Unit 4: Regression and Prediction

# 6. Multiple regression and ANOVA

(Chapter 6.3)

12/5/2018

# Recap from last time

1. For many real-world problems, lots of variables contribute a little bit
2. All of these variables affect all of the other variables in your regression model
3. Stepwise approaches try to correct for this, but there is no “one true way”

# Key ideas

1. We can check assumptions for multiple regression using plots
2. Inference for multiple regression relies on estimating how much variance is accounted for by each variable (like ANOVA)
3. Multiple Regressions and ANOVAs are two formulations of the same idea

# Which explanatory variables do not look like reliable predictors?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.6282 | 0.1720 | 26.90 | 0.00 |
| beauty | 0.1080 | 0.0329 | 3.28 | 0.00 |
| gender.male | 0.2040 | 0.0528 | 3.87 | 0.00 |
| age | -0.0089 | 0.0032 | -2.75 | 0.01 |
| formal.yes [1] | 0.1511 | 0.0749 | 2.02 | 0.04 |
| lower.yes [2] | 0.0582 | 0.0553 | 1.05 | 0.29 |
| native.non english | -0.2158 | 0.1147 | -1.88 | 0.06 |
| minority.yes | -0.0707 | 0.0763 | -0.93 | 0.35 |
| students [3] | -0.0004 | 0.0004 | -1.03 | 0.30 |
| tenure.tenure track [4] | -0.1933 | 0.0847 | -2.28 | 0.02 |
| tenure.tenured | -0.1574 | 0.0656 | -2.40 | 0.02 |

# Two approaches to model selection

1. **Forward-selection with some criterion (e.g. p-value or $R^2$ *adjusted)***:
   a. Start with regressions of response vs. each explanatory variable
   b. Pick the model with the highest $R^2_{adj}$
   c. Add the remaining variables one at a time to the existing model, and once again pick the model with the highest $R^2_{adj}$
   d. Repeat until no remaining variables increase $R^2_{adj}$

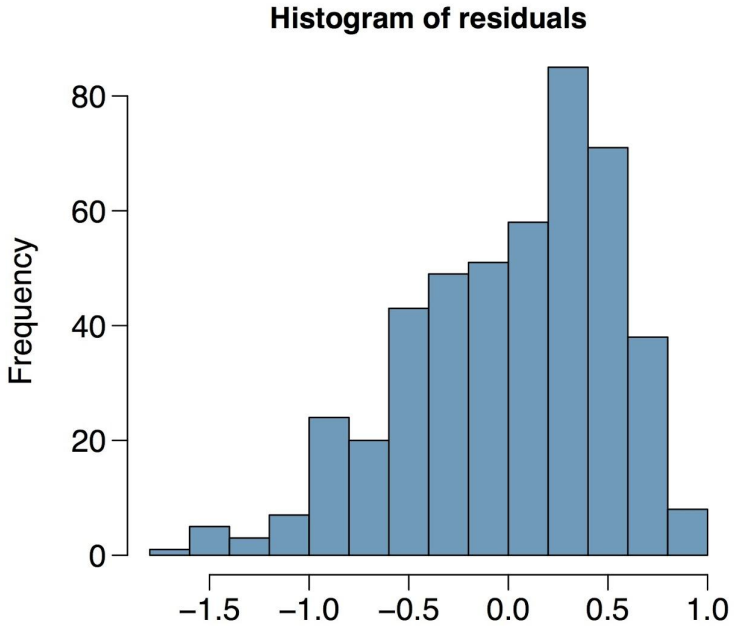
2. **Backward-selection with some criterion (e.g. p-value or $R^2$ *adjusted)***:
   a. Start with the full model
   b. Drop one variable at a time and record $R^2_{adj}$ of each smaller model
   c. Pick the model with the highest increase in $R^2_{adj}$
   d. Repeat until none of the models yield an increase in $R^2_{adj}$

