Math Workshop

Christina Walker 2024-08-14

Why Do I Need to Learn This?

- Doing your own research: Math & statistics are essential tools to design empirical studies, analyze datasets, and draw conclusions. It is useful for not only quantitative studies but also qualitative (e.g., categorizing responses/themes, triangulation, identifying patterns)
- Understanding others' research: A foundation in math and statistics will let you critically evaluate and interpret others' research

Descriptive Statistics: How researchers describe and summarize data

Imagine you are at a party with a group of people and you want to get an idea of the guests without talking to each one individually. Descriptive statistics are like taking a group photo of everyone at the party. The photo won't tell you about each person's life story, but it will give you an overview of the crowd.

Central Tendency Measures: Provide insight into the average or typical value of a dataset

Mean: The average of a dataset, calculated by summing all values and dividing by the number of observations. It is like finding the average height of everyone in the photo. It gives you a general idea, but does not tell you about individual heights.

Formula:

$$\text{Mean} = \frac{\text{Sum of all numbers}}{\text{Total number of values}}$$

Example:

Let’s find the mean of the following set of numbers: 5, 7, 9, 11, 13.

1. First, sum up all the numbers:

$$5 + 7 + 9 + 11 + 13 = 45$$

2. Next, divide by the total number of values (in this case, 5):

$$\text{Mean} = \frac{45}{5} = 9$$

Result: The mean of the set of numbers is 9.

Median: The middle value in a dataset when it is ordered from lowest to highest. This is like lining everyone up by height and finding the person in the middle. Half of the people are taller and half are shorter than this person.

1. Arrange the numbers in numerical order.
2. If there is an odd number of values, the median is the middle number.
3. If there is an even number of values, the median is the average of the two middle numbers.

Example: Find the median of the following set of numbers:

$$5, 7, 12, 15, 21, 23, 23, 40$$

1. Arrange the numbers in numerical order:

$$5, 7, 12, 15, 21, 23, 23, 40$$

(They are already in order.)

2. Since there are 8 numbers (an even number), we need to find the average of the 4th and 5th numbers.

$$(15 + 21) \div 2 = 36 \div 2 = 18$$

Result: The median of the set of numbers is **18.**

Skewness: The distribution is not symmetric because one tail is longer than the other. Whatever way the tail is longer is the direction of skew.

- Right-skewed: mean > median
- Left-skewed: mean < median

Mode: The value that appears most frequently in a dataset. If most people at the party are wearing blue, blue would be the mode or the most common color worn.

1. List the numbers in ascending order.
2. Count how many times each number appears.
3. The number(s) that appear(s) the most is the mode.

Example: Find the mode of the following set of numbers:

$$6, 9, 11, 6, 7, 9, 11, 6, 9$$

1. Arrange the numbers in numerical order:

$$6, 6, 6, 7, 9, 9, 9, 11, 11$$

2. Count how many times each number appears:

- **6** appears **3** times
- **7** appears **1** time
- **9** appears **3** times
- **11** appears **2** times

3. Identify the number(s) that appear(s) the most:

Both **6** and **9** appear **3** times, which is more frequent than any other number.

Result: The modes of the set of numbers are **6** and **9**. Since this data set has two modes, it is called **bimodal**. If there were more than two modes, the data set would be called **multimodal**.

Dispersion: Helps researchers understand the spread or variability of the data points

Range: The difference between the highest and lowest values in a dataset. This is like comparing the height of the tallest person to the shortest person at the party. The range provides a quick measure of the spread of the data, but it can be influenced by outliers or extreme values. Other measures of dispersion, like variance or standard deviation, can provide more detailed insights into the spread of data.

1. Identify the smallest number in the set.
2. Identify the largest number in the set.
3. Subtract the smallest number from the largest number.

