CompSci 190: Lecture 12: Predictions

Jeff Forbes

November 19, 2018

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.

http://bit.ly/FoDS-f18-1119-0

How Far from the Mean?

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

No matter what the distribution looks like

Standard Units

- How many SDs above average?
- z = (value - mean)/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

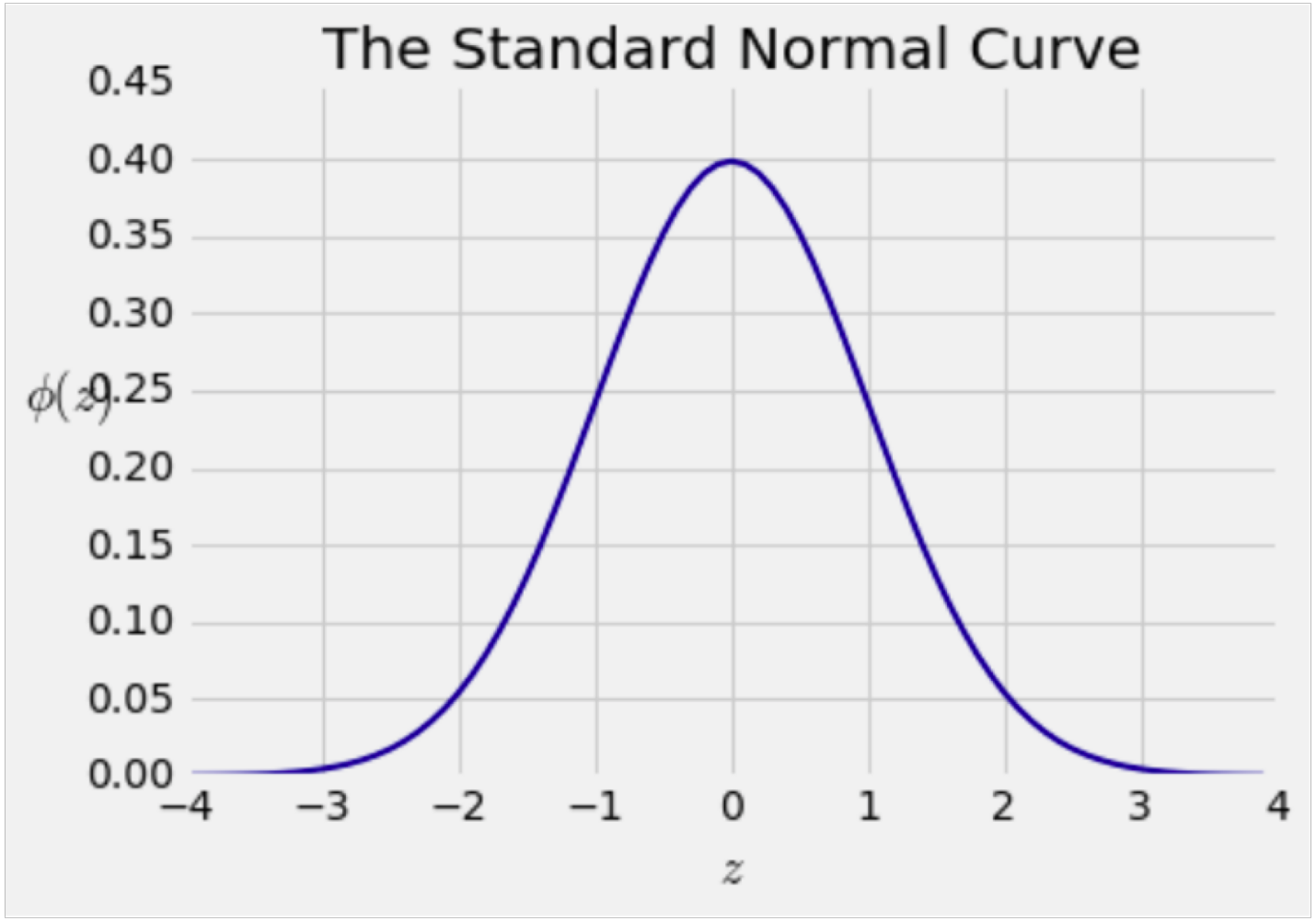
(a) the SD of the ages

(Demo)

