



# Lecture 23

---

## Center and Spread

Contributions by Vinitra Swamy ([vinitra@berkeley.edu](mailto:vinitra@berkeley.edu)) and Fahad Kamran ([fhdkmrn@berkeley.edu](mailto:fhdkmrn@berkeley.edu))  
Slides created by John DeNero ([denero@berkeley.edu](mailto:denero@berkeley.edu)) and Ani Adhikari ([adhikari@berkeley.edu](mailto:adhikari@berkeley.edu))

# **Vinitra's Thoughts on the Midterm**

# **Announcements**

# **Chance versus Confidence**

# Is This What a CI Means?

---

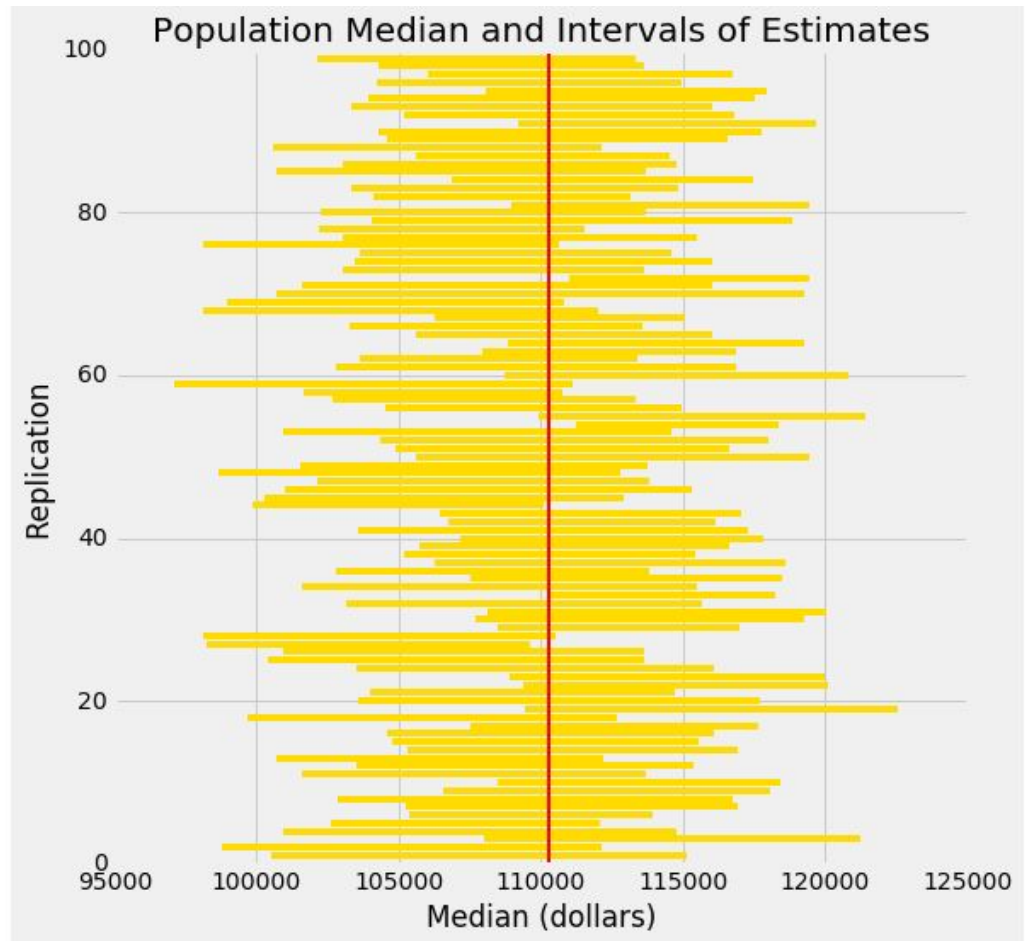
An approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

## True or False:

- There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

**Answer: False.** The average age of the mothers in the population is unknown but it's a constant. It's not random. No chances involved.

