

STATISTIC GUIDE



Created by:
AMANDA RUSSELL

Contents

What is Statistics?	5
Population vs. Sample	5
Parameter vs. Statistic	5
Types of Data	7
Qualitative Data	7
Quantitative Data	7
Levels of Measurement	7
Types of Sampling:	10
Blinding, Placebos, and Double-Blinding in Experiments	10
What is a Frequency Distribution?	14
Key Terms to Know:	14
Graphs of Frequency Distributions:	17
Different Types of Graphs	18
Measures of Central Tendency	20
Mean (Average)	20
Weighted Mean	21
Mean of Frequency Distribution	22
Skewness in Distributions	23
Measures of Variation	26
Empirical Rule (68-95-99.7 Rule)	27
Chebychev's Theorem	28
Coefficient of Variation (CV)	29
Measures of Position	30
Probability and Counting	34
The Fundamental Counting Principle	34
Types of Probability	34
Complementary Events	35
Conditional Probability	37
Independent and Dependent Events	37
Multiplication Rule	38
The Addition Rule in Probability	39
Additional Topics in Probability and Counting	41

Permutations	41
Combinations.....	41
Probability Distribution	44
Random Variables	44
Discrete Probability Distributions	44
Mean, Variance, and Standard Deviation of Discrete Random Variables	44
Binomial Distributions.....	46
Binomial Probability Formula:	46
Formulas for Mean, Variance, and Standard Deviation for Binomial:.....	47
More Discrete Probability Distributions: Geometric and Poisson Distributions.....	48
Geometric Distribution	48
Poisson Distribution	48
Normal Distribution:	51
Standard Normal Distribution:.....	51
Z-Scores:.....	52
Finding Areas Under the Curve:.....	52
What is a Normal Distribution?	54
Formula for Z-Score:	54
Normal Distributions: Finding Values	57
Steps to Find Values Using Z-Scores:.....	57
Sampling Distributions and the Central Limit Theorem.....	59
Confidence Intervals for the Mean: Population SD Known	62
Point Estimate:	62
Margin of Error (E):	62
Confidence Interval:.....	62
Level of Confidence (c):.....	62
Confidence Intervals for the Mean: Population SD Unknown.....	64
Point Estimate:	64
Margin of Error (E):	64
Confidence Interval Formula:.....	64
Confidence Intervals for Population Proportions.....	67
Steps to Calculate Confidence Intervals for Proportions:	67
What is Hypothesis Testing?	70

Types of Errors:.....	71
Significance Level (α):.....	71
Types of Tests:.....	72
Hypothesis Testing for the Mean: Population SD Known	73
Test Statistic (Z-test):	73
P-value Approach:	73
Critical Value Approach:	74
Hypothesis Testing for the Mean: Population SD Unknown	75
Hypothesis Testing for Proportions:	77
Key Terms:.....	77
Steps to Perform a Hypothesis Test for Proportions:	77
Correlation.....	80
How to Calculate the Correlation Coefficient (r):	81
Test Statistics for Correlation:	81
Critical Values:	83
Linear Regression	84
Measures of Regression and Prediction Intervals	86
Variation About a Regression Line:	86
Coefficient of Determination(r^2)	86
Standard Error of Estimate (se)	86
Prediction Intervals:	87
Z-Table (-)	96
Z-Table (+)	97

What is Statistics?

Imagine you have a big bag of jelly beans, and you want to know what flavors are inside. Statistics is like a method to help you figure that out without tasting every single jelly bean.

Definition: Statistics is the science of collecting, organizing, analyzing, and making sense of data (information).

Population vs. Sample

- **Population:** Imagine you want to know how tall everyone is in your entire country. The **population** is *every single person* in that country. But, measuring everyone's height would take forever, so instead, you might look at just a small group.
- **Sample:** A **sample** is that small group you actually measure. It's like a tiny slice of the whole cake (population). You use this sample to make guesses about the entire cake.

Parameter vs. Statistic

- **Parameter:** When we talk about a **parameter**, we're talking about a fact or number that describes the entire population. Since we often can't measure the whole population, parameters are usually unknown
**Tip: "P" for "population" and "parameter"—they go together. **
- **Statistic:** A **statistic** is like a parameter but for a sample. It's a fact or number that describes the sample, which we do have data for. We then use this statistic to make guesses about the population parameter.
** Tip: "S" for "sample" and "statistic"—they go together. **

Examples to Tie it All Together:

- Question: "The salary of each baseball player in a league—is this a population or sample?"

Answer: It's a population because it includes all players in the league (all the jelly beans in the bag).

- Question: "55% of students in a sample own a vehicle—is this a parameter or statistic?"

Answer: This is a statistic because it's based on a sample (not all students, just a few).

So, in summary:

- **Population** = everyone you're interested in.
- **Sample** = a small group from that population.
- **Parameter** = a fact about the whole population.
- **Statistic** = a fact about your sample, used to guess the parameter.

Types of Data

There are two main types of data: **Qualitative** and **Quantitative**.

Qualitative Data (L is for Labels):

- This type of data describes qualities, characteristics, or categories. It's about **what something is**.
- **Examples:** Colors of cars, types of animals, movie genres.
- **Tip:** If it involves names, labels, or descriptions (like colors or types), it's qualitative. Just remember, "**L is for Labels**"—it's about descriptions, not numbers!

Quantitative Data (N is for Numbers):

- This type of data deals with numbers and measurements. It's about **how much or how many**.
- **Examples:** Age, height, number of siblings, temperature.
- **Tip:** If you can count it or measure it (and it makes sense to do math with it), it's quantitative. Remember, "**N is for Numbers**"—it's all about numbers and measurements!

Levels of Measurement

Data can be more than just qualitative or quantitative; it can be sorted into levels based on how detailed and useful it is. Let's break it down in the simplest terms:

- **Nominal Level (Just Names or Labels):**
 - This is the simplest level. It's all about putting things into categories or giving them names. There's no order, just different groups.
 - Examples: Types of fruits (apple, banana), colors of cars (red, blue, green).
 - Tip: Think "Nominal" = "Name." It's all about labels with no special order or ranking.

- **Ordinal Level (Order Matters, but Gaps Don't):**
 - At this level, you can put things in order, like from best to worst, but you don't know how big the steps are between them. It's like lining up people by height without measuring how much taller one is than the other.
 - Examples: Movie ratings (G, PG, R), rankings (1st, 2nd, 3rd).
 - Tip: "Ordinal" = "Order." Think of a line-up where the sequence matters, but the exact difference doesn't.

- **Interval Level (Order and Gaps Matter, but No True Zero):**
 - Here, not only do things have an order, but the differences between them are meaningful. However, zero doesn't mean there's nothing. It's just a point on the scale.
 - Examples: Temperature in Celsius (0°C is a temperature, not 'no heat'), dates on a timeline.
 - Tip: "Interval" means you're looking at the spaces between numbers that matter, but zero isn't the starting point.

- **Ratio Level (Order, Gaps, and True Zero):**
 - This is the most detailed. Everything about the order and gaps matters, and zero means zero—nothing. You can say one value is twice as much as another.
 - Examples: Weight, height, time (0 seconds means no time).
 - Tip: "Ratio" means real zero. You can multiply, divide, and zero truly means none.

Simplified View:

- **Nominal:** Just names, no order. (Think "Name.")
- **Ordinal:** Order, but no exact gaps. (Think "Order.")
- **Interval:** Order, exact gaps, but no real zero. (Think "In-between," but zero isn't nothing.)
- **Ratio:** Full details: order, exact gaps, and zero means none. (Think "Real zero.")

By thinking about how much detail you have and what zero means, you can easily figure out which level your data fits into!

Types of Sampling:

Sampling is selecting a smaller group from a larger population to study. It's like tasting one piece of cake to decide if the whole cake is good.

1. **Random Sample:** Everyone has an equal chance of being chosen. It's like drawing names from a hat.
2. **Stratified Sample:** Divide the population into groups (strata) and pick some from each group. It's like picking a few red, blue, and green candies from a jar.
3. **Cluster Sample:** Divide into groups, then pick all from one or more groups. Imagine picking everyone in one classroom instead of individuals from several.
4. **Systematic Sample:** Pick every nth person, like every 10th name on a list.
5. **Convenience Sample:** Pick what's easiest, like surveying your friends. Be careful—this can be biased!

Sampling Tips:

- **Random = Fair chance for all.**
- **Stratified = A bit from every group.**
- **Cluster = All from some groups.**
- **Systematic = Every nth item.**
- **Convenience = Easy, but maybe biased.**

Blinding, Placebos, and Double-Blinding in Experiments

When conducting experiments, especially those involving people, researchers use techniques like blinding and placebos to ensure the results are accurate and not influenced by people's expectations.

Placebo:

- A placebo is a “fake” treatment that looks like the real thing but doesn’t have any active ingredients. It’s used in experiments to see if the actual treatment has a real effect beyond people’s expectations.
- **Example:** In a study testing a new painkiller, some people might get the real drug, while others get a sugar pill (placebo). This helps determine if the real drug actually works or if people feel better just because they think they’re being treated.

Blinding:

- **Single-Blind:** In a single-blind experiment, the participants do not know whether they are receiving the real treatment or the placebo. This helps prevent their expectations from affecting the results.
 - **Example:** Patients don’t know if they’re getting the real medicine or a placebo, so their reactions are unbiased.

Double-Blinding:

- In a double-blind experiment, **neither the participants nor the researchers** know who is receiving the real treatment or the placebo until after the experiment is finished. This is done to prevent both participants’ and researchers’ biases from affecting the results.
 - **Example:** In a study testing a new drug, both the patients and the doctors don’t know who gets the actual drug or the placebo. This way, doctors’ observations and patients’ responses are more objective.

Why Use These Techniques?

- **Placebos** help determine if the treatment effect is real or just psychological (thinking something will work).
- **Blinding** ensures participants' expectations don't influence their behavior or responses.
- **Double-Blinding** ensures that neither the participants' nor the researchers' biases affect the study's results, making the findings more reliable.

Simple Tips to Remember:

- **Placebo = Fake treatment.** Helps check if effects are due to the real treatment or just thinking it's real.
- **Single-Blind = Participants don't know.** They don't know if they're getting the real or fake treatment.
- **Double-Blind = Participants and researchers don't know.** Keeps everyone unbiased to get the most accurate results.

Descriptive Statistics

What is a Frequency Distribution?

A frequency distribution is a way to organize data so you can see how often different values (or ranges of values) occur. Imagine you have a lot of data points (like test scores), and you want to see which scores appear the most. A frequency distribution helps you group these scores into ranges (called classes) and shows how many times each range appears.

Key Terms to Know:

1. **Classes:** These are the groups or ranges of data. Think of them as bins that you sort your data into.
 - **Example:** If you're looking at ages, your classes might be 10-19, 20-29, etc.
2. **Frequency:** This is the count of how many data points fall into each class.
 - **Example:** If 10 people are aged 20-29, the frequency for that class is 10.
3. **Class Width:** This is the difference between the lowest number of one class and the lowest number of the next class.
 - **Tip:** Think "width" like the space between classes; how wide each class is.
4. **Midpoint:** This is the center value of each class.
 - **Formula:**
$$\text{Midpoint} = \frac{\text{Lower Class Limit} + \text{Upper Class Limit}}{2}$$
 - **Tip:** Think "middle" point.

5. Relative Frequency: This tells you the proportion (or percentage) of data points that fall into each class compared to the total number of data points.

- **Formula:** $\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Frequency}}$

6. Cumulative Frequency: This is the sum of the frequencies for a class and all previous classes.

- **Tip:** "Cumulative" means adding up as you go!

Example: Organizing Test Scores

Suppose you have the following test scores for a class:

58, 61, 62, 65, 65, 66, 68, 70, 71, 72, 73, 75, 78, 80, 82, 84, 85, 88, 90, 91 Let's create a frequency distribution for this data:

Step 1: Define Classes

- Let's make classes with a width of 10 points, starting from 50:
 - Class 1: 50-59
 - Class 2: 60-69
 - Class 3: 70-79
 - Class 4: 80-89
 - Class 5: 90-99

Step 2: Find Frequency

- Count how many scores fall into each class:
 - 50-59: **1** (58)
 - 60-69: **6** (61, 62, 65, 65, 66, 68)
 - 70-79: **6** (70, 71, 72, 73, 75, 78)
 - 80-89: **5** (80, 82, 84, 85, 88)
 - 90-99: **2** (90, 91)

Step 3: Class Width

- Class width is the difference between the lower/higher limits of consecutive classes:
 - $60 - 50 = 10$, $70 - 60 = 10$
 - So, the class width is **10**.

Step 4: Calculate Midpoints

- Use the formula: $\text{Midpoint} = \frac{\text{Lower Class Limit} + \text{Upper Class Limit}}{2}$
 - For 50-59: $\frac{50+59}{2} = 54.5$
 - For 60-69: $\frac{60+69}{2} = 64.5$
 - For 70-79: $\frac{70+79}{2} = 74.5$
 - For 80-89: $\frac{80+89}{2} = 84.5$
 - For 90-99: $\frac{90+99}{2} = 94.5$

Step 5: Calculate Relative Frequency

- Divide each class frequency by the total number of data points (20 scores):
 - For 50 – 59: $\frac{1}{20} = 0.05$ (or 5%)
 - For 60 – 69: $\frac{6}{20} = 0.3$ (or 30%)
 - For 70 – 79: $\frac{6}{20} = 0.3$ (or 30%)
 - For 80 – 89: $\frac{5}{20} = \frac{1}{4} = 0.25$ (or 25%)
 - For 90 – 99: $\frac{2}{20} = \frac{1}{10} = 0.1$ (or 10%)

Step 6: Calculate Cumulative Frequency

- Add up frequencies as you go along:
 - 50 – 59: 1 (just the first class)
 - 60 – 69: $1 + 6 = 7$
 - 70 – 79: $7 + 6 = 13$
 - 80 – 89: $13 + 5 = 18$
 - 90 – 99: $18 + 2 = 20$

Final Frequency Distribution Table

Class	Freq.	Class Width	Midpoint	Relative Freq.	Cumulative Freq
50-59	1	10	54.5	0.05 (5%)	1
60-69	6	10	64.5	0.3 (30%)	7
70-79	6	10	74.5	0.3 (30%)	13
80-89	5	10	84.5	0.25 (25%)	18
90-99	2	10	94.5	0.1 (10%)	20

Graphs of Frequency Distributions:

1. Frequency Histogram:

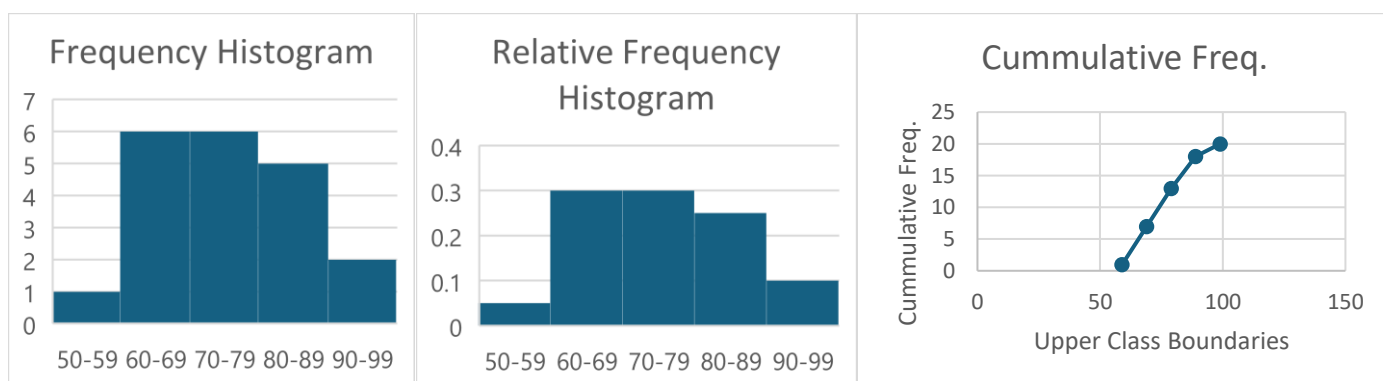
- A graph that uses bars to show the frequency of each class. The bars touch each other because there are no gaps between the classes.
- Tip: Histograms are like bar graphs, but the bars touch because the data is continuous.

2. Relative Frequency Histogram:

- Similar to a frequency histogram, but it shows relative frequencies instead of just counts.
- Tip: Same shape as a frequency histogram, but the scale on the side is in percentages.

3. Cumulative Frequency Graph (Ogive):

- This is a line graph that shows cumulative frequencies. It's like seeing a running total of frequencies.



Different Types of Graphs

1		5
2		5 5
3		5

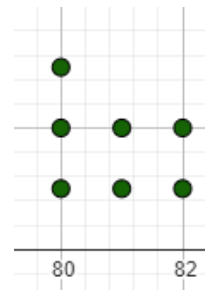
Key: 3|1 means 31

1. Stem-and-Leaf Plot:

- **What It Is:** This is a way to show numbers while keeping the original data. Each number is split into a "stem" (the first digit or digits) and a "leaf" (the last digit).
- **Example:** For the number 25, "2" is the stem and "5" is the leaf.
- **Why Use It:** It's great for seeing the shape of data and finding the smallest and largest values quickly.

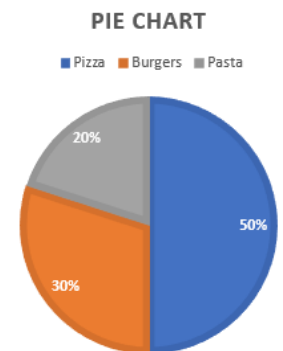
2. Dot Plot:

- **What It Is:** A graph where each data point is represented by a dot above a number line.
- **Example:** If three people scored 80 on a test, you place three dots above 80 on the line.
- **Why Use It:** Easy to see frequency and identify patterns in small data sets.



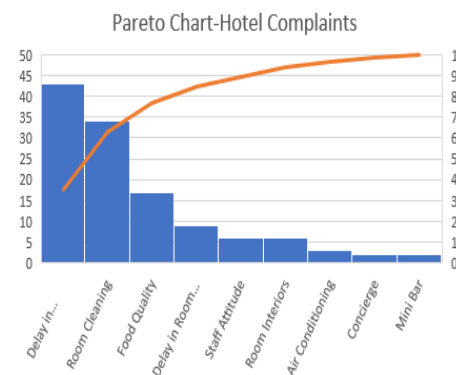
3. Pie Chart:

- **What It Is:** A circle divided into slices, where each slice represents a part of the whole.
- **Example:** If 50% of students like pizza, half of the pie will be for pizza.
- **Why Use It:** Great for showing percentages and proportions of categories.



4. Pareto Chart:

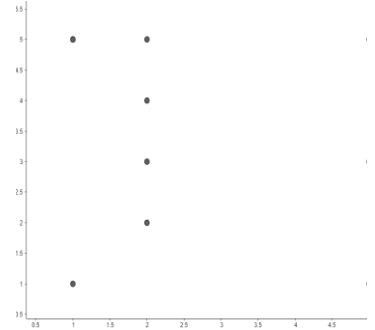
- **What It Is:** A special bar graph where bars are arranged from tallest to shortest. It often includes a line graph showing cumulative totals.



- **Example:** Used to show the most common causes of something, like customer complaints.
- **Why Use It:** Helps identify the biggest factors at a glance.

5. Scatter Plot:

- **What It Is:** A graph that shows points plotted on an x-y grid to show the relationship between two variable
- **Example:** Comparing height and weight—each point represents one person's height and weight.
- **Why Use It:** Useful for spotting trends, patterns, or correlations between variables.



6. Misleading Graphs:

- **What They Are:** Graphs that misrepresent data, often by using incorrect scales, uneven bar sizes, or other tricks.
- **Example:** A pie chart that makes one slice look bigger than it should.
- **Why Avoid Them:** They can give a false impression of the data, leading to wrong conclusions.

Tips to Remember:

- **Stem-and-Leaf Plot = Keep the Original Numbers:** Great for seeing details while keeping all data visible.
- **Dot Plot = Dots for Each Data Point:** Simple and shows frequency clearly.
- **Pie Chart = Slices of a Whole:** Use for showing parts of a whole, like market shares.
- **Pareto Chart = Bars in Order:** Shows which items are most significant; think "tallest to shortest."
- **Scatter Plot = Dots in a Grid:** Use to see if two things are related, like study time vs. grades.
- **Watch for Misleading Graphs:** Always check if the graph accurately represents the data!

Measures of Central Tendency

These measures help us understand the "center" or most typical values in a set of numbers.

Mean (Average)

- **What it is:** Add up all the numbers, then divide by how many numbers there are.
- **Formula:**
 - For a population $\mu = \frac{\sum x}{N}$ where μ (mu) is the mean, $\sum x$ is the sum of all values, and N is the total number of values.
 - For a sample, $\bar{x} = \frac{\sum x}{n}$ where \bar{x} (x-bar) is the mean, $\sum x$ is the sum of all sample values, and n is the total number of values.
- **Tip to Remember:** Think of "mean" as "mean to add and divide."

Median

- **What it is:** The middle number in an ordered set. If the set has an even number of values, the median is the average of the two middle numbers.
- **How to find it:**
 - Arrange the numbers in order.
 - If there's an odd number of values, pick the middle one.
 - If there's an even number of values, average the two middle ones.
- **Tip to Remember:** "Median" sounds like "middle."

Mode

- **What it is:** The number that appears most often in a set. A set can have no mode, one mode, or more than one mode.
- **Tip to Remember:** "Mode" sounds like "most."

Outliers

- **What it is:** An outlier is a value that is far away from the rest of the data. It can affect the mean significantly.
- **Tip:** Outliers are like the "odd one out."

Weighted Mean

- **What it is:** A mean where some numbers count more than others.
- **Formula:** $\bar{x} = \frac{\sum xw}{\sum w} = \frac{\text{Sum of the products of the entries and the weights}}{\text{Sum of the weights}}$
where x is each value, and w is its weights (how much it counts)
- **Tip to Remember:** If you see different weights, think "weighted mean."

Example: Imagine a student's grades in different categories with different weights:

- **Homework:** 85 (20% of the grade)
- **Quizzes:** 90 (10% of the grade)
- **Midterm:** 78 (30% of the grade)
- **Final Exam:** 92 (40% of the grade)

Step-by-step Calculation:

- Multiply each score by its weight (as a decimal):
 - Homework: $85 \times 0.20 = 17$
 - Quizzes: $90 \times 0.10 = 9$
 - Midterm: $78 \times 0.30 = 23.4$
 - Final Exam: $92 \times 0.40 = 36.8$
- Add the products: $17 + 9 + 23.4 + 36.8 = 86.2$
- The weights add up to $0.20 + 0.10 + 0.30 + 0.40 = 1$

So, the **weighted mean** is **86.2**.

Mean of Frequency Distribution

- **What it is:** An estimate of the mean when data is grouped into intervals.
- **Formula:** $\bar{x} = \frac{\sum xf}{n}$, where x is the midpoint of each interval, f is the frequency (how often it appears), and n is the total number of values
- **Tip:** Use midpoints of intervals for grouped data.

Example: The following table shows the ages of people in a group and how many people are in each age group. Approximate the mean of the frequency distribution

Age Group	Frequency
20-24	5
25-29	8
30-34	12
35-39	10
40-44	5

Steps to Find the Mean:

- **Find the Midpoint of Each Class:**
 - For each age group, calculate the midpoint by averaging the lower and upper bounds:
 - Midpoint for 20-24: $\frac{20+24}{2} = 22$
 - Midpoint for 25-29: $\frac{25+29}{2} = 27$
 - Midpoint for 30-34: $\frac{30+34}{2} = 32$
 - Midpoint for 35-39: $\frac{35+39}{2} = 37$
 - Midpoint for 40-44: $\frac{40+44}{2} = 42$

- **Multiply Each Midpoint by Its Frequency:**

Age Group	Midpoint x	Frequency f	xf
20-24	22	5	22(5) =110
25-29	27	8	27(8) =216
30-34	32	12	32(12) =384
35-39	37	10	37(10) =370
40-44	42	5	42(5) = 210
		$n = \sum f = 40$	$\sum xf = 1290$

- **Calculate the Mean:**
- Mean of Frequency Distribution: $\bar{x} = \frac{\sum xf}{n}$
- $\bar{x} = \frac{1290}{40} \approx 32.25$

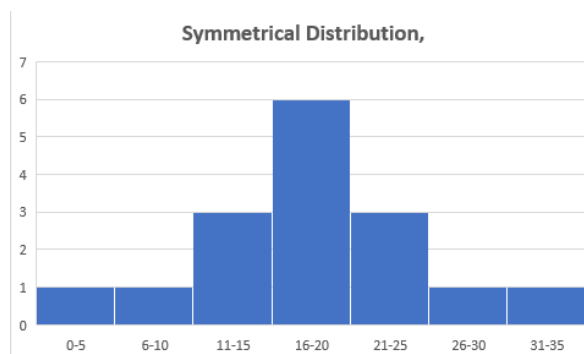
The mean age of the group is 32.25 years.

Skewness in Distributions

Skewness shows if the data is leaning more to one side. Imagine a normal distribution as a perfect bell shape. If it's not bell-shaped, it's skewed.

