

# CompSci 190: Lecture 10: Predictions

Jeff Forbes

March 26, 2019

# 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
- **Standard deviation (SD):** Measures roughly how far the data are from their mean
  - Root mean square of deviations from average
  - SD has the same units as the data

# 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  $\pm$  a few SDs”

## Chebyshev's Inequality

No matter what the shape of the distribution,  
the proportion of values in the range “average  $\pm z$  SDs” is

at least  $1 - 1/z^2$

---

# Chebyshev's Bounds

---

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

**No matter what the distribution looks like**

---

# Standard Units

---

- How many SDs above average?
- $z = (\text{value} - \text{mean})/\text{SD}$ 
  - 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

# Discussion Question

Find whole numbers  
that are close to:

- (a) the average age
- (a) the SD of the ages

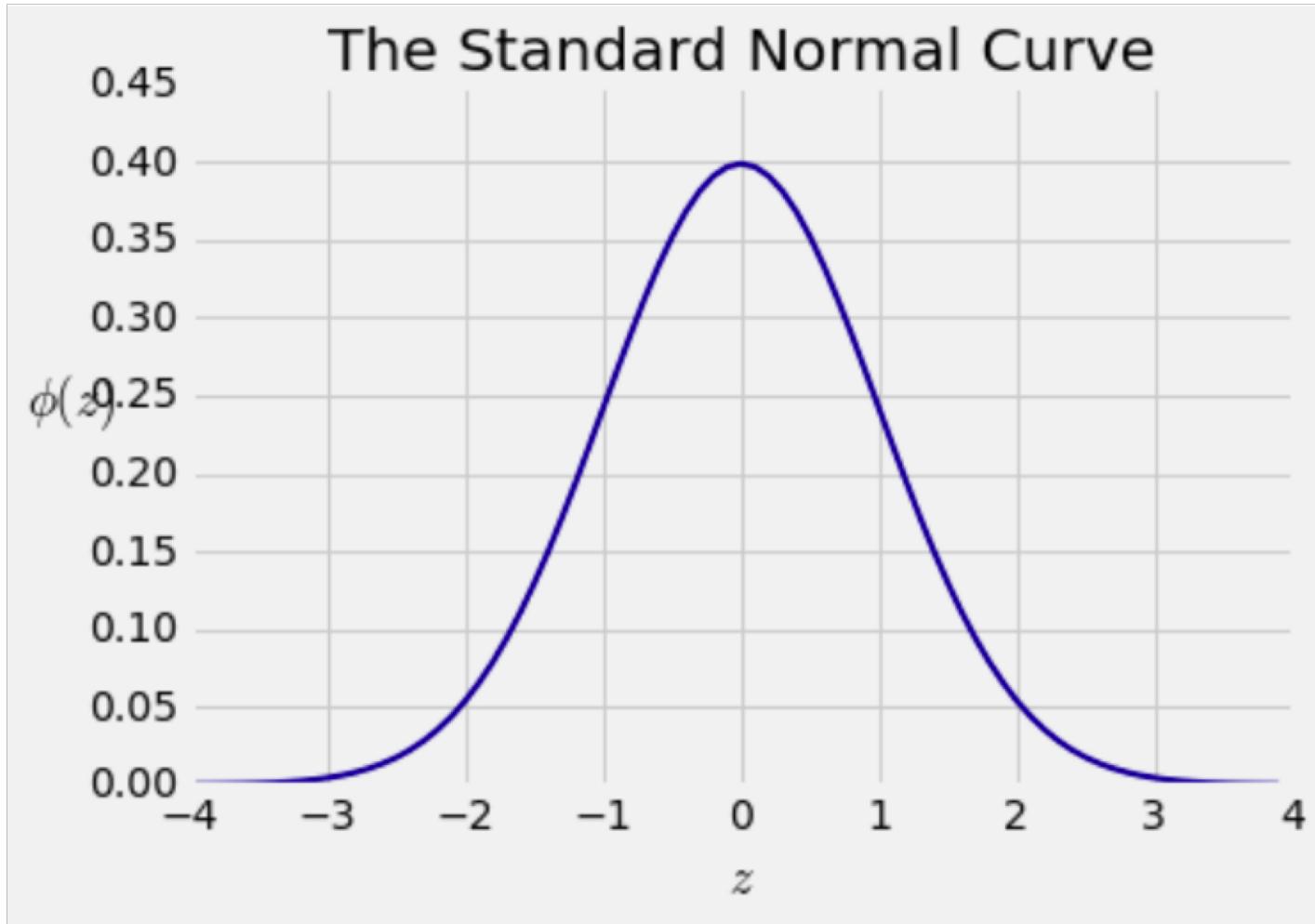
(Demo)