# Conditions for using multiple regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

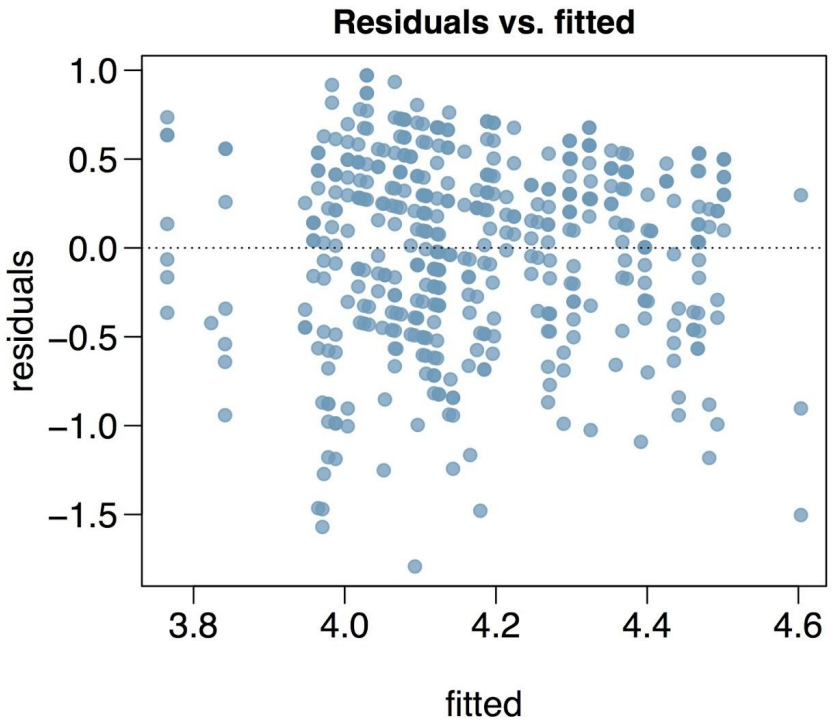
1. Residuals are nearly normal (primary concern is outliers)
2. Residuals have constant variability
3. Residuals are independent
4. Each variable is linearly related to the outcome

# Nearly normal residuals?




**Does the normal residuals condition appear to be satisfied?**

# Constant variability?




**Does constant variability appear to be satisfied?**

# Checking constant variance recap

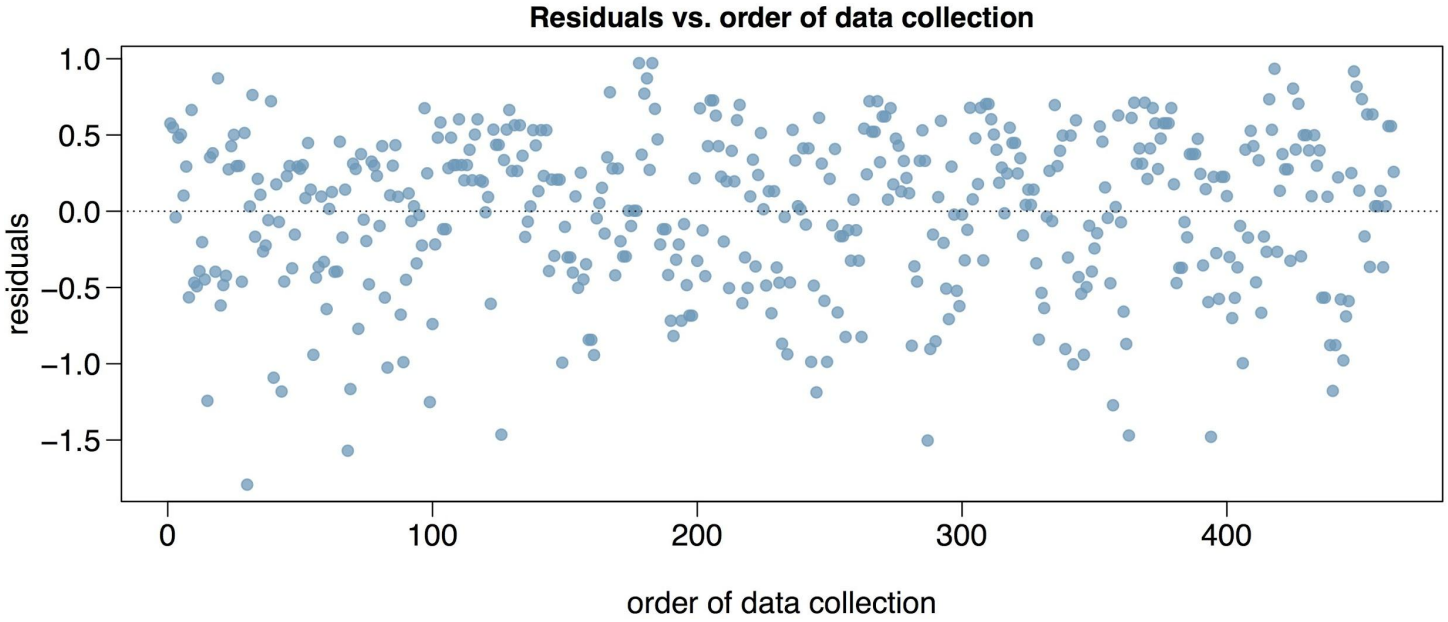
**Simple linear regression:** We check constant variability by plotting residuals vs. *x*

