BOSTON UNIVERSITY

PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 4: Central Tendency & Variability

Updates and reminders

- Homework 1 is due (via Blackboard) on Monday, September 15 at 11:59 PM.
  - Homework 1 mostly covers material from Lectures 1 - 3, with a small amount of material from today's lecture.
  - Please check **today** to make sure you know how to use R Markdown and that you can successfully knit your homework to an html file.
  - You must submit an **html** file to Blackboard. Do not submit a .Rmd file or any other file type.
  - Late homework will be penalized 3 points per day.
  - I have additional office hours tomorrow (Friday) from 10 - 11 a.m.
  - We may not be available to help you resolve last-minute technical issues.

Updates and reminders (continued)

- Exam 1 is scheduled for **Thursday, September 25** during our regular class time.
  - Exam 1 will cover material from Lectures 1 - 5 and consist of multiple choice questions.
  - After Lecture 5, I will post a review sheet on Slack that has all the topics that will be covered on the exam.
  - The exam will be closed book/notes, but you may print the provided review sheet (and write on it with your own notes) or you may bring one 8.5" x 11" sheet of paper with handwritten notes (both sides).
  - No calculator is needed.
- We will have a review session for Exam 1 on Tuesday, September 23 during our regular class time. Please bring your questions!

Central Tendency

- The central tendency describes the "center" of a dataset.
- It identifies the value that scores cluster around.
- The three common measures of central tendency are mean, median, and mode.

These are descriptive statistics.

Mean, median, and mode

Mean: The arithmetic average

Median: The middle score

Mode: The most common score

The Mean: The Arithmetic Average

Warning: Equation incoming!

The formula for the mean is:

$$M = \frac{\sum_{i=1}^{N} X_i}{N}$$

Let's break this down:

M = the mean. This is what we are trying to calculate.

$\sum$ = the summation symbol. This means "add up all the following values."

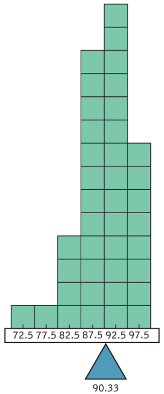
$\sum_{i=1}^{N}$ = this tells us to start with the first score (i=1) and continue adding scores until we reach the Nth score (the last score).

X_i = the value of the ith score.

Calculating the mean

$$M = \frac{\sum_{i=1}^{N} X_i}{N}$$

1. Add up all the scores.
2. Count the total number of scores.
3. Divide the sum of the scores by the total number of scores.



Nolan/Heinzen, *Statistics for the Behavioral Sciences*, 5e, © 2020 Worth Publishers

Some people like to think of the mean as a fulcrum balancing the two sides of a distribution.

The mean: Practice

$$M = \frac{\sum_{i=1}^{N} X_i}{N}$$

Five students take an exam and receive the following scores: 80, 85, 90, 95, 100. What is the mean exam score?

Answer: 90. 80 + 85 + 90 + 95 + 100 = 450. There are 5 scores, so 450 / 5 = 90.

A sixth student takes the exam and receives a score of 80. You compute the mean again:

```
90 + 80 = 170
170 / 2 = 85

Is the mean now 85?
```

Answer: No. The new mean is 86.7:

$$80 + 85 + 90 + 95 + 100 + 80 = 520$$

$$520/6 = 86.7$$

The mean: Symbols

- The mean of a sample is a **statistic**.
- The mean of a population is an estimated **parameter**.
- Typically:
  - Symbols are *italicized*. Numbers are not.
  - Latin letters are used for statistics (numbers calculated from samples).
  - Greek letters are used for parameters (numbers estimated for populations).
- The mean of a sample is denoted by M or $\bar{X}$ (X-bar).
- The mean of a population is denoted by μ (the Greek letter "mu").

The mean: Symbols

Sample Mean (but capitalize it!)

- Can be calculated directly.

Population Mean

- Usually estimated.