Example: Find the range of the following set of numbers:

$$8, 15, 3, 22, 11, 5$$

1. Identify the smallest number:

$$3$$

2. Identify the largest number:

$$22$$

3. Subtract the smallest number from the largest number:

$$22 - 3 = 19$$

Result: The range of the set of numbers is **19.**

Variance: The average of the squared differences from the mean. It provides a measure of how spread out the data is

Standard Deviation: The square root of the variance, which represents the average amount of variation or dispersion in a dataset. We measure "closeness" in terms of standard deviation units. This is like seeing how spread out everyone's height is from the average. If most people are around the same height, the spread is low. But if there are many very tall and very short people, the spread is high.

1. Calculate the mean (average) of the data set.
2. Subtract the mean from each data point and square the result.
3. Calculate the mean of these squared differences.
4. Take the square root of the result from step 3.

Formula:

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

Where:

- σ is the standard deviation.
- N is the number of observations.
- x_i is each individual observation.
- μ is the mean of the observations.

This formula calculates the standard deviation for a population. If you’re calculating the standard deviation for a sample (standard error), you’d use $\frac{1}{N-1}$ instead of $\frac{1}{N}$.

Example: Find the standard deviation of the following set of numbers:

$$4, 8, 6, 5, 9, 2$$

1. Calculate the mean:

$$(4 + 8 + 6 + 5 + 9 + 2) \div 6 = 34 \div 6 = 5.67$$

2. Subtract the mean from each data point and square the result:

$$(4 - 5.67)^2 = 2.78$$

$$(8 - 5.67)^2 = 5.44$$

$$(6 - 5.67)^2 = 0.11$$

$$(5 - 5.67)^2 = 0.44$$

$$(9 - 5.67)^2 = 11.11$$

$$(2 - 5.67)^2 = 13.44$$

3. Calculate the mean of these squared differences:

$$(2.78 + 5.44 + 0.11 + 0.44 + 11.11 + 13.44) \div 6 = 5.55$$

4. Take the square root:

$$\sqrt{5.55} = 2.36$$

(rounded to two decimal places)

Result: The standard deviation of the set of numbers is **2.36**.

Standard Error: How far errors are from the estimate, the standard deviation of the sampling distribution.

The 5-Number Summary: provides a concise description of the distribution of a dataset. It consists of five values:

1. **Minimum**: The smallest value in the dataset.
2. **First Quartile (Q1)**: The value below which 25% of the data falls.
3. **Median (Q2)**: The middle value that separates the higher half from the lower half of the dataset.
4. **Third Quartile (Q3)**: The value below which 75% of the data falls.
5. **Maximum**: The largest value in the dataset.

Steps to Find the 5-Number Summary:

1. Sort the data in ascending order.
2. Identify the minimum and maximum values.
3. Calculate the median.
4. Calculate Q1: the median of the first half of the data.
5. Calculate Q3: the median of the second half of the data.

Example: Find the 5-number summary for the following set of numbers:

$$3, 7, 8, 5, 12, 14, 21, 13, 18$$

1. Sort the data:

$$3, 5, 7, 8, 12, 13, 14, 18, 21$$

2. Identify the minimum and maximum:

Minimum: 3

Maximum: 21

3. Calculate the median:

There are 9 numbers, so the median is the 5th number: 12.

4. Calculate Q1:

The first half of the data is $3, 5, 7, 8$. The median of this set is the average of the 2nd and 3rd numbers: $(5 + 7) \div 2 = 6$.

5. Calculate Q3:

The second half of the data is $13, 14, 18, 21$. The median of this set is the average of the 2nd and 3rd numbers: $(14 + 18) \div 2 = 16$.

Result: The 5-number summary is:

- Minimum: **3**
- Q1: **6**
- Median: **12**
- Q3: **16**
- Maximum: **21**

Outliers: 1.5 * IQR above upper quartile or below lower quartile. $IQR = 16 - 6 = 10$

Bounds:

$$6 - 1.5 \times 10 = -9$$

$$16 + 1.5 \times 10 = 31$$

Correlation and Causality

Correlation measures the strength and direction of a linear relationship between two variables. It gives us an idea of how one variable changes when another variable changes.

Imagine two people, Alice and Bob, dancing together on a dance floor.

