Chapter 6: Linear combinations and multiple comparisons

The Centers for Disease Control estimates that the prevalence of HIV infection among adults and adolescents to be 447.8 per 100,000 (.448%) in the United States at the end of 2006. Approximately 75% of this population were male and 48.1% of the population were men who have sex with men. In contrast, more women are infected with HIV than men (approximately 56.5%) in sub-Saharan Africa. The differences in infection levels between women and men are most pronounced among young people (15-24 years). In this age group, 78.2% are women.

It’s widely believed that the HIV/AIDS pandemic is not homogeneous among continental-scale regions. Even at the country level, there are wide variations in infection levels between different areas. The prevalence of HIV/AIDS continues to increase in most parts of the world, and it’s often claimed that Sub-Saharan Africa is by far the worst-affected region. The best data, arguably, for global analyses are available from the World Health Organization.

Is there convincing statistical evidence that the prevalence in Africa is different from the Americas and South America? Is there convincing statistical evidence that the prevalence in Europe is different from the Americas? Is there convincing statistical evidence that the prevalence in Asia and Europe is different from the Americas and South America?

Prevalence of HIV among adults aged 15 to 49 as of 2008 are shown above for 44 countries with the greatest estimated prevalence. A dotplot (below and left) shows all countries\(^1\) \((n = 138)\) grouped by region. Apparently, African prevalence is greater than other regions and within region variability is not homogeneous.\(^2\) It’s difficult to see much detail, so a transformation to the (natural) log scale was applied to the prevalence data (below and right). The transformation appears to have substantially reduced differences among within-region sample variances and revealed additional detail.\(^3\) Mathematically, the transformation

\[^1\]A few countries reporting 0% prevalence have been omitted.
\[^2\]Lack of homogeneity is a serious problem if region means are to be compared using the analysis of variance methods of Chapter 5.
\[^3\]Since some values are 0, I’ve added a constant to every datum. The constant (.05) is half of the smallest positive value in the data set.
is $y_{ij} = \log(x_{ij} + .05)$ where $x_{ij}$ is the prevalence for the $j$th country in the $i$th region.

Before proceeding, it’s useful to assess the evidence supporting the contention that there are at least two regions that differ with respect to regional means. Let $\mu_i$ denote the mean log-prevalence in region $i$. The regional sample mean $\bar{y}_i$ computed from the country values (within region) provides the estimate of $\mu_i$.\(^4\) However, I’ll ignore this flaw to motivate and illustrate the methods of the chapter. The hypotheses of interest are

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_5$$

$$H_a: \text{at least two } \mu_i \text{ are not equal.}$$

The analysis of variance table (below) confirms the obvious: there is convincing evidence of differences between means.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sums of squares</th>
<th>d.f.</th>
<th>Mean square</th>
<th>$F$-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between regions</td>
<td>129.23</td>
<td>4</td>
<td>32.30</td>
<td>21.9</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Within regions</td>
<td>196.08</td>
<td>133</td>
<td>1.474</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>325.31</td>
<td>137</td>
<td>2.37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4For most of the regions, the sample of countries includes nearly all countries within a region. Consequently, the 10% condition stating that the sample size $n_i$ is no larger than 10% of the population size ($N_i$) obviously is not met. Not meeting the 10% condition would be a fatal flaw if this were real analysis instead of a pedagogical example. A realistic model of the data recognizes that the sampled units are a near-complete census of the population of countries and that the source of uncertainty associated with a regional mean $\bar{y}_i$ is measurement error—not sampling variability. For most statistical problems, measurement error is a small, ignored component of sampling variation.
Turning to the questions of primary interest:

1. Is there convincing statistical evidence that the prevalence in Africa is different from the Americas and South America?

2. Is there convincing statistical evidence that the prevalence in Europe is different from the Americas?

3. Is there convincing statistical evidence that the prevalence in Asia and Europe is different from the Americas and South America?

The questions I have posed constitute a *multiple comparisons* problem. The third question, for example, is asking for a comparison of mean prevalence in Asia and Europe versus Americas and South America. The number of comparisons are three (four, counting the $F$-test above), and this poses some challenges in hypothesis testing.

A general framework for these types of problems starts with the idea of a *linear combination*. Before proceeding, let $\mu_1$, $\mu_2$, $\mu_3$, $\mu_4$ and $\mu_5$ denote the means for Africa, the Americas, Asia, Europe and South America, respectively.

