# Crafting with Data

## Reality, Illusions, Truth & the Future

---

Instructor: Rob Faludi

# Plan for Today

- Errors
- ANOVA
- Overview of statistical packages
- Readings & Assignments

# Some Errors

# Kinds of Error

- How can the smoke detector be wrong?

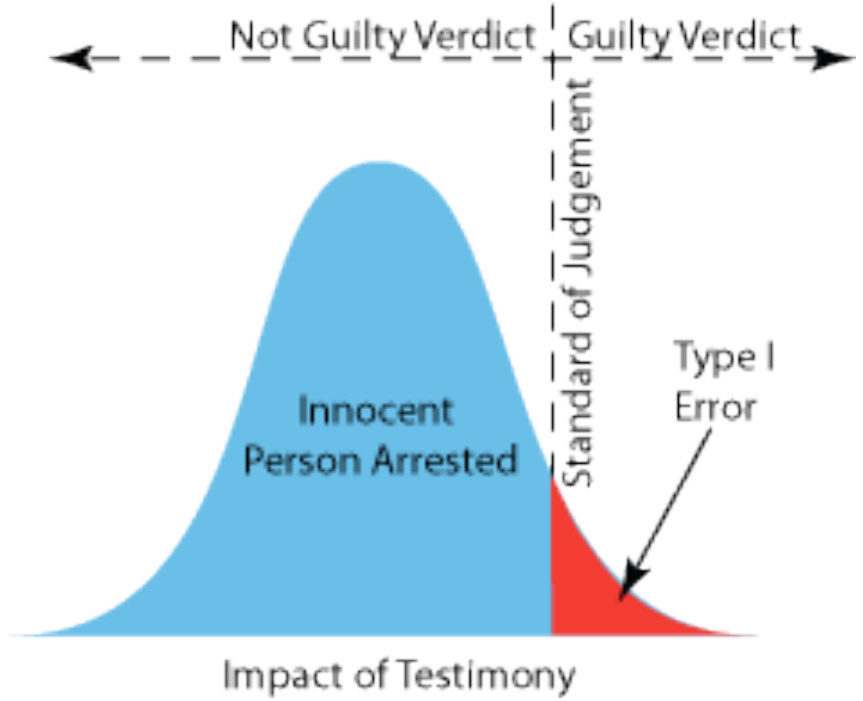
# Neyman and Pearson

- Type I (α): reject the null hypothesis when the null hypothesis is true, and
- Type II (β): fail to reject the null hypothesis when the null hypothesis is false
- "in testing hypotheses two considerations must be kept in view:
  - (1) we must be able to reduce the chance of rejecting a true hypothesis to as low a value as desired;
  - (2) the test must be so devised that it will reject the hypothesis tested when it is likely to be false.