- **Strength**: How closely the movements of one variable match the movements of another variable. Think of the synchronization between Alice and Bob’s dance steps. If they are perfectly in sync, the strength is strong. If they move somewhat in sync but not always, the strength is moderate. If they move without any connection to each other, the strength is weak or nonexistent.

- **Direction**: Whether the variables move in the same direction (positive correlation) or opposite directions (negative correlation). If Alice and Bob move forward together or backward together, their dance direction is positive. If one moves forward while the other moves backward, their dance direction is negative.

- **Perfect Positive Correlation (+1)**: This represents a perfect positive correlation of +1. When one variable increases, the other also increases by a consistent proportion. Every time Alice takes a step forward, Bob also takes a step forward. When Alice steps back, Bob does the same. They move in perfect sync, like mirror images.

- **No Correlation (0)**: This represents no correlation or a correlation of 0. The movements (or values) of one variable do not predict or relate to the movements of the other variable. Alice is dancing energetically, while Bob is just standing still or moving randomly without any connection to Alice's movements.

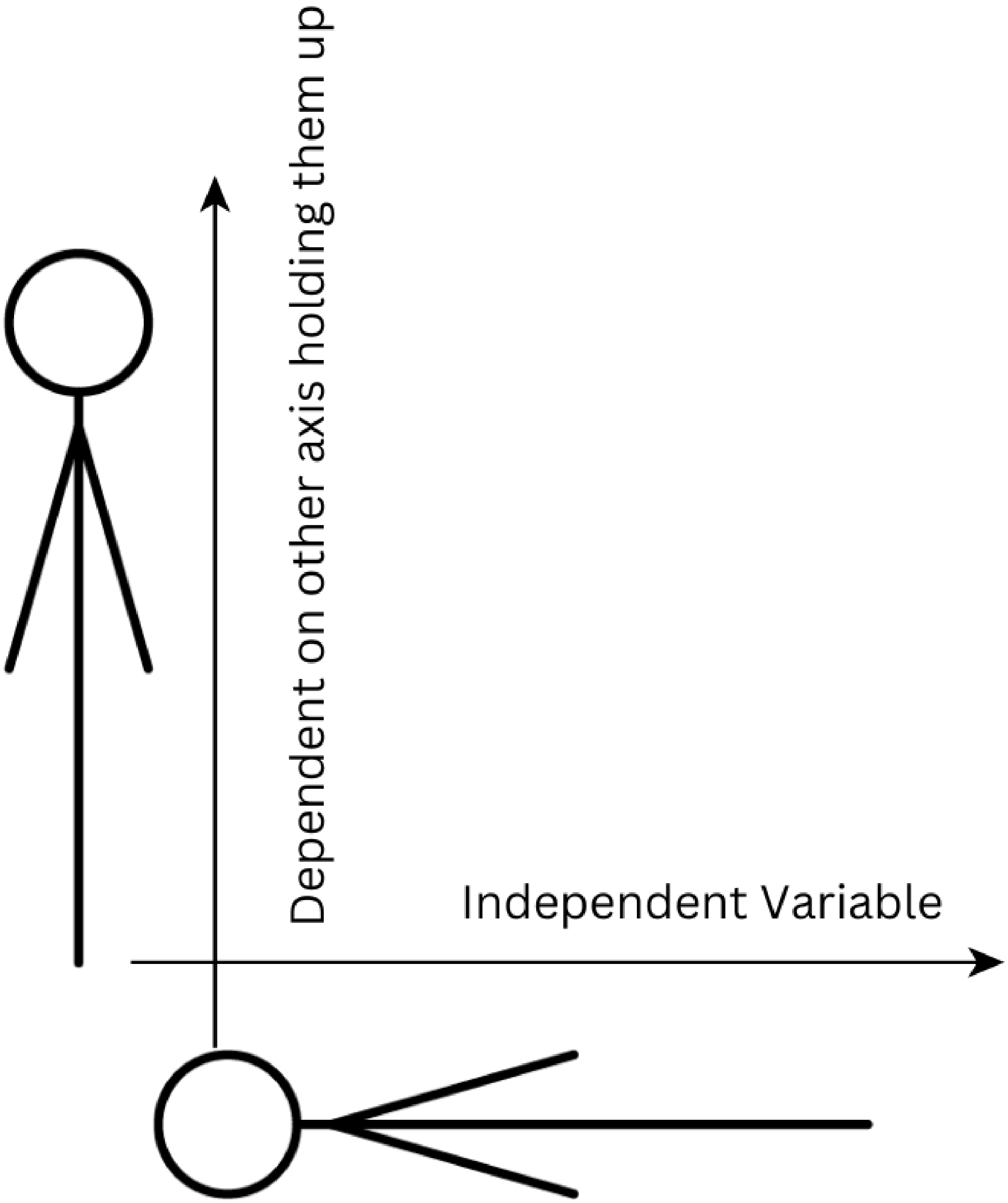
- **Perfect Negative Correlation (-1)**: This represents a perfect negative correlation of -1. When one variable increases, the other decreases by a consistent proportion. Every time Alice takes a step forward, Bob takes a step back. When Alice steps back, Bob steps forward. They move in opposite directions but are still in sync.

