Comparing Several Means

Some slides from R. Pruim  STA303/STA1002: Methods of Data Analysis II, Summer 2016
Michael Guerzhoy
The Dating World of Swordtail Fish

• In some species of swordtail fish, males develop brightly coloured swordtails

• Southern Platyfish do not

• Want to know: will female Southern Platyfish prefer males with artificial brightly-coloured swordtails?
  • If they do, that’s evidence that males in other species evolved as a result of female preference

• Experiment: multiple pair of males, one with a transparent artificial tail, one with a bright yellow artificial swordtail. Measure the percentage of time the female spends courting with the male with the yellow tail. There are 84 females in total.
Platyfish

• Eventually, we would like to know whether females spent more time with the yellow-swordtailed males. But we would like to first investigate whether there is anything else going on in the data that might affect our conclusions

• Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

• (the fish, in R)
Computing Group Means with Linear Regression

• Fit a linear regression:
  • \(Y \sim a_0 + a_{g1}I_{g1} + a_{g2}I_{g2} + \cdots + a_{gN}I_{gN}\)

  - \(Y\): the percentage of time the female spends with the yellow-tailed male
  - \(I_{gk}\): 1 if the case involves Group k, 0 otherwise

• Regression:
  • Minimize \(\sum_i (Y_i - (a_0 + a_{g1}I_{i,g1} + a_{g2}I_{i,g2} + \cdots + a_{gN}I_{i,gN}))^2\)
Computing Group Means with Linear Regression

- \( \Sigma_i (Y_i - (a_0 + a_{g1}I_{i,g1} + a_{g2}I_{i,g2} + \cdots + a_{gN}I_{i,gN}))^2 \)
  \[ = \sum_{\text{group}} \sum_{i \in \text{group}} (Y_i - a_{\text{group}})^2 \]

- \( \Sigma_{i \in \text{group}}(Y_i - a_{\text{group}})^2 \) is minimized when \( a_{\text{group}} = ? \) (show how to do this)
Computing Group Means with Linear Regression

\[ \left( \sum_{i \in \text{group}} (Y_i - a_{\text{group}})^2 \right)' = 0 \]

\[-2 \sum_{i \in \text{group}} (Y_i - a_{\text{group}}) = 0 \]

\[ \sum_{i \in \text{group}} Y_i = \sum_{i \in \text{group}} a_{\text{group}} \]

\[ a_{\text{group}} = \frac{\sum_{i \in \text{group}} Y_i}{N_{\text{group}}} \]
Computing the Means with R

• (in R)
Are the Pairs Different from Each Other?

- If we had just two pairs in which we’re interested, we could simply use a t-Test
  - Estimate the pooled variance from the entire dataset
    - Estimate the pooled variance from the entire dataset
      \[ s_p^2 = \frac{(N_{g1}-1)s_1^2 + \cdots + (N_{gN}-1)s_N^2}{(N_{g1}-1) + \cdots + (N_{gN}-1)} \]
      \( (N_{g1} - 1) + \cdots + (N_{gN} - 1) \) d.f.
    - \( \frac{\text{mean}_{g1} - \text{mean}_{g2}}{s_p \sqrt{\frac{1}{N_{g1}} + \frac{1}{N_{g2}}}} \sim t((N_{g1} - 1) + \cdots + (N_{gN} - 1)) \)
  - But we’re interested in whether any pair is different from any other pair
ANOVA

• Null Hypothesis: the means of all the groups are equal

• Notation:
  • $N$: number of individuals/observation all together
  • $X$: mean for entire data set is

• Group $i$:
  • $N_i$: number of individuals in group $i$
  • $X_{ij}$: value for individual $j$ in group $i$
  • $X_i$: mean for group $i$
ANOVA: Idea

• If all the group means are the same, the average variation within the groups should be almost as large as the average variation within the entire dataset (why almost?)

• Variation BETWEEN groups:
  • For each data value look at the difference between its group mean and the overall mean: $\sum_i N_i (\bar{X}_i - \bar{X})^2$

• Variation WITHIN groups:
  • For each data value look at the difference between the value and the group mean: $\sum_i \sum_j (X_{ij} - \bar{X}_i)^2$
ANOVA: Idea

- SSReg (Regression Sum of Squares, variation across groups): \[ \sum_i N_i (\bar{X}_i - \bar{X})^2 \] (d.f.: Ngroups-1)
- RSS (Residual Sum of Squares, variation within groups): \[ \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 \] (d.f.: Npoints-Ngroups)
- Compute the ratio of the averages:

\[ F = \frac{\sum_i N_i (\bar{X}_i - \bar{X})^2}{Ngroups-1} / \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}{Npoints-Ngroups} \]
ANOVA: Idea

• $F = \frac{\sum_i (\bar{X}_i - \bar{X})^2}{N_{groups} - 1} / \frac{\sum_i \sum_j (x_{ij} - \bar{X}_i)^2}{N_{points} - n_{groups}}$

• If “average” between-group variation is not larger than “average” within-group variation (i.e., the Null Hypothesis is true), $F \approx 1$

• If between-group variation is larger than within-group variation (i.e., the means for the different groups are different), $F > 1$

  • $\frac{\sum_i N_i (\bar{X}_i - \bar{X})^2}{\sigma^2} \sim \chi^2 (N_{groups} - 1)$
  
  • $\frac{\sum_{ij} (x_{ij} - \bar{X}_i)^2}{\sigma^2} \sim \chi^2 (N_{points} - N_{groups})$

