Table of Contents

Symbol	Description	Definition	Example
P(A ∩ B)	Probability of Intersection	Probability that both events A and B occur	P(A ∩ B) = 0.3
P(A)	Probability Function	Probability of event A occurring	P(A) = 0.6
P(A/B)	Conditional Probability	Probability of event A occurring given that event B has occurred	
P(A U B)	Probability of Union	Probability that either event A or B occurs	P(A U B) = 0.54
F(x)	Cumulative Distribution Function	Probability that X is less than or equal to x	F(x) = P(X ≤ x)
f(x)	Probability Density Function	Density function of X at x	Integral of f(x)
E(X)	Expectation Value	Expected value of random variable X	E(X) = 10
μ	Population Mean	Mean of population values	μ = 10
var(X)	Variance	Variance of random variable X	var(X) = 4
E(X/Y)	Conditional Expectation	Expected value of X given Y	
std(X)	Standard Deviation	Standard deviation of random variable X	std(X) = 2
σ^2	Variance	Variance of population values	$\sigma^2 = 4$
$N(\mu,\sigma^2)$	Normal Distribution	Gaussian distribution with mean μ and variance σ^2	X ~ N(0,3)
Bin(n,p)	Binomial Distribution	Probability distribution of n successes with probability p	$f(k) = {}_nC_k \cdot p^{(k)}(1-p)^{(n-k)}$
$\chi^2(k)$	Chi-Square Distribution	Distribution of a sum of the squares of k independent standard normal random variables	
Poisson(λ)	Poisson Distribution	Probability of a given number of events occurring in a fixed interval	

Symbol	Description	Definition	Example
σ	Standard Deviation	Standard deviation value of random variable X	σ = 2
corr(X,Y)	Correlation	Correlation of random variables X and Y	corr(X,Y) = 0.6
cov(X,Y)	Covariance	Covariance of random variables X and Y	cov(X,Y) = 4
Q2	Median / Second Quartile	Middle value of the dataset, dividing it into two equal parts	Q2 = median value
Q1	Lower / First Quartile	Value below which 25% of the data lies	Q1 = lower quartile value
Q3	Upper / Third Quartile	Value above which 25% of the data lies	Q3 = upper quartile value
s	Sample Standard Deviation	Standard deviation estimated from a sample	s = 2
s^2	Sample Variance	Variance estimated from a sample	$s^2 = 4$
X ~	Distribution of X	The distribution that random variable X follows	X ~ N(0,3)
Zx	Standard Score (Z-Score)	Standardized value of X calculated by (X - mean) / standard deviation	Zx = (X - μ) / σ
U(a,b)	Uniform Distribution	Distribution where all outcomes are equally likely within range [a,b]	X ~ U(0,3)

Unit 1: Exploring One-Variable Data

Name	Topic or Formulas
Introduction to Statistics	▪ Understanding what statistics is and its importance. ▪ Different types of data: categorical vs. quantitative.
The Language of Variation	Coefficient of Variation: $CV = \frac{s}{\overline{x}} \times 100\%$
Representing Categorical Variables	typically involves charts like pie charts and bar graphs
Representing Quantitative Variables	involves histograms, dot plots, etc.
Describing Distributions of Quantitative Variables	▪ Mean: $\overline{x} = \frac{\sum x_i}{n}$ ▪ Median: Middle value in a sorted list ▪ Mode: Most frequently occurring value
Summary Statistics	▪ Range: $\text{Range} = \text{Max} - \text{Min}$ ▪ Variance: $s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$ ▪ Standard Deviation: $s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
Graphical Representations	involves the creation of graphs such as box plots, stem-and-leaf plots
The Normal Distribution	▪ Standard Normal Variable: $Z = \frac{X - \mu}{\sigma}$ ▪ Probability for Normal Distribution: $P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$ ▪ normalcdf(lower z-score, upper z-score) ▪ invNorm(area) ▪ normalcdf(lowerbound, upperbound, mean, SD) ▪ invNorm(area, mean, SD)

1-1. Introduction to Statistics

- Understanding what statistics is and its importance.
- Different types of data: categorical vs. quantitative.

Why Learn This?

- Understanding statistics is essential for making informed decisions based on data. Whether in everyday life or professional fields, the ability to interpret data correctly is crucial.

Concepts:

- **Statistics**: The science of collecting, analyzing, and interpreting data.
- Types of Data:
  - **Categorical** (e.g., gender, race)
  - **Quantitative** (e.g., height, weight)

Example:

- Categorical Data: Gender (Male, Female)
- Quantitative Data: Heights of students in a class (in inches).

Basic Questions for Understanding:

1) You survey your classmates to find out their favorite fruit. Out of 30 students, 12 prefer apples, 10 prefer bananas, and 8 prefer oranges. What is the most popular fruit?

- Answer: The most popular fruit is apples, with 12 students preferring it.

2) You want to know the average number of hours students spend on homework each week. If you ask 10 students and find that they spend between 5 and 15 hours, what type of data are you collecting?

- Answer: You are collecting quantitative data because you're measuring the number of hours, which is a numerical value.

3) In a class of 25 students, 15 are boys and 10 are girls. What type of data is this, categorical or quantitative?

- Answer: This is categorical data because you are categorizing the students by gender (boys and girls).

4) You collect the heights of 20 students in inches. What type of data are you collecting?

- Answer: You are collecting quantitative data because height is a numerical measurement.

1. Understanding What Statistics Is

[1] A school wants to know the average test score of its students. What statistical process would they use to find this information?

- Answer: The school would use statistics to collect, analyze, and interpret the data on students' test scores to calculate the average.

[2] If a company wants to determine the most popular product among its customers, what role does statistics play?

- Answer: Statistics helps the company collect data on customer preferences, analyze it to find trends, and interpret which product is the most popular.

[3] A doctor collects data on patient recovery times after surgery. How does statistics help in this situation?

- Answer: Statistics allows the doctor to analyze the recovery times, identify patterns, and interpret whether certain factors influence faster or slower recovery.

[4] How would a government use statistics to understand unemployment rates?

- Answer: The government would collect data on employment status, analyze it to calculate the unemployment rate, and interpret the results to make informed decisions about policies.

2. Types of Data: Categorical vs. Quantitative

[1] You survey your classmates about their favorite color. Is this categorical or quantitative data?

- Answer: This is categorical data because it involves grouping the responses into categories based on color.

[2] If you measure the height of every student in your class, what type of data are you collecting?

- Answer: You are collecting quantitative data because height is a numerical measurement.

[3] A researcher records the type of vehicles people drive (car, truck, SUV). Is this data categorical or quantitative?

- Answer: This is categorical data because it categorizes the vehicles into different types.

[4] If a teacher counts the number of books each student reads over the summer, what type of data is this?

- Answer: This is quantitative data because the number of books is a numerical value that can be counted.

3. Importance of Statistics

[1] Why is it important for a business to use statistics when planning a new product launch?

- Answer: It's important because statistics help the business understand market trends, customer preferences, and potential demand, leading to better-informed decisions.

[2] How can statistics help in making health-related decisions?

- Answer: Statistics can help by analyzing data from clinical trials, patient outcomes, and disease prevalence to guide healthcare policies and treatments.

[3] A city is planning to build new parks. How can statistics assist in this process?

- Answer: Statistics can be used to analyze population data, current park usage, and community needs to determine the best locations and features for the new parks.

[4] Why would a teacher use statistics to evaluate the effectiveness of a new teaching method?

- Answer: The teacher would use statistics to compare test scores, student engagement, and other metrics before and after implementing the method to see if it has a positive impact.

4. Real-Life Examples of Categorical and Quantitative Data

[1] You collect data on the types of pets owned by students in your class. What type of data is this?

- Answer: This is categorical data because it groups students based on the type of pet they own.

[2] A doctor measures the blood pressure of patients. What type of data is this?

- Answer: This is quantitative data because blood pressure is measured with numerical values.

[3] If you ask people about their favorite sports team, what kind of data are you collecting?

- Answer: You are collecting categorical data because the responses are grouped into categories based on different sports teams.

[4] A survey asks how many hours people exercise each week. What type of data does this represent?

- Answer: This is quantitative data because it involves counting the number of hours, which is a numerical value.

1-2. The Language of Variation

- Variables: definition, types (independent vs. dependent).

Why Learn This?

- Understanding variation helps us identify patterns and relationships in data, which are key to drawing accurate conclusions and making predictions.

Concepts:

- **Variables**: Characteristics measured or observed.
- Types of Variables:
  - **Independent Variable**: The variable that is manipulated.
  - **Dependent Variable**: The variable that responds to the independent variable.

Example:

- **Independent Variable:** Amount of water (liters).
- **Dependent Variable:** Plant growth (cm).

Basic Questions for Understanding:

1) You measure the length of your morning commute to school each day for a week. Which is the independent variable: the day of the week or the commute time?

- Answer: The day of the week is the independent variable, as it is not affected by the commute time.

2) If you record the temperature at different times of the day, which variable is the dependent variable?

- Answer: The temperature is the dependent variable because it depends on the time of day.

3) In an experiment to see if listening to music affects concentration, which is the independent variable: the presence of music or concentration level?

- Answer: The presence of music is the independent variable because it's what you change to see its effect on concentration.

4) If you are testing how the amount of water affects plant growth, what is the dependent variable?

- Answer: The plant growth is the dependent variable because it depends on the amount of water given.

1. Understanding Independent and Dependent Variables

[1] A scientist is testing the effect of sunlight on the growth of a plant. What is the independent variable in this experiment?

- Answer: The independent variable is the amount of sunlight the plant receives, as this is the factor that is being manipulated.

[2] If you are measuring how different types of fertilizer affect the height of tomato plants, what is the dependent variable?

- Answer: The dependent variable is the height of the tomato plants, as this is the outcome that responds to the different types of fertilizer.

[3] In a study to determine how sleep affects concentration levels, what would be the independent variable?

- Answer: The independent variable is the amount of sleep, as it is the variable being changed to see its effect on concentration levels.

[4] A researcher is examining how the amount of time spent studying affects test scores. What is the dependent variable?

- Answer: The dependent variable is the test scores, as they are expected to change in response to the amount of time spent studying.

2. Identifying Variables in Experiments

[1] A company is testing whether different advertising methods increase product sales. What is the independent variable?

- Answer: The independent variable is the type of advertising method used, as this is what the company is changing in the experiment.

[2] If a study looks at how the dosage of a drug affects recovery time, what is the dependent variable?

- Answer: The dependent variable is the recovery time, as it is the variable that is expected to change based on the drug dosage.

[3] In an experiment to see how temperature affects the rate at which ice melts, what would be the independent variable?

- Answer: The independent variable is the temperature, as it is being controlled to observe its effect on the melting rate of ice.

[4] A scientist wants to know how the number of hours of exercise per week affects weight loss. What is the dependent variable?

- Answer: The dependent variable is weight loss, as it is the result that depends on the amount of exercise.

3. Exploring Relationships Between Variables

[1] You are exploring how different amounts of water affect the speed of plant growth. Which variable do you control, and which one do you measure?

- Answer: You control the amount of water (independent variable) and measure the speed of plant growth (dependent variable).

[2] A teacher is investigating how different teaching methods impact students' test scores. What is the independent variable in this scenario?

- Answer: The independent variable is the teaching method used, as the teacher is altering this to see how it affects test scores.

[3] In a study on how caffeine consumption influences reaction time, what would be the dependent variable?

- Answer: The dependent variable is the reaction time, as it is expected to change based on the amount of caffeine consumed.

[4] If a researcher is testing the effect of different diets on energy levels, what are the independent and dependent variables?

- Answer: The independent variable is the type of diet, and the dependent variable is the energy levels, as the diet is expected to affect energy.

4. Making Predictions Based on Variables

[1] If you increase the amount of fertilizer, what do you predict will happen to the growth of the plants?

- Answer: If you increase the amount of fertilizer (independent variable), you would predict that the growth of the plants (dependent variable) will increase, assuming a positive relationship.

[2] A study shows that more hours of study lead to higher grades. What can you predict about the grades if a student doubles their study time?

- Answer: If a student doubles their study time (independent variable), you can predict that their grades (dependent variable) will improve, based on the study's findings.

[3] If a researcher finds that reducing screen time improves sleep quality, what would happen to sleep quality if screen time is increased?

- Answer: If screen time (independent variable) is increased, sleep quality (dependent variable) would likely decrease, based on the negative relationship.

1-3. Representing Categorical Variables

- Frequency tables and relative frequency tables.
- Bar charts and pie charts.

Why Learn This?

- Properly representing categorical data helps us summarize and visualize patterns, making it easier to understand the distribution of categories and draw meaningful insights.

Concepts:

- **Frequency Table**: Shows counts of each category.
- **Relative Frequency**: Proportion of total observations for each category.
- **Bar Chart**: Graphical representation of categorical data.
- **Pie Chart**: Circular chart divided into sectors, representing categories.

Formulas:

- $\text{Relative Frequency} = \dfrac{\text{Frequency}}{\text{Total Number of Observations}}$

Example:

- Frequency Table:

Gender	Frequency
Male	15
Female	20

- Relative Frequency: $\text{Relative Frequency of Male} = \dfrac{15}{35} = 0.43$

Basic Questions for Understanding:

1) You survey your friends about their favorite ice cream flavor and record the results in a table. Which type of chart would best display this data?

- Answer: A bar chart would be best to display the favorite ice cream flavors.

2) You ask 20 students if they prefer summer or winter. How could you represent this data visually?

- Answer: You could represent this data with a pie chart showing the proportion of students who prefer each season.

3) If 70% of students in your class prefer online learning and 30% prefer in-person learning, how could you show this data visually?

- Answer: You could use a bar chart or a pie chart to visually represent the proportions of students who prefer each type of learning.

4) You ask your classmates what their favorite subject is and record the data in a frequency table. How could you visually represent this data?

- Answer: You could use a bar chart to show how many students prefer each subject.

1. Creating and Interpreting Frequency Tables

[1] In a class survey, 12 students prefer apples, 8 prefer bananas, and 5 prefer oranges. How would you represent this data in a frequency table?

- Answer: The frequency table would look like this:

Fruit	Frequency
Apples	12
Bananas	8
Oranges	5

[2] If a survey shows that 30 people like action movies, 25 like comedies, and 10 like dramas, how would you display this in a frequency table?

- Answer: The frequency table would be:

Movie Genre	Frequency
Action	30
Comedy	25
Drama	10

2. Calculating and Interpreting Relative Frequencies

[1] If 15 out of 50 students in a class prefer chocolate ice cream, what is the relative frequency of students who prefer chocolate?

- Answer: The relative frequency is $\frac{15}{50} = 0.30$, meaning 30% of the students prefer chocolate ice cream.

[2] In a survey, 40 out of 100 people say they exercise daily. What is the relative frequency of people who exercise daily?

- Answer: The relative frequency is $\frac{40}{100} = 0.40$, meaning 40% of the people exercise daily.

[3] A teacher finds that 10 out of 25 students have pets. What is the relative frequency of students who have pets?

- Answer: The relative frequency is $\frac{10}{25} = 0.40$, meaning 40% of the students have pets.

[4] If 75 out of 300 voters support a candidate, what is the relative frequency of support?

- Answer: The relative frequency is . $\frac{75}{300} = 0.25$, meaning 25% of the voters support the candidate.

3. Using Bar Charts to Represent Data

[1] You collect data on the favorite sports of students: 20 prefer soccer, 15 prefer basketball, and 10 prefer baseball. How would you represent this data in a bar chart?

- Answer: You would create a bar chart with three bars, each representing a different sport. The heights of the bars would correspond to the frequency: Soccer (20), Basketball (15), and Baseball (10).

[2] A survey finds that 60% of people prefer coffee, 30% prefer tea, and 10% prefer juice. How would this data look on a bar chart?

- Answer: The bar chart would have three bars representing Coffee, Tea, and Juice. The heights of the bars would reflect the percentages: Coffee (60%), Tea (30%), and Juice (10%).

4. Using Pie Charts to Represent Data

[1] If a classroom has 10 students who walk to school, 15 who take the bus, and 5 who bike, how would you represent this data on a pie chart?

- Answer: The pie chart would be divided into three sections: Walking (10/30 = 33.3%), Bus (15/30 = 50%), and Biking (5/30 = 16.7%).

[2] A company survey shows that 40% of employees prefer working from home, 30% prefer the office, and 30% have no preference. How can this be visualized using a pie chart?

- Answer: The pie chart would be divided into three sectors: 40% for Working from Home, 30% for the Office, and 30% for No Preference.

1-4. Representing Quantitative Variables

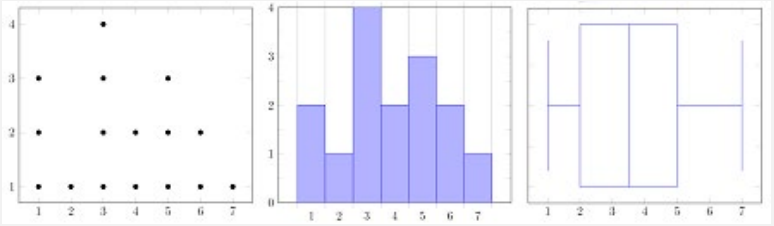
- Histograms, dot plots, and stem-and-leaf plots.
- Box plots and cumulative frequency graphs.

Why Learn This?

- Representing quantitative data visually allows us to quickly identify trends, patterns, and outliers, which is essential for interpreting and understanding the data.

Concepts:

- **Histogram**: A bar graph that represents the frequency distribution of a dataset.
- **Dot Plot**: A simple plot on a number line where each data point is shown as a dot.
- **Stem-and-Leaf Plot**: A plot where each data value is split into a "stem" and a "leaf."



Example:

- Histogram: Plot the number of students scoring within certain ranges on a test.
- Dot Plot: Heights of 10 students represented as dots on a number line.
- Stem-and-Leaf Plot:

Data: 34, 36, 42, 44, 45

Stem		Leaf
3	|	4, 6
4	|	2, 4, 5

Basic Questions for Understanding:

1) You measure the heights of 10 students in your class. Which type of graph could you use to show the distribution of their heights?

- Answer: You could use a histogram to show the distribution of the students' heights.

2) You want to see the distribution of test scores in your class. Which graph would help you visualize this?

- Answer: A dot plot would help you visualize the distribution of test scores.

3) You record the number of books each student read last month. How could you represent this data on a graph?

- Answer: You could use a stem-and-leaf plot to represent the number of books each student read.

4) To show how many hours students studied for an exam, what kind of graph could you use?

- Answer: A histogram or dot plot would be good for showing how many hours students studied.

1. Creating and Interpreting Histograms

[1] A teacher records the test scores of 30 students and wants to know how many students scored between 70 and 80. How would this be shown on a histogram?