A *linear combination* of means $\mu_1, \mu_2, \ldots, \mu_I$ is a sum

$$\gamma = C_1 \mu_1 + C_2 \mu_2 + \cdots + C_I \mu_I$$

where $C_1, C_2, \cdots, C_I$ is a set of known constants. If

$$0 = C_1 + C_2 + \cdots + C_I,$$

then the linear combination is called a *contrast*. Usually, multiple comparison problems involve several contrasts.

For example, the three questions posed above will be solved by estimating the following linear combinations:

$$\gamma_1 = \mu_1 - \frac{\mu_2 + \mu_5}{2}$$
$$\gamma_2 = \mu_2 - \mu_4$$
$$\gamma_3 = \mu_3 + \mu_4 - (\mu_2 + \mu_5).$$

If $\gamma_1 = 0$, then $\mu_1$ is equal to the average of $\mu_2$ and $\mu_5$; so, if $\gamma_1 = 0$, then the prevalence in Africa is equal to the average prevalence in the Americas and South America. The other contrasts are interpreted similarly. The coefficients for the contrasts are tabled below:
Table 2: Contrast coefficients.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
<th>$C_5$</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1$</td>
<td>1</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
<td>-0.5</td>
<td>0</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>

The hypotheses to be tested are

\[ H_0: \gamma_1 = 0 \text{ vs } H_a: \gamma_1 \neq 0, \text{ equivalently, } H_0: \mu_1 = \frac{\mu_2 + \mu_5}{2} \text{ vs } H_a: \mu_1 \neq \frac{\mu_2 + \mu_5}{2}, \]

\[ H_0: \gamma_2 = 0 \text{ vs } H_a: \gamma_2 \neq 0, \text{ equivalently, } H_0: \mu_2 = \mu_5 \text{ vs } H_a: \mu_2 \neq \mu_5, \]

\[ H_0: \gamma_3 = 0 \text{ vs } H_a: \gamma_3 \neq 0, \text{ equivalently, } H_0: \mu_3 + \mu_4 = \mu_2 + \mu_5 \text{ vs } H_a: \mu_3 + \mu_4 \neq \mu_2 + \mu_5. \]

With this structure, it’s straightforward to develop estimators and test statistics. The estimator of a linear combination

\[ \gamma = C_1\mu_1 + \cdots + C_I\mu_I \]

is

\[ \hat{\gamma} = C_1\bar{y}_1 + \cdots + C_I\bar{y}_I, \]

and the standard error is

\[ \hat{\sigma}(\hat{\gamma}) = s_p\sqrt{\frac{C_1^2}{n_1} + \cdots + \frac{C_I^2}{n_I}}, \]

where \( s_p \) is the pooled standard deviation\(^5\)

\[ s_p^2 = \frac{SS_{\text{full}}}{n - I} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - I} = 1.474. \]

The regional means and 95% confidence intervals appear in Table 3. An R function call to produce the regional means is \texttt{lm(y~Region-1)$coeff} where \( y \) contains the log-prevalence values and \texttt{Region} contains the region labels. The confidence intervals were obtained from the function call \texttt{confint(aov(y~Region-1))}.

Table 3: Regional means and 95% confidence intervals.

<table>
<thead>
<tr>
<th>Region</th>
<th>Lower bound</th>
<th>Mean</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>.239</td>
<td>.576</td>
<td>.9118</td>
</tr>
<tr>
<td>Americas</td>
<td>-.783</td>
<td>-.201</td>
<td>.3816</td>
</tr>
<tr>
<td>Asia</td>
<td>-.2089</td>
<td>-1.609</td>
<td>-1.128</td>
</tr>
<tr>
<td>Europe</td>
<td>-1.969</td>
<td>-1.557</td>
<td>-1.145</td>
</tr>
<tr>
<td>South America</td>
<td>-1.365</td>
<td>-.641</td>
<td>.0835</td>
</tr>
</tbody>
</table>