---



# 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

(Demo)

---



# 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)

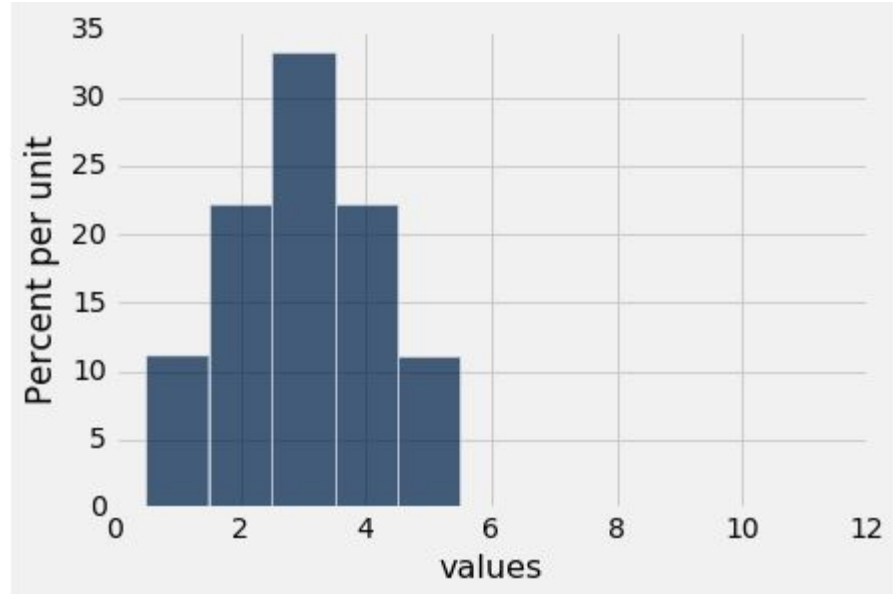
---

# Discussion Question

---

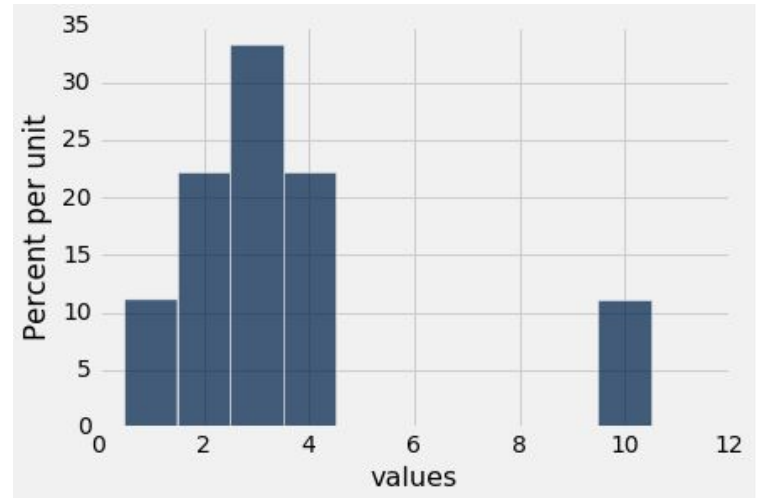
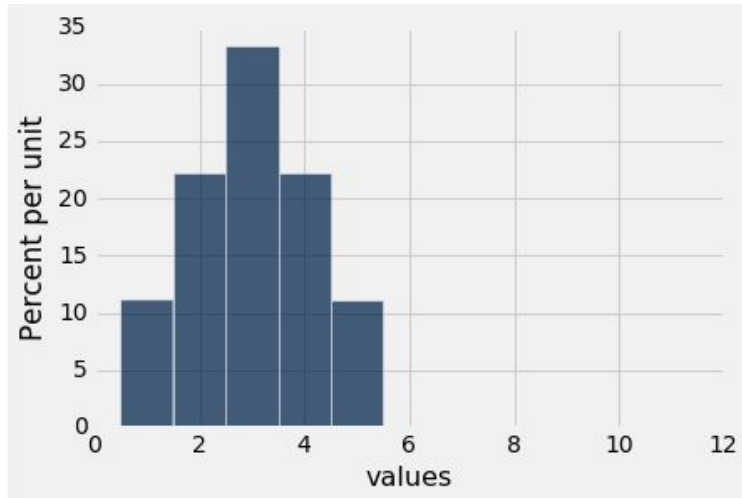
Create a data set that has this histogram. (You can do it with a short list of whole numbers.)