<b>Age in Years</b>	<b>Age in Standard Units</b>
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

... (1164 rows omitted)

# Bell Curve

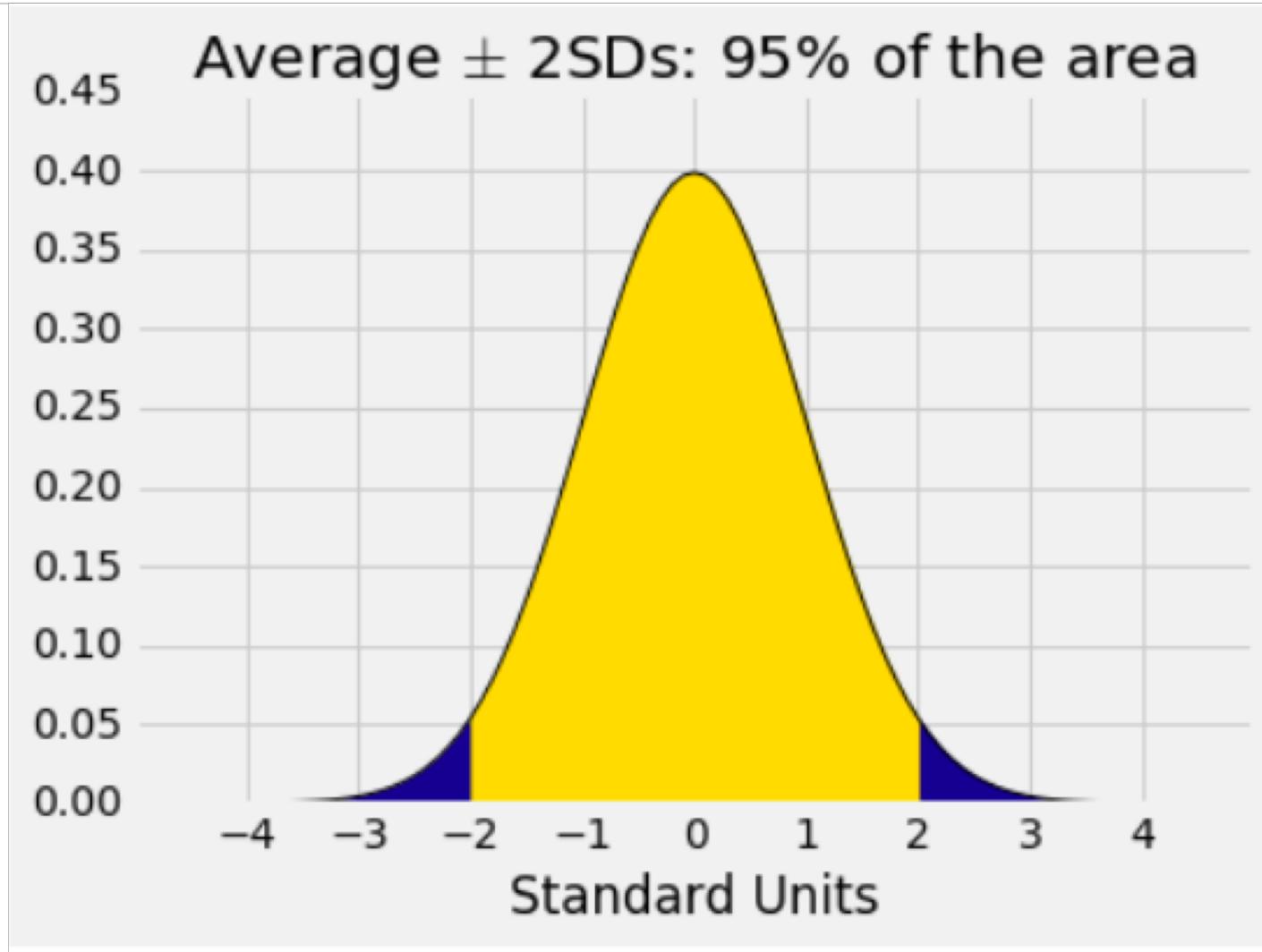
---



# Bounds and Normal Approximations

<b>Percent in Range</b>	<b>All Distributions</b>	<b>Normal Distribution</b>
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

