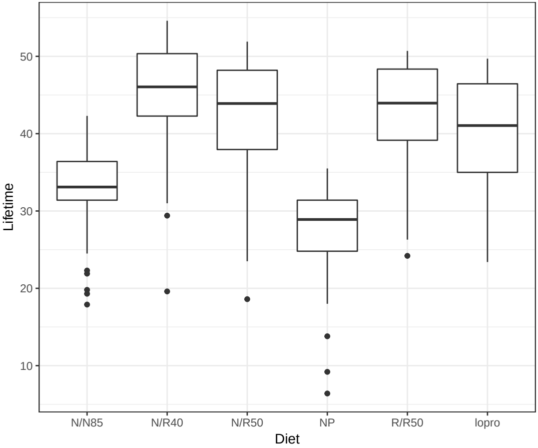
Set R06 - ANOVA and F-tests

STAT 401 (Engineering) - Iowa State University

April 3, 2017

Lifetime (months) of mice on different diets



One-way ANOVA model/assumptions

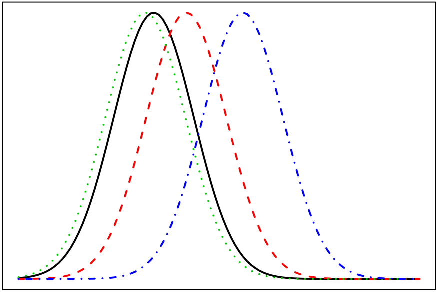
$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$.
(n_j means there can be different # of observations in each group)

Assumptions:

- Normality
  - Not skewed
  - Not heavy-tailed
- Common variance for all groups
- Independence
  - No cluster effects
  - No serial effects
  - No spatial effects

ANOVA assumptions graphically



What if you want to compare two groups?

We may still be interested in comparing two groups.

Statistical hypothesis: Is there a difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Statistical question: What is the difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Two-group analysis

Begin with the two group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

with $j = 1, 2$ and $i = 1, \ldots, n_j$
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 - 2$ degrees of freedom

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n1 + n_2 - 2}$$

What if you have more than two groups?

Multi-group analysis

The multi-group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

but now $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$
(n_j means there can be different # of observations in each group)
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_J - 1)s_J^2}{n1 + n_2 + \cdots + n_J - J}$$

Hypothesis test for comparison of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$ and we want to test the hypothesis

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

then we compute:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

where

$$SE(\overline{Y}_1 - \overline{Y}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_J - 1)s_J^2}{n_1 + n_2 + \cdots + n_J - J}.$$

Then we compare t to a t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom.

Diet effect on mice lifetime

```
# A tibble: 6  4
    Diet     n     mean       sd
  <fctr> <int>    <dbl>    <dbl>
1  N/N85    57 32.69123 5.125297
2  N/R40    60 45.11667 6.703406
3  N/R50    71 42.29718 7.768195
4     NP    49 27.40204 6.133701
5  R/R50    56 42.88571 6.683152
6  lopro    56 39.68571 6.991695
```

Test for difference in mean lifetime between NP and N/N85, i.e.

$$H_0 : \mu_4 = \mu_1 \text{ vs } H_1 : \mu_4 \neq \mu_1.$$

Showing work

$$
\begin{aligned}
\overline{Y}_1 - \overline{Y}_4 &= 32.7 - 27.4 = 5.3 \\
df &= 57 + 60 + 71 + 49 + 56 + 56 - 6 = 343 \\
s_p^2 &= \frac{(57-1)5.1^2 + (60-1)6.7^2 + (71-1)7.8^2 + (49-1)6.1^2 + (56-1)6.7^2 + (56-1)7.0^2}{57+60+71+49+56+56-6} \\
&= \frac{15314}{343} = 44.6 \\
s_p &= \sqrt{s_p^2} = \sqrt{44.6} = 6.7 \\
SE(\overline{Y}_1 - \overline{Y}_4) &= s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_4}} = 6.7\sqrt{\frac{1}{57} + \frac{1}{49}} = 1.3 \\
t &= \frac{\overline{Y}_1 - \overline{Y}_4}{SE(\overline{Y}_1 - \overline{Y}_4)} = \frac{5.3}{1.2} = 4.1 \\
p &= 2P(t_{343} < -|t|) = 2P(t_{343} < -4.1) = 0.000052
\end{aligned}
$$

So we reject the null hypothesis that there is no difference between mean lifetime of mice on the NP and N/N85 diets.