Causality implies a direct cause-and-effect relationship between two variables. In this case, Alice's action (clapping) directly causes Bob's reaction (stepping forward). Imagine a scenario where every time Alice claps her hands, Bob takes a step forward. It's consistent and predictable. Here, Alice's clapping is causing Bob's movement.

Remember: ***Correlation does not equal causation.*** Just because two variables move in sync (like our dancing partners) doesn't mean one causes the other. They might be moving together due to some other factor or pure coincidence.

- **Dependent Variable**: Y, the variable that is being measured/observed. It is the effect.

- **Independent Variable**: X, the variable that is being manipulated or categorized to observe its effect on another variable.



Sampling Methods

Random: Every individual in the population has an equal chance of being selected. Imagine you have a bowl of mixed fruit (apples, oranges, bananas, and grapes). You close your eyes and pick a fruit. Each time you pick, every fruit has an equal chance of being chosen.

Stratified Random: The population is divided into subgroups (or “strata”) based on a specific characteristic, and samples are randomly selected from each subgroup. Now, instead of picking any fruit from the bowl, you decide you want 2 of each type. So, you separate the fruits by type and pick 2 apples, 2 oranges, 2 bananas, and 2 grapes.

Cluster: The population is divided into clusters, and a random sample of clusters is chosen. All individuals within the selected clusters are then sampled. Imagine you have *multiple* bowls of mixed fruit, each from different rooms in a house. Instead of sampling fruit from every bowl, you randomly select a few bowls and sample all the fruits within those chosen bowls.

Systematic: You select a starting point and then sample every "kth" individual from the population. Back to our mixed fruit bowl. Instead of picking randomly, you decide to pick every 3rd fruit until you have enough samples.

Convenience: Individuals are chosen based on what's easiest or most convenient for the sampler, rather than any systematic or random method. You decide to sample the fruits that are closest to you in the bowl, simply because they're easy to reach.

Biases:

Sampling: Certain members of a population are more likely to be included in a sample than others, leading to a non-representative sample. For example, if you are trying to get an average height of a basketball team by measuring only the tallest players. This would give an inaccurate representation of the team's average height.

Non-response: When individuals chosen for a sample do not respond, and their non-responses are related to the property being measured. For example, you send out a survey to 100 people about their favorite ice cream flavor. Only 30 people respond, and all of them love chocolate. You might wrongly conclude that chocolate is the favorite flavor of the entire group.

Omitted Variable: Happens when a model is created without an important determinant, leading to incorrect conclusions. For example, you are trying to determine why a plant is dying. You consider sunlight and water but forget about soil quality. Your conclusions might be skewed because you omitted a crucial variable.

Voluntary: When participants self-select to be in a sample, usually those with strong opinions or feelings. A TV show asks viewers to call in and vote if they like or dislike the show. Only the most passionate viewers, either extremely positive or negative, take the time to call.

Social Desirability: Happens when respondents answer questions in a way they believe will be viewed favorably by others. In a classroom survey about homework habits, students might over-report the amount of time they study each night because they want to appear diligent.