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

... (1164 rows omitted)

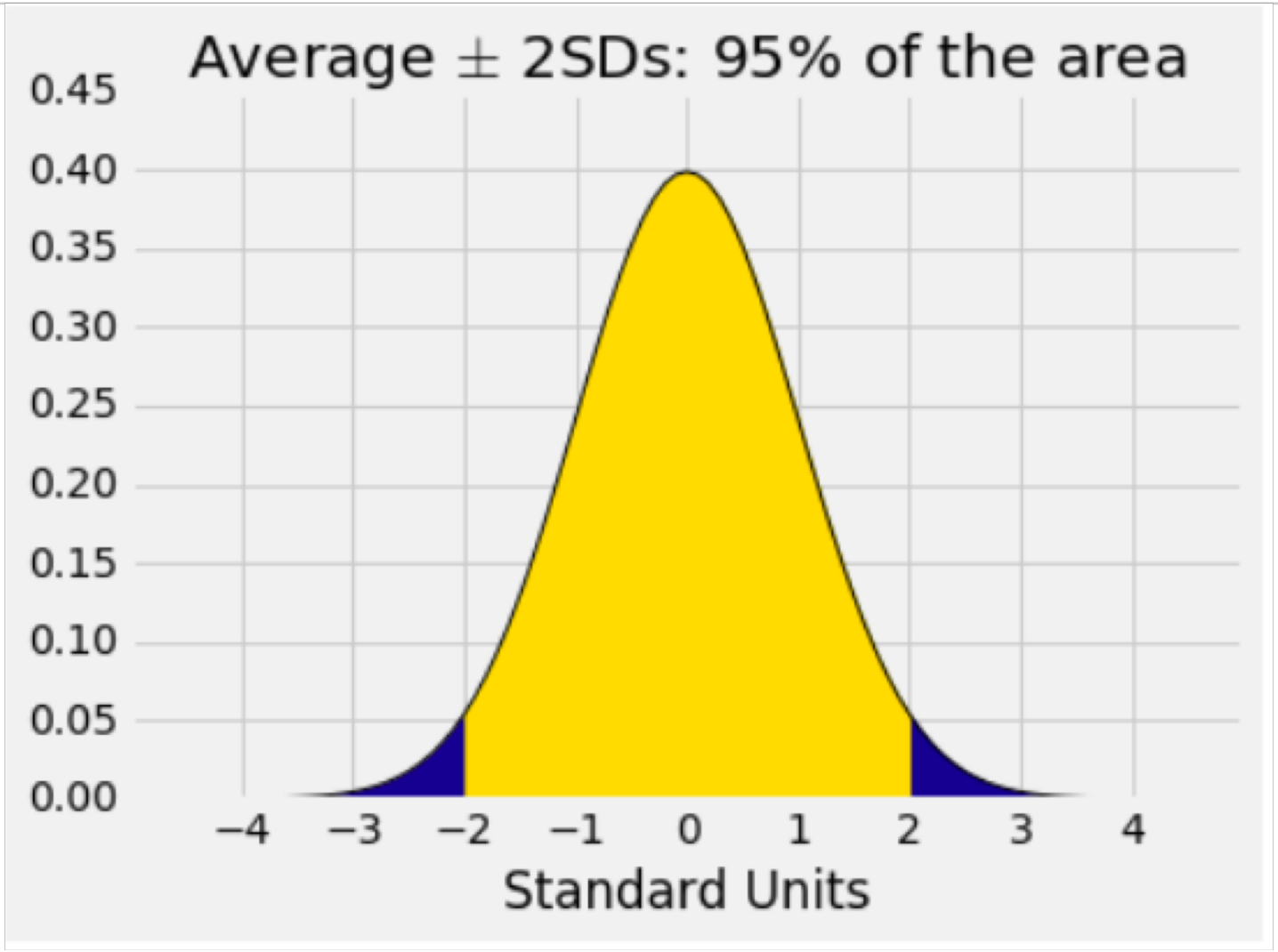
Bell Curve


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

Bounds and Normal Approximations

Percent in Range	All Distributions	Normal Distribution
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)

Why use the Standard Deviation?

