Outline

1 One-Way Anova
   - Two-sample case reconsidered
   - General case of multiple independent samples
   - Assumptions and Model
   - Levene’s test: are variances equal?

2 Comparisons among Means

3 Inference with Multiple Comparisons
   - Comparison-wise and Experiment-wise error rates
   - Making a selected set of comparisons: Bonferroni, protected t-tests
   - Comparing all treatments: Fisher’s LSD, Bonferroni, Tukey
One-way Anova

- So far we compared two treatments.
- One-way ANalysis Of VAriance (ANOVA) provides a way to compare any number of treatments. Extension of the two independent samples t-test assuming equal variances.

Key idea: break up the variation, i.e. sum of squares

\[ \sum (y_i - \bar{y})^2 \]

into variation explained by differences among treatments and variation within treatments.

First reconsider the independent two-sample case, then generalize the idea to independent multiple samples.
Two independent samples

Example: \( x: 4, 12, 8 \) and \( y: 17, 8, 11 \)

- Summary statistics:

\[
\bar{x} = 8, \quad \sum_{i=1}^{3} (x_i - \bar{x})^2 = 32, \quad \text{so} \quad s_x^2 =
\]

\[
\bar{y} = 12, \quad \sum_{i=1}^{3} (y_i - \bar{y})^2 = 42, \quad \text{so} \quad s_y^2 = \quad s_p^2 = \quad = 18.5
\]

- For testing \( H_0: \mu_1 = \mu_2 \) vs. \( H_A: \mu_1 \neq \mu_2 \), use t-test:

\[
t = \quad = 1.14
\]

on df = . p-value: \( 2 \times \mathbb{P}\{ T \geq 1.14 \} > 0.10 \).

There is no evidence against \( H_0 \).

Now Anova with same data: the idea is to partition the variation.
Sums of squares (SS)

**Total SS:** Total variation. Pretend all obs. form single sample.
Overall mean: \( = 10 \) and SSTotal is \( = 98 \)
on df =

**Treatment SS:** amount of the total variation explained by
differences between groups. Replace each observation by its
group mean.

X: 8, 8, 8 and Y: 12, 12, 12

Overall mean: still
and SSTrt is

\( = 24 \)
on df = 1.
Sums of squares (SS)

**Error SS:** amount of the total variation explained by differences within each group:

\[ \text{Error SS} = 74 \]

on df = 4.

- note: \( \frac{\text{SSError}}{\text{dfError}} = s_p^2 \).
- also:

\[
\begin{align*}
\text{SSTotal} & = \text{SSTrt} + \text{SSError} \\
\text{df Total} & = \text{df Trt} + \text{df Error}
\end{align*}
\]

An **Anova table** summarizes the information. Here MS = Mean Square = SS/df

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt</td>
<td>1</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>74</td>
<td>18.5</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>98</td>
<td></td>
</tr>
</tbody>
</table>
**F-test**

1. \( H_0 : \mu_1 = \mu_2 \) vs \( H_A : \mu_1 \neq \mu_2 \)

2. If \( H_0 \) is true, then

\[
F = \frac{\text{MSTrt}}{\text{MSError}} \sim F_{\text{dfTrt}, \text{dfError}}
\]

3. In the example, the observed \( f = 1.30 \). Compare this to an F-distribution with 1 df numerator and 4 df denominator using Table D. p-value:

\[
P\{F_{1,4} \geq 1.30\} > 0.10.
\]

4. No evidence against \( H_0 \). Do not reject \( H_0 \) at the 10% level.

Note: \( 1.30 = (1.14)^2 \) i.e \( f = t^2 \). This is special to 2 groups: Anova = t-test when only 2 groups.

The p-value is from one tail of the F distribution, even though \( H_A \) is two-sided.
The F distribution

- F-values are always $\geq 0$
- The F-distributions is located around 1, for all df’s.
Recap

SSTotal: total variation
SSTrt: variation due to treatment differences
SSErr: residual variation, within treatment groups

\[ F = \frac{\text{SSTrt}/\text{dfTrt}}{\text{SSErr}/\text{dfErr}} \]

- Small difference between group means relative to variability \(\rightarrow f \rightarrow \) p-value, and accept \( H_0 \).
- Large difference between group means relative to variability \(\rightarrow f \rightarrow \) p-value, and reject \( H_0 \).
Reduced Egg Investment Can Conceal Helper Effects in Cooperatively Breeding Birds

