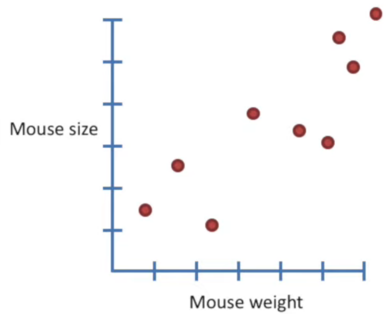
StatQuest: General Linear Models
Part 2:
t-tests and ANOVA

Implement the two-sample t-test and ANOVA with linear model method

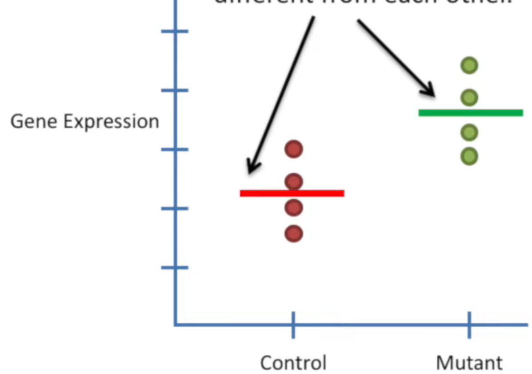


We measured **mouse weight** and **mouse size** and we wanted learn 2 things from it:

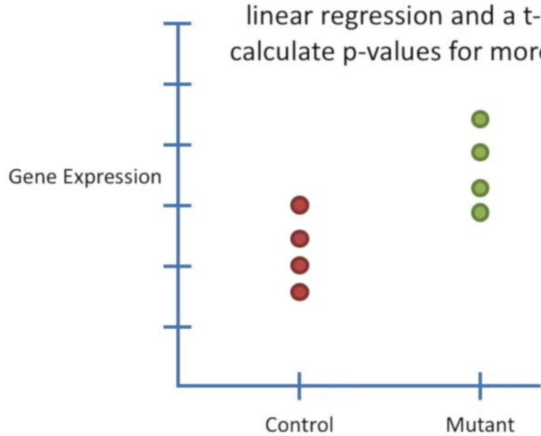
1) How useful was mouse weight for predicting mouse size? (***R***^{2} told us this).

2) Was that relationship due to chance? (The ***p*-value** told us this.)

The goal of a t-test is to compare means and see if they are significantly different from each other.

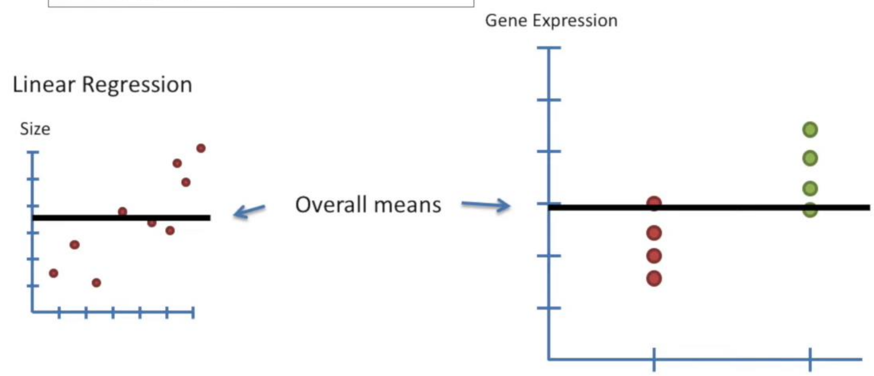


If the same method can calculate *p*-values for a linear regression and a t-test, then we can easily calculate p-values for more complicated situations.



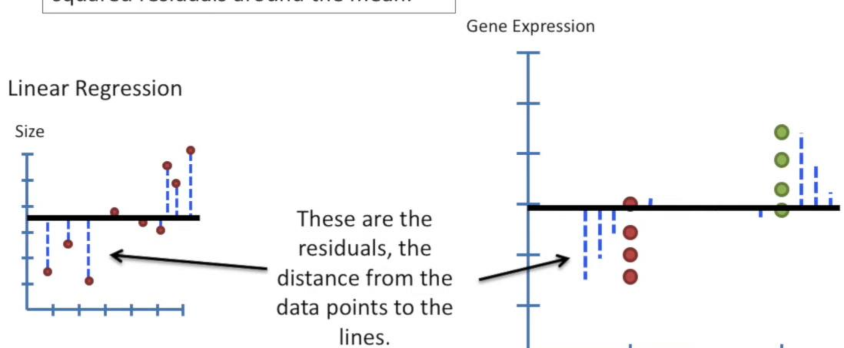
Step 1: Ignore the x-axis and find the overall mean.