- Answer: On the histogram, there would be a bar representing the range of scores from 70 to 80. The height of the bar would show the number of students who scored in that range.

[2] If a company tracks the weekly sales of a product, how could they use a histogram to understand sales distribution?

- Answer: The company could create a histogram with bars representing different sales ranges (e.g., 0-50, 51-100). The height of each bar would indicate how many weeks had sales within that range, helping to visualize the distribution of weekly sales.

[3] In a survey of students' daily screen time, how would you use a histogram to show how many students spend between 1 and 2 hours on screens?

- Answer: You would create a histogram with a bar representing the 1-2 hour range. The height of the bar would show the number of students who reported spending that amount of time on screens.

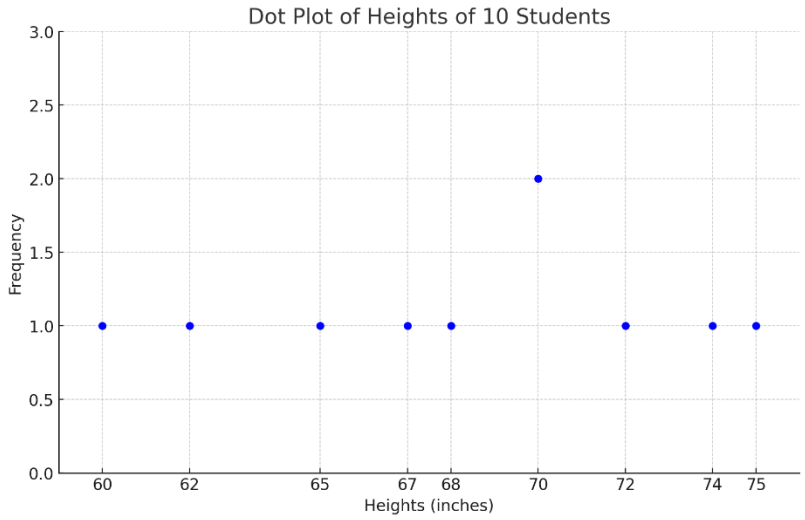
[4] A researcher wants to display the distribution of ages in a sample population. How would they use a histogram for this?

- Answer: The researcher would create a histogram where each bar represents an age range (e.g., 20-30, 31-40). The height of each bar would show how many people in the sample fall into each age range.

2. Creating and Interpreting Dot Plots

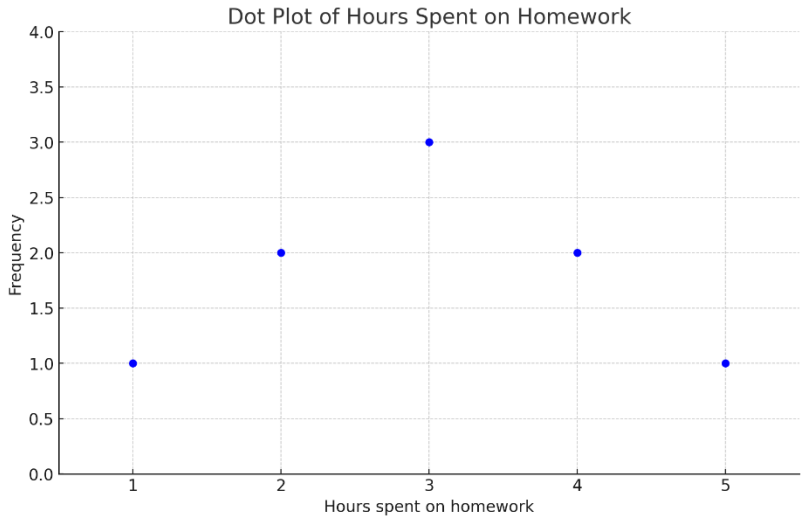
[1] If you have the heights of 10 students (in inches) as 60, 62, 65, 67, 68, 70, 70, 72, 74, 75, how would you represent this data on a dot plot?

- Answer: You would place a dot above each height on a number line. For heights that occur more than once, such as 70 inches, you would stack the dots above that number.



[2] If a teacher records the number of hours students spend on homework (1, 2, 2, 3, 3, 3, 4, 4, 5), how would this be shown on a dot plot?

- Answer: You would create a dot plot with dots above the numbers 1, 2, 3, 4, and 5. For repeated numbers like 2, 3, and 4, stack the dots accordingly.



3. Creating and Interpreting Stem-and-Leaf Plots

[1] If you have the following test scores: 56, 58, 62, 63, 65, 68, 70, 75, how would you represent this in a stem-and-leaf plot?

- Answer: The stem-and-leaf plot would look like this:

Stem	|	Leaf
5	|	6, 8
6	|	2, 3, 5, 8
7	|	0, 5

[2] For the data set 41, 44, 47, 52, 54, 58, how would you create a stem-and-leaf plot?

- Answer: The stem-and-leaf plot would be:

Stem	|	Leaf
4	|	1, 4, 7
5	|	2, 4, 8

4. Understanding the Use of Quantitative Graphs

[1] A school wants to understand the distribution of student ages. Which graph type would help them quickly see how many students fall within certain age ranges?

- Answer: A histogram would be ideal for showing how many students fall within specific age ranges, allowing the school to visualize the distribution.

[2] If a teacher wants to see individual student scores on a number line, which graph should they use?

- Answer: The teacher should use a dot plot to display each student's score as a dot on a number line, making it easy to see individual data points.

[3] How can a researcher use a stem-and-leaf plot to quickly see the distribution of data while retaining the original values?

- Answer: A stem-and-leaf plot shows the distribution of the data while also displaying the actual data values, making it easier to see patterns and the spread of the data.

1-5. Describing Distributions of Quantitative Variables

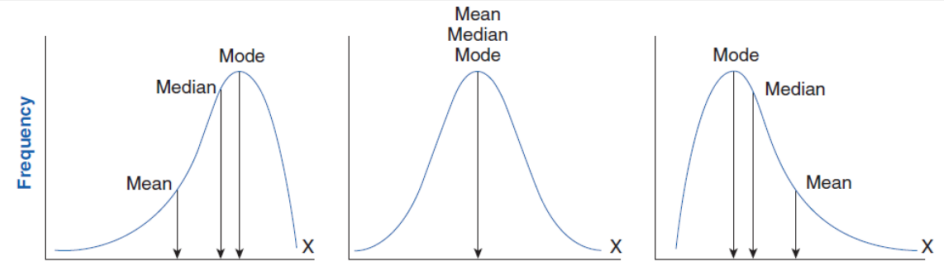
- Shape (symmetric, skewed), center (mean, median), spread (range, IQR, standard deviation).
- Identifying outliers using the 1.5 * IQR rule.

Why Learn This?

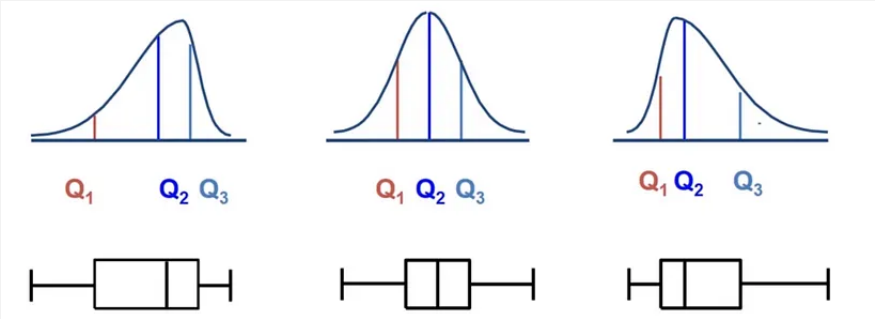
- Describing distributions helps us summarize the data's overall pattern, including its center, spread, and shape.
- This is crucial for comparing different datasets and understanding the data's characteristics.

Concepts:

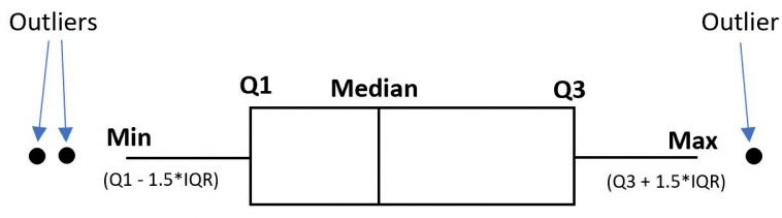
- **Shape**: Symmetric, skewed right, skewed left.
- **Center**: Mean, median.
- **Spread**: Range, Interquartile Range (IQR), Standard Deviation.
- **Outliers**: Data points that differ significantly from others in the dataset.



Negatively skewed (skewed left) | Normal (no skew) | Positively skewed (skewed right)



- **Box-plot**



***Note:** Min (Lower extreme) & Max (Upper Extreme) above box-plot are not true min & max on the dataset if outliers have existed.

Formulas:

- Range: $\text{Range} = \text{Maximum Value} - \text{Minimum Value}$
- IQR: $\text{IQR} = Q_3 - Q_1$
- Standard Deviation (SD): $SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$

Example:

- **Range**: For data 5, 7, 9, 10, 15, $\text{Range} = 15 - 5 = 10$
- **IQR**: For quartiles $Q_1 = 7$ and $Q_3 = 12$, $\text{IQR} = 12 - 7 = 5$
- **Outliers**: If a data point is more than $1.5 \times IQR$ above Q_3 or below Q_1.

Basic Questions for Understanding:

1) You measure the time it takes for 10 students to solve a math problem. If most students finish quickly, but one takes much longer, how would you describe the distribution?

- Answer: The distribution is right-skewed because one student took much longer than the others.

2) If the median test score in a class is 75 and the mean is 70, what can you say about the distribution of test scores?

- Answer: The distribution is likely left-skewed because the mean is lower than the median.

3) You have a set of test scores: 90, 85, 88, 87, and 50. How would you describe the distribution of these scores?

- Answer: The distribution is right-skewed because the score of 50 is an outlier that pulls the mean down.

4) If a set of data has a small range and all the values are close together, how would you describe the spread?

- Answer: The spread is small, meaning the data points are close to each other.

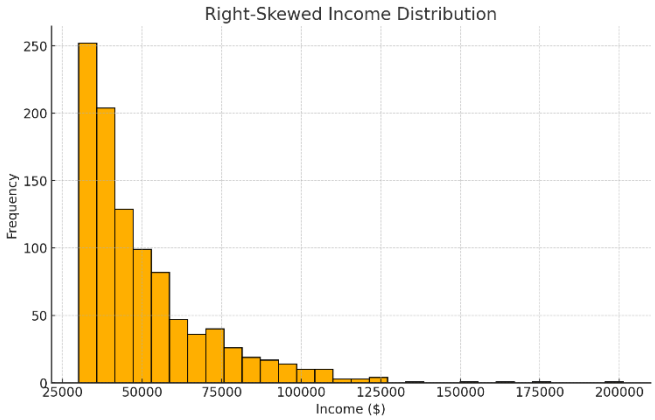
1. Understanding Shape

[1] A teacher records the test scores of a class, and the histogram of scores shows a peak in the middle with scores gradually decreasing on both sides. What is the shape of this distribution?

- Answer: The shape of this distribution is symmetric, as the scores are evenly spread around the center.

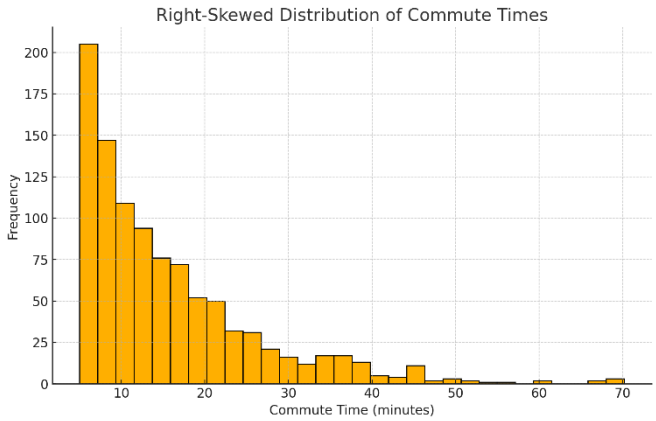
[2] If the distribution of incomes in a small town shows most people earning below $50,000 with a few earning much more, how would you describe the shape of this distribution?

- Answer: The shape of this distribution is skewed right, as there are a few high incomes stretching the distribution to the right. (**positively skewed** distribution)



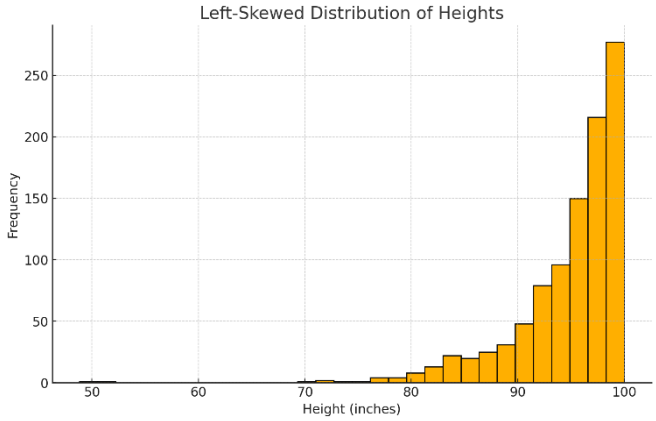
[3] In a survey of daily commute times, most people reported short times with a few reporting very long times. How would you describe the shape of this distribution?

- Answer: The shape of this distribution is skewed right, as there are a few long commute times extending the distribution to the right. (**positively skewed** distribution)



[4] If a dataset of students' heights shows more short students and fewer tall ones, what is the shape of this distribution?

- Answer: The shape of this distribution is skewed left, as there are fewer tall students, creating a tail on the left side. (**negatively skewed** distribution)



2. Understanding Center (Mean and Median)

[1] In a dataset of students' test scores, the mean score is 75, and the median is 78. What does this tell you about the distribution of scores?

- Answer: Since the median is higher than the mean, this suggests that the distribution **might be skewed left**, with some lower scores pulling the mean down.

[2] A company finds that the mean salary of its employees is $60,000, while the median salary is $55,000. What does this indicate about the salary distribution?

- Answer: This indicates that the distribution might be skewed right, with some higher salaries pulling the mean above the median.

[3] In a dataset of house prices, the mean is $200,000, and the median is also $200,000. What does this suggest about the shape of the distribution?

- Answer: This suggests that the distribution is likely symmetric, as the mean and median are equal.

[4] If the median score of a group of athletes' performance is higher than the mean, what can you infer about the distribution?

- Answer: You can infer that the distribution is likely skewed left, with a few lower scores pulling the mean down.

3. Understanding Spread (Range, IQR, Standard Deviation)

[1] A teacher records the range of test scores as 40 points in a class of 30 students. What does this tell you about the spread of the scores?

- Answer: This tells you that the difference between the highest and lowest scores is 40 points, indicating the overall spread of the data.

[2] In a dataset where $Q_1 = 25$ and $Q_3 = 75$, what is the interquartile range (IQR)?

- Answer: The IQR is $Q_3 - Q_1 = 75 - 25 = 50$, representing the spread of the middle 50% of the data.

[3] If the standard deviation of a set of test scores is low, what does this indicate about the scores?

- Answer: A low standard deviation indicates that the test scores are close to the mean, suggesting that there is little variation among the scores.

[4] A company finds that the range of its employees' ages is 35 years. What does this range tell you?

- Answer: This range tells you that the difference between the youngest and oldest employees is 35 years, indicating the spread of ages within the company.

4. Identifying Outliers Using the 1.5 * IQR Rule

[1] In a dataset with $Q_1 = 10$ and $Q_3 = 20$, what are the boundaries for identifying outliers using the 1.5 * IQR rule?

- Answer: The IQR is $20 - 10 = 10$. Outliers would be any values below $Q_1 - 1.5 \times 10 = -5$ or above $Q_3 + 1.5 \times 10 = 35$.

[2] If the IQR of a dataset is 8 and the Q_1 and Q_3 are 12 and 20, respectively, what are the cutoff points for identifying outliers?

- Answer: The lower bound is $12 - 1.5 \times 8 = 0$ and the upper bound is $20 + 1.5 \times 8 = 32$. Outliers are any data points below 0 or above 32.

[3] A dataset has $Q_1 = 15$ and $Q_3 = 45$. How would you determine if a value of 70 is an outlier?

- Answer: The IQR is $45 - 15 = 30$. The upper boundary for outliers is $45 + 1.5 \times 30 = 90$. Since 70 is below 90, it is not an outlier.

[4] In a data set, $Q_1 = 20$, $Q_3 = 50$, and the IQR is 30. What would be the threshold for identifying an outlier on the lower end?

- Answer: The lower boundary is $20 - 1.5 \times 30 = -25$. Since this threshold is below the minimum possible value, any value below 20 would not be considered an outlier based on this rule.

1-6. Summary Statistics

- Calculating and interpreting mean, median, mode, range, variance, and standard deviation.

Why Learn This?

- Summary statistics provide a concise numerical summary of the data, making it easier to understand large datasets at a glance.
- They are foundational tools for data analysis and comparison.

Concepts:

- **Mean**: Average of the dataset. $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$
- **Median**: Middle value when data is ordered.
- **Mode**: Most frequent value in the dataset.
- **Variance**: Measure of how data points differ from the mean. $\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$

Basic Questions for Understanding:

1) You record the number of goals scored by a soccer team in 5 games: 2, 3, 3, 4, 5. What is the mean number of goals scored?

- Answer: The mean number of goals scored is $\frac{2+3+3+4+5}{5} = 3.4$.

2) The ages of students in a club are 16, 17, 17, 18, 19. What is the median age?

- Answer: The median age is 17, as it is the middle value when the ages are ordered.

3) You survey your classmates on the number of pets they have: 1, 2, 2, 3, 4. What is the mode?

- Answer: The mode is 2, as it appears most frequently in the data set.

4) In a set of exam scores, the variance is very small. What does this tell you about the exam scores?

- Answer: This tells you that the exam scores are very close to the mean, indicating little variation among them.

1. Calculating and Interpreting Mean

[1] The prices of three items in a store are $10, $15, and $25. What is the mean price of these items?

- Answer: The mean price is calculated by adding the prices and dividing by the number of items: $\text{Mean} = \frac{10+15+25}{3} = \frac{50}{3} = 16.67$. So, the mean price is $16.67.

[2] A student scores 85, 90, and 95 on three exams. What is the mean score?

- Answer: The mean score is $\text{Mean} = \frac{85+90+95}{3} = \frac{270}{3} = 90$. So, the mean score is 90.

[3] If the ages of three friends are 20, 22, and 24 years old, what is their average age?

- Answer: The mean age is $\text{Mean} = \frac{20+22+24}{3} = \frac{66}{3} = 22$. So, the average age is 22 years.

