Analysis of Variance

The sample mean and variance

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (iid).

- The sample mean was defined as

  $$\bar{X} = \frac{\sum X_i}{n}$$

- The sample variance was defined as

  $$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

I haven’t spoken much about variances (I generally prefer looking at the SD), but we are about to start making use of them!
The distribution of the sample variance

If \( X_1, X_2, \ldots, X_n \) are iid Normal (mean=\( \mu \), var=\( \sigma^2 \)),
then the sample variance \( S^2 \) satisfies \( (n - 1) \frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \)

\( \xrightarrow{} \) When the \( X_i \) are not normally distributed, this is not true.

\[ \chi^2 \text{ distributions} \]

Let \( W \sim \chi^2 (df = n - 1) \)

\[ \begin{align*}
E(W) &= n-1 \\
\text{var}(W) &= 2(n-1) \\
SD(W) &= \sqrt{2(n - 1)}
\end{align*} \]

The F distribution

Let \( Z_1 \sim \chi^2_m \), and \( Z_2 \sim \chi^2_n \). Assume \( Z_1 \) and \( Z_2 \) are independent.

\( \xrightarrow{} \) Then \( \frac{Z_1/m}{Z_2/n} \sim F_{m,n} \)

\[ \text{F distributions} \]
The distribution of the sample variance ratio

Let $X_1, X_2, \ldots, X_m$ be iid Normal $(\mu_x, \sigma_x^2)$.

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal $(\mu_y, \sigma_y^2)$.

Then $(m - 1) \times \frac{S_x^2}{\sigma_x^2} \sim \chi_{m-1}^2$ and $(n - 1) \times \frac{S_y^2}{\sigma_y^2} \sim \chi_{n-1}^2$.

Hence

\[ \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{m-1,n-1} \]

or equivalently

\[ \frac{S_x^2}{S_y^2} \sim \frac{\sigma_x^2}{\sigma_y^2} \times F_{m-1,n-1} \]

Hypothesis testing

Let $X_1, X_2, \ldots, X_m$ be iid Normal $(\mu_x, \sigma_x^2)$.

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal $(\mu_y, \sigma_y^2)$.

We want to test $H_0: \sigma_x^2 = \sigma_y^2$ versus $H_a: \sigma_x^2 \neq \sigma_y^2$

\[ \rightarrow \quad \text{Under the null hypothesis} \quad S_x^2/S_y^2 \sim F_{m-1,n-1} \]
Are the variances the same in the two groups?

We want to test $H_0: \sigma_A^2 = \sigma_B^2$ versus $H_a: \sigma_A^2 \neq \sigma_B^2$

$\rightarrow$ At the 5% level, we reject the null hypothesis if our test statistic, the ratio of the sample variances (treatment group A versus B), is below 0.25 or above 4.03.

The ratio of the sample variances in our example is 2.14. We therefore do not reject the null hypothesis.
Confidence interval for the variance ratio

Let $X_1, X_2, \ldots, X_m$ be iid Normal $(\mu_x, \sigma^2_x)$.

Let $Y_1, Y_2, \ldots, Y_n$ be iid Normal $(\mu_y, \sigma^2_y)$. $X,Y$ independent.

$$\frac{S^2_x/\sigma^2_x}{S^2_y/\sigma^2_y} \sim F_{m-1,n-1}$$

Let $L$ be the 2.5th and $U$ be the 97.5th percentile of $F(m-1, n-1)$.

$$\rightarrow \quad \Pr\{L < \frac{S^2_x/\sigma^2_x}{S^2_y/\sigma^2_y} < U\} = 95\%.$$

$$\rightarrow \quad \Pr\{\frac{S^2_x/S^2_y}{U} < \frac{\sigma^2_x}{\sigma^2_y} < \frac{S^2_x/S^2_y}{L}\} = 95\%.$$

Thus, the interval $\{ \frac{S^2_x/S^2_y}{U}, \frac{S^2_x/S^2_y}{L} \}$ is a 95% confidence interval for $\sigma^2_x/\sigma^2_y$.

Example

$m = 10$; $n = 10$.