**Multiple linear regression:** We check constant variability by plotting residuals vs. fitted value.

**Why are we using different plots?**

In Multiple linear regression, there are multiple explanatory variables, so a plot of residuals vs. one of them wouldn't tell us about the whole model

# Independent residuals?




**Does it look like we have independent residuals?**

# Checking independent residuals

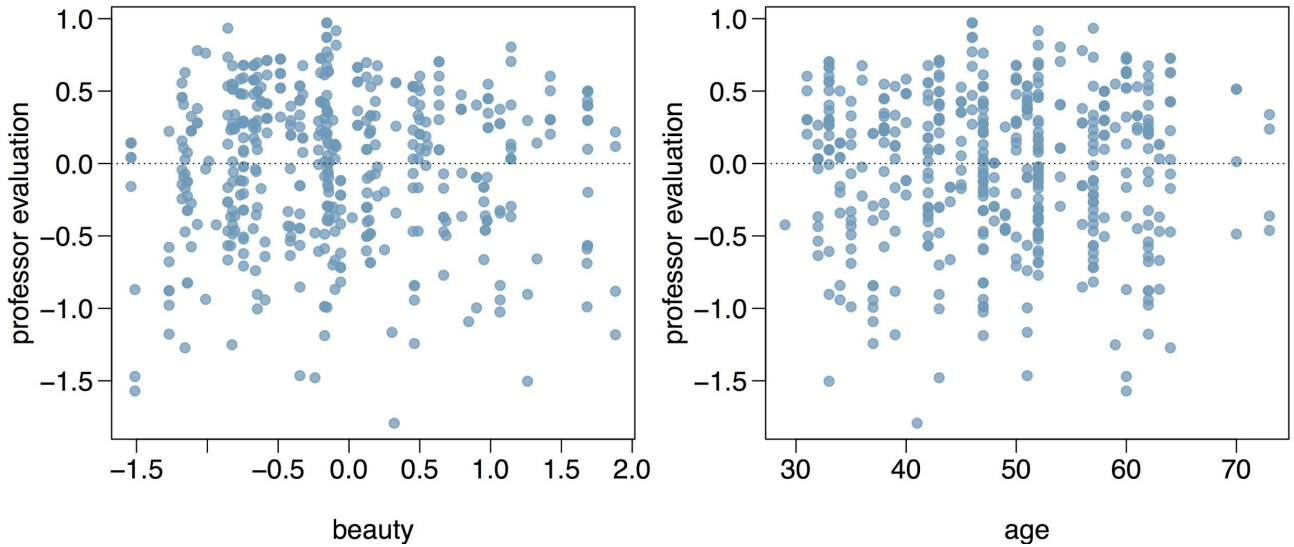
Checking independent residuals lets us indirectly check for independent observations.

If residuals (and observations) are independent, we should expect to see no trend in the scatterplot vs. order of data collection

When this condition is violated, we need to use a different method to correct for these dependencies
(e.g. something called Auto-regressive Moving Average or ARMA)

# Linearity?

Residuals vs. each (numerical) explanatory variable




**Does it look like these predictors and the evaluation are linearly related?**

(Note: One virtue of using residuals instead of the predictors on the y-axis:
We can still check for linearity without worrying about other possible violations like collinearity between the predictors.)

# Comparing means for more than two variables

Let's take a closer look at our tenure-track variables. It has three values:

- Non tenure-track, Tenure Track, Tenured

We could treat these variables as being an interval scale: e.g.

- 0 = Non tenure-track, 1 = Tenure Track, 2 = Tenured

**But should we?**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| ... | | | | |
| tenure.tenure track | -0.1933 | 0.0847 | -2.28 | 0.02 |
| tenure.tenured | -0.1574 | 0.0656 | -2.40 | 0.02 |

# Treating tenure as categorical

To compare means of 2 groups we use a **Z** or a **T**-statistic.
We could use these to ask all of the pairwise questions (e.g. tenured different from non-tenure track, tenured different from tenure-track, etc.)