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

- The Central Limit Theorem!

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

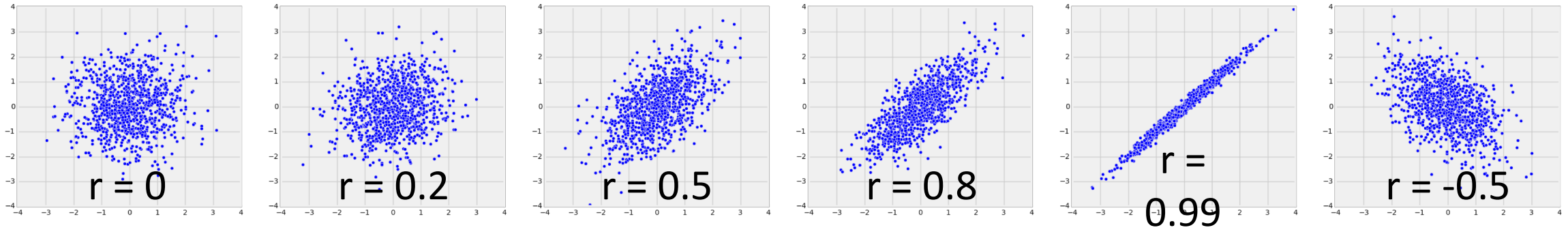
- Association
- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape”
  - Linear
  - Non-linear

Visualize then quantify

(Demo)

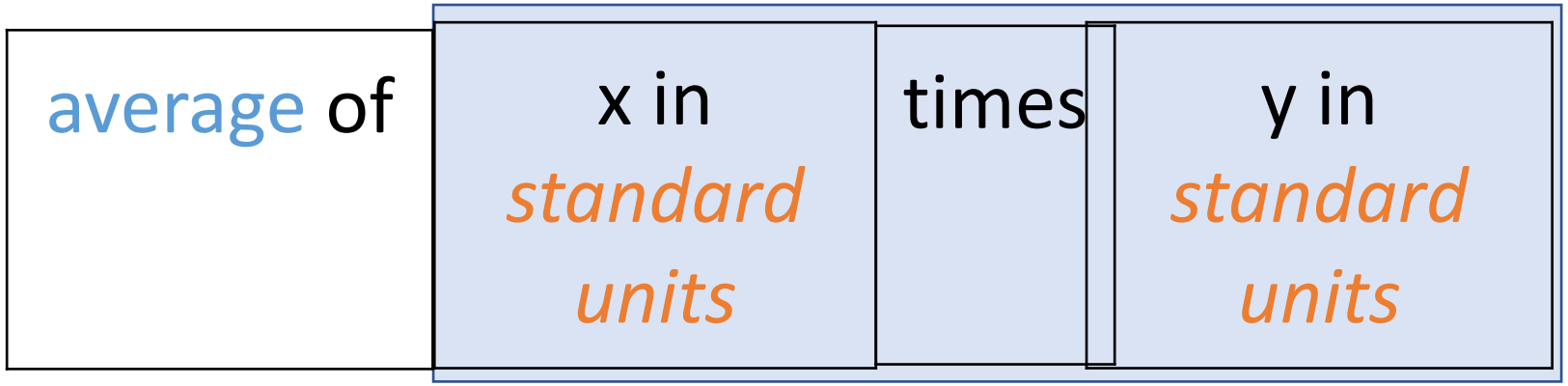
The Correlation Coefficient *r*

- Measures linear association
- Based on standard units
- $-1 \le r \le 1$
  - $r = 1$: 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*) =