2.5th and 97.5th percentiles of $F(9,9)$ are 0.248 and 4.026.

Note that, since $m = n$, $L = 1/U$.

$$s^2_x/s^2_y = 2.14$$

$$\rightarrow \quad$$ The 95% confidence interval for $\sigma^2_x/\sigma^2_y$ is

$$(2.14 / 4.026, 2.14 / 0.248) = (0.53, 8.6)$$

How about a 95% confidence interval for $\sigma_x/\sigma_y$?
## Blood coagulation time

<table>
<thead>
<tr>
<th>Diet</th>
<th>Data</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>62 60 63 59</td>
<td>61</td>
</tr>
<tr>
<td>B</td>
<td>63 67 71 64 65 66</td>
<td>66</td>
</tr>
<tr>
<td>C</td>
<td>68 66 71 67 68 68</td>
<td>68</td>
</tr>
<tr>
<td>D</td>
<td>56 62 60 61 63 64 63 59</td>
<td>61</td>
</tr>
</tbody>
</table>

**Combined**: 64
Notation

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 \( t \)
- \( \bar{Y}_\cdot \) average response (overall)

Estimating the variability

We assume that the data are random samples from four normal distributions having the same variance \( \sigma^2 \), differing only (if at all) in their means.

We can estimate the variance \( \sigma^2 \) for each treatment \( t \), using the sum of squared differences from the averages within each group.

Define, for treatment group \( t \),

\[
S_t = \sum_{i=1}^{n_t} (Y_{ti} - \bar{Y}_t)^2.
\]

Then

\[
E(S_t) = (n_t - 1) \times \sigma^2.
\]
**Within group variability**

The **within-group sum of squares** is the sum of all treatment sum of squares:

$$S_W = S_1 + \cdots + S_k = \sum_t \sum_i (Y_{ti} - \bar{Y}_t)^2$$

The **within-group mean square** is defined as

$$M_W = \frac{S_1 + \cdots + S_k}{(n_1 - 1) + \cdots + (n_k - 1)} = \frac{S_W}{N - k} = \frac{\sum_t \sum_i (Y_{ti} - \bar{Y}_t)^2}{N - k}$$

It is our first estimator of $\sigma^2$.

---

**Between group variability**

The **between-group sum of squares** is

$$S_B = \sum_{t=1}^{k} n_t (Y_{t.} - \bar{Y}.)^2$$

The **between-group mean square** is defined as

$$M_B = \frac{S_B}{k - 1} = \frac{\sum_t n_t (Y_{t.} - \bar{Y}.)^2}{k - 1}$$

It is our second estimator of $\sigma^2$.

That is, if there is no treatment effect!
Important facts

The following are facts that we will exploit later for some formal hypothesis testing:

- The distribution of $S_W/\sigma^2$ is $\chi^2(df=N-k)$
- The distribution of $S_B/\sigma^2$ is $\chi^2(df=k-1)$ if there is no treatment effect!
- $S_W$ and $S_B$ are independent

Variance contributions

$$\sum_t \sum_i (Y_{ti} - \bar{Y}_{..})^2 = \sum_t n_t (\bar{Y}_t - \bar{Y}_{..})^2 + \sum_t \sum_i (Y_{ti} - \bar{Y}_t)^2$$

$$S_T = S_B + S_W$$

$$N - 1 = k - 1 + N - k$$
## 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>

## Example

<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>228</td>
<td>3</td>
<td>76.0</td>
</tr>
<tr>
<td>within treatments</td>
<td>112</td>
<td>20</td>
<td>5.6</td>
</tr>
<tr>
<td>total</td>
<td>340</td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>
The ANOVA model

We write \( Y_{ti} = \mu_t + \epsilon_{ti} \) with \( \epsilon_{ti} \sim \text{iid } N(0, \sigma^2) \).