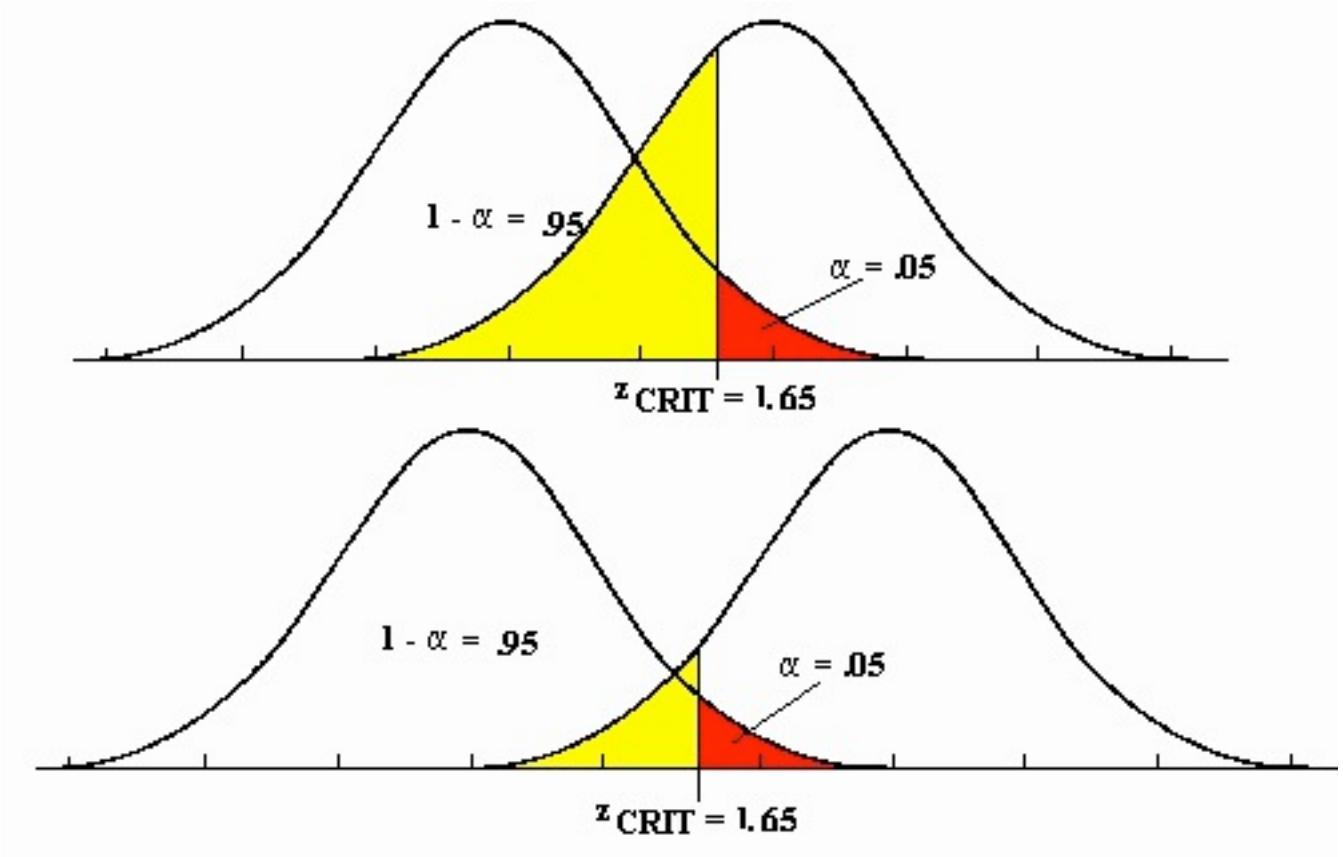
# Error table

| | innocent | guilty |
|---|---|---|
| convict | | |
| acquit | | |

# Error Graph


Not Guilty Verdict
Guilty Verdict
Standard of Judgement
Innocent
Person Arrested
Type I
Error
Impact of Testimony