The Median

- The median is the middle score when all scores are ordered from lowest to highest (ascending).
- If there is an odd number of scores, the median is the middle score.
- If there is an even number of scores, the median is the mean of the two middle scores.
- The median represents the 50th percentile of the data.

The Median

Five students take an exam and receive the following scores: 92, 85, 90, 95, 100. What is the median exam score?

Answer: 92.

First, we order the scores: 85, 90, 92, 95, 100.

The middle score is 92.

A sixth student takes the exam and receives a score of 0. What is the median exam score now?

Answer: 91. The ordered scores are now: 0, 85, 90, 92, 95, 100. The two middle scores are 90 and 92, and their mean is 91.

The Mode

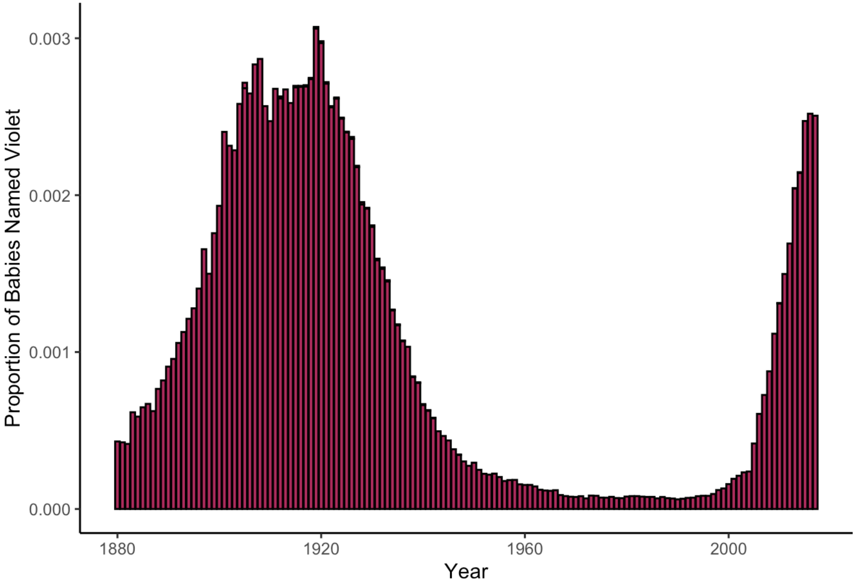
- The mode is the most common score in the dataset.
- A distribution may be unimodal (one mode), bimodal (two modes), or multimodal (many modes).
- The mode is not always useful if many values occur with the same frequency.

The Mode

- A **unimodal** distribution has one mode.
- A **bimodal** distribution has two modes.
- A **multimodal** distribution has multiple modes.

Example: Bimodal distribution

- In bimodal and multimodal distributions, the mean and median are *not* representative of the data.



Mean, median, and mode

If a distribution is positively skewed, which measure of central tendency will be the largest? Which will be the smallest?

Answer: The mean will be the largest, and the mode will be the smallest.

