Basic Statistics for Data Science

Data science is a rapidly growing field that encompasses a wide range of techniques and methods for extracting insights and knowledge from data. At the heart of data science lies statistics, which provides the foundational principles and tools to analyze and interpret data. In this article, we will explore some key concepts in basic statistics, which are crucial for understanding and working with data effectively.

Descriptive statistics

Descriptive statistics is a branch of statistics that focuses on summarizing and describing the main characteristics of a dataset. It provides a concise overview of the data, allowing analysts to gain insights into its central tendency, variability, and distribution.

Let's consider the following sample dataset for 10 cars:

Car Model Sale	Price (thousands $)	Age of Car (years)	Mileage (thousands miles)	Previous Owners
Model A	30	1	10	0
Model B	35	2	15	1
Model A	33	1	11	0
Model C	40	3	20	1
Model D	50	2	10	0
Model E	60	3	30	2
Model B	36	3	16	1
Model E	58	2	25	1
Model D	120	1	5	0
Model A	32	2	12	1

The following are the descriptive statistics of this data:

Mean: The mean (average) is found by adding all values of a specific variable and dividing by the count of those values.
For instance, the mean price is:
(30+35+33+40+50+60+36+58+120+32) / 10 = $49.4K.

Median: The median is the middle value in a sorted list of values.
If we sort the car prices, the middle values are 35 and 50, so the median is the average of these, i.e., $42.5K.

Mode: The mode is the most frequent value in a data set.
For the "Car Model," Model A appears most frequently, so it is the mode.

Geometric Mean: The geometric mean of n numbers is the nth root of their product. It's useful when comparing different products, as it tends to dampen the effect of very high or low values. For the car prices, you would multiply all prices together and take the 10th root of the product.

Regarding asymmetrical data and outliers, like the price of Model D at $120K, this is a value that deviates significantly from other values in the dataset. It affects the mean significantly by pulling it up but doesn't affect the median as much. In such cases, the median can be a more useful measure of central tendency as it is less influenced by extreme values. You could also consider trimming outliers or using robust statistical measures like the trimmed mean or Winsorized mean.

Descriptive statistics Use-Cases In EDA:

Exploratory Data Analysis (EDA) is a crucial step in data analysis where you examine and understand the dataset's characteristics and distribution. Imputation is the process of replacing missing or erroneous values in the dataset with estimated or substituted values. Mean, median, and mode are commonly used statistical measures for imputing missing data during EDA. Here's how you can use them

Consider the following dataset:

Car Model Sale	Price (thousands $)	Age of Car (years)	Mileage (thousands miles)	Previous Owners
Model A	30	1	10	0
Model B	35	2	NA	1
Model A	33	1	11	0
Model C	40	3	NA	1
Model D	50	2	10	0
Model E	NA	3	30	2
Model B	36	NA	16	1
Model E	58	2	NA	1
Model D	120	1	5	0
Model A	32	2	12	1

Mean Imputation:

Mean imputation is suitable for numerical data with a normal or approximately normal distribution. It replaces missing values with the mean of the available data. Calculate the mean of the non-missing values for the variable of interest and replace the missing values with this mean.

We can replace the mileage column NA values with the mean: 15.4

Median Imputation:

Median imputation is useful for data with outliers or skewed distributions. It's more robust to extreme values than mean imputation. Calculate the median of the non-missing values for the variable and replace the missing values with this median

We can replace the Price column's missing value with median: 36

Mode Imputation:
Mode imputation is appropriate for categorical or nominal data. It replaces missing values with the most frequently occurring category. Calculate the mode (most common category) of the variable and replace the missing values with the mode.

We can replace the Age of Car column's missing value with mode: 2

Measures of Dispersion:

Measures of dispersion, such as range, variance, and standard deviation, are used to understand the spread or variability of data. They provide insights into how data points are distributed around the central tendency, such as the mean or median. A thorough understanding of measures of dispersion is essential for assessing the variability within a dataset and making informed decisions based on the data's spread.

Range: The range is the difference between the largest and smallest values. For our sale price data: Range = Max - Min = $120k - $30k = $90k. This indicates the total spread of the sale prices.

Quartiles: Quartiles are calculated by sorting the data and finding the values that split the data into four equal parts. Here, Q1 = $32.5k, Q2 (median) = $43k, Q3 = $59k.

Interquartile Range (IQR): This is the range of the middle 50% of the values. IQR = Q3; Q1 = $59k - $32.5k = $26.5k. The IQR gives an idea of the spread of the middle half of the data.

Skewness: The skewness of our data can be seen visually, with most prices on the lower end and a few on the high end, which results in a positive skew.

Kurtosis: Similarly, the high sale price of $120k suggests a high kurtosis, indicating a distribution with heavy tails and outliers.