Or, we could compare all of them at the same time.
To compare means of 3+ groups we use a new test called **ANOVA**
and a new statistic called **F**.

$\mathbf{H_0}$: The mean outcome is the same across all categories,

$\mu_1 = \mu_2 = ... = \mu_k$, where $\mu_i$ represents the mean

$\mathbf{H_A}$: At least one mean is different than others.

# z/t vs. ANOVA - Method

**z/t test**

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

**ANOVA**

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

Large test statistics lead to small p-values.

If the p-value is small enough $H_0$ is rejected, we conclude that the population means are not equal.

# Practice Question: Hypotheses

What are the correct hypotheses for testing for a difference between the mean professor evaluation among the three tenure levels?

a) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$
$H_A: \mu_N \neq \mu_{Tr} \neq \mu_{Te}$

b) $H_0: \mu_N \neq \mu_{Tr} \neq \mu_{Te}$
$H_A: \mu_N = \mu_{Tr} = \mu_{Te}$

c) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$
$H_A$: At least one mean is different.

d) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$
$H_A: \mu_N > \mu_{Tr} > \mu_{Te}$

# Practice Question: Hypotheses

What are the correct hypotheses for testing for a difference between the mean professor evaluation among the three tenure levels?

a) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$
$H_A: \mu_N \neq \mu_{Tr} \neq \mu_{Te}$

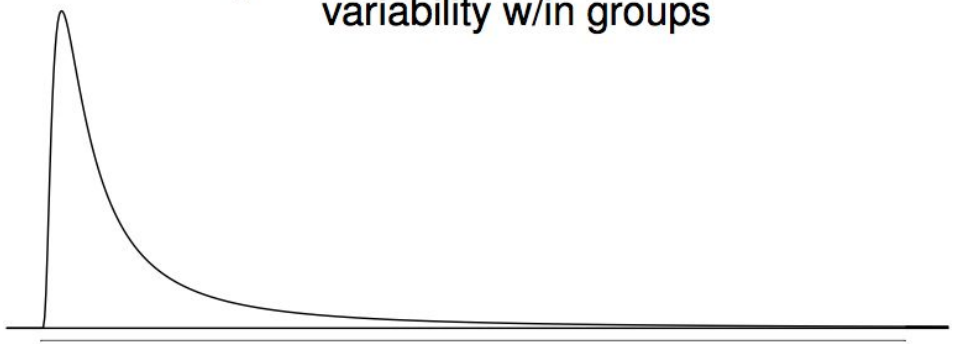
b) $H_0: \mu_N \neq \mu_{Tr} \neq \mu_{Te}$
$H_A: \mu_N = \mu_{Tr} = \mu_{Te}$

**c) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$**
**$H_A$: At least one mean is different.**

d) $H_0: \mu_N = \mu_{Tr} = \mu_{Te}$
$H_A: \mu_N > \mu_{Tr} > \mu_{Te}$

# F-distribution and p-values

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



In order to be able to reject $H_0$, we need a small p-value, which requires a large F-statistic.

In order to obtain a large F-statistic, variability between sample means needs to be greater than variability within sample means.

# ANOVA in R

```
> tenure_anova <- aov(profevaluation ~ tenure, data = prof_data)
> summary(tenure_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

# ANOVA output: Degrees of Freedom

```
> summary(tenure_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**Degrees of freedom associated with ANOVA**

- Groups: $df_G = k - 1$, where k is the number of groups
- Total: $df_T = n - 1$, where n is the total sample size
- Error: $df_E = df_T - df_G$

- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 463 - 1 = 462$
- $df_E = 462 - 2 = 460$

# ANOVA output: Sum of Squares

```
> summary(tenure_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**Sum of Squares between groups (SSG)**

measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where $n_i$ is each group size, $\bar{x}_i$ is the average for each group,

$\bar{x}$ is the overall (grand) mean.

**SSG =** $102 \times (4.284 - 4.174)^2 +$

$108 \times (4.155 - 4.174)^2 +$

$253 \times (4.139 - 4.174)^2 = 1.59$

| | mean | n |
|---|---|---|
| non-tenure | 4.284 | 102 |
| tenure track | 4.155 | 108 |
| tenured | 4.139 | 253 |
| overall | 4.174 | 463 |

# ANOVA output: Sum of Squares

