DSFA

Spring 2019

Lecture 26

Review

Announcements

- **Final Exam**
  2pm Monday, May 13. B14 Hollister Hall

- **Study Guide**

- **Personal cheat sheet**

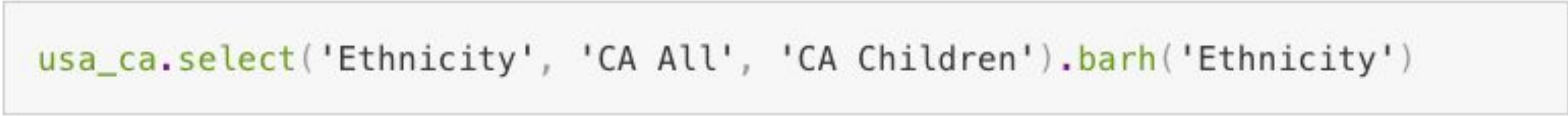
Bar Chart Versus Histogram

Bar Chart

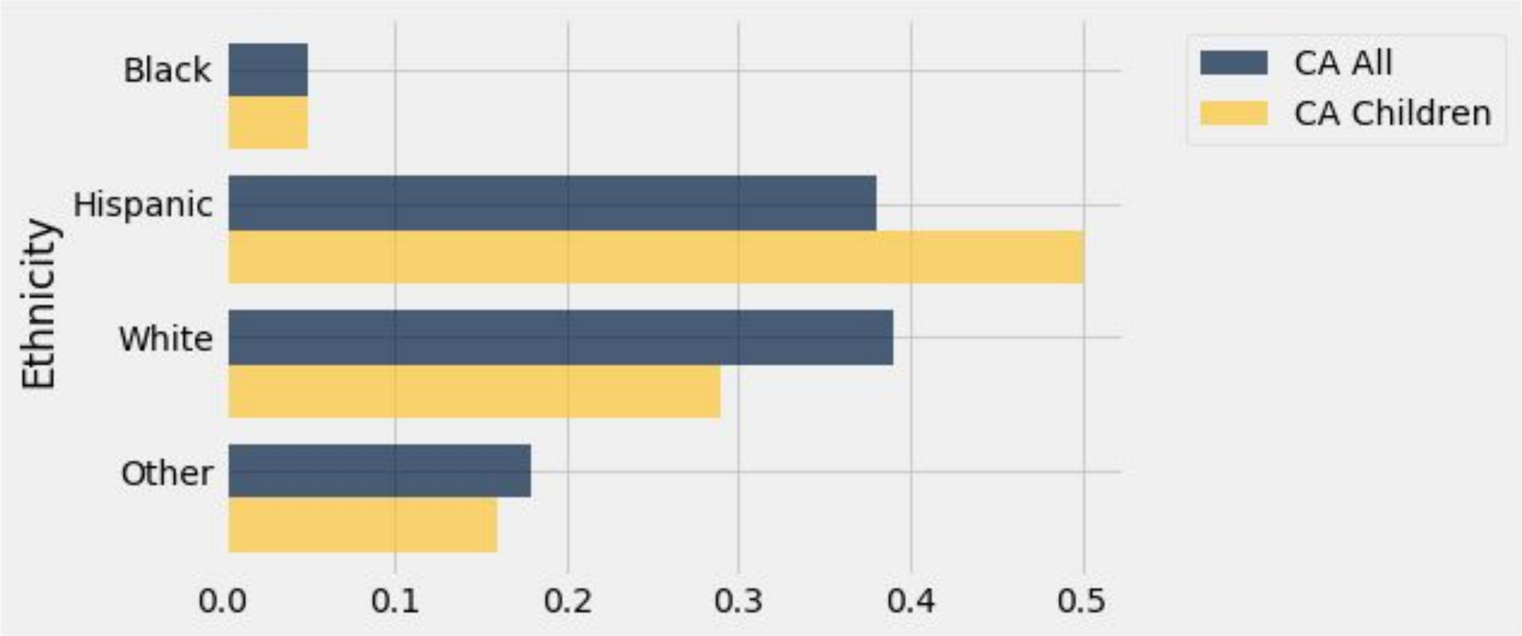
- 1 categorical axis & 1 numerical axis
- Bars have arbitrary (but equal) widths and spacings
- For distributions: height (or length) of bars are proportional to the percent of individuals

Histogram

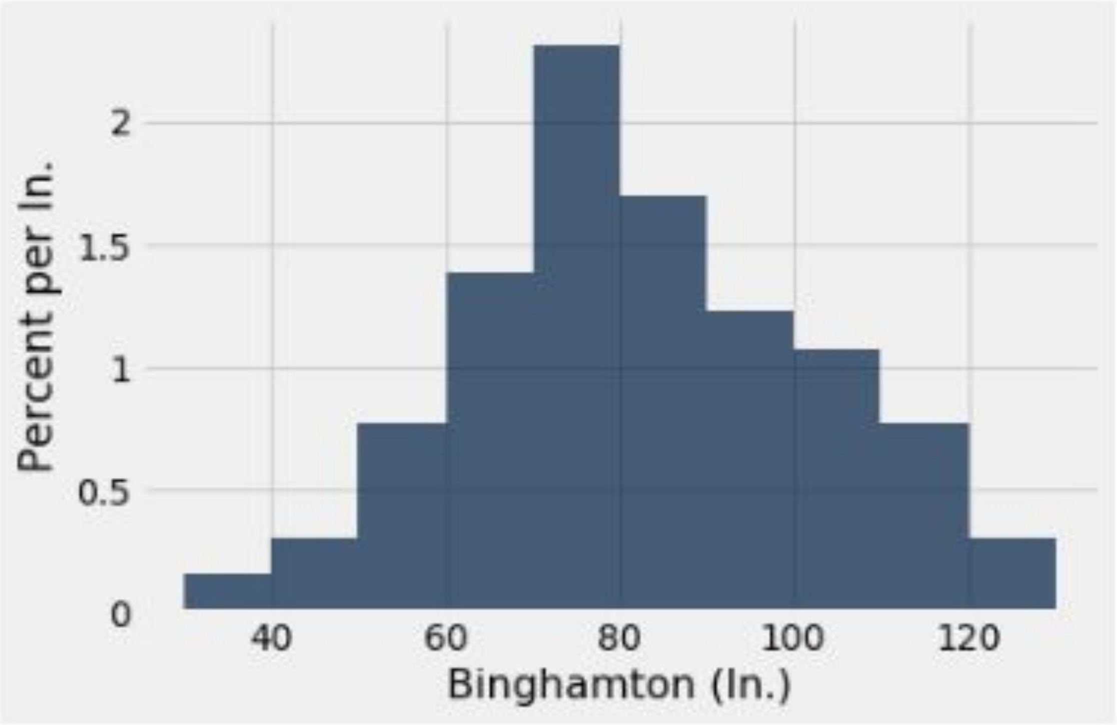
- Horizontal axis is numerical, hence to scale with no gaps
- Height measures density; areas are proportional to the percent of individuals



```
usa_ca.select('Ethnicity', 'CA All', 'CA Children').barh('Ethnicity')
```



Snowfall totals in Binghamton winters



In what proportion of winters did Binghampton receive more than 100 inches of snow?