# Type 1 and Type 2 Errors

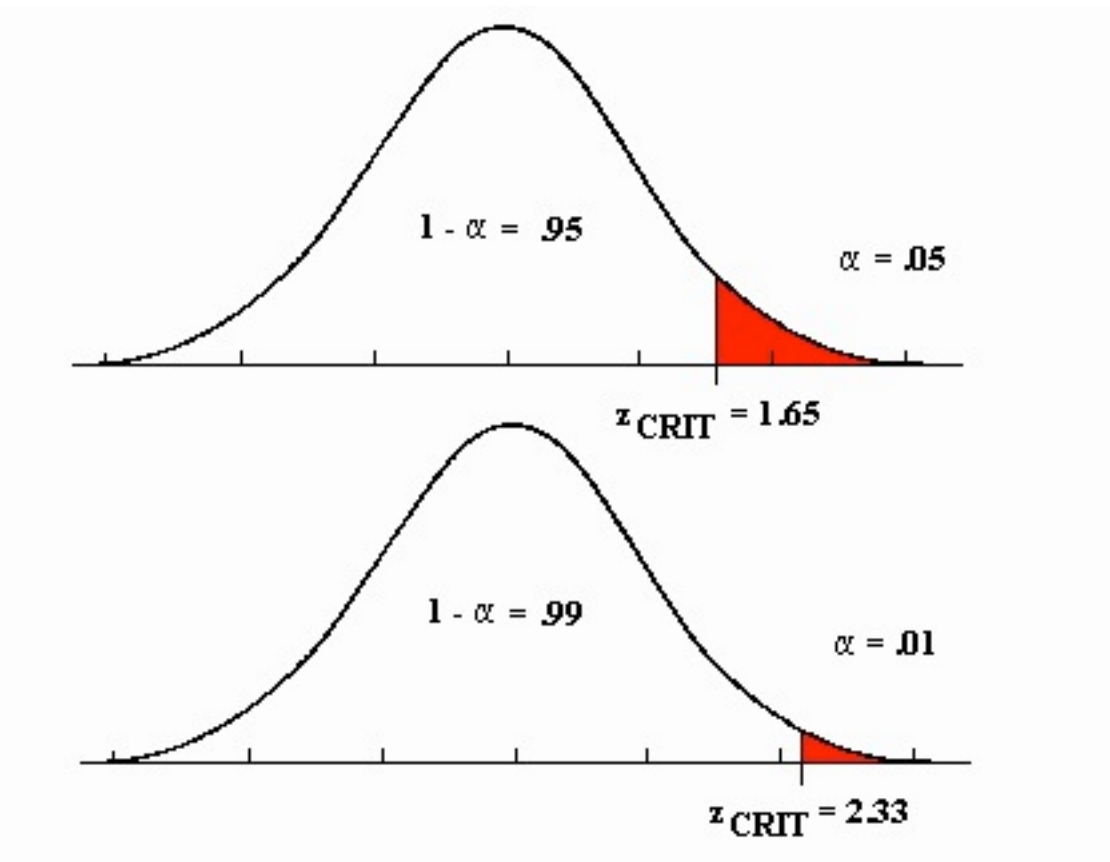


# Definitions

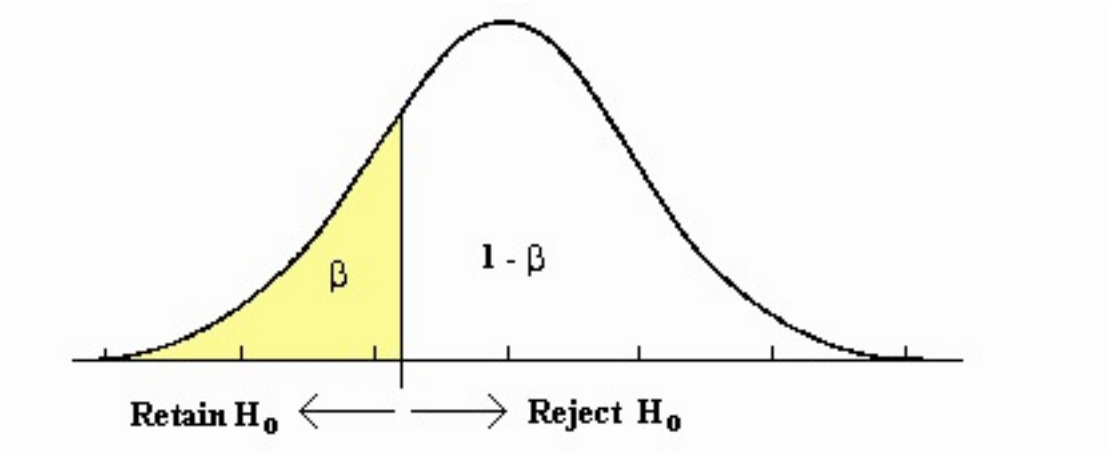
- Type I (α): reject the null hypothesis when the null hypothesis is true, and
- Type II (β): fail to reject the null hypothesis when the null hypothesis is false

