



# Analysis of Psychological Data

## Lab 3. Hooked on a Distribution: Central Tendency and Variability

Ihnwhi Heo ([iheo2@ucmerced.edu](mailto:iheo2@ucmerced.edu))

Quantitative Methods, Measurement, and Statistics

Website: <https://ihnwhiheo.github.io>

Office: <https://ucmerced.zoom.us/j/2093557522> (Thursday 3:30 - 5:30 pm)



# Some announcements

Exam 1 on February 15 (next Tuesday)

Big and red scantron forms

Pencil and eraser to fill in scantrons

Blank scratch paper

Class-approved calculator

One 8.5 x11 crib sheet (i.e., a concise set of notes, could be double-sided)



# Some announcements

## Big and red scantron forms

**ParScore® STUDENT ENROLLMENT SHEET**

**INSTRUCTOR:** Only write your lab section number or TA's name in this area

**CLASS:**

**HOUR/DAY:**

**DIRECTIONS**

- MAKE DARK MARKS
- ERASE COMPLETELY TO CHANGE
- EX.

**ID NUMBER** 100098765

**PHONE NUMBER** LEAVE THIS BLANK

**LAST NAME** CARLOS **FIRST NAME** JOSE **M.I.**

**FEED THIS DIRECTION**

**ParScore® TEST FORM**

**ID NUMBER** 100098765

**TEST FORM** 001

**EXAM #** 001

This information will be shown on the first page of your exam sheet

**DIRECTIONS**

- MAKE DARK MARKS
- ERASE COMPLETELY TO CHANGE
- EX.

You will have 50 multiple choice questions. Answer them here.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

101 T F 102 T F 103 T F 104 T F 105 T F 106 T F 107 T F 108 T F 109 T F 110 T F 111 T F 112 T F 113 T F 114 T F 115 T F 116 T F 117 T F 118 T F 119 T F 120 T F 121 T F 122 T F 123 T F 124 T F 125 T F 126 T F 127 T F 128 T F 129 T F 130 T F 131 T F 132 T F 133 T F 134 T F 135 T F 136 T F 137 T F 138 T F 139 T F 140 T F 141 T F 142 T F 143 T F 144 T F 145 T F 146 T F 147 T F 148 T F 149 T F 150 T F 151 T F 152 T F 153 T F 154 T F 155 T F 156 T F 157 T F 158 T F 159 T F 160 T F 161 T F 162 T F 163 T F 164 T F 165 T F 166 T F 167 T F 168 T F 169 T F 170 T F 171 T F 172 T F 173 T F 174 T F 175 T F 176 T F 177 T F 178 T F 179 T F 180 T F 181 T F 182 T F 183 T F 184 T F 185 T F 186 T F 187 T F 188 T F 189 T F 190 T F 191 T F 192 T F 193 T F 194 T F 195 T F 196 T F 197 T F 198 T F



# What are we going to do?

Recap to give you a big picture

Central tendency & variability

(Primers on the formula included)

Group activity

Q&A session



# Reflecting upon three weeks

Fundamentals (statistics, variables)

Some basic stat terms (population, parameter, sample, statistic)

Focus on variables (type, measurement scale)

Fiddling with variables (visualization, distribution)



# We are playing the game with the distribution

## Distribution

An intuitive way to understand how the values of a variable are spread over

We are interested in describing the distribution

What would be efficient and informative ways?



# How to describe any distributions

## Measures of central tendency

Reflect where 'values of a variable (i.e., distribution)' are centered

**Mean, median, and mode**

## Measures of variability

Reflect how much 'values of a variable (i.e., distribution)' are dispersed

Range, deviation score, sum of squares, **variance, standard deviation**



# Primers on formula

## Notation

- Uppercase Roman letters that are usually near the end of the alphabet denote variables (e.g.,  $X$  and  $Y$ )
  - A numerical subscript represents an individual value of that variable (e.g.,  $X_3$  is the third value of the variable  $X$ )
- Greek letters are used represent population parameters (e.g.,  $\mu$ ,  $\sigma^2$ ,  $\sigma$ )
- Roman letters are used to represent sample statistics (e.g.,  $\bar{X}$ ,  $s^2$ ,  $s$ )
  - An alphabetical subscript represents the corresponding variables (e.g.,  $\sigma_X^2$ ,  $s_X$ )
- Lowercase  $n$  denotes the sample size whereas the uppercase  $N$  indicates the population size