# A “Central” Area



(Demo)

# Prediction Problems

---

- Predicting one characteristic based on another:
    - Given my height, how tall will I be next year?
    - Given my height, how tall will my kid be as an adult?
    - Given my height, how much will I sleep tonight?
  - Characteristics of an example: known and unknown
  - For some sample, we know all the characteristics
-

# Relation Between Two Variables

---

- Association
- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape”
  - Linear
  - Non-linear

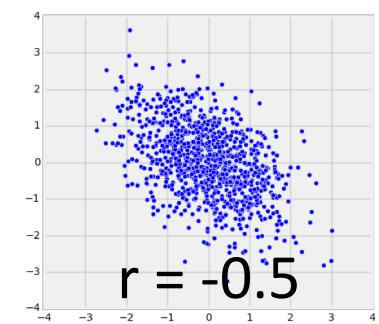
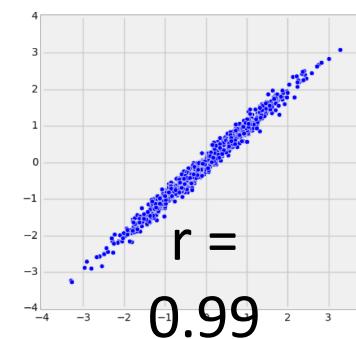
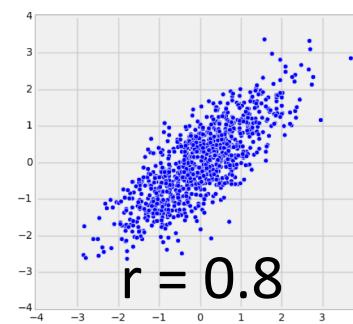
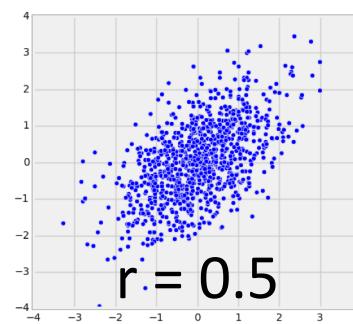
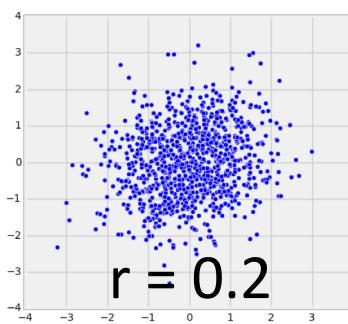
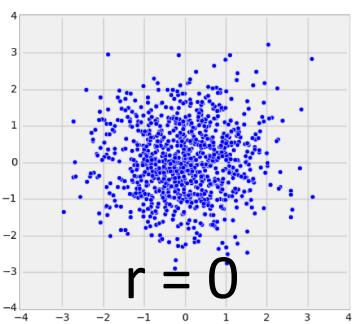
Visualize then quantify

(Demo)

# The Correlation Coefficient $r$

---

- Measures linear association
- Based on standard units
- $-1 \leq r \leq 1$ 
  - $r = 1$ : 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