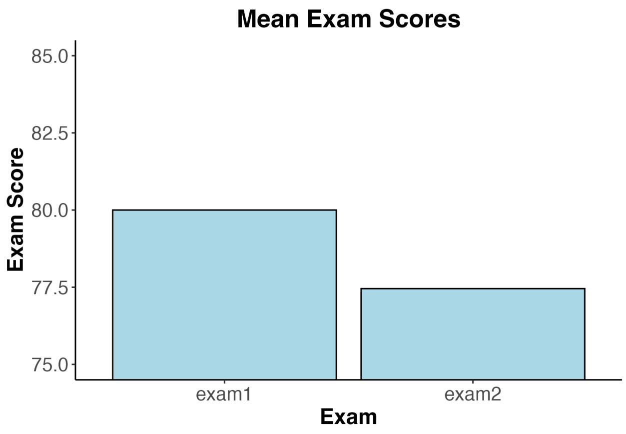
Comparing Mean, Median, and Mode

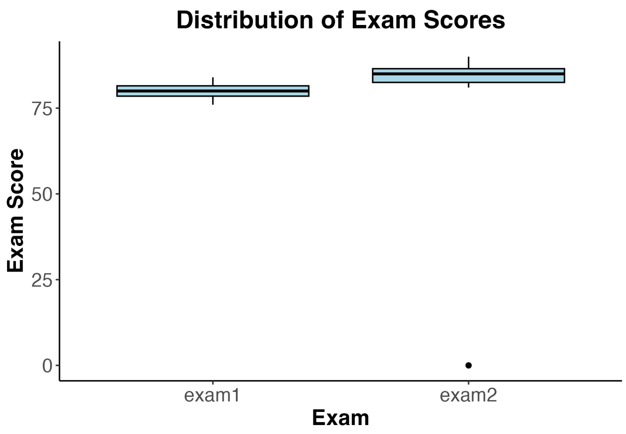
- In a normal distribution, the mean, median, and mode are equal!
- In a negatively skewed distribution, the mean is less than the median, which is less than the mode.
- In a positively skewed distribution, the mode is less than the median, which is less than the mean.

Outliers

- Outliers are extreme values that differ greatly from the rest of the data.
- Outliers can distort the mean, making it less representative of the dataset.
- The median is less affected by outliers and can be a better measure of central tendency in skewed data.

Visualizing outliers

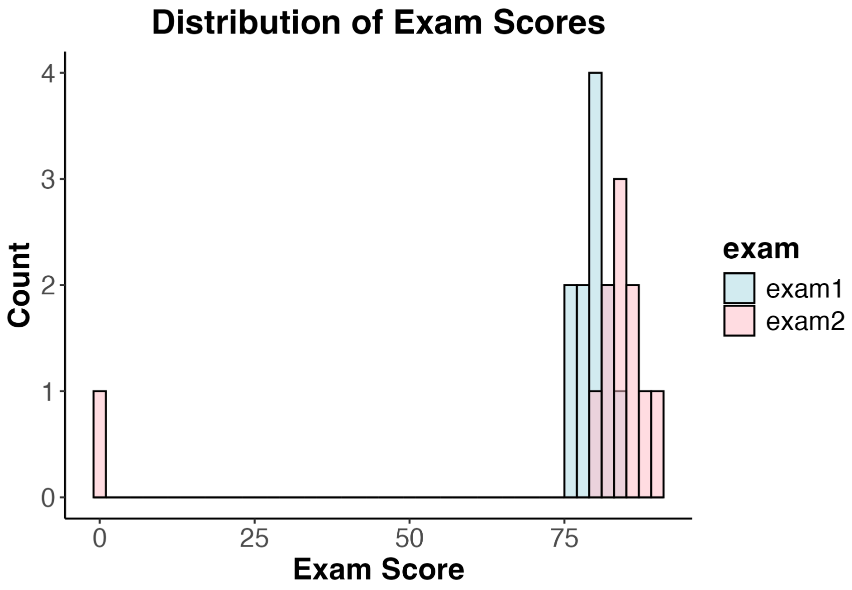




Which exam was harder?

Which exam was harder?

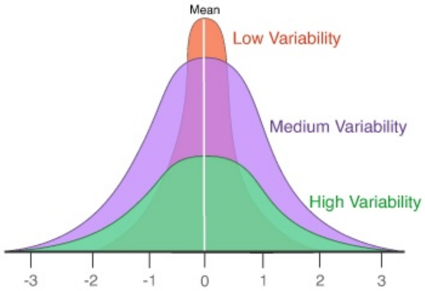
Visualizing outliers (continued)



Which exam was harder?

Variability

- Besides a dataset's center or central tendency, we also often want to know how spread out the values are.
- **Variability** describes the spread of a distribution.
- Distributions with higher variability show greater spread between scores.



How could we measure spread?

Variability

How could we measure spread?

