Week 11 – **Sample Means, CLT, Correlation**

Slides by Suraj Rampure

Fall 2017

Administrative Notes

Complete the mid semester survey on Piazza by Nov. 8!

If 85% of the class fills it out, everyone will get a bonus point (on something).

Variability of the Sample Mean

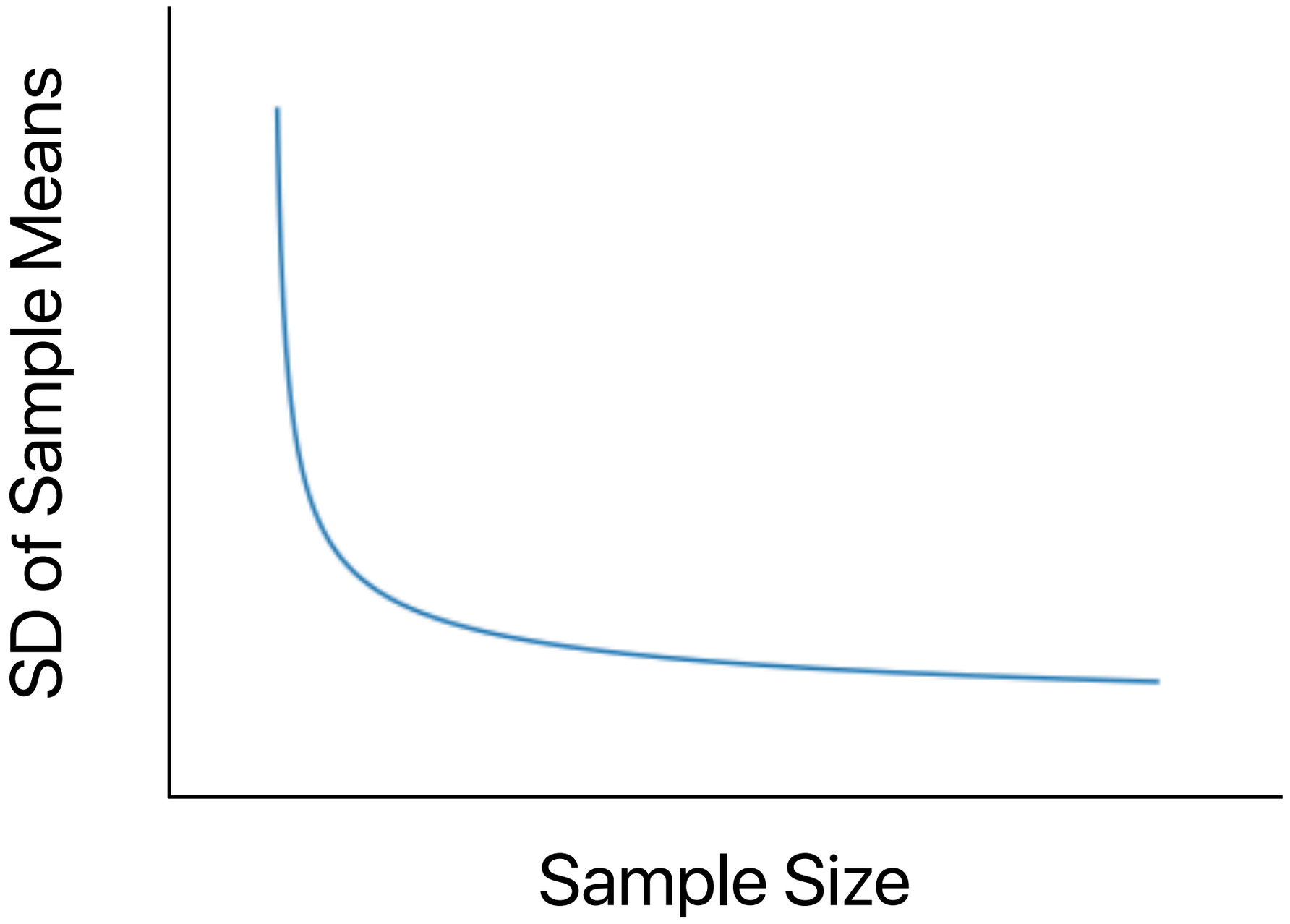
Variability of the Sample Mean

Though you've likely already seen it, this relationship will be very important when we start dealing with the Central Limit Theorem and correlation/regression.

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

Source: https://www.inferentialthinking.com/chapters/12/5/variability-of-the-sample-mean.html

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$



Overview of Data 8

Part 1	Part 2	Part 3
Tables and Visualization	Statistical Inference	**Prediction**

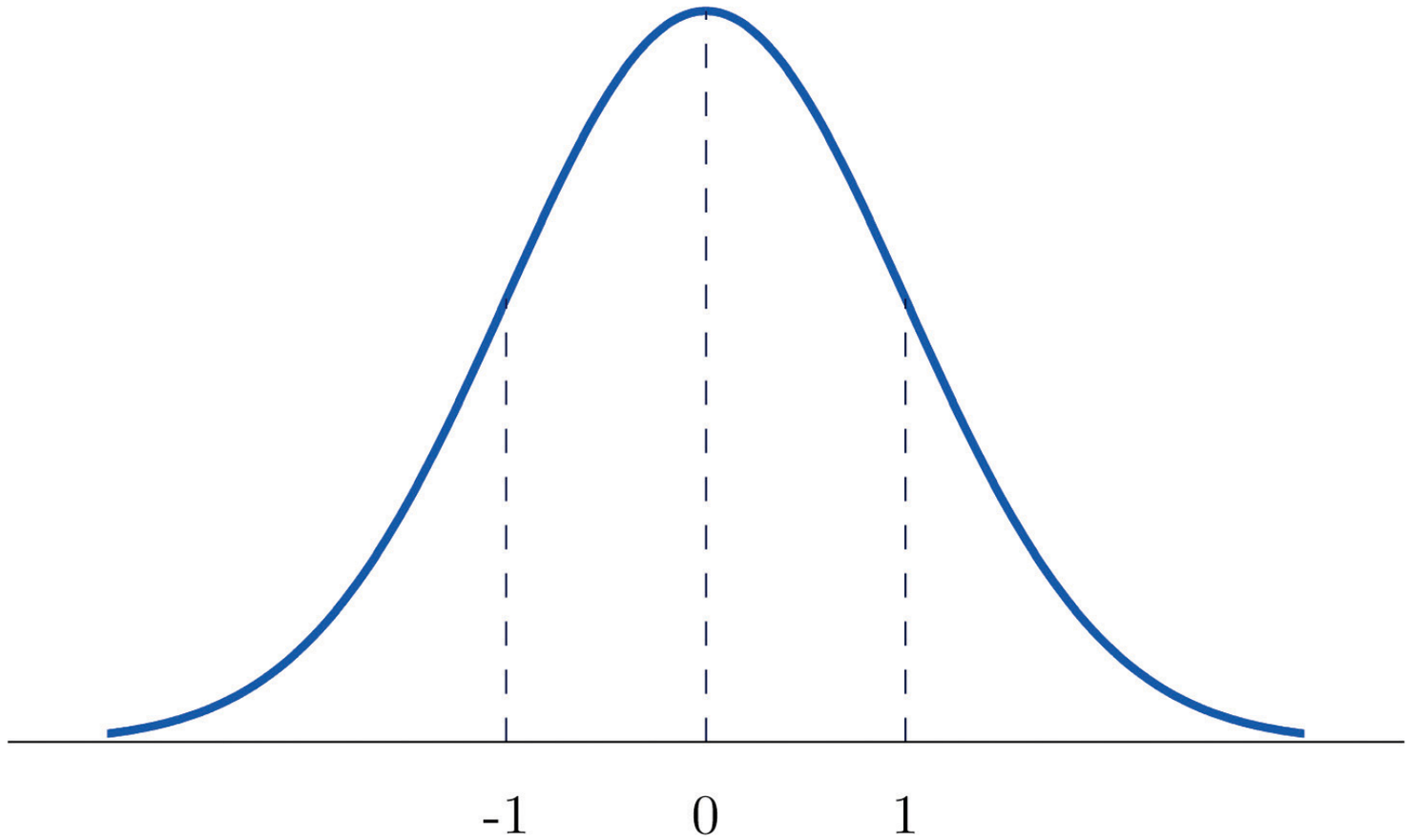
We're about to transition to Part 3.

The Normal Distribution and the Central Limit Theorem

(Review from lecture)

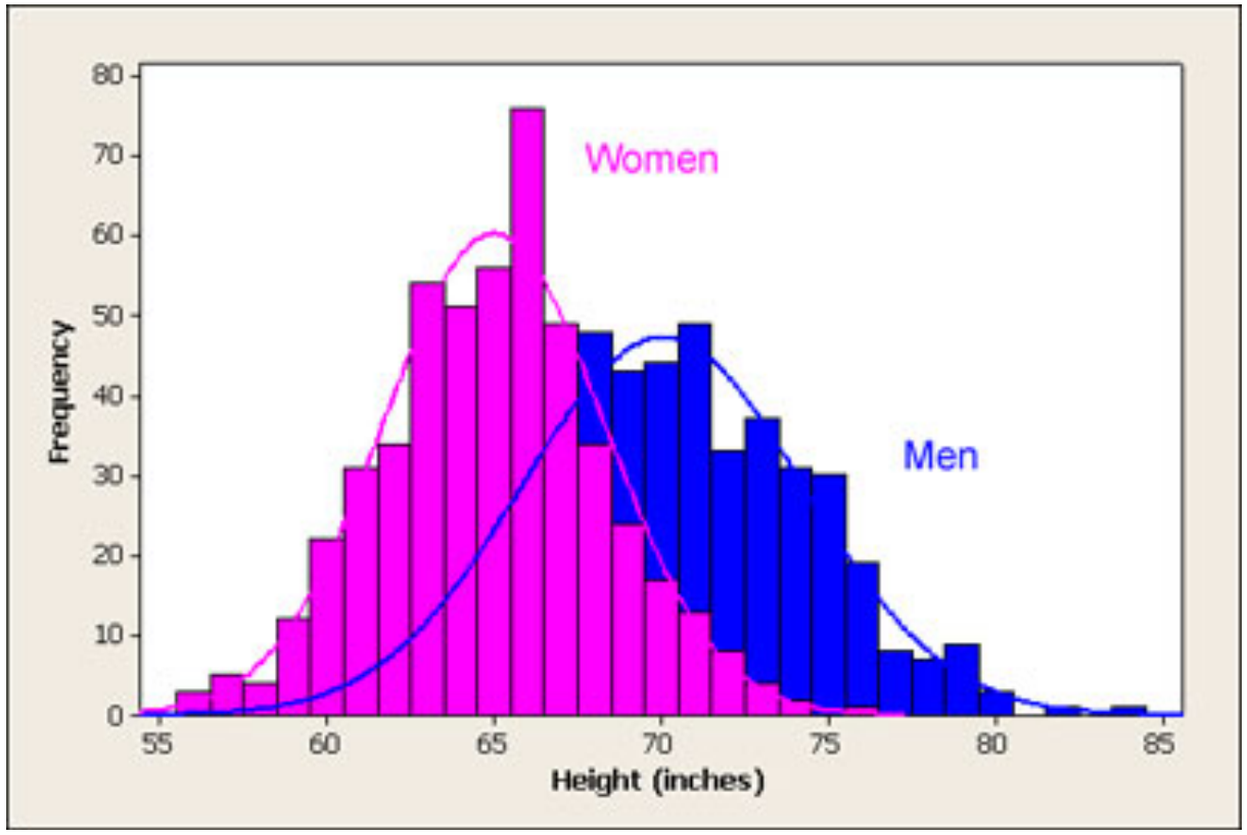
Normal Distribution

Also known as the “bell curve”



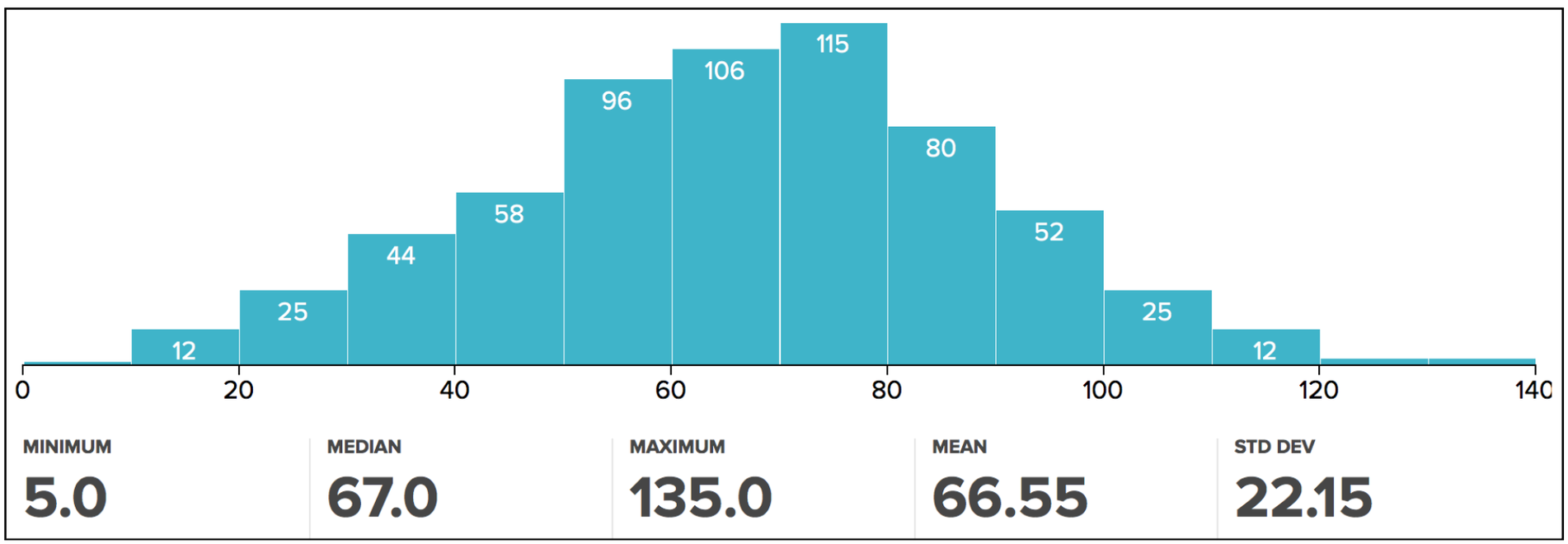
Several natural phenomena (eg. human heights and weights) are distributed normally.

Distributions are unique defined by their parameters.
The normal distribution has two parameters: the mean and the variance.



Histogram of human heights, separated into male and female

Source: http://www.usablestats.com/lessons/normal



Histogram of CS 70 midterm scores

Source: My own misery

Properties of the Normal Distribution

If some quantity has an (approximately) normal distribution, then...

68% of values are within **1 SD** of the mean

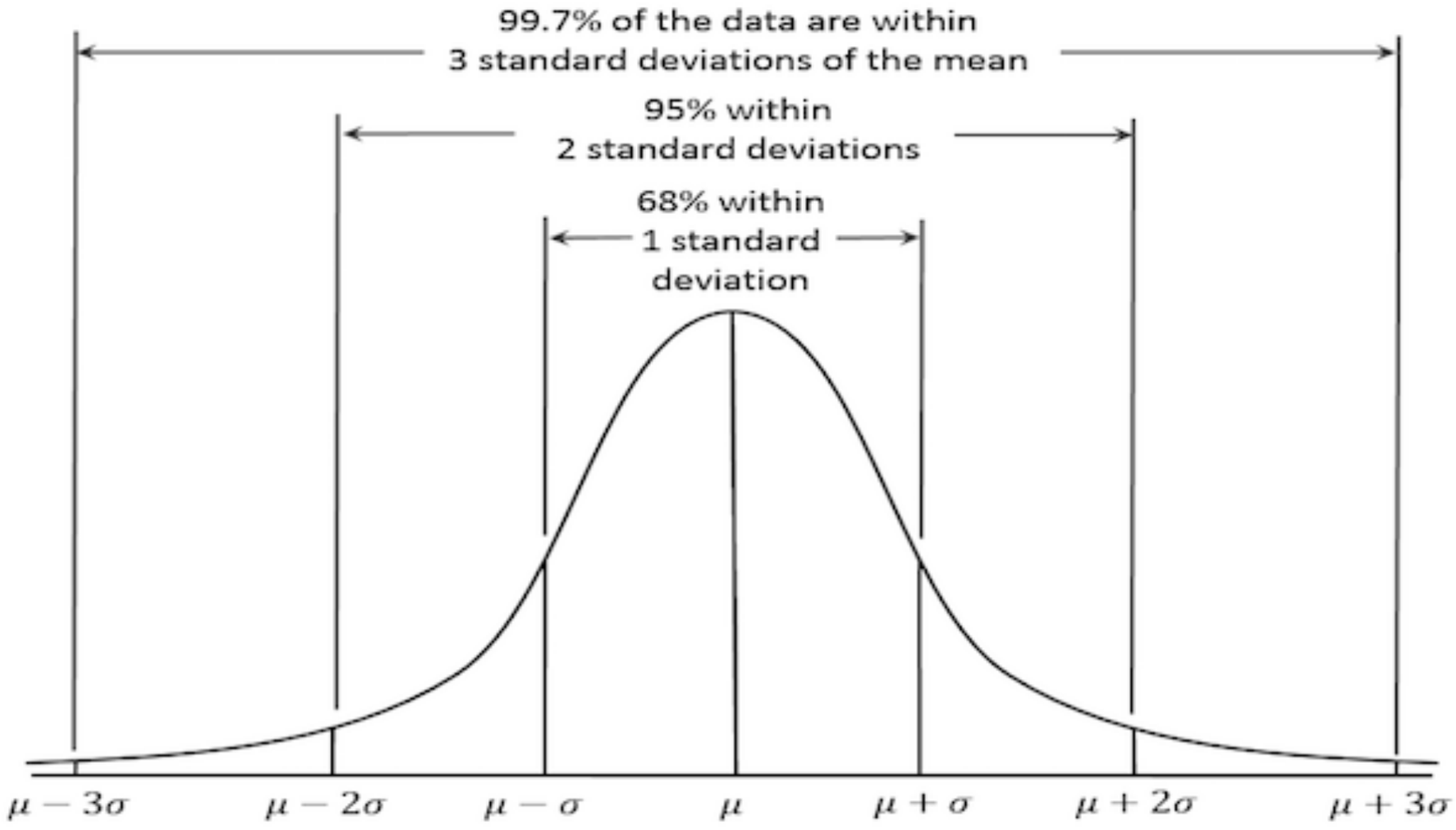
95% of values are within **2 SD** of the mean

99.7% of values are within **3 SD** of the mean

These three facts are crucial to this chapter. Remember them!

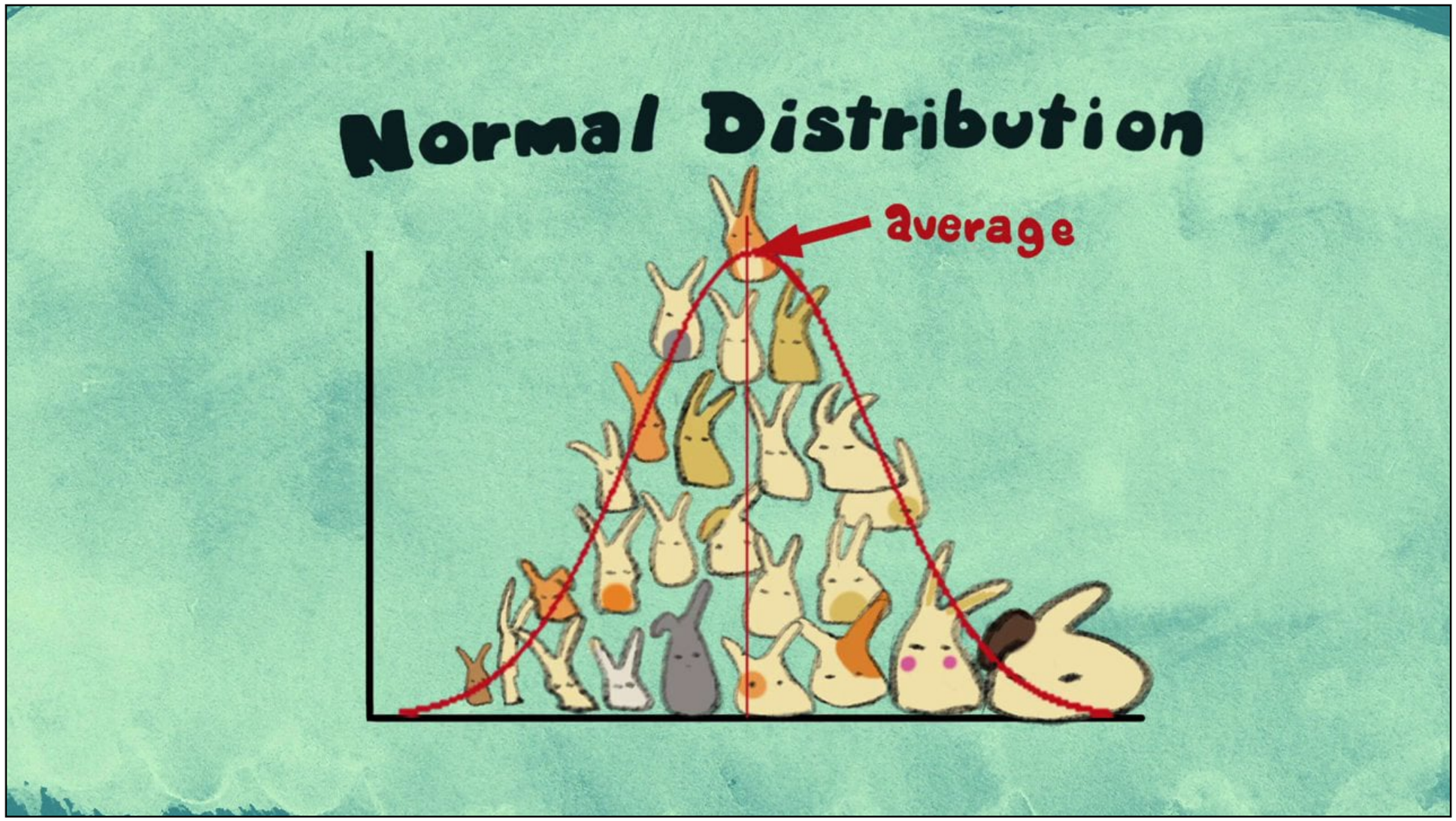
Properties of the Normal Distribution

If some quantity has an (approximately) normal distribution, then...



These three facts are crucial to this chapter. Remember them!

What is the Central Limit Theorem?



This picture appeared when I searched “Central Limit Theorem” so I thought I’d include it.

Central Limit Theorem

If you have a **sample** that is both:

LARGE

AND

DRAWN AT RANDOM WITH REPLACEMENT

THEN

regardless of the distribution of the population, the probability distribution of the sample sum/average is approximately normal.

Why do we care about the CLT?

If we have a large, random sample, and are looking to estimate means, the CLT holds.

If the CLT holds, we can assume several useful properties of the normal distribution.

We'll usually use the fact that we have a good estimate on the proportion of values that lie between **n** standard deviations of the mean. This will help us calculate confidence intervals quickly!

Confidence Intervals with the CLT

Consider the population of 30,000 of Berkeley undergraduates. As per usual, we'd like to find the mean height of all students with a high degree of confidence. Of course, we don't have the time (nor energy) to survey all 30,000 students for their heights, but we can survey a reasonably large sample. Assume the population has a standard deviation of 3 inches.

Suppose we'd like the width of our 95% confidence interval to be 1 inch. How many samples must we take in this case?

Suppose we'd like the width of our 95% confidence interval to be 1 inch. How many samples must we take in this case?

Since we're dealing with a large, random sample, the CLT holds! We know that 95% of values lie within 2 standard deviations of the mean. This means that the **width of our confidence interval will be 4 "standard deviations".**

$$4 \cdot \text{standard deviation of sample means} \leq 1$$

$$4 \cdot \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}} \leq 1$$

$$4 \cdot 3 \leq \sqrt{\text{sample size}}$$

$$\text{sample size} \geq 144$$

Therefore, we must choose a sample size of at least 144 in order to have a 95% CI of 1 inch.

Confidence Intervals with the CLT

Consider the population of 30,000 of Berkeley undergraduates. As per usual, we'd like to find the mean height of all students with a high degree of confidence. Of course, we don't have the time (nor energy) to survey all 30,000 students for their heights, but we can survey a reasonably large sample. Assume the population has a standard deviation of 3 inches.

Now, suppose that we took a sample (with replacement) of 3,650 students, and we found that the sample mean was 58 inches. What distribution will your sample means follow, and what parameters will it have?

Now, suppose that we took a sample (with replacement) of 3,650 students, and we found that the sample mean was 58 inches. What distribution will your sample means follow, and what parameters will it have?

Since we have a large, random sample, we can apply the CLT.

The sample means will follow the **normal distribution**, with parameters **mean (μ) = 58** and **standard deviation = 3/sqrt(3650)**, by the relationship between the SD of sample means and the SD of populations.

What is this concept of “standard units?”

Standard Units

If you look at graphs of several normal distributions with different means and SDs, you'll probably notice something.

Their shape is more or less the same.

In order to make calculations convenient and standardized, we always convert our values to **standard units** when dealing with normal distributions. (You'll see later we convert to standard units for other purposes, as well.)

In statistics, we deal with the **standard normal distribution**, that is, the normal distribution with **mean 0** and **standard deviation 1**.

How to Convert to Standard Units

Given a set of values $x_1, x_2, \ldots x_n$, with mean $\boldsymbol{\mu}$ and standard deviation $\mathbf{s}$, we convert each $\mathbf{x_i}$ to standard units by the rule

$$z_i = (x_i - \mu) / s$$

By converting to standard units, we're effectively changing each value from its raw quantity to the **number of standard deviations it is away from the mean**. Standard units have no units!

Sometimes, in other contexts, this is referred to as the **z-score**. We won't use that terminology in this class, but now you know!

How to Convert to Standard Units

For example, let's convert the following five weights to standard units.

120 lbs, 160 lbs, 230 lbs, 140 lbs, 180 lbs

μ = (120 + 160 + 230 + 140 + 180 lbs) / 5 = 166 lbs

`s = np.std(make_array(120, 160, 230, 140, 180))`

≈ 38 lbs

original	120 lb	160 lb	230 lb	140 lb	180 lb
std units	-1.21	-0.16	1.68	-0.68	0.37

Correlation

Correlation

Correlation gives us a way to describe the **linear relationship** between two sets of data. More specifically, the **correlation coefficient** '**r**' is a number between -1 and 1.

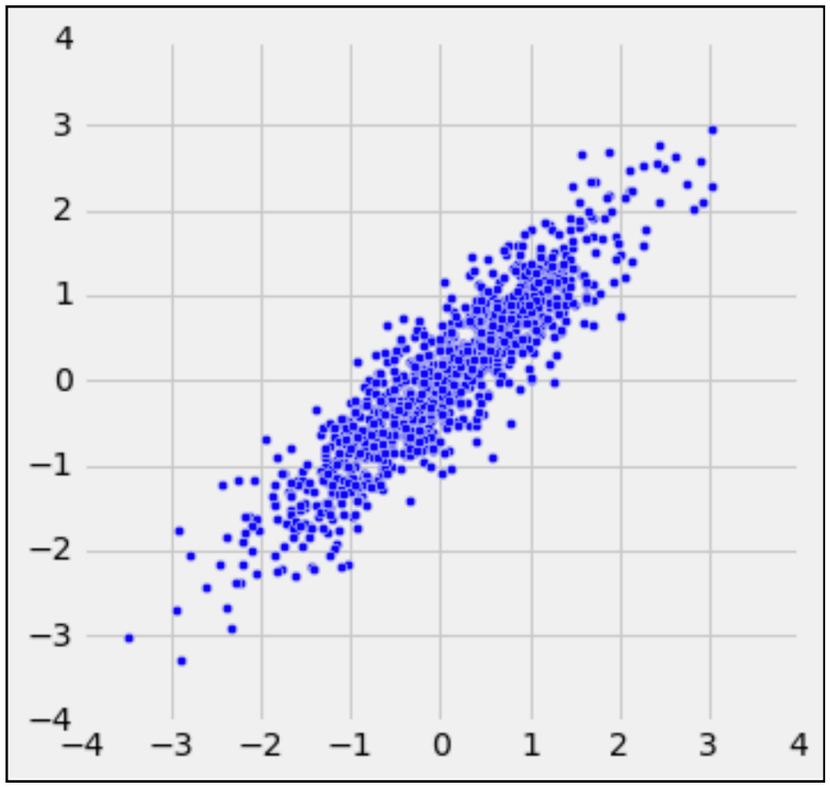
If **r = 1**, then the two sets of data lie exactly on some line with a positive slope.

If **r = 0**, then the two sets of data are uncorrelated.

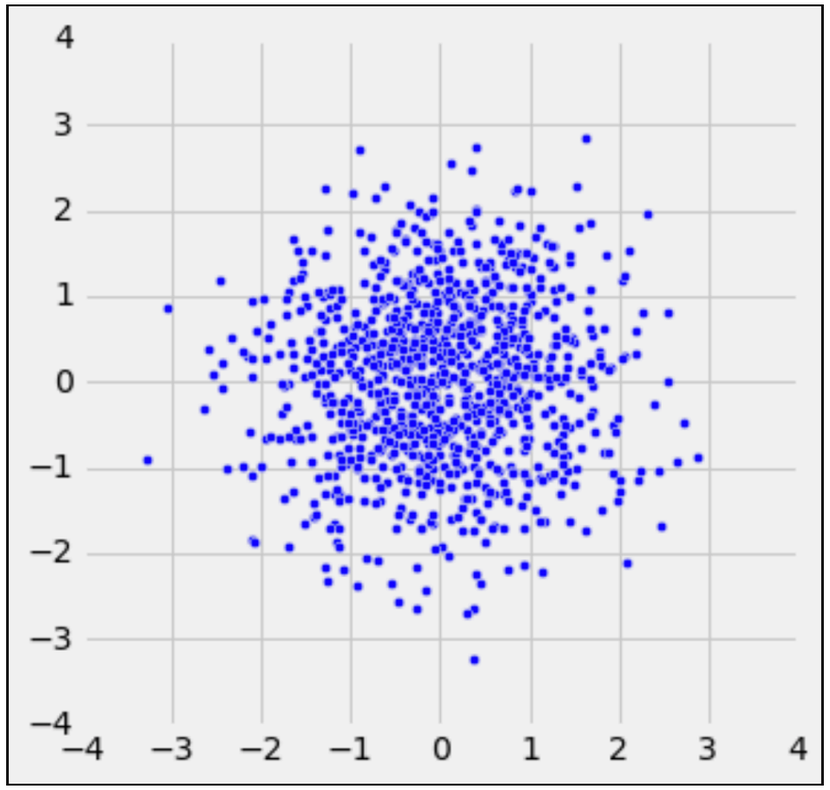
If **r = -1**, then the two sets of data lie exactly on some line with a negative slope.

Values of r in between these describe relationships that are weakly/ strongly positively/negatively correlated, depending on the value.

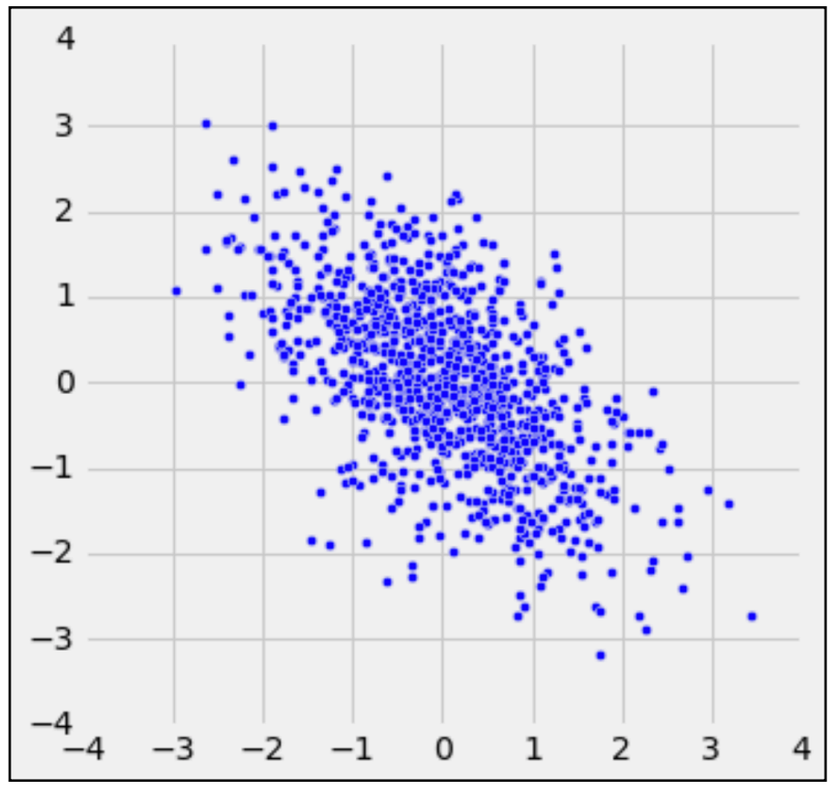
Various Values of the Correlation Coefficient



r = 0.9

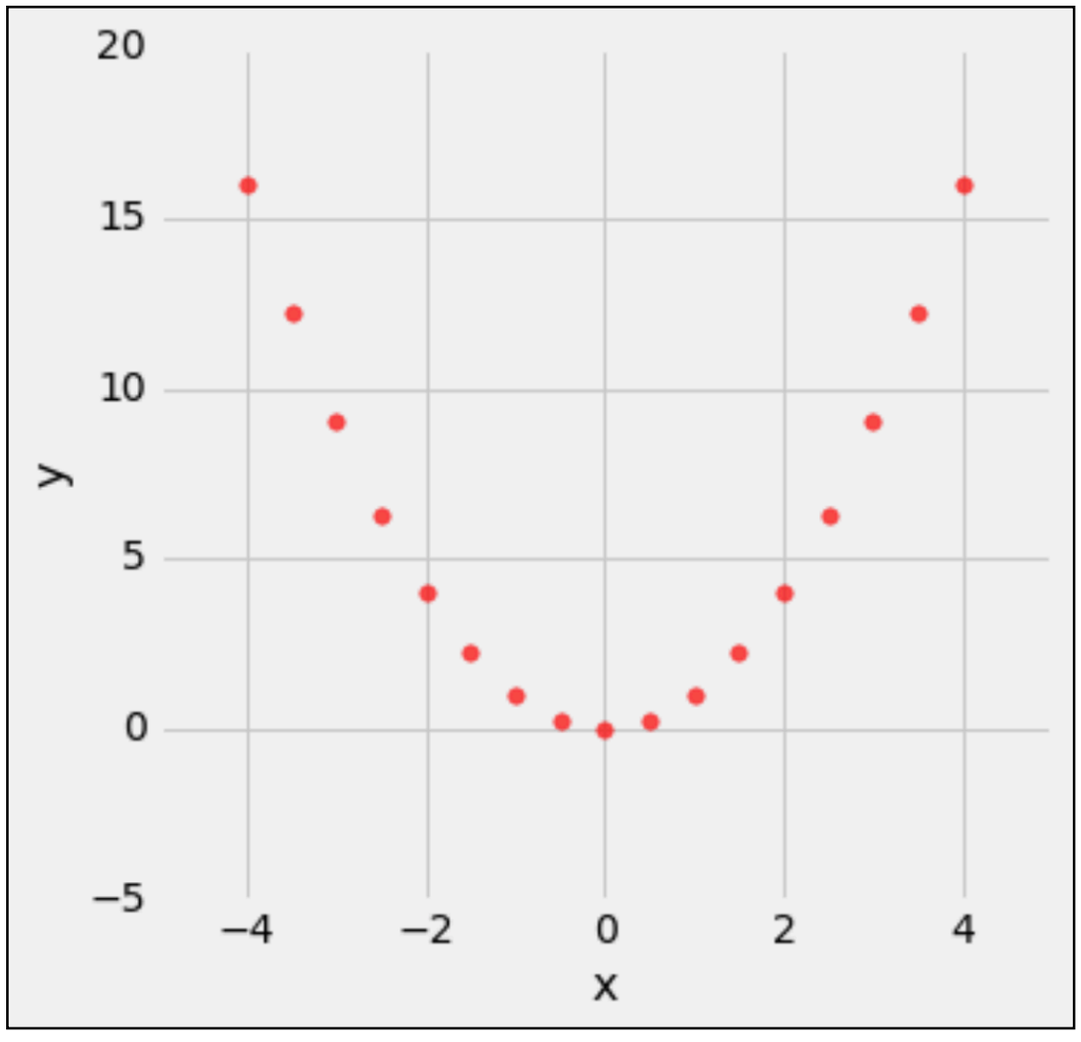


r = 0



r = -0.55

Various Values of the Correlation Coefficient



r = 0

With the **correlation coefficient**, we only look at one type of relationship – linear.

Even though the graph to the left plots points on the equation **y = x**2**, that isn't a linear relationship, so **r** = 0.

Calculation and Properties of r

r is usually calculated using a relatively complicated expression. With the functions we have at our disposal, though, it's relatively simple.

r is the **mean of the product of x and y**, when both are measured in standard units.

We first convert to standard units before calculating **r** because correlation only looks at the shape of the distribution of two variables, not the values themselves. By converting to std. units, we can compare values that are at different scales easily.

Calculation and Properties of r

r is a unit-less value. This is because it is based off of values in standard units, which themselves are unit-less values.

Remember, **r** just looks at the shape of the distribution. Therefore, you can add, subtract, multiply or divide all of your values by any constant, and the value of **r** won't change, so long as you make this change to each of your values.

`correlation(<tbl name>, <col A name>, <col B name>)`

calculates **r** for columns A, B in the specified table. This definition isn't built-in to **`datascience`**, but this is often how it will be implemented in assignments.

```
import datascience
import numpy as np
```

```
x = make_array(1, 2, 3, 4, 5)
y = make_array(3, 2, 12, 15, 25)
```

```
def std_units(x):
    return (x - np.mean(x)) / np.std(x)

def correlation(x, y):
    return np.mean(std_units(x) * std_units(y))
```

```
correlation(t.column(0), t.column(1))
```

```
0.953716136009612 8
```

```
correlation(3 * t.column(0) + 14, 100 * t.column(1) + 5)
```

```
0.953716136009612 8
```

`std_units` converts arrays to standard units, and **`correlation`** finds **r** for any two given arrays. Notice that when we shift the values of **x** and **y**, the value of **r** doesn't change.

Let **“heights”** and **“weights”** be columns in the table **people**.

Is **correlation(people, “heights”, “weights”)** always equal to **correlation(people, “weights”, “heights”)**?

Yes. **r** is the average of the product of **X, Y** in standard units, and multiplication is commutative (a • b = b • a).