Area = Percent, Height = Density

% in bin = Height x Width of bin

$$\textbf{Height} = \frac{\textbf{\% in bin}}{\textbf{Width of bin}}$$

- The height measures the percent of data in the bin ***relative to the amount of space in the bin*****.**
- So height measures crowdedness, or **density**.

Probability

Probability

- Lowest value: 0
  - Chance of event that is impossible
- Highest value: 1 (or 100%)
  - Chance of event that is certain
  - In general: 0 <= P(A) <= 1
- If an event has chance 70%, then the chance that it doesn’t happen is
  - 100% - 70% = 30%
  - 1 - 0.7 = 0.3
  - In general: P(not A) = 1 - P(A)

Equally Likely Outcomes

Assuming all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

Multiplication Rule

Chance that two events *A* and *B* both happen

= P(*A* happens)
x P(*B* happens **given that** *A* has happened)

= P(A happens) x P(B happens) if A and B **independent**

- The answer is *less than or equal to* each of the two chances being multiplied

Example: At Least One Head

- In 3 tosses:
  - Any outcome *except* TTT
  - P(TTT) = $(\frac{1}{2}) \times (\frac{1}{2}) \times (\frac{1}{2}) = (\frac{1}{2})^3 = \frac{1}{8}$
  - P(at least one head) = 1 - P(TTT) = $\frac{7}{8}$ = 87.5%

- In 10 tosses:
  - P(TTTTTTTTTT) = $(\frac{1}{2})^{10}$
  - P(at least one head) = 1 - $(\frac{1}{2})^{10}$ = 99.90%

Addition Rule

Chance that either *A or B* (inclusive)

$$\mathrm{P}(A \textit{ or } B) \;=\; \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \textit{ and } B)$$

Simplifies to $\mathrm{P}(A \textit{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$ if A and B are disjoint (mutually exclusive)

- The answer is *greater than or equal to* the chance of each individual way

Example: Roll a pair of dice

A = at least one 2

B = sum less than or equal to 4

C = double

What are P(*A*), P(*B*) and P(*C*)?

P(*A and B*), P(*A or B*), P(*B and C*), P(*B or C*)?

Sampling

Sampling

Observe some *individuals* from a *population*

a. Examine 10 rolls of a d6 (six-sided die)
b. Coat color of the first 20 people who walk through door
c. Survey 1000 students living in campus dorms, where every student living on campus is equally likely to be chosen, and ask them about their perspective on gun control

Sampling

- Deterministic sample:
  - Sampling scheme doesn’t involve chance

- Probability (random) sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population
  - Not all individuals have to have equal chance of being selected

(Demo)

Does sample look like population?

(Demo)

Large Random Samples

If the sample size is large,

then the empirical distribution of a simple random sample

resembles the population distribution,

with high probability.

Law of Large Numbers

If an experiment is repeated many times,
independently and under the same conditions,
then the proportion of times that an event occurs
gets closer to the theoretical probability of the event

Sometimes called *Law of Averages*

Terminology

Statistic

A number associated with the sample

Parameter

A number associated with the population

A statistic can be used as an **estimate** of a parameter

Bias

- **Biased estimate:** On average across all possible samples, the estimate is either too high or too low

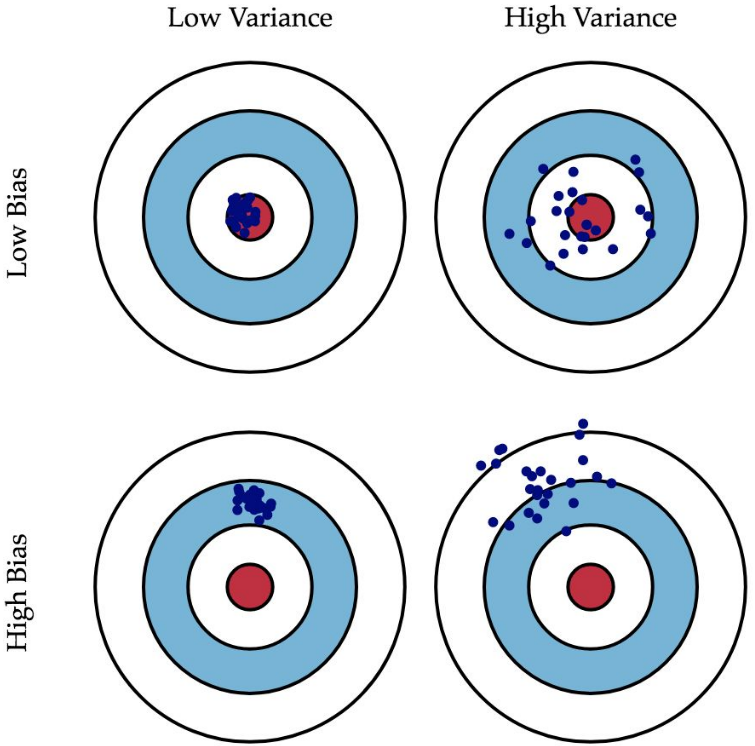
- Bias creates a systematic error in one direction

- Good estimators have low bias

Variance

- Value of an estimate **varies** from one sample to another
- High variability makes it hard to estimate accurately
- Good estimators have low variance

Bias vs Variance



http://scott.fortmann-roe.com/docs/BiasVariance.html

Distribution of a Statistic

Statistic: A quantity computed for a particular sample

Sampling distribution: Chance of each value of a statistic (computed from all possible samples)

Also known as the *probability distribution of the statistic*

Empirical distribution: Observations of a statistic (computed from some samples drawn at random)

Simulating a Statistic

Fix a sample size and choose your statistic.

- Simulate the statistic once:
    - Draw a random sample of the size you fixed.
    - Calculate the statistic and keep a record of the value
- Repeat previous step numerous times (as many times as you have patience for; thousands are good).
- You now have one value of the statistic for each repetition. Visualize the results.

(Demo)

Hypothesis Testing

Testing a Hypothesis

Step 1: The Hypotheses

- A test chooses between two views of how data were generated
- *Null hypothesis* proposes that data were generated at random
- *Alternative hypothesis* proposes some effect other than chance

Step 2: The Test Statistic

- A value that can be computed for the data and for samples

Step 3: The Sampling Distribution of the Test Statistic

- What the test statistic might be if the null hypothesis were true
- Approximate the sampling distribution by an empirical distribution

Conclusion of a Test

Resolve choice between null and alternative hypotheses

- Compare observed test statistic to its empirical distribution under the null hypothesis
- If the observed value is **consistent** with the distribution, then the test *does not* support the alternative hypothesis

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- Convention: The observed significance level (P-value)

Observed Significance Level

P-Value: The chance, under the null hypothesis,
that the test statistic is equal to the value that was observed
or is even further in the direction of the alternative.