\(^5s_p^2\) appears as \texttt{MS}_{\text{within}} in Table 1.
For $\gamma_1$,
\[
\hat{\gamma}_1 = \frac{y_1 - y_2 + y_5}{2} = \frac{.575 - (.200 - 1.56)}{2} = 1.455.
\]
and
\[
\hat{\sigma}(\gamma_2) = s_p \sqrt{\frac{C_1^2}{n_1} + \cdots + \frac{C_I^2}{n_I}} = 1.214 \sqrt{\frac{1}{51} + \frac{1/4}{17} + \frac{1/4}{11}} = 1.214 \times .204 = .2478.
\]

**The t-statistic**

Testing a hypothesis of the form $H_0 : \gamma_1 = 0$ vs $H_a : \gamma_1 \neq 0$ is straightforward because
\[
T = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}(\hat{\gamma})}
\]
has a $t_{d.f}$ distribution if $H_0$ is true. Here, d.f = $n - I$ is the degrees of freedom associated with $s_p^2$. In summary, if $H_0 : \gamma = 0$ is true, then
\[
T = \frac{\hat{\gamma}}{\hat{\sigma}(\hat{\gamma})} \sim t_{d.f}.
\]
If $H_a : \gamma \neq 0$, then unusually large or unusually small realizations of the t-statistic (1) constitute evidence against $H_0$ and in favor of $H_a$.

Similarly, a $100(1 - \alpha)$% confidence interval for $\gamma$ is
\[
\hat{\gamma} \pm t_{n-I}^* \hat{\sigma}(\hat{\gamma})
\]
where $t_{n-I}^*$ is the $\alpha/2$ quantile of the $t_{n-I}$ distribution.

Table 4 shows the results of the hypothesis tests. There is convincing evidence that the contrasts are not equal to zero for contrasts $\gamma_1$ and $\gamma_3$, and strong evidence for the contrast $\gamma_2$. In conclusion, all three comparisons reveal strong evidence of regional differences.

**Table 4: Details of the hypothesis tests involving the three contrasts.** The 95% confidence intervals were constructed using $t_{n-I}^* = t_{133}^* = -1.978$.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$\hat{\gamma}$</th>
<th>$\hat{\sigma}(\hat{\gamma})$</th>
<th>$t$</th>
<th>p-value</th>
<th>Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0 : \gamma_1 = 0$ vs $H_a : \gamma_1 \neq 0$</td>
<td>1.454</td>
<td>.248</td>
<td>8.65</td>
<td>&lt; .0001</td>
<td>[.962, 1.947]</td>
</tr>
<tr>
<td>$H_0 : \gamma_2 = 0$ vs $H_a : \gamma_2 \neq 0$</td>
<td>.4397</td>
<td>.1818</td>
<td>2.42</td>
<td>.0169</td>
<td>[.0801, .7993]</td>
</tr>
<tr>
<td>$H_0 : \gamma_3 = 0$ vs $H_a : \gamma_3 \neq 0$</td>
<td>-2.324</td>
<td>.2660</td>
<td>-8.73</td>
<td>&lt; .0001</td>
<td>[-2.850, -1.797]</td>
</tr>
</tbody>
</table>
Specific linear combinations

Often linear contrasts are used to compare two sets of means, the first of which is denoted by \( \{\mu_{1,1}, \ldots, \mu_{1,I}\} \), and the second of which is \( \{\mu_{2,1}, \ldots, \mu_{2,J}\} \) where \( J \) and \( I \) are not necessarily equal. The contrast compares the mean of the first set of means to the mean of the second set, and so the contrast is

\[
\gamma = \frac{\mu_{1,1} + \cdots + \mu_{1,I}}{I} - \frac{\mu_{2,1} + \cdots + \mu_{2,J}}{J}
\]

The coefficients associated with a particular mean will be either \( 1/I \) or \( -1/J \) or 0. A coefficient of zero is associated with a mean that is not in either set. No mean is in both sets. The coefficients sum to zero, so the comparison is a contrast of means.

Hicks\(^6\) reports data from an experiment on methods for firing naval guns. Gunners of three different physiques (slight, average, and heavy) tested two firing methods. Both methods were tested twice by each of nine teams of three gunners with supposedly identical physiques. The response was the number of rounds fired per minute. The data are available from the \texttt{nlme} library; the data are named \texttt{Gun}. The Figure to the right shows that there are very large differences associated with method and small differences associated with physique.

A formal test of the hypothesis that number of rounds fired varies with method is obtained by setting up the linear contrast coefficients in Table 5. Notice that the sum of the coefficients is 0.

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_i )</td>
<td>( \frac{1}{3} )</td>
<td>( \frac{1}{3} )</td>
<td>( \frac{1}{3} )</td>
<td>( -\frac{1}{3} )</td>
<td>( -\frac{1}{3} )</td>
<td>( -\frac{1}{3} )</td>
</tr>
<tr>
<td>( n_i )</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Let $\gamma = \sum_i C_i \mu_i$. Let

$$a = \frac{1}{3} \sum_{i=1}^{3} \mu_i \quad \text{and} \quad b = \frac{1}{3} \sum_{i=4}^{6} \mu_i$$

If method 1 produces more rounds, then $b < a$ and $0 < a - b$. Since $\gamma = a - b$, $0 < \gamma$ if method 1 produces more rounds.

A test of $H_0 : \gamma = 0$ vs $H_a : \gamma < 0$ provides a measure of evidence supporting the contention that method 1 produces more per minute than method 2. Based on the discussion above, $H_a$ is supported by large values for $\hat{\gamma}$ and the p-value will be a right tail area.

The denominator of the $t$ uses the mean within sums-of-squares ($s_p^2$), and the associated degrees of freedom are $\text{d.f.} = n - 15 = 36 - 6 = 30$. The Table below provides the formal test of significance.

Table 6: Details of the hypothesis test of age. The degrees of freedom are $\text{d.f.} = n - 15 = 473$.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$\hat{\gamma}$</th>
<th>$\hat{\sigma}(\hat{\gamma})$</th>
<th>$t$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0 : \gamma = 0$ vs $H_a : \gamma &gt; 0$</td>
<td>34.5</td>
<td>2.48</td>
<td>13.9</td>
<td>.0001</td>
</tr>
</tbody>
</table>

It is concluded that there is convincing evidence that method 1 produces more rounds per minute than method 2.

**Simultaneous inferences**

A logical (though faulty) strategy when analyzing data via the one-way analysis of variance is to test for differences among means using the data (sample means) to direct the choice of tests.

Suppose that the usual $F$-statistic supports the alternative hypothesis stating that there are at least two means that are different. Then the analyst begins looking for pairs of means that are significantly different using either the two-sample $t$-test or computing confidence intervals for the differences between pairs of population means. This approach is often described as *unplanned comparisons*. The examples above are planned comparisons since I thought about which comparison of means were most interesting (to me) before conducting the ANOVA.

The justification for pursuing unplanned comparisons is that the $F$-test has yielded evidence that at least two means are different, and so its reasonable to identify those pairs of means that are different. The problem with the strategy is that often the probability of concluding one or more pairs of means are different when they are truly equal is unacceptably
large.

If the number of means is large (e.g., \( I = 12 \)), then the number of pairs is much larger (the number of pairs is \( I \times (I-1)/2 = 66 \)). If most means are not different, and all pair-wise comparisons are made, then there is a high probability that at least one pair of means that are not truly different will be identified as significantly different. There are two types of error rates that should be identified:

1. Individual or test-wise error rate. This rate is the Type I error rate of a single test; i.e., the probability of incorrectly rejecting \( H_0 \) when an accept/reject decision is made.

2. Family- or experiment-wise error rate. This is the probability that at least one Type I error will be committed if more than one accept/reject decision is made.

A simple, artificial example supposes that 10 tests are conducted using the \( \alpha = .05 \) significance level, all 10 null hypotheses are correct, and the test statistics are statistically independent. Let \( A_i \) denote the probability that the \( i \)th test is correctly decided to be not significant. Then \( P(A_i) = 1 - \alpha = .95 \). The probability that all are correctly decided to be nonsignificant is \( P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_{10}) = .95^{10} = .5987 \) and the probability that at least one is incorrectly decided to be significant is \( 1 - P(A_1 \text{ and } A_2 \text{ and } \cdots \text{ and } A_{10}) = .4012 \). The experiment-wise error rate .401.\(^7\) In the case of \( I = 12 \) and 66 unplanned comparisons, the experiment-wise error rate (computed in this fashion) is .966.

Competent statisticians will advise the researcher contemplating conducting a large number of unplanned comparisons not to proceed. However, there are several methods that have been developed that control the experiment-wise error rate, though the actual experiment-wise error rate can only be computed if a number of unrealistic assumptions are made. The brief discussion is informative, and should not be viewed as condoning unplanned comparisons.

The discussion focuses on confidence intervals instead of hypothesis tests as the mathematics are somewhat simpler, and a 100(1 - \( \alpha \))% confidence interval for \( \mu_1 - \mu_2 \) immediately provides an \( \alpha \)-significance level test of the hypotheses \( H_0 : \mu_1 - \mu_2 = 0 \) versus \( H_a : \mu_1 - \mu_2 \neq 0 \).\(^8\)

The confidence interval methods modify the conventional confidence interval

\[
\bar{y}_1 - \bar{y}_2 \pm t^* \hat{\sigma}(\bar{y}_1 - \bar{y}_2)
\]

for \( \mu_1 - \mu_2 \) by replacing \( t^* \), the critical value from a \( t \)-distribution with degrees of freedom associated with the estimator of \( \sigma \), with a larger multiplier. The multiplier is chosen so

\(^7\)The example is artificial because tests are rarely independent, and some of the tests involve incorrect null hypotheses.

\(^8\)The decision rule is: reject \( H_0 \) if 0 is not in the interval.
that the family-wise confidence level is close to the selected individual confidence level. Two procedures that are useful for unplanned comparisons are

1. The \textit{Tukey-Kramer} procedure. Suppose that all population means $\mu_i = 0$, and let $\overline{y}_{\text{max}}$ and $\overline{y}_{\text{min}}$ denote the largest and smallest of the $I$ sample means. The interval

$$\overline{y}_{\text{max}} - \overline{y}_{\text{min}} \pm t^* \hat{\sigma}(\overline{y}_{\text{max}} - \overline{y}_{\text{min}})$$

is least likely to capture the true difference (0) between population means. The Tukey-Kramer multiplier is chosen so that the $I(I-1)/2$ confidence intervals\(^9\) will have at least a $1 - \alpha$ probability of capturing 0 in each interval. Since every other confidence interval will have a smaller difference between sample means than $\overline{y}_{\text{max}} - \overline{y}_{\text{min}}$, the procedure will succeed in controlling the family-wise error rate. Details are given on page 161 of \textit{The Statistical Sleuth}.

2. The \textit{Bonferroni} procedure is somewhat simpler to employ and usually more conservative than the Tukey-Kramer procedure as the family-wise confidence level is at least $100(1 - \alpha)$. The actual (or nominal) family-wise confidence level is unknown in practical situations. The underlying principle is to modify the individual-wise confidence level by setting it to $100(1 - \alpha/k)$ where $k$ is the number of intervals. The $t$ multiplier is the $1 - \alpha/(2k)$ percentile. For the T. Rex example, if all possible $k = I(I-1)/2 = 12 \times 11/2 = 66$ possible pair-wise comparisons, then the $1 - \alpha/(2k) = .9996212$ percentile of the $t_{40}$ distribution is 3.647.

The Bonferroni method is preferable if a large number of comparisons are conducted (e.g., $k > 10$); otherwise Tukey-Kramer is preferred.

\(^9\)In total, there are $I(I-1)/2$ possible comparisons of sample means.
Data snooping and fishing expeditions

Sometimes, many variables are measured on every sample or experimental unit simply because it’s possible. For example, a study of human performance under stress may collect blood samples at different points in a physiological stress test. Then, as many as 100 blood chemistry variables are measured on each sample.\(^{10}\) The aim is to identify which variable(s) respond substantively to physiological stress. Simplistically, the researchers may carry out one paired $t$-test for each variable and attempt to adjust the confidence intervals to control the family-wise confidence levels. The methods described above are not appropriate. In the many variables situation, some variables are at least moderately correlated whereas the multiple comparisons methods are based upon an assumption of independent tests. If the independence condition is not met then the methods fail to achieve the advertised family-wise confidence level.

A potentially disastrous procedure is to select the $t$-test with the smallest p-value, or confidence interval most distant from 0, focus on that comparison and present the p-value or confidence interval. To see why this procedure is misleading, suppose that 50 true null hypotheses are tested. The expected value of the smallest p-value among the 50 is approximately .02 and the expected smallest p-value among 100 tests is about .01. In both cases, it is likely that at least one test will be found to be significant and the alternative concluded to be correct. However, all alternatives are incorrect.

\(^{10}\)The lab may insist in measuring 5, 30 or 100.