



**190F**  
Fall 2018

# Foundations of Data Science

## Lecture 24

---

Center and Spread

# **Announcements**

# Center and Spread

# 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
- Smoothing operator: collect all the contributions in one big pot, then split evenly

(Demo)

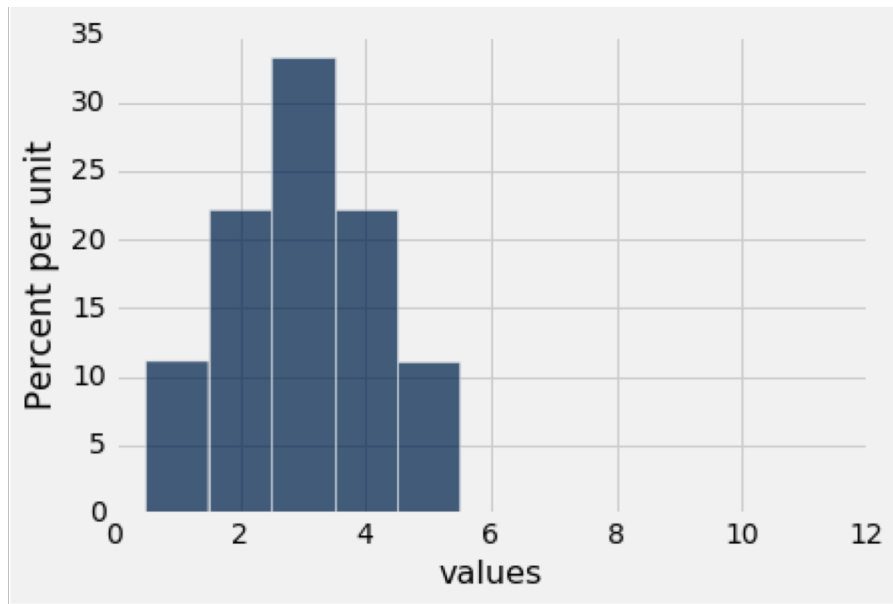
---

# Discussion Question

---

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

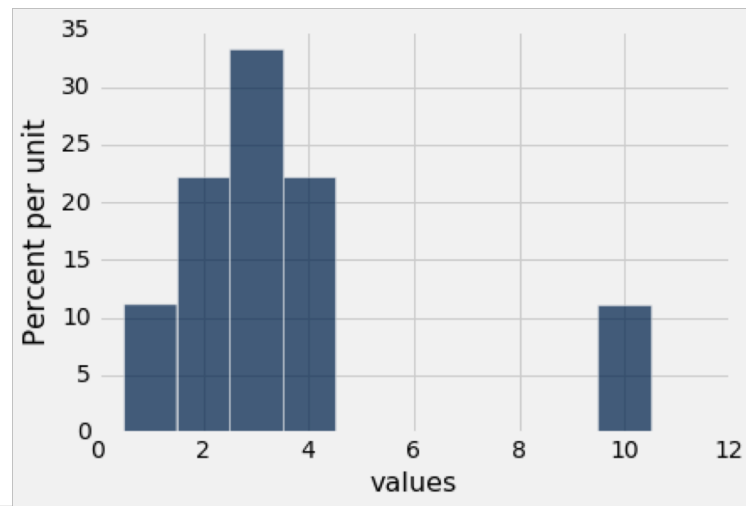
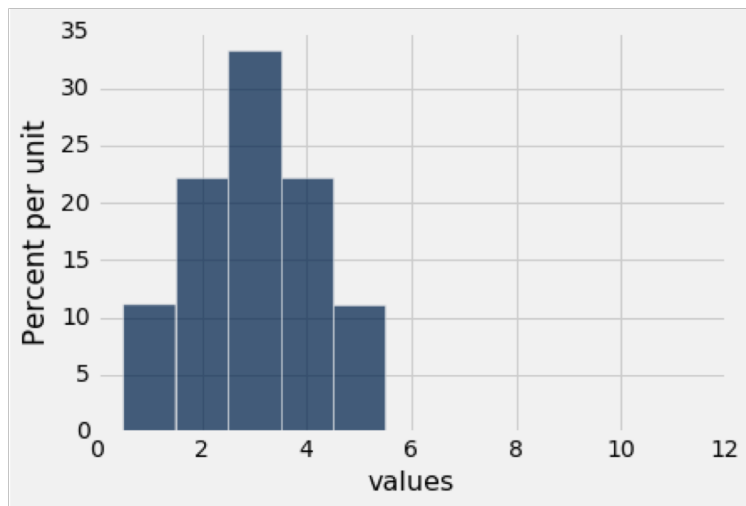
What are its median and mean?