A. F. Russell,1,2*† N. E. Langmore,3 A. Cockburn,3,4 L. B. Astheimer,5 R. M. Kilner6*

Cooperative breeding systems are characterized by nonbreeding helpers that assist breeders in offspring care. However, the benefits to offspring of being fed by parents and helpers in cooperatively breeding birds can be difficult to detect. We offer experimental evidence that helper effects can be obscured by an undocumented maternal tactic. In superb fairy-wrens (Malurus cyaneus), mothers breeding in the presence of helpers lay smaller eggs of lower nutritional content that produce lighter chicks, as compared with those laying eggs in the absence of helpers. Helpers compensate fully for such reductions in investment and allow mothers to benefit through increased survival to the next breeding season. We suggest that failure to consider maternal egg-investment strategies can lead to underestimation of the force of selection acting on helping in avian cooperative breeders.
Mass (g) of chicks, 6-8 days after hatching.

**grp-pr**: laid in groups, reared in pairs.

**control**: laid and reared by their own parents (and helpers)

**pr-grp**: laid in paired, reared in groups

![Graph showing chick mass](image-url)
## Chick mass

<table>
<thead>
<tr>
<th>grp-pr</th>
<th>control</th>
<th>pr-grp</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.24</td>
<td>5.53</td>
<td>6.77</td>
</tr>
<tr>
<td>4.94</td>
<td>6.74</td>
<td>6.11</td>
</tr>
<tr>
<td>5.27</td>
<td>6.61</td>
<td>6.06</td>
</tr>
<tr>
<td>5.93</td>
<td>7.50</td>
<td>6.70</td>
</tr>
<tr>
<td>5.88</td>
<td>6.59</td>
<td>7.10</td>
</tr>
<tr>
<td>6.38</td>
<td>6.01</td>
<td>7.46</td>
</tr>
<tr>
<td>6.07</td>
<td>6.54</td>
<td></td>
</tr>
<tr>
<td>6.64</td>
<td>5.92</td>
<td>5.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.48</td>
</tr>
<tr>
<td></td>
<td>5.88</td>
<td>6.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.08</td>
</tr>
</tbody>
</table>

| sum    | 47.36   | 82.55  | 40.2  | 170.11 |
| mean   | 5.92    | 6.35   | 6.70  | 6.30   |
$k$ independent samples

$k$ treatments, $n_i$ observations for treatment $i$.

<table>
<thead>
<tr>
<th>Trt</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>$y_{11}$</td>
<td>$y_{21}$</td>
<td>...</td>
<td>$y_{k1}$</td>
</tr>
<tr>
<td></td>
<td>$y_{12}$</td>
<td>$y_{22}$</td>
<td>...</td>
<td>$y_{k2}$</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>$y_{1n_1}$</td>
<td>$y_{2n_2}$</td>
<td>...</td>
<td>$y_{kn_k}$</td>
</tr>
<tr>
<td>Sum</td>
<td>$y_{1.}$</td>
<td>$y_{2.}$</td>
<td>...</td>
<td>$y_{k.}$</td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{y}_{1.}$</td>
<td>$\bar{y}_{2.}$</td>
<td>...</td>
<td>$\bar{y}_{k.}$</td>
</tr>
</tbody>
</table>

- Sum for the $i^{th}$ trt: $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$
- Mean for the $i^{th}$ trt: $\bar{y}_{i.} = y_{i.}/n_i$
- Grand sum: $y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k} y_{i.}$
- Grand mean: $\bar{y}_{..} = y_{..}/N$ where the total # of obs is:

$$N = \sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k.$$
Partitionning the variability (Sums of Squares)

\[ SS \text{ Total} = SS \text{ Trt} + SS \text{ Error} \]

\[ df \text{ Total} = df \text{ Trt} + df \text{ Error} \]

SS Total \[= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y^2_{..}}{N} \]
on df Total \[= N - 1 \]

SS Trt \[= \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_{..})^2 = \sum_{i=1}^{k} \frac{y_i^2}{n_i} - \frac{y^2_{..}}{N} \]
on df Trt \[= k - 1 \]

SS Error \[= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i.)^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots \]
+ \[+ (n_k - 1)s_k^2 \]
on df Error \[= N - k = (n_1 - 1) + \cdots + (n_k - 1) \]
Chick mass: SS and Anova table

\[
\text{SSTrt} = \sum_{i=1}^{3} \frac{(y_{i.})^2}{n_i} - \frac{(y.)^2}{N} = 2.1563
\]

using \(s_1 = 0.5656, s_2 = 0.5201, s_3 = 0.5477\)

\[
\text{SSErr} = 6.9956
\]

\[
\text{SSTotal} = 9.1519
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt: Parenting</td>
<td>2</td>
<td>2.15</td>
<td>1.08</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>7.00</td>
<td>0.29</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>9.15</td>
<td></td>
</tr>
</tbody>
</table>

