Linear Regression, Correlation, Bootstrap

Raguvir Kunani

Discussion 8

October 14, 2019

Simple Linear Regression Quick Review

Simple linear regression involves finding a "line of best fit" that explains the relationship between **2** variables x and y. In fancy math terms:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

All this really means is we are trying to find the values of a and b so that a line with the equation $\hat{y} = a + bx$ best fits the data.

Correlation

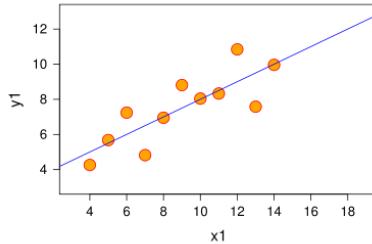Correlation is a measure of the strength of the <u>linear association</u> between 2 variables.

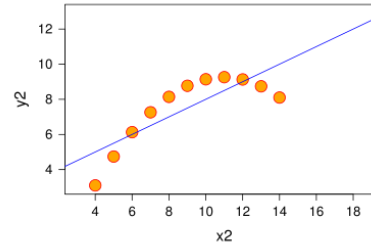| r = 0.4	r = 0	r = -0.4
Positive Correlation	No correlation	Negative
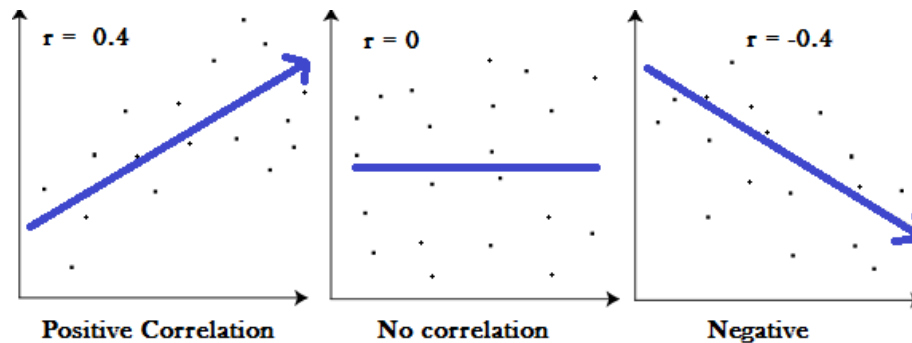
correlation coefficient: r

$-1 \leq r \leq 1$ always!

Common Pitfalls with Correlation

1. Correlation ≠ Causation *for observational studies*

2. High Correlation ≠ Strong Linear Relationship

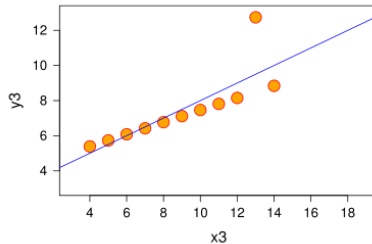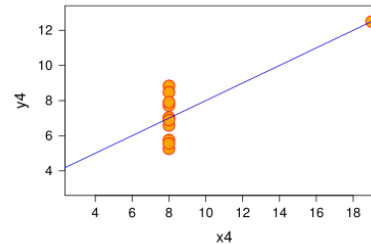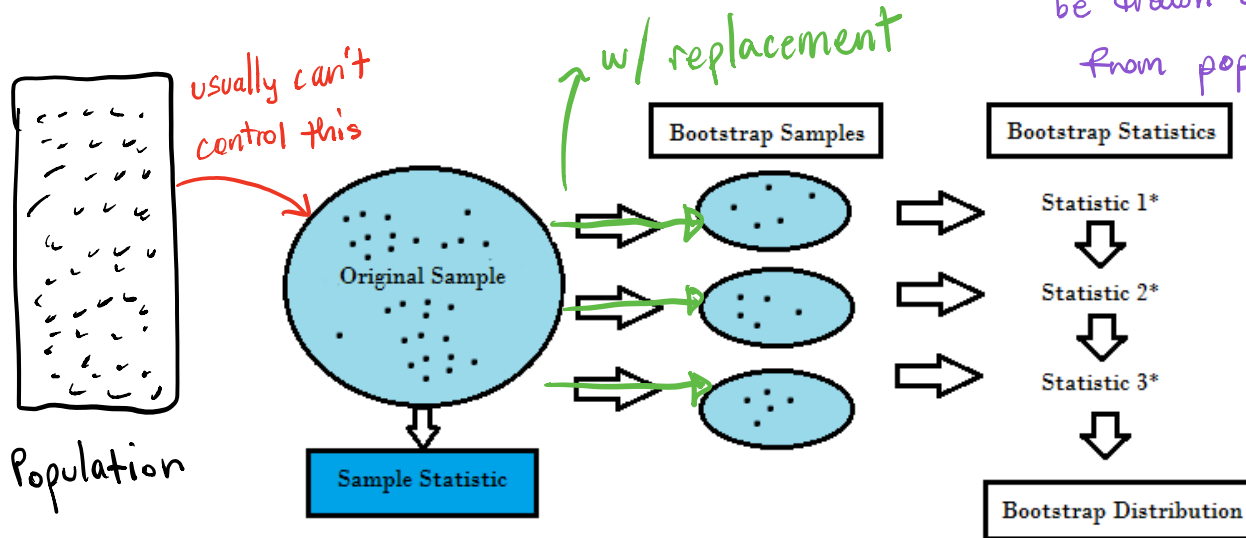*All these data have same correlation but only ① is actually linear!*

Sorry for poor wording. This should say: "treat original sample as a population and draw samples from the original sample"

Bootstrapping

Treat the sample as a population and sample from the sample!

What assumption must hold for us to do this? → *original sample must be drawn at random from population*

usually can't control this

w/ replacement

Population

Original Sample

Sample Statistic

Bootstrap Samples

Statistic 1*
Statistic 2*
Statistic 3*

Bootstrap Statistics

Bootstrap Distribution

also called bootstrap sampling distribution

Worksheet!

Hypothesis Testing Quick Review

- **Data Generation Model**: Our assumptions about how the data was generated (e.g. data is uniformly distributed)

- **Null Hypothesis**: A statement about the model saying that the observed outcome happened by chance _(under the data generation model)_

- **Test Statistic**: A function of the data; helps us determine whether to <u>reject or fail to reject</u> the null hypothesis

  NEVER accept the null hypothesis

- **Sampling distribution**: distribution of all possible values of test statistic for a fixed sample size (from data generation model)

- **p-value**: the chance, under the data generation model, that the test statistic is equal to or more extreme than the observed statistic