STA221

Neil Montgomery

Last edited: 2017-03-06 13:06

the F distributions

Call:

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SSE}{N-k}$$

the F distributions

Call:

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SSE}{N-k}$$

These are called “mean squares”, and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ “degrees of freedom”.

the F distributions

Call:

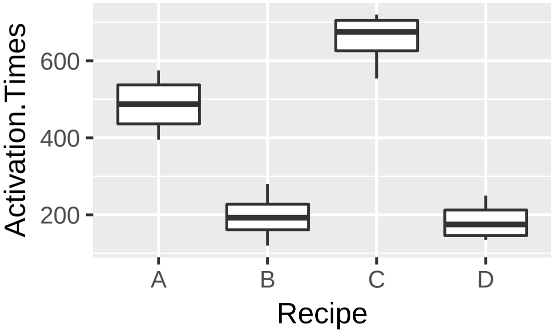
$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SSE}{N-k}$$

These are called “mean squares”, and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ “degrees of freedom”.

When the null hypothesis is true, $\frac{MS_T}{MS_E}$ lives near 1, and large values of this ratio give small p-values.

putting it all together

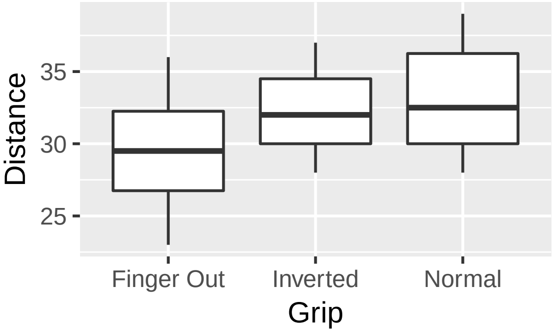
All this information is concisely displayed in what is called the “analysis of variance” table (or ANOVA table, or AOV table). Here’s the table for the Yeast example:



```
##              Df Sum Sq Mean Sq F value      Pr(>F)
## Recipe        3 638968  212989   44.74 0.000000864
## Residuals    12  57128    4761
```

putting it all together

And for the “probably not different” Frisbee example:



```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Grip          2  58.58   29.29   2.045  0.154
## Residuals    21 300.75   14.32
```

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>					
Residuals					

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k - 1$				
Residuals	$N - k$				

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k - 1$	SS_T			
Residuals	$N - k$	SS_E			

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$		
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the `R` output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the `R` output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

For example (from 25.13 “Hearing”). Four different word lists were compared for ease of hearing with background noise. 96 people were divided into four groups and the number out of 24 words understood was recorded.

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

For example (from 25.13 "Hearing"). Four different word lists were compared for ease of hearing with background noise. 96 people were divided into four groups and the number out of 24 words understood was recorded.

The sample variance for all 96 people was 70.0907895. The mean squared error was 62.371. Is there a difference between the word groups?

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

That's because the square of anything with a t_ν distribution always has an $F_{1,\nu}$ distribution.

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

That's because the square of anything with a t_ν distribution always has an $F_{1,\nu}$ distribution.

For example, consider the two-sample t test (equal variance version using pooled variance s_p^2 - section 21.3 of the text).

Q21.20 "Hard Water" Mortality rates per county in 61 counties in England and Wales classified as "North" and "South" of Derby. Is there a difference in mean mortality rate? Here's the R output:

```
##
##  Two Sample t-test
##
## data:  Mortality by Derby
## t = 6.5312, df = 59, p-value = 0.00000001673
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

mortality data via F test

From above:

```
t = 6.5312, df = 59, p-value = 0.00000001673
```

The R ANOVA output:

```
##               Df  Sum Sq Mean Sq F value       Pr(>F)
## Derby          1  886712  886712   42.66 0.0000000167
## Residuals   59 1226462   20787
```

mortality data via *F* test

From above:

```
t = 6.5312, df = 59, p-value = 0.00000001673
```

The `R` ANOVA output:

```
##             Df  Sum Sq Mean Sq F value       Pr(>F)
## Derby        1  886712  886712   42.66 0.0000000167
## Residuals   59 1226462   20787
```

Also, $6.5312^2 = 42.6565734$.

the t - F connection - II

Recall from the `Bodyfat` example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719
## Height      -0.09316    0.20119  -0.463   0.6438
##
## Residual standard error: 8.307 on 248 degrees of freedom
## Multiple R-squared:  0.0008637,  Adjusted R-squared:  -0.003165
## F-statistic: 0.2144 on 1 and 248 DF,  p-value: 0.6438
```

the t - F connection - II

Recall from the `Bodyfat` example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719
## Height      -0.09316    0.20119  -0.463   0.6438
##
## Residual standard error: 8.307 on 248 degrees of freedom
## Multiple R-squared:  0.0008637,  Adjusted R-squared:  -0.003165
## F-statistic: 0.2144 on 1 and 248 DF,  p-value: 0.6438
```

Again, $t^2 = F$ and the p-values are identical.

the t - F connection - II

Recall from the Bodyfat example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719
## Height      -0.09316    0.20119  -0.463   0.6438
##
## Residual standard error: 8.307 on 248 degrees of freedom
## Multiple R-squared:  0.0008637,  Adjusted R-squared:  -0.003165
## F-statistic: 0.2144 on 1 and 248 DF,  p-value: 0.6438
```

Again, $t^2 = F$ and the p-values are identical.

The practical downside of using F is that you lose information about the sign.

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

The main things to verify:

1. Do the groups come from a distribution with the same variance? (fatal if no)

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

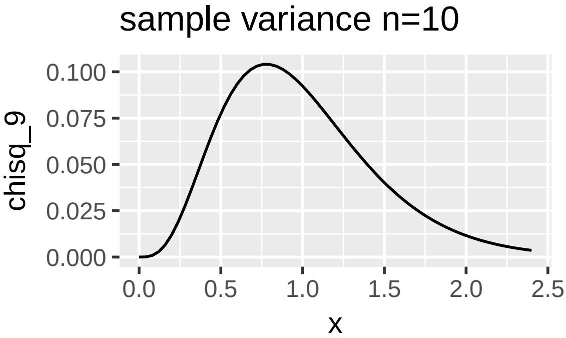
The main things to verify:

1. Do the groups come from a distribution with the same variance? (fatal if no)
2. Do the groups come from normal distributions? (OK if sample size is large enough)

Formal test for equality of variances

For the equal variance assumptions, the book says to look at plots. Other books gives a variety of heuristics. These suggestsions tend to be wildly conservative.

The problem is twofold. Within-group sample sizes tend to be small. And the sample variance itself has a very large variance.



Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

The form of the test is exactly an ANOVA, but not on the original $y_i j$. Instead, it is on the *absolute differences from the group medians*:

$$Z_{ij} = |y_{ij} - \tilde{y}_i|$$

where $\tilde{y}_i$ is the sample median of the i^{th} group.

Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

The form of the test is exactly an ANOVA, but not on the original $y_i j$. Instead, it is on the *absolute differences from the group medians*:

$$Z_{ij} = |y_{ij} - \tilde{y}_i|$$

where $\tilde{y}_i$ is the sample median of the i^{th} group.