• $F \sim F(N_{groups} - 1, N_{points} - N_{groups})$
The F distribution

- If $W_1 \sim \chi^2(k_1)$ and $W_2 \sim \chi^2(k_2)$, then
  \[ F = \frac{W_1}{W_2} \sim F(k_1, k_2) \]
ANOVA: the model

- Constant variance $\sigma^2$, (possibly) different means $\mu_i$ for the different groups
  \[ X_{ij} \sim N(\mu_i, \sigma^2) \]
- Null Hypothesis: $\mu_1 = \mu_2 = \cdots = \mu_{N\text{groups}}$
- F statistic: $F = \frac{\frac{\sum_i N_i (\bar{X}_i - \bar{X})^2}{N\text{groups} - 1}}{\frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}{N\text{points} - N\text{groups}}}$
- F-test: $P_{\mu_1=\cdots=\mu_{N\text{groups}}}(F > f)$
  - If the Null Hypothesis is true,
  \[ F \sim F(N\text{groups} - 1, N\text{points} - N\text{groups}) \]
## ANOVA table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair</td>
<td>5</td>
<td>938.7</td>
<td>187.75</td>
<td>0.7858</td>
<td>0.563</td>
</tr>
<tr>
<td>Residuals</td>
<td>78</td>
<td>18636.7</td>
<td>238.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 less than # of groups

# of data values - # of groups

(equals df for each group added together)
ANOVA table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair</td>
<td>5</td>
<td>938.7</td>
<td>187.75</td>
<td>0.7858</td>
<td>0.563</td>
</tr>
<tr>
<td>Residuals</td>
<td>78</td>
<td>18636.7</td>
<td>238.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\sum_{ij} (X_{ij} - \bar{X}_i)
\]

\[
\sum_i N_i (\bar{X}_i - \bar{X})^2
\]
### ANOVA table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair</td>
<td>5</td>
<td>938.7</td>
<td>187.75</td>
<td>0.7858</td>
<td>0.563</td>
</tr>
<tr>
<td>Residuals</td>
<td>78</td>
<td>18636.7</td>
<td>238.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
MSG = \frac{SSG}{DFG} \\
MSE = \frac{SSE}{DFE}
\]

\[
P(F > f) \sim F(DFG < DFE)
\]

\[
F = \frac{MSG}{MSE}
\]
### ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair</td>
<td>5</td>
<td>938.7</td>
<td>187.75</td>
<td>0.7858</td>
<td>0.563</td>
</tr>
<tr>
<td>Residuals</td>
<td>78</td>
<td>18636.7</td>
<td>238.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The **p-value** for the F-statistic. A measure of how compatible the data is with the hypothesis that

\[ \mu_1 = \cdots = \mu_{N\text{groups}} \]
Pairwise t-Tests

• Suppose we find (using an F-test) that there are differences between the different means. That still doesn’t tell us what the differences are
• Naively, we can run a t-Test for every pair of groups
• (in R)
Problem with Multiple Comparisons

• If we are computing a p-value and the Null Hypothesis is true, we’d get a false positive 5% of the time (1 time out of 20)
  • False positive: p-value<.05, but the Null Hypothesis is true

• If we are computing 20 p-values and the Null Hypothesis is true, what percent of the time will we get at least one false positive?
Problem with Multiple Comparisons

• If we are computing 20 p-values and the Null Hypothesis is true, what percent of the time will we get at least one false positive?

\[ 1 - (1 - 0.05)^{20} \approx 64\% \]

• If we have 7 groups, and compare each mean to each other mean, how many comparisons do we make?
  • (Show in R)
Problem with Multiple Comparisons

• N variables to do pairwise comparison on:

\[ \binom{N}{2} = \frac{N(N - 1)}{2} \text{ comparisons} \]

• Intuition:
  • See the table in R
  • For each coefficient (N) of them, compare it to every other (N-1): N(N-1) comparisons. But we compared each pair twice, so divide by two: N(N-1)/2
Bonferroni correction

• Boole’s inequality: the probability of any one of the events $E_1, E_2, \ldots, E_n$ happening is smaller than $\sum_i P(E_i)$:
  
  $P(\bigcup_i E_i) \leq \sum_i P(E_i)$

  • Idea: the probability is largest when the events are mutually exclusive, in which case the probability is $\sum_i P(E_i)$

• $P \left( \bigcup_{i=1}^{n} \left( p_i \leq \frac{\alpha}{n} \right) \right) \leq \sum_{i=1}^{n} P \left( p_i \leq \frac{\alpha}{n} \right) = \frac{n\alpha}{n} = \alpha$
Bonferroni correction

• If we want the *familywise* p-value threshold to be $\alpha$, make the individual p-value threshold be $\frac{\alpha}{n}$, where $n$ is the number of groups
• Generally, *very* conservative
  • Why?
Tukey’s Honest Significant Differences (HSD)

- Tukey’s HSD is a method of adjusting the SE estimate based on the range of the data
  - Not as conservative as using the Bonferroni correction
Confidence Intervals -- Bonferroni

• If the statistic is $t$-distributed:
  $$\hat{\theta} \pm t_{df,1-\frac{\alpha}{k}} \cdot SE(\hat{\theta})$$

• (In R)
Summary: F-test and Pairwise Comparisons

• Assuming (and checking) normal distributions with constant variance in different groups:
  • Run F-test to see if any of the means are different
  • Can follow up and check pairwise differences

• If you have a hypothesis about which group means are different *ahead of time*, that’s like running multiple studies
  • Some of your multiple studies might be wrong, of course
  • Still, okay not to adjust as long as you report that you had lots of hypotheses about which means might be different
    • Of course, if you have lots of hypotheses, people might think you’re a little bit scatterbrained