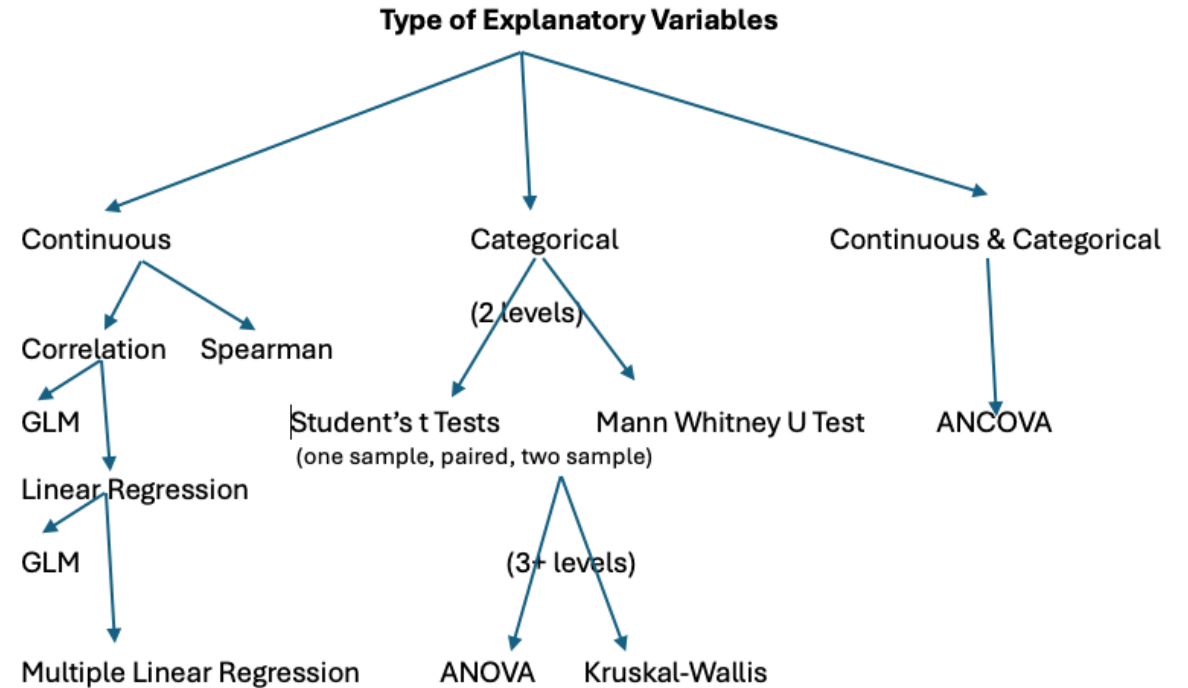# Module 3D:
ANOVA & Correlation

Assigning signal and noise to variation

Agenda:

1. ANOVA: Nuts & Bolts

2. Worked Example
   A. One way ANOVA
   B. Post-hoc tests: Tukey-Kramer
   C. Kruskal-Wallis (nonparametric)

3. **Linear Correlation**
   A. Spearman's rank

Type of Explanatory Variables

Continuous → Correlation → GLM → Linear Regression → GLM → Multiple Linear Regression

Continuous → Spearman

Categorical (2 levels) → Student's t Tests (one sample, paired, two sample) / Mann Whitney U Test

Categorical (3+ levels) → ANOVA / Kruskal-Wallis

Continuous & Categorical → ANCOVA

Correlation

Imagine you determine there is a positive correlation between the number of times individuals visit doctors (over the course of their life span) and longevity (how many years the individuals live). Would it be appropriate to conclude that frequent visits to the doctor lead to a longer life span? Why or why not?

Two Variables: Which test?

		Explanatory variable	
		Categorical	**Numerical**
Response Variable	**Categorical**	• Contingency analysis	• Logistic regression • Survival analysis
	Numerical	• t-test • Analysis of variance	•Regression • Correlation

http://www.socr.ucla.edu/htmls/SOCR_ChoiceOfStatisticalTest.html

Correlation VS Regression

- Mathematical relationship between the two methods of analysis are close: share many computational steps which are similar or the same
- Appropriate method depends on the purpose of the investigator and the nature of the variables

Regression:
   Intention: describe <u>dependence</u> of a variable Y, on an <u>independent</u> variable X.

