

<u>If</u> we reject $\boxed{H_0 : \mu_1 = \ldots = \mu_a}$ in favor of $\boxed{H_1 : \mu_i \neq \mu_k \text{ for at least one pair } i \neq k}$,

<u>then</u> carry out $\binom{a}{2}$ tests with $\boxed{H_0 : \mu_i = \mu_k}$, $\boxed{H_1 : \mu_i \neq \mu_k}$ using the rule:

Reject H_0 if $|\bar{Y}_{i\cdot} - \bar{Y}_{k\cdot}| > LSD = t_{\frac{\alpha}{2}, a(n-1)} \sqrt{\frac{2MS_E}{n}}$.

Assumptions with ANOVA

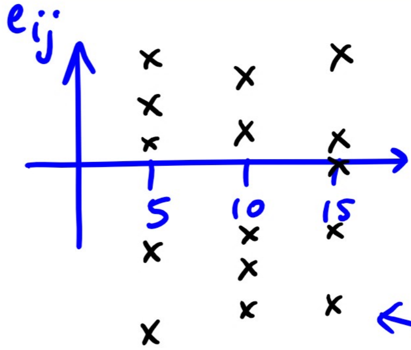
$$\varepsilon_{ij} = Y_{ij} - \bar{Y}_{i\cdot} \quad \leftarrow \text{within-group errors}$$

$$\sim N(0, \sigma^2)$$

To check <u>Normality</u> assumption: probability plot of $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$ $\leftarrow$ residuals

(Look for a straight line.)

To check that the ε_{ij} have <u>common variance</u> plot e_{ij} against e.g. treatments i.e. factor levels or treatment means



e.g. factor levels are amounts of some drug say 5mg, 10mg, 15mg

← 5 residuals given 5 observations per treatment level.

No pattern = correct assumption of common variance.

One thing that we assumed but DID NOT NEED TO ASSUME is that n_i = # observations at treatment level i

(to do the ANOVA test – needed only to have LSD)

= some common n.

Everything still works if n_i s maybe different (i.e. ANOVA works)

but formulas get more complicated:

What do we change? n vanishes, now $n_1, \dots, n_a$

$$N = \text{total \# of observations} = n_1 + n_2 + \dots + n_a$$

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}^2 - N\bar{Y}_{\cdot\cdot}^2$$

$$SS_{Treatments} = \sum_{i=1}^{a} \frac{Y_{i\cdot}^2}{n_i} - N\bar{Y}_{\cdot\cdot}^2$$

So $MS_E = SS_E / (N-a)$ ← Not $an - a = a(n-1)$

& we compare this to the critical value $f_{\alpha, a-1, N-a}$

hat happens to ~~LSD Method~~?

Multiple Comparisons?

ME (Margin of Error) = critical value for $\bar{Y}_{i\cdot} - \bar{Y}_{k\cdot}$

when testing $\left.\begin{array}{l} H_0 : \mu_i = \mu_k \\ H_1 : \mu_i \neq \mu_k \end{array}\right)$ is

$t_{\frac{\alpha}{2}, N-a} \sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}$.

<u>REVIEW</u> <u>Bias, (Estimated) Standard Error, Mean Square Error</u>

<u>Recall</u> Say $\hat{\Theta}$ is a function $h(X_1, \ldots, X_n)$ of random variables $X_1, \ldots, X_n$, which is an estimator for θ (some parameter of an underlying distribution)

<u>Bias of $\hat{\Theta}$</u> : $E(\hat{\Theta}) - \theta$ (if $E(\hat{\Theta}) = \theta$, $\hat{\Theta}$ unbiased)

<u>Standard Error of $\hat{\Theta}$</u> : $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$.

<u>Estimated Std. Error of $\hat{\Theta}$</u> : $se(\hat{\Theta})$ $\sigma_{\hat{\Theta}}$ with any parameters you need to estimate plugged in

Notice: $\sigma_{\hat{\Theta}} = se(\hat{\Theta})$ if no other parameters.

