STA221

Neil Montgomery

Last edited: 2017-07-25 19:18

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.
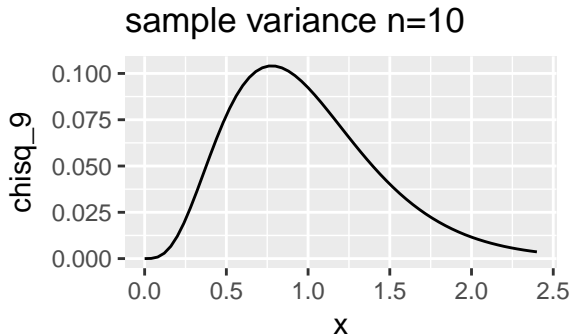
Levene's test

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_ij. 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 ith 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*:

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

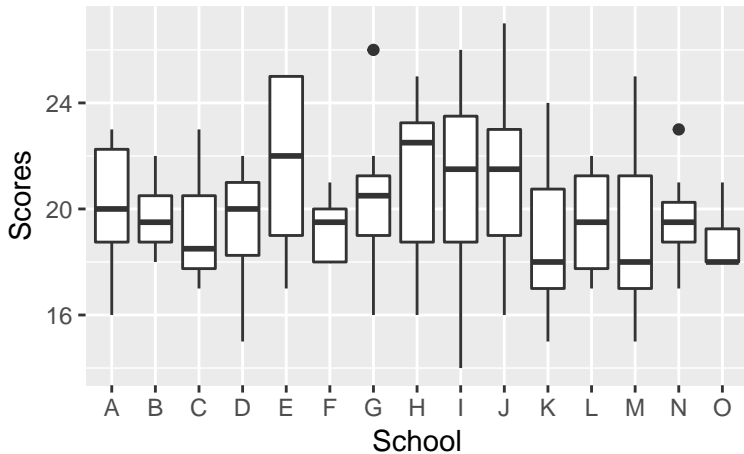where $\tilde{y}_i$ is the sample median of the ith group.

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

Levene's test example - yeast

```
## # A tibble: 2 × 3
##      Df `F value` `Pr(>F)`
## * <int>    <dbl>    <dbl>
## 1     3 0.291724 0.8306011
## 2    12       NA        NA
```

tougher example

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

tougher example - Levene

```
## # A tibble: 2 × 3
##      Df `F value`  `Pr(>F)`
## * <int>    <dbl>     <dbl>
## 1    14  1.264175 0.2422268
## 2   105       NA        NA
```

normality assumption

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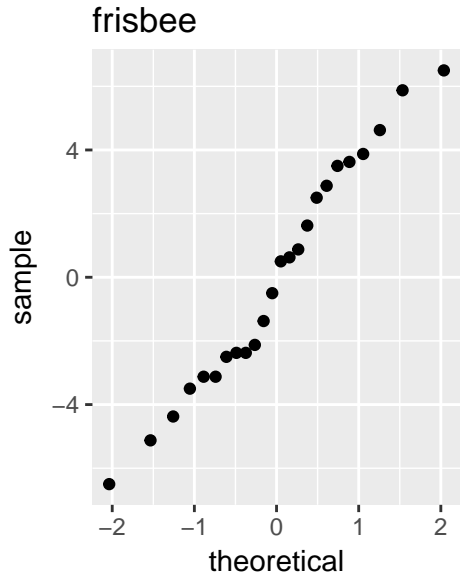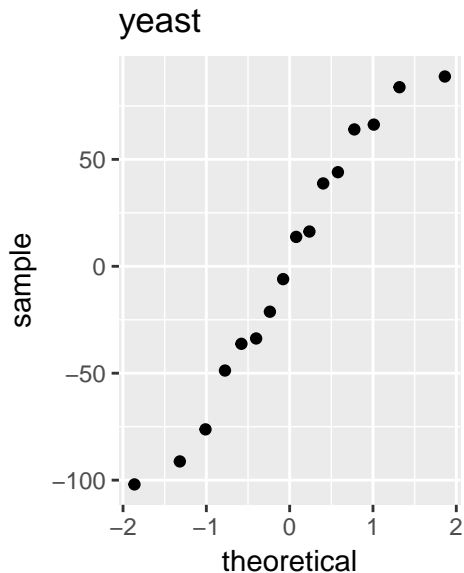
normality assumption

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. . .

pairwise comparisons

pairwise comparisons

A pairwise comparison will be a 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.

pairwise comparisons

A pairwise comparison will be a 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.