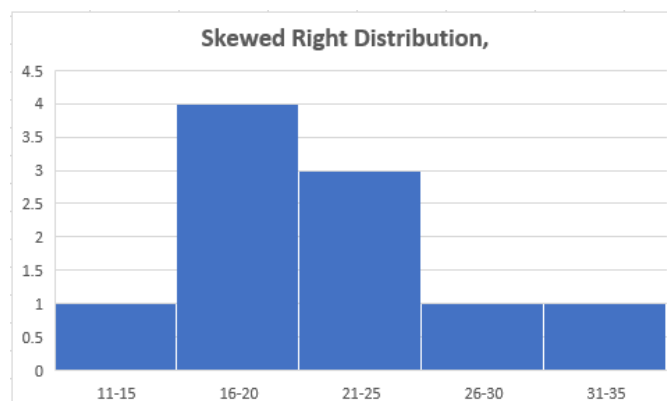
1. Symmetric Distribution (Bell Shaped)

- **What it is:** Data is evenly spread on both sides of the center.
- **Key Point:** Mean, median, and mode are all in the middle.
- **Tip:** Think of a perfect hill — balanced on both sides.



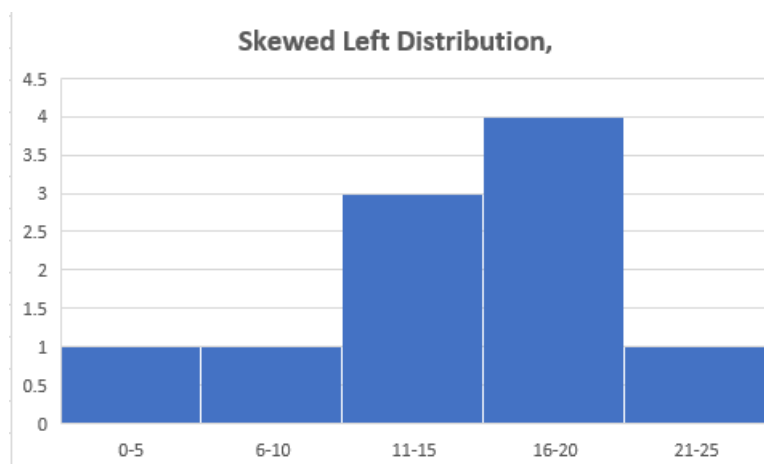
2. Skewed Right (Positive Skew)

- **What it is:** Most data is on the left side, and the tail stretches to the right.
- **Example:** Incomes, where most people earn low to moderate, but a few earn very high amounts, dragging the tail right.
- **Key Point:** Mean > Median.
- **Tip:** Imagine a hill with a gentle slope on the right — the weight is on the left.



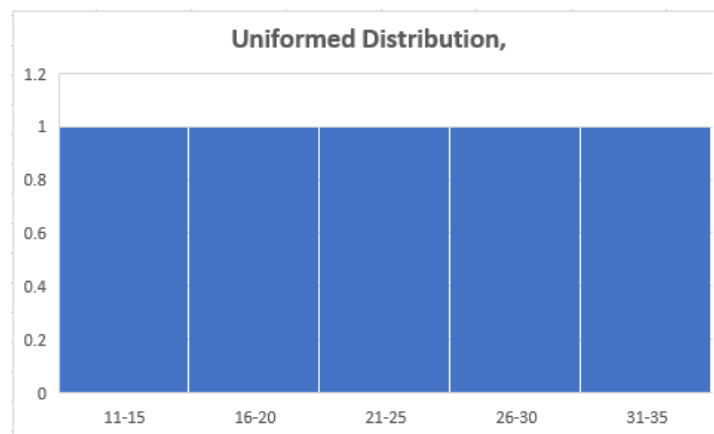
3. Skewed Left (Negative Skew)

- **What it is:** Most data is on the right side, and the tail stretches to the left.
- **Example:** Test scores where most students did well, but a few did very poorly, pulling the tail left.
- **Key Point:** Mean < Median.
- **Tip:** Imagine the opposite of right skew — gentle slope on the left, weight on the right.



4. Uniform Distribution

- **What it is:** Every value has an equal frequency. The graph looks flat because there's no peak.
- **Example:** Rolling a fair die (1-6 all have an equal chance).
- **Key Point:** All outcomes are equally likely, making the distribution look like a rectangle.
- **Tip:** Think of "uniform" as "one shape"—everything is even and balanced, like a flat line.



Visualizing Skewness:

- **Right Skew (Positive):** A steep left side and a gentle right slope.
- **Left Skew (Negative):** A gentle left slope and a steep right side.
- **Uniform (Flat and Even):** Picture a flat table or a ruler—everything is level and even.

Measures of Variation help us understand how spread out or scattered the data in a set is. Here are the key concepts:

Range

- **What it is:** The difference between the largest and smallest values in a data set.
- **Formula:** Range = Maximum Value – Minimum Value
- **Tip:** It's the simplest measure of spread but doesn't use all data points.

Variance

- **What it is:** Measures how far each data point is from the mean (average). Variance tells us how spread out the data points are.
- **Formula for Population Variance (σ^2):** $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$
 - μ : Population mean (average)
 - N: Total number of data points
 - x: Each individual data point
- **Formula for Sample variance (s^2):** $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
 - \bar{x} : Sample mean
 - n: Total number of sample data points

Standard Deviation

- **What it is:** The square root of the variance, which puts the measure back into the original units of the data (e.g., dollars, inches). It tells you how much the data typically deviates from the mean.
- **Formula for Population Standard Deviation (σ):** $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$
- **Formula for Sample Standard Deviation (s):** $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

- **Tip:** Standard deviation is a commonly used measure because it's easier to interpret. A larger standard deviation means the data is more spread out.

Standard Deviation for Grouped Data (Frequency Distributions)

- **What It Is:** When you have a large data set that's organized into groups (like frequency distributions), calculating the standard deviation helps you understand how spread out the data is around the mean.
- **Formula:** $s = \sqrt{\frac{(\bar{x} - \bar{x})^2 f}{n-1}}$ where $n = \Sigma f$
- **Tip:** Remember, the standard deviation formula for grouped data adjusts for the fact that we're using midpoints and frequencies instead of individual data points.

** How to do calculations in Excel and GeoGebra go to [page 89](#) **

Empirical Rule (68-95-99.7 Rule)

- **What it tells us:** For bell-shaped distributions (normal distributions):
 - **68% Rule:** About 68% of the data falls within **1 standard deviation** (σ) of the mean. This means most data points are close to the average.
 - Example: If the mean test score is 70 with a standard deviation of 5, the about 68% of the scores are between 65 and 75 ($70-5$ and $70+5$)
 $[\bar{x} - s \text{ \& } \bar{x} + s]$
 - **95% Rule:** About 95% of the data falls within **2 standard deviations** of the mean. This covers almost all the data.

- Using the same test score example, 95% of scores are between 60 and 80 ($70-2(5)$ and $70+2(5)$) $[\bar{x} - 2s \text{ \& } \bar{x} + 2s]$
- **99.7% Rule:** About 99.7% of the data falls within **3 standard deviations** of the mean. This covers nearly all possible data points.
 - Again, using the test score example, 99.7% of scores are between 55 and 85 ($70-3(5)$ and $70+3(5)$) $[\bar{x} - 3s \text{ \& } \bar{x} + 3s]$

Tip: Imagine a mountain with the peak at the average. The further you go from the peak, the less data there is, but most data points stay within those three-standard deviation “steps” from the center.

Chebychev's Theorem

What It Is:

- Chebychev’s Theorem is useful for **any shape of distribution**, not just bell-shaped.
- It guarantees a minimum amount of data within a certain number of standard deviations from the mean.

Key Points:

- **At Least 75% of Data Within 2 Standard Deviations:**
 - No matter the shape of your data, at least 75% of the data points will lie within 2 standard deviations from the mean.
 - This rule is a safety net when the data isn't perfectly bell-shaped.
- **At Least 88.9% of Data Within 3 Standard Deviations:**
 - At least 88.9% of the data points are within 3 standard deviations from the mean.

General Formula:

- The theorem states: At least $1 - \frac{1}{k^2}$ of data falls within k standard deviations from the mean, where $k > 1$
- - For example, if $k = 2$:
$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 0.75 \text{ or } 75\%$$
 - For $k = 3$:
$$1 - \frac{1}{3^2} = 1 - \frac{1}{9} = 0.889 \text{ or } 88.9\%$$

Tip: Think of Chebychev's Theorem as a broad safety measure — it works no matter how your data looks, even if it's not a nice bell curve. It's there to guarantee that a decent chunk of your data will always be within a certain range from the average.

Coefficient of Variation (CV)

- **What it is:** Describes the standard deviation as a percentage of the mean. It's useful for comparing variability between data sets with different units or means.
- **Formula:** $CV = \frac{\sigma}{\mu} * 100\%$ (for populations)
 $CV = \frac{s}{\bar{x}} * 100\%$ (for samples)

Key Symbols:

- | | |
|--|----------------------------------|
| • μ : Population mean | • s^2 : Sample variance |
| • \bar{x} : Sample mean | • f : Frequency |
| • σ : Population standard deviation | • N : Number of pop. entries |
| • s : Sample standard deviation | • n : Number of sample entries |
| • σ^2 : Population variance | |

Measures of Position

Measures of position tell us where a particular data value stands within a set of data. They help us understand the relative standing of a data point compared to others.

Key Concepts:

Quartiles:

- **What it is:** Quartiles divide data into four equal parts.
 - **Q1 (First Quartile):** 25% of data falls below this point.
 - **Q2 (Second Quartile or Median):** 50% of data falls below this point.
 - **Q3 (Third Quartile):** 75% of data falls below this point.
- **Tip:** Think of quartiles as road markers showing how far along you are in a journey through the data.

Interquartile Range (IQR):

- **What it is:** Measures the spread of the middle 50% of the data.
- **Formula:** $IQR = Q3 - Q1$
- **Tip:** Use IQR to find out how "spread out" the middle half of your data is. It's useful for spotting outliers.

Percentiles:

- **What it is:** Divides data into 100 equal parts. For example, the 90th percentile means 90% of data is below that value.
- **Formula:** Percentiles of $x = \frac{\text{Number of data entries less than } x}{\text{Total number of data entries}} * 100$
- **Tip:** Percentiles are like checkpoints in a race. The higher your percentile, the farther you are in the race compared to others.

****** How to do Quartiles and IQR in Excel and GeoGebra go to [page 91](#) - 93******

Z-Scores

- **What it is:** A z-score, also known as a standard score, tells you how far away a data point is from the mean of a data set in terms of standard deviations.
- **Formula:** $z = \frac{x-\mu}{\sigma} \rightarrow z = \frac{\text{Value}-\text{Mean}}{\text{Standard deviation}}$

How to Read Z-Scores:

- **Positive Z-Score:** The data point is above the mean.
 - Example: A z-score of +2 means the value is 2 standard deviations above the mean.
- **Negative Z-Score:** The data point is below the mean.
 - Example: A z-score of -1 means the value is 1 standard deviation below the mean.
- **Z-Score of 0:** The data point is exactly at the mean.

Interpreting Z-Scores:

- **Between -1 and 1:** The data point is close to the mean and considered typical.
- **Beyond ± 2 :** The data point is further from the mean and might be considered unusual.
- **Beyond ± 3 :** Very unusual, as it falls outside the typical 95% range according to the Empirical Rule.

Box-and-Whisker Plot

What is a Box-and-Whisker Plot?

- A Box-and-Whisker Plot (Box Plot) visually represents the distribution of a data set using the five-number summary: minimum, Q1, median (Q2), Q3, and maximum.
- It provides a clear picture of the data spread and helps identify outliers.

Key Parts of a Box-and-Whisker Plot:

1. Box:

- Extends from Q1 to Q3, showing the middle 50% of the data.
- The line inside the box shows the median (Q2).

2. Whiskers:

- Extend from the box to the minimum and maximum values that are not outliers.

3. Outliers:

- Data points that fall outside of 1.5 times the IQR (interquartile range) from Q1 and Q3 are often marked with dots or asterisks.

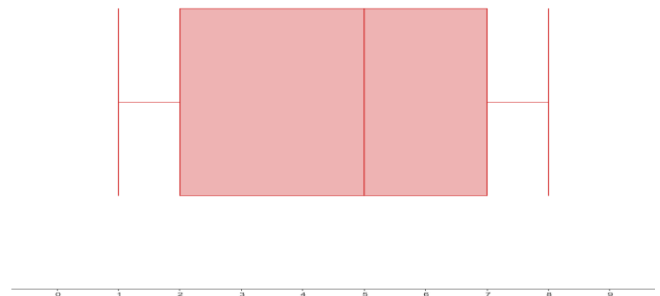
Formula: $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$

Reading a Box-and-Whisker Plot:

- **Center:** The median line inside the box shows the central value of the data.
- **Spread:** The length of the box (IQR) shows the spread of the middle half of the data.
- **Whiskers:** Show the variability outside the middle 50% of the data.
- **Outliers:** Points outside the whiskers indicate unusual values.

Five-Number Summary:

- This includes:
 1. **Minimum:** Lowest value
 2. **Q1:** First quartile
 3. **Median (Q2):** Middle value
 4. **Q3:** Third quartile
 5. **Maximum:** Highest value



- **Tip:** These numbers summarize the data quickly and can be visualized with a **Box-and-Whisker Plot**.

Probability

Probability and Counting

Probability Experiments

- **What It Is:** A probability experiment is an action or trial that produces specific results, such as tossing a coin, rolling a die, or picking a card.
- **Sample Space:** The set of all possible outcomes. For example:
 - **Coin Toss:** Sample space = {Heads, Tails}.
 - **Rolling a Die:** Sample space = {1, 2, 3, 4, 5, 6}.
- **Event:** Any subset of the sample space. For instance:
 - Getting an even number when rolling a die: {2, 4, 6}.
 - Drawing a heart from a deck of cards.

The Fundamental Counting Principle

- **What It Is:** A method to find the total number of ways multiple events can occur together.
- **Key Formula:** If one event can happen in m ways, and another can happen in n ways, then both can happen in $m \times n$ ways.
 - **Example:** Choosing a shirt (3 options) and pants (2 options): Total outfits = $3 \times 2 = 6$.

Types of Probability

- **Classical (Theoretical) Probability**
 - **Definition:** When each outcome in a sample space is equally likely to occur.
 - **Formula:** $P(E) = \frac{\text{Number of outcomes in Event E}}{\text{Total number of outcomes in sample space}}$
 - **Example:** Probability of rolling a 4 on a die: $P(4) = \frac{1}{6}$

- **Empirical (Statistical) Probability**
 - **Definition:** Based on observations or experiments rather than theory.
 - **Formula:** $P(E) = \frac{\text{Frequency of event E}}{\text{Total frequency}}$
 - **Example:** If it rained 20 out of 100 days, the probability of rain is $\frac{20}{100} = 0.2$
- **Subjective Probability**
 - **Definition:** Based on personal belief or experience rather than exact data.
 - **Example:** Estimating a 70% chance of your favorite team winning based on recent performance.

Complementary Events

- **Definition:** The complement of an event is everything that is not in the event.
- **Formula:** $P(E') = 1 - P(E)$
- **Example:** If the probability of rain today is 0.3, the probability of no rain is $1 - 0.3 = 0.7$

Tips to Remember

- **Sample Space:** Think of it as the "menu" of all possible results.
- **Event:** A "selection" from the menu of the sample space.
- **Complement:** What's left over after your event.

Formulas Recap

- **Classical Probability:** $P(E) = \frac{\text{Number of outcomes in event}}{\text{Total outcomes}}$
- **Empirical Probability:** $P(E) = \frac{\text{Frequency of event}}{\text{Total frequency}}$
- **Complementary Events:** $P(E') = 1 - P(E)$

Example Problem

- **Problem:** There are 26 defective DVRs out of 200. What is the probability of selecting a non-defective DVR?
- **Solution:** Non – defective DVRs = $200 - 26 = 174$.
Probability = $\frac{174}{200} = 0.87$

Conditional Probability

Conditional probability is the probability of an event occurring, given that another event has already occurred. It helps us figure out the likelihood of something happening when we know something else has happened.

Formula:

The probability of event B occurring given that event A has occurred is written as:

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

This reads as "the probability of B given A."

Tip:

- Think of it as "narrowing down" the possibilities because you already know some information (like knowing it's raining, which affects the probability you'll need an umbrella).

Independent and Dependent Events

- **Independent Events:** The occurrence of one event does not affect the probability of the other. For example, flipping a coin and rolling a die are independent; the result of one doesn't change the result of the other.
 - If the event A and B are independent, then $P(B|A) = P(B)$
- **Dependent Events:** The occurrence of one event affects the probability of the other. For example, drawing cards from a deck without replacement means each draw affects the next.

Multiplication Rule

The Multiplication Rule helps you find the probability of two events happening together (in sequence).

Formulas:

1. For Dependent Events: $P(A \text{ and } B) = P(A) \times P(B|A)$

This is used when one event affects the other

2. For Independent Events: $P(A \text{ and } B) = P(A) \times P(B)$

This is used when the events do not affect each other

Tip:

- If you're multiplying probabilities, remember to check if the events affect each other. If they do, use the dependent formula.

Examples

1. **Independent Events:** If you flip a coin (probability of heads is 0.5) and roll a 6-sided die (probability of getting a 3 is $\frac{1}{6}$), the probability of getting heads and a 3 is:

$$P(\text{Heads and } 3) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

2. **Dependent Events:**

Drawing two cards from a deck without replacement:

- Probability of drawing a club (13 clubs out of 52 cards)

$$\frac{13}{52} = \frac{1}{4}$$

- After drawing a club, there are 51 cards left, and still 13 hearts. Probability of drawing a heart next = $\frac{13}{51}$

- So, the combined probability is: $P(\text{Club and then Heart}) = \frac{1}{4} * \frac{13}{51} \approx 0.064$

The Addition Rule in Probability

The Addition Rule helps you find the probability that at least one of two events happen. It applies whether the events are mutually exclusive (they can't happen at the same time) or not. Here's how it works:

Key Concepts:

- 1. Mutually Exclusive Events:** Event A and B are mutually exclusive if they cannot occur at the same time. For example, rolling a 3 or a 4 on a single die roll – they can't both happen at once. When events are mutually exclusive, $P(A \text{ and } B) = 0$
- 2. Not Mutually Exclusive Events:** Events A and B are not mutually exclusive if they can occur at the same time. For example, drawing a red card and a face card from a deck of cards.

The Addition Rule Formulas:

- For Mutually Exclusive Event

$$P(A \text{ or } B) = P(A) + P(B)$$

Since A and B cannot both happen at the same time, we simply add the probabilities of A and B

- For Non-Mutually Exclusive Events:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

We subtract $P(A \text{ and } B)$ to avoid double-counting the outcomes where both A and B occur

Tips to Remember:

- If events don't overlap at all (mutually exclusive), just add the probabilities.
- If events can overlap, add the probabilities but subtract the overlap (common outcomes) to avoid counting them twice.

Example:

Suppose you're finding the probability of drawing a card that is either a King or a heart from a standard deck of 52 cards:

- There are 4 Kings in the deck, so $P(\text{King}) = \frac{4}{52} = \frac{1}{13}$
- There are 13 hearts in the deck, so $P(\text{Heart}) = \frac{13}{52} = \frac{1}{4}$
- There is 1 King of Hearts, so $P(\text{King and Heart}) = \frac{1}{52}$

Since these events are not mutually exclusive (you can draw a King that is also a heart), use the formula for non-mutually exclusive events:

$$\begin{aligned} P(\text{King and Hearts}) &= P(\text{King}) + P(\text{Heart}) - P(\text{King and Hearts}) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} \end{aligned}$$

Additional Topics in Probability and Counting

In this section, we'll explore **Permutations**, **Combinations**, and some practical tips on how to calculate them using your calculator.

Permutations

- **Definition:** A permutation is an ordered arrangement of objects. Order matters in permutations.
- **Formula:** To find the number of permutations of n objects taken r at a time use: $P(n, r) = \frac{n!}{(n-r)!}$ Where $n!$ (n factorial) means multiplying all whole numbers from n down to 1 (*e.g.*, $4! = 4 * 3 * 2 * 1 = 24$)
- **Calculator Tip:** On a typical calculator, you might find this under the " nPr " function:
 1. Enter the total number of items (n).
 2. Press the " nPr " button.
 3. Enter the number of items to arrange (r).
 4. Press "=" to get the result.

Distinguishable Permutations

- **Definition:** This is used when some objects are identical, and you want to find unique arrangements.
- **Formula:** $\frac{n!}{n_1! * n_2! * \dots * n_k!}$ where n is the total number of items, and n_1, n_2, \dots, n_k are the counts of each type of identical items

Combinations

- **Definition:** A combination is a selection of items where order does NOT matter.
- **Formula:** To find the number of combinations of n objects taken r at a time, use $C(n, r) = \frac{n!}{r!(n-r)!}$

- **Calculator Tip:** On a typical calculator, you might find this under the "nCr" function:
 1. Enter the total number of items (n).
 2. Press the "nCr" button.
 3. Enter the number of items to choose (r).
 4. Press "=" to get the result.

Examples

- **Permutations Example:** If you have 5 books labeled A,B,C,D,E and want to arrange 3 of them on a shelf in order:
 - Use $P(5,3) = \frac{5!}{(5-3)!} = \frac{5*4*3*2*1}{2*1} = \frac{120}{2} = 60$
- **Combinations Example:** A class has 26 students. In how many ways can six students form a group for an activity:
 - Use $C(26,6) = \frac{26!}{6!(26-6)!} = 230230$

Key Tips

- Permutations = **Order matters.**
- Combinations = **Order does NOT matter.**
- Always check whether the problem is asking for arrangements (permutations) or selections (combinations).

Discrete Probability Distributions

Probability Distribution describes how the probabilities are distributed over the values of a random variable.

Random Variables

- **Random Variable (x):** A variable that takes numerical values determined by the outcome of a probability experiment.
 - **Discrete Random Variable:** Can take a finite or countable number of values (like rolling a die).
 - **Continuous Random Variable:** Can take an infinite number of values within a range (like measuring height).

Discrete Probability Distributions

- Lists each possible value of a discrete random variable along with its probability.
- Must satisfy:
 1. Each probability is between 0 and 1: $0 \leq P(x) \leq 1$
 2. Sum of all probabilities is 1: $\sum P(x) = 1$

Mean, Variance, and Standard Deviation of Discrete Random Variables

- **Mean (μ):** Also called the expected value. It's the "average" value you expect over many trials.
 - **Formula:** $\mu = \sum xP(x)$
 - Multiply each value by its probability and add them all up.
- **Variance (σ^2):** Measures how much the values of the random variable differ from the mean.
 - **Formula:** $\sigma^2 = \sum (x - \mu)^2 P(x)$
 - Subtract the mean from each value, square it, multiply by the probability, and add them up.
- **Standard Deviation (σ):** The square root of the variance; it shows the spread of the distribution.
 - **Formula:** $\sigma = \sqrt{\sigma^2}$

Expected Value

- Represents the average result if the experiment is repeated many times.
- **Formula:** $E(x) = \mu = \sum xP(x)$
- It is useful for understanding the long-term behavior of the random variable.

Tips and Key Points

- **Discrete Variables** are countable (e.g., number of children, dice rolls).
- **Continuous Variables** are measurable (e.g., time, height).
- The sum of probabilities in a distribution must always be 1.
- **Mean** gives you the central tendency, while **Variance** and **Standard Deviation** tell you about the spread.

Example Calculation

Imagine a probability distribution of rolling a fair die:

- Values (x): 1, 2, 3, 4, 5, 6
- Probabilities P(x): $\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$

Mean (Expected Value) Calculation:

$$\mu = \left(1 * \frac{1}{6}\right) + \left(2 * \frac{1}{6}\right) + \left(3 * \frac{1}{6}\right) + \left(4 * \frac{1}{6}\right) + \left(5 * \frac{1}{6}\right) + \left(6 * \frac{1}{6}\right) = 3.5$$

Variance Calculation:

$$\sigma^2 = \left[(1 - 3.5)^2 * \frac{1}{6}\right] + \left[(2 - 3.5)^2 * \frac{1}{6}\right] + \dots + \left[(6 - 3.5)^2 * \frac{1}{6}\right] \approx 2.92$$

Standard Deviation Calculation:

$$\sigma = \sqrt{2.92} \approx 1.71$$

This breakdown helps to understand how the average (mean) and variability (variance/standard deviation) of outcomes are calculated for random variables.

Binomial Distributions

What is a Binomial Distribution? A binomial distribution is used when you are dealing with a situation where there are only two outcomes: success or failure. Examples include flipping a coin (heads or tails) or taking a true/false quiz. A binomial distribution helps you figure out the probability of having a certain number of successes (like getting heads) in a set number of trials (like flipping the coin 10 times).

Key Terms:

- **n:** Number of trials (how many times you do something).
- **p:** Probability of success in one trial (chance of success each time).
- **q:** Probability of failure in one trial (chance of failure each time). It is calculated as $q = 1 - p$
- **x:** Number of successes you are interested in finding the probability for.

Conditions for a Binomial Experiment:

1. Fixed number of trials (n).
2. Each trial is independent (one trial does not affect the other).
3. Only two possible outcomes (success or failure).
4. Probability of success (p) is the same for each trial.

Binomial Probability Formula: The probability of exactly x successes in n trials is given by:

$$P(x) = \binom{n}{k} p^x q^{n-x}$$

Where:

- $\binom{n}{k}$ is a combination formula and calculates how many ways x successes can occur in n trials, It is read “n choose x”
- p^x is the probability of success raised to the power of x (how many successes you want)
- q^{n-x} is probability of failure raised to the power of remaining trials

Formulas for Mean, Variance, and Standard Deviation for Binomial:

- **Mean (μ):** $\mu = np$
- **Variance (σ^2):** $\sigma^2 = npq$
- **Standard Deviation (σ):** $\sigma = \sqrt{npq}$

Tips for Using a Calculator or GeoGebra:

- In calculators like GeoGebra, you can find a "Binomial Calculator" which makes these calculations easier without doing them by hand
- To find the probability of exactly x successes, you can input the values of n, p, and x into the tool.
- For calculating mean, variance, or standard deviation, input n and p directly into their respective formulas.

Example Problem: Suppose you flip a coin 10 times. What is the probability of getting exactly 6 heads?

- Here, $n=10$, $p=0.5$ (since it's a fair coin), and $x=6$.
- Use the formula: $P(6) = \binom{10}{6} (0.5)^6 (0.5)^{10-6} \approx .2051$

** To Calculate Binomial Distribution in GeoGebra go to [page 94](#) **

More Discrete Probability Distributions: Geometric and Poisson Distributions

Geometric Distribution

- **Definition:** A geometric distribution is used when you repeat a trial until the first success occurs.
- **Conditions:**
 1. A trial is repeated until a success occurs.
 2. The trials are independent of each other.
 3. The probability of success (p) is the same for each trial.
 4. The random variable x represents the number of the trial in which the first success occurs.
- **Formula:** The probability that the first success will occur on trial number x is given by: $P(x) = p * q^{x-1}$
- **Example:** If the probability of success on each trial is $p=0.2$, and you want to find the probability that the first success occurs on the third trial, plug in the values:
$$P(3) = 0.2 * (0.8)^{3-1} = 0.2 * 0.64 = 0.128$$

Poisson Distribution

- **Definition:** The Poisson distribution is used when you want to find the probability of a certain number of events happening in a fixed interval of time, area, or volume.
- **Conditions:**
 1. You're counting the number of times x an event occurs in a given interval.
 2. The probability of the event occurring is the same for each interval.
 3. The events occur independently of each other.
- **Formula:** The probability of exactly x occurrences in an interval is given by $P(x) = \frac{\mu^x e^{-\mu}}{x!}$

Where

- μ is the average number of occurrences in the interval.
- e is approximately equal to 2.71828.

Example: If $\mu=0.8$ and you want to find the probability that exactly 2 events occur:

- $$P(2) = \frac{0.8^2 * e^{-0.8}}{2!} \approx 0.1438$$

Summary of Discrete Probability Distributions

- **Binomial Distribution:** Used when there are fixed trials, each with two outcomes (success or failure).
- **Geometric Distribution:** Used when trials are repeated until the first success.
- **Poisson Distribution:** Used when the probability of a certain number of events occurring within a fixed interval of time.

Using Technology for Calculations (reference [pg. 95](#) for step by step explanation):

- For Geometric and Poisson calculations, you can use software like GeoGebra:
 1. **Geometric Distribution:** Use the formula input to calculate the probability of the first success occurring on a specific trial.
 2. **Poisson Distribution:** Use the Poisson function in GeoGebra to calculate the probability of a given number of occurrences.

Normal Probability Distribution

Normal Distribution:

- A **normal distribution** is a type of continuous probability distribution for a random variable that is symmetrically distributed around its mean (average).
- The graph of a normal distribution is called a **bell curve** or **normal curve** because of its shape: it's highest in the middle and tapers off equally on both sides.
- **Key properties of the normal distribution:**
 1. The mean, median, and mode of the distribution are all equal.
 2. The curve is symmetric about the mean.
 3. The total area under the curve is 1 (this represents the total probability).
 4. The curve approaches the x-axis but never actually touches it.
 5. The points where the curve changes from curving upwards to downwards are called **inflection points**, located at one standard deviation (σ) away from the mean (μ).

Standard Normal Distribution:

- The **standard normal distribution** is a special case of the normal distribution that has a mean (μ) of 0 and a standard deviation (σ) of 1.
- It allows us to compare different normal distributions by converting them into a common scale (z-scores).

Z-Scores:

- A **z-score** tells you how many standard deviations a value (x) is from the mean (μ).
- The formula for calculating a z-score is: $z = \frac{x - \mu}{\sigma}$ where
 - x = value in the dataset
 - μ = mean of the dataset
 - σ = standard deviation of the dataset
- Z-scores can be used to find probabilities or areas under the normal curve.

Properties of the Standard Normal Distribution:

- The **total area** under the standard normal curve is 1.
- The **area to the left of $z = 0$** (the mean) is 0.5.
- As z-scores move away from 0 (in either direction), the cumulative area under the curve gets closer to 1 or 0.

Finding Areas Under the Curve:

- You can use a **z-table** (Standard Normal Table) or technology like a calculator or software to find the area (probability) under the curve corresponding to a specific z-score.
- **Steps to find the area:**
 1. **Left of a z-score:** Use the z-table to find the area directly.
 2. **Right of a z-score:** Find the area to the left and subtract it from 1.
 3. **Between two z-scores:** Find the area to the left of each z-score and subtract the smaller area from the larger.

Tips for Understanding Normal Distributions:

- Always sketch the curve and mark the mean and standard deviations.
- Remember that most of the data (about 68%) falls within one standard deviation of the mean.
- Use the Empirical Rule for a quick estimate:
 - 68% of data within 1 standard deviation ($\mu \pm \sigma$)
 - 95% within 2 standard deviations ($\mu \pm 2\sigma$)
 - 99.7% within 3 standard deviations ($\mu \pm 3\sigma$)

What is a Normal Distribution?

A normal distribution, also called a bell curve, is a way to describe data that is spread out symmetrically around a mean (average). The mean, median, and mode are all the same and located at the center of the distribution.

Key Points of a Normal Distribution:

- 1. Symmetrical Shape:** It's bell-shaped and symmetrical around the mean.
- 2. Mean, Median, Mode:** All three are located at the center of the curve.
- 3. Total Area:** The total area under the curve equals 1 (or 100%).
- 4. Standard Deviation:** The curve's width depends on the standard deviation (σ); a larger standard deviation means a wider curve.

What You Need to Find Probability:

- **Mean (μ):** The average value.
- **Standard Deviation (σ):** Measures the spread of the data.
- **Z-score:** A measure that describes how many standard deviations a value (x) is from the mean.

Formula for Z-Score:

$$Z = \frac{x - \mu}{\sigma}$$

- **x:** the value you're looking at.
- **μ :** the mean.
- **σ :** the standard deviation.

Finding Probabilities:

1. **Convert Values to Z-Scores:** Use the z-score formula to convert your value(s) into z-scores.
2. **Use a Z-Table or Technology:** The z-score tells you how far from the mean your value is. Use a z-table or tools like Excel or GeoGebra to find the probability associated with that z-score.

Steps to Find Probability:

1. **Identify Mean and Standard Deviation:** Find or be given these values.
2. **Calculate Z-Scores for Boundaries:** If looking for probability between two values, find the z-scores for both.
3. **Look Up Z-Scores:** Use a standard normal table (or technology) to find the probability corresponding to your z-scores.
4. **Subtract for Ranges:** If you're looking between two values, subtract the smaller probability from the larger.

Example Problem:

Suppose you want to find the probability that a student scores between 71 and 81 on a test where the mean is 87 and the standard deviation is 5.

1. Calculate Z-Scores:

- For 81: $z = \frac{81-87}{5} = -1.2$
- For 71: $z = \frac{71-87}{5} = -3.2$

2. Find Probabilities Using Z-Scores:

- Look up the probabilities corresponding to these z-scores:
 - For $z=-1.2$, the probability is **0.1151**.
 - For $z=-3.2$, the probability is **0.0007**.

3. Calculate the Probability:

- **Subtract:** $0.1151 - 0.0007 = \mathbf{0.1144}$

So, the probability of scoring between 71 and 81 is about 0.1144 or 11.44%.

Tips:

- **Excel:** Use =NORM.DIST(x, mean, standard deviation, TRUE) for cumulative probability.
- **GeoGebra:** Use the normal distribution calculator tool for visualizations and calculations.

Normal Distributions: Finding Values

Overview: In statistics, a normal distribution is a bell-shaped curve where most values are clustered around the mean (average), and the probability of values decreases as you move away from the mean. Finding values in a normal distribution often involves using the z-score, which tells you how many standard deviations a value is from the mean.

Key Concepts:

Z-Score Formula: The z-score is calculated using the formula: $z = \frac{x - \mu}{\sigma}$

where:

- x = the value you're examining
- μ = mean of the distribution
- σ = standard deviation of the distribution

Transforming Z-Scores to X-Values: To find a specific value (x) given a z-score, you can rearrange the z-score formula: $x = \mu + z\sigma$

Steps to Find Values Using Z-Scores:

- 1. Identify the given values:** You'll usually have the mean (μ), standard deviation (σ), and either the z-score or probability.
- 2. Find the Z-Score Using a Z-Table:**
 - A z-table shows the cumulative probability for a z-score.
 - To find a value, locate the cumulative area (probability) in the table and find the corresponding z-score.
- 3. Convert Z-Score to X-Value:**
 - Use the formula: $x = \mu + z\sigma$
 - Plug in the mean, standard deviation, and z-score to find the x-value.

Example: If you need to find the value representing the 95th percentile of a distribution with a mean of 100 and a standard deviation of 15:

- Look up 0.95 in the z-table to get a z-score (approximately 1.645).
- Use the formula: $x = 100 + 15(1.645) = 124.675$

Tips:

- **Using Excel:** Use =NORM.S.INV(probability) for z-scores and =NORM.INV(probability, mean, stdev) for x-values.
- **Using GeoGebra:** Input probability values to find z-scores or x-values by using built-in normal distribution tools.

Using the Z-Table:

- **Left side:** Find the first two digits of your z-score.
- **Top row:** Find the second decimal place.
- **Intersect:** Find the cumulative probability at the intersection of these two values.

Remember:

- Use the z-score to standardize any normal distribution.
- Areas under the curve to the left of the z-score correspond to probabilities (percentiles).

Sampling Distributions and the Central Limit Theorem

Sampling Distributions

- **Definition:** A sampling distribution is the probability distribution of a sample statistic, like the sample mean, obtained from random samples of size n from a population.
- **Key Idea:** When you take multiple samples from a population, the sample means will vary. The distribution of these sample means forms the sampling distribution of the sample mean.

Central Limit Theorem (CLT)

- **Definition:** The Central Limit Theorem states that if you take a large enough sample size ($n \geq 30$) from any population, the distribution of the sample means will be approximately normal (bell-shaped), even if the population itself is not normal.
- **Key Points:**
 1. If $n \geq 30$, or the population is already normally distributed, the sample means will be normally distributed.
 2. The mean of the sampling distribution ($\mu_{\bar{x}}$) equals the population mean (μ)
 3. The standard deviation of the sampling distribution ($\sigma_{\bar{x}}$), also called the standard error is $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation

Why It's Important

- The CLT allows us to make inferences about a population mean using sample data, which is crucial in statistics because we rarely have access to the entire population.

Formulas to Remember

- **Mean of Sample Means:** $\mu_{\bar{x}} = \mu$
- **Standard Error (Standard Deviation of Sample Means):**
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
- **Z-score for Sample Means:** $Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ This formula is used to find the probability of a sample mean lying within a certain range

Tips

- **Check Sample Size:** If $n \geq 30$, you can use the CLT to assume normality for the sampling distribution of the mean.
- **Use Technology:** Excel, GeoGebra, or statistical calculators can be used to find probabilities and values for sampling distributions.
- **Remember:** The larger the sample size, the closer the sampling distribution will be to a normal distribution.

How to Use in Excel/GeoGebra

- **Excel:** Use functions like NORM.DIST or NORM.S.INV to find probabilities or values related to normal distributions.
- **GeoGebra:** Provides tools for visualizing distributions and calculating probabilities using sampling distributions and the CLT.

Confidence Intervals

Confidence Intervals for the Mean: Population SD Known

When we talk about confidence intervals, we're basically trying to estimate the range where we think the true population mean (μ) lies, based on a sample mean (\bar{x}). Confidence intervals give us a way to say "We're X% sure that the true mean is somewhere between this lower value and this upper value"

Key Concepts:

Point Estimate:

- The sample mean (\bar{x}) is our best guess (point estimate) for the population mean (μ)

Margin of Error (E):

- This tells us how much we expect our sample mean (\bar{x}) to vary from the true population mean (μ)
- **Formula:** $E = z_c * \frac{\sigma}{\sqrt{n}}$
 - z_c = critical value from the Z-table corresponding to the confidence level (like 1.96 for 95% confidence)
 - σ = known population standard deviation.
 - n = sample size.