- Type I = false alarm
- Type 2 = failure to alarm

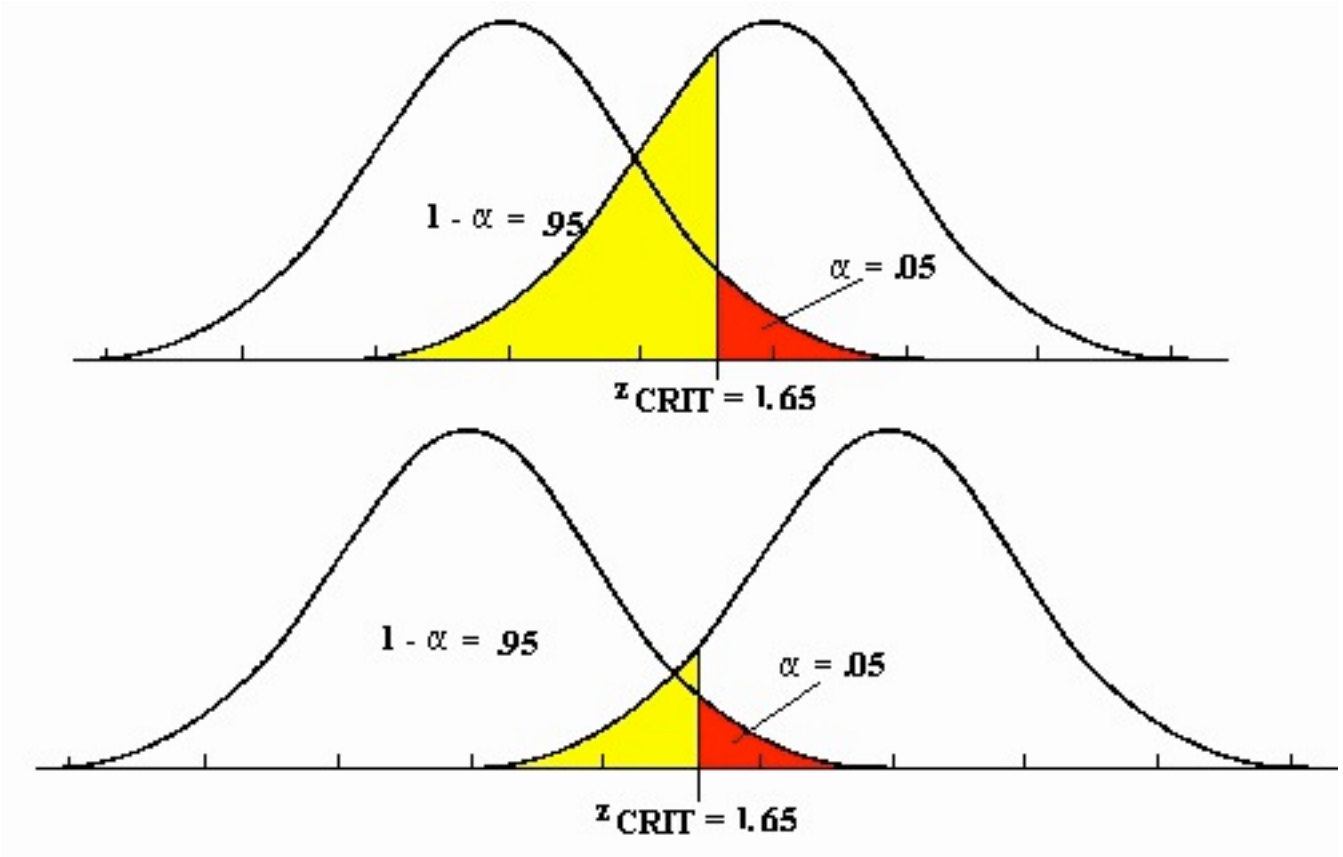
# Type 1 Error



# Type 2 Error

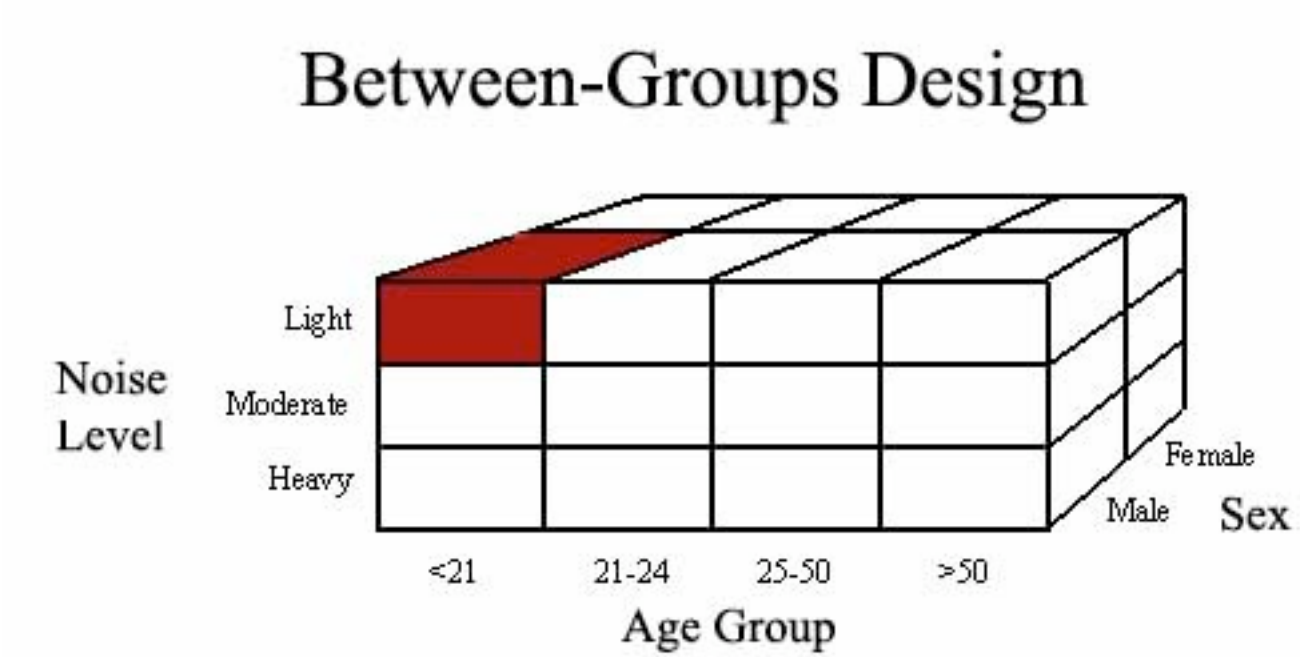


# Type 1 and Type 2 Errors



# ANOVA: Analysis of Variance

# ANOVA: Analysis of Variance



# ANOVA: Analysis of Variance

- When you want to compare many factors

$$F = \frac{\text{Var Between Means}}{\text{Var Within Groups}} = \frac{MS_{Bet}}{MS_{Within}}$$

$$SS_T = \sum x^2 - \frac{(\sum x_T)^2}{N}$$

$$SS_b = \sum \frac{(\sum x_\cdot)^2}{n} - \frac{(\sum x_T)^2}{N}$$

$$SS_w = SS_T - SS_b$$

$$df_b = \text{(number of groups - 1)}$$

$$df_T = \text{(number of subjects - 1)}$$

$$df_w = df_T - df_b$$

$$MS_b = \frac{SS_b}{df_b}$$

$$MS_w = \frac{SS_w}{df_w}$$

$$F = \frac{MS_b}{MS_w}$$

- It's just a ratio, then look the value up in a table
- http://www.psych.utah.edu/stat/introstats/anovaflash.html
- Multivariate analysis of variance: MANOVA