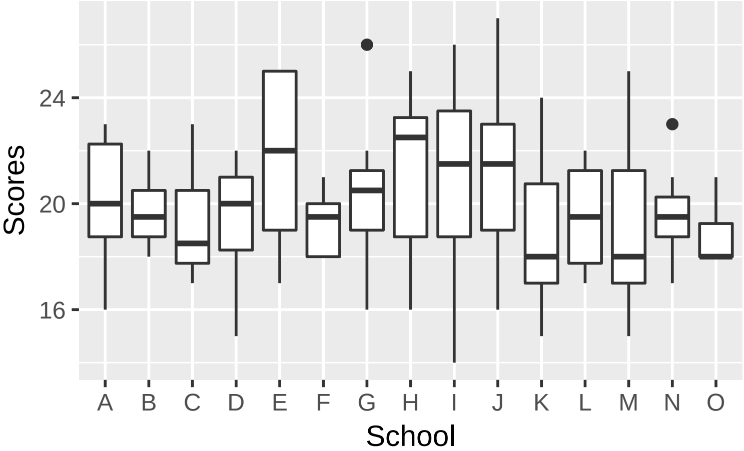
Plugging the Z_{ij} into the ANOVA formulae gives an approximate $F_{k-1,N-k}$ distribution.

Levene's test example - yeast

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.2917 0.8306
##       12
```

tougher example

From textbook question 25.18 “School System”. 15 schools selected. 8 students per school.



tougher example - Levene

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  14  1.2642 0.2422
##       105
```

normality assumption

Technically, all the groups have to be normal. But the samples sizes are usually too small.

normality assumption

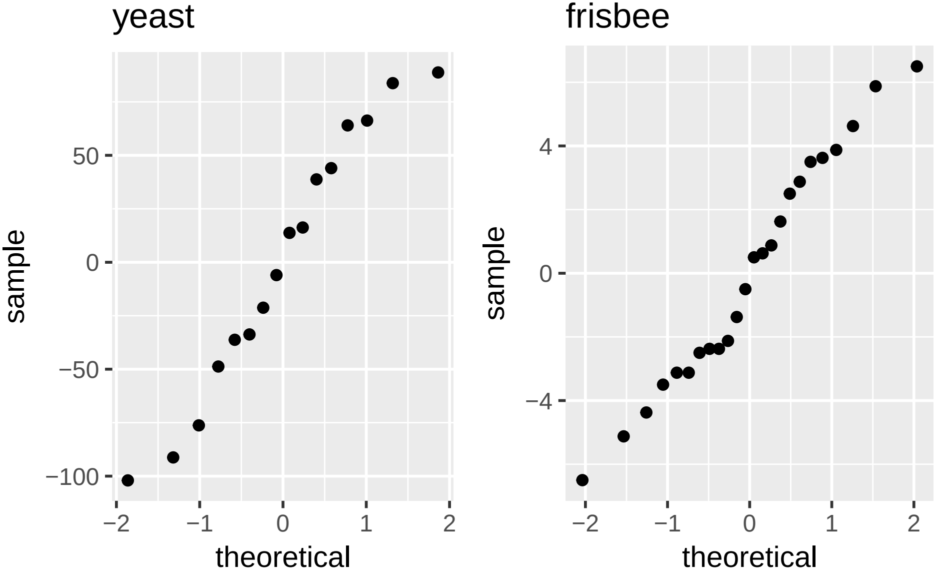
Technically, all the groups have to be normal. But the samples sizes are usually too small.

If the equal variance assumption has been satisfied (do that first), then the method is to pool all the *residuals* together:

$$y_{ij} - \overline{y}_i$$

and look at a normal quantile plot.

normal assumption verification examples



pairwise comparisons

The ANOVA F-test is an “omnibus” test—it tells you if there are *any* differences, without giving information about where the differences might be.

pairwise comparisons

The ANOVA F-test is an “omnibus” test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

pairwise comparisons

The ANOVA F-test is an “omnibus” test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

Or, there may be some differences between groups that are noticed after collecting data, which is gets us into dangerous territory!

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

Or, there may be some differences between groups that are noticed after collecting data, which is gets us into dangerous territory!

The approach in any case will be to perform multiple pooled two-sample t procedures, using the overall MSE in place of the usual pooled variance:

$$\frac{\overline{y}_i - \overline{y}_j}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k}$$

The usual technique is to produce confidence intervals for each desired pair. But at what confidence level? The usual 95% level leads to a problem...