DSC 10, Spring 2018
Lecture 17

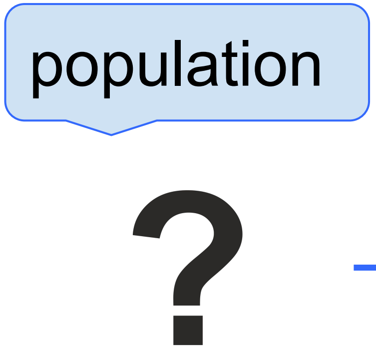
Center and Spread

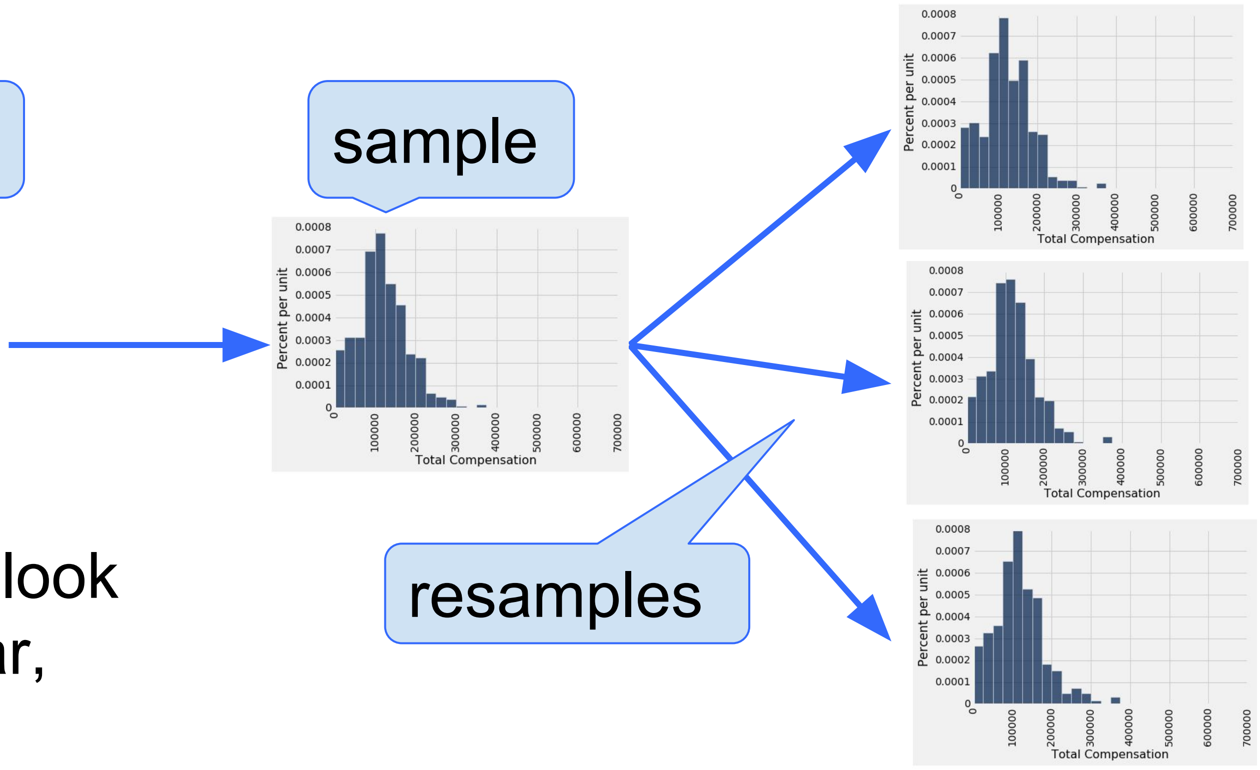
sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Review: Bootstrapping and Confidence Intervals

Inference Using the Bootstrap





All of these look pretty similar, most likely.