```
> summary(tenure_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**Sum of Squares between groups (SST)**

measures the variability across all observations

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

**SST =** $(4.7 - 4.174)^2 + (4.6 - 4.174)^2 + (4.1 - 4.174)^2 + (4.5 - 4.174)^2 + \ldots$

**Sum of Squares error (SSE)**

measures the variability within groups

$$SSE = SST - SSG$$

# ANOVA output: Mean Square Error

```
> summary(tenure_anova)

             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**Mean Square Error (MSE)**

Calculated as sum of squares divided by the degrees of freedom.

$MSG = SSG / DF_g = 1.59/2 = .7946$

$MSE = SSE / DF_E = 135.07/460 = .2936$

# ANOVA output: F-value

```
> summary(tenure_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**Test statistic - F**

The ratio between within group variability and between group variability

$$F = \frac{MSG}{MSE}$$

$$F = \frac{.7946}{.2936} = 2.706$$

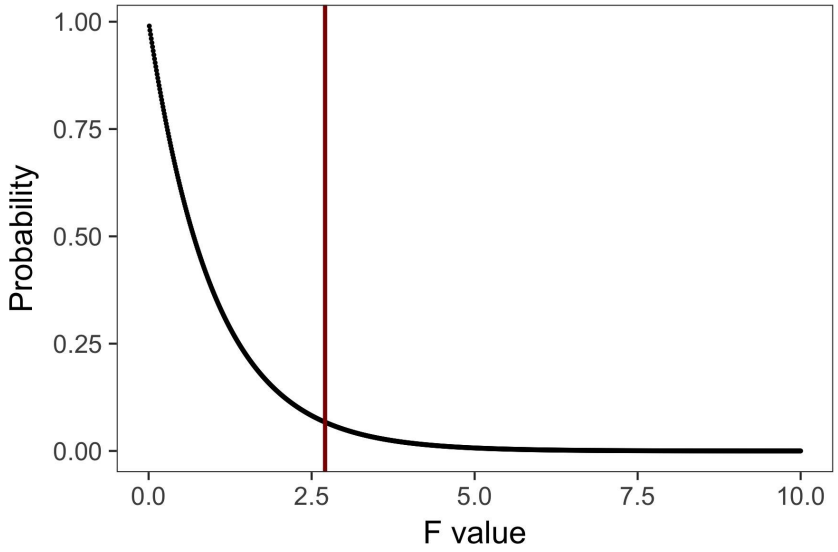
# ANOVA output: p-value

```
> summary(tenure_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

**p-value**

probability of at least as large a ratio between the "between group" and "within group" variability, if the means of all groups are equal.

It's calculated as the area under the F-curve, with degrees of freedom $df_G$ and $df_E$, above the observed F-statistic.



# Why the different values?

```
> summary(tenure_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
tenure        2   1.59  0.7946   2.706 0.0679 .
Residuals   460 135.07  0.2936
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| ... | | | | |
| tenure.tenure track | -0.1933 | 0.0847 | -2.28 | 0.02 |
| tenure.tenured | -0.1574 | 0.0656 | -2.40 | 0.02 |

**Two things:**

1. The regression model is doing pairwise comparison
2. The regression model includes other variables.
   Tenure is fit against the residuals from those variables

# Another look at $R^2$

$R^2$ can be calculated in two ways:

1. Squaring the correlation coefficient of standardized *x* and *y* (*R*)
2. Based on the definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

This lets us use **ANOVA** to calculate the explained variability and total variability

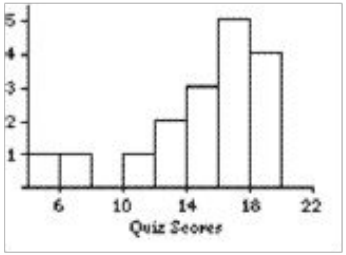
# Key ideas

1. We can check assumptions for multiple regression using plots
2. Inference for multiple regression relies on estimating how much variance is accounted for by each variable (like ANOVA)
3. Multiple Regressions and ANOVAs are two formulations of the same idea

# Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) Test


https://apps3.cehd.umn.edu/artist/caos.html

e.g. For this graphical display of Quiz Scores, which estimates of the mean and median are most plausible?




a. median = 13.0 and mean = 12.0

b. median = 14.0 and mean = 15.0

**C.** median = 16.0 and mean = 14.3

d. median = 16.5 and mean = 16.2

You will take a CAOS Pre and Post Test. *These will be graded for completion, not correctness*.

# Statistics and the absurd

"Man stands face to face with the irrational.
He feels within him his longing for happiness and for reason.
The absurd is born of this confrontation between the human need and the unreasonable silence of the world."