Pooled variance

MS Error = pooled estimate of variance \(s^2_p\)
Chick mass: the F test

- $H_0$: “all population means are equal” vs. $H_A$: “not all population means are equal”.
- Observed test statistic:
  \[ f = \frac{\text{MSTr}}{\text{MSErr}} = 3.70 \]

- Compare this with $F_{2,24}$ from Table D: at 5% $f_{2,24} = 3.40$, and at 1% $f_{2,24} = 5.61$, so

  \[ < p\text{-value} < \]

Reject $H_0$ at level
Moderate evidence that there is a treatment effect on chick mass.
R commands: `lm` and `anova`

```r
> chickmass = read.table("chickmass.dat", header=T)
> chickmass
   mass parents
 1   6.24  grp-pr
 2   4.94  grp-pr
  ... 
26   7.10  pr-grp
27   7.46  pr-grp
>
> fit = lm(mass~parents, data=chickmass)
> anova(fit)

Analysis of Variance Table
Response: mass

    Df  Sum Sq Mean Sq F value    Pr(>F)
parents  2 2.1563 1.0782  3.6989 0.03979 *
Residuals 24 6.9956 0.2915

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```
Assumptions

1. For each treatment, a random sample $Y_{ij} \sim N(\mu_i, \sigma_i^2)$.
2. Equal variances $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.
3. Independent samples across treatments.

The F-test

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ versus $H_A$: Not all $\mu_i$’s are equal.

Under $H_0$, the test statistic $F = \frac{\text{MSTrt}}{\text{MSError}} \sim F_{dfTrt, dfError}$

Parameter estimation

- Estimate $\sigma^2$ by $S_p^2$, i.e. MS Error.
- Estimate $\mu_i$ by $\bar{Y}_i$.
- Or if all are assumed equal, estimate $\mu$ by $\bar{Y}$.
## Review

<table>
<thead>
<tr>
<th>One-Sample</th>
<th>Two-Sample Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$\mathcal{N}$</strong></td>
<td><strong>$\mathcal{D}$</strong></td>
</tr>
<tr>
<td>$H_0: \mu = \mu_0$, CI for $\mu$. $\sigma^2$ is known ($Z$) or unknown ($T_{n-1}$)</td>
<td>$H_0: \mu = \mu_0$, CI for $\mu$ ($Z$ if large $n$)</td>
</tr>
<tr>
<td>Paired $H_0: \mu_D = 0$, CI for $\mu_D$ ($Z$ or $T_{n-1}$)</td>
<td>Paired $H_0: \mu_D = 0$ (Signed rank)</td>
</tr>
<tr>
<td>2 ind samples $H_0: \mu_1 = \mu_2$ (t-test), CI for $\mu_1 - \mu_2$</td>
<td>2 ind samples $H_0: \mu_1 = \mu_2$ (Mann-Whitney)</td>
</tr>
<tr>
<td>$k$ ind samples $H_0: \mu_1 = \cdots = \mu_k$ (Kruskal-Wallis)</td>
<td>$k$ ind samples $H_0: \sigma_1^2 = \cdots = \sigma_k^2$ (Levene’s)</td>
</tr>
<tr>
<td>2 ind samples $H_0: \sigma_1^2 = \sigma_2^2$ (Levene’s)</td>
<td>2 ind samples $H_0: \sigma_1^2 = \sigma_2^2$ (Levene’s)</td>
</tr>
</tbody>
</table>
**Assumptions**

Always address model assumptions via detection, correction, and robustness.

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>normal scores plot (or histogram)</td>
</tr>
<tr>
<td>Independence</td>
<td>Study design</td>
</tr>
<tr>
<td>Equal variance</td>
<td>Levene’s test</td>
</tr>
</tbody>
</table>
Detecting non-normality: QQ plot of residuals

Residuals: deviations from sample means \( r_{ij} = y_{ij} - \bar{y}_i \).

<table>
<thead>
<tr>
<th>( Y_1 )</th>
<th>res.</th>
<th>( Y_2 )</th>
<th>res.</th>
<th>( Y_3 )</th>
<th>res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>18</td>
<td>22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \bar{y}_1 = 6 \)
\( \bar{y}_1 = 12.5 \)
\( \bar{y}_1 = 17 \)

Normal QQplot for
3d sample values

Normal QQplot for
3d sample residuals
Detecting non-normality: QQ plot of residuals

1st sample residuals

2nd sample residuals

3rd sample residuals

All residuals
R commands - Kruskal-Wallis test

- Generalization of Mann-Whitney test to more than 2 independent samples: uses sums of ranks.
- Alternative to one-way Anova for non-normal data.