   Regression equations are employed to support hypotheses regarding possible **causation** of changes in Y by changes in X; to predict Y in terms of X

   A line (in linear regression) is fit to the data

   *Study the **effects** of **X** on **Y***
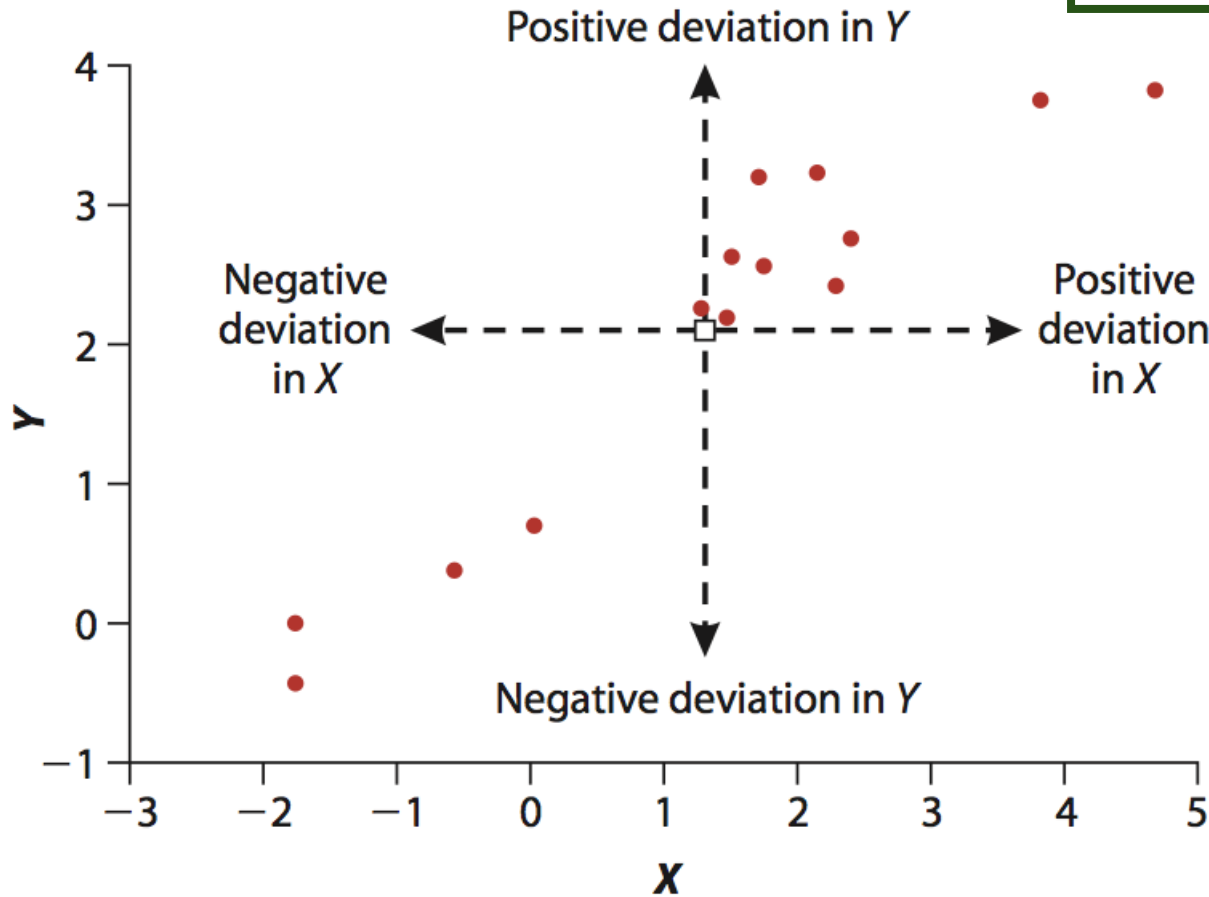
Correlation:

- Measures the amount/degree of linear association between two **numerical** variables