Variance: The variance is the average of the squared differences from the mean. Here's how we'd calculate it for our data:

1. Compute the mean (average): $49.4k
2. Subtract the mean from each price and square the result to get the squared differences.
3. Find the average of those squared differences.

For simplicity, let's say the squared differences add up to 4473 (in thousands of dollars squared). So, Variance = 4473 / 10 = 447.3 (in thousands of dollars squared). This value tells us how spread out the prices are, squared.

Standard Deviation: This is simply the square root of the variance. So, in our case, the standard deviation = sqrt (447.3) = $21.1k. This value tells us how spread out the prices are in the same unit as the original prices (thousands of dollars).

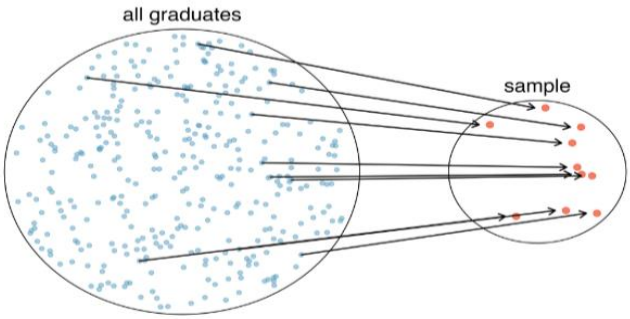
In data science, these measures help us understand our data. For example, a high standard deviation would tell us that the car prices vary a lot. If we were building a predictive model, knowing our data is skewed might t lead us to transform our data to achieve better model performance. If we were investigating the relationship between variables, understanding our quartiles might help us identify non-linear relationships. In essence, these metrics give us the context we need to apply machine learning techniques effectively.

Sample Vs Population

Consider the following three research questions:
What is the average mercury content in swordfish in the Atlantic Ocean?

The research question refers to a target **population**. In this question, the target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Oftentimes, it is not feasible to collect data for every case in a population. Collecting data for an entire population is called a **census**. A census is difficult because it is too expensive to collect data for the entire population, but it might also be because it is difficult or impossible to identify the entire population of interest! Instead, a sample is taken. A **sample** is the data you have. Ideally, a sample is a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and to answer the research question.

We might try to estimate the time to graduation for Duke undergraduates in the last five years by collecting a sample of graduates. All graduates in the last five years represent the population, and graduates who are selected for review are collectively called the sample. In general, we always seek to randomly select a sample from a population



Sampling Techniques, Biases, and Variables

Sampling techniques involve selecting a subset of individuals or data points from a larger population to gather information and make inferences about the population as a whole. It is crucial to employ appropriate sampling techniques to ensure the validity and representativeness of the data collected. Biases, such as selection bias or response bias, can skew the results and lead to erroneous conclusions. Additionally, variables can be classified into different types, such as categorical, ordinal, and continuous, which impacts the choice of statistical techniques used for analysis.

Sampling Types

1. **Simple Random Sampling**: Each car sale from the data set has an equal chance of being included in the sample. For example, out of 100 car sales, we could randomly select 20 sales for analysis.
2. **Systematic Sampling:** We select every nth car sale from the data set. For example, we could select every 10th car sale for a total of 10 sales.
3. **Stratified Sampling**: We divide the data set into strata (subgroups) based on certain characteristics like car model or car make, and randomly sample from each stratum. For example, we could stratify the data by the car model and then select 5 sales from each model.
4. **Cluster Sampling:** We divide the data set into clusters (groups), randomly select a few clusters, and include all sales from those clusters in the sample. For example, if the data are grouped by the month of the sale, we could select 2 months and include all sales from those months in our sample.

Variable Types

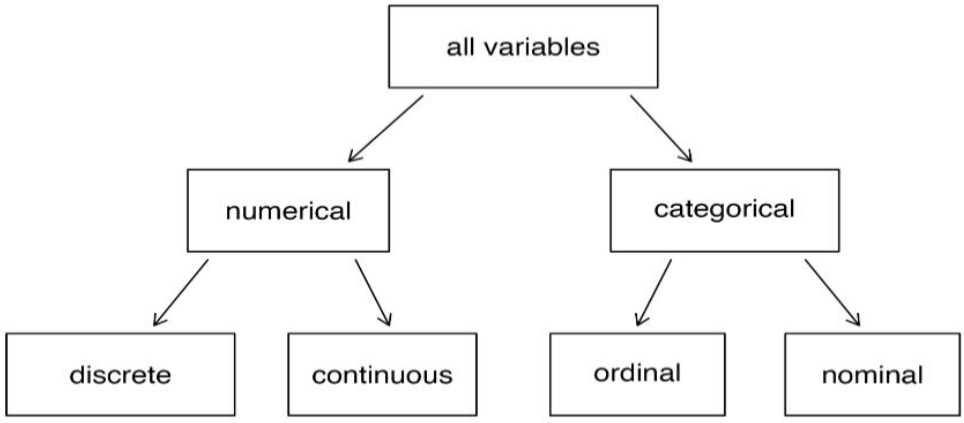


Figure 1.1: Breakdown of variables into their respective types.