Confidence interval for the difference of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{n_1+n_2+\cdots+n_J-J}(1-\alpha/2) SE(\overline{Y}_1 - \overline{Y}_2)$$

where the t critical value, $t_{n_1+n_2+\cdots+n_J-J}(1-\alpha/2)$, needs to be calculated using a statistical software.

A 95% confidence interval for the difference in mean lifetime for N/N85 minus NP ($\mu_1 - \mu_4$) is

$$5.3 \pm 1.96 \times 1.3 = (2.8, 7.8).$$

The statistical conclusion would be

> *In this study, mice on the N/N85 diet lived an average of 5.3 months longer than mice on the NP diet (95% CI (2.8,7.8)).*

One-way ANOVA F-test

Are any of the means different?

Hypotheses in English:

H_0: all the means are the same

H_1: at least one of the means is different

Statistical hypotheses:

$$H_0: \quad \mu_j = \mu \text{ for all } j \qquad\qquad Y_{ij} \stackrel{iid}{\sim} N(\mu, \sigma^2)$$
$$H_1: \quad \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j' \qquad\qquad Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

An ANOVA table organizes the relevant quantities for this test and computes the pvalue.

ANOVA table

A start of an ANOVA table:

Source of variation	Sum of squares	d.f.	Mean square
Factor A (Between groups)	$SSA = \sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2$	$J-1$	$\frac{SSA}{J-1}$
Error (Within groups)	$SSE = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2$	$n-J$	$\frac{SSE}{n-J}\ (= s_p^2)$
Total	$SST = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}\right)^2$	$n-1$	

where

- J is the number of groups,
- n_j is the number of observations in group j,
- $n = \sum_{j=1}^{J} n_j$ (total observations),
- $\overline{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ (average in group j),
- and $\overline{Y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}$ (overall average).

ANOVA table

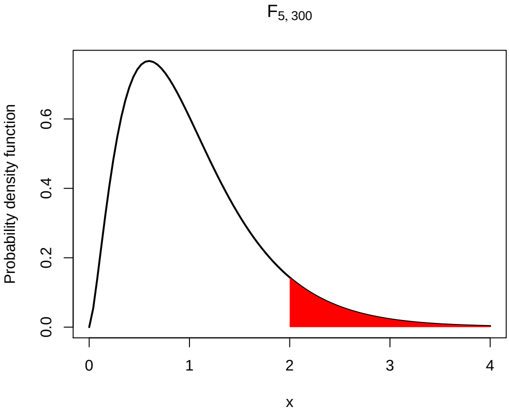
An easier to remember ANOVA table:

Source of variation	Sum of squares	df	Mean square	F-statistic	p-value
Factor A (between groups)	SSA	$J-1$	MSA = SSA/$J-1$	MSA/MSE	(see below)
Error (within groups)	SSE	$n-J$	MSE = SSE/$n-J$		
Total	SST=SSA+SSE	$n-1$			

Under H_0,

- the quantity MSA/MSE has an F-distribution with $J-1$ numerator and $n-J$ denominator degrees of freedom,
- larger values of MSA/MSE indicate evidence against H_0, and
- the p-value is determined by $P(F_{J-1,n-J} > MSA/MSE)$.