```r
> kruskal.test(mass~parents, data=chickmass)

Kruskal-Wallis rank sum test

data:  mass by parents
Kruskal-Wallis chi-squared=5.9655, df=2, p-value=0.05065
```
Detect unequal variance

- Plot treatment residuals versus treatment means (= “fitted values”). There should be no pattern (right plot).

- Or use an extension of Levene’s test for

\[ H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2. \]

Same main idea, except that a one-way ANOVA is used instead of a two-sample t-test.
Levene’s test

<table>
<thead>
<tr>
<th>grp-pr</th>
<th>control</th>
<th>pr-grp</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>99</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>124</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

1. Find the median for each sample:

\[ \hat{y}_{\text{grp-pr}} = \]
\[ \hat{y}_{\text{control}} = \]
\[ \hat{y}_{\text{pr-grp}} = \]

2. Take absolute values of deviations from median:

- **grp-pr**: 1.1 .7 .1 .1 .1 .2 .4 .6
- **control**: 1 .6 .6 .6 .5 .4 0 0 .1 .1 .2 .3 1
- **pr-grp**: .65 .65 .05 .05 .35 .75
Levene’s test

3 In samples with odd sample size, remove 1 zero.

control: 1 .6 .6 .6 .5 .4 0 .1 .1 .2 .3 1

4 Perform a one-way ANOVA f-test on the new data

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>2</td>
<td>0.008</td>
<td>0.004</td>
<td>0.036</td>
<td>( p &gt; 0.20 )</td>
</tr>
<tr>
<td>Error</td>
<td>23</td>
<td>2.632</td>
<td>0.114</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>2.640</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Outline

1. One-Way Anova
   - Two-sample case reconsidered
   - General case of multiple independent samples
   - Assumptions and Model
   - Levene’s test: are variances equal?

2. Comparisons among Means

3. Inference with Multiple Comparisons
   - Comparison-wise and Experiment-wise error rates
   - Making a selected set of comparisons: Bonferroni, protected t-tests
   - Comparing all treatments: Fisher’s LSD, Bonferroni, Tukey
Comparisons among Means

- In one-way ANOVA, if we reject $H_0$, then we know that not all treatment means are the same.
- But this may not be informative enough. We now consider particular comparisons of treatment means.
- We will consider contrasts and all pairwise comparisons.
Chick mass continued

We had $k = 3$ trts (grp-pr, control, pr-grp). Can we test $H_0 : \mu_1 = \mu_3$ vs. $H_A : \mu_1 \neq \mu_3$?

- The observed test statistic is

$$t = -2.67$$

on df = .

p-value $= 2 \times \mathbb{P}\{T \geq 2.67\}$ is between 0.01 and 0.02.

- Or confidence interval for $\mu_1 - \mu_3$:

$$(\bar{y}_1 - \bar{y}_3) \pm t_{df, \alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_3}}$$

95% CI for $\mu_{pr-grp} - \mu_{grp-pr}$: we use $t_{24,0.025} = 2.064$ and

which is $0.78 \pm 0.60$ or $[0.18 \text{ g}, 1.38 \text{ g}]$. 
Now test $H_0 : \mu_2 - \frac{1}{2}(\mu_1 + \mu_3) = 0$ vs. $H_A : \mu_2 - \frac{1}{2}(\mu_1 + \mu_3) \neq 0$.

- Estimate $\mu_2 - \frac{1}{2}(\mu_1 + \mu_3)$ by $= 0.04$ g because $\mathbb{E}\left(\bar{Y}_2. - \frac{1}{2}(\bar{Y}_1. + \bar{Y}_3.)\right) = \mu_2 - \frac{1}{2}(\mu_1 + \mu_3)$

- We will see that

$$S_{\bar{Y}_2. - \frac{1}{2}(\bar{Y}_1. + \bar{Y}_3.)} = S_p \sqrt{\frac{1}{n_2} + \frac{1}{4n_1} + \frac{1}{4n_3}}$$

here: $= 0.21$ g

We now generalize this situation.
A contrast is a quantity of the form

\[ \sum_{i=1}^{k} c_i \mu_i \]

where \( k \) = # of trts, \( \mu_i = i^{th} \) trt mean, \( c_i = i^{th} \) contrast coefficient.

We require that \( \sum_{i=1}^{k} c_i = 0 \).

We have seen two contrasts already:

\( \mu_1 - \mu_3 \) is a contrast with \( c_1 = 1, c_2 = 0, c_3 = -1 \)

\( \mu_2 - \frac{1}{2}(\mu_1 + \mu_3) \) is a contrast with \( c_1 = \), \( c_2 = \), \( c_3 = \).
Contrasts

Estimate $\sum_{i=1}^{k} c_i \mu_i$ by $X = \sum_{i=1}^{k} c_i \bar{Y}_i$.

- Here $\mathbb{E} X = \sum_{i=1}^{k} c_i \mu_i$, because

- Variance of $X$:

$$\text{var} \left( \sum_{i=1}^{k} c_i \bar{Y}_i. \right) =$$

so we can estimate $\text{sd}(X)$ by $S_X = S_p \sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_k^2}{n_k}}$

Ex: for $\mu_1 - \mu_3$ we get $s_X =$
Contrasts

The statistic

\[ T = \frac{\sum_{i=1}^{k} c_i \bar{Y}_i - \sum_{i=1}^{k} c_i \mu_i}{S_p \sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}} \]

has a T distribution on df = \( N - k \) (df Error) if the data meet the conditions for Anova.

- To test \( H_0 : \sum c_i \mu_i = 0 \) we use the test statistic

\[ t = \frac{\sum_{i=1}^{k} c_i \bar{Y}_i}{S_p \sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}}} \]

- To build a confidence interval for the contrast \( \sum c_i \mu_i \) we use

\[ \sum_{i=1}^{k} c_i \bar{Y}_i \pm t_{df \, Err, \alpha/2} \, S_p \sqrt{\sum_{i=1}^{k} \frac{c_i^2}{n_i}} \]
Chick mass: inference on $\mu_{\text{control}} - \left(\mu_{\text{pr-grp}} + \mu_{\text{grp-pr}}\right)/2$

For this contrast, $c_1 = \ldots$, $c_2 = \ldots$, $c_3 = \ldots$, so $s_X = \ldots = 0.21$

For testing $H_0 : \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0$, the observed t-value is $t = \ldots = 0.19$

on df = \ldots. p-value: $2 \times \mathbb{P}\{T \geq 0.19\} = 0.85$.

For a **95% confidence interval**: use $t = 2.064$

i.e. $0.04 \pm 0.43$ or $[-0.39, 0.47]$ g.
Contrasts in R