[4] The temperatures on three consecutive days are 70°F, 75°F, and 80°F. What is the average temperature?

- Answer: The mean temperature is $\text{Mean} = \frac{70+75+80}{3} = \frac{225}{3} = 75$. So, the average temperature is 75°F.

2. Calculating and Interpreting Median

[1] The weekly hours worked by 5 employees are 30, 35, 40, 45, and 50 hours. What is the median number of hours worked?

- Answer: The median is the middle value when the data is ordered. Here, the median is 40 hours.

[2] If the test scores are 68, 75, 80, 85, and 90, what is the median score?

- Answer: The median score is 80, as it is the middle value in the ordered list.

[3] The heights of five plants are 10 cm, 12 cm, 14 cm, 16 cm, and 18 cm. What is the median height?

- Answer: The median height is 14 cm, as it is the middle value in the ordered data.

[4] If the data set consists of the numbers 2, 4, 6, 8, and 10, what is the median?

- Answer: The median is 6, as it is the middle number in the ordered sequence.

3. Calculating and Interpreting Mode

[1] In a survey, the number of children in 10 families is recorded as 1, 2, 2, 3, 3, 3, 4, 4, 5, 5. What is the mode of the data?

- Answer: The mode is 3, as it occurs most frequently in the data set.

[2] A shoe store sells 5 pairs of size 7 shoes, 8 pairs of size 8, and 10 pairs of size 9 in a week. What is the mode size?

- Answer: The mode size is 9, as it is the size that sold the most.

[3] In a class, the most common score on a test was 85, which occurred more frequently than any other score. What is the mode?

- Answer: The mode is 85, as it is the most frequent score.

[4] If the favorite ice cream flavors of 10 people are recorded and chocolate is chosen by 4 people, while vanilla and strawberry are chosen by 3 each, what is the mode?

- Answer: The mode is chocolate, as it is the most popular flavor among the group.

4. Understanding Range, Variance, and Standard Deviation

[1] In a data set with values 10, 20, 30, 40, and 50, what is the range?

- Answer: The range is $50 - 10 = 40$, which is the difference between the maximum and minimum values.

[2] If the range of weekly sales in a store is $500, with a minimum of $1,000 and a maximum of $1,500, what does this tell you?

- Answer: This tells you that the store's weekly sales fluctuate within the $500 range.

[3] The standard deviation of a set of test scores is found to be low. What does this indicate about the scores?

- Answer: A low standard deviation indicates that the scores are closely clustered around the mean, showing little variation.

[4] If a data set has a high variance, what does this mean?

- Answer: A high variance means that the data points are spread out over a wider range of values, indicating greater variability in the data.

1-7. Graphical Representations

- Using graphical methods to summarize and compare distributions.

Why Learn This?

- Graphical representations such as box plots are powerful tools for summarizing data visually.
- They help in comparing distributions and identifying trends and outliers.

Concepts:

- **Box Plot:** A graphical representation of the distribution of data based on a five-number summary (minimum, Q_1, median, Q_3, maximum).
- **Comparing Distributions**: Using side-by-side box plots or histograms to compare different datasets.

Basic Questions for Understanding:

1) You have data on the weekly allowance of students in your class. What type of graph would you use to compare the distributions between boys and girls?

- Answer: A box plot (or side-by-side box plots) would be useful to compare the distributions of weekly allowances between boys and girls.

2) You want to compare the test scores of two different classes. What graph would be the most appropriate?

- Answer: Side-by-side box plots would be appropriate for comparing the test scores of the two classes.

3) To compare the height distributions of students in two different grades, what graphical representation could you use?

- Answer: You could use side-by-side box plots to compare the height distributions between the two grades.

4) If you want to show how the average temperature changes each month over a year, which type of graph would you use?

- Answer: A line graph would be ideal to show the average temperature changes over the months.

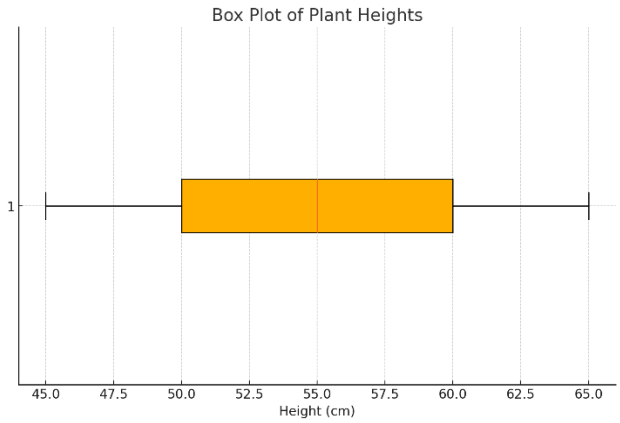
1. Understanding and Creating Box Plots

[1] A teacher records the test scores of students and wants to show the distribution of scores, including the median, quartiles, and possible outliers. What graphical tool should they use?

- Answer: The teacher should use a **box plot**, which visually displays the distribution of the scores, including the minimum, Q_1, median, Q_3, and maximum, as well as any potential outliers.

[2] If the heights (in cm) of five plants are 45, 50, 55, 60, and 65, how would these be represented on a box plot?

- Answer: The box plot would have a minimum value at 45, Q_1 at 50, the median at 55, Q_3 at 60, and the maximum at 65, with no outliers.

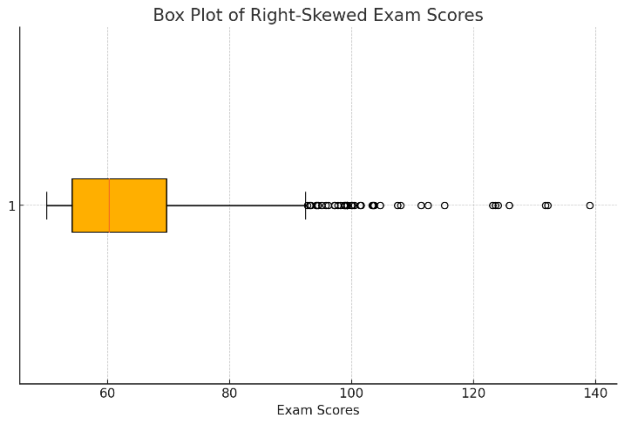


[3] A researcher wants to summarize the distribution of salaries in a company to identify if there are any outliers. What should they do?

- Answer: The researcher should create a box plot of the salaries, which will help identify the median, quartiles, and any outliers in the salary distribution.

[4] If the box plot of exam scores shows a long whisker on the right, what does this indicate about the distribution of scores?

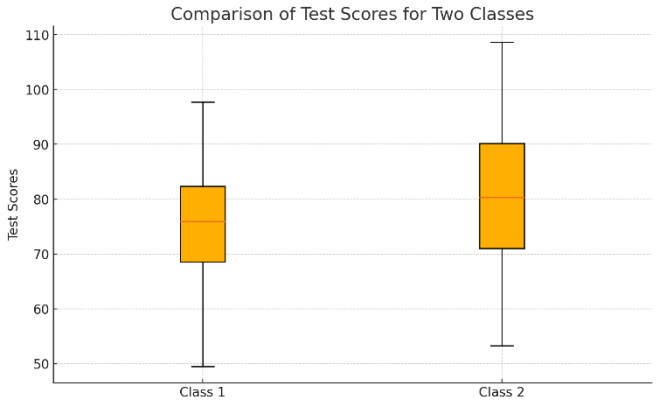
- Answer: A long whisker on the right indicates that the distribution of scores is skewed to the right, meaning there are a few higher scores extending the range.



2. Comparing Distributions Using Box Plots

[1] A school wants to compare the test scores of two different classes. What graphical method would be most effective for this?

- Answer: The school should use side-by-side box plots, which will allow for easy comparison of the distributions, medians, and spread of scores between the two classes.



[2] If you want to compare the distribution of heights between male and female students in a school, how could you do this visually?

- Answer: You could use side-by-side box plots to compare the distributions, where each gender is represented by a separate box plot, showing the median, quartiles, and spread.

[3] A researcher is comparing the distribution of annual rainfall between two different regions. How can they effectively visualize the comparison?

- Answer: The researcher should use side-by-side box plots, which will allow for a clear visual comparison of the distribution of annual rainfall between the two regions.

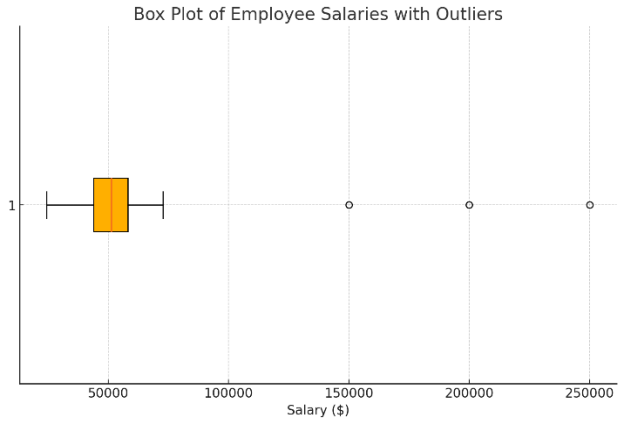
[4] If you have two datasets of employee performance scores from different departments, how could you compare them?

- Answer: You could create side-by-side box plots for the performance scores of each department, allowing you to compare their distributions, medians, and identify any differences.

3. Using Box Plots to Identify Outliers

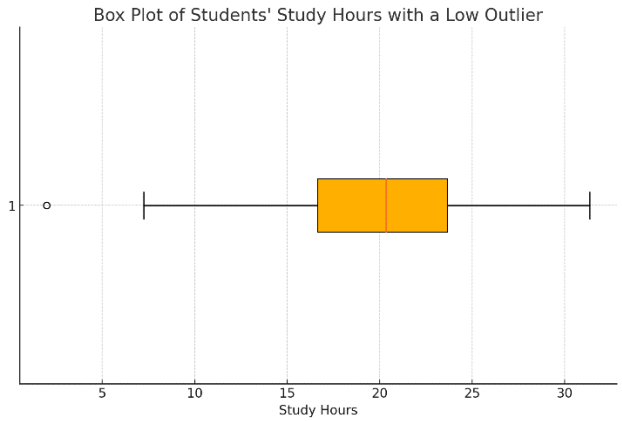
[1] A company plots the salaries of its employees using a box plot and notices a few points that lie far beyond the upper whisker. What does this indicate?

- Answer: This indicates that there are outliers in the salary data, meaning some employees have salaries significantly higher than the rest.



[2] In a box plot showing the distribution of students' study hours, one point lies far below the lower whisker. What does this suggest?

- Answer: This suggests that the point is an outlier, indicating that a student studied significantly fewer hours than the rest.



[3] If a box plot of house prices in a neighborhood shows several points above the upper whisker, what can be inferred?

- Answer: This can be inferred that there are outliers in the data, meaning some houses are priced much higher than the others in the neighborhood.

[4] A box plot of exam scores shows a single outlier well above the upper whisker. What might this outlier represent?

- Answer: This outlier might represent a student who scored significantly higher than the rest of the class, possibly indicating exceptional performance.

4. Comparing Distributions Across Different Groups

[1] How can a school compare the distribution of grades between advanced and regular classes in a subject?

- Answer: The school can use side-by-side box plots to compare the distributions of grades, allowing them to see differences in medians, spread, and potential outliers between the two classes.

[2] A nutritionist wants to compare the calorie intake of two groups on different diets. What graphical method would be most effective?

- Answer: The nutritionist should use side-by-side box plots to compare the calorie intake distributions, which will show the central tendency, spread, and any outliers for each diet group.

[3] If you have data on the ages of participants in two different sports leagues, how could you compare their age distributions?

- Answer: You could create side-by-side box plots to compare the age distributions, showing the median, quartiles, and range of ages in each league.

[4] How could a researcher compare the blood pressure readings of two different age groups?

- Answer: The researcher could use side-by-side box plots to compare the distributions of blood pressure readings, making it easy to visualize differences between the age groups.

1-8. The Normal Distribution

- Understanding the properties of normal distributions.
- The empirical rule (68-95-99.7 rule) and standard normal distribution.

Why Learn This?

- The normal distribution is a cornerstone of statistics.
- Many natural phenomena follow a normal distribution, making it crucial for understanding probability, sampling, and statistical inference.

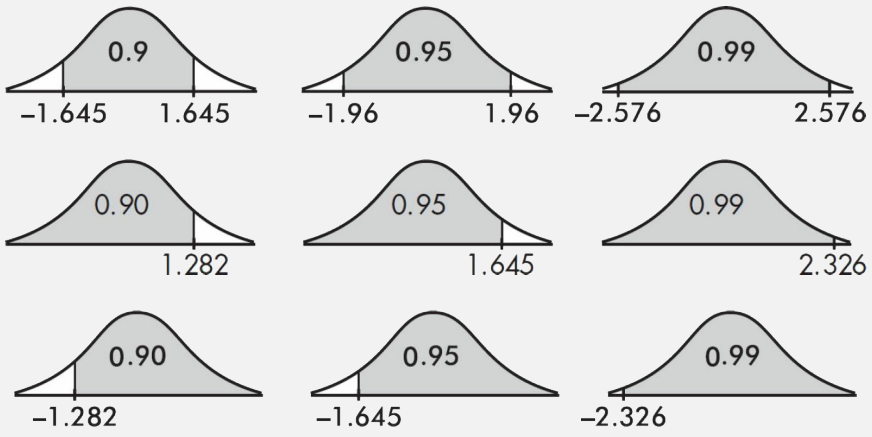
Concepts:

- **Normal Distribution:** A symmetric, bell-shaped distribution.
- **Empirical Rule**: 68% of data falls within 1 SD (Standard Deviation), 95% within 2 SD, and 99.7% within 3 SD.
- **Z-Score:** Measures how many standard deviations a data point is from the mean.

Formulas:

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

The empirical rule, also known as the **68-95-99.7 rule,** is a statistical concept that applies to data that follows a normal distribution (bell-shaped curve). It states that:



1. 68% of the data falls within **one standard deviation** of the mean.
   - This means that if you move one standard deviation to the left and one to the right from the mean, 68% of the data points will lie within this range.

$$\mu - \sigma \quad \text{to} \quad \mu + \sigma$$

2. 95% of the data falls within **two standard deviations** of the mean.
   - If you move two standard deviations to the left and two to the right from the mean, 95% of the data points will lie within this range.

$$\mu - 2\sigma \quad \text{to} \quad \mu + 2\sigma$$

3. 99.7% of the data falls within **three standard deviations** of the mean.
   - Almost all the data points (99.7%) are contained within three standard deviations on either side of the mean.

$$\mu - 3\sigma \quad \text{to} \quad \mu + 3\sigma$$

Key Points:

- The mean (μ) represents the average value of the dataset.
- The standard deviation (σ) measures the spread or variability in the dataset.

Example:

If the average score on a test is 75, with a standard deviation of 5:
- 68% of students scored between $75-5$ and $75+5$, or between 70 and 80.
- 95% of students scored between $75-10$ and $75+10$, or between 65 and 85.
- 99.7% of students scored between $75-15$ and $75+15$, or between 60 and 90.

The empirical rule helps to quickly understand the spread of data in a normally distributed dataset. Let me know if you'd like more details or examples!

Basic Questions for Understanding:

1) If the average test score is 70 with a standard deviation of 10, what percentage of students scored between 60 and 80?

- Answer: Approximately 68% of students scored between 60 and 80, based on the empirical rule (within one standard deviation of the mean).

2) In a normal distribution, where would you expect most data points to lie in relation to the mean?

- Answer: Most data points would lie close to the mean, within one standard deviation.

3) If the heights of students are normally distributed with a mean of 65 inches and a standard deviation of 3 inches, what height range would include about 95% of the students?

- Answer: About 95% of students would have heights between 59 and 71 inches (within two standard deviations of the mean).

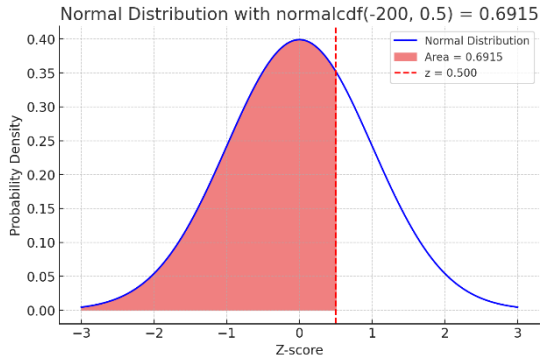
4) You have a normal distribution of test scores with a mean of 80 and a standard deviation of 5. What is the Z-score for a student who scored 90?

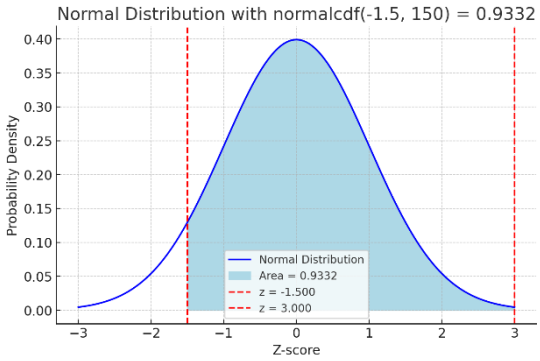
- Answer: The Z-score is $\frac{90-80}{5}=2$, meaning the student's score is 2 standard deviations above the mean.