F-distribution



One-way ANOVA F-test (by hand)

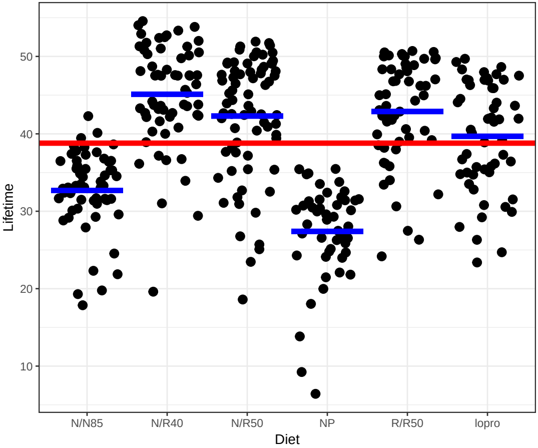
```
# A tibble: 6  4
    Diet     n     mean       sd
  <fctr> <int>    <dbl>    <dbl>
1  N/N85    57 32.69123 5.125297
2  N/R40    60 45.11667 6.703406
3  N/R50    71 42.29718 7.768195
4     NP    49 27.40204 6.133701
5  R/R50    56 42.88571 6.683152
6  lopro    56 39.68571 6.991695
    n     mean
1 349 38.79713
```

So

$$
\begin{aligned}
SSA = & \; 57 \times (32.7-38.8)^2 + 60 \times (45.1-38.8)^2 + 71 \times (42.3-38.8)^2 + 49 \times (27.4-38.8)^2 \\
& +56 \times (42.9-38.8)^2 + 56 \times (39.7-38.8)^2 = 12734 \\
SST = & \; (35.5-38.8)^2 + (35.4-38.8)^2 + (34.9-38.8)^2 + \cdots + (19.6-38.8)^2 + (47.6-38.8)^2 = 28031 \\
SSE = & \; SST - SSA = 28031 - 12734 = 15297 \\
J-1 = & \; 5 \\
n-J = & \; 349 - 6 = 343 \\
n-1 = & \; 348 \\
MSA = & \; SSA/J-1 = 12734/5 = 2547 \\
MSE = & \; SSE/n-J = 15297/343 = 44.6 = s_p^2 \\
F = & \; MSA/MSE = 2547/44.6 = 57.1 \\
p = & \; P(F_{5,343} > 57.1) < 0.0001
\end{aligned}
$$