```r
> library(gregmisc)
# install.packages("gregmisc")  # only once to install the package

> attach(chickmass)

> fit = lm(mass ~ parents)
> levels(parents)
[1] "control" "grp-pr" "pr-grp"
> control.vs.switched = rbind(" Control-Switch" = c(1,-1/2,-1/2))
> control.vs.switched
[,1] [,2] [,3]
Control-Switch 1 -0.5 -0.5

> fit.contrast(fit, parents, control.vs.switched)

 # Estimate Std.Err t value Pr(>|t|)
parents Control-Switch 0.0414 0.209 0.198 0.845

> fit.contrast(fit, parents, control.vs.switched, conf.int=.95)

 # Estimate Std.Err t value Pr(>|t|) lowerCI upperCI
parents Control-Switch 0.0414 0.209 0.198 0.845 -0.39 0.473

> detach(chickmass)
```
Recap on Standard errors

- One sample: $S_{\bar{Y}} = S\sqrt{\frac{1}{n}}$
- Two samples: $S_{\bar{Y}_1 - \bar{Y}_2} = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- Multiple samples: $S_{\sum_{i=1}^{k} c_i \bar{Y}_i} = S_p\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_k^2}{n_k}}$

The design is **balanced** if all groups have same # observations, i.e. all $n_i = n$. Then

$$\text{var}\left(\sum_{i=1}^{k} c_i \bar{Y}_i\right) = \frac{\sigma^2}{n} \left(c_1^2 + \cdots + c_k^2\right).$$
Outline

1. One-Way Anova
   - Two-sample case reconsidered
   - General case of multiple independent samples
   - Assumptions and Model
   - Levene’s test: are variances equal?

2. Comparisons among Means

3. Inference with Multiple Comparisons
   - Comparison-wise and Experiment-wise error rates
   - Making a selected set of comparisons: Bonferroni, protected t-tests
   - Comparing all treatments: Fisher’s LSD, Bonferroni, Tukey
Concerns with Multiple Comparisons

Multiple comparisons: making several comparisons simultaneously.

- **Comparison-wise error rate (CWER)** is the Type I error rate $\alpha$ for each comparison.

- **Experimentwise error rate (EWER)** is the probability of getting at least one false rejection among multiple comparisons, given that all $H_0$’s are true.

Ex: We do 100 comparisons at $\alpha = 0.05$ each. If all 100 $H_0$’s are true, we expect 5 of them to be said “significant”. There may be up to a 99% chance that at least one is said significant: $\text{EWER} \approx 99\%$ (bad!)

Ex: 6 comparisons at $\alpha = .05$ each, EWER may be 26%.

- Problem: EWER can be much larger than CWER. In practice, control CWER, or EWER, or find a compromise.
Example: HIV vaccine trial

Science 318:1048 (13 November 2007)

AIDS RESEARCH

Did Merck’s Failed HIV Vaccine Cause Harm?


The vaccine had worked on monkeys.
ing from the unexpected failure in September of the most promising vaccine candidate in clinical trials, met here last week to explore an even more alarming finding: The vaccine, made by Merck and Co., may actually have increased the risk of HIV infection in some study participants.

Working with the academic-based HIV Vaccine Trials Network (HVTN) and the U.S. National Institutes of Health (NIH) in Bethesda, Maryland, Merck researchers stopped the multicountry study after an interim analysis revealed that the vaccine did not work (*Science*, 5 October, p. 28). Now further analysis suggests that the vaccine may have helped HIV infect a subset of participants who at the trial’s start had high levels of antibody to adenovirus 5 (Ad5), which causes the common cold and is also a component of the vaccine. “This is the worst possible outcome in a vaccine trial,” said AIDS researcher Eric Hunter of Emory University in Atlanta, Georgia, one of the trial results, Merck researchers and their partners reported that, as of 17 October, HIV had infected 83 people in the placebo-controlled

When the researchers subsequently examined the high-Ad5-antibody group, they were startled to find 21 infections in vaccinees versus nine in the placebo group.

The statistical analysis is ambiguous. Typically, researchers deem a difference as significant if it has a 95% probability of not being due to chance—a P value of less than 0.05. By these standards, the finding, with a P value of 0.029, was significant. But Steven Self, HVTN’s head statistician at the University of Washington (UW), Seattle, cautioned that this comparison merits a more stringent cutoff for significance, between 0.025 and 0.0025, because the study was not designed to assess potential harm, nor did investigators plan to evaluate a subset of the study population. Still, Self said this “trend” deserves close examination.

### Cumulative HIV Infections (males)*

<table>
<thead>
<tr>
<th>Ad5 Levels</th>
<th>All Ad5 Levels</th>
<th>49 Vaccine</th>
<th>33 Placebo</th>
<th>P value = 0.077</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative HIV Infections</td>
<td>0</td>
<td>26</td>
<td>35</td>
<td>Weeks</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>21</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>27</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>33</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>39</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
</tr>
</tbody>
</table>

* Cases accrued as of 17 October 2007.

### Cumulative HIV Infections (Ad5 > 200 Units)

<table>
<thead>
<tr>
<th>21 Vaccine</th>
<th>9 Placebo</th>
<th>P value = 0.029</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative HIV Infections</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>31</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>36</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>41</td>
<td>45</td>
<td>50</td>
</tr>
</tbody>
</table>

Double trouble. The vaccine clearly failed (*left*), but in men with high Ad5 antibodies (*right*), it may have increased their risk of infection. (Women were excluded from this analysis because only one became infected during the study.)
Selection bias

Consider \( k > 2 \) treatment comparisons in the following way:
- Take the largest and the smallest treatment means
- Compare them at \( \alpha \) level.

The actual Type I error rate is larger than \( \alpha \), because we selected the test that has the highest chance of leading to rejection.