Statistically Significant: The P-value is less than 5%

Highly Statistically Significant: The P-value is less than 1%

Mendel’s Peas

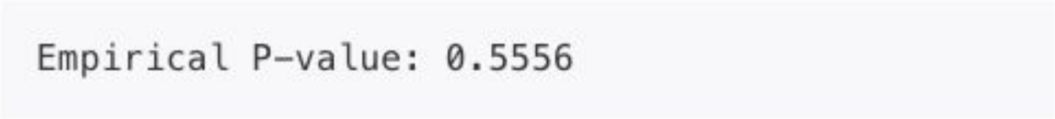
Mendel’s Null hypothesis: 75% of pea plants will have purple flowers

Alternative hypothesis: Mendel’s hypothesis is wrong

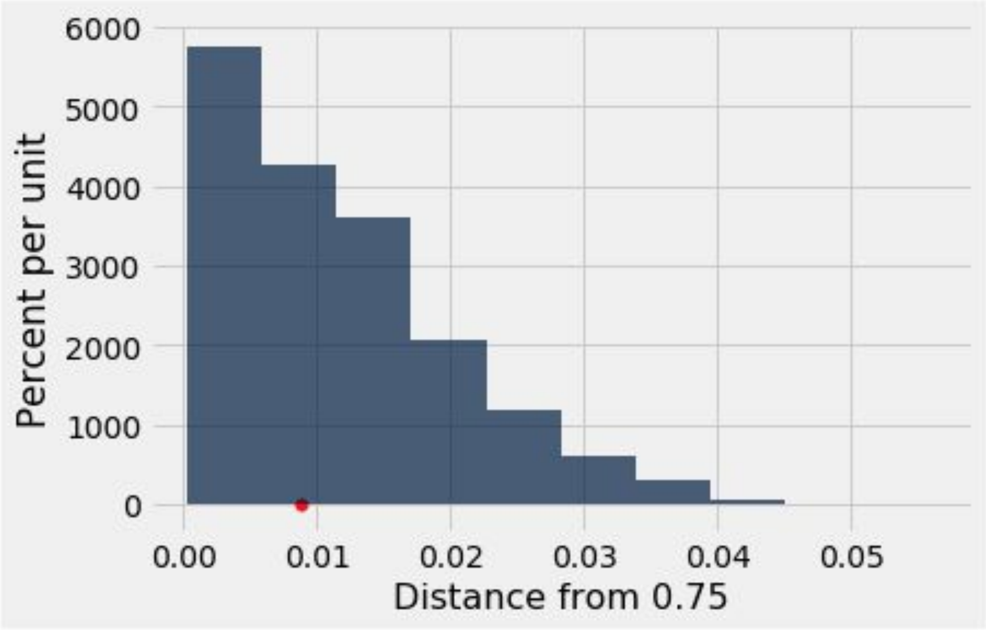
Test statistic: | sample proportion purple - 0.75 |

Observed value: | 0.75888 - 0.75 | = 0.00888 (n = 929)

Simulated null distribution of the test-statistic

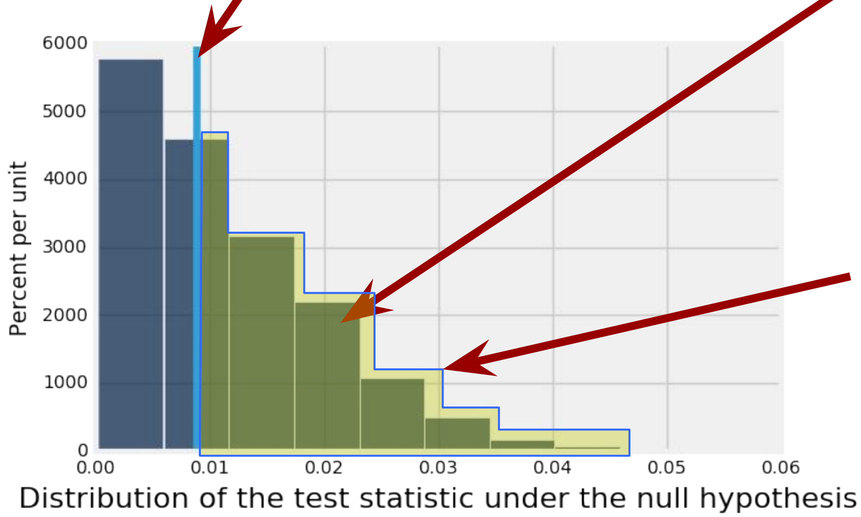


```
Empirical P-value: 0.5556
```



Quantifying Conclusions

P(the test statistic would be equal to or more extreme than the observed test statistic under the null hypothesis)



Evaluating Mendel's pea flower hypothesis

This area is the P-value (approximately)

Can the Conclusion be Wrong?

Yes.

	Null is true	Alternative is true
Test rejects the null	✗	✓
Test doesn’t reject the null	✓	✗

An Error Probability

- The cutoff for the P-value is an error probability.

- If:
  - your **cutoff is 5%**
  - and the **null hypothesis happens to be true**
  - (but you don’t know that)

- then there is about a **5% chance** that **your test will reject the null hypothesis anyway**.

The Bootstrap

Variability of the Estimate

- One sample ➜ One estimate
- But the random sample could have come out differently
- And so the estimate could have been different
- Main question:
  - **How different could the estimate have been?**

Where to Get Another Sample?

- One sample ➜ One estimate
- To get many values of the estimate, we needed many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

