ANOVA assumptions

- Data in each group are a random sample from some population.
- Observations within groups are independent.
- Samples are independent.
- Underlying populations normally distributed.
- Underlying populations have the same variance.

Diagnostics

- QQ plot within each group
- QQ plot of all residuals, $y_{ti} - \bar{y}_t$.
- Plot residuals, $y_{ti} - \bar{y}_t$, against fitted values, $\bar{y}_t$.
- Plot SD versus mean for each group.
- Plot the residuals against other factors. (e.g., order of measurements, weight or age of mouse).

Key idea: plot everything you can think of, though generally with particular goals in mind (i.e., looking for particular types of artifacts).
ANOVA Tables

Original scale / 1000:

<table>
<thead>
<tr>
<th>source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>between strains</td>
<td>33</td>
<td>20</td>
<td>1.69</td>
<td>1.70</td>
<td>0.042</td>
</tr>
<tr>
<td>within strains</td>
<td>124</td>
<td>125</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>157</td>
<td>145</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$log_{10}$ scale:

<table>
<thead>
<tr>
<th>source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>between strains</td>
<td>3.35</td>
<td>20</td>
<td>0.167</td>
<td>2.25</td>
<td>0.0036</td>
</tr>
<tr>
<td>within strains</td>
<td>9.29</td>
<td>125</td>
<td>0.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>12.63</td>
<td>145</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within-group QQ-plots : IL10
Within-group QQ-plots : $\log_{10} \text{IL10}$

QQ plots of all residuals

IL10

$log_{10} \text{IL10}$
Residuals vs fitted values

SDs vs means
Homogeneity of variances

One of the ANOVA assumptions was homogeneity of the group variances. This can formally be tested with Bartlett’s test.

Assume we have $k$ treatment groups.

- $n_t$: number of cases in treatment group $t$.
- $N$: number of cases (overall).
- $Y_{ti}$: response $i$ in treatment group $t$.
- $\bar{Y}_t$: average response in treatment group $t$.
- $S_t^2$: the sample variance in treatment group $t$.

Bartlett’s test

We want to test $H_0: \sigma_1^2 = \cdots = \sigma_k^2$ versus $H_a: H_0$ is false.

- Calculate the pooled sample variance:
  \[ S^2 = \frac{\sum_t (n_t - 1) \times S_t^2}{\sum_t (n_t - 1)} = \frac{\sum_t (n_t - 1) \times S_t^2}{N - k} \]

- Calculate the test statistic
  \[ X^2 = (N - k) \times \log(S^2) - \sum_t (n_t - 1) \times \log(S_t^2) \]

- Calculate the following correction factor:
  \[ C = 1 + \frac{1}{3(k - 1)} \left[ \sum_t \frac{1}{n_t - 1} - \frac{1}{\sum_t (n_t - 1)} \right] \]

If $H_0$ is true, then
\[ X^2/C \sim \chi^2(df=k-1) \]
Example

- For the example data, there are 21 strains with between 5 and 10 observations per strain.
- The pooled sample variance on original scale / 1000 is 0.99.
- The pooled sample variance on log\textsubscript{10} scale is 0.074.
- The test statistics were 79.9 and 34.0.
- The correction factor ended up being 1.07.
- Thus we look at the values 79.9 / 1.07 = 74.8 and 34.0 / 1.07 = 31.8.
- Since there are 21 strains, we refer to the $\chi^2$(df = 20) distribution.
- We end up with P-values of $2.9 \times 10^{-8}$ and 0.045.

The R function \texttt{bartlett.test()} can be used to do these calculations.

Another example

Rate of growth in fish eggs from different mothers
# ANOVA Table

<table>
<thead>
<tr>
<th>source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>between moms</td>
<td>12757</td>
<td>7</td>
<td>1822</td>
<td>13.5</td>
<td>$4 \times 10^{-16}$</td>
</tr>
<tr>
<td>within moms</td>
<td>73510</td>
<td>546</td>
<td>135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>86267</td>
<td>553</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
QQ plot of all residuals

QQ plots within each group
Possible transformations

- Logarithm
- Square root
- No transformation

Why transform?

- Obtain approximate normality
- Stabilize variation
- More informative graphs
- Obtain symmetry

Highly skewed data: take logs
Why take logs?

- Statistics better behaved
- Stabilize SD (esp. if coefficient of variation constant)

Note: \( \text{mean}\{ \log X \} = \log [ \text{geo. mean}\{ X \} ] \)

(As for the IL10 cytokine measurements back at the beginning of this lecture.)

Counts: take square root
Ratios: take logs

If you are interested in ratios of average responses, it might be better to look at the log ratios.

This insures that as much importance is given to ratios $< 1$ as to ratios $> 1$.

Ratios: $$(0,1) (1,\infty)$$

Log ratios: $$(-\infty,0) (0,\infty)$$

The next figure should have been made on the log scale.
Outliers

“Outlier”: an odd-looking data point (far away from the others). (This requires some sense of the scale.)

Q: Is it an error?

Q: Does it have undo influence on the results?

Example: (see data on next page)

With outlier: $P = 0.29$; 95% CI for mean diff. = ($-58, 116$)

Without outlier: $P = 0.029$; 95% CI for mean diff. = (16, 82)
What to do with outliers

• Look at your data.
• Determine the cause.
• Determine the influence of the outlier.
• Consider a method of analysis that is resistant to the effects of outliers (aka robust). (This requires more than a tiny amount of data.)
• Delete outliers with great care, and report it if you do.
A spike in the distribution

Survival time in 120 intercross mice, following infection with *Listeria monocytogenes*

Phenotype by genotype

**D5Mit357**

**D7Mit105**
Phenotype by genotype

D5Mit357

D1Mit355

D13Mit147

Genotype

Survival time (hrs)

AA AB BB

91% 57% 58%

Genotype

Survival time (hrs)

AA AB BB

40% 71% 97%

Genotype

Survival time (hrs)

AA AB BB

79% 70% 61%

Genotype

Survival time (hrs)

AA AB BB

91% 57% 58%