BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 20 – Standard Deviation Normal Distributions Correlation

Announcements

- Project 2:
  - due Monday 04/18
- No Lab this week
- Homework 7 - Confidence Intervals, Resampling, the Bootstrap, and the Central Limit Theorem
  - Due Thursday 04/07
- Dropping 1 homeworks and 1 lab
- Speak up!!
  - More posts on ed-stem – great job!

Data Science in this course

- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference
  - Make reliable conclusions about the world
  - Statistics is useful
- Prediction
  - Informed guesses about unseen data

Center & Spread

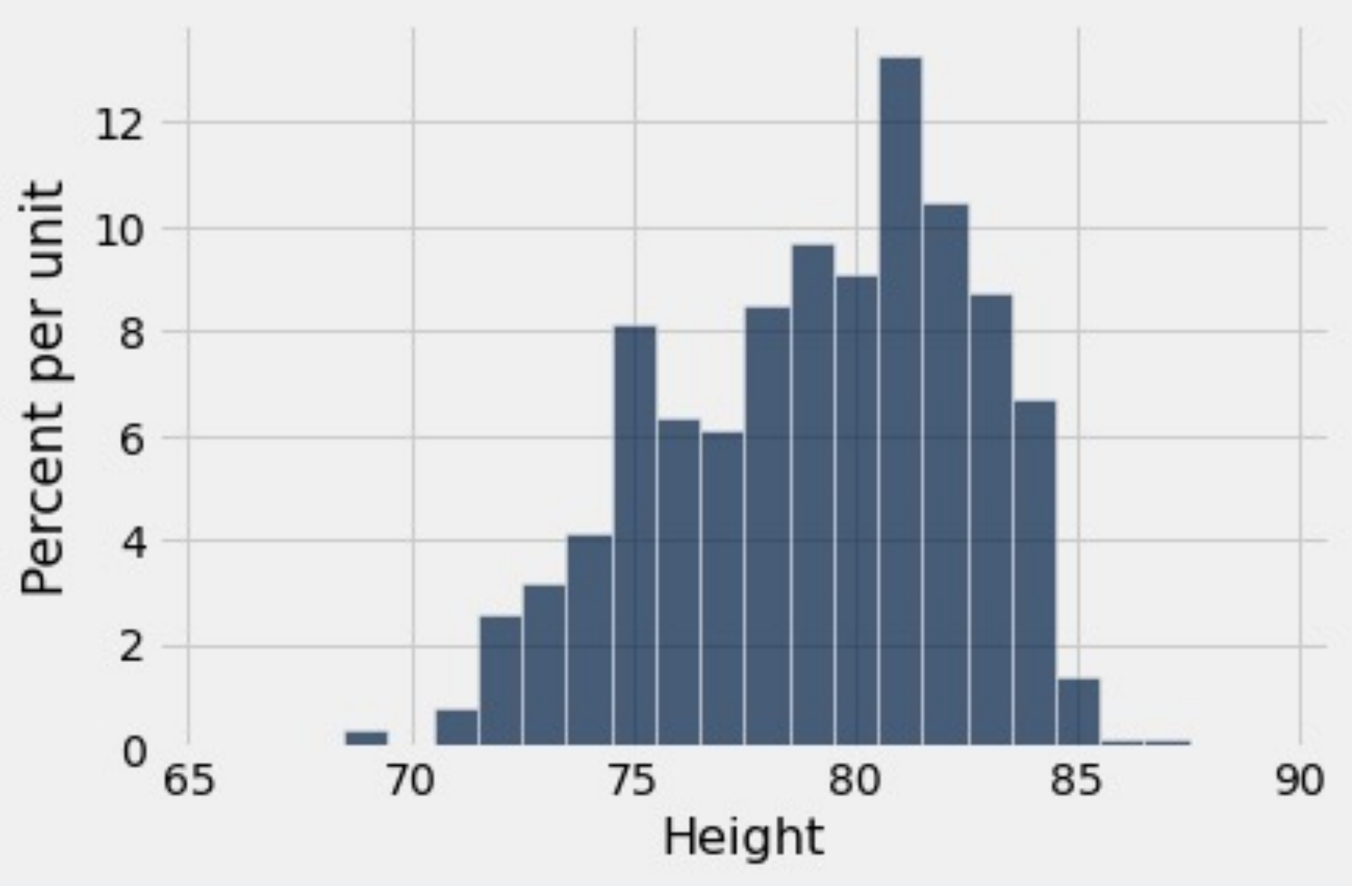
Questions/Goals

- How can we quantify natural concepts like "center" and "variability"?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

Average and Median

Question

- Which is bigger, median or mean?



Comparing Mean and Median

- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from the median in the direction of the tail.

Standard Deviation

Defining Variability

- **Plan A:** “biggest value - smallest value”
  - Doesn’t tell us much about the shape of the distribution
  - In other words, doesn’t tell us where most values are
- **Plan B**:
  - Measure variability around the mean
  - Need to figure out a way to quantify this

How far from the average?

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average

Steps: 5 4 3 2 1

- SD has the same units as the data

Why use Standard Deviation

- There are two main reasons.
- **The first reason:**
  - No matter what the shape of the distribution, the bulk of the data are in the range “average plus or minus a few SDs”
- **The second reason:**
  - Relation with the bellshaped curve
  - Discuss this later

Q: How big are most values?

A: Chebyshev’s Inequality

How big are most values?

No matter what the shape of the distribution,
the bulk of the data are in the range "average ± a few SDs"

Chebyshev's Inequality

No matter what the shape of the distribution,
the proportion of values in the range "average ± *z* SDs" is

at least 1 - 1/*z*2

Chebyshev's Bounds

Range	Proportion

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)

Chebyshev's Bounds

Range	Proportion
average ± 2 SDs	at least 1 - 1/4 (75%)
average ± 3 SDs	at least 1 - 1/9 (88.888...%)
average ± 4 SDs	at least 1 - 1/16 (93.75%)
average ± 5 SDs	at least 1 - 1/25 (96%)

True no matter what the distribution looks like

Understanding HW05 Results

Statistics:
Minimum: 7.5
Maximum: 29.0
Mean: 24.55
Median: 25.0
Standard Deviation: 3.96

- At least 50% of the class had scores between 20.59 and 28.51

- At least 75% of the class had scores between 16.62 and 32.47

Standard Units

Standard Units

- How many SDs above average?
- ***z* = (value - average)/SD**
  - Negative z: value below average
  - Positive z: value above average
  - z = 0: value equal to average
- When values are in standard units: average = 0, SD = 1
- Chebyshev: At least 96% of the values of *z* are between -5 and 5

Question

What whole numbers are closest to

(1) Average age

(2) The SD of ages

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

Answers

(1) Average age is close to 27 (standard unit here is close to 0)

(2) The SD is about 6 years (standard unit at 33 is close to 1. 33 – 27 = 6)

Age in Years	Age in Standard Units
27	-0.0392546
33	0.992496
28	0.132704
23	-0.727088
25	-0.383171
33	0.992496
23	-0.727088
25	-0.383171
30	0.476621
27	-0.0392546

The SD and the Histogram

- Usually, it's not easy to estimate the SD by looking at a histogram.