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- It generates a “good” interval about 95% of the time.
    - “good” means it contains the parameter

When *Not* to Use The Bootstrap

- If you're trying to estimate very high or very low percentiles, or min and max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)
- If the original sample is very small or not random

(Demo)

Can You Use a C.I. Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

A: True
B: False
C: I’m lost

Percentiles

Percentiles

- The data: numerical values
- The *p*th percentile is:
  - the smallest value in a set
  - that is at least as large as
  - *p*% of the elements in the set

The median (50%) of 4, 7, 9, 10, 15 is 9

Computing Percentiles

- The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

Computing Percentiles

- The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

- The 80th percentile is the 4th ordered element:

```
(80/100)*5 = 4
```

Percentile

Size of set

Which ordered element (counting from 1)

Computing Percentiles

- The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

- The 80th percentile is the 4th ordered element:

```
(80/100)*5 = 4
```

Percentile

Size of set

Which ordered element (counting from 1)

- For a percentile that does not exactly correspond to an element, take the next greater element instead

The `percentile` Function

- The *p*th percentile is the value in a set that is at least as large as *p*% of the elements in the set
- Function in the `datascience` module:

  `percentile(p, values)`

- **`p`** is between 0 and 100
- Returns the *p*th percentile of the array

Discussion Question

Which of the following are `True`, when `s = [1, 7, 3, 9, 5]`?

```
1.  percentile(10, s) == 0
2.  percentile(39, s) == percentile(40, s)
3.  percentile(40, s) == percentile(41, s)
4.  percentile(50, s) == 5
```

A. 1 and 2
B. 2 and 3
C. 2 and 4
D. 3 and 4
E. None of the above combinations

(Demo)

Average

The Average (The Mean)

Data: 2, 3, 3, 9 **Average = (2+3+3+9)/4 = 4.25**

- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

(Demo)