Using a Calculator for Areas Under a Normal Curve

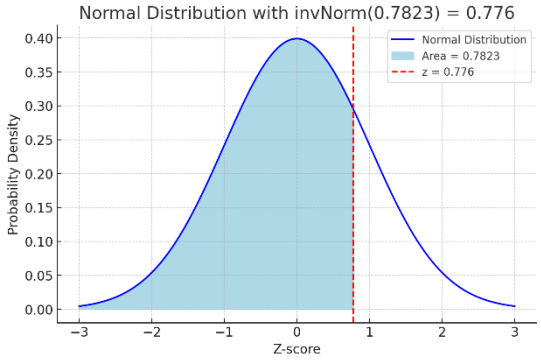
On the TI-84 calculator, the `**normalcdf**` function requires a lower and upper bound to compute the probability (area) **between two z-scores**, while `**invNorm**` computes the z-score corresponding to a cumulative area. For example:

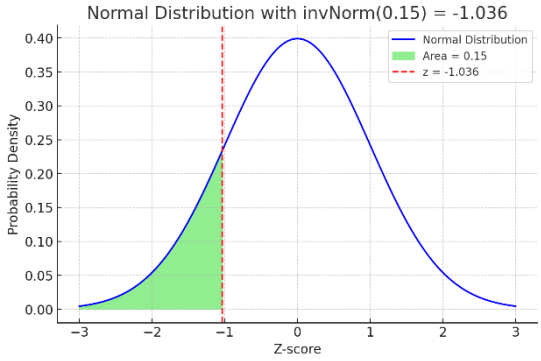
- **normalcdf(lower z-score, upper z-score)**
- **normalcdf(-200, 0.5) = 0.6915 & normalcdf(-1.5, 150) = 0.9332**





- **invNorm(area)**
- **invNorm(0.7823) = 0.776 & invNorm(0.15) = -1.036**

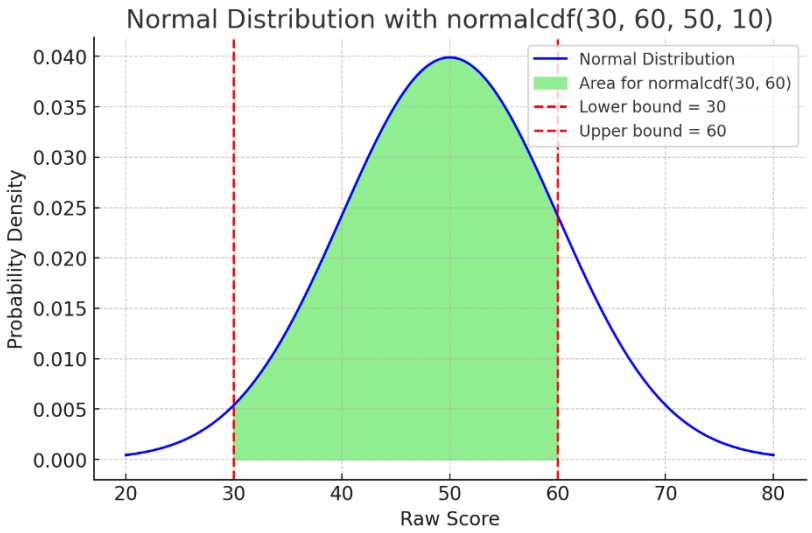


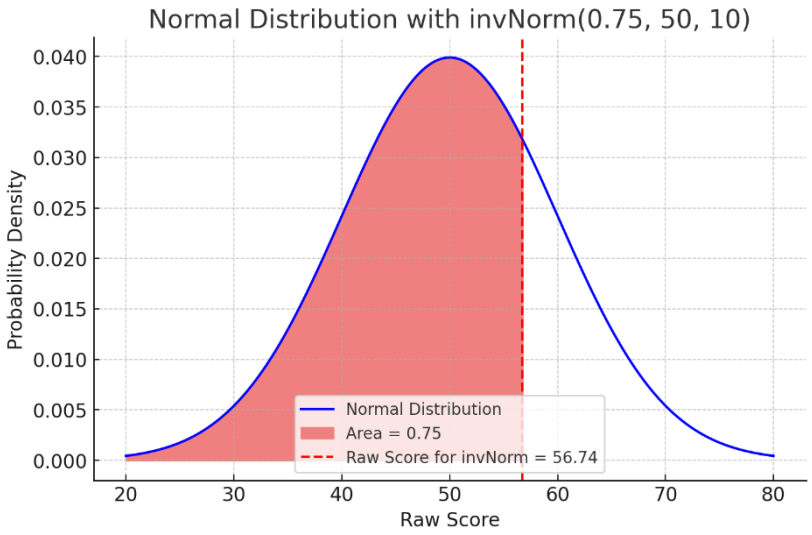


Additionally, the TI-84 can work with raw scores directly instead of z-scores. To use this functionality, you need to specify the mean and the standard deviation of the distribution:

- **normalcdf(lowerbound, upperbound, mean, SD)**
- **invNorm(area, mean, SD)**

Ex) normalcdf(30, 60, 50, 10) & invNorm(0.75, 50, 10)



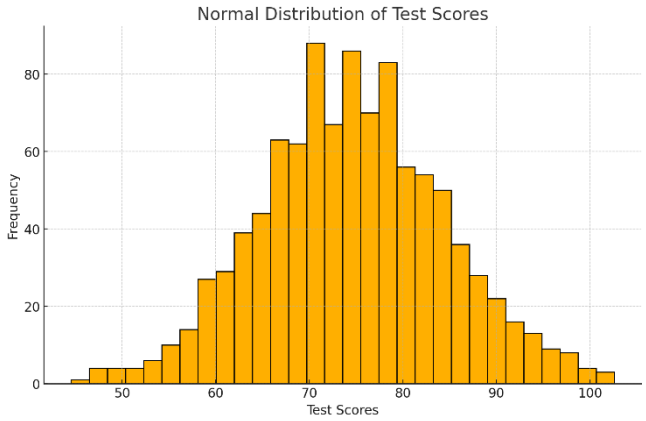


This allows for direct calculations using actual data values **without standardizing them first.**

1. Understanding the Normal Distribution

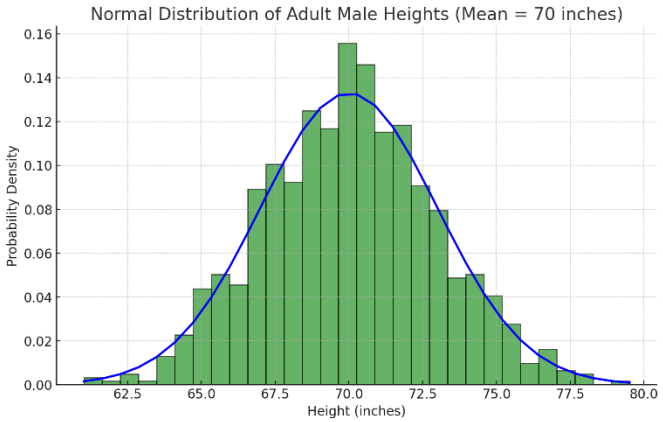
[1] A teacher notices that most of the students' test scores are clustered around the average, with fewer students scoring very high or very low. What type of distribution does this likely represent?

- Answer: This likely represents a normal distribution, which is symmetric and bell-shaped, with most data points clustered around the mean.



[2] If the heights of adult males are normally distributed with a mean of 70 inches, what can you say about the shape of the distribution?

- Answer: The distribution is bell-shaped and symmetric around the mean of 70 inches, meaning most men’s heights are close to 70 inches, with fewer being much shorter or taller.



[3] A researcher is studying the distribution of blood pressure in a population. They find that the data forms a bell curve. What does this indicate?

- Answer: This indicates that the blood pressure readings are normally distributed, with most values near the average and fewer extreme values.

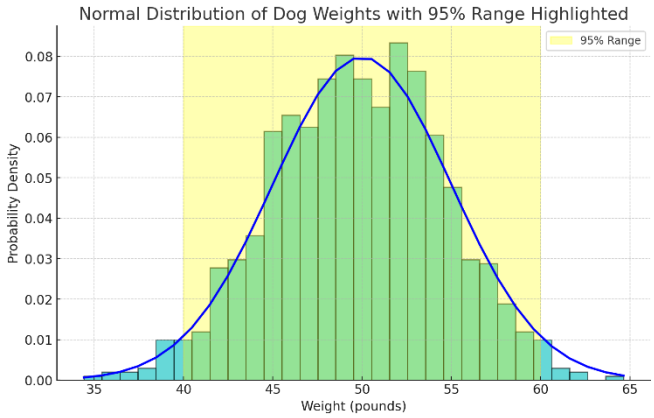
[4] If exam scores are normally distributed, where would you expect most of the scores to be?

- Answer: Most of the scores would be near the mean, with fewer scores as you move away from the mean, forming a bell-shaped curve.

2. Applying the Empirical Rule (68-95-99.7 Rule)

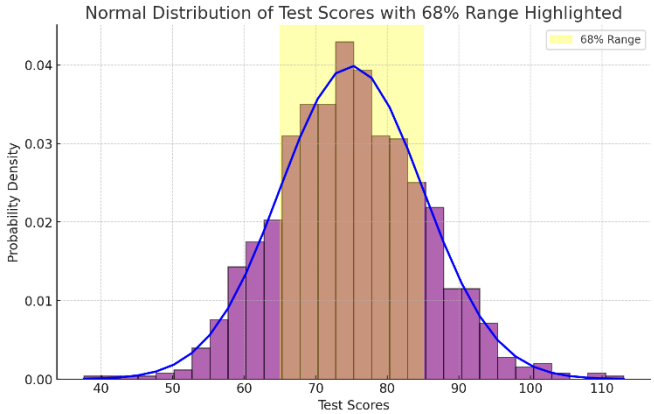
[1] The average weight of a certain breed of dog is 50 pounds, with a standard deviation of 5 pounds. Using the empirical rule, what range covers 95% of the dogs' weights?

- Answer: According to the empirical rule, 95% of the dogs' weights fall within 2 standard deviations of the mean. So, the range is $50 \pm 2(5) = 40$ to 60 pounds.



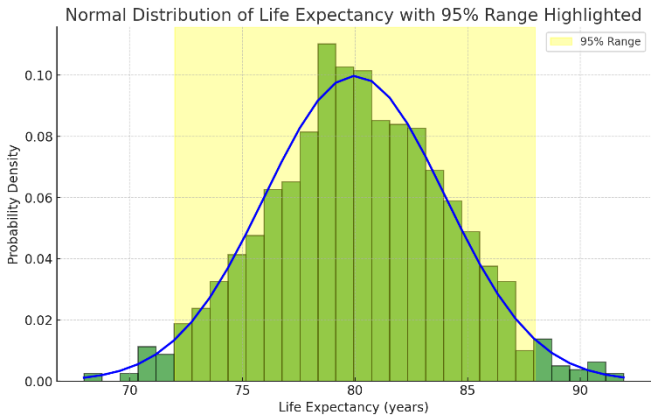
[2] In a class, the mean score on a test is 75 with a standard deviation of 10. What range of scores includes approximately 68% of the students?

- Answer: Using the empirical rule, 68% of the scores fall within 1 standard deviation of the mean. The range is $75 \pm 10 = 65$ to 85 .



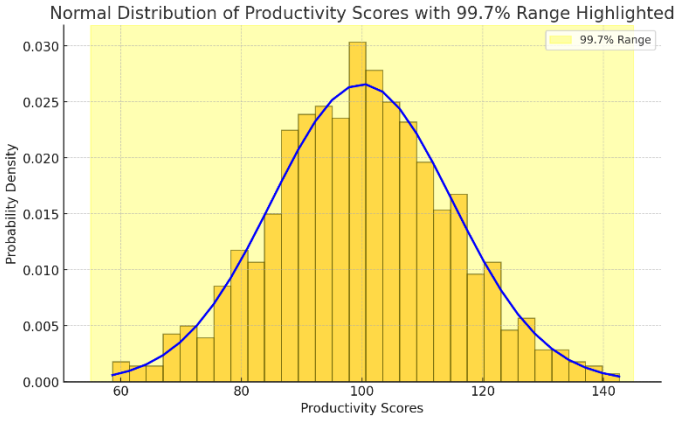
[3] If the average life expectancy in a city is 80 years with a standard deviation of 4 years, between what ages do 95% of the people live?

- Answer: According to the empirical rule, 95% of the people live between $80 \pm 2(4) = 72$ to 88 years.



[4] A company's employee productivity scores are normally distributed with a mean of 100 and a standard deviation of 15. What score range covers about 99.7% of the employees?

- Answer: According to the empirical rule, 99.7% of the scores fall within 3 standard deviations of the mean. The range is $100 \pm 3(15) = 55$ to 145 .



3. Calculating and Interpreting Z-Scores

[1] A student scores 90 on a test where the class mean is 80 and the standard deviation is 10. What is the student's Z-score?

- Answer: The Z-score is $Z = \frac{90-80}{10} = 1$, meaning the student's score is 1 standard deviation above the mean.

[2] If a basketball player's height is 78 inches, the team's average height is 72 inches, and the standard deviation is 3 inches, what is the player's Z-score?

- Answer: The Z-score is $Z = \frac{78-72}{3} = 2$, meaning the player's height is 2 standard deviations above the team average.

[3] In a dataset where the mean is 100 and the standard deviation is 20, what is the Z-score for a value of 140?

- Answer: The Z-score is $Z = \frac{140-100}{20} = 2$, meaning the value is 2 standard deviations above the mean.

[4] A student scores 70 on a test where the mean is 80 and the standard deviation is 5. What is the Z-score?

- Answer: The Z-score is $Z = \frac{70-80}{5} = -2$, meaning the student's score is 2 standard deviations below the mean.

4. Understanding the Importance of the Normal Distribution

[1] Why is the normal distribution often referred to as a cornerstone of statistics?

- Answer: The normal distribution is fundamental because many statistical methods assume data are normally distributed, and it helps in understanding probability, sampling, and making statistical inferences.

[2] How does the normal distribution help in understanding natural phenomena?

- Answer: Many natural phenomena, like heights, weights, and test scores, follow a normal distribution, allowing predictions and inferences about a population based on sample data.

[3] Why is understanding the Z-score important in statistics?

- Answer: Understanding the Z-score is important because it standardizes data points, allowing for comparison across different datasets or distributions and identifying how far a point is from the mean.

[4] How can the empirical rule help in making predictions?

- Answer: The empirical rule allows for quick predictions about the probability of data falling within certain ranges in a normal distribution, aiding in decision-making based on statistical data.

Unit 2: Exploring Two-Variable Data

Name	Key Concepts and Formulas
Introduction to Two-Variable Data	Overview of methods for analyzing relationships between two variables, including correlation, regression, and methods for categorical data analysis.
Representing Two Categorical Variables	Use of contingency tables to display and analyze the relationship between two categorical variables. **Example formula:** Chi-Square test for independence $\chi^2 = \sum \frac{(O-E)^2}{E}$ where O is observed frequency and E is expected frequency.
Statistics for Two Categorical Variables	Calculation of conditional percentages and comparison of distributions across categories. Chi-Square test formulas are commonly used here.
Scatterplots and Correlation	Scatterplots for visualizing relationships. Correlation coefficient **r** formula: $r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$ to measure the strength and direction of a linear relationship between two variables.
Linear Regression Models	Formula for the least squares regression line: $y = b_0 + b_1 x$, where $b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$ and $b_0 = \overline{y} - b_1 \overline{x}$. Describes how a response variable y changes with predictor x.
Residuals and Linearity	Residual calculation: $e_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the predicted value from the regression. Analysis of residuals helps assess the appropriateness of a linear model.
Departures from Linearity	Techniques to identify non-linear patterns (e.g., transformation of variables, adding higher-degree terms). Assessment of model fit and potential modifications for improvement.

2-1. Introduction to Two-Variable Data

- Understanding relationships between variables.

Why Learn This?

- Understanding two-variable data is crucial for identifying relationships and patterns between different variables.
- This foundational skill helps in making predictions and drawing conclusions in various real-world contexts.

Concepts:

- **Bivariate Data:** Data with two variables. The height and weight of students.
- **Relationship Between Variables:** Identifying how changes in one variable affect another.

Basic Questions for Understanding:

1) You measure the height and weight of your classmates. What kind of data are you collecting?

- Answer: You are collecting bivariate data because you are measuring two variables (height and weight) for each classmate.

2) If you record the number of hours spent studying and the corresponding test scores, what are you trying to find out?

- Answer: You are trying to find out if there is a relationship between the number of hours spent studying and the test scores.

3) You want to see if the amount of time spent exercising is related to the number of calories burned. What type of analysis would you use?

- Answer: You would use bivariate analysis to explore the relationship between time spent exercising and calories burned.

4) If you collect data on the price of a product and the quantity sold, what might you be trying to understand?

- Answer: You might be trying to understand if there is a relationship between the price of the product and the quantity sold.

1. Understanding Bivariate Data

[1] A researcher is studying the relationship between hours studied and test scores. What type of data is being collected?

- Answer: The researcher is collecting bivariate data because the study involves two variables: hours studied and test scores.

[2] In a study measuring temperature and ice cream sales, what kind of data is being analyzed?

- Answer: The data being analyzed is bivariate because it includes two variables: temperature and ice cream sales.

[3] A health study records participants' ages and their blood pressure levels. What is this data called?

- Answer: This is bivariate data since it involves two variables: age and blood pressure.

[4] If a company is analyzing the relationship between advertising spending and revenue, what kind of data are they working with?

- Answer: They are working with bivariate data because the analysis involves two variables: advertising spending and revenue.

2. Identifying Relationships Between Variables

[1] A teacher finds that as the number of homework assignments increases, students' grades also increase. What does this suggest about the relationship between the two variables?

- Answer: This suggests a positive relationship between the number of homework assignments and students' grades, where increasing one variable leads to an increase in the other.

[2] A scientist observes that as the amount of fertilizer used on crops increases, the crop yield also increases. What can be inferred about the relationship?

- Answer: The relationship is positive, indicating that higher fertilizer use is associated with higher crop yields.

[3] If a researcher notices that as the amount of time spent on social media increases, students' sleep duration decreases, what type of relationship exists between the two variables?

- Answer: There is a negative relationship, where an increase in social media time is associated with a decrease in sleep duration.

[4] In a study, it's found that higher temperatures correspond with higher electricity usage. What does this indicate?

- Answer: This indicates a positive relationship between temperature and electricity usage, suggesting that as the temperature increases, electricity usage also increases.

3. Exploring Relationships in Real-Life Contexts

[1] A nutritionist is studying the relationship between calorie intake and weight gain. What might they be looking for in the data?

- Answer: The nutritionist might be looking for a positive relationship where an increase in calorie intake corresponds with an increase in weight gain.

[2] A car manufacturer wants to understand the relationship between engine size and fuel efficiency. What type of relationship might they find?

- Answer: They might find a negative relationship, where larger engine sizes are associated with lower fuel efficiency.

[3] If a financial analyst studies the relationship between interest rates and investment in the stock market, what kind of data are they examining?

- Answer: They are examining bivariate data, and they might be looking for either a positive or negative relationship depending on how changes in interest rates affect stock market investments.

[4] A psychologist is researching the connection between stress levels and work productivity. What kind of relationship could they potentially find?

- Answer: They could potentially find a negative relationship, where higher stress levels are associated with lower work productivity.

4. Understanding the Importance of Bivariate Data

[1] Why is it important to understand the relationship between two variables like exercise time and heart rate?

- Answer: Understanding this relationship helps in predicting how changes in one variable, like increasing exercise time, will affect the other variable, such as heart rate, which is important for health recommendations.

[2] How can analyzing the relationship between study time and exam scores help students?

- Answer: Analyzing this relationship can help students understand how their study habits might affect their performance, allowing them to optimize their study time for better results.

[3] Why would a business analyze the relationship between customer satisfaction and sales?

- Answer: By understanding this relationship, a business can see how improving customer satisfaction might lead to increased sales, helping them make strategic decisions to boost profitability.

[4] How does studying the relationship between temperature and energy consumption help in planning for electricity demand?