Three main **descriptive statistics** are used to measure variability:

1. Range: the difference between the highest and lowest score
2. Variance: average square deviation from the mean
3. Standard deviation: the square root of the variance

Range

Equation time!

$$Range = X_{max} - X_{min}$$

- The range is the difference between the highest and lowest score.
- For example, if the highest quiz score is 90 and the lowest is 70, the range is 20.
- The range is simple to compute, but it only depends on two scores and can be distorted by outliers.

Interquartile Range

- The interquartile range (IQR) measures the distance between the 25th percentile (Q1) and 75th percentile (Q3).
- The IQR represents the middle 50% of the data.
- Because it is less influenced by outliers, the IQR is often a more robust measure of variability.

Practice: Computing the interquartile Range

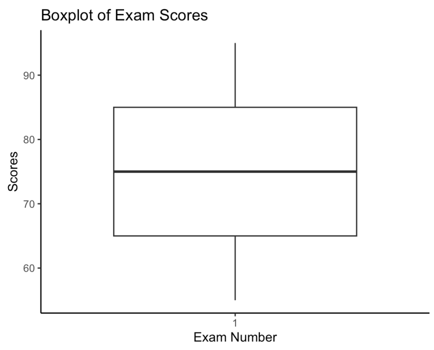
- The following are the scores of 9 students on an exam: 55, 60, 65, 70, 75, 80, 85, 90, 95. What is the interquartile range (IQR) of these scores?

1. Order the scores (they are already ordered).
2. Find the median (75).
3. Split the data into two halves: lower half (55, 60, 65, 70) and upper half (80, 85, 90, 95).
4. Find Q1 (the median of the lower half) = (60 + 65) / 2 = 62.5.
5. Find Q3 (the median of the upper half) = (85 + 90) / 2 = 87.5.
6. Compute the IQR: IQR = Q3 - Q1 = 87.5 - 62.5 = **25.**

Boxplots revisited

- Boxplots visually represent the median, quartiles, and potential outliers in a dataset.
- The box represents the interquartile range (IQR), with a line at the median.
- "Whiskers" extend to the smallest and largest values within 1.5 * IQR from the quartiles.
- Points outside this range are considered potential outliers.

```
#make list of exam scores
exam_scores <- c(55, 60, 65, 70, 75, 80, 85, 90, 95)

#put in dataframe
exam_scores_df <- data.frame(exam_scores)
exam_scores_df$exam <- factor(1)  #add variable for x axis

#use ggplot to make boxplot
ggplot(exam_scores_df, aes(x = exam, y = exam_scores)) +
  geom_boxplot() +
  labs(title = "Boxplot of Exam Scores", y = "Scores")
```



Variance

What if we want a measure of spread influenced by ALL scores?

- We can compute the distance each score is from the mean (the deviation), and take the average of that.
- But... if we add up all the deviations, they will always equal zero!

Example

- Scores: 70, 80, 90
- Mean: 80; Deviations: -10, 0, +10
- Sum of deviations: -10 + 0 + 10 = 0

To solve this, we square each deviation (to make them positive) before summing them!

Variance (continued)

Variance (continued)

Sample Variance

Things are about to get tricky...

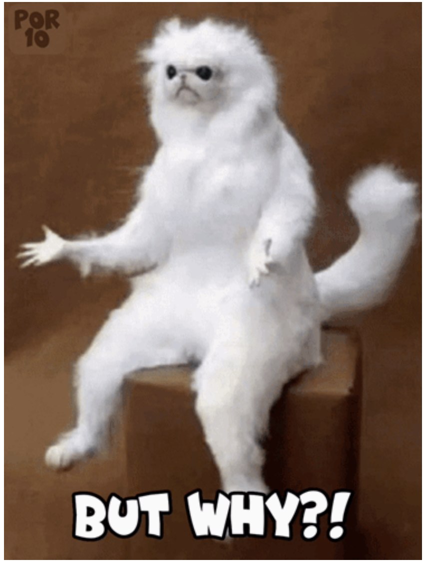
- The formula for the variance of a **SAMPLE** is:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - M)^2}{N - 1}$$

What changed?