We will adjust the confidence level $100 \cdot (1 - \alpha)\%$ when multiple comparisons take place.

pairwise comparisons

A pairwise comparison will be a 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.

We will adjust the confidence level $100 \cdot (1 - \alpha)\%$ when multiple comparisons take place.

Recall: a lower α gives a *wider* confidence interval. In the t case the full formula is:

$$(\overline{y}_i - \overline{y}_j) \pm t_{N-k,\alpha/2}\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

When you subject one dataset to multiple hypothesis tests at the same "level" α, you are exposing yourself to that α probability over and over again.

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

When you subject one dataset to multiple hypothesis tests at the same "level" α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

When you subject one dataset to multiple hypothesis tests at the same "level" α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

But, it is possible to put an *upper bound* on the overall effect.

if you torture your data long enough, it will tell you anything

From "classical" hypothesis testing there was the notion of a "Type I Error", which is *rejecting H_0 when it is true*, aka "false positive".

The probability of a Type I Error is often called α and is set to be something low like 0.05. This is also called the "level" of the test.

When you subject one dataset to multiple hypothesis tests at the same "level" α, you are exposing yourself to that α probability over and over again.

Also, all the different hypothesis tests are usually not independent, so calculating the overall effect of multiple hypothesis tests is usually impossible.

But, it is possible to put an *upper bound* on the overall effect.

Definition: the *experimentwise error rate* is the probability of *any* Type I Errors among all tests done on the dataset from one experiment.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

You might recall the expression $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

bounding the experimentwise error rate - I

Suppose we're going to do m hypothesis tests on a dataset.

Denote by $A_1, A_2, \ldots, A_m$ the events where A_i means "a Type I Error occurred when hypothesis test i took place", and $P(A_i) = \alpha$.

The goal is to put one overall upper bound on the experimentwise error rate, which we'll call α^*.

You might recall the expression $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This implies $P(A \cup B) \leq P(A) + P(B)$. You can extend this to any number of events, i.e.:

$$P(A_1 \cup A_2 \cup \cdots \cup A_m) \leq P(A_1) + P(A_2) + \cdots + P(A_m)$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\alpha^* = P(\text{any Type I Errors})$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\alpha^* = P(\text{any Type I Errors})$$
$$= P(A_1 \cup A_2 \cup \cdots A_m)$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\begin{aligned}
\alpha^* &= P(\text{any Type I Errors}) \\
&= P(A_1 \cup A_2 \cup \cdots A_m) \\
&\leq P(A_1) + P(A_2) + \cdots + P(A_m)
\end{aligned}$$

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\begin{aligned}
\alpha^* &= P(\text{any Type I Errors}) \\
&= P(A_1 \cup A_2 \cup \cdots A_m) \\
&\leq P(A_1) + P(A_2) + \cdots + P(A_m)
\end{aligned}$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\alpha^* = P(\text{any Type I Errors})$$
$$= P(A_1 \cup A_2 \cup \cdots A_m)$$
$$\leq P(A_1) + P(A_2) + \cdots + P(A_m)$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

This is called a "Bonferroni correction".

bounding the experimentwise error rate - II

How could the individual tests all be adjusted so that α^* remains small?