# SPSS, SAS, Excel, Docs, Java, R and free stuff


The Predictive Analytics Company




*Free Statistics*

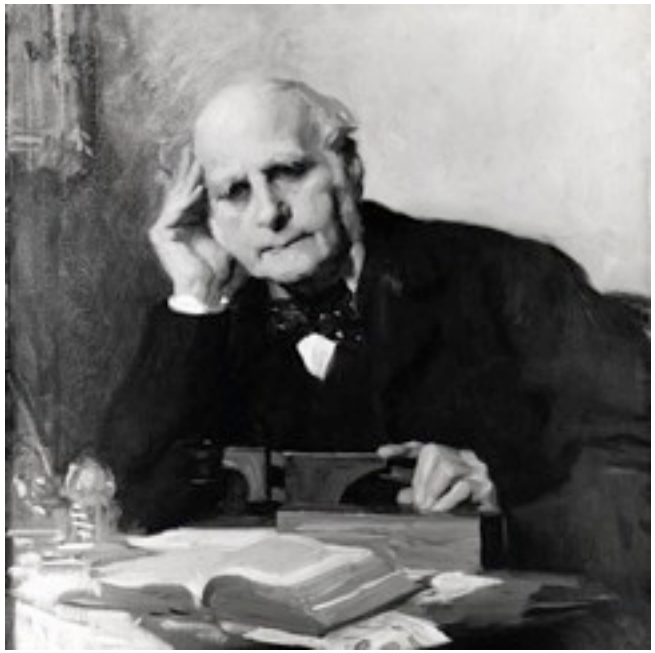
www.freestatistics.info

http://www.r-project.org/

http://www.freestatistics.info/stat.php

# Regression

- What's a better word for this?
- Galton
- author, explorer, psychologist, fingerprints, pioneer statistician



- also, founder of eugenics
- ethics aren't always obvious

# Dependent vs. Independent Variables

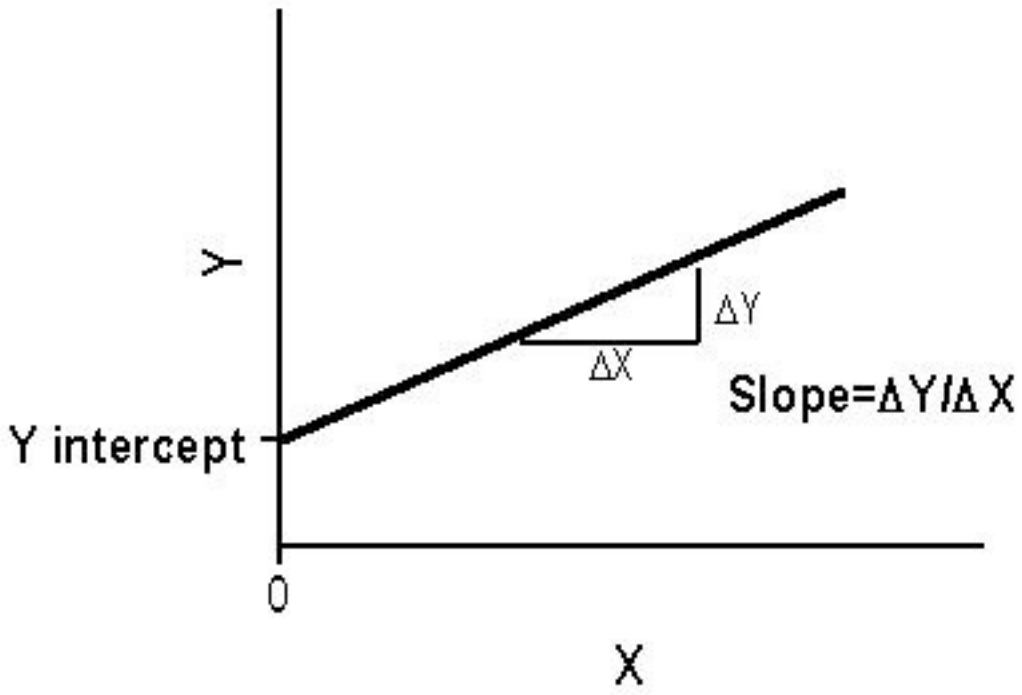
- Dependent variable is what we're interested in

- Independent will tell us where to look for it

# Linear Regression

- Y' = bX + A

- where X is the variable represented on the abscissa (X-axis), b is the slope of the line, A is the Y intercept, and Y' consists of the predicted values of Y for the various values of X



# Standard Error of the Estimate

- The standard error of the estimate is a measure of the accuracy of predictions made with a regression line.