- Estimate the degree to which variables **covary**

- We <u>do not</u> express one variable as a function of the other variable

- <u>No</u> distinction between dependent & independent variables
  - In fact, we usually assume that they both stem from a common cause -- with a pair of variables whose correlation is studied, one may very well be the cause of the other but we neither know nor assume!
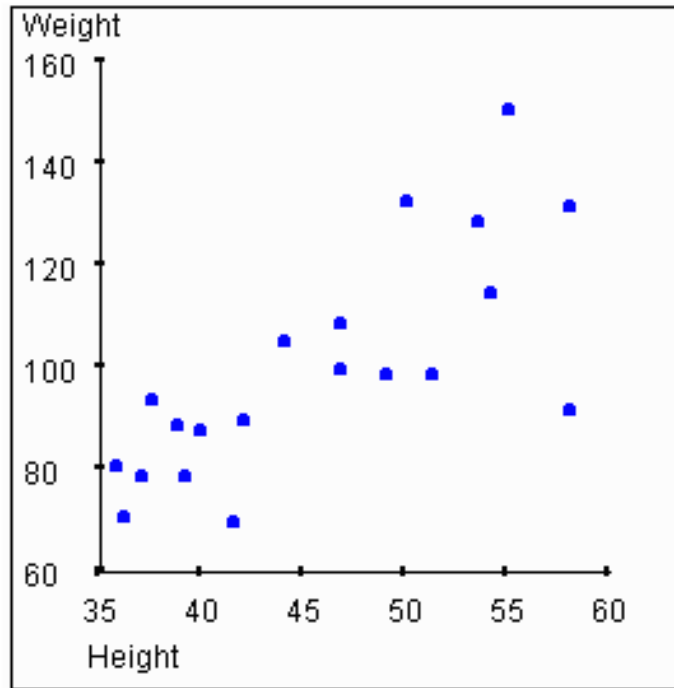
(Pearson) Correlation Coefficient

$$r = \frac{\overset{\circ}{a}(X - \bar{X})(Y - \bar{Y})}{\sqrt{\overset{\circ}{a}(X - \bar{X})^2} \sqrt{\overset{\circ}{a}(Y - \bar{Y})^2}} = \frac{Co\,var\,iance(X,Y)}{s_x s_y}$$
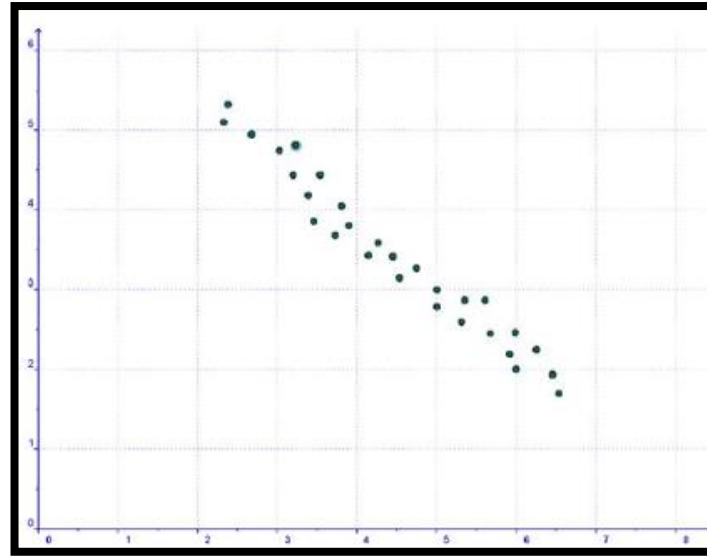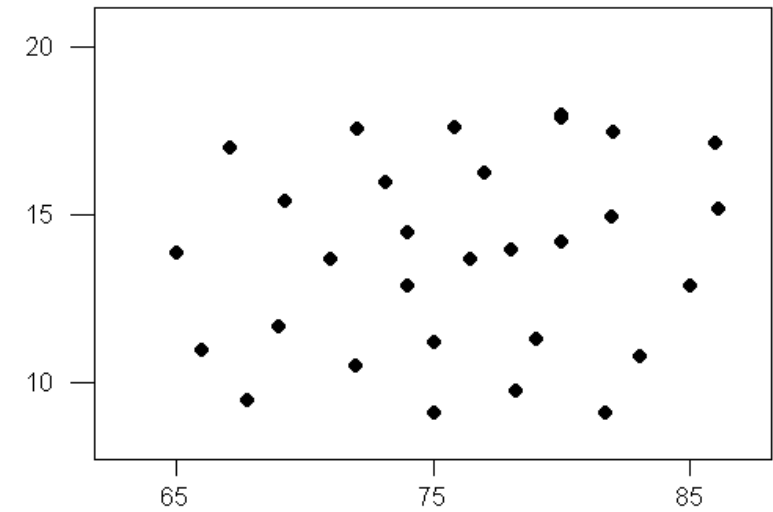
NOISE

Positive Correlation

Negative Correlation

No Correlation

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$

$$-1 < r, \rho < 1$$

This ONLY describes LINEAR correlation. If two variables are tightly correlated, but nonlinear: