DSFA

Spring 2019

Lecture 20

Center and Spread

Announcements

- Project 2: Final submission April 16
- Prelim 2:   In-class. Tuesday, April 16
  - Sample questions will be posted
  - Study guide and personal cheat sheet
  - Covers material in chapters 9 through 11
- Homework 7 posted on Tuesday

Questions

- 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

The Average (or Mean)

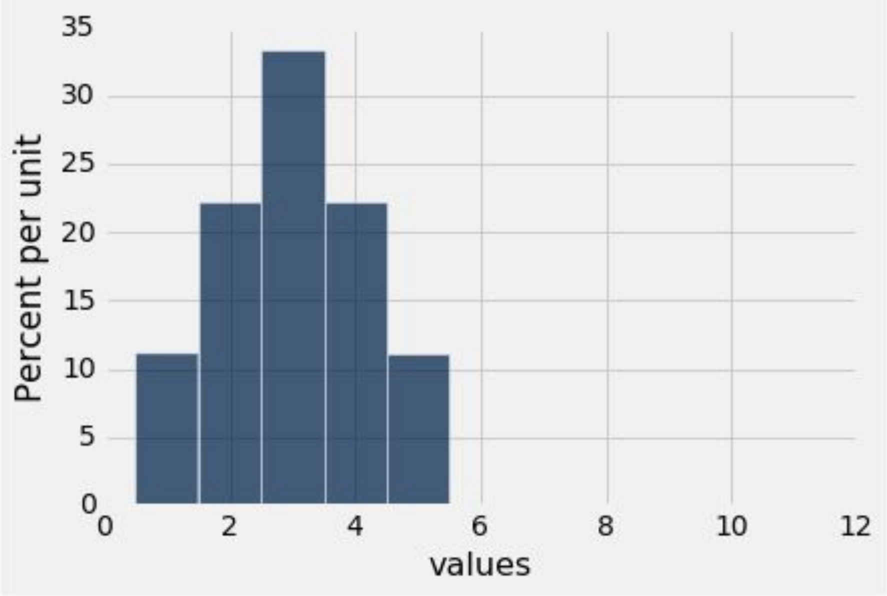
Data: 2, 3, 3, 9   **Average = (2+3+3+9)/4 = 4.25**

- 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

Discussion Question

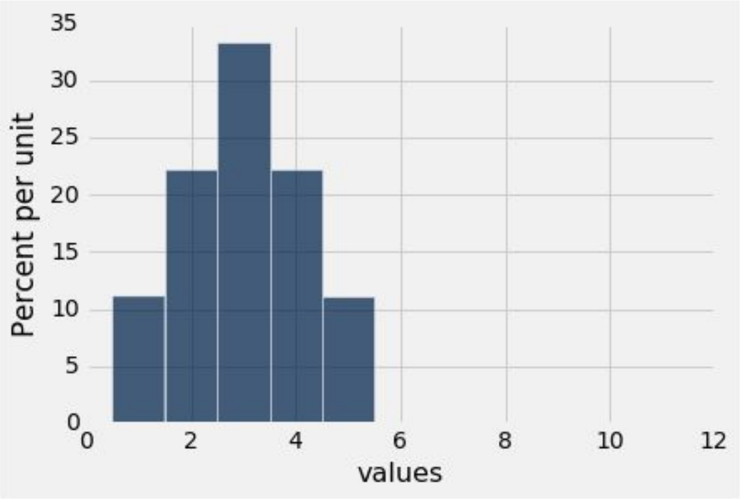
Create a data set that has this histogram. (You can do it with a short list of whole numbers.)

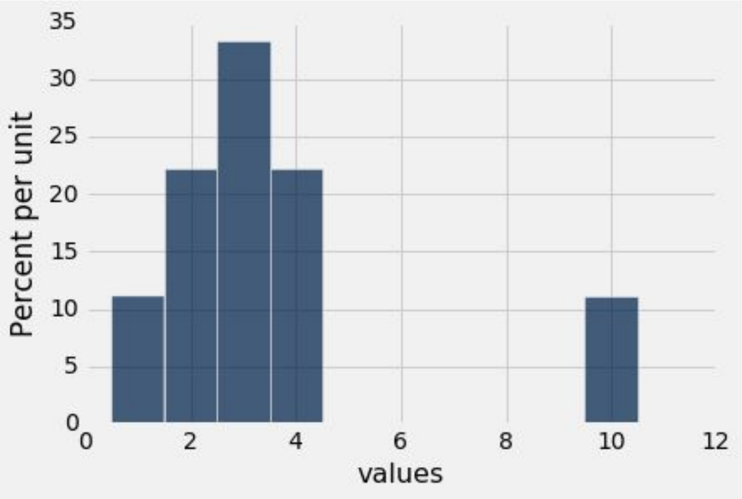
What are its median and mean?