- The sum of the errors of prediction is zero, so just like in confirmatory statistics, we square the errors so we can deal with them mathematically

- Standard Error of the Estimate: $$\sigma_{est} = \sqrt{\frac{\sum (Y-Y')^2}{N}}$$

  - ...where N is the number of pairs of (X,Y) points, Y is each dependent variable and Y' (y-prime) is the predicted value

- We don't typically know population values so: $$s_{est} = \sqrt{\frac{\sum (Y-Y')^2}{N-2}}$$

# Sums of Squares

- Similar to ANOVA, we need to partition the error terms

- SSY = SSY' + SSE

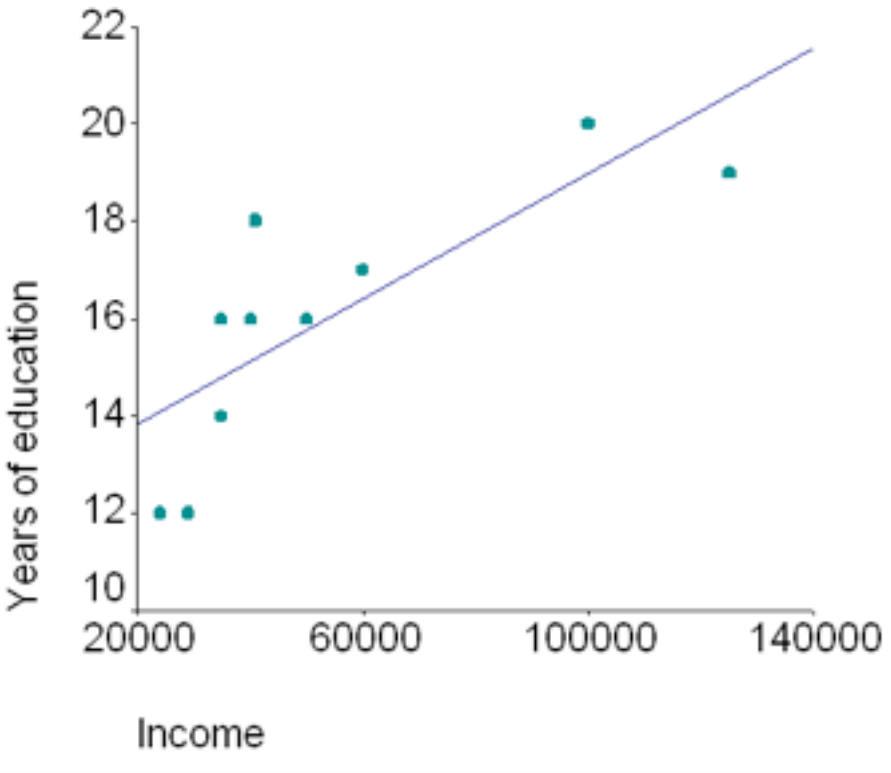
- $r^2$ = SSY'/SSY

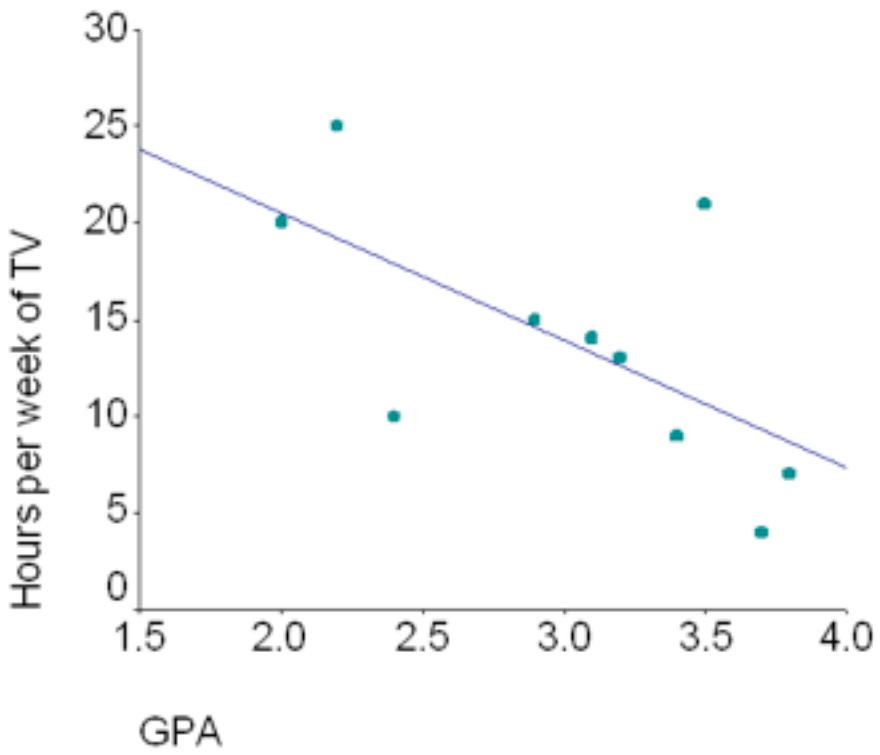
- Significance test: $$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- Look up *t* in a table and you're done.