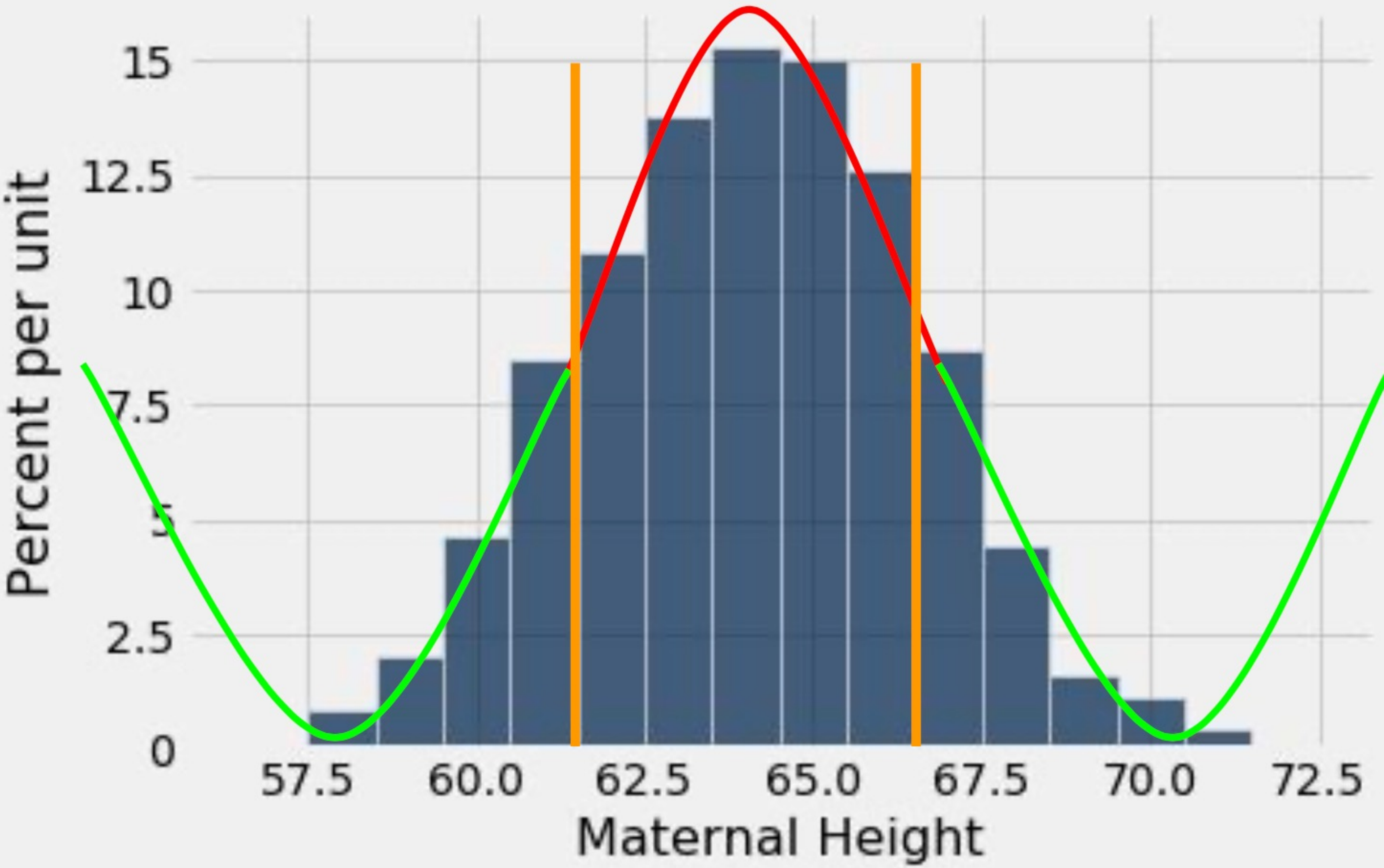
- But if the histogram has a bell shape, then you can