Albert Camus, The Myth of Sisyphus

To understand statistics is to embrace the absurd: *There is no certainty, only degrees of doubt*

# Statistics connect scientific theories to the world

The artifacts of science are models

*All models are wrong, but some are useful*

George Box

Because there is no certainty, no model can be *True*.

Statistics is a set of tools for helping us to figure which ones are more useful.

# Statistics are an expression of liberty

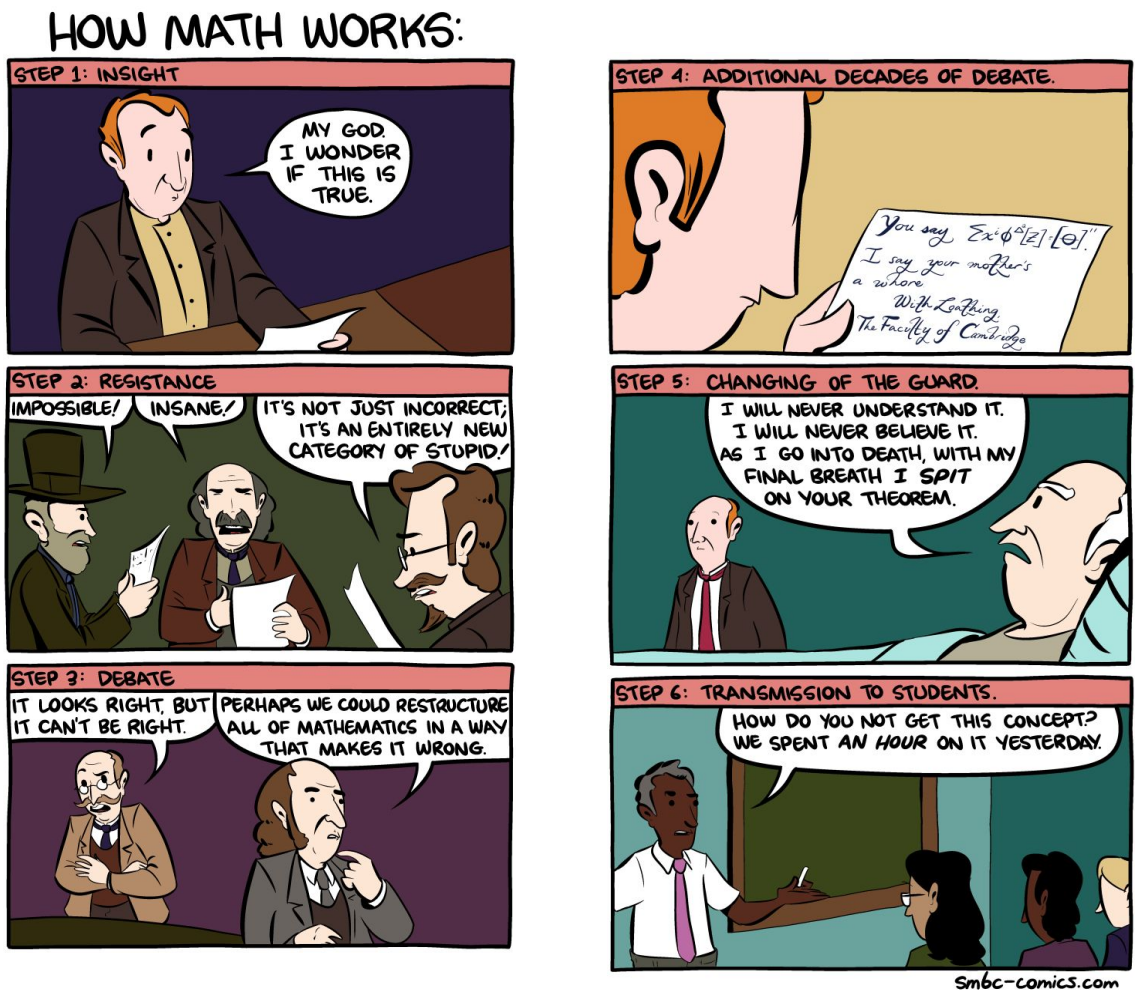
The fundamental premise of inferential statistics: You could be wrong!

The practice of statistics is *doubt* of authority

**Ubi dubium ibi libertas**

Thanks to John Kruschke

# The Curse of Knowledge




- These ideas are challenging
- If you don't understand them right away, don't worry!
- They took centuries to develop