# Correlation

- how do variables relate to each other?



# Correlation Coefficient

- Sample correlation: $$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

- where x-bar and y-bar are the sample means of X and Y , sx and sy are the sample standard deviations of X and Y

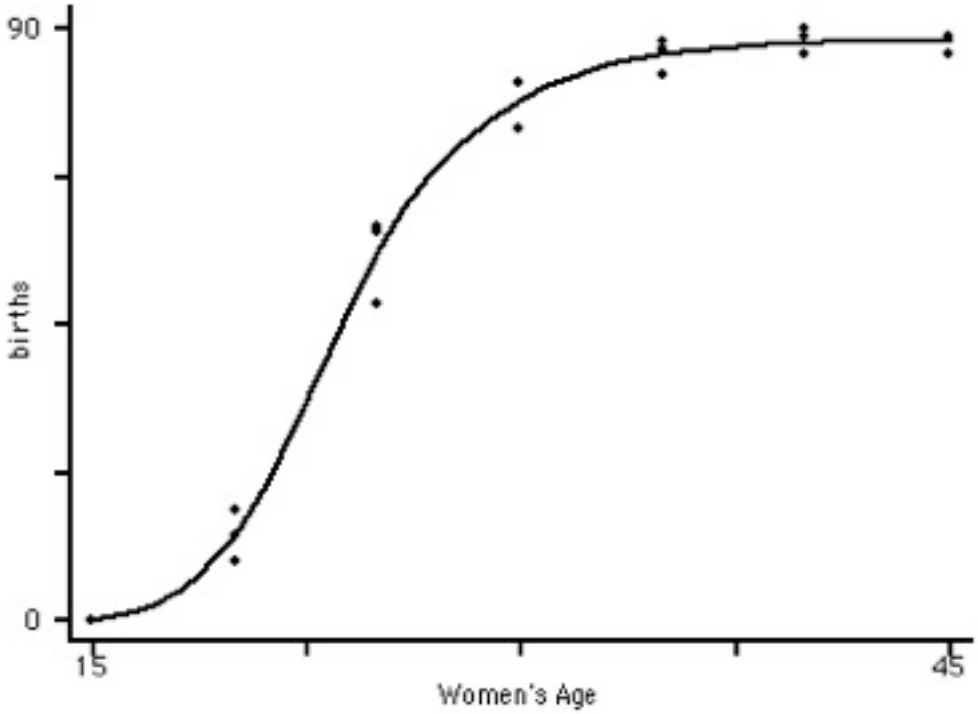
- Phew! So how do we interpret *r* ?