So to find these you typically have to calculate

$$E(\hat{\Theta}) \text{ and } V(\hat{\Theta}), \quad \text{for } \hat{\Theta} = h(X_1, \ldots, X_n).$$

<u>Example</u> <u>Question</u> 2 machines with same underlying mean ($\mu_1 = \mu_2$), but with different variances: $\sigma_2^2 = \frac{1}{4}\sigma_1^2$.

If n_i observations are taken from machine i find the bias and standard error of $\hat{\mu} = \frac{1}{4}\bar{X}_1 + \frac{3}{4}\bar{X}_2$, estimator of μ $(= \mu_1 = \mu_2)$.

<u>Solution</u> We need to find bias of $\hat{\mu} = E(\hat{\mu}) - \mu$ and $\sigma_{\hat{\mu}} = \sqrt{V(\hat{\mu})}$.

So we need to find $E(\hat{\mu})$ and $V(\hat{\mu})$, where $\hat{\mu} = \frac{1}{4}\bar{X}_1 + \frac{3}{4}\bar{X}_2$.

$$\Rightarrow \bar{\mu} = \frac{1}{4}\left(\frac{X_{11} + X_{12} + \ldots + X_{1n_1}}{n_1}\right) + \frac{3}{4}\left(\frac{X_{21} + \ldots + X_{2n_2}}{n_2}\right)$$

where X_{ij} is the jth observation from machine i

$$= \frac{1}{4n_1}(X_{11} + \ldots + X_{1n_1}) + \frac{3}{4n_2}(X_{21} + \ldots + X_{2n_2}).$$

$X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$

So really the question here is about $E(h(\underline{\mathbf{X}}))$ & $V(h(\underline{\mathbf{X}}))$

$X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$

where $h(\underline{\mathbf{X}})$ is a linear combination of random variables.

Recall: If $h(\overset{X_1,\dots,X_k}{\cancel{\underline{X}}}) = c_1 X_1 + c_2 X_2 + \dots + c_k X_k$

then $E(h(\overset{X_1,\dots,X_k}{\cancel{\underline{X}}})) = c_1 E(X_1) + c_2 E(X_2) + \dots + c_k E(X_k)$

$V(h(\overset{X_1,\dots,X_k}{\cancel{\underline{X}}})) = c_1^2 V(X_1) + c_2^2 V(X_2) + \dots + c_k^2 V(X_k)$.

2 most useful special cases are: TOTALS & MEANS
where X_i all have same mean μ
and same variance σ^2.

TOTALS $X_1 + X_2 + \dots + X_k$

$$E(X_1 + \dots + X_k) = E(X_1) + \dots + E(X_k)$$
$$= k\mu.$$

$$V(X_1 + \dots + X_k) = V(X_1) + \dots + V(X_k)$$
$$= k\sigma^2.$$

MEANS $\bar{X} = \frac{1}{k}(\underbrace{X_1 + \dots + X_k}) \rightarrow$ a single r.v.

$$E(\bar{X}) = \frac{1}{k} E(X_1 + \dots + X_k) = \frac{1}{k} \cdot k\mu = \mu$$

$$V(\bar{X}) = \frac{1}{k^2} V(X_1 + \dots + X_k) = \frac{1}{k^2} \cdot k\sigma^2 = \frac{\sigma^2}{k}.$$

Back to the question: $h(\overset{X_{11},\dots,X_{1n_1},X_{21},\dots,X_{2n_2}}{\cancel{\bar{X}_1, \bar{X}_2}}) = \frac{1}{4}\bar{X}_1 + \frac{3}{4}\bar{X}_2 = \hat{\mu}$.

not wrong, but lazy!

$$E(h(\overset{X_{11},\dots,X_{1n_1},X_{21},\dots,X_{2n_2}}{\cancel{\underline{X}}})) = \frac{1}{4}E(\bar{X}_1) + \frac{3}{4}E(\bar{X}_2) = \frac{1}{4}\mu + \frac{3}{4}\mu = \mu$$