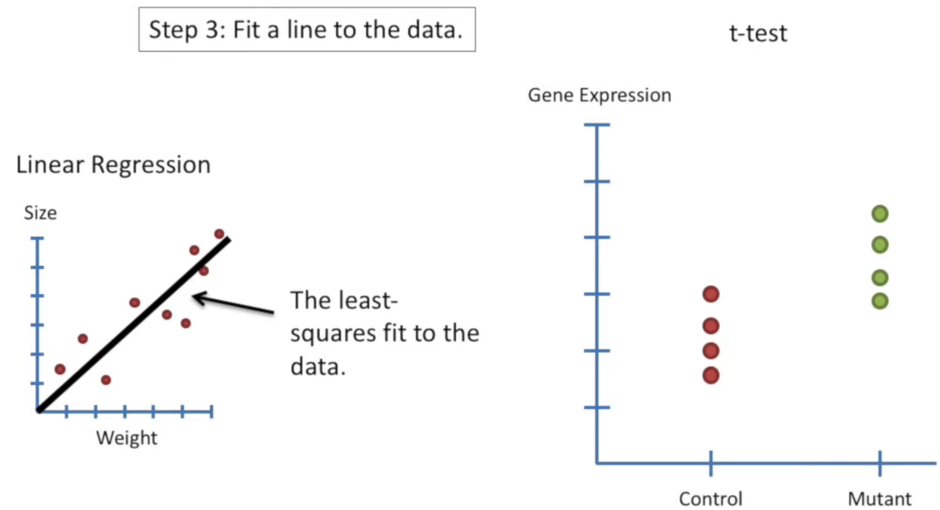
t-test

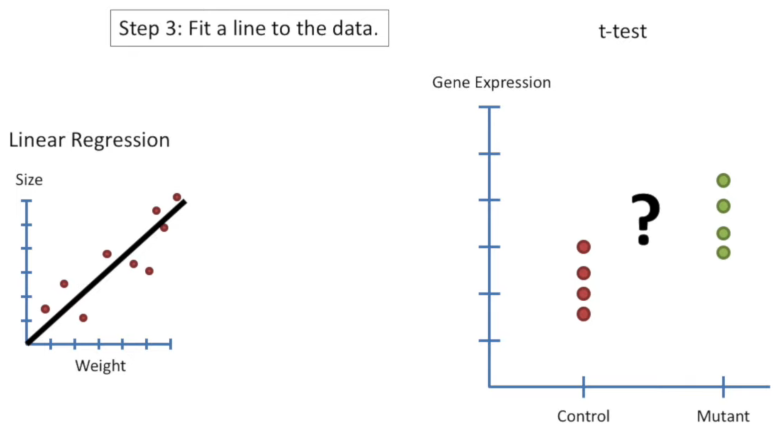


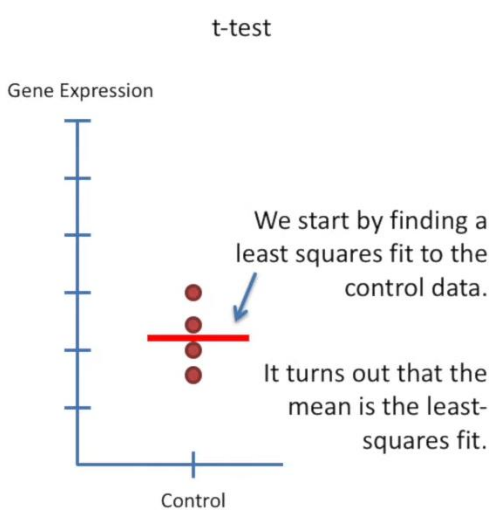
Step 2: Calculate SS(mean), the sum of squared residuals around the mean.