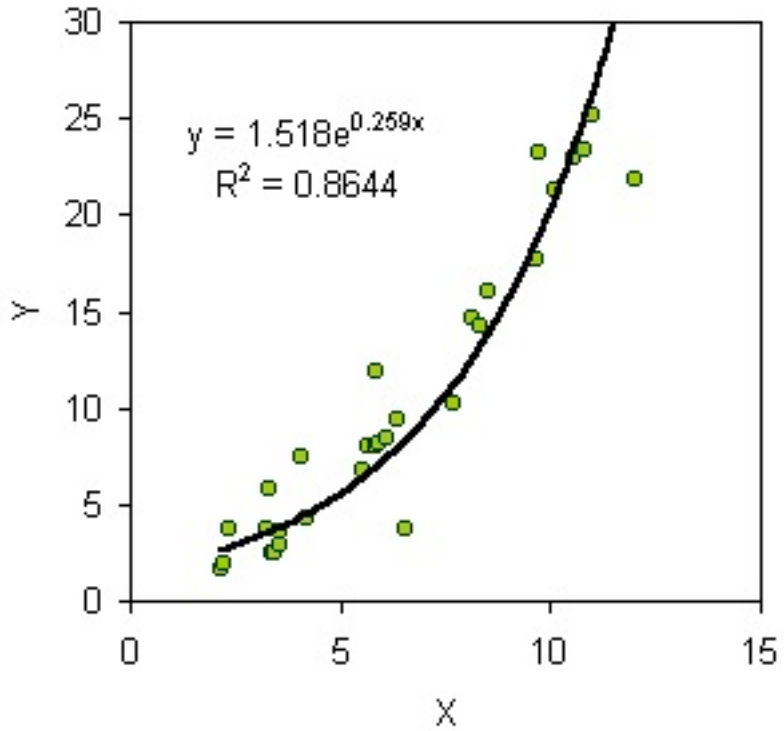
- large ≠ important, usually = trivial

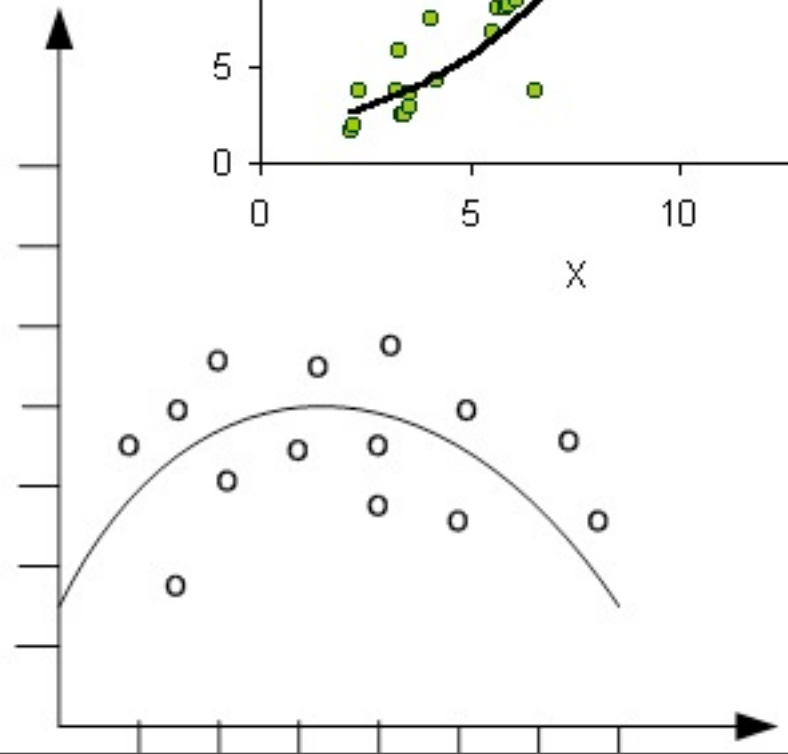
| | | |
|---|---|---|
| small | 0.1 | 0.3 |
| medium | 0.3 | 0.5 |
| large | 0.5 | 1.0 |

# Non-linear regression

- deals with curves



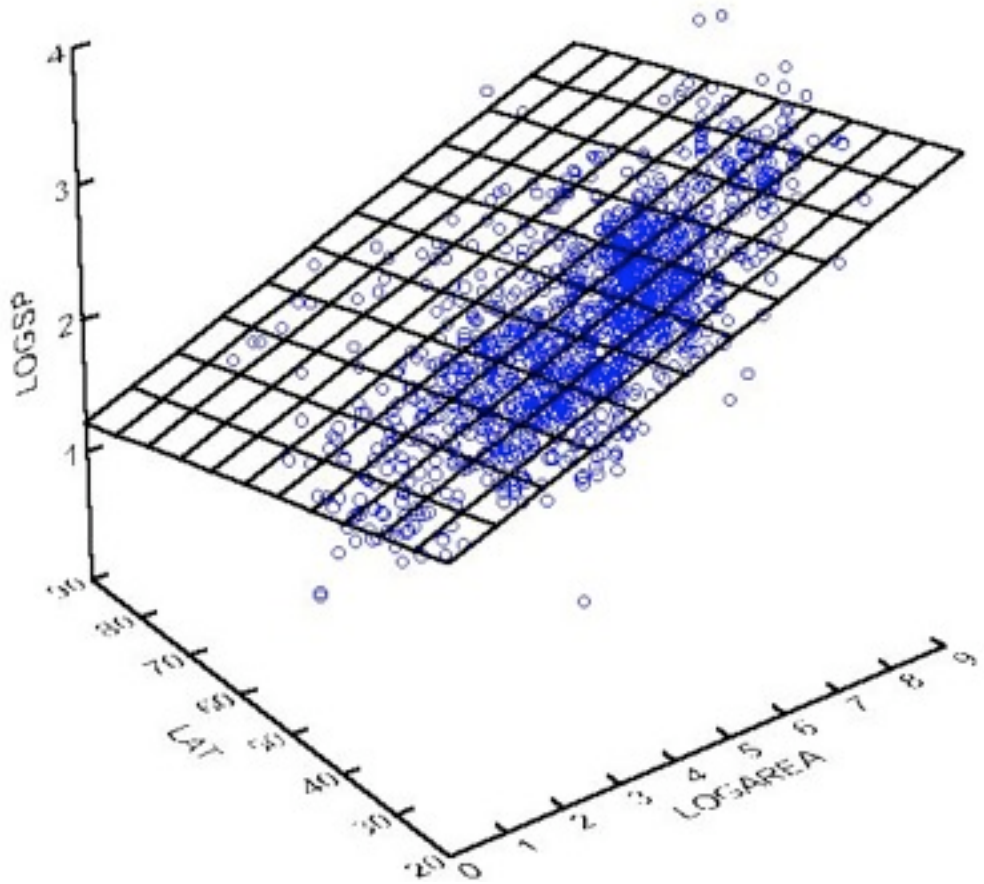



# Multiple regression

- Deals with multiple independent variables

- fitting a plane



# Readings and Assignments

- Readings
    - Lady Tasting Tea
- Assignments
    - be ready to present anything unpresented