Confidence Interval:

- It's the range where we think the true mean (μ) lies.
- Formula: $\bar{x} - E < \mu < \bar{x} + E$
- So, the confidence interval is $\bar{x} \pm E$

Level of Confidence (c):

- This tells us how confident we are that the interval contains the true mean (μ).
- Common levels: 90%, 95%, 99%.
- The higher the confidence, the wider the interval.

Steps to Construct a Confidence Interval:

1. Identify the Sample Statistics:

- Sample mean (\bar{x}) and sample size (n)

2. Find the Critical Value:

- Use a Z-table to find Z_c based on your desired confidence level
- For 95% confidence, $Z_c \approx 1.96$

3. Calculate the Margin of Error (E):

- Use the formula $E = z_c * \frac{\sigma}{\sqrt{n}}$

4. Find the Confidence Interval:

- Plug in the values into $\bar{x} \pm E$ to find the lower and upper bounds

Example:

Let's say we have:

- $\bar{x} = 50$
- $\Sigma = 8$
- $n = 25$
- Confidence Level = 95%

1. Critical Value: $z_c = 1.96$

2. Margin of Error: $E = 1.96 * \frac{8}{\sqrt{25}} = 1.96 * \frac{8}{5} = 3.136$

3. Confidence Interval: $50 - 3.136 < \mu < 50 + 3.136$
 $= 46.864 < \mu < 53.136$

This means we're 95% confident that the true population mean (μ) is between 46.864 and 53.136.

Tips:

- **Use Technology:** You can use Excel or GeoGebra to calculate the critical values or even the entire confidence interval.

Confidence Intervals for the Mean: Population SD Unknown

When constructing confidence intervals for a population mean (μ) with an unknown population standard deviation (σ), the process involves using the t-distribution rather than the normal distribution. This is because the standard deviation is estimated from the sample, which adds more variability to the estimate.

Key Concepts:

Point Estimate: The best point estimate of the population mean (μ) is the sample mean (\bar{x})

Margin of Error (E): The margin of error tells you how much you can expect your sample mean to differ from the population mean due to sampling variability.

- **Formula:** $E = t_c * \frac{s}{\sqrt{n}}$
- **t_c :** The critical value from the t-distribution for the desired confidence level (found using degrees of freedom: d.f = n-1)
- **s:** Sample standard deviation.
- **n:** Sample size.

Confidence Interval Formula:

- $\bar{x} - E < \mu < \bar{x} + E$
- This interval means that if we repeated the sampling process many times, a certain percentage (e.g., 95%) of the intervals would contain the true population mean.

Steps to Construct a Confidence Interval:

1. Identify Key Values:

- Sample Mean (\bar{x})
- Sample standard deviation (s)
- Sample size (n)
- Desired confidence level (c)

2. Determine the Critical Value (t_c):

- Use a t-distribution table or software (like Excel or GeoGebra) to find the critical value corresponding to your confidence level and degrees of freedom (d.f.=n-1).

3. Calculate the Margin of Error (E):

- Use the formula $E = t_c * \frac{s}{\sqrt{n}}$

4. Construct the Confidence Interval:

- Plug your values into the formula $\bar{x} \pm E$

Tips:

- **Using Excel:** You can use the function for two tailed T.INV.2T (probability, degrees freedom) to find the critical value.
- **Using GeoGebra:** Use the command InverseTDistribution(degrees of freedom, probability) to find the critical value.
- **T-distribution:** It is used instead of the normal distribution because it accounts for the increased variability when estimating the population standard deviation from the sample.

Example:

If you have a sample mean of 13.2, a sample standard deviation of 2.0, and a sample size of 6, and you want a 95% confidence interval:

1. Find t_c for 95% confidence with $d.f. = 5$, $t_c = 2.571$
2. Calculate $E = 2.571 * \frac{2.0}{\sqrt{6}} \approx 2.1$
3. Construct the interval: 13.2 ± 2.1 , which is (11.1,15.3)
This interval means you are 95% confident that the true population mean falls between 11.1 and 15.3.

Confidence Intervals for Population Proportions

What is a Confidence Interval? A confidence interval gives a range of values that is likely to contain the true value of a population parameter (like a mean or proportion) based on a sample of data. It tells us how confident we can be that the interval includes the true value.

Key Terms:

- **Point Estimate:** This is your best guess of the population parameter from your sample data. For a population proportion, it's represented by \hat{p} , calculated as $\hat{p} = \frac{x}{n}$, where x is the number of successes and n is the total number of trials or sample size
- **Margin of Error (E):** The margin of error measures how much you can expect your sample results to differ from the true population value. For proportions, it's calculated using the formula $E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}$. Here, z_c is the critical value from the standard normal distribution $\hat{q} = 1 - \hat{p}$ and n is the sample size

Steps to Calculate Confidence Intervals for Proportions:

- **Find the point estimate \hat{p} :** Calculate the sample proportion by dividing the number of successes by the sample size
- **Check the conditions:** Ensure that $np \geq 5$ and $nq \geq 5$ to confirm that the sampling distribution can be approximated by a normal distribution
- **Find the Critical Value z_c :** Use a z-table or technology to find the critical value corresponding to your confidence level (e.g., 1.96 for 95% confidence).
- **Calculate the Margin of Error (E):** Use the margin of error formula with the values of \hat{p} , \hat{q} , and n
- **Construct the Confidence Interval:** Add and subtract the margin of error from your point estimate to find the interval: $\hat{p} - E < p < \hat{p} + E$

Example: Suppose you surveyed 2429 people, and 1367 said they made a New Year's resolution. Calculate a 90% and 95% confidence interval for the population proportion.

- **Calculate \hat{p}**

$$\hat{p} = \frac{1367}{2429} \approx 0.563$$

- **Check Conditions:**

$$np = 2429 \times 0.563 \approx 1367.5 \text{ and } np = 2429 \times 0.437 \approx 1078.08$$

Both are greater than 5, so conditions are met

- **Find Critical Values:**

- For 90% confidence: $z_c = 1.64$

- For 95% confidence: $z_c = 1.96$

- **Calculate Margin of Error (E):**

- For 90%: $E = 1.64 \sqrt{\frac{0.563 \times 0.437}{2429}} \approx 0.017$

- For 95%: $E = 1.96 \sqrt{\frac{0.563 \times 0.437}{2429}} \approx 0.020$

- **Construct Confidence Intervals:**

- 90%: $0.563 - 0.017 < p < 0.563 + 0.017$
 $0.546 < p < 0.580$

- 95%: $0.563 - 0.020 < p < 0.563 + 0.020$
 $0.543 < p < 0.583$

Tips:

- The higher the confidence level, the wider the interval.
- To reduce the margin of error, you can increase the sample size.
- Using software like Excel, GeoGebra, or a calculator with statistical functions can simplify these calculations.

Hypothesis Testing with One Sample

What is Hypothesis Testing?

- Hypothesis testing is a way to use data (like results from experiments or surveys) to decide if something is likely true or not.
- It's like checking if a claim someone makes (like a commercial saying a car gets 50 miles per gallon) can be trusted based on sample data.

Key Terms:

- **Null Hypothesis (H_0):** This is a starting point that assumes there's no effect or no difference. It's like saying, "Nothing unusual is happening"
- **Alternative Hypothesis (H_a):** This is what you want to probe. It suggests there is an effect or a difference, like saying, "Yes, something is happening here"

Steps to Follow:

- **State Your Hypotheses:**
 - The **null hypothesis (H_0)** usually includes statements with "=" (equal), " \leq " (less than or equal), " \geq " (greater than or equal)
 - The **alternative hypothesis (H_a)** involves " \neq " (not equal), "<" (less than), ">" (greater then). This represents what you are testing for.
 - **Tip:** Think of H_0 as the "**boring**" option (nothing is happening) and H_a as the "**interesting**" option (something is happening)
- **Collect Data:** You gather data from experiments or surveys. The data helps you make decisions.
- **Choose a Significance Level (α):** This is like setting your standard for "How sure do I need to be before I reject H_0 ?"

- **Common values** are 0.05 (5%), 0.01 (1%), or 0.10 (10%).
- A lower α means you need stronger evidence to reject H_0
- **Calculate the Test Statistic:**
 - This number comes from your data and depends on the type of test you are doing (like using a z-score, t-score, etc.).
 - **P-Value:** This tells you how likely your data would occur if H_0 were true.
 - **Tip:** A small P-value (less than α) means it's unlikely your data would happen by chance, so you would reject H_0
- **Make a Decision:**
 - **Reject H_0** if the P-value is less than or equal to α . This means there is strong evidence for H_a .
($Pvalue < \alpha$: *Reject H_0*)
 - **Fail to Reject H_0** if the P-value is greater than α . This means you don't have enough evidence to support H_a , but it doesn't prove H_0 is true.
($Pvalue > \alpha$: *Fail to Reject H_0*)

Types of Errors:

- **Type I Error:** Rejecting H_0 when it's actually true. Imagine thinking the product is bad when it's actually good
- **Type II Error:** Failing to reject H_0 when H_a is true. Imagine missing a bad product because the test didn't catch the issue

Significance Level (α):

- This is your tolerance for making a Type I error. Common values are 0.05, 0.01, or 0.10. If $\alpha = 0.05$, that means you are okay with a 5% chance of wrongly rejecting the null hypothesis.

P-Value:

- The p-value tells you how likely your sample data is, given that the null hypothesis is true. A small p-value ($\leq \alpha$) means the data is unlikely under the null hypothesis, so you reject H_0 .

Types of Tests:

- **One-Tailed Test:** Test for an effect in one direction (e.g., H_a uses " $>$ " or " $<$ "; tests for a difference in one specific direction only)
- **Two-Tailed Test:** Test for an effect in both directions (e.g., H_a uses " \neq " (not equal); test for any difference.)

Tips to Remember:

- Think of H_0 as “the boring, default option.”
- H_a is “what you’re trying to prove” (the exciting option)
- Small P-value means your results are surprising enough to doubt H_0
- Compare P-value to α :
 - **Less than α** = Reject H_0
 - **Greater than α** = Stick with H_0
 - This way, you’re just comparing numbers: if the P-value is small enough (less than α), you go with H_a . If not, you stick with H_0
- **Type I Error** is like a false alarm—saying something is there when it's not.
- **Type II Error** is missing the alarm—failing to see something that is there.

Hypothesis Testing for the Mean: Population SD Known

Hypotheses:

- **Null Hypothesis (H_0):** This usually states that there is no effect or difference. For example, $H_0: \mu = \mu_0$, where μ_0 is the hypothesized population mean
- **Alternative Hypothesis (H_a):** This states what you want to test for. It can be:
 - **Two tailed:** $H_a: \mu \neq \mu_0$ (testing for any difference)
 - **Left-tailed:** $H_a: \mu < \mu_0$ (testing for a decrease)
 - **Right tailed:** $H_a: \mu > \mu_0$ (testing for an increase)

Test Statistic (Z-test):

- You can use the z-test when the population standard deviation (σ) is known. The formula is $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- \bar{x} is the sample mean, μ is the hypothesized mean, σ is the population standard deviation, and n is the sample size

P-value Approach:

- **Decision Rule:**
 - If the **P-value** is less than or equal to the significance level (α), **reject H_0**
 - If the **P-value** is greater than α , **fail to reject H_0**
- A low P-value indicates strong evidence against H_0 , so you reject it

Critical Value Approach:

- **Rejection Regions** are determined by the significance level and the test type:
- **Left-tailed:** Reject H_0 if $z < -z_\alpha$
Here, $-z_\alpha$ is the critical value to the left of the mean, marking the start of the statistic z is less than this critical value, it falls in the rejection region
- **Right-tailed:** Reject H_0 if $z > z_\alpha$
Here, z_α is the critical value to the right of the mean, marking the start of the rejection region. If your test statistic z is greater than this critical value, it falls in the rejection regions
- **Two-tailed:** *Reject H_0 if $|z| > \frac{z_\alpha}{2}$*
Here, $|z|$ means you're considering the absolute value of z , so both directions (left or right) are considered. The critical value $\frac{z_\alpha}{2}$ splits the significance level equally into both tails (half on the left, half on the right), creating two rejection regions

Tips for Hypothesis Testing

- **Two-tailed tests** are used when the hypothesis involves "not equal" (\neq). It tests for any kind of difference.
- **One-tailed tests** are used when the hypothesis is directional (" $<$ " or " $>$ "). Choose right-tailed for increases and left-tailed for decreases.
- Remember: **Right = Increase, Left = Decrease** in the context of the test direct

Hypothesis Testing for the Mean: Population SD Unknown

When you don't know the population standard deviation (σ), you use the t-test for a mean. This test is helpful when you're dealing with a small sample size (less than 30) or when the population standard deviation is not given.