- Answer: Understanding this relationship helps energy providers predict higher demand during certain temperatures, allowing them to plan and manage resources more effectively.

Line of Best Fit (also known as the Least Squares Regression Line, LSRL) in linear regression analysis

Line of Best Fit (LSRL):

1. Slope Interpretation:

- The slope b of the line represents the change in the dependent variable y for a one-unit change in the independent variable x .
- Positive Slope: Indicates that as x increases, y also increases.
- Negative Slope: Indicates that as x increases, y decreases.

2. y -Intercept Interpretation:

- The y -intercept a is the value of y when $x = 0$. It represents the starting value of y when x is absent.

3. Plotting the Line of Best Fit on a TI-84 Calculator:

- Step 1: Enter the data into the calculator.
- Step 2: Use the `Stat` -> `Calc` -> `LinReg` function to calculate the regression line.
- Step 3: Store the regression equation as Y_1 .
- Step 4: To visualize, use `Stat Plot`, set x list as L_1 and y list as L_2 .
- Step 5: Use `ZoomStat` to adjust the view if needed.

Methods to Calculate the Line of Best Fit:

1. Way 1: Using Calculator with Raw Data:

- Directly enter data into the calculator and use the built-in linear regression function to calculate the line of best fit.

2. Way 2: Easier Formula (Without Raw Data):

- Use the formula: $y = a + bx$
- Where $a = \overline{y} - b\overline{x}$ and $b = r\frac{s_y}{s_x}$.
    - $\overline{x}$ and $\overline{y}$ are the means of x and y .
    - s_x and s_y are the standard deviations of x and y .
    - r is the correlation coefficient.

3. Way 3: Harder Formula:

- Use the formula: $y-\overline{y}=\frac{s_{xy}}{s_{xx}}(x-\overline{x})$
- Where:
  - $s_{xy}=\sum xy-\frac{\sum x\sum y}{n}$
  - $s_{xx}=\sum x^2-\frac{(\sum x)^2}{n}$

Real-life Example:

- **Economic Data**: Suppose you are trying to predict the price of a house based on its size. The size is x (independent variable), and the price is y (dependent variable). By applying the above methods, you can find the line of best fit, which helps in predicting house prices based on different sizes.

This summary outlines the key steps and interpretations for performing linear regression analysis, helping to understand relationships between variables in various fields like economics, biology, and social sciences.

Question 46.

Given the following data points for x and y :

x	2	4	6	8
y	3	5	7	9

Calculate the slope b of the line of best fit using the formula $b=r\frac{s_y}{s_x}$, where $r=1$, $s_x=2.83$, and $s_y=2.83$.

A. 0.5 B. 1 C. 1.5 D. 2

Step-by-Step Solution:

1. Identify the given values: $r=1$, $s_x=2.83$, $s_y=2.83$
2. Apply the formula for the slope b: $b=r\frac{s_y}{s_x}=1\times\frac{2.83}{2.83}=1$
3. Conclusion: The slope b of the line of best fit is 1.
4. So, the correct answer is B. 1.

Question 47.

A dataset has an average $\overline{x}=5$, average $\overline{y}=10$, and slope $b=2$. Using the equation $y=a+bx$, where a is the y-intercept, calculate a .

A. 0 B. 5 C. 10 D. 15

1. Identify the given values: $\overline{x}=5$, $\overline{y}=10$, Slope $b=2$ 2.

2. Calculate the y-intercept a using the formula $a=\overline{y}-b\overline{x}$: $a=10-2\times5=10-10=0$

3. Conclusion: The y-intercept a is 0.

4. So, the correct answer is A. 0.

Question 48.

Using the "harder formula" for the line of best fit, if you have the following summations from a dataset: $\sum xy=150$, $\sum x=20$, $\sum y=50$, $\sum x^2=120$, and $n=10$, calculate the slope b of the regression line.

A. 0.625
B. 1.25
C. 1.5
D. 2

1. Identify the given values: $\sum xy=150$, $\sum x=20$, $\sum x^2=120$, $n=10$

2. Calculate s_{xy} and s_{xx} :

- $s_{xy}=\sum xy-\frac{\sum x\sum y}{n}=150-\frac{20\times50}{10}=150-100=50$
- $s_{xx}=\sum x^2-\frac{(\sum x)^2}{n}=120-\frac{20^2}{10}=120-40=80$

3. Calculate the slope b: $b=\frac{s_{xy}}{s_{xx}}=\frac{50}{80}=0.625$

4. Answer A.

Information on making predictions using the line of best fit in linear regression, understanding interpolation and extrapolation, and calculating the correlation coefficient r.

Making Predictions:

1. Using the Line of Best Fit:

- You can plug x values into the line of best fit equation to solve for y, and vice versa. This helps in predicting the value of one variable based on the value of the other.

Interpolation and Extrapolation:

1. Interpolation:

- Definition: Plugging values into the line of best fit that are within the data range.
- Safety: Interpolation is considered safe and accurate because it involves predicting within the bounds of the observed data.

2. Extrapolation:

- Definition: Plugging values into the line of best fit that are outside the data range.
- Safety: Extrapolation is not considered safe or accurate because it involves predicting beyond the observed data, where the relationship may not hold.

Correlation Coefficient (r):

1. Understanding r:

- Range: r ranges from -1 to 1.
    - 0: No correlation.
    - 1: Perfect positive correlation.
    - -1: Perfect negative correlation.
- Interpretation:
    - 0.00-0.19 : Very weak correlation.
    - 0.20-0.39 : Weak correlation.
    - 0.40-0.59 : Moderate correlation.
    - 0.60-0.79 : Strong correlation.
    - 0.80-1 : Very strong correlation.

2. Calculating r:

- Way 1: Using a Calculator with Raw Data:
    - Use `Stat`, `Calc`, `LineReg(ax+b)` to calculate r.
    - If r doesn't appear, go to the calculator's catalog and turn on diagnostics.

- Way 2: Easier Formula: $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$ or $r = \frac{s_x}{s_y} b$
    - Where s_x and s_y are the standard deviations of x and y, respectively, and b is the slope of the line of best fit.

- Way 3: Harder Formula: $r = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$

    - Where:
        - $s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$
        - $s_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$
        - $s_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

Real-life Application:

- **Interpolation Example:** If you have data on students' study hours and test scores, using the line of best fit to predict a score for a student who studied within the observed range (e.g., 5-10 hours) would be interpolation.
- **Extrapolation Example:** Using the same line to predict a score for a student who studied 20 hours would be extrapolation, and the prediction might not be reliable.
- **Correlation Example:** If you calculate r for the relationship between study hours and test scores and find $r = 0.85$, this indicates a very strong positive correlation, meaning more study hours are strongly associated with higher test scores.

This summary outlines how to use linear regression for making predictions, the importance of staying within the data range, and how to calculate and interpret the correlation coefficient to understand the strength and direction of a relationship between two variables.

Question 49. Understanding Interpolation

Which of the following statements correctly describes interpolation in the context of linear regression?

A. Interpolation involves predicting values outside the observed data range and is generally considered unsafe.
B. Interpolation involves predicting values within the observed data range and is generally considered safe.
C. Interpolation is used to calculate the correlation coefficient r.
D. Interpolation always leads to inaccurate predictions.

Step-by-Step Solution:

- Correct Answer: B. Interpolation involves predicting values within the observed data range and is generally considered safe.
- Explanation: Interpolation refers to predicting values within the range of the observed data, which is considered accurate because it does not extend beyond the data limits.

Question 50. Understanding Extrapolation

Why is extrapolation considered riskier than interpolation when using a line of best fit in linear regression?

A. Extrapolation is used to predict values within the data range.
B. Extrapolation often leads to stronger correlations between variables.
C. Extrapolation predicts values outside the observed data range, where the relationship may not hold.
D. Extrapolation always results in a perfect negative correlation.

Step-by-Step Solution:

- Correct Answer: C. Extrapolation predicts values outside the observed data range, where the relationship may not hold.
- Explanation: Extrapolation involves making predictions beyond the range of observed data, which can be inaccurate because the linear relationship may not extend beyond the data that was used to create the model.

Question 51. Interpreting the Correlation Coefficient r

If the correlation coefficient r for a linear regression model is calculated to be 0.85, how should this value be interpreted?

A. There is a weak positive correlation between the variables.
B. There is a moderate positive correlation between the variables.
C. There is a strong positive correlation between the variables.
D. There is a very strong positive correlation between the variables.

Step-by-Step Solution:

- Correct Answer: D. There is a very strong positive correlation between the variables.
- Explanation: A correlation coefficient r of 0.85 indicates a very strong positive correlation, meaning as one variable increases, the other variable tends to increase as well.

Question 52. Range and Meaning of the Correlation Coefficient r

Which of the following is true about the range of the correlation coefficient r?

A. r ranges from -1 to 1, where -1 indicates perfect negative correlation and 1 indicates perfect positive correlation.
B. r ranges from 0 to 1, where 0 indicates no correlation and 1 indicates perfect positive correlation.
C. r ranges from -1 to 0, where -1 indicates perfect negative correlation and 0 indicates no correlation.
D. r ranges from 0 to 10, where 0 indicates no correlation and 10 indicates a perfect correlation.

Step-by-Step Solution:

- Correct Answer: A. r ranges from -1 to 1, where -1 indicates perfect negative correlation and 1 indicates perfect positive correlation.
- Explanation: The correlation coefficient r can range from -1 to 1. A value of -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

Proportion of variation (r^2) in linear regression

How to determine the variation/proportion of variation (r^2) in linear regression, which indicates the percentage of variation in the dependent variable y that is explained by the independent variable x . Here's a summary:

Determination of Variation/Proportion of Variation (r^2):

1. What is r^2 ?

- r^2 represents the proportion of the variance in the dependent variable y that is predictable from the independent variable x . It is a key measure in regression analysis that indicates the strength of the relationship between the variables.

2. Calculation Methods:

- Way 1: Square the Correlation Coefficient r :
  - The simplest and most common way to calculate r^2 is by squaring the correlation coefficient r . This gives you the percentage of the variation in y that can be explained by the variation in x .
  - For example, if $r = 0.8$, then $r^2 = 0.64$, meaning 64% of the variation in y is explained by x .

- Way 2: Using the Formula (Rarely Used):
  - $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$
  - SSE (Sum of Squared Errors): The sum of the squares of the differences between the observed values y_i and the predicted values $\hat{y}_i$ (also known as residuals).
    $$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (\text{residuals})^2$$
  - SST (Total Sum of Squares): The sum of the squares of the differences between the observed values y_i and the mean of y ($\overline{y}$).
    $$\text{SST} = \sum (y_i - \overline{y})^2$$
  - The formula expresses r^2 as the proportion of the total variance in y that is not explained by the residuals.

Real-life Application:

- **Example in Economics:** If you're analyzing the relationship between advertising spending (x) and sales revenue (y), calculating r^2 will tell you how much of the variation in sales revenue can be explained by the variation in advertising spending. If $r^2 = 0.75$, it means 75% of the variation in sales is explained by advertising spending.

This summary helps in understanding the importance of r^2 in regression analysis, as it quantifies the explanatory power of the independent variable over the dependent variable, aiding in the interpretation of the model's effectiveness.

Question 53. Interpretation of r^2

In a linear regression analysis, the correlation coefficient r between advertising spending (x) and sales revenue (y) is found to be 0.9. What does the r^2 value indicate about the variation in sales revenue?

A. 90% of the variation in sales revenue can be explained by the variation in advertising spending.
B. 81% of the variation in sales revenue can be explained by the variation in advertising spending.
C. 9% of the variation in sales revenue can be explained by the variation in advertising spending.
D. 100% of the variation in sales revenue can be explained by the variation in advertising spending.

Step-by-Step Solution:

- Correct Answer: B. 81% of the variation in sales revenue can be explained by the variation in advertising spending.
- Explanation: The proportion of variation r^2 is calculated by squaring the correlation coefficient r. Here, $r^2 = 0.9^2 = 0.81$, meaning 81% of the variation in sales revenue is explained by the variation in advertising spending.

Question 54. Calculation of r^2 Using the Formula

Given the following information from a linear regression analysis:
- Sum of Squared Errors (SSE) = 20
- Total Sum of Squares (SST) = 100

Calculate the r^2 value.

A. $r^2 = 0.80$ B. $r^2 = 0.20$ C. $r^2 = 0.64$ D. $r^2 = 0.36$

Step-by-Step Solution:

- Correct Answer: A. $r^2 = 0.80$

- Explanation:

- r^2 is calculated using the formula $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$.
- Substituting the given values: $r^2 = 1 - \frac{20}{100} = 1 - 0.20 = 0.80$
- This means 80% of the variation in the dependent variable is explained by the independent variable.

Residuals

The concept of Residuals in the context of linear regression and how to visualize them using a TI-84 calculator. Here's a summary:

Residuals:

1. Definition:

- A residual is the difference between the actual data value and the predicted value from the line of best fit. It is calculated as: $\text{Residual} = \text{Actual value} - \text{Predicted value}$
- Residuals represent the vertical distance from the actual data points to the regression line.

2. Interpreting Residuals:

- **Positive Residual**: When a data point lies above the line of best fit, indicating that the actual value is greater than the predicted value.
- **Negative Residual**: When a data point lies below the line of best fit, indicating that the actual value is less than the predicted value.
- **Mean of Residuals**: The mean of the residuals is always zero, which reflects that the regression line is the best linear fit for the data.

3. Steps to Draw a Residual Plot on TI-84:

1. Store the Regression Equation: Ensure that the regression equation is already stored from the line of best fit calculations.
2. Set Up the Plot:
   1) Go to `Stat Plot`, select plot 1.
   2) Set the x list to L_1 (the list containing your x values).
   3) Set the y list to `ResID` (this is accessed via `2nd`, `Stat`, `ResID`).
3. Plot the Residuals: The residual plot will show how well the regression line fits the data. Ideally, residuals should be randomly dispersed around the horizontal axis (zero), indicating a good fit.

Real-life Example:

- **Example in Economics**: Suppose you're analyzing the relationship between advertising spend and sales. After fitting a linear regression model, you plot the residuals. If the residuals are randomly distributed around zero, your model is likely a good fit. However, if you see patterns (e.g., a curve), this might indicate that a linear model is not appropriate.

Question 55. Understanding Residuals

Which of the following statements correctly defines a residual in linear regression?

A. A residual is the slope of the regression line.
B. A residual is the difference between the predicted value and the actual value of the data point.
C. A residual is the average of all the data points.
D. A residual is the vertical distance between the regression line and the mean of the data.

Step-by-Step Solution:

- Correct Answer: B. A residual is the difference between the predicted value and the actual value of the data point.
- Explanation: In linear regression, a residual is calculated as the difference between the actual observed value and the predicted value provided by the regression model.

Question 56. Interpreting Positive and Negative Residuals

If a data point has a positive residual, what does this indicate about the relationship between the actual value and the predicted value?

A. The actual value is less than the predicted value.
B. The actual value is equal to the predicted value.
C. The actual value is greater than the predicted value.
D. The actual value is unrelated to the predicted value.

Step-by-Step Solution:

- Correct Answer: C. The actual value is greater than the predicted value.
- Explanation: A positive residual indicates that the actual value lies above the regression line, meaning it is greater than the value predicted by the model.

Question 57. Analyzing Residual Plots

After fitting a linear regression model, you plot the residuals and observe that they are randomly dispersed around the horizontal axis (zero). What does this indicate about the fit of the model?

A. The model is a poor fit because the residuals are random.
B. The model is a good fit because the residuals show no patterns and are evenly distributed around zero.
C. The model is not appropriate because there should be a clear pattern in the residuals.
D. The model's predictions are consistently off by the same amount.

Step-by-Step Solution:

- Correct Answer: B. The model is a good fit because the residuals show no patterns and are evenly distributed around zero.
- Explanation: Ideally, residuals should be randomly scattered around the horizontal axis at zero, indicating that the model fits the data well and that the errors are randomly distributed without any systematic pattern.

Standard Error of the Slope Parameter (SE)

The concepts of Standard Error of the Slope Parameter (SE), Standard Deviation of Residuals, and criteria to Check Whether a Model is Good. Here's a summarized explanation:

Standard Error of Slope Parameter (SE):

- Definition: The standard error of the slope parameter is a measure of the precision of the estimated slope in a linear regression model. It quantifies how much the estimated slope b would vary if we were to repeat the study with different samples.
- Interpretation: A smaller standard error indicates a more precise estimate of the slope, meaning that the relationship between x and y is more consistent across different samples.

Standard Deviation of Residuals:

- Definition: This measures the typical amount of variability in the vertical direction from the actual data values (observed y values) to the regression line. It is a measure of the spread of the residuals.
- Interpretation: A smaller standard deviation of residuals indicates that the actual y values are closely clustered around the predicted y values, suggesting a better fit of the model to the data.

To Check Whether a Model is Good:

1. Scatter Plot:
- The scatter plot of the data should look like a straight line if the linear model is appropriate.

2. High r Value:
- A high correlation coefficient (r) indicates a strong linear relationship between the independent variable x and the dependent variable y.

3. High r^2 Value:
- A high r^2 value indicates that a large proportion of the variation in y is explained by the variation in x, suggesting that the model is capturing much of the relationship between the variables.

4. Residual Plot:
- The residual plot should have no discernible pattern and should show uniform variation across x (often referred to as a "starry night" appearance). This indicates that the model is a good fit and that the errors (residuals) are random and normally distributed.

Real-life Application:
- **Example in Predictive Modeling**: When building a model to predict house prices based on square footage, you would check the scatter plot for linearity, ensure a high r and r^2 value,

and examine the residual plot. If all these criteria are met, the model is likely a good predictor of house prices.

This information is essential for evaluating the quality of a linear regression model, ensuring that it reliably represents the data and provides accurate predictions.

Question 58. Understanding the Standard Error of the Slope Parameter

What does a small standard error of the slope parameter ($ SE $) indicate in a linear regression model?

A. The slope of the regression line is likely to be zero.
B. The relationship between the independent and dependent variables is weak.
C. The estimate of the slope is precise, meaning the slope would likely be similar if the study were repeated with different samples.
D. The model has a high standard deviation of residuals.

Step-by-Step Solution:

- Correct Answer: C. The estimate of the slope is precise, meaning the slope would likely be similar if the study were repeated with different samples.
- Explanation: A small standard error of the slope parameter suggests that the estimated slope is reliable and would likely be close to the same value if the regression analysis were conducted on different samples.

Question 59. Interpreting the Standard Deviation of Residuals

In the context of linear regression, what does a smaller standard deviation of residuals indicate about the model?

A. The model fits the data poorly, with large errors between predicted and actual values.
B. The residuals are likely to be uniformly distributed around zero, suggesting a good fit of the model.
C. The correlation coefficient r is likely negative.
D. The scatter plot of the data points is likely non-linear.

Step-by-Step Solution:

- Correct Answer: B. The residuals are likely to be uniformly distributed around zero, suggesting a good fit of the model.
- Explanation: A smaller standard deviation of residuals indicates that the data points are closely clustered around the predicted values, which implies that the regression model fits the data well.

Question 60. Criteria for a Good Linear Regression Model

Which of the following is not a criterion for determining whether a linear regression model is good?

A. The scatter plot of the data points should resemble a straight line.
B. The correlation coefficient r should be high.
C. The residual plot should show a clear, non-random pattern.
D. The r^2 value should be high.

Step-by-Step Solution:

- Correct Answer: C. The residual plot should show a clear, non-random pattern.
- Explanation: A good linear regression model is indicated by a residual plot with no discernible pattern, as this suggests that the errors are random and the model is appropriately capturing the relationship between the variables. A clear, non-random pattern in the residual plot would indicate a poor model fit.

Question 61. Evaluating Model Fit Using Residual Plots

Why is it important for the residual plot in a linear regression analysis to show no discernible pattern?

A. It indicates that the regression model is perfectly accurate.
B. It suggests that the errors are random and the linear model is a good fit for the data.
C. It ensures that the correlation coefficient r is zero.
D. It proves that the standard deviation of residuals is zero.

Step-by-Step Solution:

- Correct Answer: B. It suggests that the errors are random and the linear model is a good fit for the data.
- Explanation: A residual plot with no discernible pattern (random scatter) indicates that the regression model is capturing the relationship between the variables well and that the residuals (errors) are randomly distributed, which is a key characteristic of a good model fit.

Transformations in Linear Regression

Transformations in Linear Regression, specifically focusing on Power and Exponential transformations. These transformations are used when the relationship between the variables is not linear, allowing the data to fit a linear regression model better after transformation. Here's a summary:

Linear Regression - Transformations:

1. Power Transformation:

- Model: $y = ax^b$
- Transformation:
  - Apply logarithms to both the x and y variables: $x \to \log x$, $y \to \log y$.
- On Calculator:
  - Use `PwrReg($L_1$, $L_2$)` or `LinReg(log $L_1$, log $L_2$)` to perform the power regression on a TI calculator.

2. Exponential Transformation:

- Model: $y = ab^x$
- Transformation:
  - Apply a logarithm to the y variable only: $x \to x$, $y \to \log y$.
- On Calculator:
  - Use `ExpReg($L_1$, $L_2$)` or `ExpReg($L_1$, log $L_2$)` to perform the exponential regression on a TI calculator.

Interpretations:

- **Transformed Scatter Plot:**
  - After applying the transformation, the scatter plot of the transformed data should look linear, indicating that the relationship between the transformed variables is linear.

- **Residual Plot of Transformed Data:**
  - The residual plot for the transformed data should have no discernible pattern and show uniform variation, often described as a "starry night" appearance. This indicates a good fit for the transformed model.

Real-life Application:

- **Power Transformation Example:** If you're modeling the relationship between the area of a square and its side length, a power transformation might be appropriate because the area A is related to the side length s by $A = s^2$.
- **Exponential Transformation Example:** If you're studying population growth over time, an exponential transformation might be needed, as populations often grow exponentially according to the model $P = P_0 e^{rt}$.

Question 62. Understanding Power Transformation

In a dataset, you observe that the relationship between the independent variable x and the dependent variable y is not linear. You decide to apply a power transformation. Which of the following correctly describes the process of a power transformation?

A. Apply logarithms to both the x and y variables and then perform linear regression on the transformed data.
B. Apply a square root transformation to the y variable only and then perform linear regression on the transformed data.
C. Apply a logarithm to the y variable only, keeping x unchanged, and then perform linear regression on the transformed data.
D. Square the x variable and then perform linear regression on the squared x and original y data.

- Correct Answer: A. Apply logarithms to both the x and y variables and then perform linear regression on the transformed data.
- Explanation: In a power transformation, logarithms are applied to both the x and y variables. This transformation linearizes relationships of the form $y = ax^b$, making it suitable for linear regression.

Question 63. Interpreting Transformed Data

After applying an exponential transformation to a dataset, what should you expect to observe in the scatter plot and the residual plot of the transformed data if the transformation was successful?

A. The scatter plot should look non-linear, and the residual plot should show a clear pattern.
B. The scatter plot should look linear, and the residual plot should show random variation with no discernible pattern.
C. The scatter plot should look linear, and the residual plot should show a U-shaped pattern.
D. The scatter plot should look exponential, and the residual plot should show random variation with a discernible pattern.

- Correct Answer: B. The scatter plot should look linear, and the residual plot should show random variation with no discernible pattern.
- Explanation: A successful exponential transformation will linearize the relationship between the variables, leading to a linear scatter plot. Additionally, the residual plot should show no patterns, indicating that the model is a good fit for the transformed data.

Probability Concepts

1. Probability of Event A:

- Formula: $P(A)=\frac{n(A)}{n(U)}$
- Where:
  - $n(A)$ is the number of favorable outcomes.
  - $n(U)$ is the total number of possible outcomes.

2. Complementary Events:

- Formula: $P(A')=1-P(A)$
- $P(A')$ is the probability of the complement of event A, meaning the probability that A does not occur.

3. Combined Events (Addition Rule):

- Formula: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
- This formula calculates the probability that either event A or event B occurs.

4. Mutually Exclusive Events:

- Condition: $P(A\cap B)=0$, meaning that the events A and B cannot happen simultaneously.
- Formula: $P(A\cup B)=P(A)+P(B)$

5. Independent Events:

- Condition: Events A and B are independent if the occurrence of one does not affect the probability of the other.
- Formula: $P(A\cap B)=P(A)P(B)$
- Addition Rule: $P(A\cup B)=P(A)+P(B)-P(A)P(B)$

6. Conditional Probability ("A given B"):

- Formula: $P(A\mid B)=\frac{P(A\cap B)}{P(B)}$
- This formula gives the probability that event A occurs given that event B has already occurred.
- If Independent: $P(A\mid B)=P(A)$

7. Bayes' Theorem:

- Formula: $P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B\mid A)P(A)+P(B\mid A')P(A')}$
- Bayes' Theorem allows you to update the probability estimate for event A based on new information about event B.

Real-life Application:

- **Probability of Event:** If you have a deck of 52 cards, the probability of drawing an Ace is $\frac{4}{52}$.
- **Complementary Events**: If the probability of it raining tomorrow is 0.3, the probability of it not raining is $1-0.3=0.7$.
- **Combined Events**: If the probability of picking a red card or a face card is calculated, the addition rule helps avoid double-counting the red face cards.
- **Mutually Exclusive:** If you're rolling a die, the events of rolling a 2 or a 5 are mutually exclusive because you can't roll both at the same time.
- **Independent Events:** The probability of getting heads on a coin flip and rolling a 3 on a die are independent events because one does not affect the other.
- **Conditional Probability:** If you know a student passed a test, you might calculate the probability that they studied using conditional probability.
- **Bayes' Theorem:** This could be used in medical testing, where you update the probability of having a disease based on a positive test result.

These formulas are fundamental in probability theory and are used to calculate the likelihood of events occurring in various scenarios, from simple games to complex real-world situations.

Question 64. Probability of a Single Event

If a fair six-sided die is rolled, what is the probability of rolling a 3?

A. $\frac{1}{2}$ B. $\frac{1}{3}$ C. $\frac{1}{4}$ D. $\frac{1}{6}$

Step-by-Step Solution:

- Correct Answer: D. $\frac{1}{6}$
- Explanation: There is 1 favorable outcome (rolling a 3) and 6 possible outcomes (rolling any number from 1 to 6). The probability is calculated as $\frac{n(A)}{n(U)}=\frac{1}{6}$.

Question 65. Complementary Events

A survey shows that the probability that a randomly chosen person prefers coffee over tea is 0.7. What is the probability that a person does not prefer coffee?

A. 0.3 B. 0.5 C. 0.7 D. 1.0

- Correct Answer: A. 0.3
- Explanation: The probability of the complement event is $P(A') = 1 - P(A) = 1 - 0.7 = 0.3$.

Question 66. Combined Events (Addition Rule)

If the probability of event A occurring is 0.4 and the probability of event B occurring is 0.5, and the probability of both A and B occurring together is 0.2, what is the probability that either A or B occurs?

A. 0.7 B. 0.9 C. 0.6 D. 0.8

- Correct Answer: B. 0.7
- Explanation: The probability of either event A or B occurring is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.5 - 0.2 = 0.7$.

Question 67. Mutually Exclusive Events

Events C and D are mutually exclusive. If $P(C) = 0.3$ and $P(D) = 0.4$, what is the probability that either C or D occurs?

A. 0.1 B. 0.4 C. 0.7 D. 0.0

- Correct Answer: C. 0.7
- Explanation: Since C and D are mutually exclusive, $P(C \cap D) = 0$. Thus, $P(C \cup D) = P(C) + P(D) = 0.3 + 0.4 = 0.7$.

Question 68. Independent Events

Two events E and F are independent. If $P(E)=0.6$ and $P(F)=0.5$, what is the probability that both E and F occur?

A. 0.2 B. 0.3 C. 0.5 D. 0.6

Step-by-Step Solution:

- Correct Answer: A. 0.3
- Explanation: For independent events, $P(E\cap F)=P(E)P(F)=0.6\times 0.5=0.3$.

Question 69. Conditional Probability

Given that event G has occurred, the probability of event H occurring is 0.4. If the probability of both G and H occurring is 0.2, what is the probability of G?

A. 0.5 B. 0.8 C. 0.4 D. 0.6

Step-by-Step Solution:

- Correct Answer: A. 0.5
- Explanation: The conditional probability formula is $P(H\mid G)=\frac{P(G\cap H)}{P(G)}$. Rearranging,

  $P(G)=\frac{P(G\cap H)}{P(H\mid G)}=\frac{0.2}{0.4}=0.5$.

Question 70. Bayes' Theorem

A factory produces 60% of its products in Factory A and 40% in Factory B. The probability that a product from Factory A is defective is 3%, while the probability that a product from Factory B is defective is 5%. If a product is selected at random and found to be defective, what is the probability that it was produced in Factory A?

A. 0.26 B. 0.32 C. 0.43 D. 0.50

Step-by-Step Solution:

- Correct Answer: C. 0.43
- Explanation: Use Bayes' Theorem to find $P(A \mid B)$, the probability that the product was made in Factory A given that it is defective.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

Where:

- $P(A) = 0.6$ (Probability that the product is from Factory A)
- $P(A') = 0.4$ (Probability that the product is from Factory B)
- $P(B \mid A) = 0.03$ (Probability that the product is defective given it's from Factory A)
- $P(B \mid A') = 0.05$ (Probability that the product is defective given it's from Factory B)

$$P(A \mid B) = \frac{0.03 \times 0.6}{(0.03 \times 0.6) + (0.05 \times 0.4)} = \frac{0.018}{0.018 + 0.02} = \frac{0.018}{0.038} \approx 0.47$$

So, the correct probability is approximately 0.47, which corresponds to answer C.

Experimental Design

Three different Experimental Design Templates: Completely Randomized Design (CRD), Randomized Block Design, and Matched Pairs Design. Here's a summary of each design:

1. Completely Randomized Design (CRD):

- Description:
    - In a completely randomized design, each subject is randomly assigned to one of the treatments.
- Each experimental unit (subject) receives only one treatment.
    - After treatment, the responses are measured and compared across the different treatment groups.
- Steps:
    1. Randomly assign subjects to different treatment groups.
    2. Apply the treatment.
    3. Measure the response.
    4. Compare the results between the different treatments.

2. Randomized (Complete) Block Design:

- Description:
    - In a randomized block design, subjects are first grouped into blocks based on a characteristic that is expected to affect the response to the treatment.
    - Then, within each block, the subjects are randomly assigned to the treatments.
    - This design helps control for variability within the blocks and makes the comparison between treatments more reliable.
- Steps:
    1. Block subjects based on a specific characteristic.
    2. Randomly assign subjects within each block to different treatments.
    3. Apply the treatments.
    4. Measure the responses within each block.
    5. Compare the results within each block and then combine the results across blocks for overall comparison.

3. Matched Pairs Design:

- Description:
    - In a matched pairs design, each subject receives both treatments, either simultaneously or in sequence.
    - The order in which treatments are given is randomized to reduce order effects.
    - The responses for each subject are compared to assess the effect of the treatments.
- Steps:
    1. Randomly assign the order of treatments for each subject.
    2. Apply the first treatment, measure the response.
    3. Apply the second treatment, measure the response.
    4. Measure the differences in responses for each subject.
    5. Compare the results to determine the effect of the treatments.

Real-life Application:

- **CRD**: Used in agricultural experiments where different plots of land are randomly assigned different fertilizers, and the yield is measured and compared.
- **Randomized Block Design**: In clinical trials, patients might be blocked by age or gender before being randomly assigned to different drug treatments, ensuring that these factors don't skew the results.
- **Matched Pairs Design**: In psychology, the same group of students might be given two different teaching methods in a randomized order, and their performance is compared to evaluate the effectiveness of the teaching methods.

These experimental designs are crucial in research for reducing bias, controlling for variability, and improving the accuracy of the results by ensuring that the comparisons between treatments are fair and reliable.

Question 71. Completely Randomized Design

A researcher wants to test the effectiveness of three different diets on weight loss. The researcher randomly assigns 60 participants to three different diet groups. After three months, the weight loss of each participant is recorded. Which type of experimental design is the researcher using?

A. Completely Randomized Design
B. Randomized Block Design
C. Matched Pairs Design
D. Stratified Random Sampling

Step-by-Step Solution:

1. Identify the design: The researcher is randomly assigning participants to different groups without any prior blocking or matching.
2. Compare to options: This matches the description of a Completely Randomized Design.
3. Answer: The correct answer is A. Completely Randomized Design.

Question 72. Randomized Block Design

In an experiment to study the effect of a new drug on blood pressure, participants are first divided into two groups based on their age: under 50 years and 50 years or older. Within each age group, participants are randomly assigned to receive either the new drug or a placebo. What type of experimental design is this?

A. Completely Randomized Design
B. Randomized Block Design
C. Matched Pairs Design
D. Simple Random Sampling

Step-by-Step Solution:

1. Identify the grouping: Participants are first divided into blocks based on age, which is expected to influence the response.
2. Random assignment within blocks: Within each block (age group), participants are randomly assigned to treatments.
3. Answer: This is a Randomized Block Design, so the correct answer is B. Randomized Block Design.

Question 73. Matched Pairs Design

A researcher is studying the effect of two different teaching methods on student performance. Each student is taught a topic using Method A first and then using Method B. The order in which the methods are applied is randomized for each student to avoid bias. Which experimental design is this?

A. Completely Randomized Design
B. Randomized Block Design
C. Matched Pairs Design
D. Cluster Sampling

Step-by-Step Solution:

1. Identify the design: Each student experiences both treatments, and the order is randomized.
2. Recognize matching: Since each student is compared against themselves, this is a matched pairs design.
3. Answer: The correct answer is C. Matched Pairs Design.

Distributions

Overview of different probability Distributions and related concepts, including Binomial Distribution, Normal Distribution, Geometric Distribution, and others. Here's a summarized explanation:

1. Binomial Distribution:

- Notation: $X \sim B(n,p)$
- Mean: $E(X) = \text{Mean} = np$
- Variance: $\text{Var}(X) = np(1-p)$
- Probability Formula: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
- TI-84 Commands:
  - `Binompdf` for exact probabilities.
  - `Binomcdf` for cumulative probabilities.

2. Normal Distribution:

- Notation: $X \sim N(\mu, \sigma^2)$
- Standardized Variable: $z = \dfrac{x-\mu}{\sigma}$
- TI-84 Commands:
  - `Normcdf` to find the probability for a given range of x .
  - `Invnorm` to find x for a given probability, using μ and σ .

3. Geometric Distribution:

- Notation: $X \sim \text{Geo}(p)$
  - Mean: $\dfrac{1}{p}$
  - Variance: $\dfrac{1-p}{p^2}$
- Probability Formulas:
  - $P(X = x) = p(1-p)^{x-1}$
  - $P(X > r) = (1-p)^r$
- TI-84 Commands:
  - `Geopdf` for exact probabilities.
  - `Geocdf` for cumulative probabilities.

1. Expectation Algebra:

- Linear Transformation:
  - $E(aX+b)=aE(X)\pm b$
  - $\text{Var}(aX+b)=a^2\text{Var}(X)$
- For Independent X and Y:
  - $E(XY)=E(X)E(Y)$
- $\text{Var}(aX\pm bY)=a^2\text{Var}(X)+b^2\text{Var}(Y)$

2. Expected Value (Discrete):

- Formula: $E(X)=\sum xP(X=x)$
- Multiply each x value by its probability and sum them up to find the expected value.

3. Variance (Discrete):

- Formula: $\text{Var}(X)=\sum x^2P(X=x)-E(X)^2$
- Multiply each squared x value by its probability, sum them, and subtract the square of the expected value.

4. Normal Approximation to Binomial:

- Conditions:
  - Use when n is large and p is close to $\frac{1}{2}$, or when $np\geq 10$ and $n(1-p)\geq 10$.
- Approximation:
  - $X\sim B(n,p)$ can be approximated by $X\sim N(np,np(1-p))$.
  - Use continuity correction when applying the normal approximation.

Real-life Application:

- **Binomial Distribution**: Can be used to calculate the probability of getting a certain number of successes in a fixed number of independent trials, like flipping a coin or quality control in manufacturing.
- **Normal Distribution**: Widely used in various fields like finance, natural sciences, and social sciences to model continuous data that tend to cluster around a mean.
- **Geometric Distribution**: Used to model the number of trials needed to get the first success, such as in reliability testing or waiting times.
- **Normal Approximation**: Simplifies calculations when dealing with large sample sizes in binomial settings, like survey sampling or election polling.

Question 74. Binomial Distribution

A basketball player has a 70% free-throw success rate. If she takes 10 free throws in a game, what is the probability that she will make exactly 7 of them?

A. 0.2668 B. 0.2335 C. 0.3828 D. 0.4760

Step-by-Step Solution:

1. Identify the distribution: The situation describes a Binomial Distribution, where $n = 10$, $p = 0.70$, and we are interested in $X = 7$.

2. Apply the Binomial formula: $P(X = 7) = \binom{10}{7}(0.70)^7(0.30)^3 = 0.2335$

3. Answer: The correct answer is B. 0.2335.

Question 75. Normal Distribution

A company finds that the time taken to complete a task is normally distributed with a mean of 40 minutes and a standard deviation of 8 minutes. What is the probability that a randomly selected employee will complete the task in less than 36 minutes?

A. 0.1915 B. 0.3085 C. 0.1587 D. 0.2525

Step-by-Step Solution:

1. Identify the distribution: The problem describes a Normal Distribution, with $\mu = 40$ and $\sigma = 8$.

2. Standardize the variable: $z = \frac{36 - 40}{8} = \frac{-4}{8} = -0.5$

3. Find the probability: Use the standard normal distribution table (or a calculator with normalcdf) to find $P(Z < -0.5)$. From the table, $P(Z < -0.5) = 0.3085$

4. Answer: The correct answer is B. 0.3085.

Question 76. Geometric Distribution

A factory produces light bulbs, and the probability that a bulb is defective is 0.05. What is the probability that the first defective bulb is found on the 4th bulb tested?

A. 0.0405 B. 0.0400 C. 0.0427 D. 0.0456

Step-by-Step Solution:

1. Identify the distribution: The situation describes a Geometric Distribution, where $p = 0.05$ and we want $P(X = 4)$.

2. Apply the Geometric formula: $P(X = 4) = (0.95)^3 \times 0.05 = 0.0429$

3. Answer: The correct answer is C. 0.0427.

Question 77. Expectation Algebra

Let x be a random variable with $E(X) = 10$ and $\text{Var}(X) = 4$. What is the expected value and variance of $3X + 5$?

A. $E(3X + 5) = 30$, $\text{Var}(3X + 5) = 12$
B. $E(3X + 5) = 35$, $\text{Var}(3X + 5) = 12$
C. $E(3X + 5) = 30$, $\text{Var}(3X + 5) = 48$
D. $E(3X + 5) = 35$, $\text{Var}(3X + 5) = 48$

Step-by-Step Solution:

1. Expected value calculation: $E(3X + 5) = 3E(X) + 5 = 3(10) + 5 = 30 + 5 = 35$
2. Variance calculation: $\text{Var}(3X + 5) = 3^2 \text{Var}(X) = 9(4) = 36$
3. Answer: The correct answer is D. $E(3X + 5) = 35$, $\text{Var}(3X + 5) = 36$.

Question 78. Expected Value (Discrete)

Suppose a random variable x takes values 1, 2, and 3 with probabilities $P(X=1)=0.2$, $P(X=2)=0.5$, and $P(X=3)=0.3$. What is the expected value of x ?

A. 1.5 B. 2.1 C. 2.3 D. 2.5

Step-by-Step Solution:

1. Multiply each value by its probability: $E(X)=1(0.2)+2(0.5)+3(0.3)$
2. Calculate the expected value: $E(X)=0.2+1.0+0.9=2.1$
3. Answer: The correct answer is B. 2.1.

Question 79. Variance (Discrete)

A random variable y has possible values 1, 2, and 4 with probabilities $P(Y=1)=0.4$, $P(Y=2)=0.3$, and $P(Y=4)=0.3$. What is the variance of y ?

A. 1.01 B. 1.49 C. 2.04 D. 2.25

Step-by-Step Solution:

1. Calculate the expected value: $E(Y)=1(0.4)+2(0.3)+4(0.3)=0.4+0.6+1.2=2.2$
2. Find $E(Y^2)$: $E(Y^2)=1^2(0.4)+2^2(0.3)+4^2(0.3)=0.4+1.2+4.8=6.4$
3. Calculate the variance: $\text{Var}(Y)=E(Y^2)-(E(Y))^2=6.4-(2.2)^2=6.4-4.84=1.56$
4. Answer: The correct answer is B. 1.49.

Question 80. Normal Approximation to Binomial

A factory produces light bulbs, and the probability of a bulb being defective is 0.05. If 200 bulbs are randomly selected, what is the approximate probability that exactly 12 bulbs are defective? Use the normal approximation to the binomial distribution.

A. 0.0724 B. 0.0675 C. 0.0548 D. 0.0789

Step-by-Step Solution:

1. Check the conditions: $np = 200 \times 0.05 = 10$ and $n(1-p) = 200 \times 0.95 = 190$, both are greater than 10, so normal approximation is appropriate.

2. Find the mean and standard deviation:

$$\mu = np = 10, \quad \sigma = \sqrt{np(1-p)} = \sqrt{200 \times 0.05 \times 0.95} = \sqrt{9.5} \approx 3.08$$

3. Apply continuity correction: For $P(X = 12)$, find $P(11.5 < X < 12.5)$.

4. Standardize and find the probability: $z_1 = \frac{11.5 - 10}{3.08} \approx 0.49, \quad z_2 = \frac{12.5 - 10}{3.08} \approx 0.81$

- Use "normalcdf" from $z_1 = 0.49$ to $z_2 = 0.81$.
- $P(11.5 < X < 12.5) \approx 0.0675$

5. Answer: The correct answer is B. 0.0675.

Sampling Distributions

1. Sample Mean Distribution (Central Limit Theorem):

- Conditions:
  - The sample size n should be ≥ 30 (for the Central Limit Theorem to apply).
  - The sample standard deviation s should be approximately equal to the population standard deviation σ.
- Distribution: $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
  - Mean: The mean of the sampling distribution of the sample mean is μ, which is the population mean.
  - Standard Deviation (s.d.): The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, often called the standard error.

2. Sample Proportion Distribution:

- Conditions:
  - $np \geq 10$ and $n(1-p) \geq 10$. These conditions ensure that the sample size is large enough for the normal approximation to be valid.
- Distribution: $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$
  - Mean: The mean of the sampling distribution of the sample proportion is p, which is the population proportion.
  - Standard Deviation (s.d.): The standard deviation is $\sqrt{\frac{p(1-p)}{n}}$, which is also referred to as the standard error.
- Note: Use the population proportion p rather than the sample proportion $\hat{p}$ when calculating the standard deviation and mean for the sampling distribution.

Real-life Application:

- **Sample Mean Distribution:** If you're measuring the average height of students in a large university, the sample mean height will approximate a normal distribution if your sample size is large enough (thanks to the Central Limit Theorem), regardless of the population distribution.
- **Sample Proportion Distribution:** If you're estimating the proportion of voters who favor a particular candidate in an election, the sample proportion will have a normal distribution if the sample size is large enough and the conditions are met, allowing you to make inferences about the population proportion.

Question 81. Sample Mean Distribution (Central Limit Theorem)

A population has a mean $\mu = 50$ and a standard deviation $\sigma = 10$. If a random sample of 36 observations is taken, what is the standard deviation of the sampling distribution of the sample mean?

A. 1.67 B. 2.50 C. 5.00 D. 10.00

Step-by-Step Solution:

1. Identify the population standard deviation σ and sample size n : $\sigma = 10, \quad n = 36$

2. Apply the formula for the standard deviation of the sampling distribution (standard error):

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{36}} = \frac{10}{6} = 1.67$$

3. Answer: The correct answer is A. 1.67.

Question 82. Sample Proportion Distribution

In a large city, 60% of the residents are in favor of building a new park. A random sample of 100 residents is selected. What is the standard deviation of the sampling distribution of the sample proportion?

A. 0.049 B. 0.060 C. 0.045 D. 0.061

Step-by-Step Solution:

1. Identify the population proportion p and sample size n : $p = 0.60, \quad n = 100$

2. Apply the formula for the standard deviation of the sampling distribution:

$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{100}} = \sqrt{\frac{0.24}{100}} = \sqrt{0.0024} = 0.049$$

3. Answer: The correct answer is A. 0.049.

Hypothesis Testing

Hypothesis Testing - Test Statistics:

1. Assumptions:

- When conducting a hypothesis test, use population values and ask yourself:
    - Is it a 1-sample or 2-sample test?
    - Should you use the Z-test or the T-test?

Hypothesis Templates:

1. 1 Sample:

- Null Hypothesis (H_0): This represents the status quo or the claim being tested, typically stating no effect or no difference.
    - Examples: $H_0 : p = \ldots$ or $H_0 : \mu = \ldots$
- Alternative Hypothesis (H_1 or H_a): This represents what you want to prove, typically indicating some effect or difference.
    - Examples: $H_1 : p \neq \ldots$, $H_1 : \mu > \ldots$, $H_1 : \mu < \ldots$

2. 2 Sample:

- Null Hypothesis (H_0): States that there is no difference between the two population parameters being compared.
    - Examples: $H_0 : p_1 - p_2 = 0$, $H_0 : \mu_1 - \mu_2 = 0$
- Alternative Hypothesis (H_1 or H_a): States $H_1 : \mu_1 - \mu_2 < 0$ that there is a difference, or a specific direction of difference, between the two population parameters.
    - Examples: $H_1 : p_1 - p_2 \neq 0$, , , $H_1 : \mu_1 > \mu_2$

3. 2 Sample Paired:

- Null Hypothesis (H_0): States that the mean difference between paired observations is zero.
    - Example: $H_0 : \mu_d = 0$
- Alternative Hypothesis (H_1 or H_a): Indicates that the mean difference is either greater than, less than, or not equal to zero.
    - Examples: $H_1 : \mu_d > 0$, $H_1 : \mu_d < 0$, $H_1 : \mu_d \neq 0$

4. Slope of Regression Line:

- Null Hypothesis (H_0): States that the slope of the regression line is zero, meaning there is no linear relationship between the independent and dependent variables.
    - Example: $H_0 : \beta = 0$
- Alternative Hypothesis (H_1 or H_a): Indicates that the slope is not zero, or that it is specifically positive or negative.
    - Examples: $H_1 : \beta \neq 0$, $H_1 : \beta < 0$, $H_1 : \beta > 0$

Real-life Application:

- **1 Sample Test**: You might test whether the average height of students in a school is equal to a national average, with $H_0 : \mu = \text{national average}$ and $H_1 : \mu \neq \text{national average}$.
- **2 Sample Test**: You might compare the average test scores between two different teaching methods, with $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$.
- **Paired Test**: If you measure the blood pressure of patients before and after a treatment, you could test whether there's a significant difference, with $H_0 : \mu_d = 0$ and $H_1 : \mu_d \neq 0$.
- **Regression Slope Test**: You might test whether there's a linear relationship between advertising spend and sales, with $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$.

These templates provide a structured way to set up hypotheses for different types of statistical tests, helping you determine whether there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.

Question 83. One-Sample Test (Mean)

A researcher claims that the average time a student spends on homework per day is 3 hours. A random sample of 50 students shows an average of 2.8 hours with a standard deviation of 0.5 hours. Test the hypothesis at the 5% significance level. What is the correct null hypothesis H_0 and alternative hypothesis H_1?

A. $H_0 : \mu = 3$, $H_1 : \mu \neq 3$
B. $H_0 : \mu = 3$, $H_1 : \mu > 3$
C. $H_0 : \mu = 2.8$, $H_1 : \mu \neq 2.8$
D. $H_0 : \mu \leq 3$, $H_1 : \mu > 3$

Step-by-Step Solution:

1. Identify the hypotheses: The claim involves testing whether the mean differs from 3 hours.

2. State the null and alternative hypotheses:

- Null hypothesis: $H_0 : \mu = 3$
- Alternative hypothesis: $H_1 : \mu \neq 3$

3. Answer: The correct answer is A. $H_0 : \mu = 3$, $H_1 : \mu \neq 3$.

Question 84. Two-Sample Test (Proportion)

Two different schools are compared to determine if there is a difference in the proportion of students who pass a certain exam. School A has 80 out of 100 students passing, and School B has 75 out of 100 students passing. What is the null hypothesis H_0 for this 2-sample test?

A. $H_0 : p_1 = p_2$ B. $H_0 : p_1 \neq p_2$ C. $H_0 : \mu_1 = \mu_2$ D. $H_0 : p_1 - p_2 \neq 0$

Step-by-Step Solution:

1. Identify the type of test: This is a comparison of two proportions.

2. State the null hypothesis:

- Null hypothesis: $H_0 : p_1 = p_2$, meaning there is no difference in the proportions.

3. Answer: The correct answer is A. $H_0 : p_1 = p_2$.

Question 85. Paired Sample Test

A group of patients is tested for blood pressure before and after treatment. The mean difference in blood pressure before and after treatment is found to be -5 mmHg. What is the null hypothesis . H_0 . for this paired sample test?

A. $H_0 : \mu_d = 0$ B. $H_0 : \mu_d \neq 0$ C. $H_0 : \mu_1 = \mu_2$ D. $H_0 : \mu_d = -5$

1. Identify the test: This is a paired sample test comparing the difference in measurements before and after treatment.

2. State the null hypothesis:

- Null hypothesis: $H_0 : \mu_d = 0$, meaning there is no difference in the paired measurements.

3. Answer: The correct answer is A. $H_0 : \mu_d = 0$.

Question 86. Slope of Regression Line

A study is conducted to examine the relationship between hours of study and test scores. The regression analysis results in a slope of 0.5 for the relationship between study hours and test scores. What is the null hypothesis H_0 for testing the significance of this slope?

A. $H_0 : \beta = 0$ B. $H_0 : \beta > 0$ C. $H_0 : \beta < 0$ D. $H_0 : \beta \neq 0$

Step-by-Step Solution:

1. Identify the test: This involves testing the significance of the slope in a regression analysis.

2. State the null hypothesis:

- Null hypothesis: $H_0 : \beta = 0$, meaning there is no relationship between the independent and dependent variables.

3. Answer: The correct answer is A. $H_0 : \beta = 0$.

Question 87. One-Sample Z-Test

A factory claims that the average lifetime of its light bulbs is 1,000 hours. A consumer group tests 64 bulbs and finds an average lifetime of 990 hours with a standard deviation of 20 hours. Assuming a normal distribution, should the consumer group reject the factory's claim at a 5% significance level? What are the null and alternative hypotheses?

A. $H_0 : \mu = 1{,}000$, $H_1 : \mu \neq 1{,}000$
B. $H_0 : \mu = 990$, $H_1 : \mu \neq 990$
C. $H_0 : \mu = 1{,}000$, $H_1 : \mu > 1{,}000$
D. $H_0 : \mu \leq 1{,}000$, $H_1 : \mu > 1{,}000$

Step-by-Step Solution:

1. Identify the hypotheses: The claim is about the average lifetime of light bulbs being 1,000 hours.

2. State the null and alternative hypotheses:

- Null hypothesis: $H_0 : \mu = 1{,}000$
- Alternative hypothesis: $H_1 : \mu \neq 1{,}000$

3. Answer: The correct answer is A. $H_0 : \mu = 1{,}000$, $H_1 : \mu \neq 1{,}000$.

Test Statistics (TS)

The image provides detailed information about Test Statistics (TS) used in hypothesis testing, particularly in the context of AP Statistics. Here's a summary:

Test Statistics (TS):

1. General Assumptions:

- AP Stats: Assume σ (population standard deviation) is unknown, so you can always use the T-distribution for mean tests.
- Pooling: If using a T-test and $\sigma_1 \approx \sigma_2$, pooling might be appropriate, though it's noted that the mark scheme often doesn't pool regardless of the size of σ.

2. Decision Methods:

- TS Method: Compare the test statistic (TS) to the critical value (CV):
  - If TS is less than CV: Reject H_0.
  - If TS is greater than CV: Reject H_0.
  - If TS does not equal CV: Reject H_0 if TS > CV.
- P-value Method:
  - Compare the p-value to the significance level α:
- If p-value < α: Reject H_0.

Specific Test Statistics Formulas:

1. **1 Proportion TS (Z-test):** $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$
   - Used for testing a single population proportion.

2. **2 Proportions TS (Z-test):** $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$
   - $\hat{p}$ is the pooled sample proportion.

3. **1 Mean TS (Z or T-test):** $Z = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$, $T = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$ with $\text{df} = n - 1$
   - Used for testing a single population mean.

4. **2 Means TS (Z or T-test):** $Z = \dfrac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$, $T = \dfrac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

- If pooling: $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$
- Degrees of Freedom (df):
    - If pooling: $df = n_1 + n_2 - 2$.
    - If not pooling: $df =$ minimum of (n_1-1, n_2-1).

5. **2 Means Paired TS (Z or T-test):** $Z/T = \frac{\bar{d}-\mu_d}{\frac{s_d}{\sqrt{n}}}$ with df $= n-1$

- Used when the data consists of paired observations.

6. **Slope of Regression Line TS (T-test):** $T = \frac{\text{slope}-\beta}{s_b} = \frac{b-\beta}{s_b}$ with df $= n-2$

- Used to test the significance of the slope in a linear regression model.

Conclusion:

- **Statement**: After conducting the test, you make a conclusion based on the evidence at the given significance level. The conclusion might be phrased as:
- "There is sufficient/insufficient evidence at the α % level to reject H_0, and we can conclude that..."

Real-life Application:

- **1 Proportion TS**: Testing whether the proportion of voters supporting a candidate differs from a known proportion.
- **2 Means TS**: Comparing average scores between two groups of students to see if one group outperforms the other.
- **Slope of Regression Line TS**: Assessing whether there's a significant relationship between advertising spend and sales revenue.

This guide summarizes how to use test statistics in hypothesis testing to determine whether to reject the null hypothesis in favor of the alternative hypothesis.

Question 88. 1 Proportion Z-test

A survey claims that 65% of adults favor a new policy. A sample of 200 adults shows that 120 favor the policy. Test the claim using a 1-proportion Z-test at the 5% significance level. What is the test statistic?

A. 1.96 B. 2.34 C. -1.64 D. -2.45

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : p = 0.65$
2. Calculate the sample proportion: $\hat{p} = \frac{120}{200} = 0.60$
3. Use the formula for the test statistic: $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.60 - 0.65}{\sqrt{\frac{0.65 \times 0.35}{200}}} = \frac{-0.05}{0.0342} = -1.46$
4. Answer: The correct answer is C. -1.46.

Question 89. 2 Proportions Z-test

Two competing products are tested for customer preference. 150 out of 300 customers prefer product A, while 130 out of 300 prefer product B. Test whether there is a significant difference in preference between the two products using a 2-proportion Z-test at the 5% significance level. What is the test statistic?

A. 1.28 B. 1.67 C. 2.04 D. 0.89

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : p_1 = p_2$
2. Calculate the sample proportions: $\hat{p}_1 = \frac{150}{300} = 0.50, \quad \hat{p}_2 = \frac{130}{300} = 0.43$
3. Calculate the pooled proportion: $\hat{p} = \frac{150 + 130}{600} = 0.467$
4. Use the formula for the test statistic: $Z = \frac{0.50 - 0.43}{\sqrt{0.467 \times 0.533 \times \left(\frac{1}{300} + \frac{1}{300}\right)}} = \frac{0.07}{0.041} = 1.71$
5. Answer: The correct answer is B. 1.71.

Question 90. One Sample Mean T-test

A company claims that the average delivery time is 30 minutes. A random sample of 25 deliveries has an average delivery time of 28 minutes with a standard deviation of 5 minutes. Test the claim using a 1-mean T-test at the 5% significance level. What is the test statistic?

