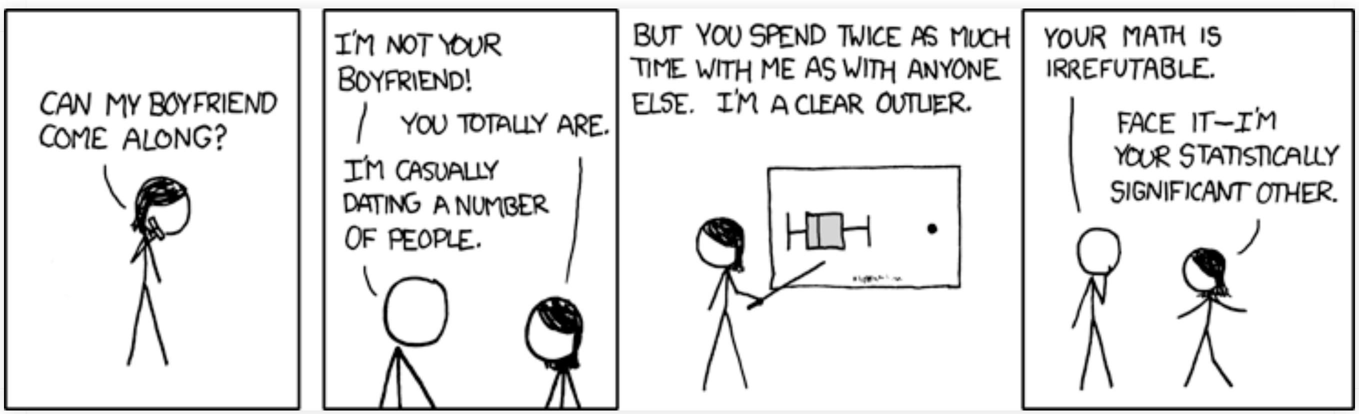
Errors in Hypothesis Testing; Linear Regression

M. Drew LaMar

September 07, 2016



https://xkcd.com/539/

Introduction to Quantitative Biology, Fall 2016

Class announcements

- Upcoming Booz Allen Events (http://www.boozallen.com/)
  - **Diversity Brunch**: September 9, 10 AM - 11:30 AM in the Cohen Career Center
  - **Fall Career & Internship Fair**: September 9, 12 PM - 4 PM in the Sadler Center
  - **Meet the Firms Friday**: September 16, 11 AM - 2 PM in Miller Hall
  - **Corporate Presentation/Info Session**: September 27, 6 PM - 8 PM in Miller Hall

Interval estimates

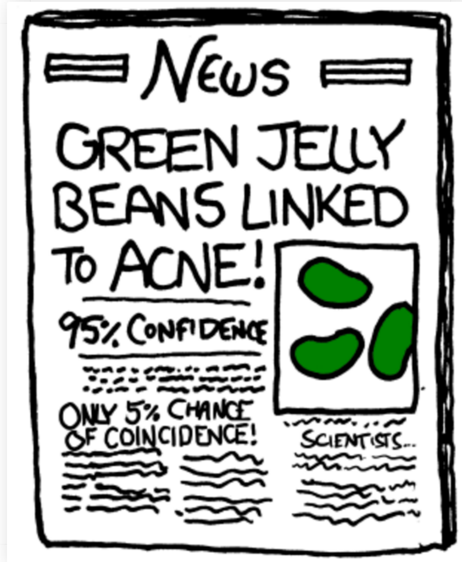
http://www.smbc-comics.com/comic/2011-11-23

Hypothesis testing

So, P = 0.04. Is that good?

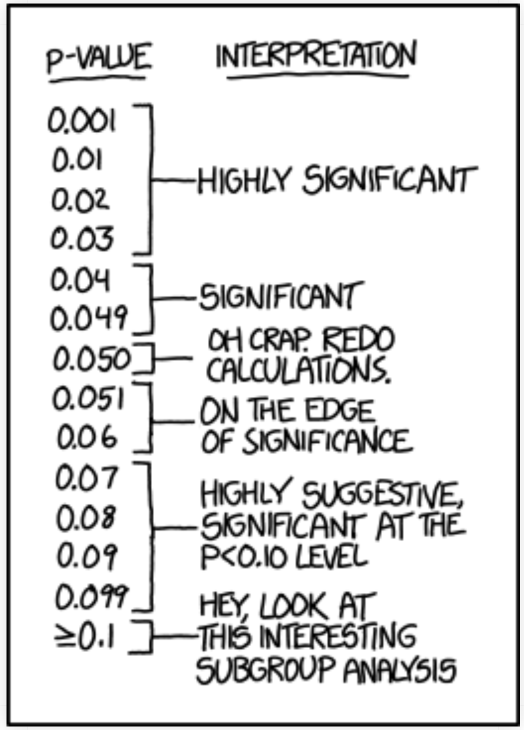
https://youtu.be/7jSE3JANx14?t=4m29s

Jelly Beans



https://xkcd.com/882/

P-Values



https://xkcd.com/1478/

Hypothesis Testing

https://xkcd.com/892/

Errors in Hypothesis Testing

	Reality	
Conclusion	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

Definition: ***Type I error*** is rejecting a true null hypothesis. The probability of a Type I error is given by

$$\Pr[\text{Reject } H_0 \mid H_0 \text{ is true}] = \alpha$$

Definition: ***Type II error*** is failing to reject a false null hypothesis. The probability of a Type II error is given by

$$\Pr[\text{Do not reject } H_0 \mid H_0 \text{ is false}] = \beta$$

Errors in Hypothesis Testing - Power

	Reality	
Conclusion	H_0 true	H_0 false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