# Primers on formula

## Notation

- $\Sigma$  (the uppercase Greek letter sigma) indicates the operation of summation (i.e., the addition of scores)
  - $\Sigma_{i=1}^5 X_i = X_1 + X_2 + X_3 + X_4 + X_5$
- Numerical exponents indicate multiplying that number itself multiple times
  - $X^5 = X \times X \times X \times X \times X$
- (Positive) square root
  - $\sqrt{X^2} = (X^2)^{\frac{1}{2}} = X$
- Summation and exponents combined
  - $\Sigma_{i=1}^5 (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 + (X_4 - \bar{X})^2 + (X_5 - \bar{X})^2$   
 $= (X_1 - \bar{X}) \times (X_1 - \bar{X}) + (X_2 - \bar{X}) \times (X_2 - \bar{X}) + \dots + (X_5 - \bar{X}) \times (X_5 - \bar{X})$



# Central tendency

## Summary

	<b>Sample</b>	<b>Population</b>
Mean	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	$\mu_X = \frac{\sum_{i=1}^N X_i}{N}$
Median	(When aranged) The $\frac{n+1}{2}$ th value	(When aranged) The $\frac{N+1}{2}$ th value
Mode	The most frequent value	The most frequent value

## Key ideas

- How these measures are related to distributions?
- Which measures are useful and when?
- Calculation

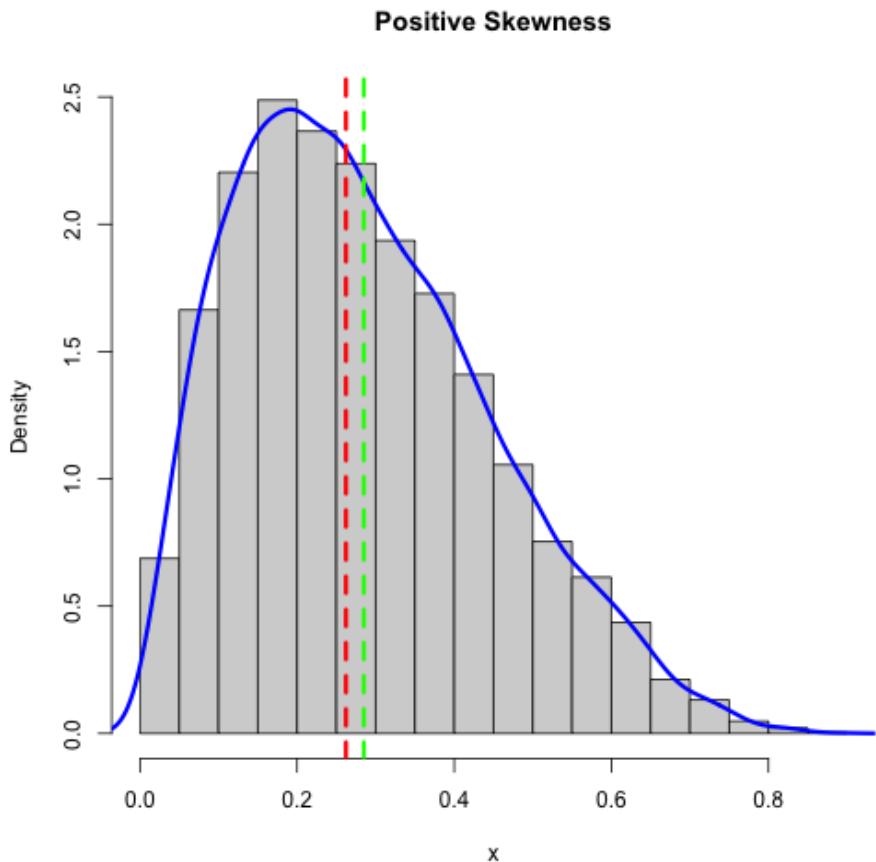


# Central tendency

## Recall Positive Skewness

```
set.seed(322)
x = rbeta(10000, 2, 5)
hist(x, main="Positive Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
abline(v = c(mean(x), median(x)),
       col=c("green", "red"),
       lty=c(2,2), lwd=c(3, 3))
```