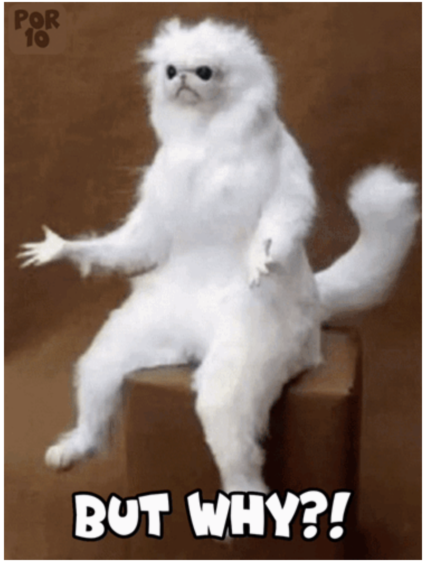
1. We use M (the sample mean) instead of μ (the population mean).
2. We divide by $N - 1$ instead of N.
3. We use s^2 (the sample variance) instead of σ^2 (the population variance).

Why do we divide by N-1 instead of N?



Why N-1?

Why do we divide by N-1 instead of N?



- When we use the sample mean (M) instead of the population mean (μ), we are using an estimate that is **based on the sample data.**
- The sample mean tends to be closer to the sample scores than the true population mean would be.
- This can lead to an underestimation of the true population variance.
- Dividing by N-1 instead of N provides a better estimate of the population variance, especially for small sample sizes.

Why N-1? (Continued)

Suppose we have a population with the following scores: 70, 75, 80, 85, 90.

1. Compute the population mean.
2. Compute the population variance.

Now imagine we take a sample of 3 scores from this population: 70, 75, 85.

3. Compute the sample mean.
4. Compute the sample variance using N in the denominator.
5. Compute the sample variance using N-1 in the denominator.

Code to the rescue!

```
# Population data
population_scores <- c(70, 75, 80, 85, 90)
population_mean <- mean(population_scores)
population_variance <- sum((population_scores - population_mean)^2) / length(population_scores)

# Sample data
sample_scores <- c(70, 75, 85)
sample_mean <- mean(sample_scores)
sample_variance_N <- sum((sample_scores - sample_mean)^2) / length(sample_scores)
sample_variance_N_minus_1 <- sum((sample_scores - sample_mean)^2) / (length(sample_scores) - 1)
```

- Population Variance: 50
- Sample Variance (N): 38.89
- Sample Variance (N-1): 58.33

Standard Deviation

- The standard deviation is the square root of the variance.
- The formula for the standard deviation of a population is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

- The formula for the standard deviation of a sample is:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - M)^2}{N - 1}}$$

- It represents the typical amount that each score deviates from the mean.
- Larger standard deviations indicate greater spread, while smaller ones indicate that scores cluster more closely around the mean.

Standard deviation vs. variance

- The standard deviation is often more interpretable than the variance because it is in the same units as the original data.
- For example, if exam scores are measured in points, the standard deviation will also be in points, while the variance will be in points squared.
- Both the variance and standard deviation provide valuable information about the spread of a dataset, but the standard deviation is often preferred for its interpretability.

Variability in R

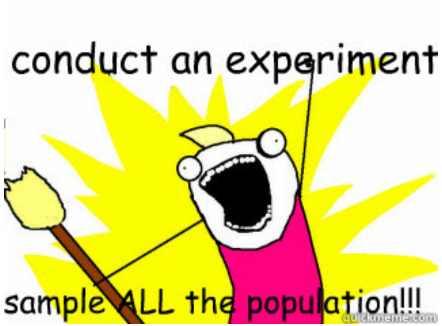
- R actually has built-in functions to compute variance and standard deviation.

```r
# Sample data
scores <- c(70, 75, 80, 85, 90)

# Compute sample variance
sample_variance <- var(scores)

# Compute sample standard deviation
sample_sd <- sd(scores)
```

Note: The `var()` function in R computes the sample variance (dividing by N-1), and the `sd()` function computes the sample standard deviation.