Definition: The ***power*** of a statistical test (denoted $1 - \beta$) is given by

$$\begin{aligned} \Pr[\text{Reject } H_0 \mid H_0 \text{ is false}] &= 1 - \beta \\ &= 1 - \Pr[\text{Type II error}] \end{aligned}$$

Probability of errors in hypothesis testing

	Reality	
Conclusion	H_0 true	H_0 false
Reject H_0	α	$1-\beta$
Do not reject H_0	$1-\alpha$	β

- α is the *significance level*
- $1-\beta$ is the *power*

Statistical power example
https://qubeshub.org/tools/statpowerviz/

Power analysis

Power of a statistical test is a function of
- Significance level α
- Variability of data
- Sample size
- Effect size

- Desired power is set by researcher (typically 80%)
- Significance level set by researcher
- Data variability and effect size can be estimated by previous studies or pilot studies
- Sample size is then calculated to achieve desired power given previous fixed attributes

Regression

Definition: ***Regression*** is the method used to predict values of one numerical variable (response) from values of another (explanatory).

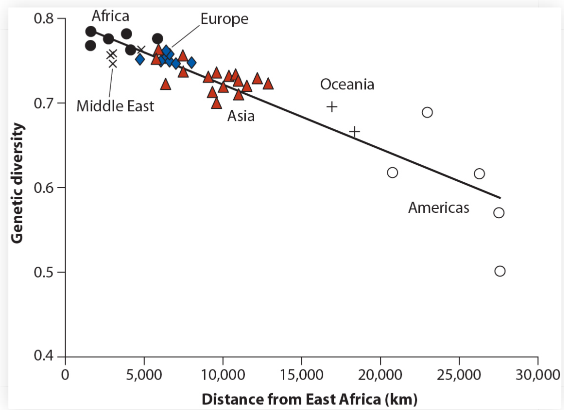
Note: Regression can be done on data from an observational or experimental study.

We will discuss 3 types:

- Linear regression
- Nonlinear regression
- Logistic regression

Linear regression

Definition: ***Linear regression*** draws a straight line through the data to predict the response variable from the explanatory variable.



Slope determines *rate of change* of response with explanatory - humans lose 0.076 units of genetic diversity with every 10,000 km from East Africa.

Formula for the line

Definition: For the ***population***, the regression line is

$$Y = \alpha + \beta X,$$

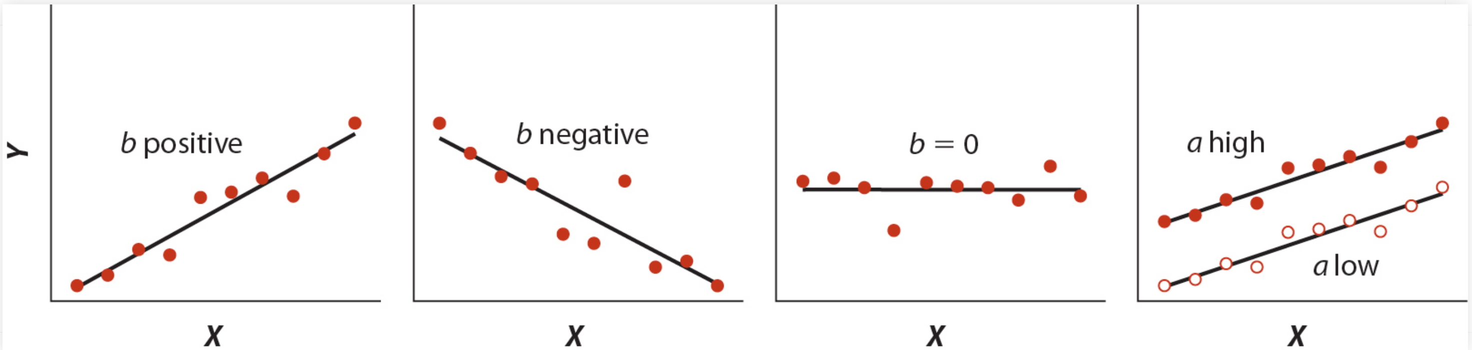
where α (the ***intercept***) and β (the ***slope***) are population parameters.

Definition: For a ***sample***, the regression line is

$$Y = a + bX,$$

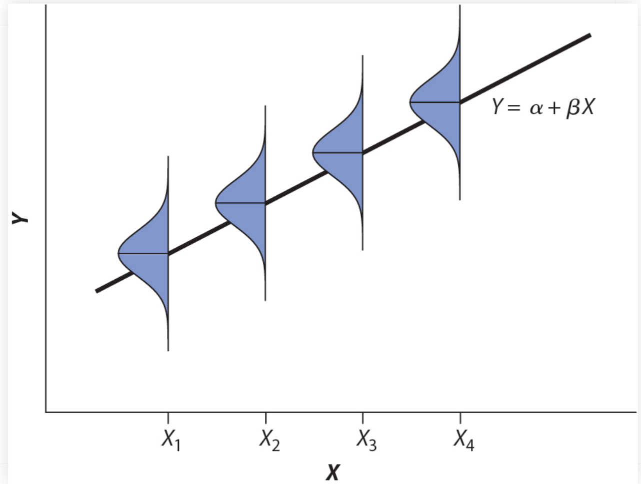
where a and b are estimates of α and β, respectively.

Graph of the line



- a: intercept
- b: slope

Assumptions of linear regression



Note: At each value of X, there is a population of Y-values whose *mean* lies on the true regression line (this is the *linear* assumption).

Assumptions of linear regression

- Linearity
- Residuals are normally distributed
- Constant variance of residuals
- Independent observations