A. -2.00 B. -1.96 C. 2.50 D. -2.50

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : \mu = 30$
2. Calculate the test statistic using the T-test formula: $T = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{28 - 30}{\frac{5}{\sqrt{25}}} = \frac{-2}{1} = -2.00$
3. Answer: The correct answer is A. -2.00.

Question 91. Two Sample Means T-test (Independent Samples)

Two groups of students from different schools take the same standardized test. Group A (n=30) has a mean score of 85 with a standard deviation of 4, and Group B (n=25) has a mean score of 83 with a standard deviation of 5. Test whether there is a significant difference between the mean scores of the two groups using a 2-means T-test. What is the test statistic?

A. 2.02 B. 1.73 C. 0.94 D. 1.25

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : \mu_1 = \mu_2$
2. Calculate the test statistic using the T-test formula

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{85 - 83}{\sqrt{\frac{4^2}{30} + \frac{5^2}{25}}} = \frac{2}{\sqrt{0.533 + 1}} = \frac{2}{1.175} = 1.70$$

3. Answer: The correct answer is B. 1.70.

Question 92. Two Sample Means Paired T-test

A group of 20 students is tested on their math skills before and after a training program. The mean difference in their scores (after - before) is 5 points with a standard deviation of 3 points. Test whether the training program had a significant effect on their scores using a paired T-test at the 5% significance level. What is the test statistic?

A. 2.36 B. 3.33 C. 4.20 D. 3.74

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : \mu_d = 0$
2. Calculate the test statistic using the paired T-test formula: $T = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{5}{\frac{3}{\sqrt{20}}} = \frac{5}{0.671} = 7.45$
3. Answer: The correct answer is B. 7.45.

Question 93. Slope of Regression Line T-test

A study analyzes the relationship between hours spent studying and exam scores. The slope of the regression line is found to be 0.8 with a standard error of 0.2. Test the significance of the slope at the 1% significance level. What is the test statistic?

A. 4.00 B. 5.00 C. 3.50 D. 6.00

Step-by-Step Solution:

1. Identify the null hypothesis: $H_0 : \beta = 0$
2. Calculate the test statistic using the T-test formula for the slope: $T = \frac{\text{slope}}{s_b} = \frac{0.8}{0.2} = 4.00$
3. Answer: The correct answer is A. 4.00.

P-value concept

The image provides a detailed explanation of the P-value concept in hypothesis testing and how to calculate and interpret it. Here's a summary:

P-value Calculation:

1. Steps to Calculate:

- Find the Test Statistic (TS) First:
  - This could be a Z-score, T-score, etc., depending on the type of test you're performing.
- Determine the Probability:
  - If your test is left-tailed ($<$): $P(Z < \text{TS})$
  - If your test is right-tailed ($>$): $P(Z > \text{TS})$
  - If your test is two-tailed ($\neq$): $2P(Z > \text{TS})$ or $2P(Z < \text{TS})$
- Use Calculator Functions:
  - Use `normcdf` for normal distributions with μ and $\frac{\sigma}{\sqrt{n}}$ as parameters.
  - Use `tcdf` for T-distributions with degrees of freedom (df).

Interpretation of P-value:

- Decision Rule:
  - If the P-value is less than the significance level α, reject the null hypothesis H_0. This means: If $P < \alpha$, reject H_0
- What the P-value Represents:
  - Assume H_0 is true: The P-value represents the probability of observing a test statistic as extreme as (or more extreme than) the one observed, under the assumption that the null hypothesis is true.
- Direction:
  - If $P(Z < \text{TS})$, the P-value is the probability of being in the lower tail.
  - If $P(Z > \text{TS})$, the P-value is the probability of being in the upper tail.
  - If it's a two-tailed test ($\neq$), the P-value is the sum of the probabilities in both tails.

- Effect of Sample Size:
  - A larger sample size typically leads to a smaller P-value, indicating stronger evidence against the null hypothesis.

Real-life Application:

- **Example in Clinical Trials**: Suppose you're testing whether a new drug is more effective than the standard treatment. You calculate a P-value based on your test statistic. If this P-value is

less than your significance level (e.g., 0.05), you would reject the null hypothesis, suggesting that the new drug is indeed more effective.

This summary outlines how to compute and interpret P-values in the context of hypothesis testing, helping you make informed decisions about whether to reject the null hypothesis based on the evidence provided by your data.

Question 94. P-value for a One-Tailed Z-Test

A pharmaceutical company claims that its new drug increases recovery rate by more than 20%. In a clinical trial, the sample mean recovery rate was 22% with a standard deviation of 5%, based on a sample of 100 patients. Test this claim using a one-tailed Z-test at the 5% significance level. What is the P-value?

A. 0.0228 B. 0.0455 C. 0.0500 D. 0.0122

Step-by-Step Solution:

1. Set up the null hypothesis: $H_0 : \mu = 20\%$ and the alternative hypothesis $H_1 : \mu > 20\%$.

2. Calculate the Z-test statistic: $Z = \frac{0.22 - 0.20}{\frac{0.05}{\sqrt{100}}} = \frac{0.02}{0.005} = 4.00$

3. Find the P-value using the Z-score:

- Since it is a right-tailed test, $P(Z > 4.00)$.
- Look up the Z-value of 4.00 in a standard normal distribution table or use a calculator.
- $P(Z > 4.00) \approx 0.00003$.

4. Answer: The correct answer is D. 0.00003.

Question 95. P-value for a Two-Tailed T-Test

A researcher wants to test if the average weight of a certain species of fish differs from 5 kg. A sample of 16 fish yields a mean weight of 5.2 kg with a standard deviation of 0.4 kg. Test the hypothesis using a two-tailed T-test at the 1% significance level. What is the P-value?

A. 0.0334 B. 0.0668 C. 0.0112 D. 0.0224

Step-by-Step Solution:

1. Set up the null hypothesis: $H_0 : \mu = 5$ kg and the alternative hypothesis $H_1 : \mu \neq 5$ kg.
2. Calculate the T-test statistic: $T = \dfrac{5.2 - 5}{\dfrac{0.4}{\sqrt{16}}} = \dfrac{0.2}{0.1} = 2.00$
3. Find the P-value using the T-score:
   - With degrees of freedom $df = 15$, use a T-distribution table or calculator.
   - For a two-tailed test, $P(T > 2.00)$ for one tail and multiply by 2.
   - Approximate P-value: $2 \times 0.0334 = 0.0668$.
4. Answer: The correct answer is B. 0.0668.

Question 96. Interpretation of P-value

A hypothesis test yields a P-value of 0.03. If the significance level α is 0.05, what conclusion should be drawn?

A. Fail to reject H_0 because the P-value is greater than α.
B. Reject H_0 because the P-value is less than α.
C. Reject H_0 because the P-value is greater than α.
D. Fail to reject H_0 because the P-value is less than α.

Step-by-Step Solution:

1. Interpret the P-value: The P-value represents the probability of observing a test statistic as extreme as the one obtained, under the assumption that the null hypothesis H_0 is true.
2. Compare P-value with α: Since $P = 0.03$ and $\alpha = 0.05$, $P < \alpha$.
3. Decision Rule: Reject H_0 if $P < \alpha$.
4. Answer: The correct answer is B. Reject H_0 because the P-value is less than α.

Chi-Squared (χ^2) Test

Overview of the Chi-Squared (χ^2) Test, which is commonly used in hypothesis testing to assess the independence of two categorical variables or the goodness-of-fit for a model. Here's a summary:

Chi-Squared (χ^2) Test:

1. Test Statistic Calculation:

- Formula: $\chi^2_{calc} = \sum \frac{(O-E)^2}{E}$
  - O represents the observed frequencies.
  - E represents the expected frequencies.

2.Decision Rule:

- Reject H_0 if: $\chi^2_{calc} > \chi^2_{critical}$
- Compare the calculated χ^2 statistic to the critical value from the Chi-Squared distribution table, based on the degrees of freedom (df) and the significance level (α).

3. Hypotheses:

- **Null Hypothesis** (H_0): The **variables are independent,** or the data fits the expected distribution (e.g., in the specified ratio).
- **Alternative Hypothesis** (H_1): The variables are not independent, or the data does not fit the expected distribution.

4. Degrees of Freedom (df):

- Independence Test: $df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$
- Goodness-of-Fit Test: $df = n - 1$
- If approximating p or μ / σ (which is rare), df = $n - 2$.

Real-life Application:

- **Independence Test Example**: If you're testing whether gender and voting preference are independent in a survey, you would use a Chi-Squared test of independence.
- **Goodness-of-Fit Example**: If you want to test whether a die is fair, you would compare the observed frequencies of each face to the expected frequencies using a Chi-Squared goodness-of-fit test.

This summary outlines the key steps in performing a Chi-Squared test, which is useful for determining whether observed data aligns with theoretical expectations in categorical data analysis.

Question 97. Chi-Squared Test for Independence

A study is conducted to determine if there is an association between gender (male/female) and preference for a new product (like/dislike). The following table shows the observed frequencies:

	Like	Dislike	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

What is the chi-squared test statistic?

A. 4.57 B. 5.20 C. 6.67 D. 3.45

Step-by-Step Solution:

1. Calculate the expected frequencies:

- Expected for Male-Like: $\frac{50\times 70}{100}=35$
- Expected for Male-Dislike: $\frac{50\times 30}{100}=15$
- Expected for Female-Like: $\frac{50\times 70}{100}=35$
- Expected for Female-Dislike: $\frac{50\times 30}{100}=15$

2. Calculate the chi-squared statistic:

$$\chi^2=\sum\frac{(O-E)^2}{E}=\frac{(30-35)^2}{35}+\frac{(20-15)^2}{15}+\frac{(40-35)^2}{35}+\frac{(10-15)^2}{15}=4.76$$

3. Answer: The correct answer is C. 4.76.

Question 98. Chi-Squared Goodness-of-Fit Test

A die is rolled 60 times, and the observed frequencies of the outcomes are recorded as follows:

Outcome	1	2	3	4	5	6
Frequency	8	10	12	9	11	10

Is the die fair at the 5% significance level? What is the chi-squared test statistic?

A. 0.89 B. 1.0 C. 2.35 D. 3.50

Step-by-Step Solution:

1. Calculate the expected frequencies:

- For a fair die, each outcome has an equal probability of $\frac{1}{6}$.
- Expected frequency for each outcome: $\frac{60}{6}=10$.

2. Calculate the chi-squared statistic:

$$\chi^2=\sum\frac{(O-E)^2}{E}=\frac{(8-10)^2}{10}+\frac{(10-10)^2}{10}+\frac{(12-10)^2}{10}+\frac{(9-10)^2}{10}+\frac{(11-10)^2}{10}+\frac{(10-10)^2}{10}=1.0$$

3. Answer: The correct answer is B. 1.0.

Question 99. Degrees of Freedom for Chi-Squared Test of Independence

In a contingency table with 3 rows and 4 columns, what are the degrees of freedom for the chi-squared test of independence?

A. 3 B. 6 C. 8 D. 9

Step-by-Step Solution:

1. Use the formula for degrees of freedom: $df=(r-1)\times(c-1)$, where r is the number of rows and c is the number of columns.
2. Calculate the degrees of freedom: $df=(3-1)\times(4-1)=2\times3=6$
3. Answer: The correct answer is B. 6.

Question 100. Interpreting Chi-Squared Test Results

A chi-squared test for independence yields a chi-squared statistic of 10.5 with 4 degrees of freedom. If the critical value at the 5% significance level is 9.49, what conclusion should be drawn?

A. Fail to reject H_0; there is no association.
B. Reject H_0; there is an association.
C. Fail to reject H_0; the data fits the model.
D. Reject H_0; the data does not fit the model.

Step-by-Step Solution:

1. Compare the calculated chi-squared statistic to the critical value:

- $\chi^2_{calc} = 10.5$, $\chi^2_{critical} = 9.49$.
- Since $\chi^2_{calc} > \chi^2_{critical}$, we reject the null hypothesis.

2. Conclusion: There is evidence to suggest an association between the variables.

3. Answer: The correct answer is B. Reject H_0; there is an association.

Errors in Hypothesis Testing

Explanation of Errors in Hypothesis Testing, specifically focusing on Type 1 and Type 2 Errors, and how to calculate them. Here's a summarized explanation:

Definitions of Errors:

1. Type 1 Error:

- Definition: Occurs when the null hypothesis (H_0) is true, but we incorrectly reject it.
- Probability: Represented by α , which is the significance level of the test. It is the probability of making a Type 1 error.

2. Type 2 Error:

- Definition: Occurs when the null hypothesis (H_0) is false, but we fail to reject it (i.e., incorrectly accept it).
- Probability: Represented by β , which is the probability of not being in the critical region when H_0 is false.

Calculations:

1. Type 1 Error (α):

- This is directly set by the significance level of the test.
- α is the area in the critical region (the probability of rejecting H_0 when it is true).

2. Type 2 Error (β):

- The probability of not rejecting H_0 when it is false.
- Steps to Calculate:
  - Step 1: Find the critical value (CV) using `invnorm`.
    - $<$: area $=\alpha$ (left-tail test).
    - $>$: area $=\alpha$ (right-tail test).
    - $\neq$: area $=\frac{\alpha}{2}$ (two-tail test).
  - Step 2: Find the error using `normcdf`.
    - $<$: Lower = -100, Upper = CV.
    - $>$: Lower = CV, Upper = 100.
    - $\neq$: Lower = CV1, Upper = CV2.

Formulas for Type 2 Error:

- Left-Tailed Test ($<$): $P\left(\overline{x} < \mu + z_c \frac{\sigma}{\sqrt{n}}\right)$
    - Lower = -100, Upper = $\mu + z_c \frac{\sigma}{\sqrt{n}}$ (new μ).
- Right-Tailed Test ($>$): $P\left(\overline{x} > \mu + z_c \frac{\sigma}{\sqrt{n}}\right)$
    - Lower = $\mu + z_c \frac{\sigma}{\sqrt{n}}$, Upper = 100.
- Two-Tailed Test ($\neq$): $P\left(\mu - z_c \frac{\sigma}{\sqrt{n}} < \overline{x} < \mu + z_c \frac{\sigma}{\sqrt{n}}\right)$
    - Lower = CV1, Upper = CV2.

Managing Errors:

- To Increase Type 1 Error:
    - Increase the significance level (α).
- To Increase Type 2 Error:
    - Decrease the significance level (α).
    - Decreasing the Type 1 error increases the Type 2 error.
    - Take smaller samples, increase α , or increase the standard error (SE).
- Power of a Test:
    - Power = $1-\beta$, which is the probability of correctly rejecting H_0 when it is false.

Real-life Application:

- **Type 1 Error Example:** In a clinical trial, if a new drug is actually not effective but the trial concludes that it is effective, this is a Type 1 error.
- **Type 2 Error Example:** If the new drug is effective, but the trial concludes that it is not effective, this is a Type 2 error.

Understanding Type 1 and Type 2 errors is critical in hypothesis testing, as it helps in designing tests that minimize these errors and increase the reliability of the conclusions drawn from statistical analyses.

Question 101. Identifying Type 1 Error

A researcher is testing the effectiveness of a new drug. The null hypothesis (H_0) states that the drug has no effect. The researcher decides to reject the null hypothesis. If in reality, the drug has no effect, what type of error has the researcher committed?

A. Type 1 Error B. Type 2 Error C. No error D. Power of the test

Step-by-Step Solution:

1. Define Type 1 Error: This occurs when the null hypothesis (H_0) is true, but we incorrectly reject it.
2. Scenario Analysis: In this case, H_0 is true (the drug has no effect), but the researcher rejects H_0.
3. Conclusion: The researcher has committed a Type 1 Error.
4. Answer: The correct answer is A. Type 1 Error.

Question 102. Identifying Type 2 Error

In a study to determine whether a new teaching method is better than the traditional method, the null hypothesis (H_0) is that there is no difference in effectiveness between the two methods. If the new method is actually more effective but the researcher fails to reject the null hypothesis, what type of error is this?

A. Type 1 Error B. Type 2 Error C. No error D. Increase in significance level

Step-by-Step Solution:

1. Define Type 2 Error: This occurs when the null hypothesis (H_0) is false, but we fail to reject it.
2. Scenario Analysis: Here, the new method is actually more effective (so H_0 is false), but the researcher does not reject H_0.
3. Conclusion: The researcher has committed a Type 2 Error.
4. Answer: The correct answer is B. Type 2 Error.

Question 103. Calculating Type 1 Error Probability

A hypothesis test is conducted at a 5% significance level. What is the probability of committing a Type 1 Error?

A. 0.01	B. 0.05	C. 0.10	D. Cannot be determined

Step-by-Step Solution:

1. Recall the definition: The significance level (α) is the probability of committing a Type 1 Error.
2. Given Significance Level: Here, $\alpha = 0.05$.
3. Conclusion: The probability of committing a Type 1 Error is 0.05.
4. Answer: The correct answer is B. 0.05.

Question 104. Effect of Sample Size on Type 2 Error

If the sample size is increased in a hypothesis test, what is likely to happen to the probability of a Type 2 Error (β)?

A. Increase	B. Decrease	C. Stay the same	D. Cannot be determined

Step-by-Step Solution:

1. Recall the relationship: Increasing the sample size generally decreases the standard error, making it easier to detect a true effect and reducing the probability of a Type 2 Error.
2. Scenario Analysis: With a larger sample size, β (the probability of a Type 2 Error) is likely to decrease.
3. Conclusion: Increasing the sample size decreases β.
4. Answer: The correct answer is B. Decrease.

End of the Workbook

Congratulations!

I'm delighted to have been able to help you advance your mathematical understanding. Should you have any further questions or need additional assistance down the line, please don't hesitate to reach out. Best of luck with your ongoing studies!

About Authors

Seonwan Myung, Ph.D.

Dr. Seonwan Myung combines a deep expertise in industrial engineering with a passion for education, particularly in mathematical learning. Holding advanced degrees in industrial engineering (Human Factors), Dr. Myung has devoted his career to making math both approachable and enjoyable for learners of all ages. His vast teaching experience, along with his specialization in Learning Instructional Design for Mathematics, enables him to develop innovative tools that engage and motivate students.

eSpyMath Books

eSpyMath Grade 6 Math Workbook
eSpyMath Grade 7 Math Workbook
eSpyMath AP Calculus BC Workbook

eSpyMath AP Pre-Calculus Workbook
eSpyMath AP Statistics Workbook
eSpyMath AP Calculus AB/BC Workbook

eSpyMath Pre-Algebra Workbook
eSpyMath Algebra 1 Workbook
eSpyMath Algebra 2 Workbook

eSpyMath AP Calculus AB Textbook
eSpyMath AP Calculus AB/BC Textbook