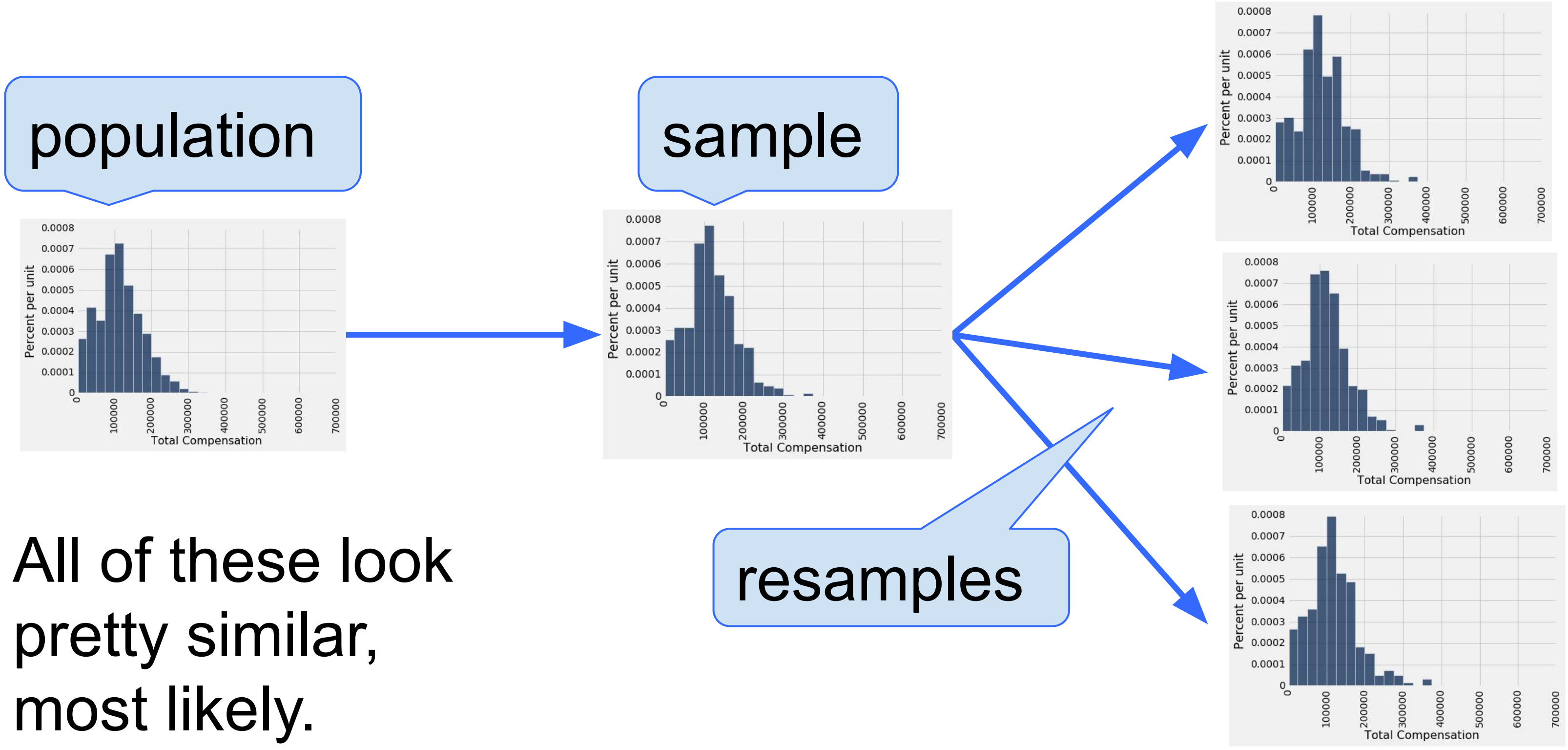
The Bootstrap

- A technique for simulating repeated random sampling

- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population

- So we sample at random from the original sample!

Why the Bootstrap Works



population

sample

resamples

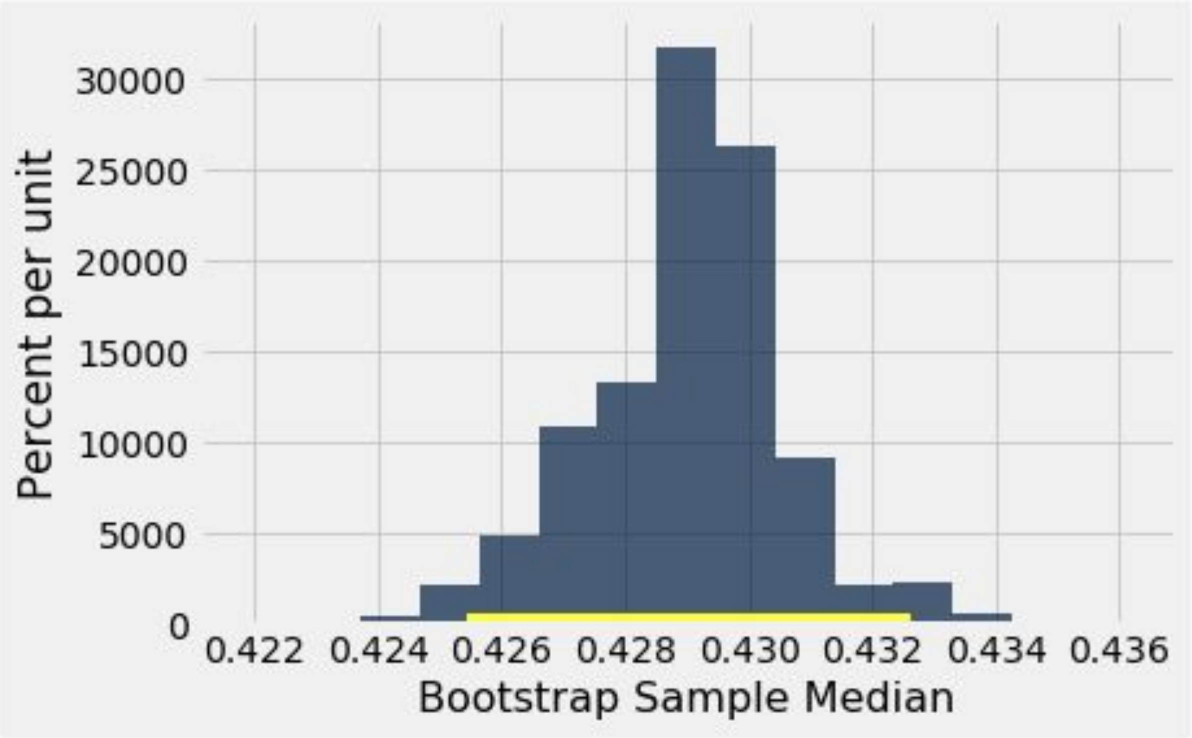
All of these look pretty similar, most likely.

Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained

- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

95% (bootstrap) confidence interval



Interval extends from the 2.5 percentile to the 97.5 percentile of the bootstrap distribution

