The ANOVA model

Three different ways to describe the model:

A. \( Y_{ti} \) independent with \( Y_{ti} \sim N(\mu_t, \sigma^2) \)

B. \( Y_{ti} = \mu_t + \epsilon_{ti} \) where \( \epsilon_{ti} \sim iid N(0, \sigma^2) \)

C. \( Y_{ti} = \mu + \tau_t + \epsilon_{ti} \) where \( \epsilon_{ti} \sim iid N(0, \sigma^2) \) and \( \sum_t \tau_t = 0 \)
Example

For each of 8 mothers and 8 fathers, we observe (estimates of) the number of crossovers, genome-wide, in a set of independent meiotic products.

Question:

Do the fathers (or mothers) vary in the number of crossovers they deliver?
ANOVA tables

Female meioses:

<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 families</td>
<td>1485</td>
<td>7</td>
<td>212.2</td>
<td>4.60</td>
<td>0.0002</td>
</tr>
<tr>
<td>within families</td>
<td>3873</td>
<td>84</td>
<td>46.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>5358</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Male meioses:

<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 families</td>
<td>114</td>
<td>7</td>
<td>16.3</td>
<td>1.23</td>
<td>0.30</td>
</tr>
<tr>
<td>within families</td>
<td>1112</td>
<td>84</td>
<td>13.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>1226</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Permutation test

The P-values calculated above rely on the assumption that the measurements in the underlying populations are normally distributed.

Alternatively, one may use a permutation (aka randomization) test to obtain P-values.

1. Permute (shuffle) the XO counts relative to the family IDs.
2. Re-calculate the F statistic.
3. Repeat (1) and (2) many times (e.g., 1000 or 10,000 times).
4. Estimate the P-value as the proportion of the F statistics from permuted data that are $\geq$ the observed F statistic.

Permutation dist'n : Females

$\hat{P} = 0$

Observed
Another example

Are the population means the same?

By now, we know two ways of testing that:

- two-sample t-test
- ANOVA with two treatments

But do they give similar results?
### ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between treatments</td>
<td>$S_B = \sum_t n_t (\bar{Y}<em>t - \bar{Y}</em>{..})^2$</td>
<td>$k - 1$</td>
<td>$M_B = S_B / (k - 1)$</td>
</tr>
<tr>
<td>Within treatments</td>
<td>$S_W = \sum_t \sum_i (Y_{ti} - \bar{Y}_t)^2$</td>
<td>$N - k$</td>
<td>$M_W = S_W / (N - k)$</td>
</tr>
<tr>
<td>Total</td>
<td>$S_T = \sum_t \sum_i (Y_{ti} - \bar{Y}_{..})^2$</td>
<td>$(N - 1)$</td>
<td></td>
</tr>
</tbody>
</table>

### ANOVA for two groups

The ANOVA test statistic is $M_B/M_W$, with

$$M_B = n_1 (\bar{Y}_1 - \bar{Y}_{..})^2 + n_2 (\bar{Y}_2 - \bar{Y}_{..})^2$$

and

$$M_W = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2}$$
Two-sample t-test

The test statistic for the two sample t-test is

\[
t = \frac{\bar{Y}_1 - \bar{Y}_2}{s \sqrt{1/n_1 + 1/n_2}}
\]

with

\[
s^2 = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2}
\]

This also assumes equal variance within the groups!

Result

\[
\frac{M_B}{M_W} = t^2
\]
Reference distributions

If there was no difference in means, then

\[ \frac{M_B}{M_W} \sim F_{1, n_1 + n_2 - 2} \]

\[ t \sim t_{n_1 + n_2 - 2} \]

Now does this mean \( F_{1, n_1 + n_2 - 2} = (t_{n_1 + n_2 - 2})^2 \) ?

A few facts

\[ F_{1,k} = t_k^2 \]

\[ F_{k,\infty} = \frac{\chi_k^2}{k} \]

\[ N(0,1)^2 = \chi_1^2 = F_{1,\infty} = t_{\infty}^2 \]
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>⋯</th>
<th>$k_1$</th>
<th>⋯</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>⋮</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>$k_2$</td>
<td>$t_{k_2}^2$</td>
<td>.</td>
<td>$F_{k_1,k_2}$</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>⋮</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>⋮</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>$\infty$</td>
<td>$t_{\infty}^2$</td>
<td>.</td>
<td>$\chi_{k_1}^2$</td>
<td>$\frac{k_1}{k_1}$</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

**Underlying group dist'ns**

$\mu_1$

$\mu_2$

$\mu_3$

$\mu_4$

$\mu_5$

$\mu_6$

$\mu_7$

$\mu_8$

**Standard ANOVA model**

**Data**
The random effects model

Two different ways to describe the model:

A. \( \mu_t \sim \text{iid N}(\mu, \sigma_A^2) \)
   
   \[ Y_{ti} = \mu_t + \epsilon_{ti} \text{ where } \epsilon_{ti} \sim \text{iid N}(0, \sigma^2) \]

B. \( \tau_t \sim \text{iid N}(0, \sigma_A^2) \)
   
   \[ Y_{ti} = \mu + \tau_t + \epsilon_{ti} \text{ where } \epsilon_{ti} \sim \text{iid N}(0, \sigma^2) \]

\[ \rightarrow \text{ We add another layer of sampling.} \]
Hypothesis testing

In the standard ANOVA model, we considered the $\mu_t$ as fixed but unknown quantities.

We test the hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ using the statistics $M_B/M_W$ from the ANOVA table and the comparing this to an $F(k - 1, N - k)$ distribution.

In the random effects model, we consider the $\mu_t$ as random draws from a normal distribution with mean $\mu$ and variance $\sigma^2_A$.

We seek to test the hypothesis $H_0 : \sigma^2_A = 0$ versus $H_a : \sigma^2_A > 0$.

As it turns out, we end up with the same test statistic and same null distribution.

Estimation

For the random effects model it can be shown that

$$E(M_B) = \sigma^2 + n_0 \times \sigma^2_A$$

where

$$n_0 = \frac{1}{k - 1} \left( N - \sum_t n_t^2 \right)$$

Recall also that $E(M_W) = \sigma^2$.

Thus, we may estimate $\sigma^2$ by $\hat{\sigma}^2 = M_W$.

And we may estimate $\sigma^2_A$ by $\hat{\sigma}^2_A = (M_B - M_W)/n_0$

(provided that this is $\geq 0$).
The samples sizes for the 8 families were (14, 12, 11, 10, 10, 11, 15, 9), for a total sample size of 92.

Thus, \( n_0 \approx 11.45 \).

For the \textbf{female meioses}, \( M_B = 212 \) and \( M_W = 46 \). Thus

\[
\hat{\sigma} = \sqrt{46} = 6.8 \quad \text{(Note: overall sample mean = 40.3)}
\]

\[
\hat{\sigma}_A = \sqrt{\frac{212 - 46}{11.45}} = 3.81.
\]

For the \textbf{male meioses}, \( M_B = 16.3 \) and \( M_W = 13.2 \). Thus

\[
\hat{\sigma} = \sqrt{13.2} = 3.6 \quad \text{(Note: overall sample mean = 22.8)}
\]

\[
\hat{\sigma}_A = \sqrt{\frac{16.3 - 13.2}{11.45}} = 0.52.
\]