- Trailing off toward the right end
- Mode < Median < Mean
- Score distribution of difficult exams



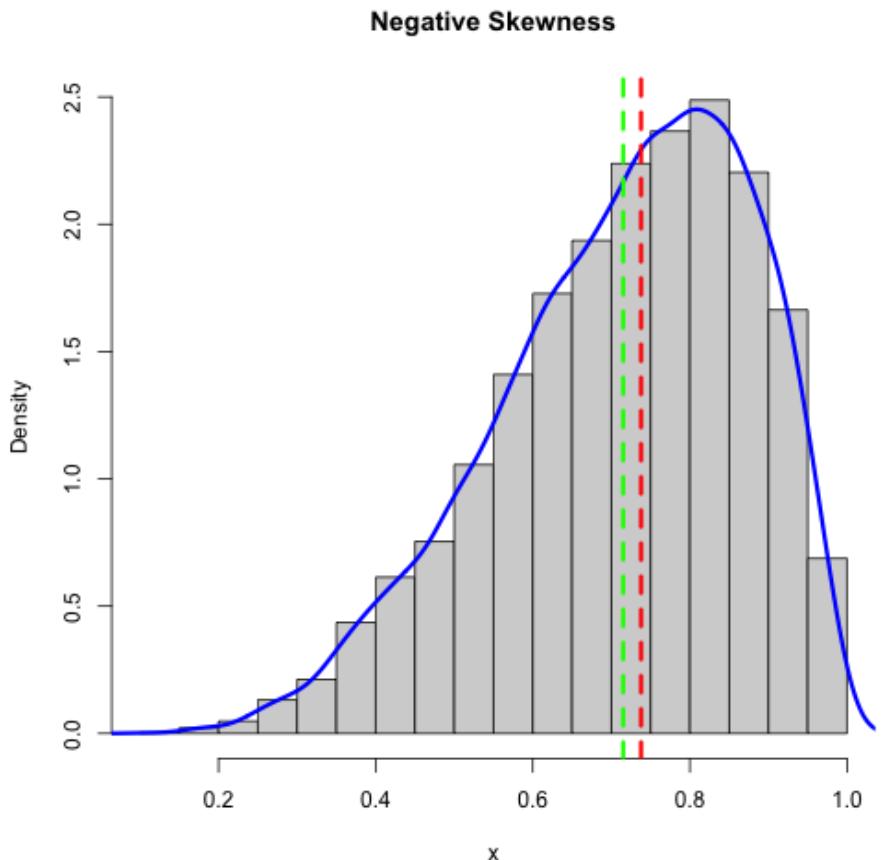


# Central tendency

## Recall Negative Skewness

```
set.seed(322)
x = rbeta(10000, 5, 2)
hist(x, main="Negative Skewness", freq=FALSE)
lines(density(x), col='blue', lwd=3)
abline(v = c(mean(x), median(x)),
       col=c("green", "red"),
       lty=c(2,2), lwd=c(3, 3))
```

- Trailing off toward the left end
- Mean < Median < Mode
- Score distribution of easy exams





# Central tendency

## Mean vs. median

When **outliers** (i.e., extreme values) exist, consider using the median as a central tendency measure

Imagine Elon Musk enters this room when we calculate our monthly income...

## Mode

Useful for variables measured on nominal and ordinal scales

Do you remember the distributional forms such as unimodal and bimodal?



# Central tendency

Medians are resistant to the outliers... and hey means, don't be mean...