The SD and Bell Shaped Curves

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

Points of Inflection



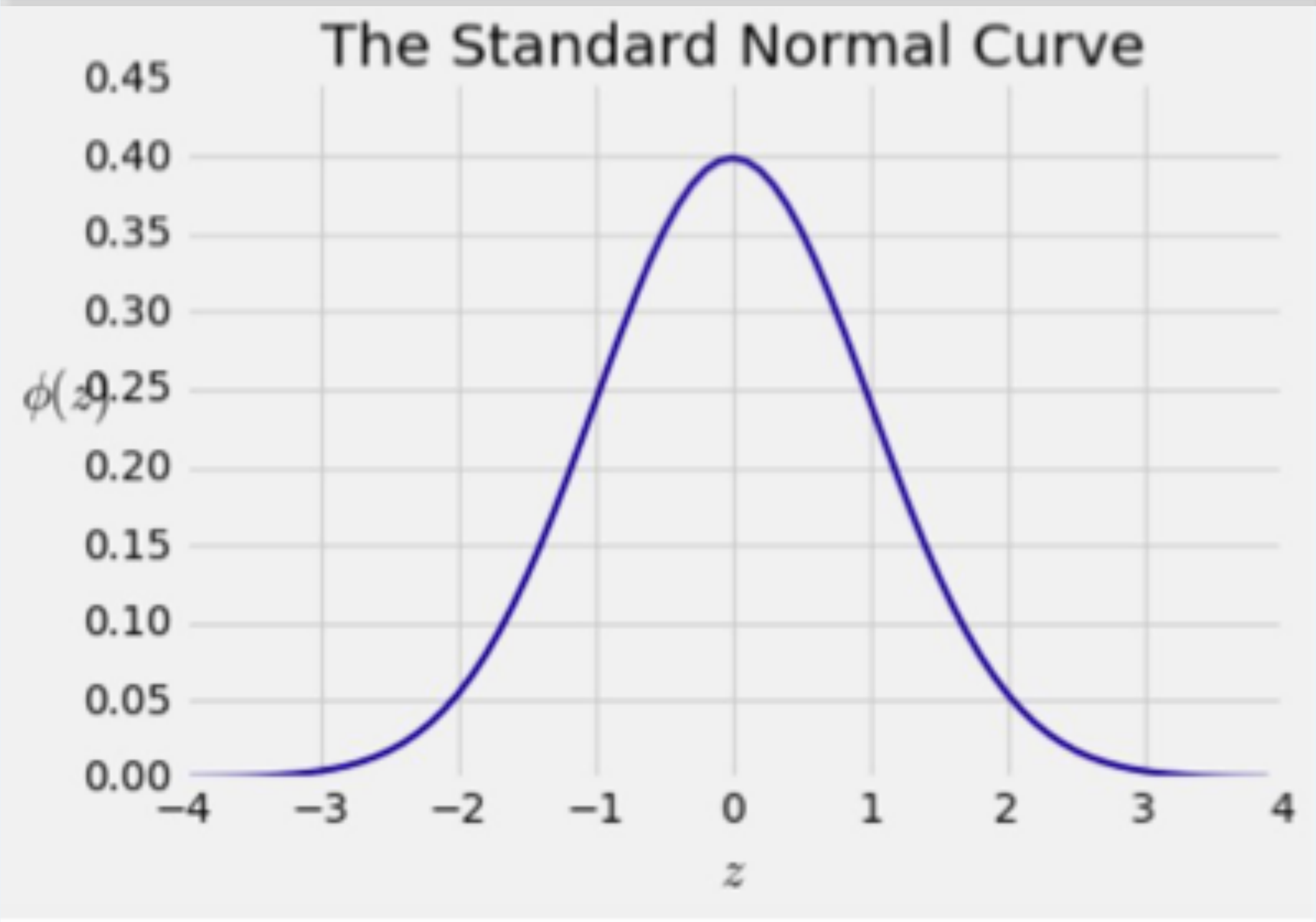
Normal Distribution

Standard Normal Curve

Equation for the normal curve

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

Bell Curve


The Standard Normal Curve
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
φ(z)
−4
−3
−2
−1
0
1
2
3
4
z

How Big are Most of the Values

No matter what the shape of the distribution,

the bulk of the data are in the range “average ± a few SDs”