An easy method uses:

$$\begin{aligned} \alpha^* &= P(\text{any Type I Errors}) \\ &= P(A_1 \cup A_2 \cup \cdots A_m) \\ &\leq P(A_1) + P(A_2) + \cdots + P(A_m) \end{aligned}$$

So can achieve $\alpha^* = \alpha$ simply by dividing each of these α by m.

This is called a "Bonferroni correction".

It's not a bad idea to apply a Bonferroni correct to any situation in which you are subjecting a dataset to lots of hypothesis tests.

full example, including some pairwise comparisons

From the "Hearing" example there were four lists of words and 96 people. The % of words understood by each person was recorded.

full example, including some pairwise comparisons

From the "Hearing" example there were four lists of words and 96 people. The % of words understood by each person was recorded.

Suppose List4 was some sort of "default list", and the other three lists were new word lists being evaluated. So it will be particularly interesting to investigate these three pairwise differences:

$$\mu_1 - \mu_4$$
$$\mu_2 - \mu_4$$
$$\mu_3 - \mu_4$$

full example, including some pairwise comparisons

From the "Hearing" example there were four lists of words and 96 people. The % of words understood by each person was recorded.

Suppose List4 was some sort of "default list", and the other three lists were new word lists being evaluated. So it will be particularly interesting to investigate these three pairwise differences:

$$\mu_1 - \mu_4$$
$$\mu_2 - \mu_4$$
$$\mu_3 - \mu_4$$

We will fix the experimentwise error rate at $\alpha^* = 0.05$ for the multiple comparisons.

hearing full example - I

First, look at a plot:
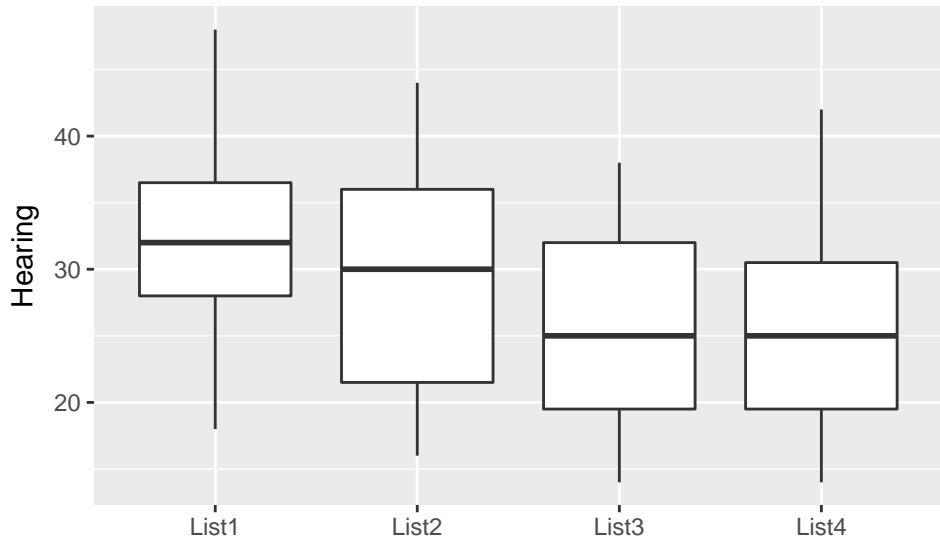
hearing full example - II

Next, verify the model assumptions starting with Levene's test:
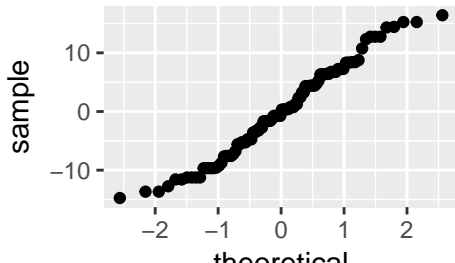
```
## # A tibble: 2 × 3
##      Df `F value`  `Pr(>F)`
## * <int>     <dbl>     <dbl>
## 1      3 0.5701226 0.6360853
## 2     92        NA        NA
```

hearing full example - II

Next, verify the model assumptions starting with Levene's test:

```
## # A tibble: 2 × 3
##       Df `F value`  `Pr(>F)`
## * <int>     <dbl>     <dbl>
## 1     3 0.5701226 0.6360853
## 2    92        NA        NA
```