Framing: Occurs when the way information is presented or framed alters the perception of that information, influencing decisions or judgments. Imagine how you would respond if asked, “Do you support the freedom of peaceful assembly?” versus “Do you support allowing large groups to block city streets?”

Probability

Probability is a measure between 0 and 1 that indicates the likelihood of an event occurring. A probability of 0 means the event won’t happen, while 1 means it’s certain.

Applications in Political Science

1. **Election Forecasting**: Using probability to predict election outcomes based on polling data, historical trends, and other factors.

2. **Policy Analysis**: Evaluating the potential impact of policies, such as the likelihood that a new policy will reduce unemployment rates.

3. **International Relations**: Estimating the probability of international events like wars, treaties, or diplomatic negotiations.

4. **Public Opinion**: Gauging the likelihood of shifts in public opinion in response to various stimuli.

5. **Voter Behavior**: Analyzing the probability of voters supporting specific candidates or policies based on demographics.

6. **Legislative Success**: Estimating the chances of a bill passing in a legislative body.

7. **Conflict Resolution**: Assessing the likelihood of successful peace negotiations.

Basic Probability Rules

1. $\text{P}(\text{event occurring}) \geq 0$ and ≤ 1.
2. $\text{P}(\text{event occurring}) = 1 - \text{P}(\text{event not occurring})$.
3. The sum of probabilities for all possible events is 1.

Core Concepts

Conditional Probability: The probability of an event (A) occurring given that another event (B) has already occurred. Represented as $\text{P}(\text{A}|\text{B})$.

$$\text{P}(\text{A}|\text{B}) = \frac{\text{P}(\text{A} \cap \text{B})}{\text{P}(\text{B})}$$

Where:

- $\text{P}(\text{A}|\text{B})$ is the probability of event A occurring given that event B has occurred.
- $\text{P}(\text{A} \cap \text{B})$ is the probability of both events A *and* B occurring.
- $\text{P}(\text{B})$ is the probability of event B occurring.

Example: Suppose you have a deck of 52 playing cards. Let's find the probability of drawing an Ace given that the card drawn is red.

Let:

- A be the event that the card drawn is an Ace.
- B be the event that the card drawn is red.

From a standard deck:

- There are 2 red Aces (Ace of Hearts and Ace of Diamonds).
- There are 26 red cards in total (Hearts and Diamonds).

Using the formula:

$$\text{P}(\text{A}|\text{B}) = \frac{\text{P}(\text{A} \cap \text{B})}{\text{P}(\text{B})}$$

$$\text{P}(\text{A}|\text{B}) = \frac{2/52}{26/52} = \frac{2}{26} = \frac{1}{13}$$

So, the probability of drawing an Ace given that the card is red is $\frac{1}{13}$.

Independence: Two events are independent if the occurrence of one does not affect the occurrence of the other. Represented as $P(A \text{ and } B) = P(A) \times P(B)$. For example, tossing a coin and getting heads will not influence the likelihood of you then rolling a fair six-sided die and getting a 5.

Mutually Exclusive: Events that cannot occur simultaneously. For example, in a single coin toss if it shows heads (Event A) then it cannot show tails (Event B) at the same time.

Central Limit Theorem: Given a population with **any** distribution and taking random independent samples of size ‘n’ from that distribution, the sample means of those independent samples will be approximately normally distributed. Going back to our basket of fruits, if you make enough smoothies (taking random samples of the fruit) and plot the taste (or mean) of each smoothie on a graph, the distribution of the tastes (means) of all the smoothies will start to look like a bell curve (a normal distribution), regardless of the original distribution of fruits in the basket.

Bayes’ Theorem: Describes the probability of an event based on prior knowledge. Given by the formula:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Hypothesis Testing

This is a statistical method use to make inferences or draw conclusions about a population based on a sample.

Imagine two politicians debating a policy. One claims it will benefit the majority (null hypothesis), while the other believes it won’t (alternative hypothesis). The debate’s outcome, based on evidence and arguments, determines which claim is more likely.

p-value: The probabiliy of observing the sample data if the null hypothesis is true. Think of it as the audience’s reaction. If the p-value is low, it means the audience found the alternative hypothesis more convincing, given the evidence. If the p-value is high, it indicates that the audience found insufficient evidence to challenge the null hypothesis, suggesting they believe the policy might indeed benefit the majority.

Confidence Intervals

A confidence interval is a range of values that is used to estimate an unknown population parameter. It provides an estimate about the precision of the sample statistic.

Imagine you conduct a poll to estimate the percentage of voters who support a particular candidate. After surveying a random sample of 1,000 eligible voters, you find that 55% of respondents support the candidate. You also calculate a 95% confidence interval for this proportion, which is (52%, 58%).

This means that if you were to conduct many similar polls with different random samples of voters and calculate the confidence interval for each, 95% of those intervals would include the actual proportion of all voters in the city who support the candidate.

Formula: $\bar{x} - 1.96 * SE, \bar{x} + 1.96 * SE$

Ways to Determine Significance:

1. Is p ≤ .05?
2. Is t greater than 1.96 in a two-tailed test?
3. Do the confidence interval bounds have the same sign?

Level of Confidence	Alpha	Since…	T-Stat
90% (.90)	.10	.90 = 1-.10	1.64
95% (.95)	.05	.95 = 1-.05	1.96
99% (.99)	.01	.99 = 1-.01	2.58

Key Points on Confidence Intervals:

- Confidence Intervals quantify uncertainty.
- Conditions for Valid CIs:

  1. Random sample
  2. Normality
  3. Independence

Regression Analysis

Linear Regression: You might want to know if spending more on a campaign leads to more votes. Linear regression can help quantify this relationship.

Logistic Regression: You might want to predict whether a candidate will win or lose based on factors like campaign spending, demographics, and previous voting patterns. To do this, you would use logistic regression.

Variable Types:

- **Variable**: An empirical measure of a concept/characteristic that varies across observations.

Variable Type	Definition	Example
Continuous	Exact Number	Exact Weight
Discrete	An estimation/categorization of the true value	Weight Rounded
Categorical	Observations belong to a discrete set of categories	Race, Gender
Nominal	No Order	Name, PhD Program
Ordinal	Ordered	Socioeconomic Status
Interval Level	Change from one category to the next is identical across all values	Temperature

Data Visualization

Bar Charts: Used to compare quantities of different categories. For example: comparing data across categories, showing frequency or proportions, comparing parts of a whole. Imagine a chart showing how many seats each party won in different states. Each state is a category and the number of seats is represented by the height of the bar.

Line Graphs: Display data points over a continuous interval or time span. Used when observing trends over time, comparing changes over the same period for more than one group, analyzing patterns and relationships. For example, a line graph could show how voter turnout has changed over several election years, helping to spot trends or significant changes.

Scatter Plots: Displays values for two variables for a set of data. Used when investigating the relationship between two variables, observing and showing correlations, identifying outliers or anomalies in the data. For example, you could plot each city’s average income against the percentage of voters supporting a particular party. This scatter plot might reveal if wealthier cities tend to support a specific party.

Common Statistical and OLS Regression Symbols:

Symbol	Description
n	sample size
N	population size
μ	population mean
$\bar{x}$	sample mean
δ^2	population variance
δ	population standard deviation
$\hat{\delta}$	sample standard deviation
H_0	null hypothesis
H_1 or H_a	alternative hypothesis
α	P(Type I Error) or regression intercept
β	P(Type II Error) or regression slope coefficient
ϵ	error or residual
δ_ϵ	standard error of regression

Symbol	Description
df	degrees of freedom
r	correlation coefficient
R^2	coefficient of determination
F	F-statistic
t	t-statistic
p	probability value
χ^2	chi-square statistic
λ	eigenvalue
ϱ	population correlation coefficient
$\hat{p}$	sample proportion
π	population proportion
s^2	sample variance
σ^2	population variance
σ	population standard deviation
z	z-score
b_0	y-intercept in regression
b_1	slope in regression