Weights

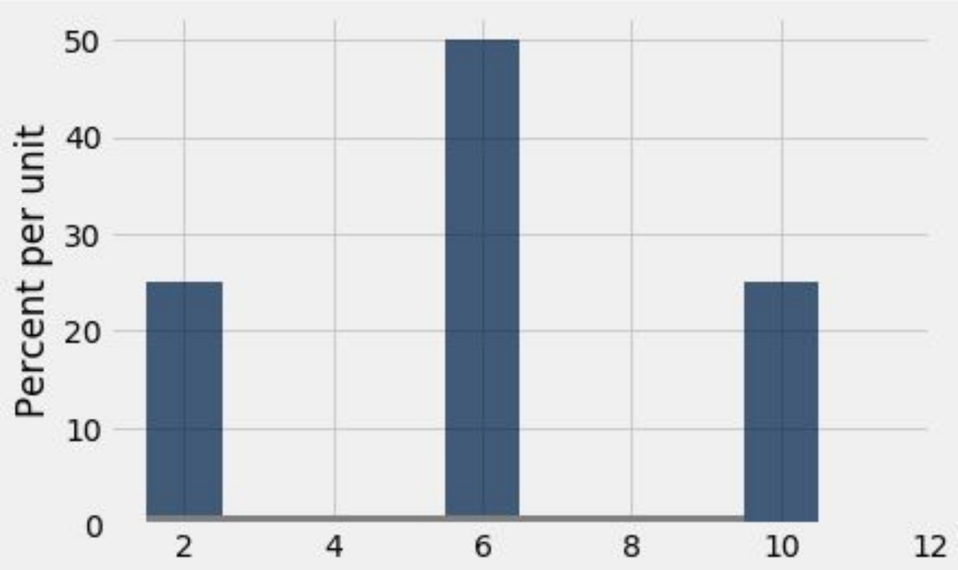
Data: 2, 3, 3, 9

$$4.25 = \frac{2 + 3 + 3 + 9}{4}$$

$$= 2*(1/4) + 3*(1/4) + 3*(1/4) + 9*(1/4)$$

$$= 2*(1/4) + 3*(1/2) + 9*(1/4)$$

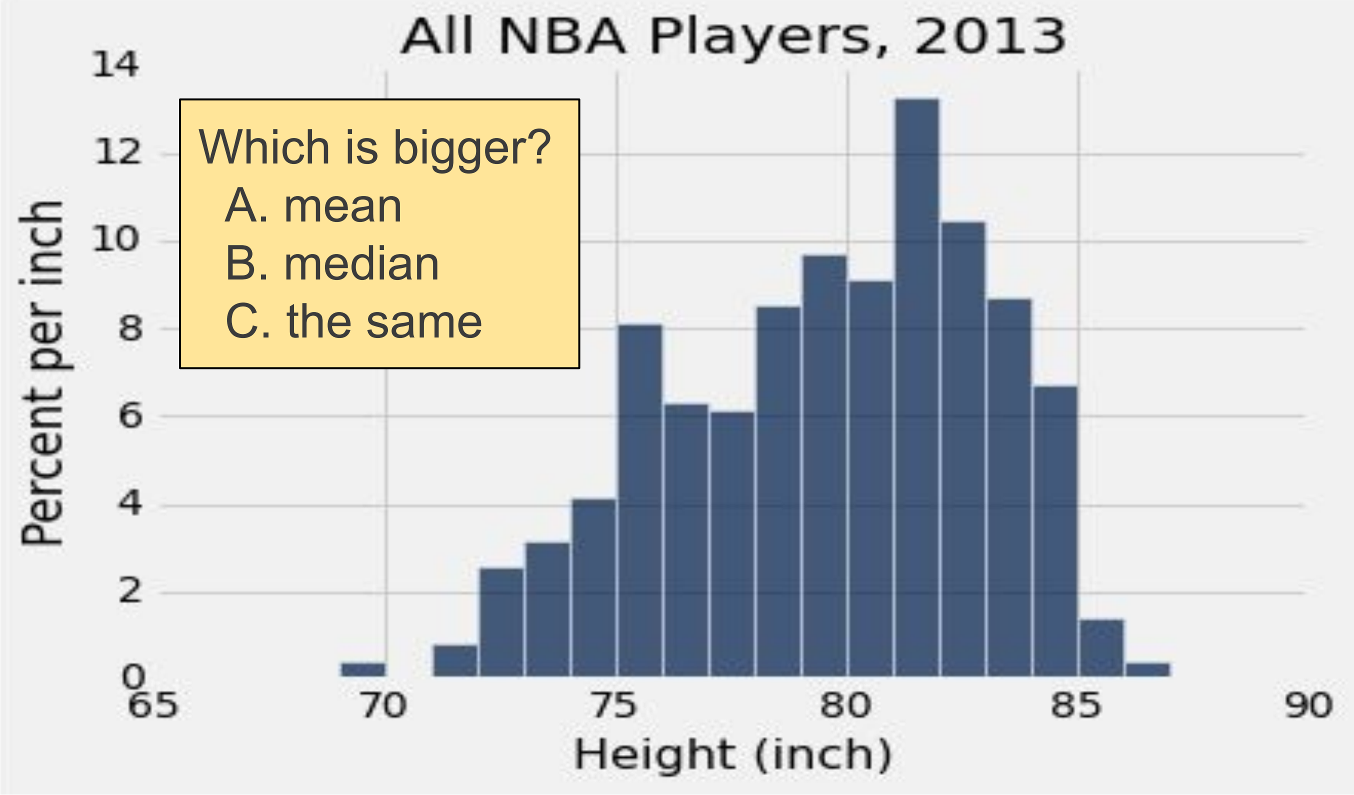
Discussion Question



How can you calculate the mean?