As a picture

```
Warning:  Ignoring unknown aesthetics:  y
```



R code and output for one-way ANOVA

```
m <- lm(Lifetime~Diet, case0501)
anova(m)


Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value    Pr(>F)
Diet        5  12734  2546.8  57.104 < 2.2e-16 ***
Residuals 343  15297    44.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

General F-tests

The one-way ANOVA F-test is an example of a general hypothesis testing framework that uses F-tests. This framework can be used to test

- composite alternative hypothesesor, equivalently,
- a full vs a reduced model.

The general idea is to balance the amount of variability remaining when moving from the reduced model to the full model measured using the sums of squared errors (SSEs) relative to the amount of complexity, i.e. parameters, added to the model.

Simple vs Composite Hypotheses

Suppose

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

for $j = 1, \ldots, 3$ then a simple hypothesis is

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

and a composite hypothesis is

- $H_0 : \mu_1 = \mu_2 = \mu_3$
- $H_1 : \mu_j \neq \mu_{j'}$ for some j and j'

since there are four possibilities under H_1

- $\mu_1 = \mu_2 \neq \mu_3$
- $\mu_2 = \mu_3 \neq \mu_1$
- $\mu_3 = \mu_1 \neq \mu_2$
- none of μ_1, μ_2, μ_3 are equal

Testing Composite hypotheses

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$ and we want to test the composite hypothesis

- $H_0 : \mu_j = \mu$ for all j
- $H_1 : \mu_j \neq \mu_{j'}$ for some j and j'

think about this as two models:

- $H_0 : Y_{ij} \stackrel{ind}{\sim} N(\mu, \sigma^2)$ (reduced)
- $H_1 : Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ (full)

We can use an F-test to calculate a p-value for tests of this type.

Nested models: full vs reduced

Definition

Two models are nested if the reduced model is a special case of the full model.

For example, consider the full model

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2).$$

One special case of this model occurs when $\mu_j = \mu$ and thus

$$Y_{ij} \stackrel{ind}{\sim} N(\mu, \sigma^2).$$

is a reduced model and these two models are nested.

Calculating the sum of squared residuals (errors)

Model	$Full$	$Reduced$
Assumption	$H_1 : Y_{ij} \overset{ind}{\sim} N\left(\mu_j, \sigma^2\right)$	$H_0 : Y_{ij} \overset{iid}{\sim} N(\mu, \sigma^2)$
Mean	$\hat{\mu}_j = \overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$	$\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} Y_{ij}$
Residual	$r_{ij} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \overline{Y}_j$	$r_{ij} = Y_{ij} - \hat{\mu} = Y_{ij} - \overline{Y}$
SSE	$\sum_{j=1}^{J}\sum_{i=1}^{n_j} r_{ij}^2$	$\sum_{j=1}^{J}\sum_{i=1}^{n_j} r_{ij}^2$

General F-tests

Do the following

1. Calculate

   Extra sum of squares =
   Residual sum of squares (reduced) - Residual sum of squares (full)

2. Calculate

   Extra degrees of freedom =
   # of mean parameters (full) - # of mean parameters (reduced)

3. Calculate F-statistics

$$\text{F} = \frac{\text{Extra sum of squares / Extra degrees of freedom}}{\hat{\sigma}^2_{full}}$$

4. Pvalue is $P(F_{ndf,ddf} > F)$
   - numerator degrees of freedom (nnd) = Extra degrees of freedom
   - denominator degrees of freedom (ddf) = n - # of mean parameters (full)

Example

Recall the mice data set.
Consider the hypothesis that all diets have a common mean lifetime except NP.

Let

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

with $j = 1$ being the NP group then the hypotheses are

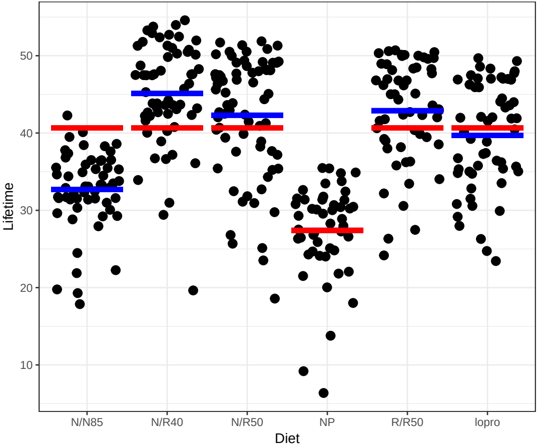
- $H_0 : \mu_j = \mu$ for $j \neq 1$
- $H_1 : \mu_j \neq \mu_{j'}$ for some $j, j' = 2, \ldots, 6$

As models:

- $H_0 : Y_{i1} \sim N(\mu_1, \sigma^2)$ and $Y_{ij} \sim N(\mu, \sigma^2)$ for $j \neq 1$
- $H_1 : Y_{ij} \sim N(\mu_j, \sigma^2)$

As a picture

```
Warning: Ignoring unknown aesthetics: y
Warning: Ignoring unknown aesthetics: y
```



Making R do the calculations

```
case0501$NP = factor(case0501$Diet == "NP")

modR = lm(Lifetime~NP,   case0501)
modF = lm(Lifetime~Diet, case0501)
anova(modR,modF)


Analysis of Variance Table

Model 1: Lifetime ~ NP
Model 2: Lifetime ~ Diet
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    347 20630
2    343 15297  4    5332.2 29.89 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Are there differences in means amongst low calorie diets?

Let Y_{ij} be the lifetime in months for mouse i in group j where the groups are N/N85 (j=1), N/R40 (j=2), N/R50 (j=3), NP (j=4), R/R50 (j=5), and lopro (j=6). Assume

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

and test the hypotheses

H_0: $\mu_2 = \mu_3 = \mu_5 = \mu_6$

H_1: at least one of $\mu_2, \mu_3, \mu_5, \mu_6$ is different from the rest

Implicitly, we are allowing μ_1 and μ_4 to be different from the others.