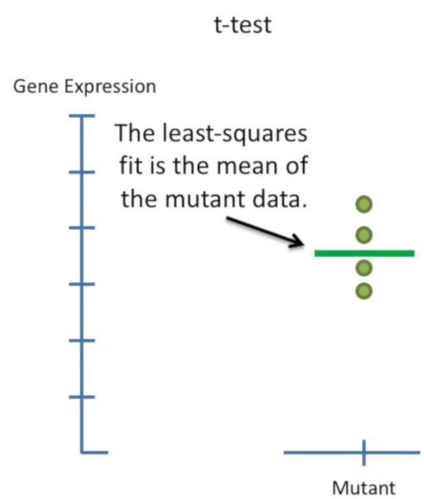
t-test




Step 3: Fit a line to the data.
t-test
Gene Expression
Linear Regression
Size
The least-squares fit to the data.
Weight
Control
Mutant


Step 3: Fit a line to the data.
t-test
Gene Expression
Linear Regression
Size
?
Weight
Control
Mutant

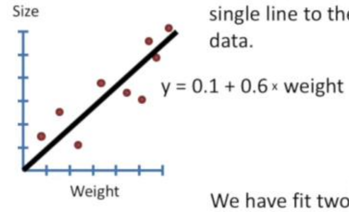

t-test
Gene Expression
We start by finding a least squares fit to the control data.
It turns out that the mean is the least-squares fit.
Control


t-test
Gene Expression
The least-squares fit is the mean of the mutant data.
Mutant

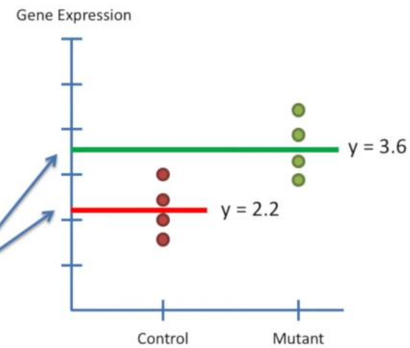
Step 3: Fit a line to the data.

t-test

Linear Regression



Originally we fit a single line to the data.



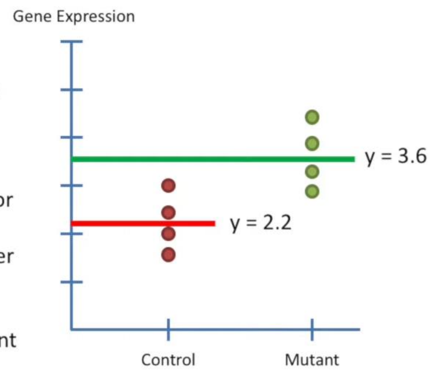
We have fit two lines to the data.

t-test

However, there is a way to combine these two lines into a single equation.

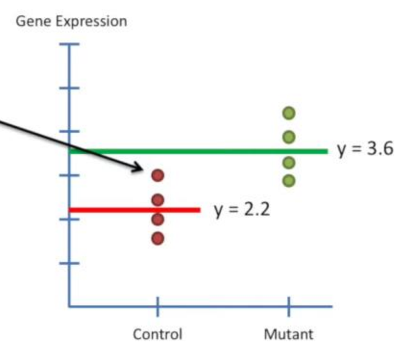
This will make the steps for computing "*F*" the exact same for the regression and the t-test, which, in turn, means a computer can do it automatically.

This is key, because we don't want to do this by hand, ever….



$y = 1 \times \textcolor{red}{\mathbf{2.2}} + 0 \times \textcolor{green}{\mathbf{3.6}} + \text{the residual}$

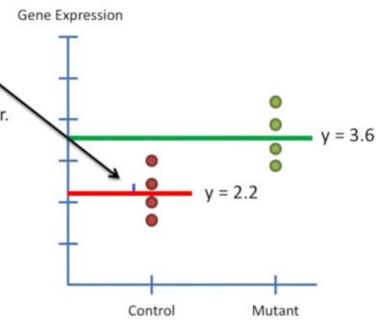
This is the equation (which combines both lines) for this point



y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

The only difference is the residual. This one is smaller.

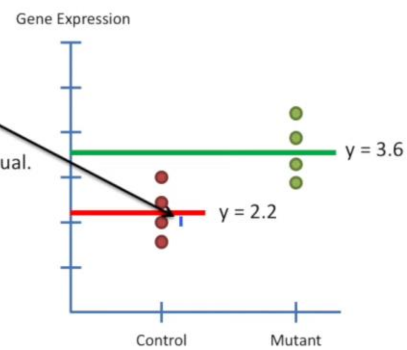


y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

Again, the only difference is the residual.



y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

y = 1×**2.2** + 0×**3.6** + the residual

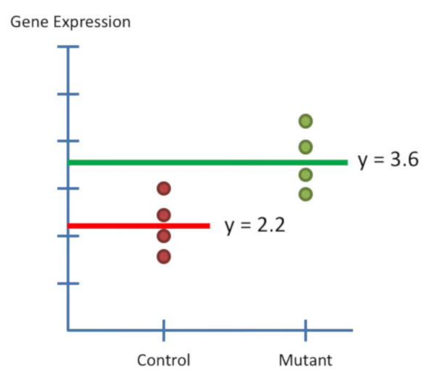
y = 0×**2.2** + 1×**3.6** + the residual

y = 0×**2.2** + 1×**3.6** + the residual

y = 0×**2.2** + 1×**3.6** + the residual

y = 0×**2.2** + 1×**3.6** + the residual

They function like on/off switches for the two means.



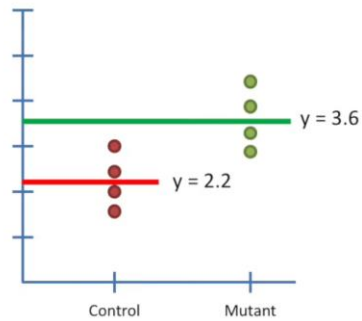
$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

When we isolate the 1s and 0s, they form a matrix called a “design matrix”.

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The design matrix can be combined with an abstract version of the equation to represent a “fit” to the data.

y = column1 × **2.2** + column2 × **3.6**



y = column1 × **2.2** + column2 × **3.6**

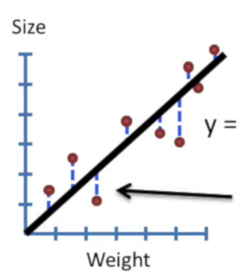
In practice, the role of each column is assumed, and the equation is written out like this:

$y = \textbf{mean}_{\textbf{control}} + \textbf{mean}_{\textbf{mutant}}$

Now that we have the “fit” for the control and mutant data down to a single equation (plus design matrix). We can move on to calculating *F* and the p-value.

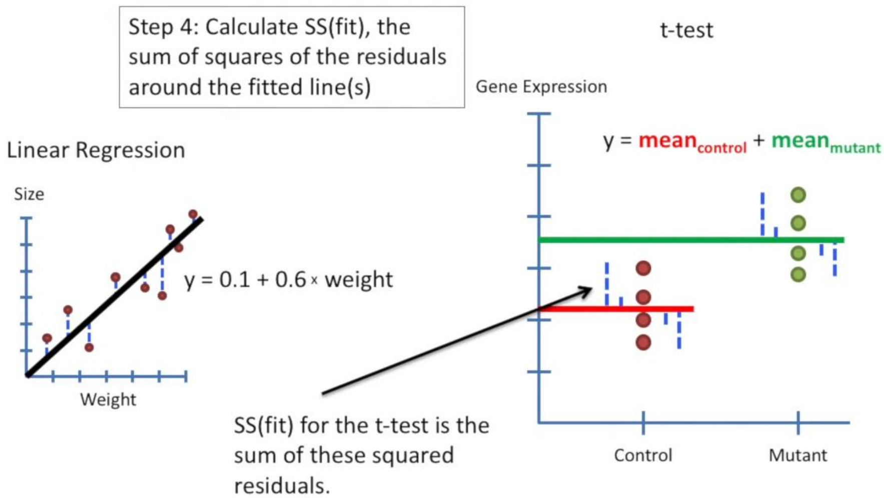
Step 4: Calculate SS(fit), the sum of squares of the residuals around the fitted line(s)