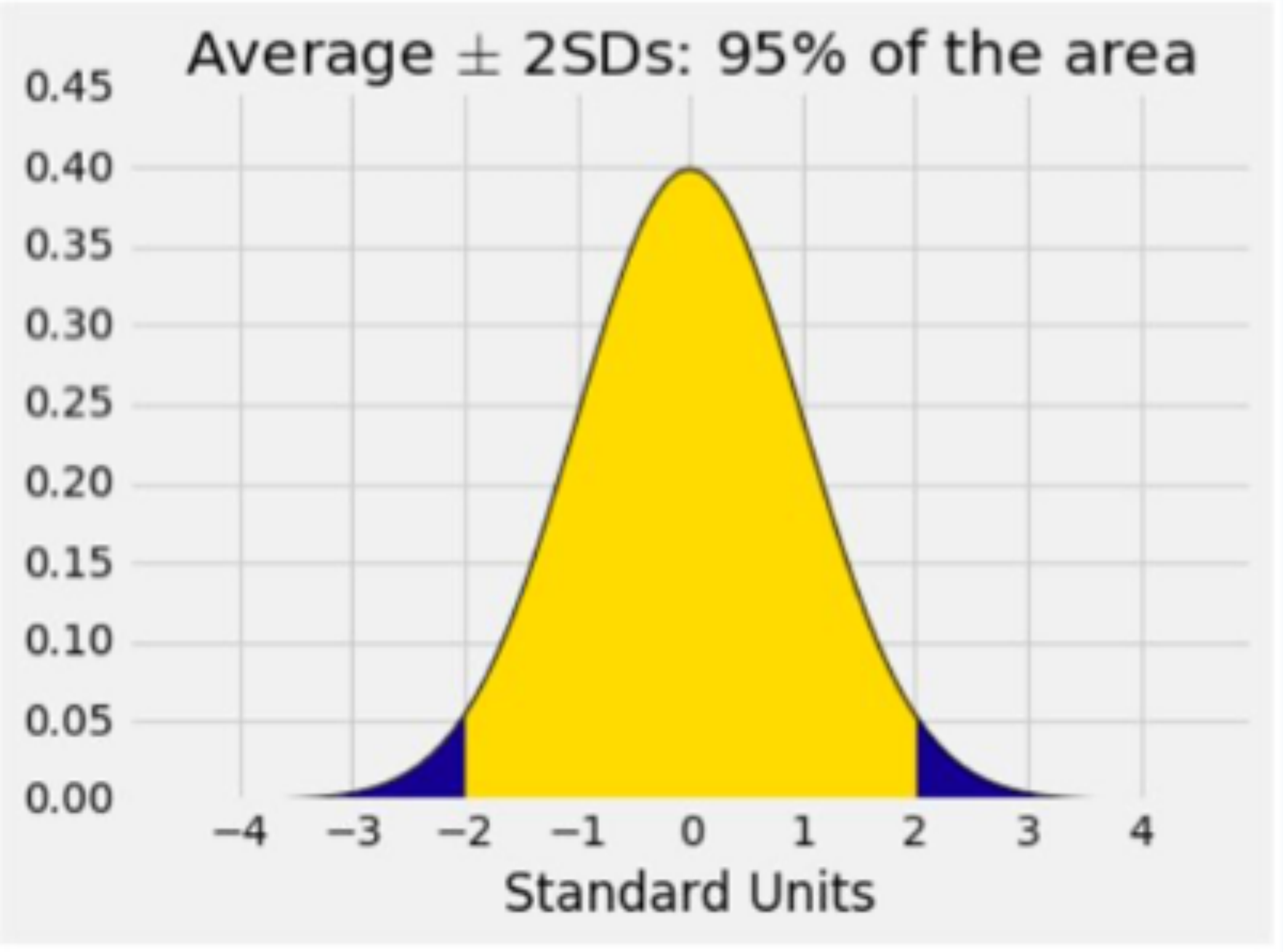
If a histogram is bell-shaped, then

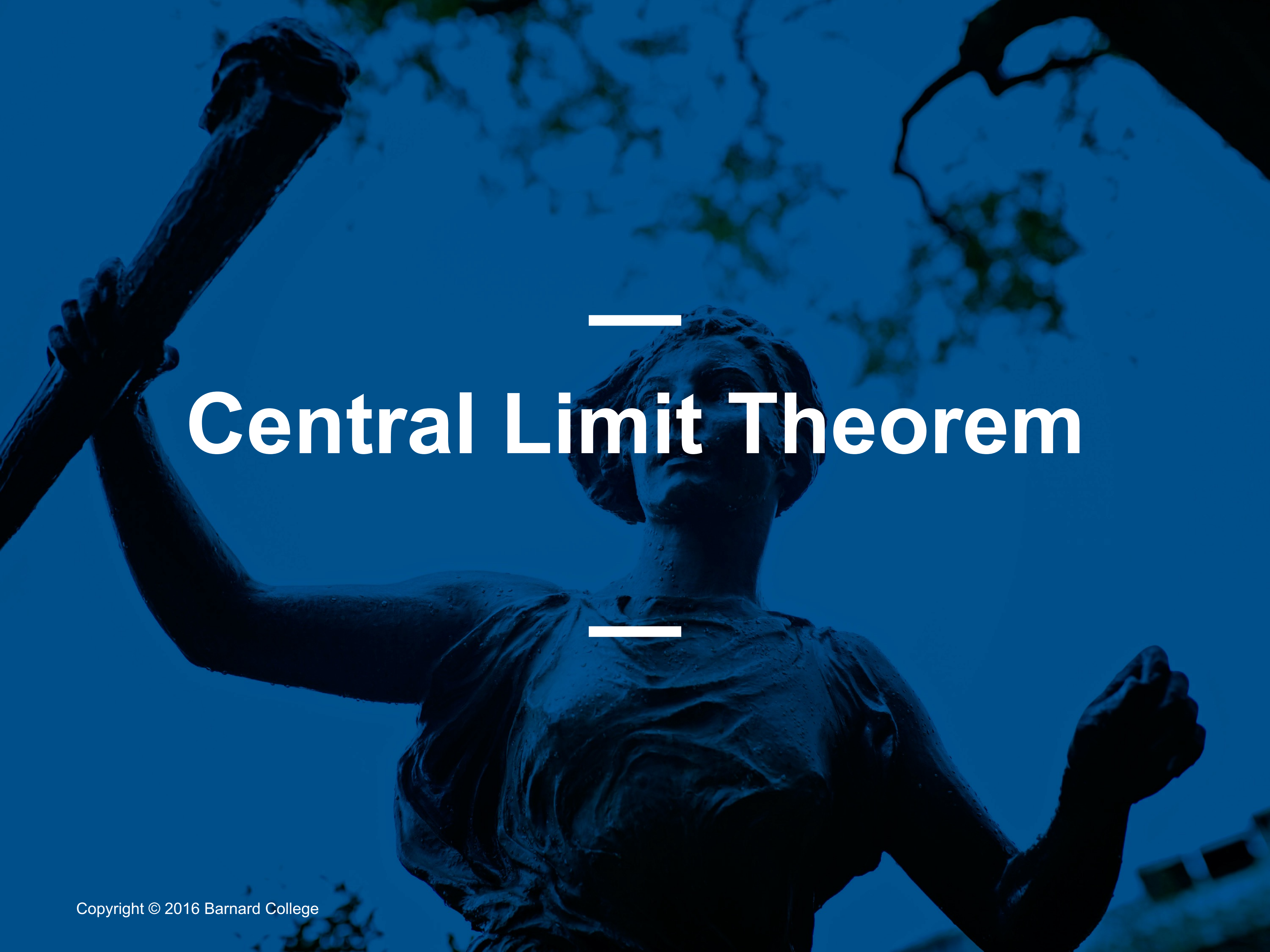
- Almost all of the data are in the range “average ± 3 SDs

Bounds and Approximations

Percent in Range	All Distributions	Normal Distributions
Average +- 1 SD	At least 0%	About 68%
Average +- 2 SDs	At least 75%	About 95%
Average +- 3 SDs	At least 88.888…%	About 99.73%

A "Central" Area





Central Limit Theorem

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample sum (or the sample average) is roughly normal

Sample Average

- We often only have a sample
- We care about sample averages because they estimate population averages.
- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.
- CLT allows us to make inferences based on averages of random samples

Correlation

Prediction

- To predict the value of a variable:
  - Identify (measurable) attributes that are associated with that variable
  - Describe the relation between the attributes and the variable you want to predict
  - Then, use the relation to predict the value of a variable

Visualizing Two Numerical Variables

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

Visualize, then quantify

The Correlation Coefficient *r*

- Measures **linear** association
- Based on standard units
- $-1 \le r \le 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*

Definition of *r*

Correlation Coefficient (r) =

average of product of standard(x) and standard(y)

Steps: 4 3 2 1

Measures how clustered the scattered data are around a straight line

Operations that leave *r* unchanged

R is not affected by:

- Changing the units of the measurement of the data
  - Because *r* is based on standard units
- Which variable is plotted on the x- and y-axes
  - Because the product of standard units is the same

Interpreting *r*

Causal Conclusion

Be careful …

- Correlation measures linear association
- Association doesn't imply causation
- Two variables might be correlated, but that doesn't mean one causes the other

Nonlinearity and Outliers

Both can affect correlation

- Draw a scatter plot before computing *r*

Ecological Correlation

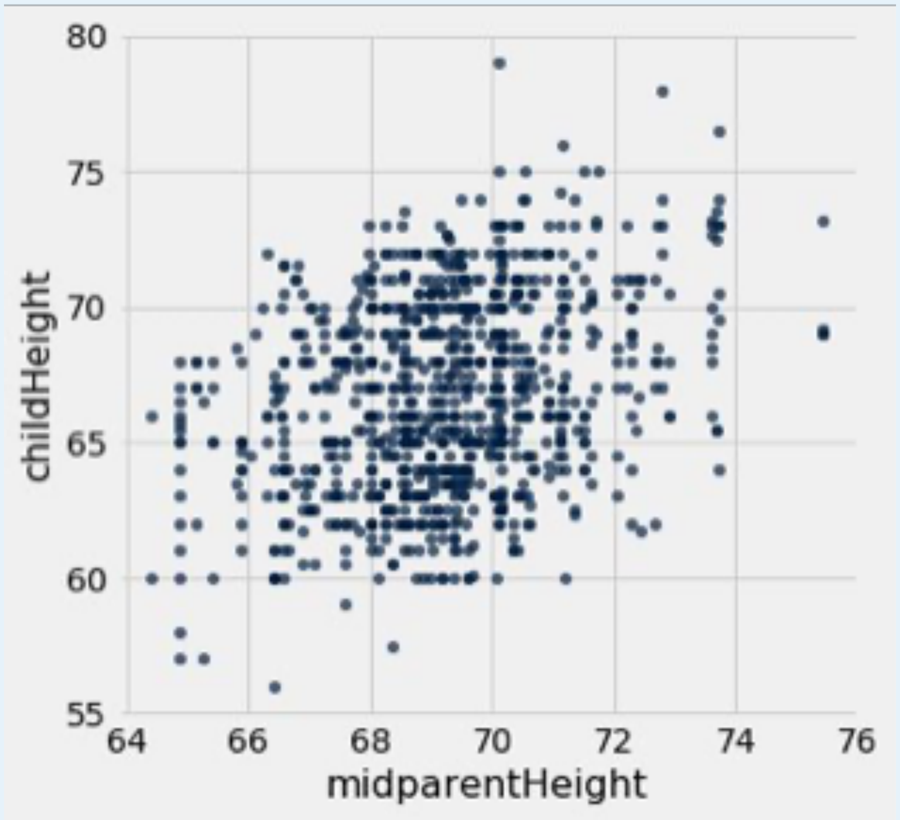
- Correlations based on groups or aggregated data
- Can be misleading:
  - For example, they can be artificially high