. . . followed by the normal quantile plot of the residuals:

Next we do the overall ANOVA *F* test:

```
##            Df Sum Sq Mean Sq F value  Pr(>F)
## ListID      3    920  306.82   4.919 0.00325
## Residuals  92   5738   62.37
```

hearing full example - III

Next we do the overall ANOVA *F* test:

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## ListID        3    920  306.82   4.919 0.00325
## Residuals    92   5738   62.37
```

And since the p-value is low **we may proceed with the pairwise comparisons.**

hearing full example - IV

To make the three confidence intervals we need the estimated mean differences and the group sample sizes:

```
## # A tibble: 4 × 3
##   ListID     n     mean
##   <fctr> <int>    <dbl>
## 1 List1     24 32.75000
## 2 List2     24 29.66667
## 3 List3     24 25.25000
## 4 List4     24 25.58333
```

We are doing three comparisons at an experimentwise error rate of 0.05, so we'll produce the $(1 - 0.05/3) \cdot 100\% = 98.33\%$ confidence intervals.

hearing full example - IV

To make the three confidence intervals we need the estimated mean differences and the group sample sizes:

```
## # A tibble: 4 × 3
##   ListID     n     mean
##   <fctr> <int>    <dbl>
## 1 List1     24 32.75000
## 2 List2     24 29.66667
## 3 List3     24 25.25000
## 4 List4     24 25.58333
```

We are doing three comparisons at an experimentwise error rate of 0.05, so we'll produce the $(1 - 0.05/3) \cdot 100\% = 98.33\%$ confidence intervals.

The value of $t_{92,0.0167}$ is 2.4385416.

The three pairwise comparisons of interest can be made using these confidence intervals:

Comparison	Estimate	Margin of Error	Lower	Upper
$\mu_1 - \mu_4$	7.167	5.559	1.607	12.726
$\mu_2 - \mu_4$	4.083	5.559	-1.476	9.643
$\mu_3 - \mu_4$	-0.333	5.559	-5.893	5.226

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

Trick: use Bonferroni's correction *assuming you were going to look at all the comparisons in advance.*

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

Trick: use Bonferroni's correction *assuming you were going to look at all the comparisons in advance*.

With k groups there will be $k(k-1)/2$ such comparisons.

post-hoc comparison trick

Dangerous territory: perform a comparison *after looking at the data*.

Trick: use Bonferroni's correction *assuming you were going to look at all the comparisons in advance*.

With k groups there will be $k(k-1)/2$ such comparisons.

Example using question 25.19 "Fertilizers". There are $k = 10$ fertilizers being compared with $n = 10$ mung bean sprouts each. After a week, the bean heights are measured.

"Fertilizers" example
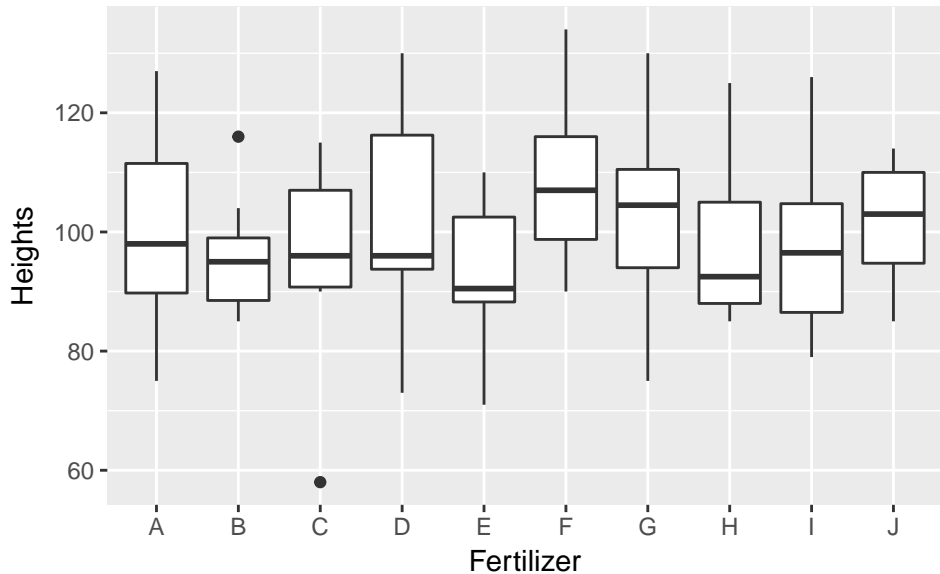
"Fertilizers" example

E and F look interesting. I think I'll test that pairwise difference at the end.
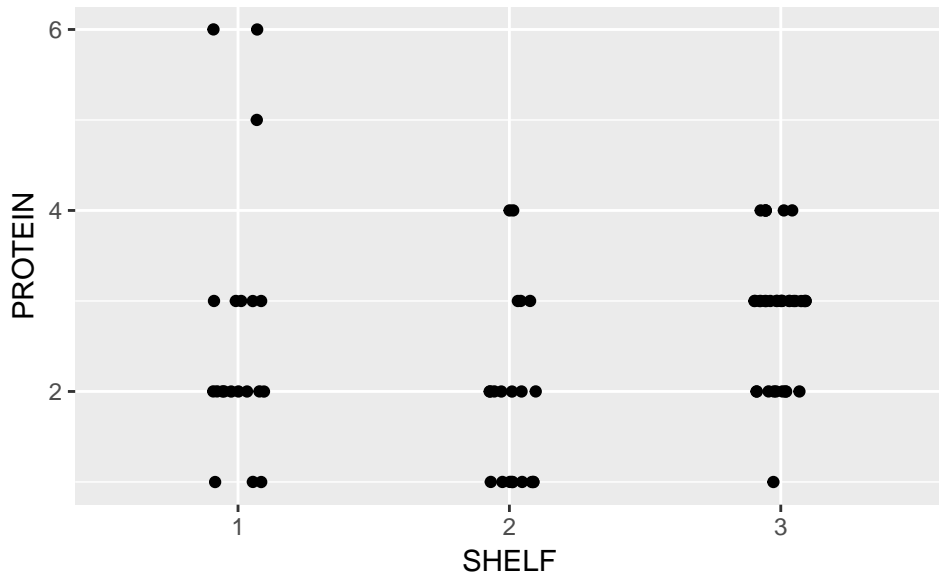
"Fertilizers" example

E and F look interesting. I think I'll test that pairwise difference at the end.

I need to run the ANOVA and verify the assumptions:

```
## # A tibble: 2 × 5
##      Df `Sum Sq` `Mean Sq` `F value` `Pr(>F)`
## * <int>    <dbl>     <dbl>     <dbl>    <dbl>
## 1     9     2074       230      1.19     0.31
## 2   110    21331       194        NA       NA


## # A tibble: 2 × 3
##      Df `F value` `Pr(>F)`
## * <int>     <dbl>    <dbl>
## 1     9     0.742     0.67
## 2   110        NA       NA
```

OK, so then let's look at the "Cereals" data from Q25.21

"Cereals redux"

I'd also like to see if there is a difference between shelves 2 and 3.
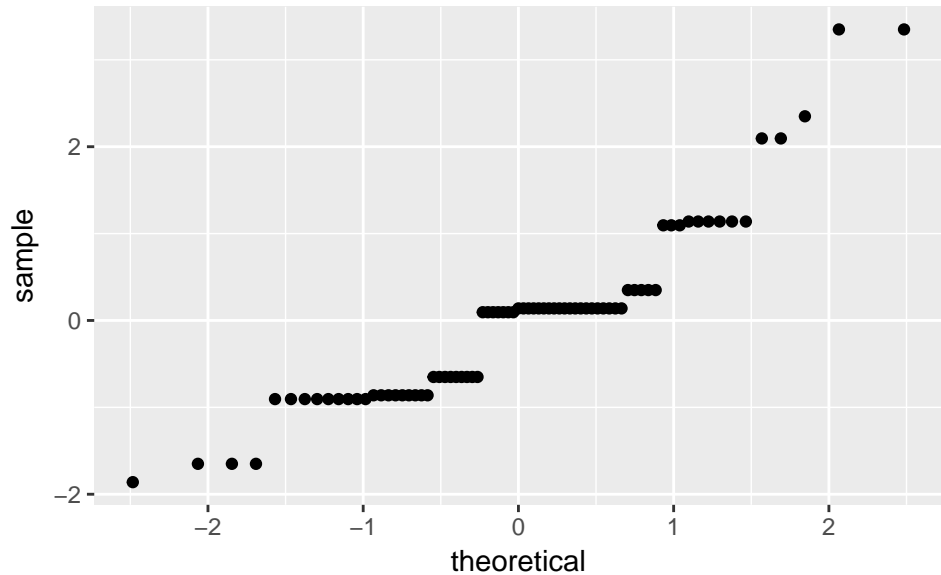
"Cereals redux"

I'd also like to see if there is a difference between shelves 2 and 3.

Start with the analysis and assumption verification:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## SHELF        2   12.4    6.21    5.84 0.0044
## Residuals   74   78.7    1.06

## # A tibble: 2 × 3
##       Df `F value` `Pr(>F)`
## * <int>     <dbl>    <dbl>
## 1     2      2.34    0.103
## 2    74        NA       NA
```

"Cereals redux"

comparing shelves 2 and 3

There are 3 *possible* comparisons, so a Bonferroni correction with $m = 3$ will be needed, even though I'm only actually doing one comparison.

comparing shelves 2 and 3

There are 3 *possible* comparisons, so a Bonferroni correction with $m = 3$ will be needed, even though I'm only actually doing one comparison.

Computer says: $t_{74, 0.05/6} = 2.45$

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## SHELF          2   12.4    6.21    5.84 0.0044
## Residuals     74   78.7    1.06

## # A tibble: 3 × 3
##     SHELF     n  mean
##    <fctr> <int> <dbl>
## 1       1    20  2.65
## 2       2    21  1.90
## 3       3    36  2.86
```

"All pairwise comparisons"

Sometimes it is valuable to simply summarize all possible pairwise comparisons to determine which groups are the same and which are different.

"All pairwise comparisons"

Sometimes it is valuable to simply summarize all possible pairwise comparisons to determine which groups are the same and which are different.

Here is an efficient algorithm for performing this task for cases when the group sample sizes are all the same (equal to some n). Let's look at the Yeast example again.

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Recipe        3 638968  212989    44.7 0.00000086
## Residuals    12  57128    4761
```

```
## # A tibble: 4 × 3
##   Recipe     n  mean
##   <fctr> <int> <dbl>
## 1      D     4   184
## 2      B     4   196
## 3      A     4   486
## 4      C     4   656
```

"all pairwise" with Yeast

"all pairwise" with Yeast

Computer says: $t_{12, 0.05/12} = 3.153$

```
## # A tibble: 4 × 3
##   Recipe     n  mean
##   <fctr> <int> <dbl>
## 1      D     4   184
## 2      B     4   196
## 3      A     4   486
## 4      C     4   656
```

"all pairwise" with Yeast

Computer says: $t_{12, 0.05/12} = 3.153$

```
## # A tibble: 4 × 3
##   Recipe     n  mean
##   <fctr> <int> <dbl>
## 1      D     4   184
## 2      B     4   196
## 3      A     4   486
## 4      C     4   656
```

The "margin of error" is 153.82.