x in  
*standard  
units*

times

y in  
*standard  
units*

Measures how clustered the scatter is around a straight line

---

# Interpreting $r$

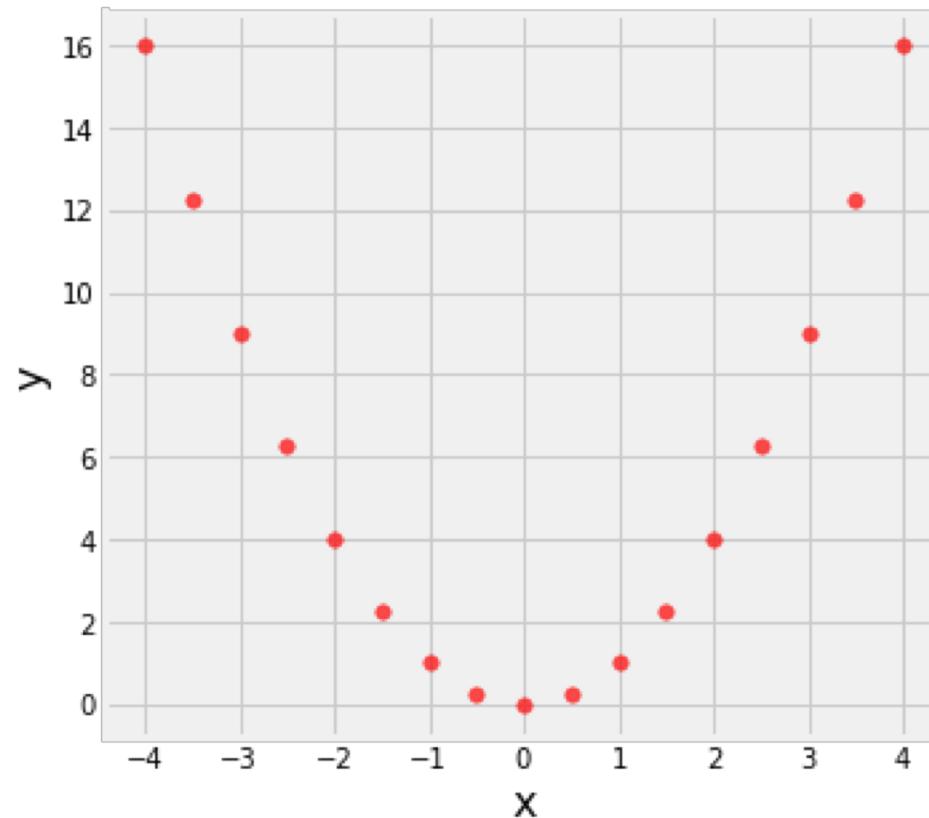
---

Don't jump to conclusions about causality

# Interpreting $r$

---

Watch out for non-linearity.

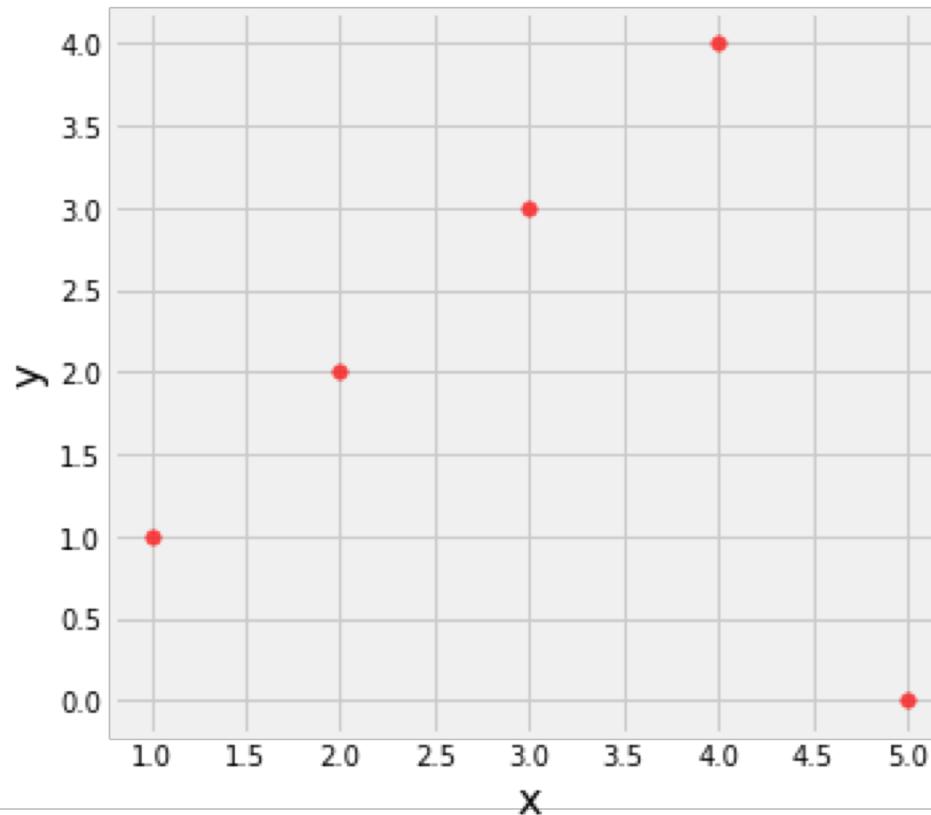


$$r = 0.0$$

# Interpreting $r$

---

Watch out for outliers.