Key Concepts

T-Test Formula:

- Use the formula: $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

Where:

- \bar{x} = sample mean
- μ = hypothesized population mean
- s = sample standard deviation
- n = sample size
- This formula compares the sample mean (\bar{x}) to the hypothesized population mean (μ)

Degrees of Freedom (d.f.):

- The degrees of freedom are calculated as $n-1$, where n is the sample size.

Critical Values:

- Critical values come from the t-distribution table based on the level of significance (α) and degrees of freedom.
- If your test is:
 - Left-tailed: Use the "One Tail, α " column with a negative sign.
 - Right-tailed: Use the "One Tail, α " column with a positive sign.
 - Two-tailed: Use the "Two Tails, α " column with both positive and negative signs

Decision Rule:

- Compare the calculated t-value to the critical value from the t-distribution table:

- **Reject H_0** if the calculated t-value falls in the rejection region
- **Fail to reject H_0** if the t-value does not fall in the rejection region

Tips to Remember

- If the population standard deviation is **unknown** and the sample size is **small**, use the **t-test**.
- Always calculate the degrees of freedom as $n-1$.
- Use a t-table or Excel to find critical values for the given significance level and degrees of freedom.

How to Perform the T-Test in Excel

1. **Input your data:** Make sure your sample data is organized in a column.
2. **Go to the 'Data' tab** and click on 'Data Analysis'.
3. **Select 't-Test: Two-Sample Assuming Unequal Variances':**
 - For a one-sample t-test, you can compare your sample data against a hypothesized mean (adjust for one sample using data and hypothetical mean values).
4. **Input the range** of your data and the hypothesized mean.
5. **Select the significance level** (like 0.05) and whether it's one-tailed or two-tailed.
6. **Excel will provide the t-statistic, P-value, and critical value:** Use these to determine whether to reject or fail to reject H_0

Example

- **Claim:** $\mu > 73$, $\alpha = 0.05$
- Sample statistics: $\bar{x} = 74.6$, $s = 3.5$, $n = 23$
Using the t-test formula: $t = \frac{74.6 - 73}{3.5 / \sqrt{23}} \approx 2.19$
- From the t-table, for 22 degrees of freedom (since $n-1=22$) and $\alpha=0.05$ find the critical value.
- If $t > \text{critical value}$, reject H_0

Hypothesis Testing for Proportions:

What is it? Hypothesis testing for proportions helps you determine if a certain proportion in a population is significantly different from a claimed proportion. This is useful when you're dealing with percentages, like the proportion of people who support a new policy or the proportion of defective items in a batch.

Key Terms:

- **Proportion (p):** This is the percentage or fraction of the population that has a specific characteristic.
- **Sample Proportion (\hat{p}):** This is the proportion found in your sample, calculated by dividing the number of successes (x) by the sample size (n). Formula: $\hat{p} = \frac{x}{n}$

Steps to Perform a Hypothesis Test for Proportions:

- **Set Up Your Hypotheses:**
 - **Null Hypothesis (H_0):** This states that the population proportion equals a certain value ($p = p$)
 - **Alternative Hypothesis (H_a):** This suggests the proportion is different (not equal, greater than, or less than p)
- **Check Assumptions:**
 - Ensure that both $np \geq 5$ and $nq \geq 5$ (where $q = 1 - p$) to use normal approximation
- **Calculate the Test Statistic (z):**
 - Use the formula $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ where \hat{p} is your sample proportion, p is the population proportion from your null hypothesis, $q = 1 - p$, and n is the sample size

- **Find the Critical Value or P-value:**
 - Compare your calculated z-value against critical values from the z-table based on your significance level (α).
 - For P-value: Look up your z-value in the z-table to find the probability.
- **Make a Decision :**
 - **Reject H_0 :** If the z-value falls into the rejection region or if the P-value is less than α
 - **Fail to Reject H_0 :** If the z-value does not fall into the rejection region or if the P-value is greater than α

Types of Tests:

- **Left-tailed test:** Use when testing if the proportion is less than a certain value ($H_a: p < p_0$)
- **Right-tailed test:** Use when testing if the proportion is greater than a certain value ($H_a: p > p_0$)
- **Two-tailed test:** Use when testing if the proportions is not equal to certain value ($H_a: p \neq p_0$)

(p_0 : The hypothesized population proportion (the value you're testing against).)

Correlation and Regression

Correlation

The concept of **correlation** deals with understanding the relationship between two variables. When we talk about correlation, we're asking, "Do these two variables move together in some predictable way?"

Key Points of Correlation:

1. Correlation Coefficient (r):

- This is a value between -1 and 1 that tells us how strongly two variables are related.
- If $r = 1$ it's a perfect positive correlation (as one variable increases, the other does too).
- If $r = -1$, it's a perfect negative correlation (as one variable increases, the other decreases).
- If $r = 0$, there's no linear relationship between the variables.

2. Scatter Plot:

- A scatter plot is a graph with points plotted to show a possible relationship between two sets of data.
- The x-axis represents the independent (or explanatory) variable, and the y-axis represents the dependent (or response) variable.

3. Types of Correlation:

- Positive Correlation: As one variable increases, the other also increases (r is positive).
- Negative Correlation: As one variable increases, the other decreases (r is negative).
- No Correlation: There's no consistent pattern between the two variables (r is close to 0).

4. Strength of Correlation:

- Strong Correlation: Values of r close to -1 or 1.
- Weak Correlation: Values of r close to 0.

How to Calculate the Correlation Coefficient (r):

The correlation coefficient r can be calculated using the formula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Where:

- n = number of data pairs
- x and y are the variables.

Hypothesis Testing for Correlation:

When testing if a correlation is significant (meaning it's not due to random chance), we can use a hypothesis test:

- **Null Hypothesis (H_0):** There is no significant correlation ($p = 0$)
- **Alternative Hypothesis (H_a):** There is a significant correlation ($p \neq 0$)

Using Technology:

- You can use calculators, software, GeoGebra, or Excel to find the correlation coefficient easily.
- Excel: Use the =CORREL (array1, array2) function to calculate the correlation coefficient.

Test Statistics for Correlation:

When you want to test whether a correlation is statistically significant, you can use a **t-test for the correlation coefficient**. This helps determine if the observed correlation in your sample reflects a true correlation in the population.

t-Test for the Correlation Coefficient:

The test statistic t is calculated using the formula:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Where:

- r is the correlation coefficient.
- n is the number of data pairs (sample size).
- $n-2$ represents the degrees of freedom (because we lose two degrees of freedom when estimating both the mean of x and y).

Steps to Perform the Test:

1. State the Hypotheses:

- **Null Hypothesis H_0 :** The population correlation coefficient is zero ($\rho=0$); no linear relationship
- **Alternative Hypothesis (H_a):** The population correlation coefficient is not zero ($\rho \neq 0$); there is a linear relationship

2. Calculate the Test Statistic t :

Use the formula above with your sample data.

3. Determine the Critical Value:

- Use the t-distribution table with $n-2$ degrees of freedom.
- Check the critical value for your chosen significance level (α), commonly 0.05 or 0.01.

4. Decision Rule:

- **Reject H_0** if the absolute value of the test statistics t is greater than the critical value from the t-table
- **Fail to Reject H_0** if t is less than or equal to the critical values

Critical Values:

- Critical values are the threshold values that determine whether to reject the null hypothesis.
- They depend on the significance level (α) and the degrees of freedom ($n-2$).

Finding Critical Values:

1. **Determine α :** Common values are 0.05 or 0.01.
2. **Find Degrees of Freedom (df):** $n-2$.
3. **Use a t-Distribution Table:** Locate the row for your df and the column for your α to find the critical value.

Example:

Suppose you have 10 pairs of data points ($n=10$), and you found a correlation coefficient of $r=0.6$. You want to test if this correlation is significant at the 0.05 level.

1. **Calculate the test statistic t:**

$$t = \frac{0.6}{\sqrt{\frac{1 - 0.6^2}{10 - 2}}} = \frac{0.6}{\sqrt{\frac{1 - 0.36}{8}}} = \frac{0.6}{\sqrt{\frac{0.64}{8}}} = \frac{0.6}{\sqrt{0.08}} \approx 2.12$$

2. **Determine the critical value:**

- With $df=10-2=8$ and $\alpha=0.05$, the two-tailed critical value from the t-table is approximately ± 2.306 .

3. **Compare t to the critical value:**

- Since $|2.12| < 2.306$, you fail to reject H_0 . There is not enough evidence to say the correlation is significant

Using Excel:

- You can use Excel to calculate t and find critical values using functions like T.INV.2T(alpha, df) for critical values and =CORREL(range1, range2) for the correlation coefficient.

Linear Regression

What is Linear Regression?

- Linear regression is a method used to find the relationship between two variables by fitting a straight line (regression line) through the data points on a scatter plot.
- This line can be used to predict values of one variable (dependent variable, y) based on the other variable (independent variable, x).

Key Concepts:

1. **Regression Line:** The line that best fits the data, minimizing the sum of the squares of the vertical distances (residuals) between the data points and the line.
2. **Equation of the Line:** The equation of the regression line is $\hat{y}=mx+b$, where:
 - **m** is the **slope** (how steep the line is),
 - **b** is the **y-intercept** (where the line crosses the y-axis),
 - \hat{y} is the predicted value of y for a given x
3. **Slope m:** Indicates how much y changes for a unit change in x . If m is positive, y increases as x increases, and if m is negative, y decreases as x increases.

Calculating the Line:

- The slope m and y-intercept b can be calculated using formulas that involve the means of x and y , and sums of their products and squares.
- However, it's often easier to use software or tools like Excel, GeoGebra, or a graphing calculator.

Using Excel to Find the Regression Line:

1. Input your data in two columns (e.g., x in column A and y in column B).
2. Highlight the data, go to the "Insert" tab, and select "Scatter Plot."
3. Click on the plot, and add a trendline by selecting "Add Trendline" and then "Linear."
4. Check the "Display Equation on chart" option to see the line's equation.

Practical Tips:

- **Outliers:** Be cautious of data points that are far from the rest; they can affect the line significantly.
- **Extrapolation:** Predictions outside the range of the original data (extrapolation) can be unreliable.
- **Software Tools:** Besides Excel, tools like GeoGebra and graphing calculators can automate regression calculations, making it easy to visualize and compute the regression line without manual calculations.

Measures of Regression and Prediction Intervals

Key Concepts:

Variation About a Regression Line:

- **Total Variation:** The overall measure of how much the data points deviate from the mean of the response variable (\bar{y})
- **Explained Variation:** The part of the total variation that is explained by the regression line. It shows how much of the data's variability can be explained by the relationship between x and y.
- **Unexplained Variation:** The portion of the variation that the regression line does not explain. It shows how much variation is due to other factors or randomness.

Coefficient of Determination(r^2)

- This measures the proportion of the variance in the dependent variable that is predictable from the independent variable. It is calculated as $r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$
- For example, if $r^2 = 0.81$, it means 81% of the variability in y is explained by x, and the remaining 19% is unexplained

Standard Error of Estimate (s_e)

- This measures the average distance that the observed values fall from the regression line, it is calculated as $s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$
- A smaller s_e indicates that the data points are closer to the regression line.

Prediction Intervals:

- A prediction interval gives a range of values that is likely to contain the value of y for a given x . It accounts for the variability in the data and is calculated using: $\hat{y} - E < y < \hat{y} + E$
- Here, E (margin of error) considers the standard error and the confidence level.

Using Excel or GeoGebra:

- Excel: Use the LINEST function to get the regression statistics, including slope, intercept, r^2 , and standard error
- For prediction intervals, you can use the regression output and additional statistical functions to calculate the required values.
 - **GeoGebra**: Input the data points into GeoGebra and use the regression tools to find the line of best fit and the coefficient of determination.