A. (2 + 6 + 10)/3
B. (2 + 6 + 10)/4
C. (2 + 6 + 6 + 10)/3
D. (2 + 6 + 6 + 10)/4
E. None of the above

Discussion Question



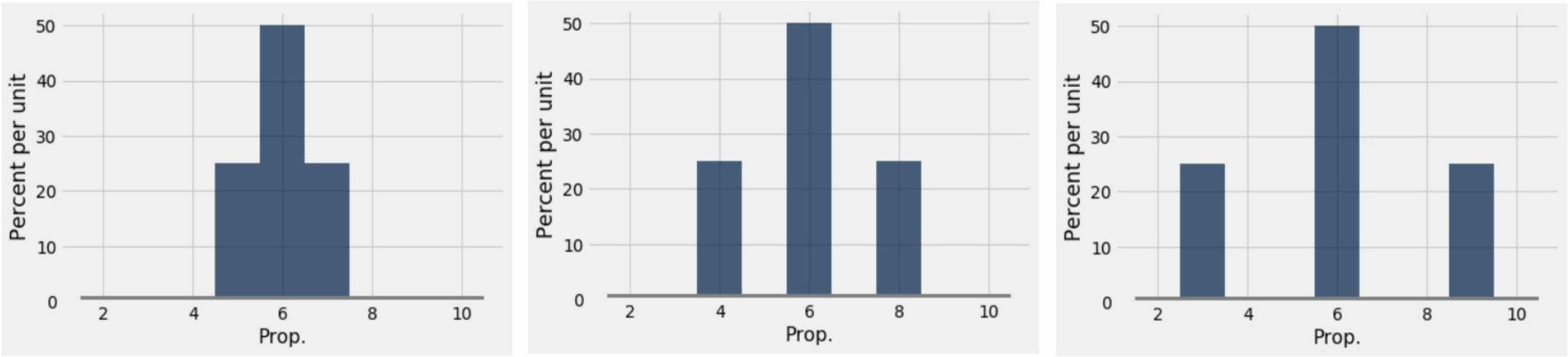
Properties of the Mean

- Balance point of the histogram
- Not the “halfway point” of the data; the mean is not the median...
- If the distribution is symmetric about a point, then that point 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

Measuring Variability

Center and Spread

- The mean is a measure of **center**.
    - An alternative measure of center is the median.
- Different data sets can have the same mean, but different **spread** or **variability** around that mean.



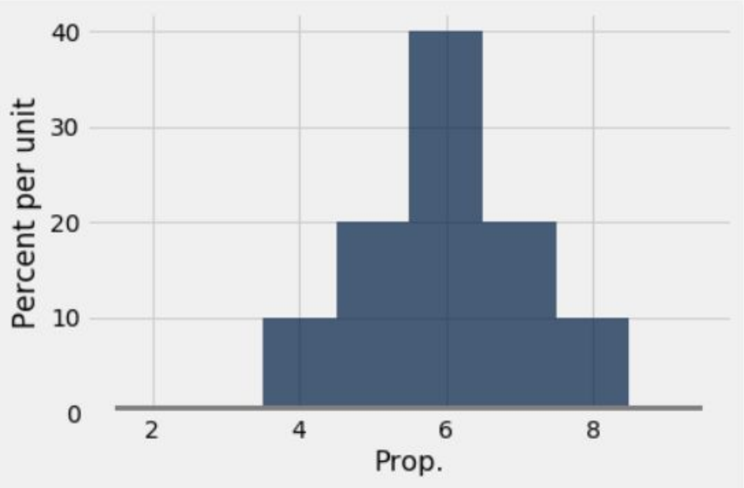
Defining Variability

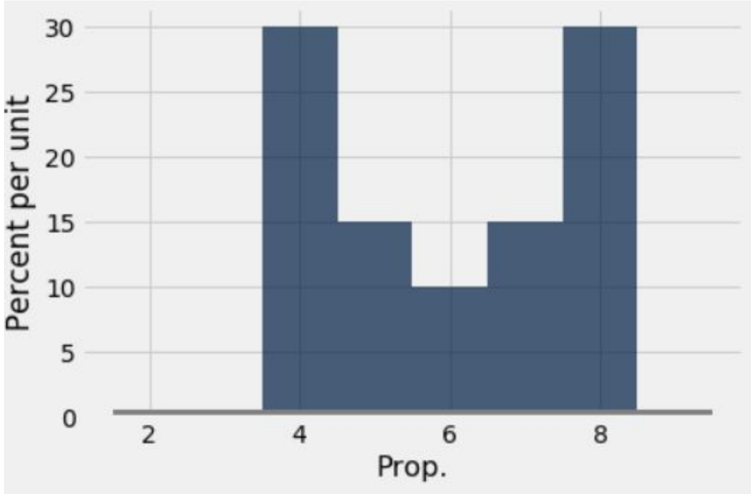
Plan A: “largest value - smallest value”