Consider 1000 genes in a micro-array experiment. Their expression levels are measured in liver tissues and in brain tissues. Pick the gene with strongest evidence for differential expression and compare its p-value to \( \alpha = 0.05 \).
In a study of five varieties of barley, the weight of roots of \( n = 7 \) plants per variety is recorded. Group means:

\[
\bar{y}_1 \cdot \bar{y}_2 \cdot \bar{y}_3 \cdot \bar{y}_4 \cdot \bar{y}_5
\]

\[
16.3 \quad 19.3 \quad 14.7 \quad 20.3 \quad 18.5
\]

ANOVA table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt</td>
<td>4</td>
<td>145.94</td>
<td>36.48</td>
<td>5.09</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>214.74</td>
<td>7.16</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>360.68</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- Case I: a few selected comparisons, chosen in advance.
- Case II: all pairwise comparisons.
Bonferroni idea

Consider two comparisons, each with CWER = \( \alpha \). Let

\[ E_1: \text{“a Type I error is made on the first comparison”} \]

\[ E_2: \text{“a Type I error is made on the second comparison”} \]

\[
\text{EWER} = \mathbb{P}\{\text{at least one Type I error is made}\} \\
\leq \mathbb{P}(E_1) + \mathbb{P}(E_2) = 2\alpha
\]

If the two tests are independent, \( \mathbb{P}(E_1 \text{ and } E_2) = \alpha^2 \). In that case, EWER = .0975 with \( \alpha = .05 \).

Usually \( \mathbb{P}\{E_1 \text{ and } E_2\} \) is small and thus

\[ \text{EWER} \approx 2\alpha. \]

The Bonferroni inequality: with \( r \) comparisons

\[
\text{EWER} \leq r\alpha \quad \text{always} \\
= 1 - (1 - \alpha)^r \quad \text{if comparisons are independent.}
\]
Bonferroni correction on a selected set of comparisons

Example: consider making the two comparisons

\[ H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \]
\[ H_0 : (\mu_2 + \mu_3)/2 = (\mu_4 + \mu_5)/2 \]

and we want the EWER to be 0.05.

- Bonferroni: EWER \( \approx 2\alpha \) and thus we choose \( \alpha = \) for each of the two comparisons.
- In general, if \( r \) comparisons are to be made (chosen in advance) and we want EWER to be 0.05, i.e.

for each comparison \( H_0 \), perform a t-test and reject \( H_0 \) if the p-value is \( \leq 0.05/r \).
The Bonferroni correction

When making $r$ comparisons,

$$\text{EWER} \approx r \times \text{CWER}$$

Thus EW p-value = $r \times$ CW p-value.
Reject each comparison if its p-value is $< \alpha / r$, or equivalently, if its corrected EW p-value is $< \alpha$.

- Example: I made 5 comparisons and the most significant one has CW p-value $0.005 < p < 0.01$, then the EW p-value is actually $0.025 < p < 0.05$.
- Example: 1000 genes in a micro-array experiment. The gene with smallest p-value has $p = 0.0005$.

- The Bonferroni method is quite conservative, especially when $r$ is large.
Protected t-test

- Goal: control CWER, but with protection.
- Procedure: perform an overall F-test for

\[ H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \]

and if the F-test is significant at \( \alpha \), then perform a usual t-test at \( \alpha \) for each comparison.

- Past simulation studies show that the protected t-test performs well, even though it is somewhat liberal.
Making all pairwise comparisons

Barley root example: F-test gave $p < .01$. Now we want to compare all pairs of group means.

<table>
<thead>
<tr>
<th>Group</th>
<th>3</th>
<th>1</th>
<th>5</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>14.7</td>
<td>16.3</td>
<td>18.5</td>
<td>19.3</td>
<td>20.3</td>
</tr>
</tbody>
</table>

Among many approaches, we consider three approaches:
- Fisher’s least significant difference (LSD).
- Bonferroni test.
- Tukey’s method.

We also focus on methods available to balanced data with $n_1 = n_2 = \cdots = n_k$. 
Fisher’s Least Significant Difference (LSD)

For balanced data: $n$ observations for each treatment.

1. Protected LSD: Do F-test first. Proceed if $p < \alpha$ only.
2. Find the critical distance $d_L$ so that this distance leads exactly to a p-value of $\alpha$:
   \[
   \frac{d_L}{s_p \sqrt{\frac{2}{n}}} = t_{\alpha/2, \text{dfErr}}
   \]
   and thus the LSD is:

   Fishers’s LSD

   \[
   d_L = t_{\alpha/2, \text{dfErr}} \times s_p \sqrt{\frac{2}{n}}.
   \]

Fisher’s LSD

- Barley roots: F-test gave p-value < 0.01, we can proceed.
- \( n = \), \( s_p^2 = \), \( \text{dfErr} = \) and \( t_{0.025,30} = 2.042 \), so we have

\[
d_L = 2.92
\]

and:

Group: 3 1 5 2 4
Mean: 14.7 16.3 18.5 19.3 20.3

Two group means that are within 2.92 of each other are connected with a line and are not significantly different.

Interpretation is not transitive.

Alternative displays are possible. See the textbook.
Bonferroni tests

- Barley roots: the total number of comparisons is
- Bonferroni tests: Make all pairwise t-tests, using $s_p$ for all comparisons:
  \[ t_{ij} = (\bar{y}_i - \bar{y}_j) / \sqrt{s_p^2 2/n} \]

\[
\begin{array}{cccc}
1 & 5 & 2 & 4 \\
3 & 0.27 & 0.012 & 0.003 & 0.0005 \\
1 & 0.134 & 0.044 & 0.009 \\
5 & 0.58 & 0.22 \\
2 & 0.49 \\
\end{array}
\]

- declare a comparison significant if $p < 0.005$.
  i.e. if difference exceeds $d_B = t_{0.025/10,dfErr} \sqrt{s_p^2 2/n} = 4.33$

Group: 3 1 5 2 4
Mean: 14.7 16.3 18.5 19.3 20.3

- Often, it is too conservative and not useful.
Tukey’s method

- Also known as studentized range, or Q-method, or HSD for honestly significant difference.
- Controls the selection bias, from making $k(k - 1)/2$ comparisons among $k$ samples. Involves the distribution of the maximum of all t-values from all t-tests:

$$\max\{T_1, \ldots, T_{k(k-1)/2}\}$$

- Find

$$D_Q = Q_{k, \text{dfErr, } \alpha} \times S_p \sqrt{\frac{1}{n}}$$

and use it to compare sample means.

If $\mu_i$'s are all equal (i.e. all $H_0$'s are true) there is a probability $\alpha$ that the largest group mean is different from the smallest group mean by $D_Q$. 
Tukey’s method

- $Q_{k, \text{dfErr}, \alpha}$ is called the Q-score. Use a table, or R:

  ```r
  > qtukey(.05, nmeans=5, df=30, lower.tail=F)
  > 4.102079
  ```

- Barley roots: $k = \ldots$, $n = \ldots$, $s_p^2 = \ldots$, dfErr = \ldots, and at $\alpha = 0.05$, $Q_{5,30,0.05} = 4.11$ so

  $$d_Q = 4.16$$

  and thus

  Group: 3 1 5 2 4
  Mean: 14.7 16.3 18.5 19.3 20.3

- Tukey’s method tends to be conservative.
Associated confidence intervals

- Difference to declare significance:
  
  Tukey: \( d_Q = Q_{k,\text{dfErr},\alpha} \, s_p \sqrt{1/n} \)  
  LSD: \( d_L = t_{\alpha/2,\text{dfErr}} \, s_p \sqrt{2/n} \)  
  Bonferroni: \( d_B = t_{\alpha/(2r),\text{dfErr}} \, s_p \sqrt{2/n} \) 

- Associated confidence intervals: \( \bar{y}_j - \bar{y}_k \pm d_{Q,L,\text{or } B} \).

- Unequal sample sizes: both LSD and QD require adjustment
  
  \[
  d_Q = Q_{k,\text{dfErr},\alpha} \, s_p \sqrt{(1/n_1 + 1/n_2)/2} \\
  d_L = t_{\text{dfErr},\alpha/2} \, s_p \sqrt{1/n_1 + 1/n_2} \\
  d_B = t_{\text{dfErr},\alpha/(2r)} \, s_p \sqrt{1/n_1 + 1/n_2}
  \]

  (for instance)

  Complication: not the same distances \( d \) are to be used to compare *all* pairs of means.
R commands: `pairwise.t.test`