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

---

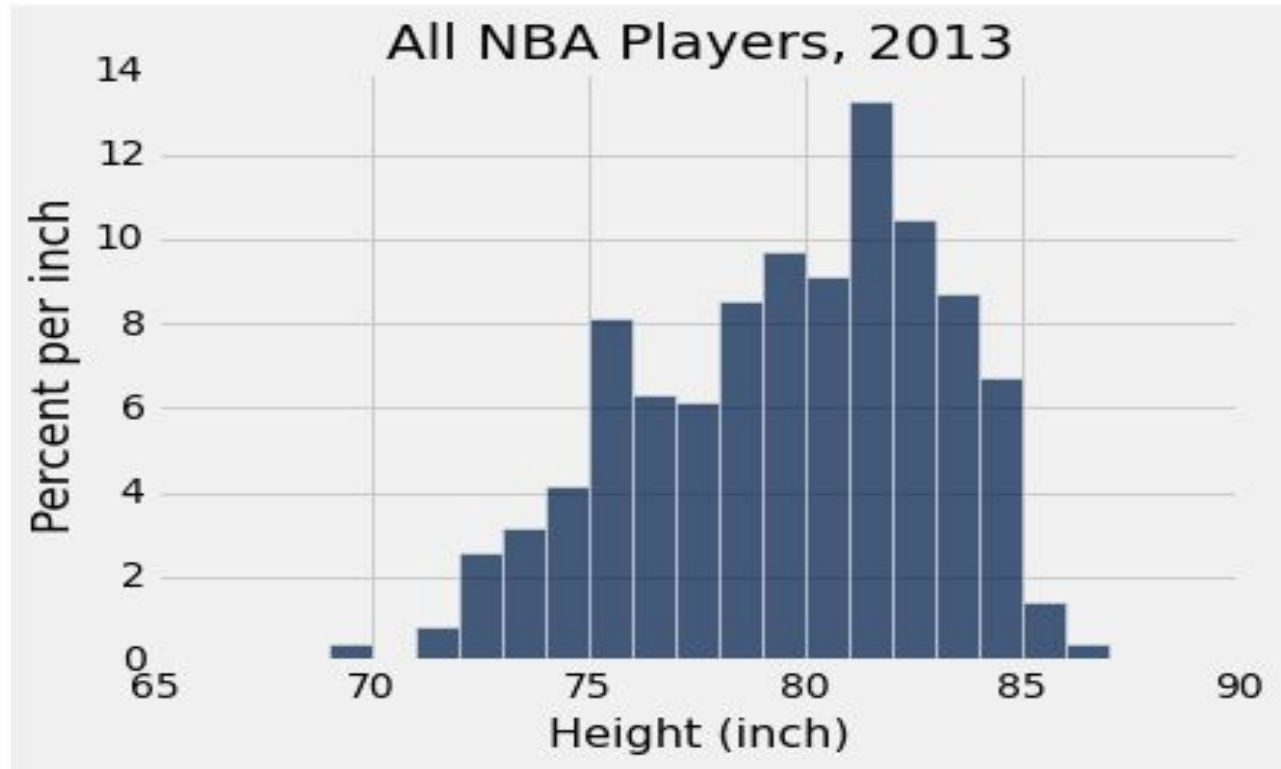
- **Mean:** Balance point of the histogram
  - **Median:** Half-way point of data; half the area of histogram is on either side of median
  - If the distribution is symmetric about a value, then that value is both the average and the median.
  - If the histogram is skewed, then the mean is pulled away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

---



# Chebyshev's Bounds

---

Range	Proportion
average $\pm$ 2 SDs	3/4
average $\pm$ 3 SDs	8/9
average $\pm$ 4 SDs	15/16
average $\pm$ 5 SDs	24/25

**No matter what the  
distribution looks like**