Defining Variability

Plan A: “largest value - smallest value”

- Doesn’t provide information about the shape of the distribution





Defining Variability

Plan A: “largest value - smallest value”

- Doesn’t provide information about the shape of the distribution

Plan B:

- Measure how far the data is from the mean
- Need a precise way to quantify this

(Demo)

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

  5 4 3 2 1

- SD has the same units as the data; hence OK to say "average plus or minus a few SDs"

Discussion Question

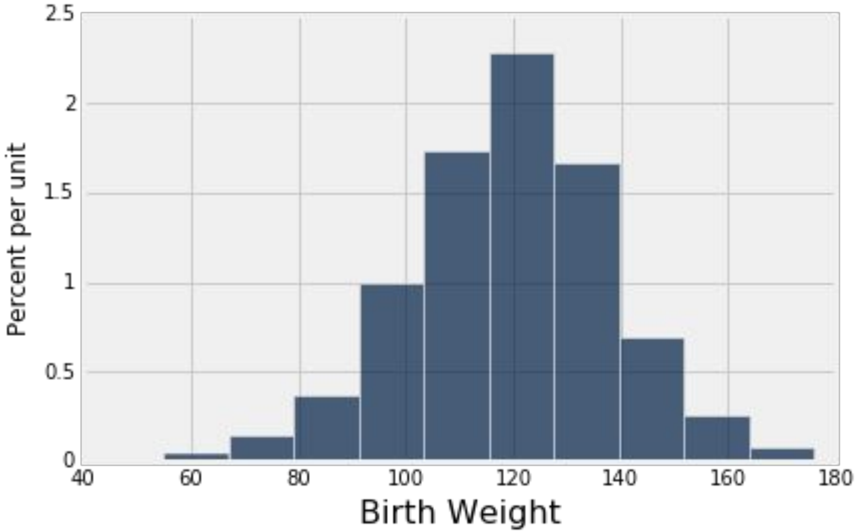
Three gorilla siblings are 2, 3, and 4 years old.

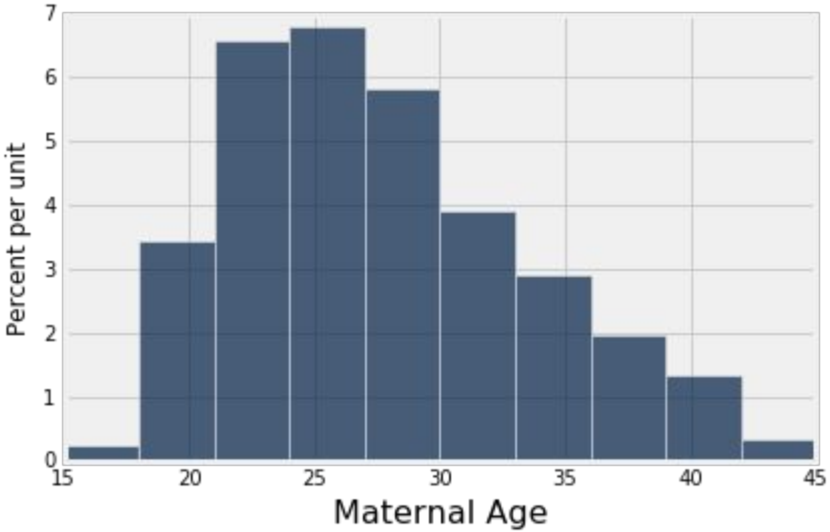
What is the standard deviation of gorilla ages?

A. 1
B. 2
C. sqrt(2)
D. sqrt(⅔)
E. None of the above.

SD = root mean square of deviations from average

Which Has Larger SD?





A. Birth Weight (Left)
B. Maternal Age (Right)
C. Cannot tell from the histograms

(Demo)

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
- Most values of *z* are between -5 and 5 (later)

(Demo)

Chebyshev's Inequality

How Big are Most of the Values?

No matter what the shape of the distribution,
<u>the bulk</u> of the data falls in the range "average ± <u>a few</u> 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

(Demo)