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($\mu, \sigma^2$)
then the sample variance $s^2$ satisfies $(n - 1) s^2/\sigma^2 \sim \chi^2_{n-1}$

When the $X_i$ are not normally distributed, this is not true.

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

$E(W) = n - 1$

$\text{var}(W) = 2(n - 1)$

$SD(W) = \sqrt{2(n - 1)}$
The F distribution

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

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

\[ \text{F distributions} \]

\[ \text{df}=20,10 \]
\[ \text{df}=20,20 \]
\[ \text{df}=20,50 \]

The distribution of the sample 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(\( \mu_y, \sigma^2_y \)).

Then \( (m - 1) \times \frac{s^2_x}{\sigma^2_x} \sim \chi^2_{m-1} \) and \( (n - 1) \times \frac{s^2_y}{\sigma^2_y} \sim \chi^2_{n-1} \).

Hence

\[ \frac{s^2_x/\sigma^2_x}{s^2_y/\sigma^2_y} \sim F_{m-1,n-1} \]

or equivalently

\[ \frac{s^2_x}{s^2_y} \sim \frac{\sigma^2_x}{\sigma^2_y} \times F_{m-1,n-1} \]
Hypothesis testing

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(\( \mu_y, \sigma^2_y \)).

We want to test \( H_0: \sigma^2_x = \sigma^2_y \) versus \( H_a: \sigma^2_x \neq \sigma^2_y \)

Under the null hypothesis \( s^2_x / s^2_y \sim F_{m-1,n-1} \)

Critical regions

- If the alternative is \( \sigma^2_x \neq \sigma^2_y \), we reject if the ratio of the sample variances is unusually large or unusually small.

- If the alternative is \( \sigma^2_x > \sigma^2_y \), we reject if the ratio of the sample variances is unusually large.

- If the alternative is \( \sigma^2_x < \sigma^2_y \), we reject if the ratio of the sample variances is unusually small.
Are the variances the same in the two groups?

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

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 \( \sigma_x^2 / \sigma_y^2 \)

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\((\mu_y, \sigma_y^2)\).

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

Let \( L = 2.5\text{th %ile} \) and \( U = 97.5\text{th %ile} \) of \( F(m-1, n-1) \).

Then \( \Pr[L < (s_x^2 / \sigma_x^2) / (s_y^2 / \sigma_y^2) < U] = 95\% \).

Thus \( \Pr[(s_x^2 / s_y^2) / U < \sigma_x^2 / \sigma_y^2 < (s_x^2 / s_y^2) / L] = 95\% \).

Thus, the interval \((s_x^2 / s_y^2) / U, (s_x^2 / s_y^2) / L)\)

is a 95\% confidence interval for \( \sigma_x^2 / \sigma_y^2 \).

Example

\( m = 10; n = 10 \).

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

(Note that, since \( m = n \), \( L = 1/U \).)

\[
s_x^2 / s_y^2 = 2.14
\]

The 95\% confidence interval for \( \sigma_x^2 / \sigma_y^2 \) is

\[
(2.14 / 4.026, 2.14 / 0.248) = (0.53, 8.6)
\]

How about a 95\% confidence interval for \( \sigma_x / \sigma_y \)?
<table>
<thead>
<tr>
<th>Diet</th>
<th>Coagulation Time</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
Assume we have k treatment groups.

\begin{align*}
    n_t & \quad \text{number of cases in treatment group } t \\
    N & \quad \text{number of cases (overall)} \\
    Y_{ti} & \quad \text{response } i \text{ in treatment group } t \\
    \bar{Y}_t & \quad \text{average response in treatment group } t \\
    \bar{Y} & \quad \text{average response (overall)}
\end{align*}

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 estimate of \( \sigma^2 \).

Between group variability

The between-group sum of squares is

\[ S_B = \sum_{t=1}^{k} n_t (\bar{Y}_t - \bar{Y}_{..})^2 \]

The between-group mean square is defined as

\[ M_B = \frac{S_B}{k - 1} = \frac{\sum_t n_t (\bar{Y}_t - \bar{Y}_{..})^2}{k - 1} \]

It is our second estimate 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>$S_B = \sum_t n_t (\bar{Y}_t - \bar{Y})^2$</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>

**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

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

Hypothesis testing

We assume

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

[equivalently, \( Y_{ti} \sim \text{independent N}(\mu_t, \sigma^2) \)]

We want to test

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

[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)

\[ H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \]
\[ H_a : H_0 \text{ is false.} \]

\[ M_B = 76, M_W = 5.6, \text{ therefore } M_B/M_W = 13.57. \] 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.

\[ \begin{align*}
0 & 2 4 6 8 10 12 14 \\
M_B & = 76, M_W = 5.6, \text{ therefore } M_B/M_W = 13.57.
\end{align*} \]

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

Now what did we do...?

\[
\begin{pmatrix}
62 & 63 & 68 & 56 \\
60 & 67 & 66 & 62 \\
63 & 71 & 71 & 60 \\
59 & 64 & 67 & 61 \\
65 & 68 & 63 \\
66 & 68 & 64 \\
63 \\
59
\end{pmatrix} =
\begin{pmatrix}
64 & 64 & 64 & 64 \\
64 & 64 & 64 & 64 \\
64 & 64 & 64 & 64 \\
64 & 64 & 64 & 64 \\
63 \\
64 \\
64 \\
64
\end{pmatrix} +
\begin{pmatrix}
-3 & 2 & 4 & -3 \\
-3 & 2 & 4 & -3 \\
-3 & 2 & 4 & -3 \\
-3 & 2 & 4 & -3 \\
2 & 4 & -3 \\
2 & 4 & -3 \\
2 & 4 & -3 \\
2 & 4 & -3
\end{pmatrix} +
\begin{pmatrix}
1 & -3 & 0 & -5 \\
1 & -3 & 0 & -5 \\
1 & -3 & 0 & -5 \\
1 & -3 & 0 & -5 \\
-1 & 1 & -2 & 1 \\
2 & 5 & 3 & -1 \\
-2 & -2 & -1 & 0 \\
-1 & 0 & 2 \\
0 & 0 & 3 \\
2 \\
-2
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>observations</th>
<th>grand average</th>
<th>treatment deviations</th>
<th>residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_{ii})</td>
<td>(\bar{y}<em>{..}) + (\bar{y}<em>t - \bar{y}</em>{..}) + (y</em>{ii} - \bar{y}_t)</td>
<td>(T) + (R)</td>
<td></td>
</tr>
</tbody>
</table>

Vector \(\mathbf{Y} = \mathbf{A} + \mathbf{T} + \mathbf{R}\)

Sum of Squares \(98,644 = 98,304 + 228 + 112\)

D's of Freedom \(24 = 1 + 3 + 20\)