# 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

---

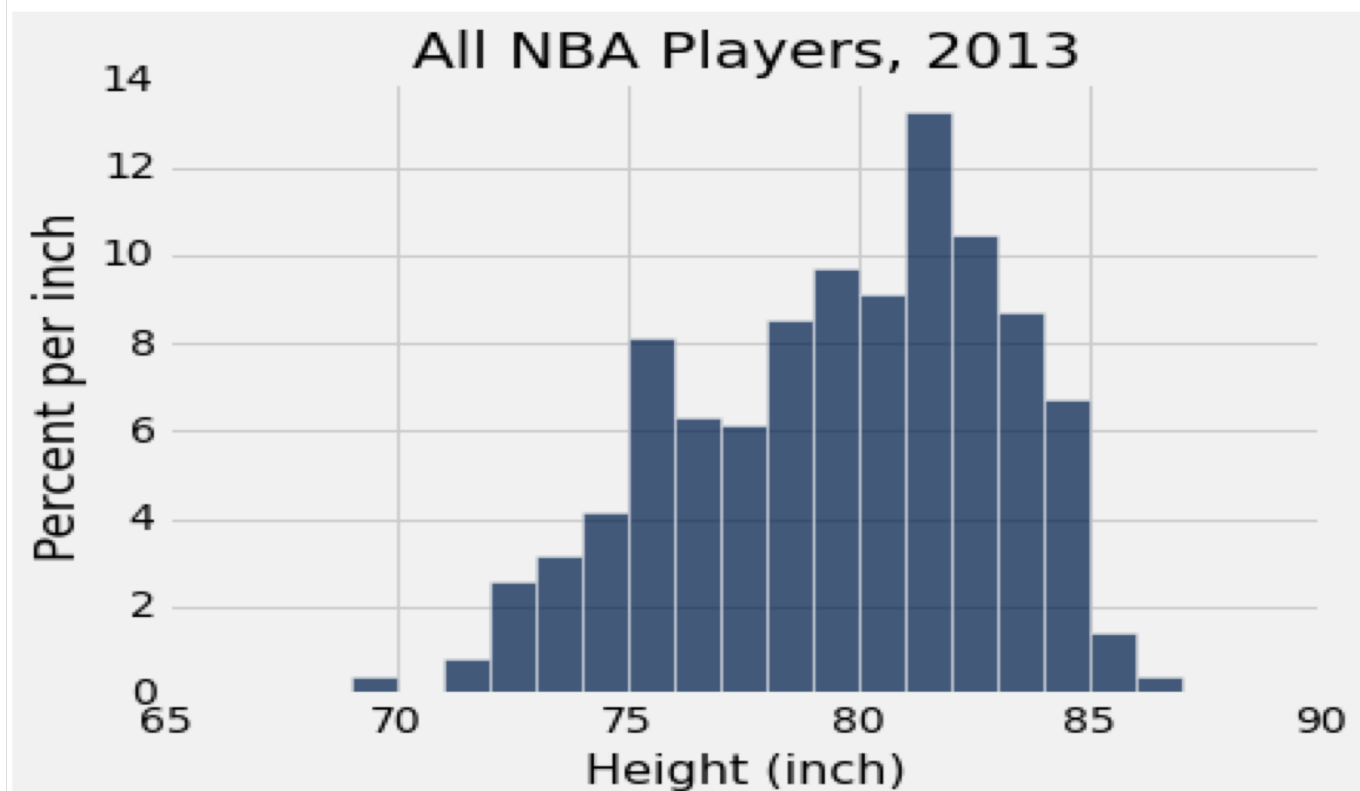
- **Mean:** Balance point of the histogram
  - **Median:** Half-way point of data; half the area of histogram is on either side of median
  - If the distribution is symmetric about a value, then that value is both the average and the median.
  - If the histogram is skewed, then the mean is pulled away 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  
5      4      3                      2                      1
  - 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  $\pm$  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  $\pm$  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

# Standard Units

---

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

# 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)

# The SD and the Histogram

---

- Usually, it's not easy to estimate the SD by looking at a histogram.
  - But if the histogram has a bell shape, then you can.
-

# The SD and Bell-Shaped Curves

---

If a histogram is bell-shaped, then

- the average is at the center
- the SD is the distance between the average and the points of inflection on either side

(Demo)

---