Review of the Topics for the Final (continued)

IV. Comparing the Means of Many Independent Samples

In many applications we are interested in comparing the means of several populations. The statistical method for this analysis is called \textit{analysis of variance} or \textit{ANOVA}. In analysis of variance we compute \textit{between} the groups and \textit{within} the groups variations.

The simple one-way \textit{analysis of variance model} is:

\[ y_{ij} = \mu + \tau_i + \epsilon_{ij} \]

where \( \mu \) is the grand mean, \( \tau_i \) is the main effect of the \( i \)-th treatment and \( \epsilon_{ij} \) is the error. The \textit{global null hypothesis} in analysis of variance is: \( H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \) and the alternative hypothesis is \( H_A : \) Not all \( \mu_i \) ‘s are equal. To test this hypothesis we use the \textit{F test}. The F distribution depends on two parameters \textit{numerator degrees of freedom} and \textit{denominator degrees of freedom}. The F distribution is given in Table 10.

The \textit{F Statistics} is calculated as

\[ F_s = \frac{\text{MS(between)}}{\text{MS(within)}} \]
To compare the means simultaneously we use *multiple comparison procedures*. The Bonferroni and Tukey methods are among the popular multiple comparison procedures.

**Example:** Problem #3 of the sample final.

---

**V. Linear Regression and Correlation**

Simple *linear regression* is one of the most widely used statistical techniques for developing a mathematical relationship between a *dependent* variable and an *independent* variable.

The *simple linear regression model* is:

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

The Least Squares Estimation:

Let \( \hat{y} = b_0 + b_1 x \) denote the estimated regression line. Find \( b_0 \) and \( b_1 \) such that \( SS(\text{resid}) = \sum (y - \hat{y})^2 \) is minimized.

The *residuals* are defined as observed minus estimated values.

**Example:** Problem #1 of the sample final.
Assessing the Model

A 100(1-\(\alpha\))% confidence interval for \(\beta_1\) is:

Example: Problem #1 of the sample final.

To test \(H_0: \beta_1 = 0\), compute \(t = \ldots\) and use the chart.

Example: Problem #1 of the sample final.

Given a specific value of the independent variable \(x\), say \(x_g\), a 100(1-\(\alpha\))% prediction interval for \(y\) is:

\[
y_{\hat{}} \pm t_{\alpha/2} S_e \sqrt{1 + 1/n + (x_g - \bar{x})^2 / \Sigma(x - \bar{x})^2}
\]

a 100(1-\(\alpha\))% confidence interval for the expected value of \(y\) is:

\[
y_{\hat{}} \pm t_{\alpha/2} S_e \sqrt{1/n + (x_g - \bar{x})^2 / \Sigma(x_g - \bar{x})^2}
\]

where \(y_{\hat{}} = y_{\hat{}} = b_0 + b_1 x\).

Example: Problem #1 of the sample final.
Coefficient of Correlation

The *Pearson coefficient of correlation* is a measure of the strength of linear relationship between the *response* variable $y$ and *explanatory* variable $x$. The *sample correlation coefficient* is defined as:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

**Example:** Problem #1.

To test the null hypothesis $H_0: \rho = 0$, we use the test statistic

$$t = r \sqrt{(n-2)/(1-r^2)}$$

which is distributed as $t$ with $(n-2)$ degrees of freedom.

**Example:** Problem #1 of the sample final.

The *coefficient of determination* is defined as:

$$r^2 = 1 - \text{SS(resid) / SS(total)}$$

**Example:** Problem #1 of the sample final.
Another example of analysis of variance table:

The regression equation is
Weight = 98.8 + 5.91 Shoe Size

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>StDev</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>98.83</td>
<td>45.15</td>
<td>2.19</td>
<td>0.065</td>
</tr>
<tr>
<td>Shoe Siz</td>
<td>5.913</td>
<td>4.708</td>
<td>1.26</td>
<td>0.249</td>
</tr>
</tbody>
</table>

S = 11.26       R-Sq = 18.4%   R-Sq(adj) = 6.7%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>200.0</td>
<td>200.0</td>
<td>1.58</td>
<td>0.249</td>
</tr>
<tr>
<td>Residual Error</td>
<td>7</td>
<td>888.0</td>
<td>126.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>1088.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A Practice Problem:

In an insurance study, the following data were collected on the number of times eight shipments were transferred from one aircraft to another aircraft and the number of broken items in each shipment.

<table>
<thead>
<tr>
<th>Number of Transfers (x)</th>
<th>Number of Broken Items (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

We would like to fit a linear regression model to this data and use it to predict the number of broken items in shipments.

(For partial solution of this problem see the solution for the sample final)