YData: Introduction to Data Science

Lecture 29: correlation

Overview

Review: Sampling distributions

Confidence intervals for a mean and proportions revisited

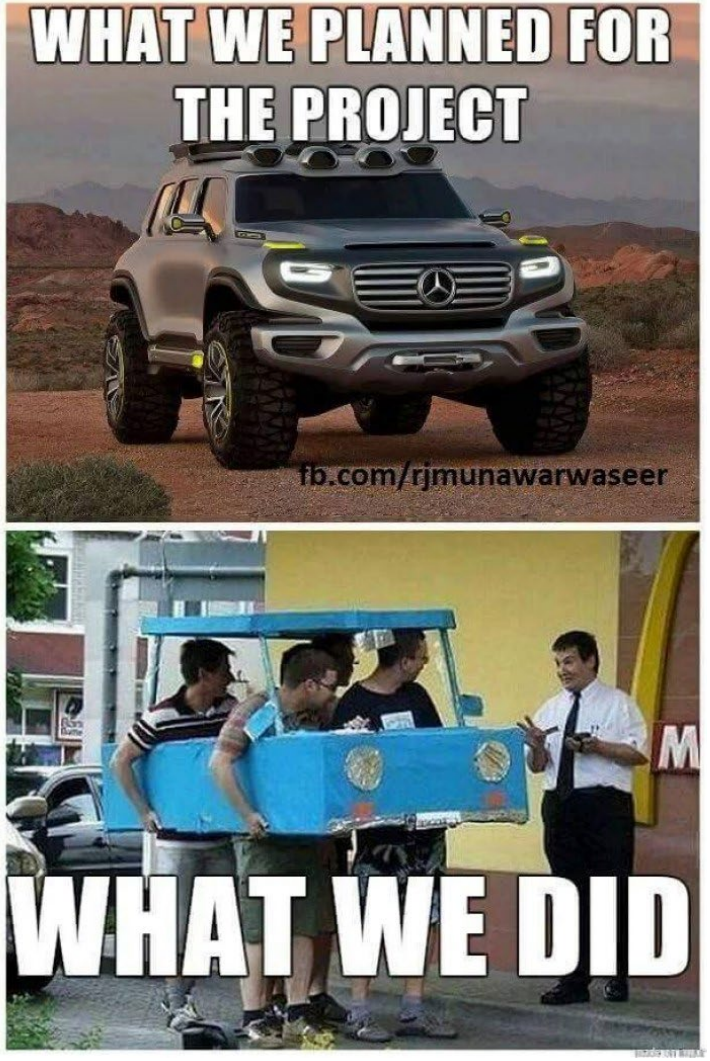
Correlation

- Predictions
- Associations
- The correlation coefficient
- Correlation cautions