```r
squirrel = read.table("squirrelProteomic",header=T)
attach(squirrel)
plot(spot687 ~ group)
plot(log(spot687) ~ group)
# a log transformation seems appropriate, to have similar variances

fit = lm(log(spot687) ~ group)
plot(fit$residuals~fit$fitted)  # check equal variances
qqnorm(fit$residuals)         # check normality of residuals
anova(fit)    # significant F test. The following t-tests are protected.

# pairwise comparisons using the pooled SD and various adjustments
# for multiple comparisons:
pairwise.t.test(log(spot687), group)
pairwise.t.test(log(spot687), group, p.adjust="none")
pairwise.t.test(log(spot687), group, p.adjust="bonferroni")
pairwise.t.test(log(spot687), group, p.adjust="fdr")

# 'summary' gives some pairwise comparisons, but not adjusted.
summary(fit)
```
R commands: TukeyHSD

# Pairwise ‘separate’ t-tests, i.e. not using the pooled SD:
pairwise.t.test(log(spot687), group, p.adjust="none", pool.sd=FALSE)

# Checking that results are the same with separate t-tests:
t.test(log(spot687)[group=="SA I/R"], log(spot687)[group=="SA"])
t.test(log(spot687)[group=="SA"], log(spot687)[group=="IBA"])

# pairwise comparisons with Tukey’s method: use ‘aov’ to perform anova:
fit = aov(log(spot687) ~ group)
anova(fit)  # same results as before, when ‘lm’ was used.

# then use ‘TukeyHSD’:
TukeyHSD(fit)
Which method?

- **Newman-Keuls method** is intermediate approach between LSD and QD. **Duncan’s multiple range** test is another sequential method commonly used.
- These methods represent a tradeoff between controlling Type I error and power.
  
  Conservative $\rightarrow$ liberal

  Bonferroni, QD, NK, DMRT, LSD

- Choice of method depends on objectives and experiences.