Making R do the calculations

```
case0501$local = ifelse(case0501$Diet=='N/N85', 1, 2) # NP is 2 here
case0501$local[case0501$Diet=='NP'] = 0               # now NP is 1
case0501$local = factor(case0501$local)
mod1 = lm(Lifetime~1,     case0501)
modR = lm(Lifetime~local, case0501)
modF = lm(Lifetime~Diet,  case0501)
anova(mod1, modR, modF)


Analysis of Variance Table

Model 1: Lifetime ~ 1
Model 2: Lifetime ~ local
Model 3: Lifetime ~ Diet
  Res.Df   RSS Df Sum of Sq        F    Pr(>F)
1    348 28031
2    346 16163  2   11868.5 133.0585 < 2.2e-16 ***
3    343 15297  3     865.4   6.4682 0.0002863 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


anova(modF) # To get the pooled estimate of the variance for the full model


Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value    Pr(>F)
Diet        5  12734  2546.8  57.104 < 2.2e-16 ***
Residuals 343  15297    44.6
---
```

Lack-of-fit F-test

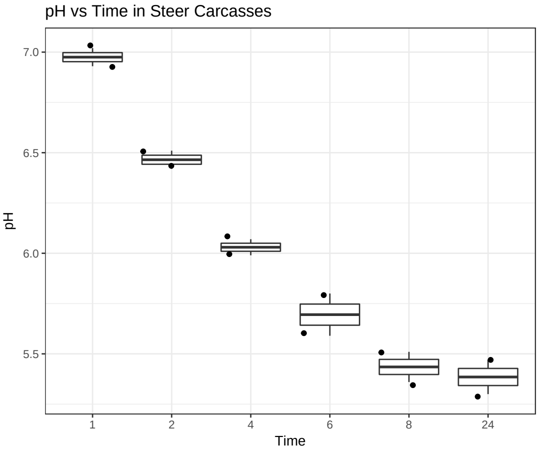
Let Y_{ij} be the i^{th} observation from the j^{th} group where the group is defined by those observations having the same explanatory variable value (X_j).

Two models:

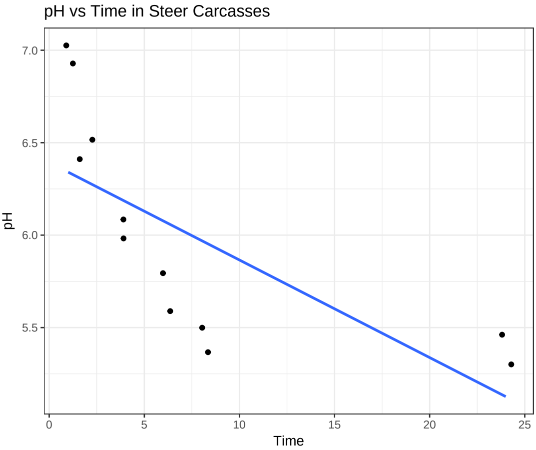
$$\begin{array}{lll} \text{ANOVA:} & Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2) & \text{(full)} \\ \text{Regression:} & Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2) & \text{(reduced)} \end{array}$$

- Regression model is reduced:
  - ANOVA has J parameters for the mean
  - Regression has 2 parameters for the mean
  - Set $\mu_j = \beta_0 + \beta_1 X_j$.
- Small pvalues indicate a lack-of-fit, i.e. the reduced model is not adequate.
- Lack-of-fit F-test requires multiple observations at a few X_j values!

pH vs Time - ANOVA



pH vs Time - Regression



```
# Use as.factor to turn a continuous variable into a categorical variable
m_anova = lm(pH ~ as.factor(Time), Sleuth3::ex0816)
m_reg   = lm(pH ~           Time , Sleuth3::ex0816)
anova(m_reg, m_anova)


Analysis of Variance Table

Model 1: pH ~ Time
Model 2: pH ~ as.factor(Time)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     10 1.97289
2      6 0.05905  4    1.9138 48.616 0.0001048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

- Use t-tests for simple hypothesis tests and CIs
- Use F-tests for composite hypothesis tests
    - One-way ANOVA F-test
    - General F-tests
    - Lack-of-fit F-tests

Think about F-tests as comparing models.