Announcements

Project 2 is due tonight

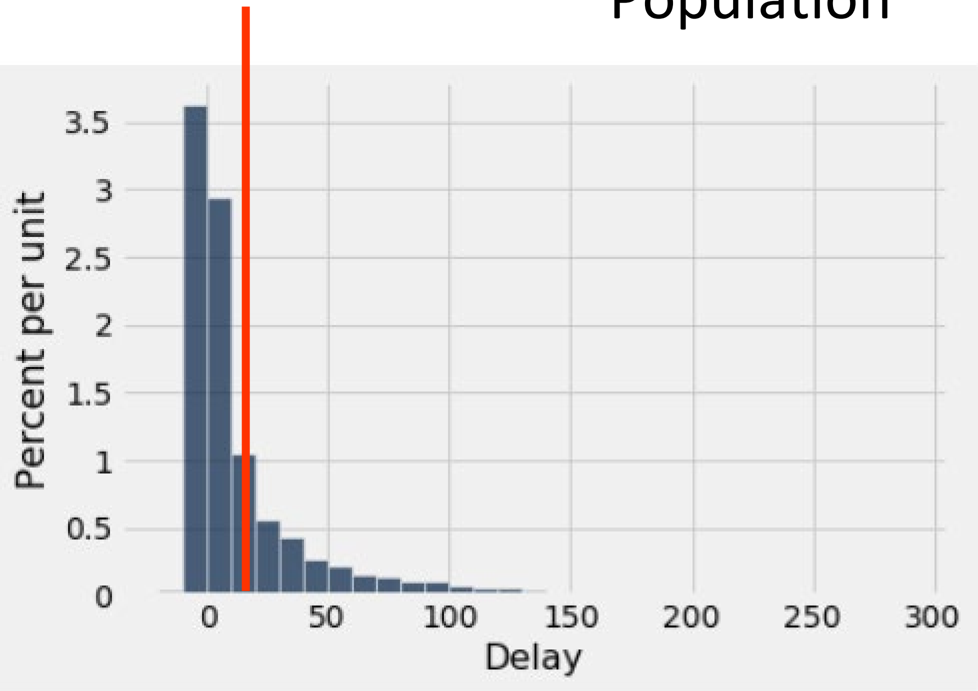
Homework 8 is due on Sunday



Review: sampling distributions

μ = 16.7

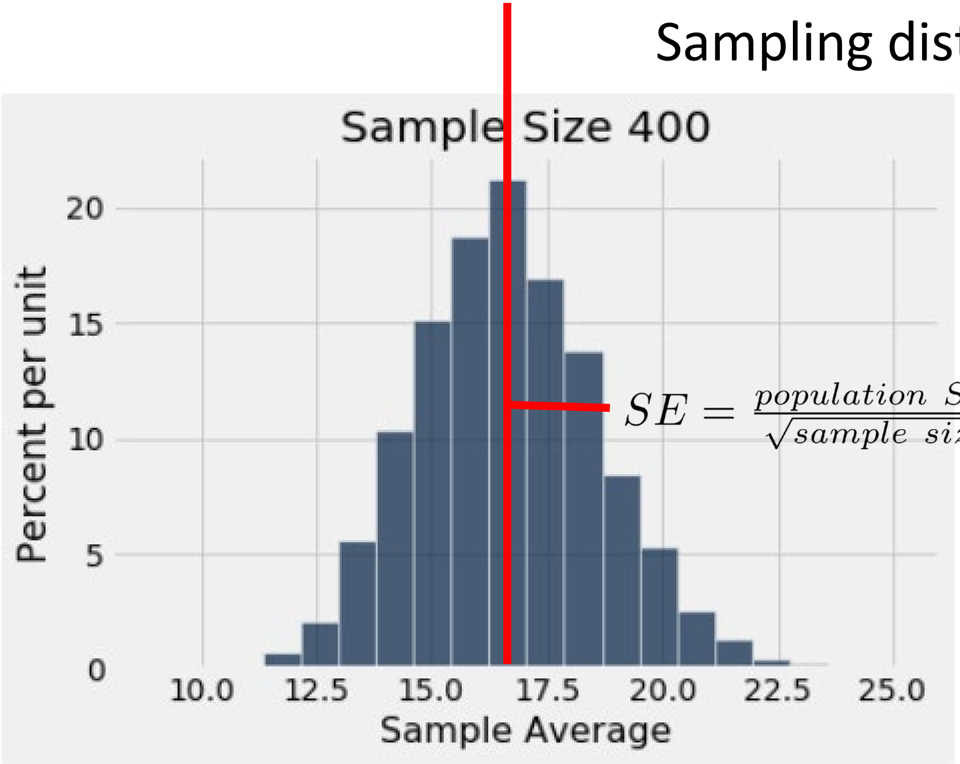
Population



Take many sample of size n = 400 and calculate means ($\bar{x}$'s)

μ = 16.7

Sampling dist



$SE = \frac{population\ SD}{\sqrt{sample\ size}}$

By the CLT, the sampling distribution of $\bar{x}$'s is roughly normal:

- Center: the population average (μ)
- Spread: $SE = \frac{population\ SD}{\sqrt{sample\ size}}$

Two approximate sampling distributions

For the population of flight delays, we had:

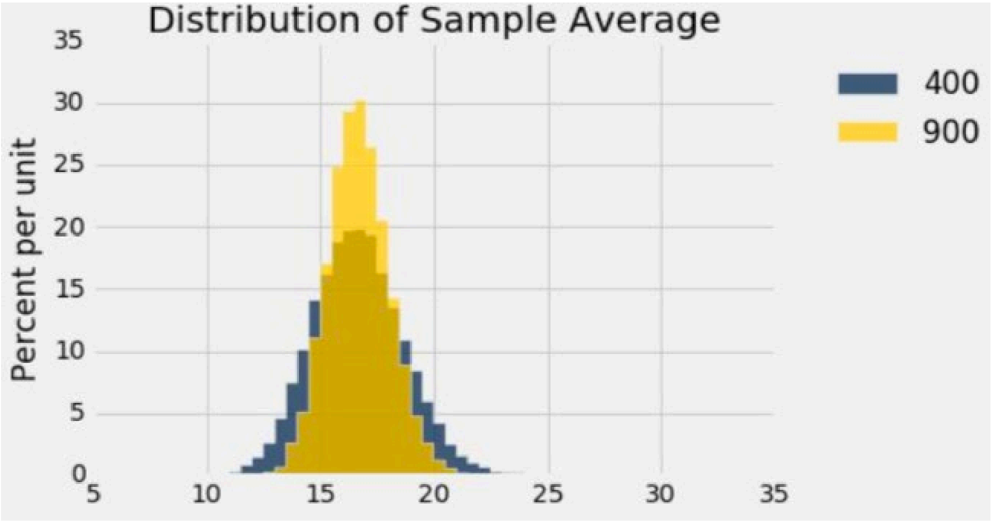
- μ = 16.7
- σ = 39.5

For a sample of size n = 400, we had:

- np.mean(means_400) = 16.7
- np.std(means_400) = 1.98 # SE from samp dist
- σ/sqrt(400) = 1.97 # SE based on equation

For a sample of size n = 900, we had:

- np.mean(means_900) = 16.7
- np.std(means_900) = 1.31 # SE from samp dist
- σ/sqrt(900) = 1.32 # SE based on equation



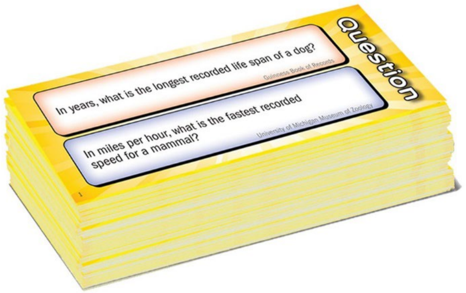
$$SE = \frac{population\ SD}{\sqrt{sample\ size}}$$

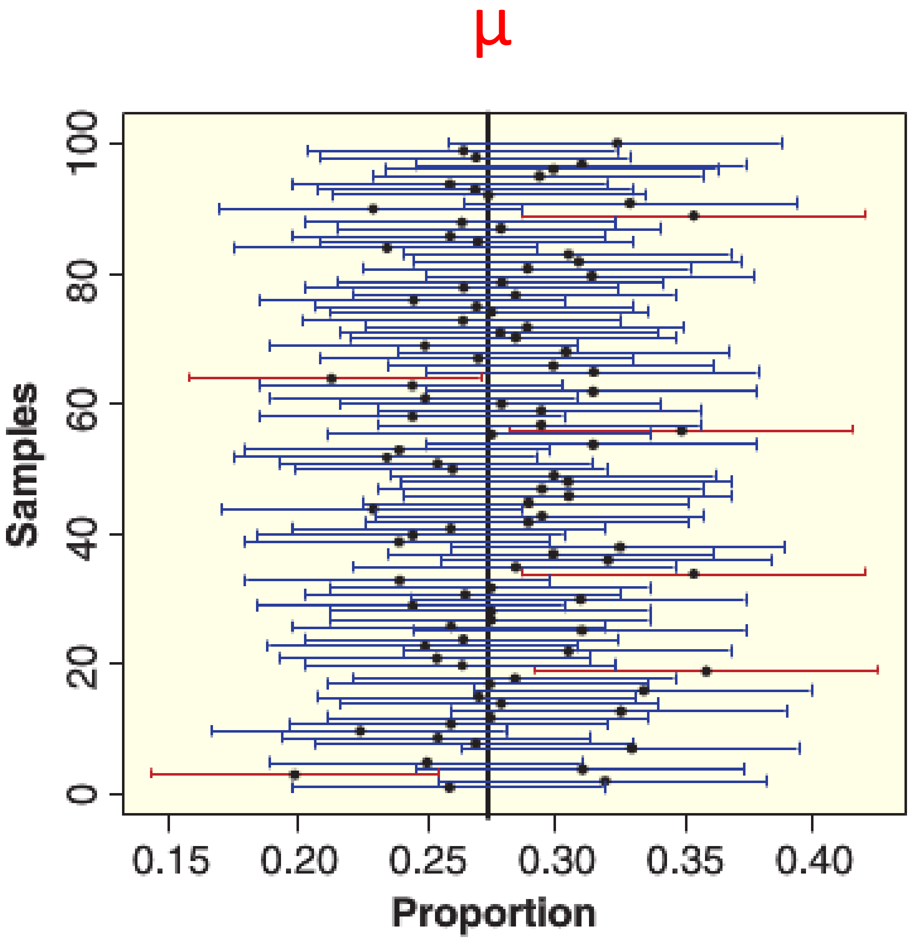
Confidence intervals

Recall: confidence intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

The **confidence level** is the percent of all intervals that contain the parameter





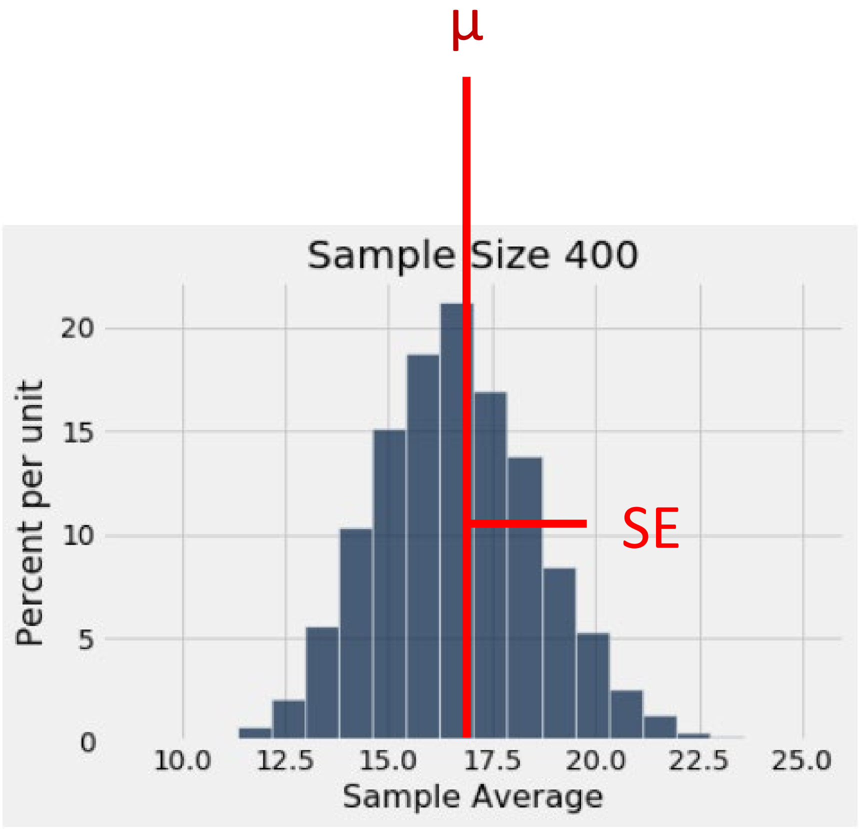
Variability of the sampling distribution

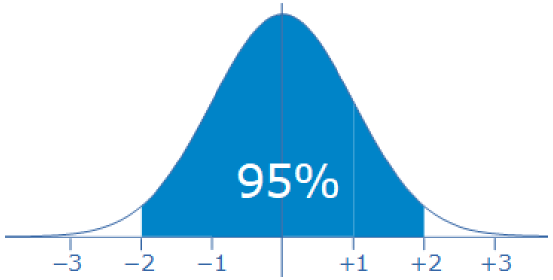
Recall our sampling distribution is roughly normal:

- Center: the population average (μ)
- Spread: $SE = \frac{population\ SD}{\sqrt{sample\ \ size}}$

What percent of our statistics lie within 2 standard deviations (i.e., 2 SE) of the mean?

- 95% of our statistics in the sampling distribution lie within 2 SE of the mean





Constructing confidence intervals

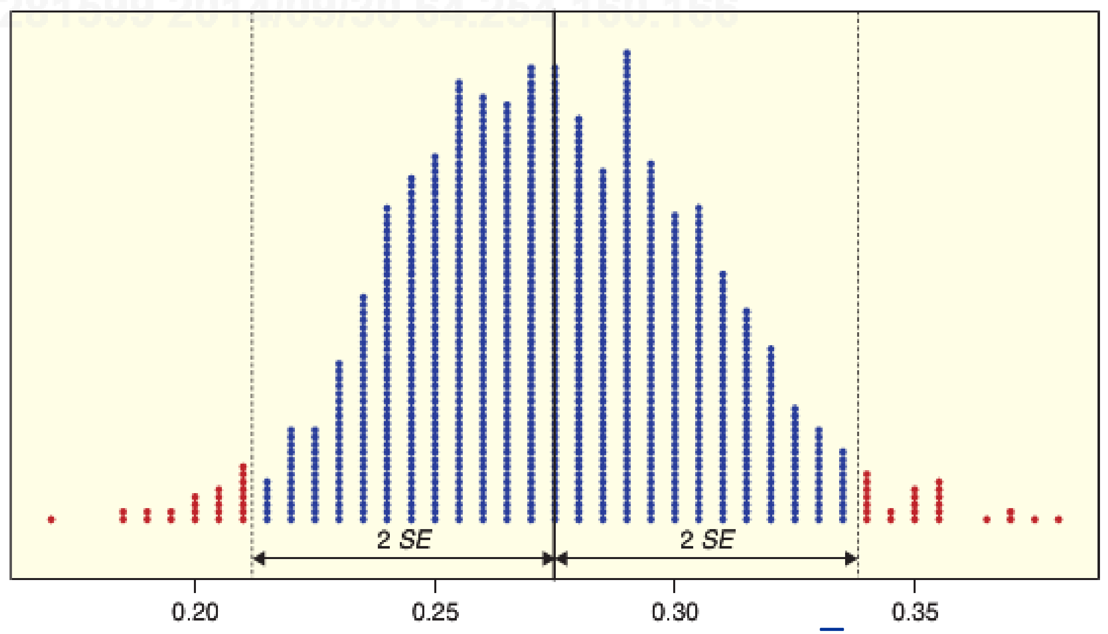
We can construct 95% confidence intervals for a population mean μ using:

$\bar{x} \pm 2 \cdot SE$

Why does this work?

Sampling distribution

μ



$\bar{x}$

2 · SE   2 · SE

Confidence interval

95% percent of the sample means $\bar{x}$ we collect will be within $\pm\ 2 \cdot SE$ of the population mean μ

- So $\bar{x} \pm 2 \cdot SE$ will overlap with μ 95% of the time

Let's explore this in Jupyter!

Sample proportions

Proportions are averages

Suppose we had the following data and we wanted to calculate the proportion of cats ($\hat{p}_{cat}$):

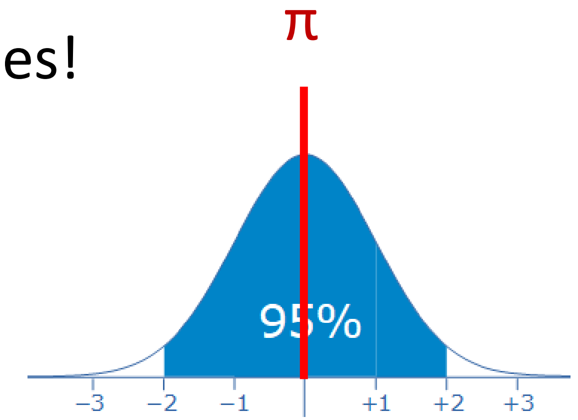
Categorical data: "dog", "cat", "fish", "dog", "cat", "dog", "cat", "cat", "fish", "dog"

We can code data: 0 1 0 0 1 0 1 1 0 0

We can calculate the proportion based on taking the average of the coded data

Since we are dealing with averages, the central limit theorem applies!

A conservative estimate for the SE is: $SE = \frac{.5}{\sqrt{n}}$



Let's explore this in Jupyter!

Prediction

Guess the future

Predictions are based on incomplete information

One way to predict an outcome for an individual

- Find others who are like that individual and whose outcomes you know
- Use those outcomes as the basis of your prediction

What examples of predictions have we seen in this class already?

- Class 9…
- Galton, predicting children's heights based on their parents' heights

Let's explore this in Jupyter!

Association

Two numerical variables

When we have two quantitative variables, we can explore trends in our data that are useful for making predictions

- Usual to visualize trends, and then to quantify them

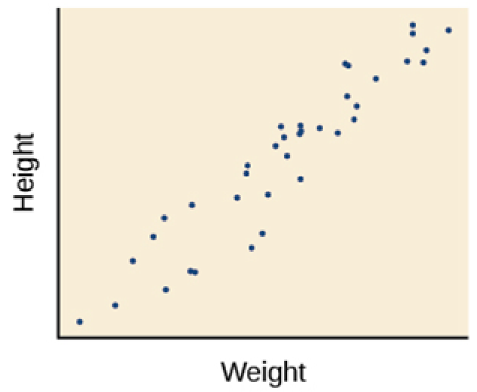
Trend

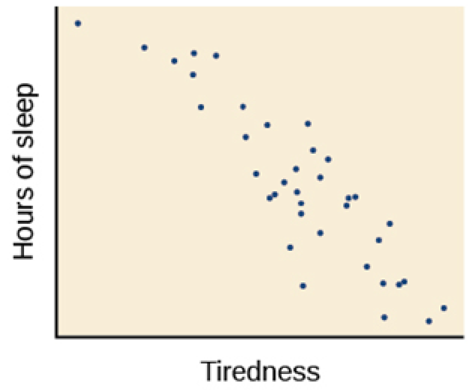
- Positive association
- Negative association

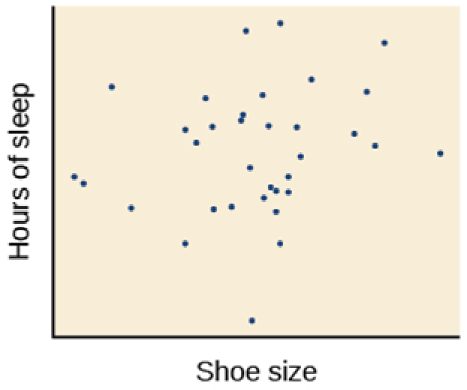
Pattern

- Any discernible "shape" in the scatter
- Linear
- Non-linear

Let's explore this in Jupyter!







Correlation coefficient

The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables

- The statistic is denoted with the symbol r
- The parameter is denoted with the symbol ρ (rho)

Based on standard units

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

It is always between -1 and 1:

- r = 1: scatter is perfect straight line sloping up
- r = -1: scatter is perfect straight line sloping down
- r = 0: No linear association; uncorrelated

Let's explore this in Jupyter!

Correlation cautions

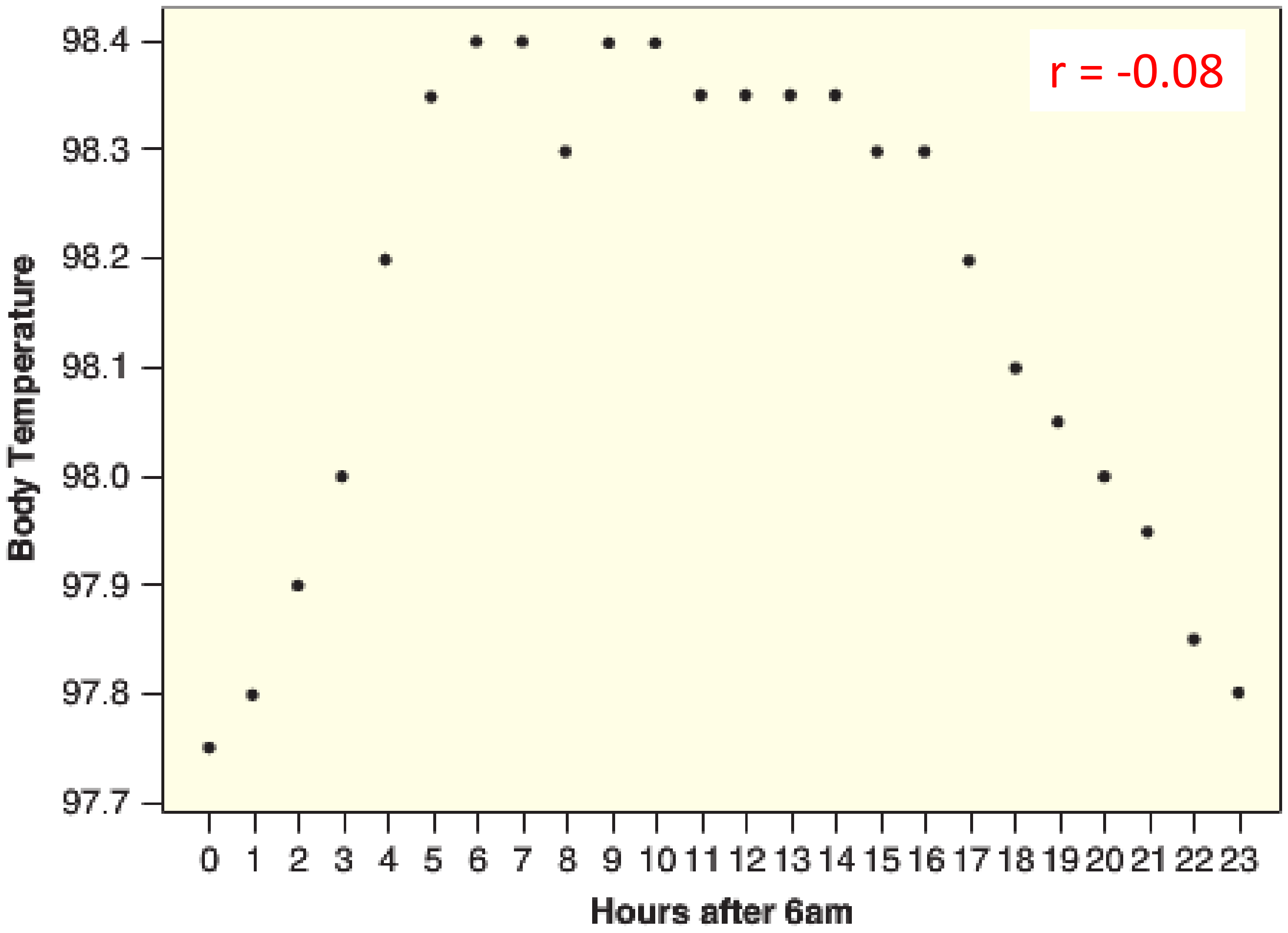
Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables

Correlation caution #2

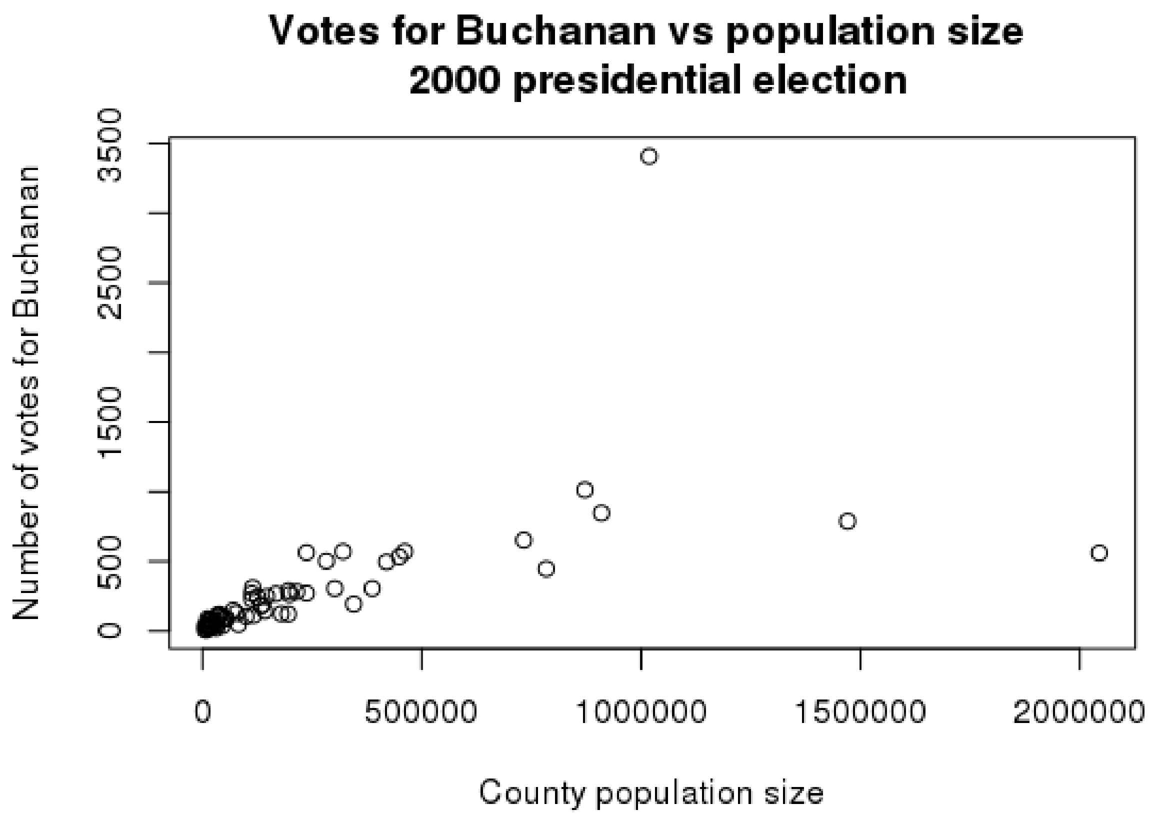
A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day



Correlation caution #3

Correlation can be heavily influenced by outliers. Always plot your data!



With Palm Beach
r = 0.61

Without Palm Beach
r = .78

Let's explore this in Jupyter!