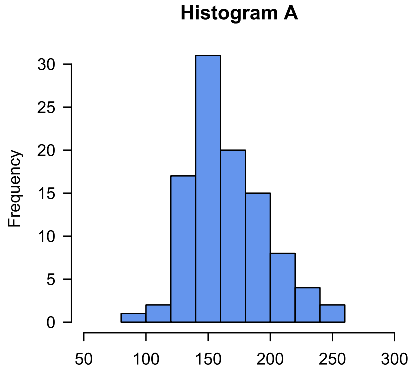
Thought question: Why would the built in 'var' and 'sd' functions compute the sample variance and standard deviation instead of the population versions?

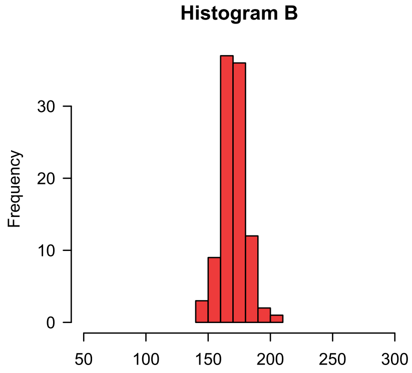
Answer: In practice, we often work with samples rather than entire populations. Therefore, R's built-in functions are designed to compute sample statistics by default.



Variability: Practice

Which histogram displays data with higher variance?



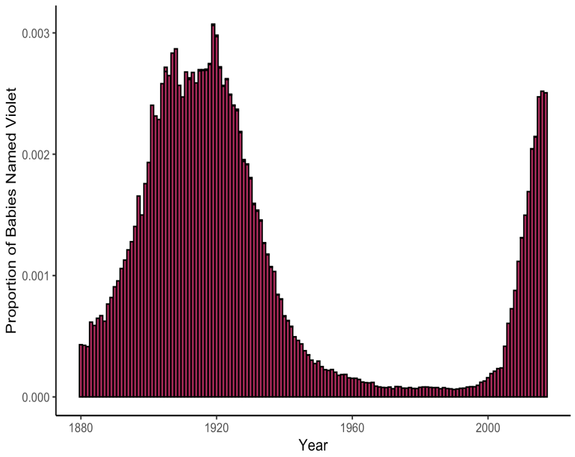


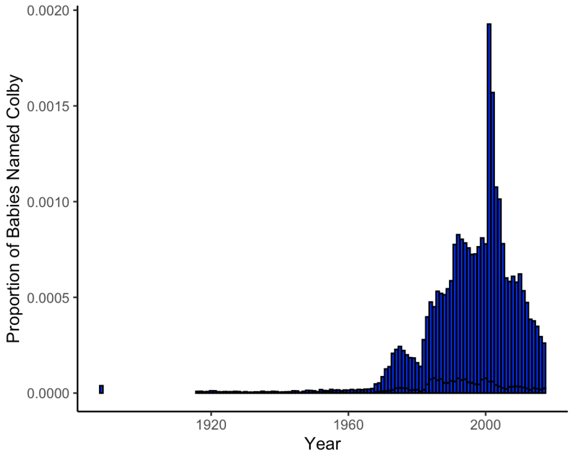
Answer: The histogram on the left has higher variance because the scores are more spread out from the mean.

Which measure of variability would be most affected by an outlier?

Variability: Practice (Continued)

Which distribution likely has a greater standard deviation?





Answer: The "Violet" distribution likely has a greater standard deviation because its scores are more spread out from the mean.

That’s all for today!

Remember, Homework 1 is due Monday at 11:59 PM.