Name	State	Pop2017	POP_Change	Unemployment_rate	Median_edu
Autauga County	Alabama	55504	1.48	3.86	Some_college
Baldwin County	Alabama	212628	9.19	3.99	Some_college
Barbour County	Alabama	25270	-6.22	5.90	Hs_diploma
Bibb County	Alabama	22668	0.73	4.39	Hs_diploma
Blount County	Alabama	58013	0.68	4.02	Hs_diploma

First consider **unemployment_rate**, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area code as numerical since the average, sum, and difference of area codes does not have any clear meaning. Instead, we would consider area codes as a categorical variable.

The **pop2017** variable is also **numerical**, although it seems to be a little different than unemployment_rate. This variable of the population count can only take whole non-negative numbers (0, 1, 2, …). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the unemployment rate variable is said to be **continuous**.

The variable state can take up to 51 values after accounting for Washington, DC: AL, AK, …, and WY. Because the responses themselves are categories, state is called a categorical variable, and the possible values (states) are called the variable's levels (e.g., DC, AL, AK, etc.)

Finally, consider the **median_**edu variable, which describes the median education level of county residents and takes values below_hs, hs_diploma, some_college, or bachelors in each county. This

variable seems to be a hybrid: it is a categorical variable, but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any categorical variable in this book will be treated as a nominal (unordered) categorical variable.

Biases

1. **Selection Bias:**
   Selection bias occurs when the selection process for participants or data points is not random, leading to a non-representative sample.
   Example: In a survey about online shopping preferences, if you only collect responses from people who are active on a particular online shopping platform, you may miss the opinions of those who prefer other platforms.

2. **Sampling Bias:**
   Sampling bias happens when the sample used for analysis is not representative of the entire population, leading to skewed results.
   Example: If you want to estimate the average income of people in a city and you survey people only in affluent neighborhoods, your estimate will be biased and likely higher than the true average.

3. **Non-Response Bias:**
   Non-response bias occurs when individuals who do not participate in a survey or study differ systematically from those who do, leading to an inaccurate representation of the population.
   Example: If a political poll only reaches respondents who are enthusiastic about politics and willing to participate, it may not accurately reflect the opinions of less engaged voters.

4. **Measurement Bias:**
   Measurement bias occurs when the method of measurement or data collection systematically underestimates or overestimates a variable of interest.
   Example: Using a faulty thermometer to measure body temperature in a medical study may introduce measurement bias, leading to inaccurate temperature readings.

5. **Confirmation Bias:**
   Confirmation bias refers to the tendency to search for, interpret, and remember information in a way that confirms one's preexisting beliefs or hypotheses, leading to a skewed analysis.
   Example: A researcher analyzing the effects of a new teaching method may focus on evidence that supports its effectiveness while ignoring or downplaying contradictory data.

6. **Observer Bias:**
   Observer bias occurs when the person collecting or analyzing data has preconceived notions or expectations that influence their observations or interpretations.
   Example: In a clinical trial, a researcher who expects a certain treatment to be effective may unintentionally interpret patient outcomes more favorably for that treatment.
7. **Publication Bias:**
   Publication bias happens when studies with statistically significant results are more likely to be published than studies with non-significant results, leading to an overrepresentation of significant findings in the literature.
   Example: If pharmaceutical companies only publish studies where their drug shows positive effects, it may create a misleading impression of the drug's overall effectiveness.
8. **Survivorship Bias:**
   Survivorship bias occurs when data is only collected from surviving or successful subjects, ignoring those that did not survive or were unsuccessful.
   Example: Analyzing the characteristics of successful businesses without considering those that failed can lead to unrealistic expectations and strategies.
9. **Response Bias:**
   Response bias occurs when respondents provide inaccurate or biased information due to social desirability, misinterpretation of questions, or other factors.
   Example: In a survey about alcohol consumption, respondents may underreport their drinking habits due to the stigma associated with excessive drinking.
10. **Time-Interval Bias:**
    Time-interval bias occurs when data is collected over different time periods, and the choice of time intervals influences the results.
    Example: If you analyze sales data for a retail store over a year but compare it to data from the previous year with a different economic climate, it may lead to misleading conclusions about performance.