# Variability

## Range

$$\max - \min$$

## Deviation score

$$X_i - \bar{X}$$

## Sum of squares

$$\sum_{i=1}^n (\text{deviation score})^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$



# Variability

## Variance

- Sample variance =  $s^2 = \frac{\sum_{i=1}^n (\text{deviation score})^2}{n-1} = \frac{\text{sum of squares}}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$
- Population variance =  $\sigma^2 = \frac{\sum_{i=1}^N (\text{deviation score})^2}{N} = \frac{\text{sum of squares}}{N} = \frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}$

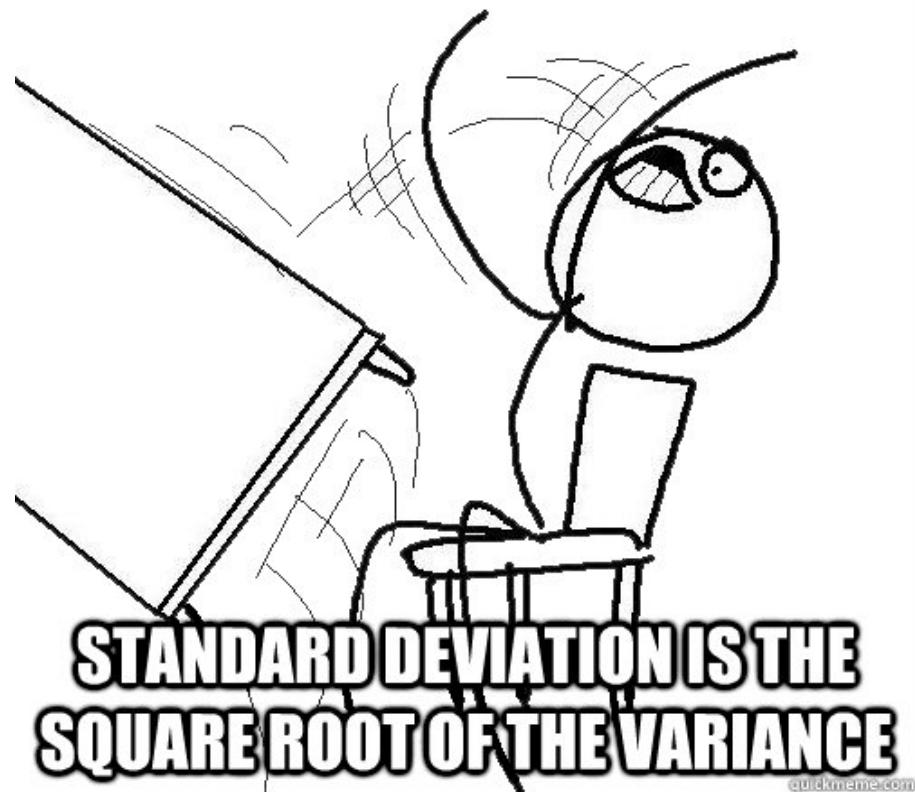
## Standard deviation

- Sample standard deviation =  $\sqrt{\text{sample variance}} = \sqrt{s^2} = s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$
- Population standard deviation =  $\sqrt{\text{population variance}} = \sqrt{\sigma^2} = \sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}}$



# Variability

Remember!





# Variability

## Some tips on interpreting the standard deviation

### Average distance from the mean

→ Say, the average height of males in the Netherlands is 6 ft, and the standard deviation is 2.5 inches. What do these central tendency and variability measures mean?

What if the standard deviation is zero?



# Central tendency and variability

## Summary

	<b>Sample</b>	<b>Population</b>
Mean	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	$\mu = \frac{\sum_{i=1}^N X_i}{N}$
Median	(When aranged) The $\frac{n+1}{2}$ th value	(When aranged) The $\frac{N+1}{2}$ th value
Mode	The most frequent value	The most frequent value
Variance	$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}$
Standard deviation	$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N}}$



# Group activity

## Case study

David, a statistician at the United States Air Force, has collected the data on pilots' satisfaction with aviation training. With the data collected, his research team would like to evaluate the current training system and know how to improve it in the future. See the collected data following:

25, 35, 18, 20, 15, 5, 25, 10, 22, 25

## Can you answer below?

- What are the mean, the median, and the mode?
- What are the variance, and the standard deviation?
- What is the relative frequency of the score 25?
- What is the cumulative frequency of the score equal to 18 or lower?



# Q&A session





# Before you go home...

Lab materials are available at

<https://github.com/IhnwhiHeo/PSY010>

Any questions or comments?

Office hours or my email



# Thanks! Good luck with your exam!