Prediction

Guess the future

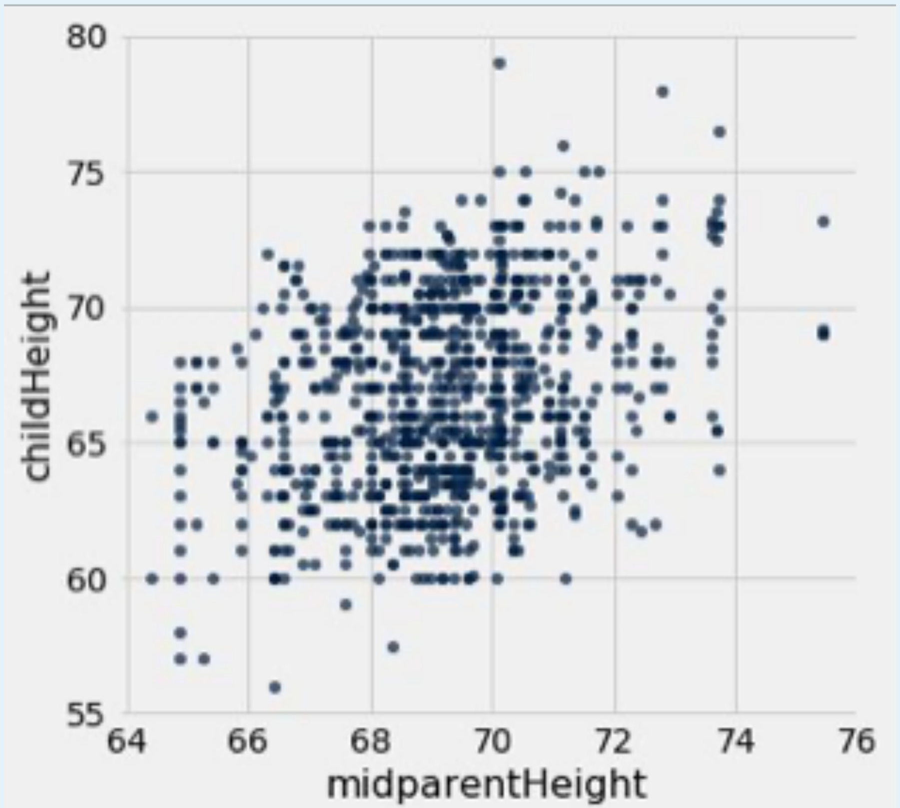
- Based on incomplete information
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

Galton's Heights



Goal: Predict the height of a new child, based on that child's midparent height

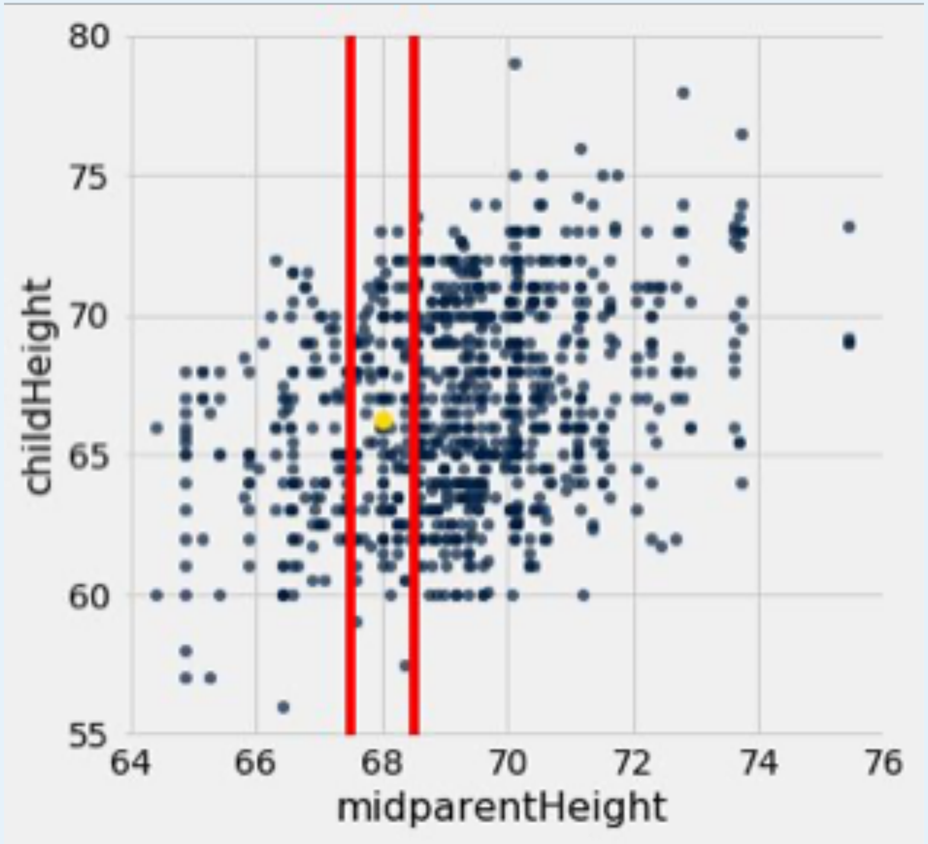
Galton's Heights



How can we predict a child's height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

