

# YData: An Introduction to Data Science

## Lecture 25: Center and Spread

Elena Khusainova & John Lafferty  
Statistics & Data Science, Yale University  
Spring 2021

Credit: [data8.org](https://data8.org)



# Announcements

- Assignment 08 posted; due 4/12
- Project 2 posted; checkpoint next Friday
- Reminder: Let us know if you want help finding a partner
- Next two weeks will be busy!

# Chance versus Confidence

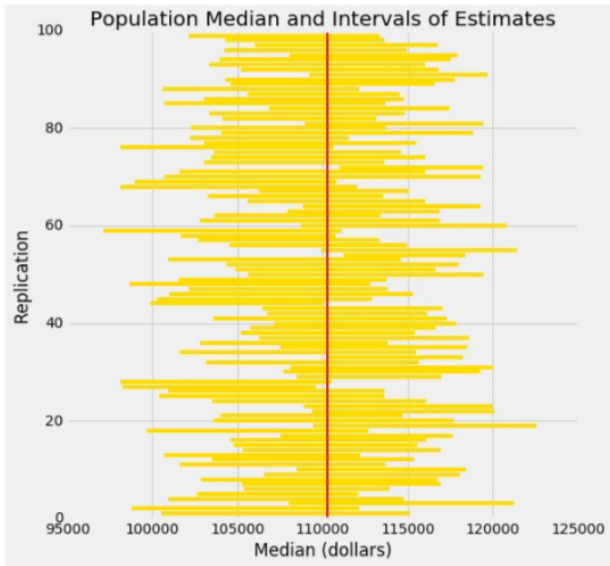
# Is This What a CI Means?

By our calculation, an approximate 95% confidence interval for the median salary of a San Francisco city employee is \$97,588 to \$112,767.

## True or False:

- There is a 0.95 probability that the median salary of San Francisco employees is in the range \$97,588 to \$112,767.

Answer: **False**. The median salary in the population is unknown but it's a constant. It's not random.



# Confidence Intervals For Testing

# Using a CI for Testing

- Null hypothesis: **Population average =  $x$**
- Alternative hypothesis: **Population average  $\neq x$**
- Cutoff for P-value:  $p\%$
- Method:
  - Construct a  $(100-p)\%$  confidence interval for the population average
  - If  $x$  is not in the interval, reject the null
  - If  $x$  is in the interval, can't reject the null

# 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 =  $(2+3+3+9)/4 = 4.25$**

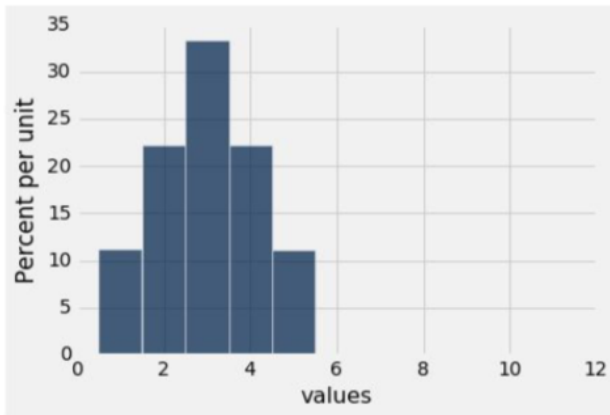
- Need not be a value in the collection
- Need not be an integer even if the data are integers
- Somewhere between min and max, but not necessarily halfway in between
- Same units as the data
- Smoothing operator: collect all the contributions in one big pot, then split evenly

(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?

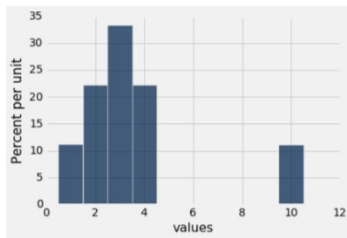
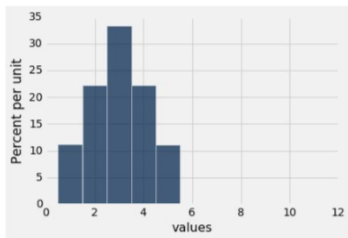


# Discussion Question

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.



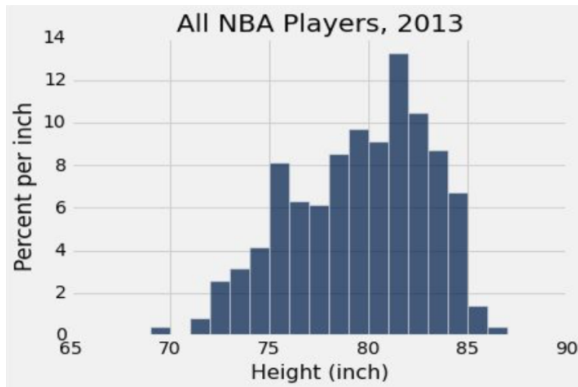
# Comparing Mean and Median

- **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



(DEMO)

# 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
- $SD = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$   
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

$$\text{at least } 1 - 1/z^2$$

# 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**