So bias $\hat{\mu} = E(\hat{\mu}) - \mu = \mu - \mu = 0$.

$$V(\hat{\mu}) = \frac{1}{16}V(\bar{X}_1) + \frac{9}{16}V(\bar{X}_2)$$

$$= \frac{1}{16}\frac{\sigma_1^2}{n_1} + \frac{9}{16}\frac{\sigma_2^2}{n_2} = \frac{1}{16}\frac{\sigma_1^2}{n_1} + \frac{9}{16}\left(\frac{1}{4}\sigma_1^2\right)/n_2$$

$$= \frac{\sigma_1^2}{16}\left(\frac{1}{n_1} + \frac{9}{4n_2}\right) \qquad \text{So } \sigma_{\hat{\mu}} = \sqrt{V(\hat{\mu})}$$

$$= \frac{\sigma_1}{4}\sqrt{\frac{1}{n_1} + \frac{9}{4n_2}}.$$

But if we hadn't remembered those shortcuts using sample mean expectation & variance, just work through it: $E(\hat{\mu}) = E\left(\frac{1}{4n_1}(X_{11} + \ldots + X_{1n_1}) + \frac{3}{4n_2}(X_{21} + \ldots + X_{2n_2})\right)$

$$= \frac{1}{4n_1}E(\overbrace{X_{11} + \ldots + X_{1n_1}}^{\text{this is a r.v.}}) + \frac{3}{4n_2}E(\overbrace{X_{21} + \ldots + X_{2n_2}}^{\text{also a r.v.}})$$

$$= \frac{1}{4n_1}\left(E(X_{11}) + \ldots + E(X_{1n_1})\right) + \frac{3}{4n_2}\left(E(X_{21}) + \ldots + E(X_{2n_2})\right)$$

$$= \frac{1}{4n_1}(\underbrace{\mu_1 + \ldots + \mu_1}_{n_1 \text{ times}}) + \frac{3}{4n_2}(\underbrace{\mu_2 + \ldots + \mu_2}_{n_2 \text{ times}}) \qquad \text{— but } \mu = \mu_1 = \mu_2!$$

$$= \frac{1}{4\cancel{n_1}}\cdot \cancel{n_1}\cdot\mu + \frac{3}{4\cancel{n_2}}\cdot\cancel{n_2}\cdot\mu = \frac{1}{4}\mu + \frac{3}{4}\mu = \mu. \text{ So Bias} = 0.$$

While $V(\hat{\mu}) = V\left(\frac{1}{4n_1}(X_{11} + \ldots + X_{1n_1}) + \frac{3}{4n_2}(X_{21} + \ldots + X_{2n_2})\right)$

$$= \frac{1}{16n_1^2}V(X_{11} + \ldots + X_{1n_1}) + \frac{9}{16n_2^2}V(X_{21} + \ldots + X_{2n_2})$$

$$= \frac{1}{16n_1^2}\left(V(X_{11}) + \ldots + V(X_{1n_1})\right) + \frac{9}{16n_2^2}\left(V(X_{21}) + \ldots + V(X_{2n_2})\right)$$

$$= \frac{1}{16n_1^2}(\cancel{n_1}\cdot\sigma_1^2) + \frac{9}{16n_2^2}(\cancel{n_2}\cdot\sigma_2^2) = \frac{\sigma_1^2}{16n_1} + \frac{9}{16n_2}\left(\frac{1}{4}\sigma_1^2\right) = \frac{\sigma_1^2}{16}\left(\frac{1}{n_1} + \frac{9}{4n_2}\right)$$

So standard error $= \sigma_{\hat{\mu}} = \sqrt{V(\hat{\mu})} = \frac{\sigma_1}{4}\sqrt{\frac{1}{n_1} + \frac{9}{4n_2}}$.