Using \( \tau_t = \mu_t - \mu \) we can also write

\[
Y_{ti} = \mu + \tau_t + \epsilon_{ti}.
\]

The corresponding analysis of the data is

\[
y_{ti} = \bar{y}_. + (\bar{y}_t - \bar{y}_.) + (y_{ti} - \bar{y}_t).
\]

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 \( \epsilon_{ti} \sim \text{iid } N(0, \sigma^2) \)

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

We assume

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

Equivalently, \( Y_{ti} \sim \text{independent } \mathcal{N}(\mu_t, \sigma^2) \)

We want to test

\[ H_0 : \tau_1 = \cdots = \tau_k = 0 \quad \text{versus} \quad H_a : H_0 \text{ is false}. \]

Equivalently, \( H_0 : \mu_1 = \cdots = \mu_k \)

For this, we use a one-sided F test.

Another fact

It can be shown that

\[ E(M_B) = \sigma^2 + \frac{\sum t n_t \tau_t^2}{k-1} \]

Therefore

\[ E(M_B) = \sigma^2 \quad \text{if } H_0 \text{ is true} \]

\[ E(M_B) > \sigma^2 \quad \text{if } H_0 \text{ is false} \]
**Recipe for the hypothesis test**

Under $H_0$ we have

$$\frac{M_B}{M_W} \sim F_{k-1, N-k}.$$ 

Therefore

- Calculate $M_B$ and $M_W$.
- Calculate $M_B/M_W$.
- Calculate a p-value using $M_B/M_W$ as test statistic, using the right tail of an F distribution with $k - 1$ and $N - k$ degrees of freedom.

**Example (cont)**

$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ versus $H_a : H_0$ is false.

$M_B = 76, M_W = 5.6$, therefore $M_B/M_W = 13.57$.

Using an F distribution with 3 and 20 degrees of freedom, we get a pretty darn low p-value. Therefore, we reject the null hypothesis.

The R function `aov()` does all these calculations for you!
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.

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

Female meioses
Male meioses

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 are based on the assumption that the measurements in the underlying populations are normally distributed.

Alternatively, one may use a permutation 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 (1000 or 10,000 times, say).
4. Estimate the P-value as the proportion of the F statistics from permuted data that are bigger or equal to the observed F statistic.

Female meioses

Permutation dist’n : Females

\[ \hat{P} = 0 \]

F statistic

Observed
Male meioses

Permutation dist’n : Males

$P = 32\%$

Another example

Are the population means the same?

By now, we know two ways of testing that:
Two-sample t-test, and 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>$S_B = \sum_t n_t(\bar{Y}_t - \bar{Y}.)^2$</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

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

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

\[
N(0,1)^2 = \chi_1^2 = F_{1,\infty} = t_\infty
\]
Fixed effects

Underlying group dist'ns

Standard ANOVA model

Data

Random effects

Dist'n of group means

Underlying group dist'ns

Observed underlying group means

Random effects model

Data
The random effects model

Two different ways to describe the model:

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

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

\[ \longrightarrow \text{ We add another layer of sampling.} \]

Hypothesis testing

\[ \rightarrow \text{ In the standard ANOVA model, we considered the } \mu_t \text{ as fixed but unknown quantities.} \]

We test the hypothesis \( H_0 : \mu_1 = \cdots = \mu_k \) (versus \( H_0 \) is false) using the statistic \( \frac{M_B}{M_W} \) from the ANOVA table and the comparing this to an \( F(k – 1, N – k) \) distribution.

\[ \rightarrow \text{ In the random effects model, we consider the } \mu_t \text{ as random draws from a normal distribution with mean } \mu \text{ and variance } \sigma^2_A. \]

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. \( \text{ For one-way ANOVA, that is!} \)
**Estimation**

For the random effects model it can be shown that

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

where

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

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 \(\hat{\sigma}_A^2 = (M_B - M_W)/n_0\)
(provided that this is \(\geq 0\)).

**The first example**

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 female meioses, \(M_B = 212\) and \(M_W = 46\). Thus

\[
\hat{\sigma} = \sqrt{46} = 6.8 \quad \rightarrow \text{overall sample mean} = 40.3
\]

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

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

\[
\hat{\sigma} = \sqrt{13.2} = 3.6 \quad \rightarrow \text{overall sample mean} = 22.8
\]

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