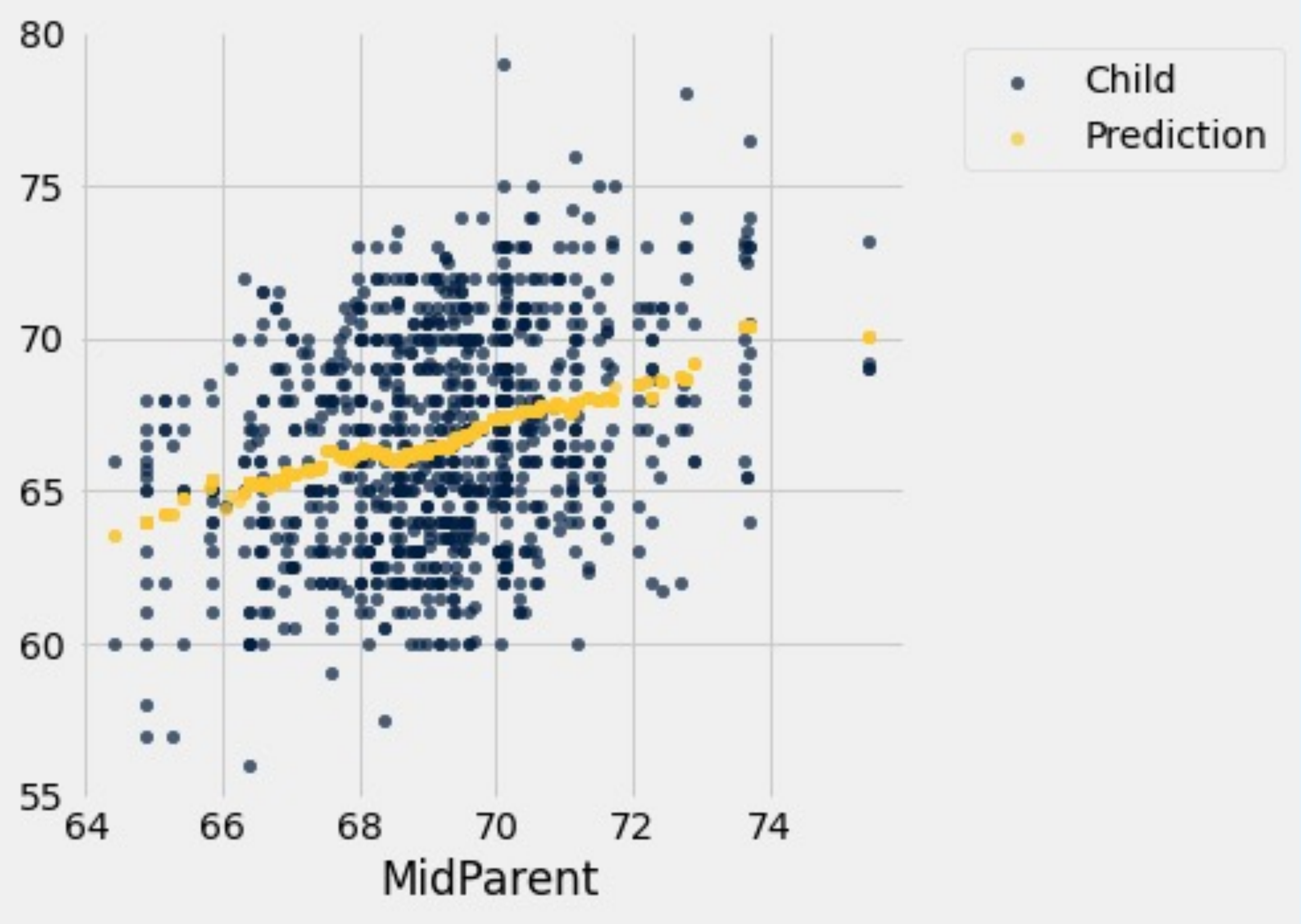
Galton’s Heights



How can we predict a child’s height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

Predicted Heights



Graph of Average

For each x value, the prediction is the average of the y values in its nearby group.

The graph of these predictions is the

graph of averages

If the association between x and y is linear, then points in the graph of averages tend to fall on a line. The line is called the **regression line**

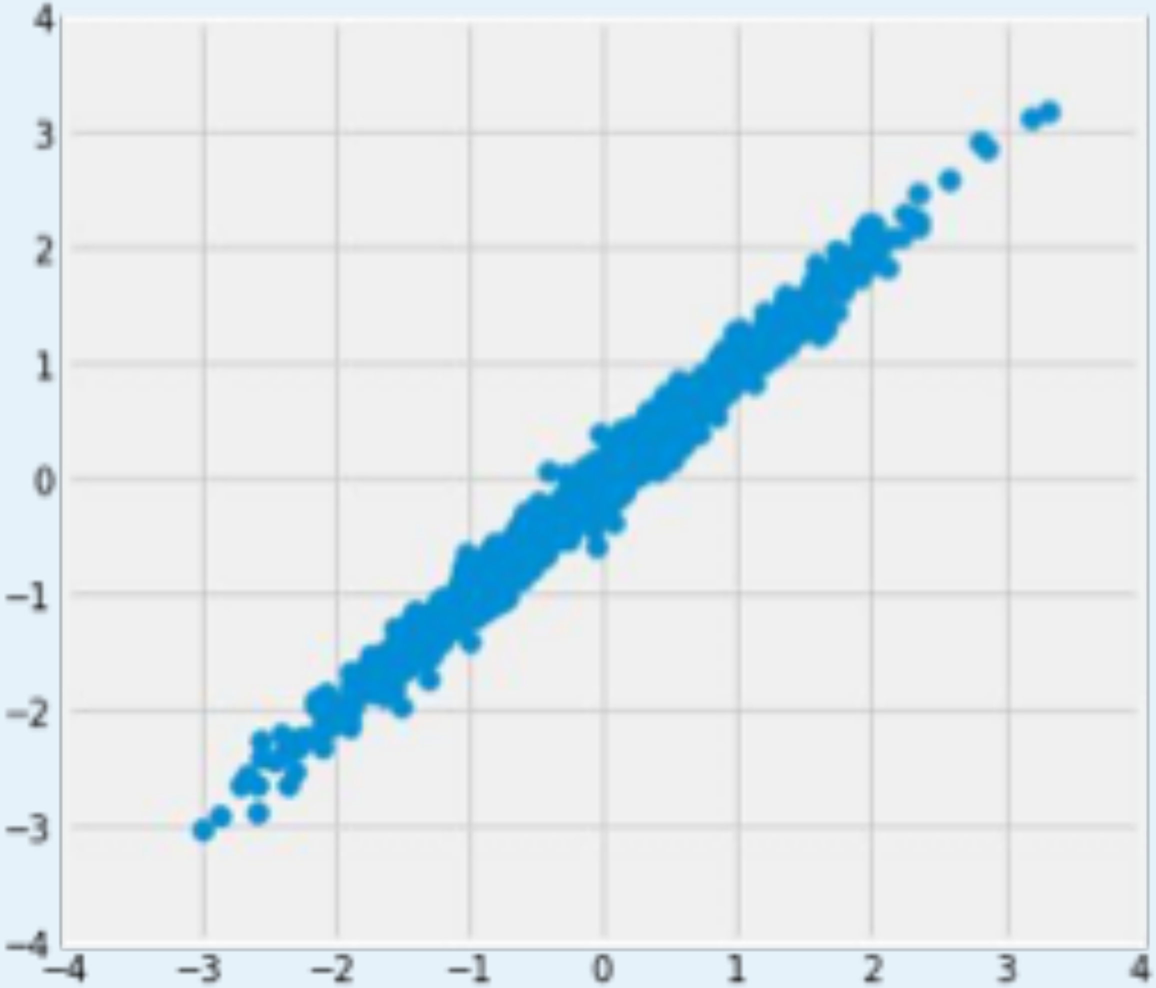
Nearest Neighbor Regression

A method for predicting a numerical y, given a value of x:

- Identify the group of points where the values of x are close to the given value
- The prediction is the average of the y values for the group

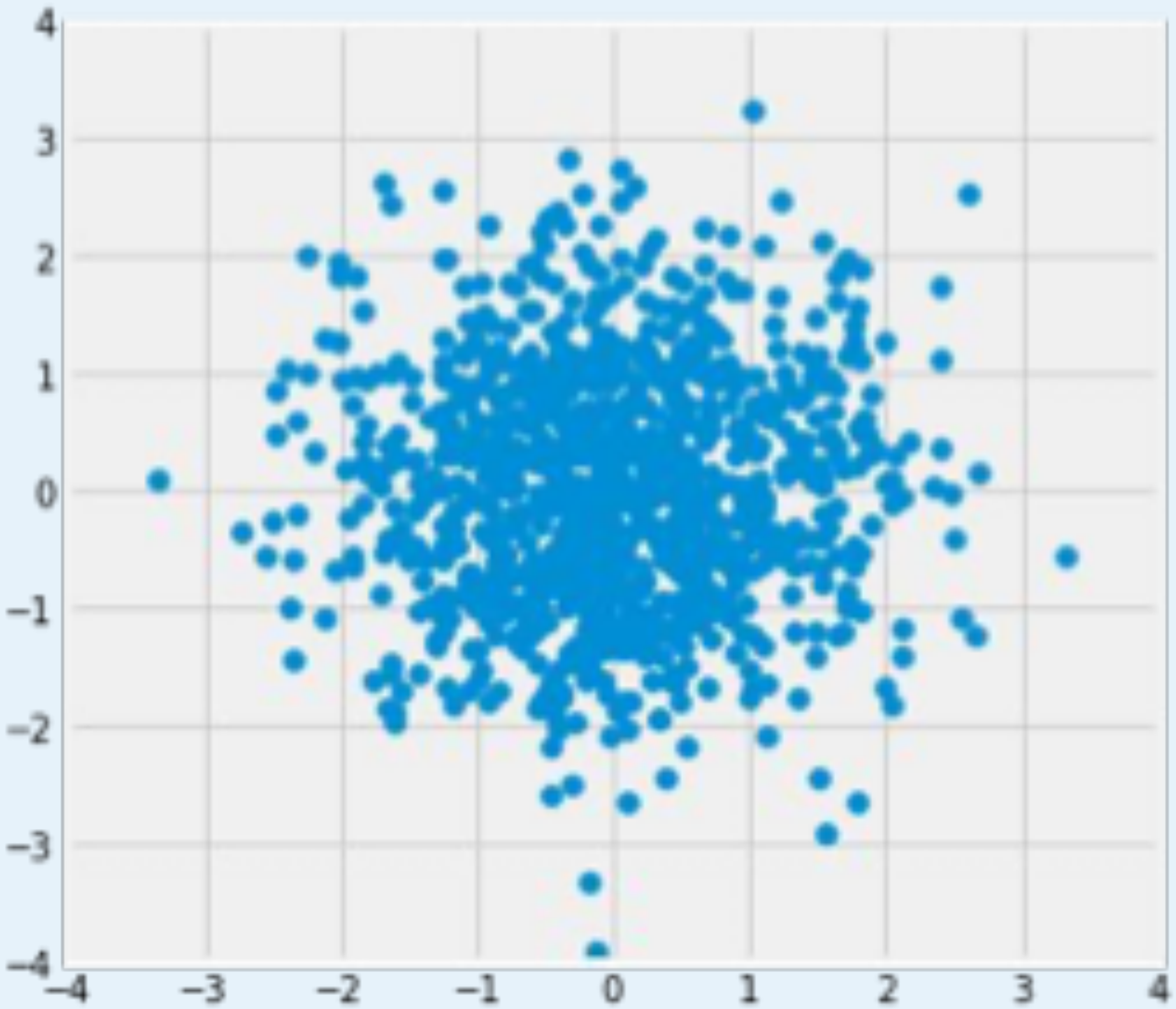
Linear Regression

Where is the prediction line?