Mean versus Median

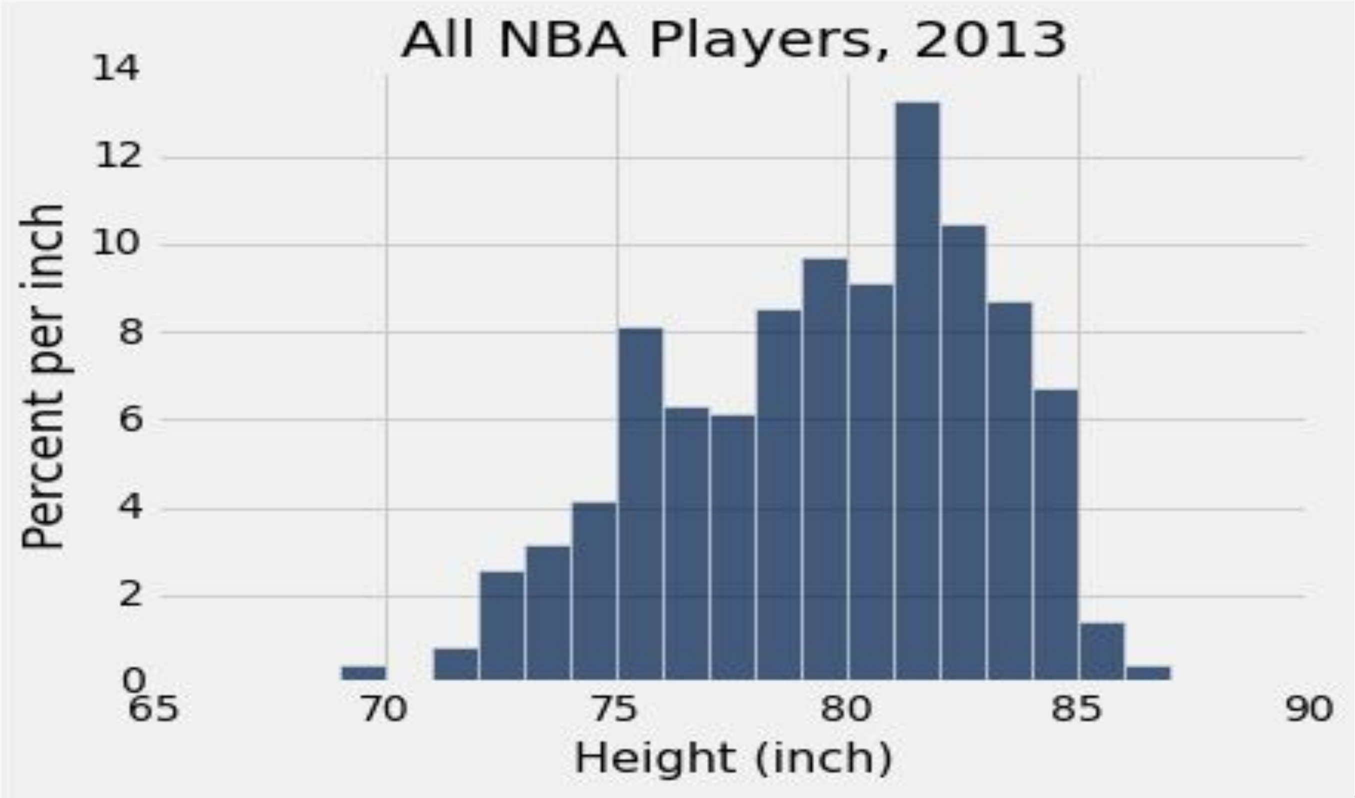
- **Mean:** Balance point of the histogram
- **Median:** Halfway point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from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

Discussion Question

Which is bigger?

(a) mean

(b) median



Standard Deviation

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

  Example: Sample: 2, 3, 3, 8, Average/Mean: 4.0

$$SD = \sqrt{\frac{1}{4}\left[(2-4)^2 + (3-4)^2 + (3-4)^2 + (8-4)^2\right]}$$

- SD has the same units as the data

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

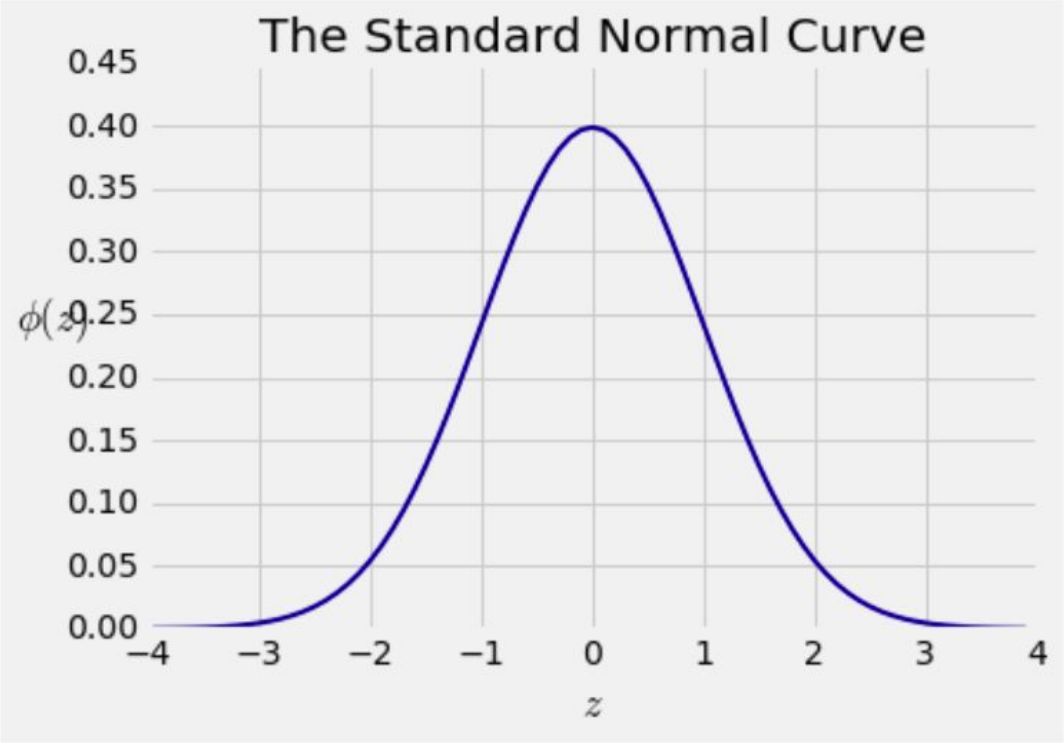
Standard Units

Standard Units

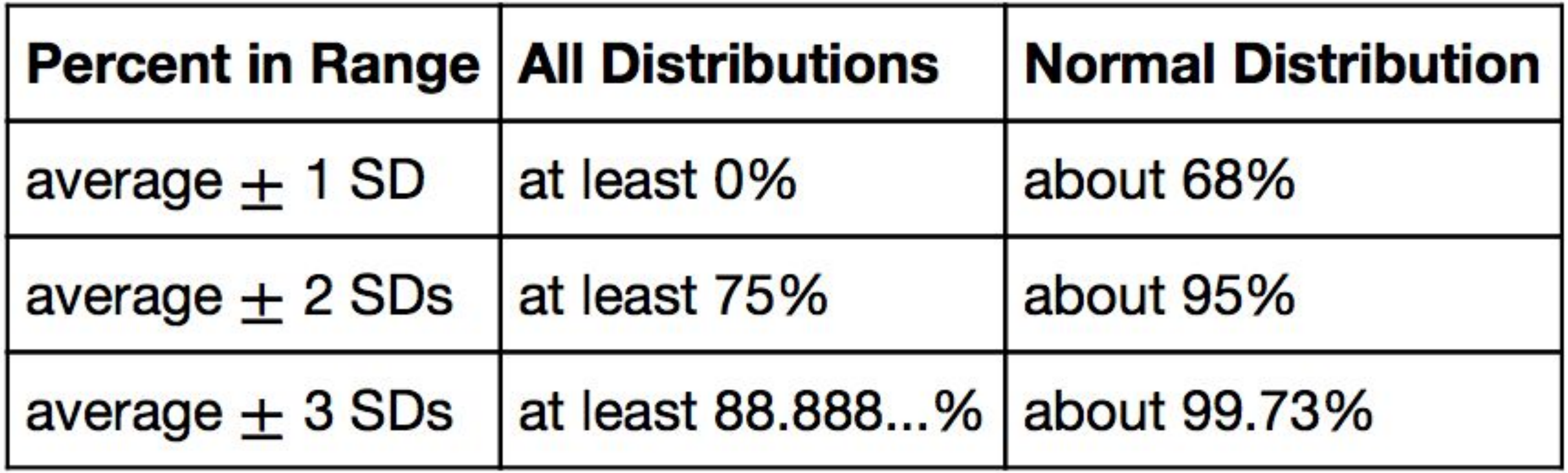
- How many SDs above average?
- ***z* = (value - mean)/SD**
  - Negative z: value below average
  - Positive z: value above average
  - z=0: value equal to average
- When values are in standard units: average = 0, SD = 1