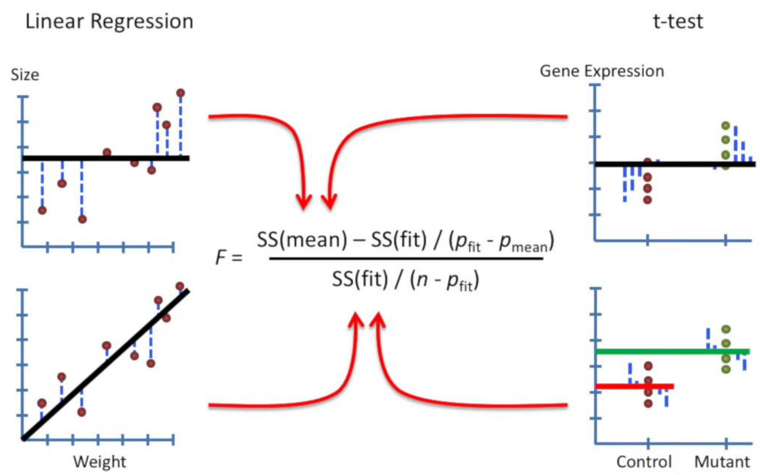
Linear Regression



y = 0.1 + 0.6 × weight

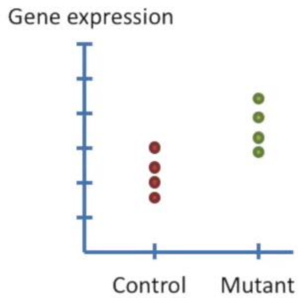
SS(fit) is the sum of these squared residuals.



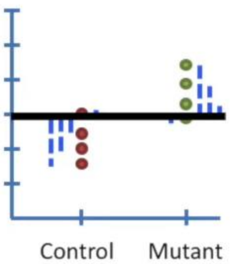


Summary and ANOVA

The original data.



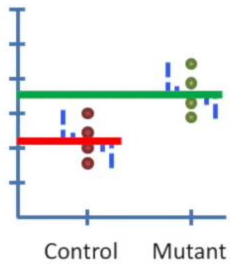
Calculate SS(mean)



y = overall mean

$p_{\text{mean}} = 1$

Calculate SS(fit)



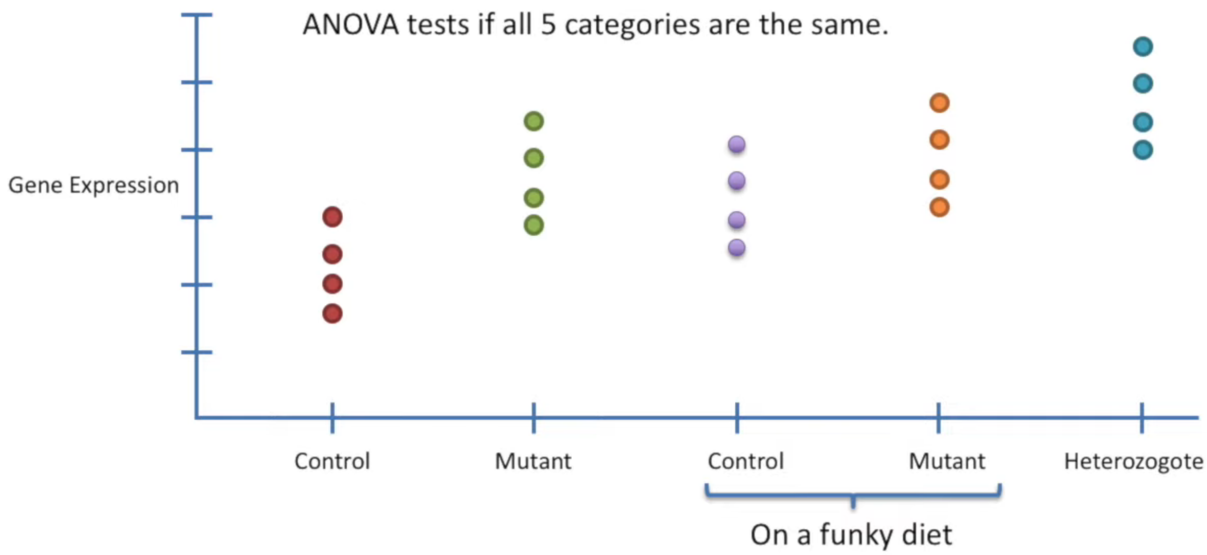
$y = \textbf{mean}_{\textbf{control}} + \textbf{mean}_{\textbf{mutant}}$

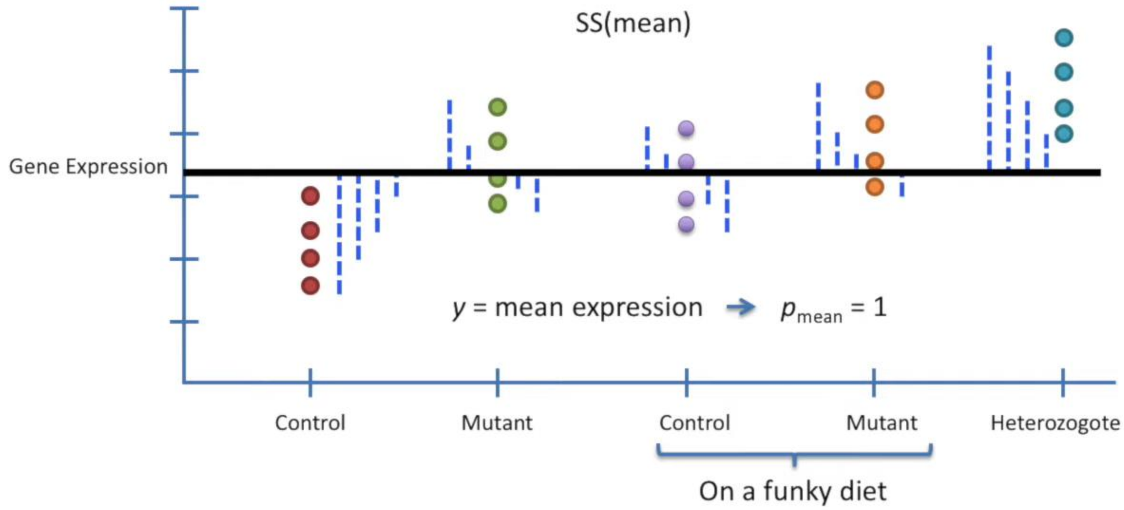
$$F = \frac{\text{SS(mean)} - \text{SS(fit)} / (p_{\text{fit}} - p_{\text{mean}})}{\text{SS(fit)} / (n - p_{\text{fit}})}$$

$p_{\text{fit}} = 2$

Now let's do an ANOVA!!

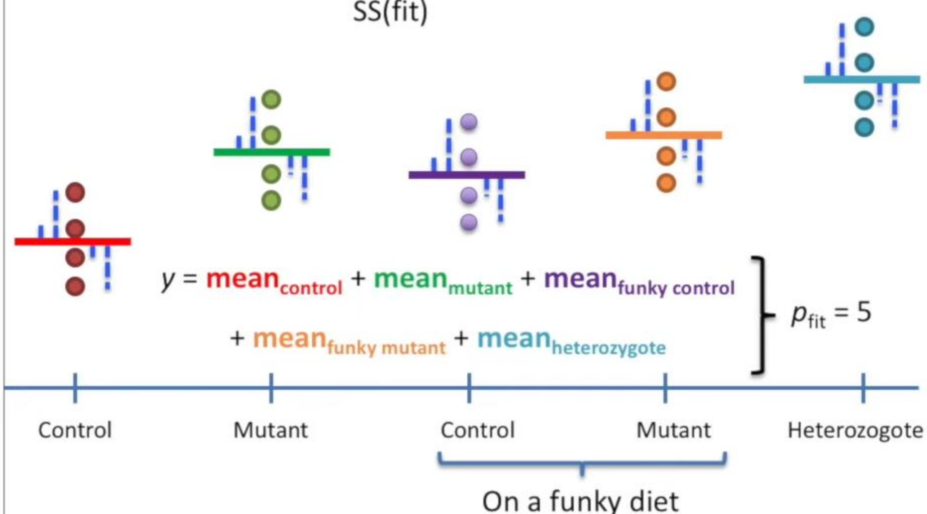
ANOVA tests if all 5 categories are the same.





1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

Here's what the design matrix looks like - one column per category.



$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$