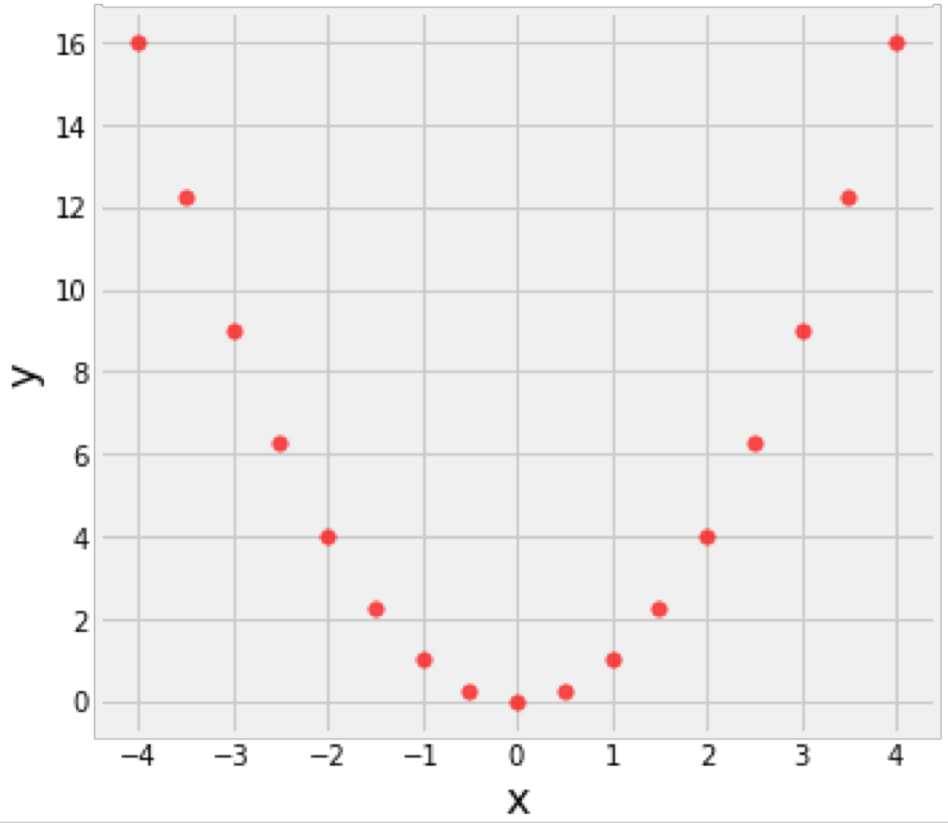
Measures how clustered the scatter is around a straight line

Interpreting *r*

Don't jump to conclusions about causality

Interpreting *r*

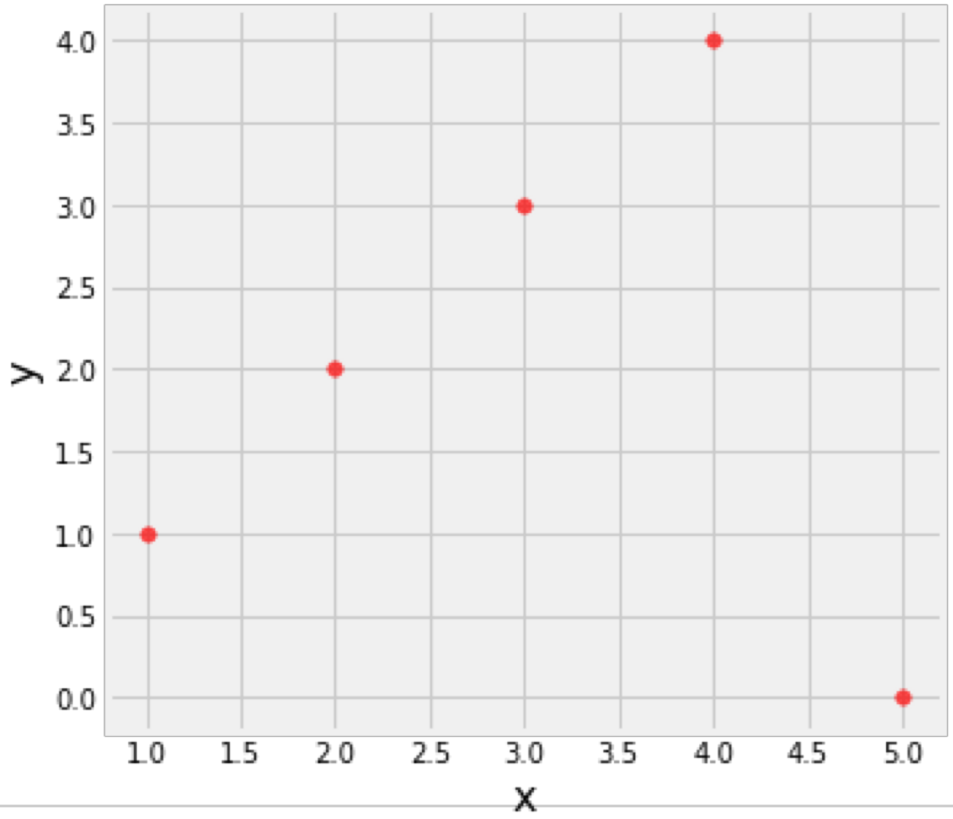
Watch out for non-linearity.



r = 0.0

Interpreting *r*

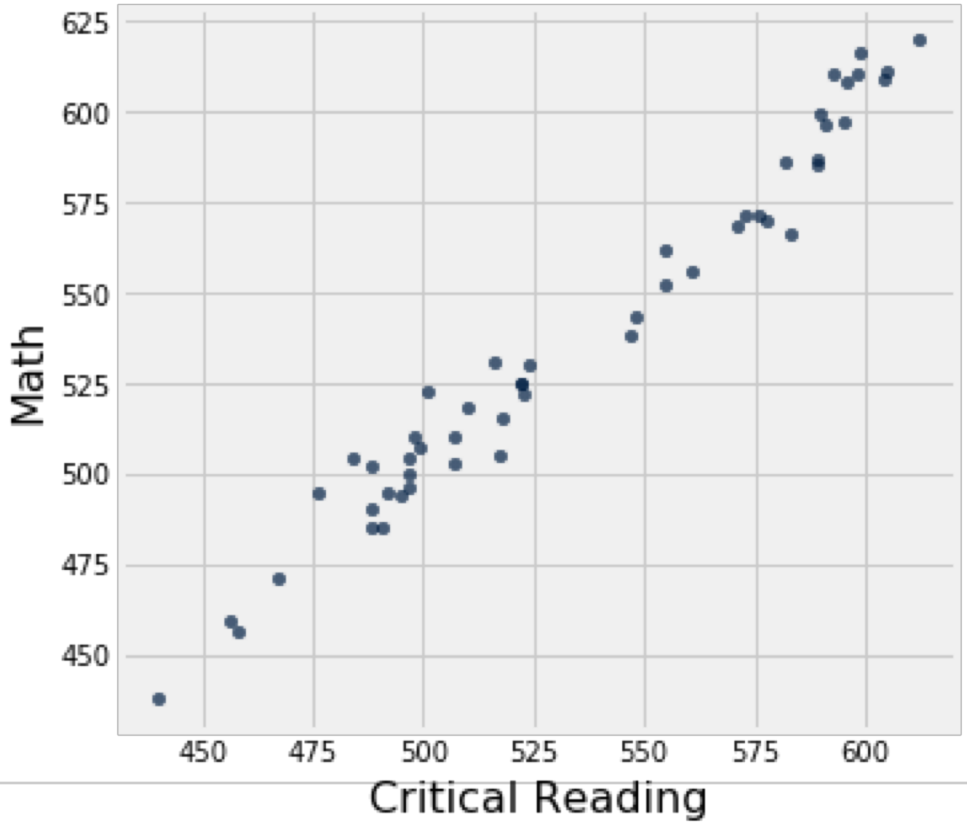
Watch out for outliers.



r = 0.0

Interpreting *r*

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



r = 0.98