The Normal Distribution

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 ± 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%

Central Limit Theorem

Second Reason for Using the SD

If the sample is

- large, and
- drawn at random with replacement,

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

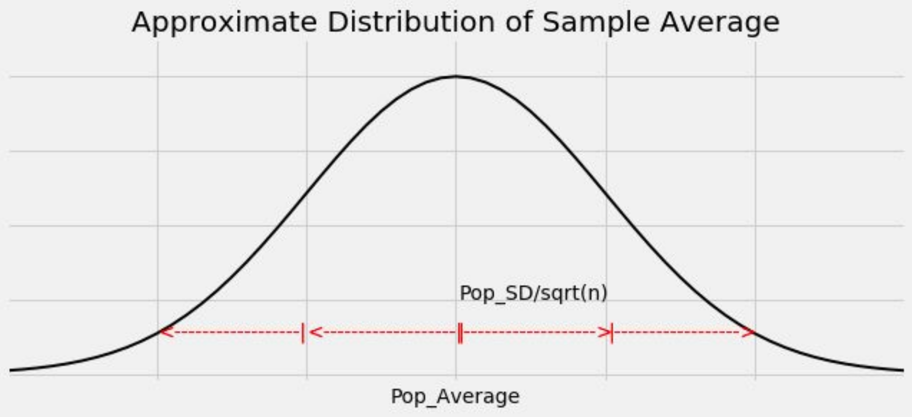
the probability distribution of the sample sum (or of the sample average) is roughly bell-shaped

Distribution of the Sample Average

- Fix a large sample size.
- Draw many random samples of that size.
- Compute the average of each sample.
- You’ll end up with a lot of averages.
- The distribution of those is called the *(sampling) distribution of the sample average (or mean).*
- It’s roughly normal, centered at the population average.
- SD = (population SD) / $\sqrt{\text{sample size}}$

Confidence Intervals

The Key to 95% Confidence



- For about 95% of all samples, the sample average and population average are within **2 SDs** of each other.

- **SD** = SD of sample average

  = (population SD) / $\sqrt{\text{sample size}}$

Constructing the Interval

For 95% of all samples,

- If you stand at the population average and look two **SD**s on both sides, you will find the sample average.
- Distance is symmetric.
- So if you stand at the sample average and look two **SD**s on both sides, you will capture the population average.

Width of the Interval

Total width of a 95% confidence interval for the population average

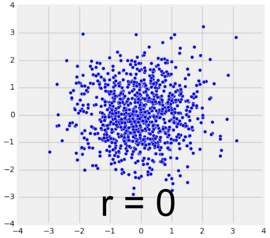
= 4 * SD of the sample average

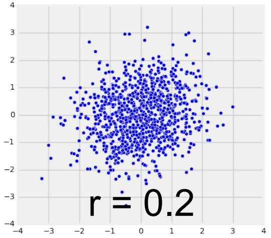
= 4 * (population SD) / $\sqrt{\text{sample size}}$

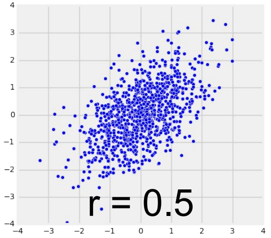
What sample is required for the width to be less than *w*?

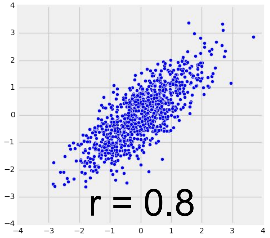
The Correlation Coefficient *r*

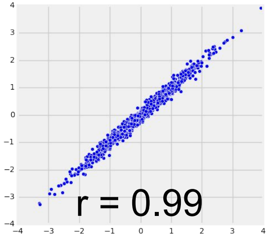
- Measures linear association
- Based on standard units
- -1 ≤ *r* ≤ 1
    - ○ *r* = 1: scatter is perfect straight line sloping up
    - ○ *r* = -1: scatter is perfect straight line sloping down
- *r* = 0: No linear association; *uncorrelated*

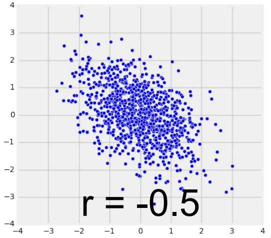












Definition of *r*

Correlation Coefficient (*r*) =

average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

Properties of *r*

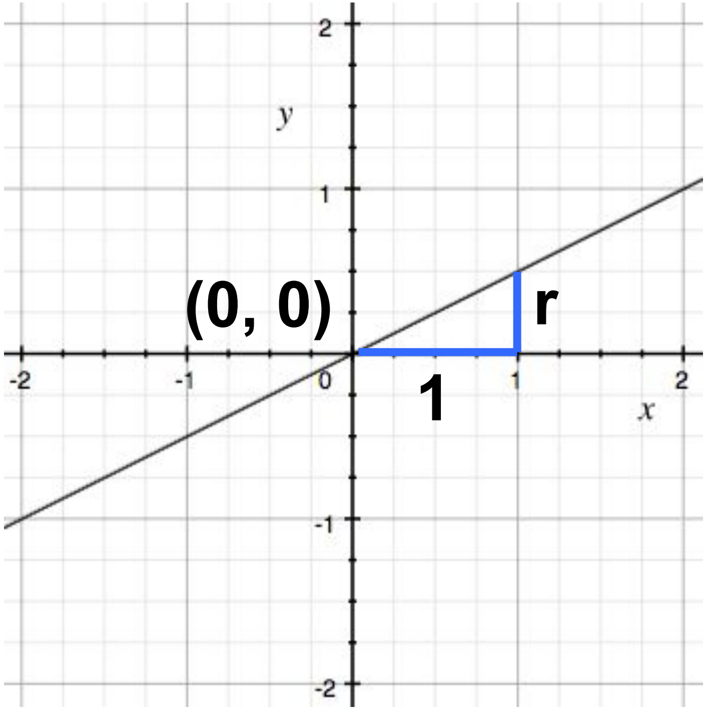
- *r* is a pure number, with no units
- *r* is not affected by changing units of measurement
- *r* is not affected by switching the horizontal and vertical axes (symmetric in *x* and *y*)

Prediction

- **Problem:** given a known *x* value, predict *y*, where both are in standard units
- **Solution:**
  - Compute *r*
  - Predict that $y = r * x$
- Why is that a line?