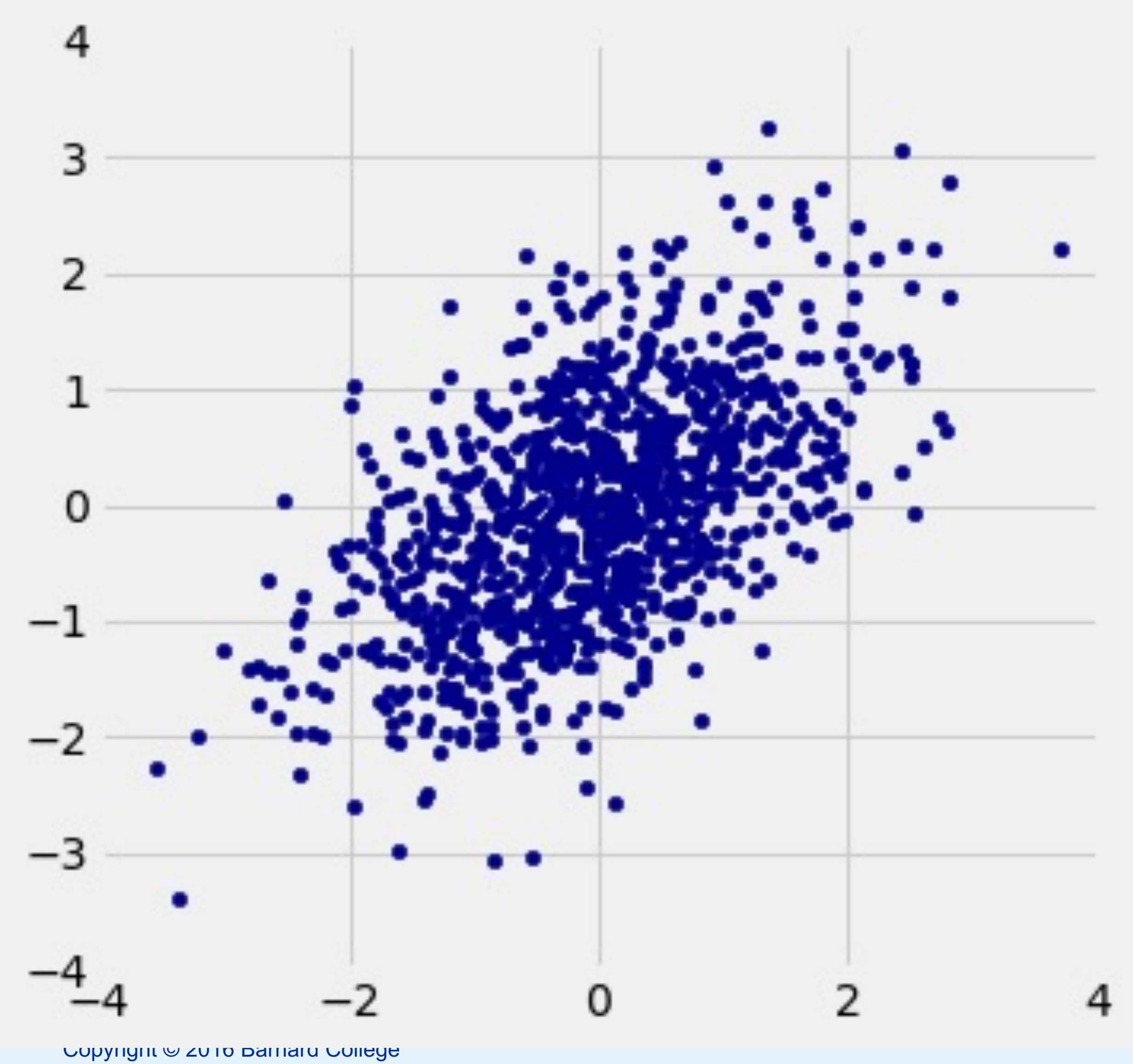
$r = 0.99$

Where is the prediction line?



$r = 0.0$

Where is the prediction line?



$r = 0.5$

Identifying the Line

- If the scatter plot is oval shaped, then we can spot an important feature of the regression line

Linear Regression

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y’s)

On average,

y deviates from 0 less than x deviates from 0

$$y_{su} = r \times x_{su}$$

Slope and Intercept

Regression Line Equation

In original units, the regression line has this equation:

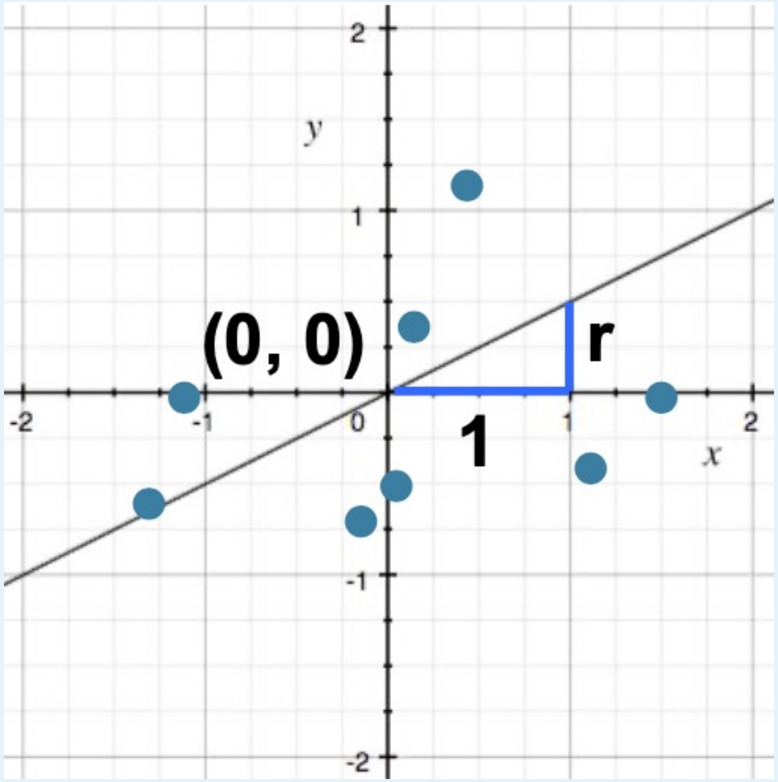
$$\frac{estimate\ of\ y\ - mean(y)}{SD\ of\ y} = r \times \frac{given\ x\ - mean(x)}{SD\ of\ x}$$

Lines can be expressed by *slope* & *intercept*

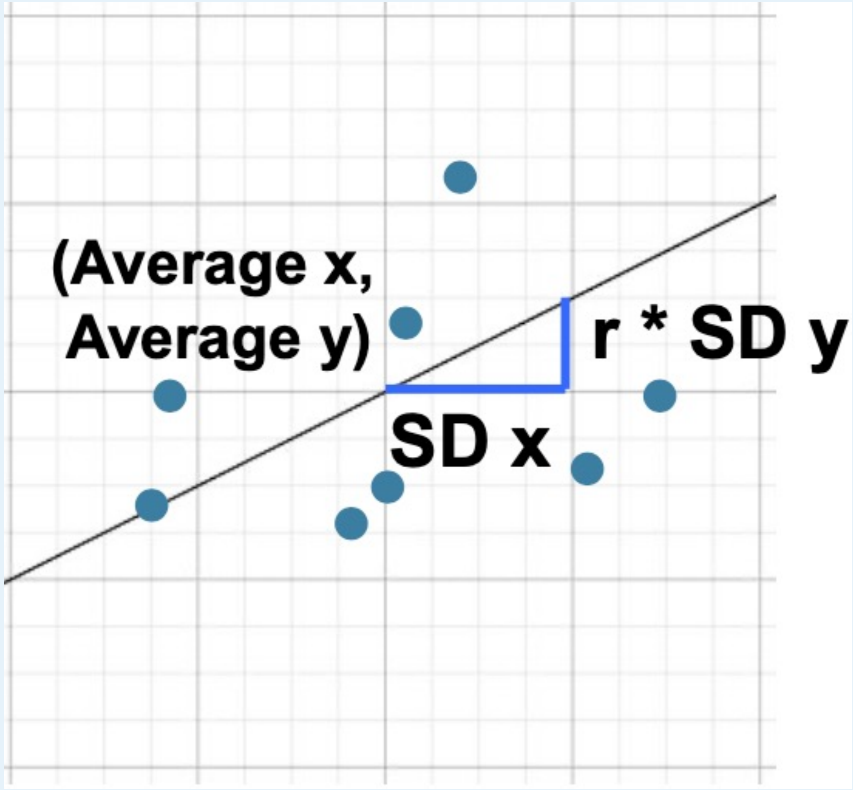
$$y = slope \times x + intercept$$

Regression Line

Standard Units



Original Unites



Slope and Intercept

$$estimate\ of\ y = slope\ * x + intercept$$

$$\boldsymbol{slope\ of\ the\ regression\ line}$$

$$r\ * \frac{SD\ of\ y}{SD\ of\ x}$$

$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

$$mean(y) - slope \times mean(x)$$