Discussion Question

Are the medians of these two distributions the same or different? Are the means the same or different? If you say “different,” then say which one is bigger.





Comparing Mean and Median

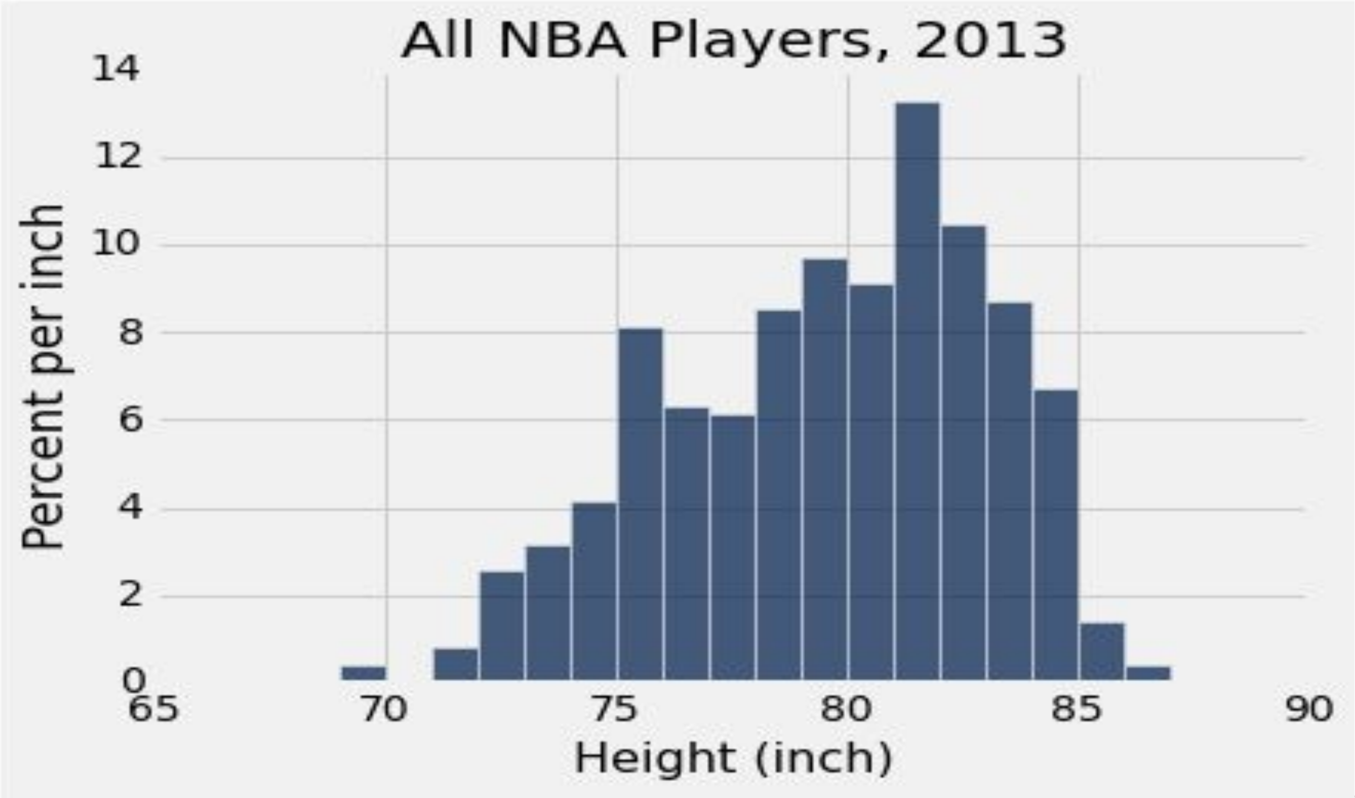
- **Mean:** Balance point of the histogram
- **Median:** Half-way point of data; half the area of histogram is on either side of median
- If the distribution is symmetric about a value, then that value is both the average and the median.
- If the histogram is skewed, then the mean is pulled away from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

Discussion Question

Which is bigger?

(a) mean

(b) median



Standard Deviation

Defining Variability

Plan A: “biggest value - smallest value”

- Doesn’t tell us much about the shape of the distribution

Plan B:

- Measure variability around the mean
- Need to figure out a 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

  Example:

  Sample: 2, 3, 3, 9

  Average/Mean: 4.25

$$\mathrm{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- SD has the same units as the data

Why Use the SD?

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 ± a few SDs”

- **The second reason:**

Coming up in the next lecture.

Chebyshev's Inequality

The Mathematician’s Name

- Chebyshev
- Chebychev
- Chebishov
- Čebyšev
- Tchebichev
- Tchebicheff
- Tschebyscheff
- Tschebyschew
- **Чебышёв**

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

(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
- Chebyshev: At least 96% of the values of *z* are between -5 and 5

(Demo)