Algebra review:

Equation of a Line



$$y = r * x$$

In general:

$$y = a * x + b$$

(a is slope, b is intercept)

Prediction

- **Predict** *y(su) = r * x(su)*

- Example:
  - A course has a typical prelim (mean=70, std=10), and a hard final (mean=50, std=12)
  - The scores on the exams look linearly related when visualized, with *r* = .75
  - **Predict** a student's final exam score, given that their prelim score was 90

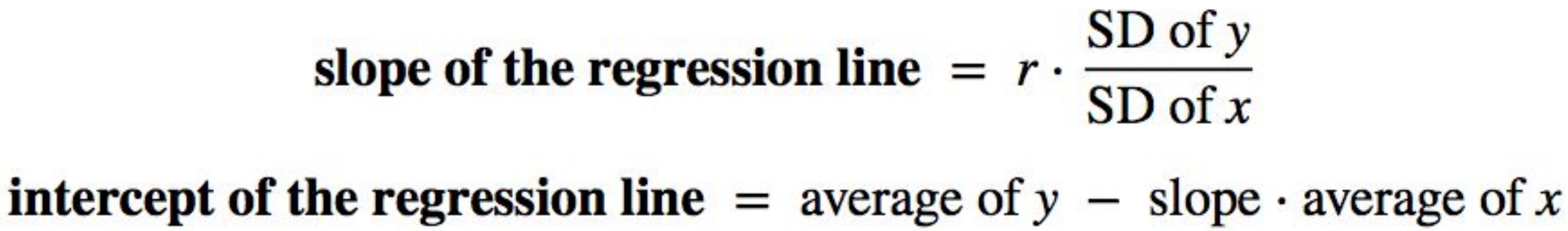
Equation for regression line

$$\frac{\text{y - mean(all y)}}{\text{std(all y)}} = r * \frac{\text{x - mean(all x)}}{\text{std(all x)}}$$

Do some algebra to put that in the form y = slope * x + intercept...

Slope and Intercept

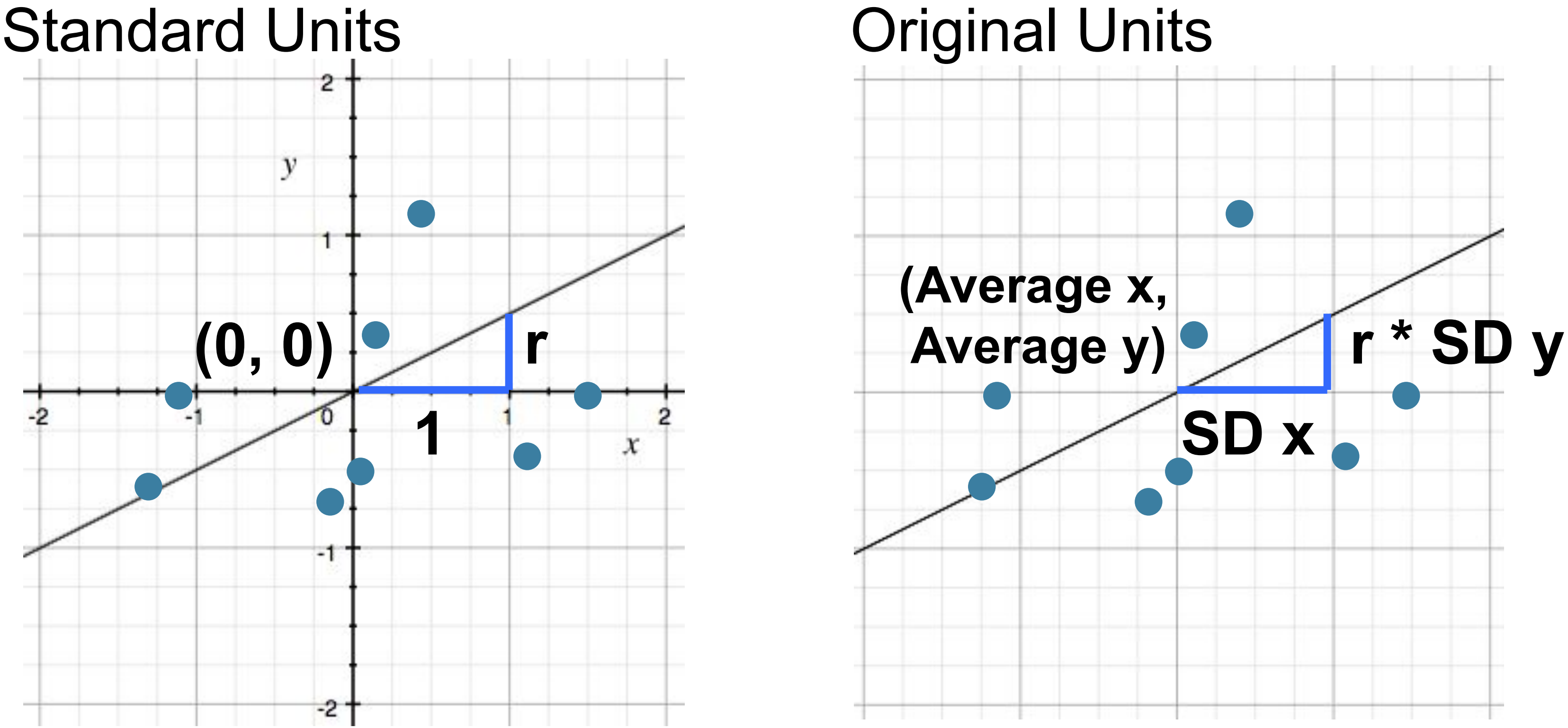
$$y \;=\; \text{slope} * x \;+\; \text{intercept}$$