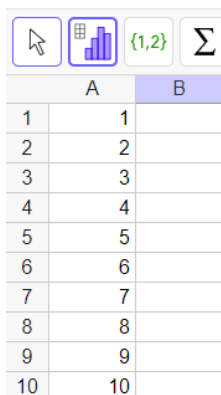
References

How to find Range, Variance, and Standard Deviation in Excel

	A	B	C	D	E	F
1	1		RANGE	MAX(A1:A10)-MIN(A1:A10)		
2	2		Pop. Variance	VAR.P(A1:A10)		
3	3		Sample Variance	VAR.S(A1:A6)		
4	4		Pop. Stdev	STDEV.P(A1:A10)		
5	5		Sample Stdev	STDEV.S(A1:A6)		
6	6					
7	7					
8	8					
9	9					
10	10					

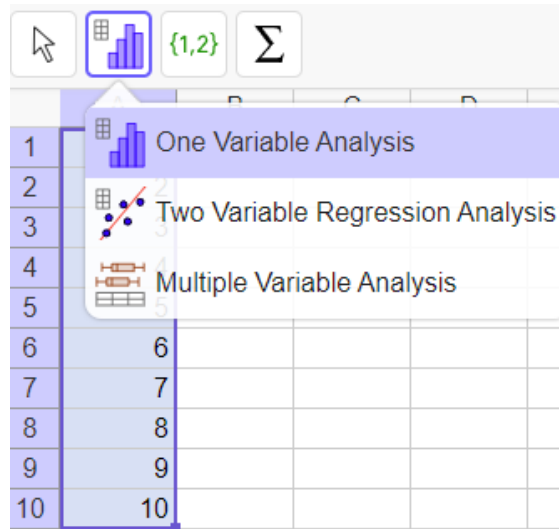
How to find Range, Variance, and Standard Deviation in GeoGebra

Step 1: Copy your data in GeoGebra Spreadsheet



	A	B
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	

Step 2: Highlight your data and then Click on the 2nd box and choose ONE VARIABLE ANALYSIS



Step 3: Look on the right side until you find this symbol Σx and then click on it



Step 4: Just read the data and use what you need

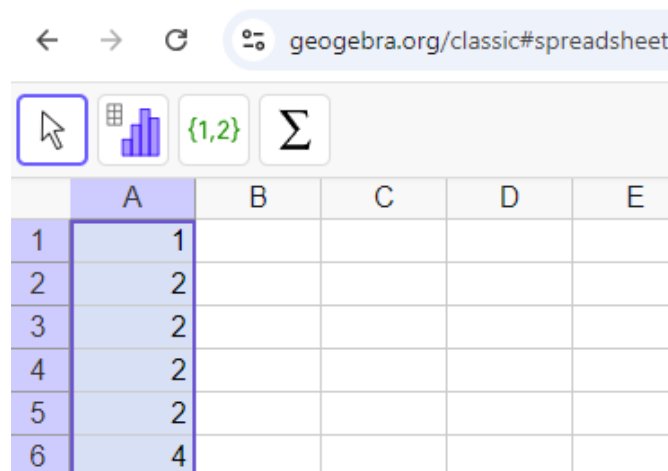
Statistics	
n	10
Mean	5.5
σ	2.8723
s	3.0277
Σx	55
Σx^2	385
Min	1
Q1	3
Median	5.5
Q3	8
Max	10

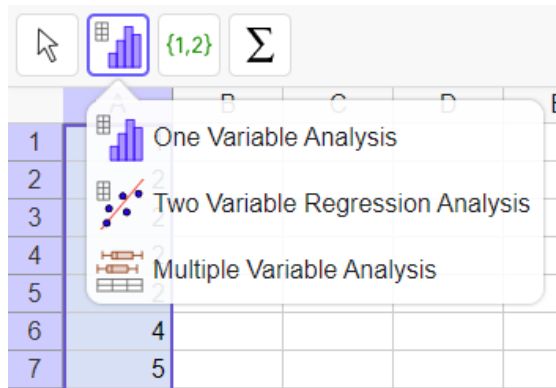
How to find Quartiles/IQR and Percentiles in Excel

	A	B	C	D
1	1		First Quartile	QUARTILE.EXC(A1:A15,1)
2	2			
3	2		Second Quartile (Median)	QUARTILE.EXC(A1:A15,2)
4	2			
5	2		Third Quartile	QUARTILE.EXC(A1:A15,3)
6	4			
7	5		IQR (Q3-Q1)	D5-D1
8	5			
9	5		90th Percentile	PERCENTILE.EXC(A1:A15,0.90)
10	6			
11	6		80th Percentile	PERCENTILE.EXC(A1:A15,0.8)
12	7			
13	8			
14	8			
15	8			

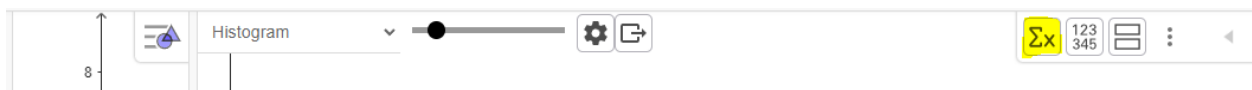
How to find Quartiles/IQR

Step 1: Go to GeoGebra classic and pull up the spreadsheet and enter your data and highlight your data





Step 3: Then look at the right side of the screen and find the symbol Σx and click on it

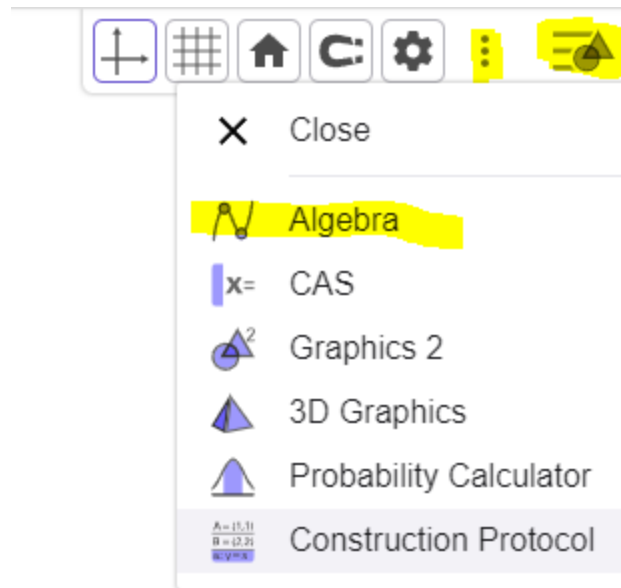


Step 4: Read the data that they gave you as you see they give Q1, Q2 (Median), Q3

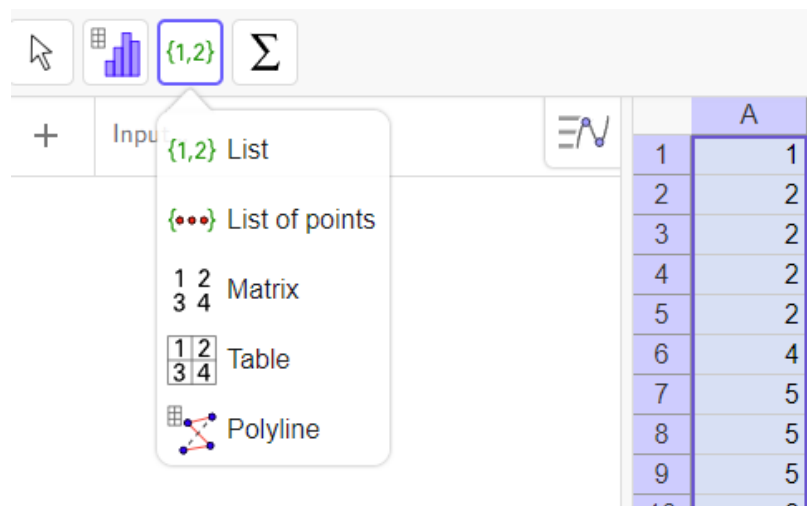
Statistics	
n	15
Mean	4.7333
σ	2.3795
s	2.4631
Σx	71
Σx^2	421
Min	1
Q1	2
Median	5
Q3	7
Max	8

How to find Percentile in GeoGebra

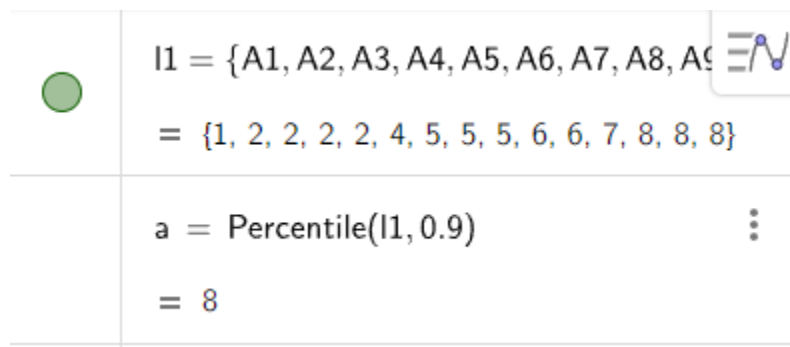
Step 1: With your spreadsheet still up, you need to also pull up the Algebra Calculator. In order to do that you are going to look at the right side and click on the figure that has a circle and triangle together and after that you are going to click on the three vertical dots and then choose the Algebra Calculator



Step 2: Highlight your data and then click on the third box and click list

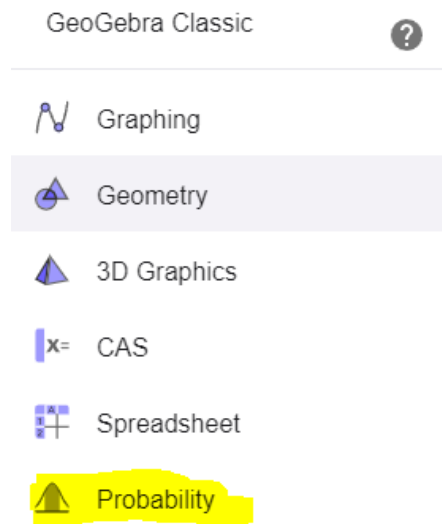


Step 3: On the Algebra Calculator type Percentile (Data list, percentile (decimal))

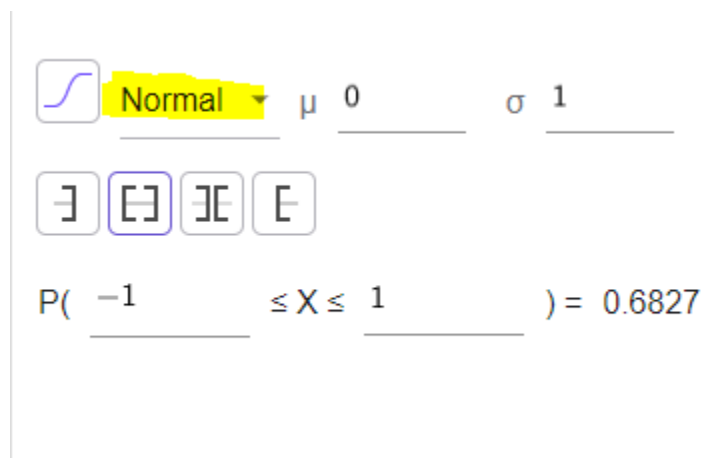


How to find Binomial Distribution in GeoGebra

Step 1: Go onto GeoGebra Classic and on the right-hand side click on Probability



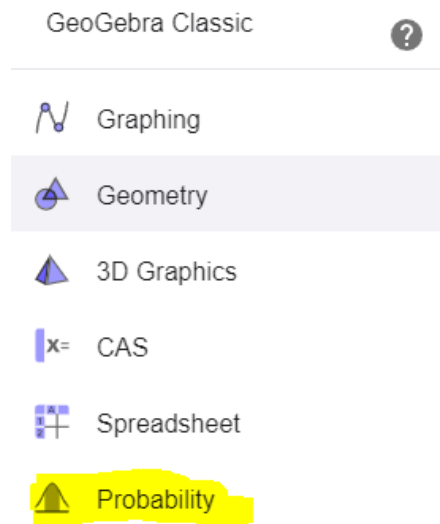
Step 2: Now look at the bottom left corner there is a drop-down tab that says normal click on that change it to binomial



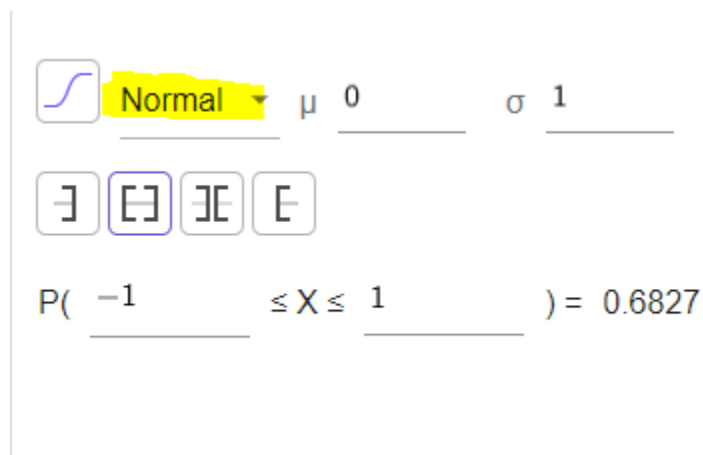
Step 3: You just fill in your information like the mean, standard deviation and at the range you need

How to find Poisson/Geometric Distribution in GeoGebra

Step 1: Go onto GeoGebra Classic and on the right-hand side click on Probability



Step 2: Now look at the bottom left corner there is a drop-down tab that says normal click on that change it to Poisson/Geometric



Step 3: You just fill in your information like that is need for each of the distributions

Z-Table (-)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Z-Table (+)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998