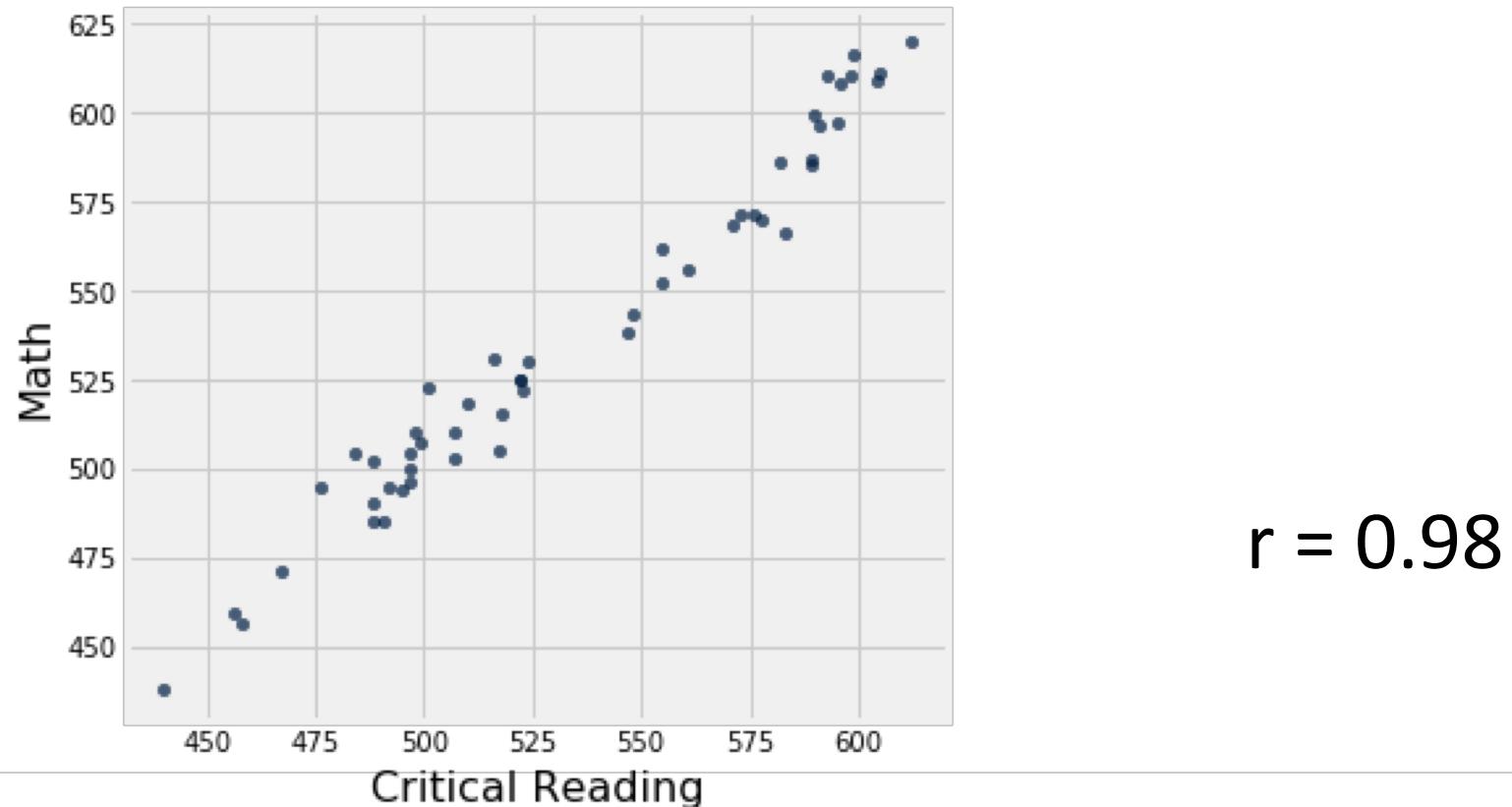


$$r = 0.0$$

# Interpreting $r$

---

Watch out for ecological correlations, based on aggregates or averaged data.



# Lab 8

---



J. Michael Shaughnessy and Maxine Pfannkuch

## **Is Old Faithful?**

**Statistical Thinking: A Story of Variation and Prediction**

- 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 of the sample average) is roughly bell-shaped**