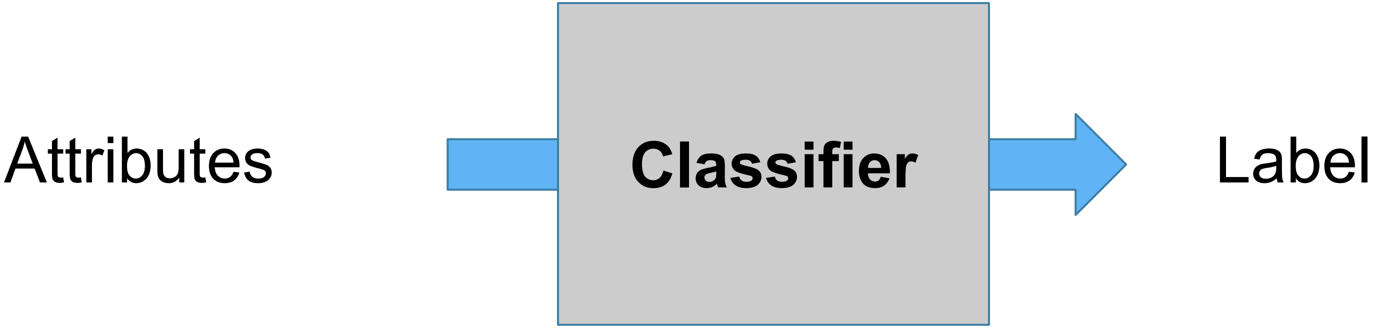
$$\textbf{slope of the regression line} \;=\; r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} \;=\; \text{average of } y \;-\; \text{slope} \cdot \text{average of } x$$

Regression Line



Classifier



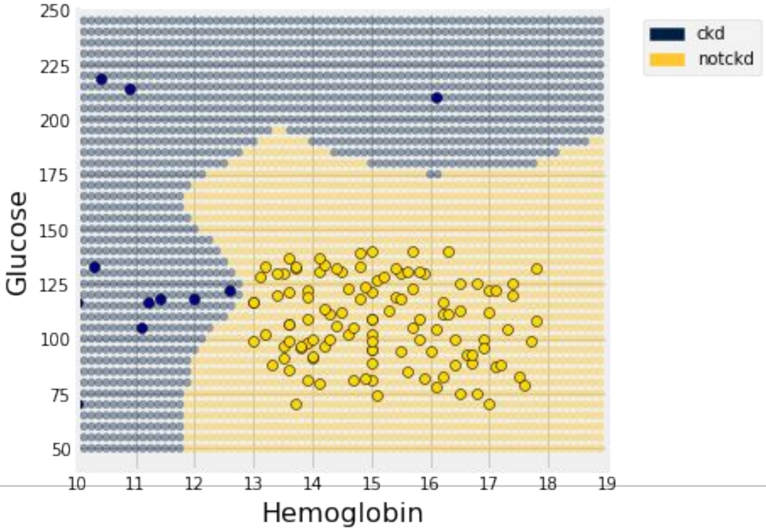
Nearest Neighbor

How to classify a new individual:

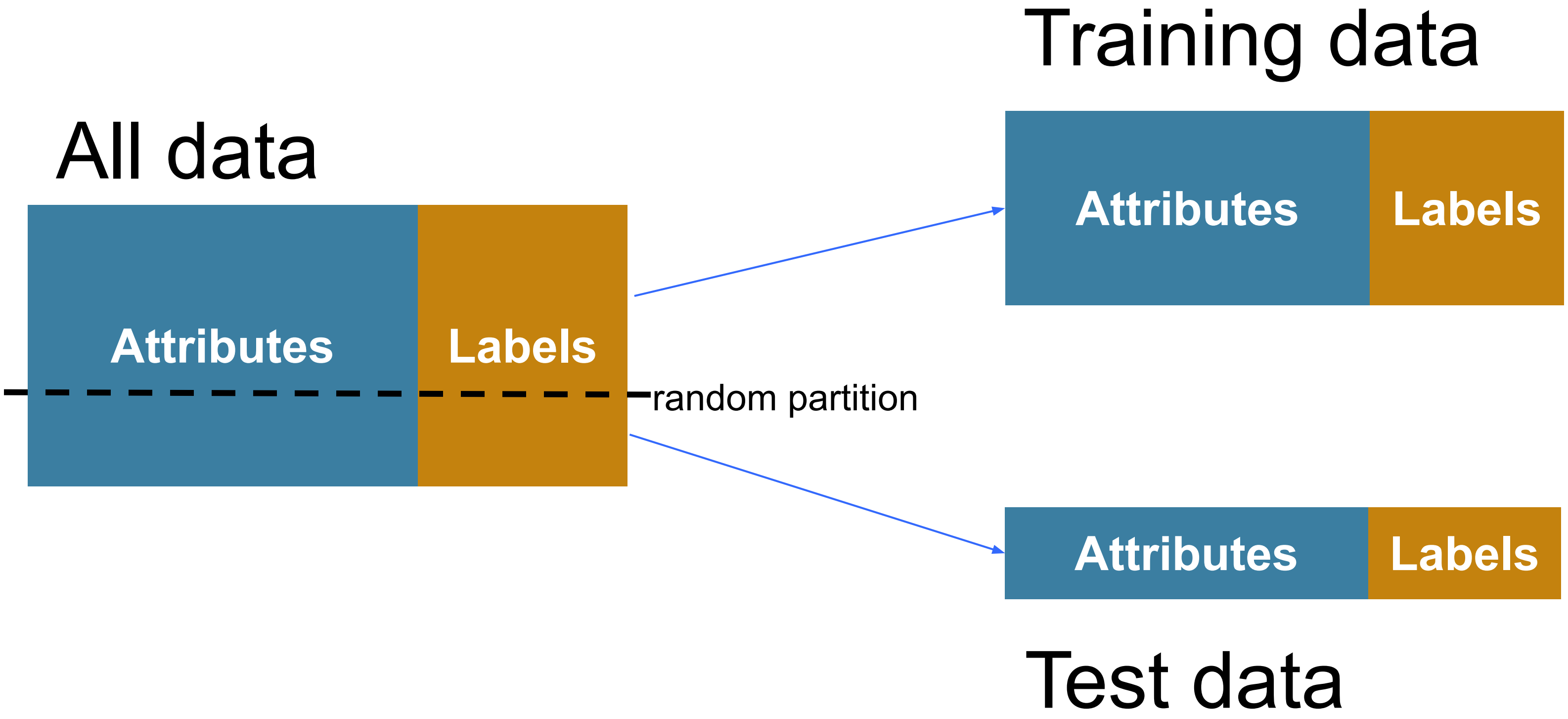
- Find their nearest neighbor: the individual closest to them in the data set
- Assign the new individual the **same** label as that nearest neighbor

Decision Boundary

- Partition between the two classes
- Computer figured out that boundary, instead of humans having to “hard code” it: machine learning



Train vs. Test



Train vs. Test

- Use training data to create the classifier
- Use test data to evaluate the finished classifier

- **Never** allow classifier to see test data until the very end: think of classifier as a cheater who would be happy to just memorize the answers

Multiple Neighbors

- If data are noisy, asking just the closest neighbor might not be ideal for accuracy
- Instead, ask the *k* closest neighbors, and take the majority label