Regression Analysis

Assume you've run the regression in Excel using the 'Data Analysis' add-in, with 'Sale Price' as your dependent variable and 'Age of Car', 'Mileage', and 'Previous Owners' as independent variables. The output table will give you a list of coefficients, their standard errors, t-Stats, P-values, and confidence intervals for each independent variable, as well as various summary statistics like R Square, Adjusted R Square, etc. Let's focus on the coefficients and some of these summary statistics.

	Coefficients	Standard Error	T Stat	P-value
Intercept	60.0	2.5	24.0	0.000
Age of Car	-1.5	0.3	-5.0	0.000
Mileage	-0.8	0.1	-8.0	0.000
Previous Owners	2.0	0.8	2.5	0.015

R Square	0.85
Adjusted R Square	0.83
Standard Error	2.6
Observations	100

The coefficients table tells us that for each additional year of the car's age, the Sale Price decreases by $1.5k, holding other variables constant. For each additional thousand miles on the car, the Sale Price decreases by $0.8k, holding other variables constant. And for each additional previous owner, the Sale Price increases by $2k, holding other variables constant.

The sign of each coefficient indicates the direction of the relationship with the Sale Price. Negative coefficients suggest an inverse relationship (as Age and Mileage increase, the Sale Price decreases) while a positive coefficient suggests a direct relationship (as the number of Previous Owners increases, the Sale Price increases, though this might be counterintuitive).

The P-values are all less than 0.05, suggesting that all these relationships are statistically significant at the 5% level.

The R-Square value of 0.85 indicates that 85% of the variance in Sale Price can be explained by our model, i.e., by the Age of Car, Mileage, and Previous Owners. Adjusted R-Square takes into account the number of predictors in the model and can be a better measure when comparing different models.

Finally, the Standard Error of 2.6 is a measure of the accuracy of the predictions. It's an estimate of the standard deviation of the error term ε, and it's in the same units as the dependent variable.

By interpreting these indicators, we gain insight into the relationships between the independent and dependent variables and the q

As for the non-linear relationship, a multiple linear regression model assumes a linear relationship between the independent and dependent variables. However, in real-world scenarios, this relationship is often not perfectly linear. For instance, the effect of Mileage on Sale Price might not be the same at low mileage versus high mileage. Perhaps at low mileage, the Sale Price decreases slowly with increasing mileage, but after a certain point, the Sale Price might start decreasing rapidly.

We can sometimes handle non-linearity by transforming the data or the model. For instance, we might take the logarithm of one or more variables, or we might add interaction terms or polynomial terms to our model to capture the non-linear relationship. Careful exploratory data analysis can often help us identify non-linear relationships, choose appropriate transformations, and build better models.

Confidence Interval, Significance level and p value.

Confidence intervals provide a range of values within which we can be reasonably confident that the true population parameter lies. It quantifies the uncertainty associated with estimating population parameters from sample data. The significance level, often denoted as alpha, determines the threshold below which we reject or accept a null hypothesis in hypothesis testing. The p-value represents the probability of obtaining results as extreme as the observed data,

assuming the null hypothesis is true. It helps us make decisions about the statistical significance of our findings.

1. **Confidence Interval (CI):** Represents how confident are you about your model?
   A confidence interval gives an estimated range of values which is likely to include an unknown population parameter. In the context of multiple linear regression, we often compute confidence intervals for the coefficients of the independent variables. For instance, if the coefficient of 'Car Age' is -500 with a 95% confidence interval of [-700, -300], we would be 95% confident that the true population effect of 'Car Age' on 'Sale Price' is between -$700 and -$300.
2. **Significance Level:** $\alpha = 1 -$ confidence interval
   The significance level, denoted as α, is a threshold that we set for deciding whether the null hypothesis should be rejected. The null hypothesis for each independent variable in a multiple regression is typically that the variable has no effect (i.e., a coefficient of zero). A common choice of α is 0.05, meaning that we would need evidence strong enough to occur by chance less than 5% of the time in order to reject the null hypothesis.

   Imagine you're playing a game where you have to find hidden treasures. To win, you need to be very sure that you've found a real treasure, not just something shiny. The significance level is like deciding how sure you want to be.
   Imagine you and your friends are playing a game to find hidden gems in a park. You want to make sure that when you say you found a real gem, it's a genuine one, not just a shiny rock.

   Significance Level: Let's say you set a significance level of 5%, which is the same as 0.05. This means you want to be 95% sure that what you found is a real gem before you get excited.
   Hypothesis: Your hypothesis is that a gem is hidden in the park.
   Test: You and your friends search and find a shiny object. You want to test if it's a real gem.
   Comparison: You compare your find to the 5% level. If there's less than a 5% chance that what you found is just a shiny rock (based on tests and evidence), you get to declare it as a gem and celebrate your discovery.

   So, if the evidence strongly suggests that what you found is rare and precious (less than 5% chance of being wrong), you'll be confident in calling it a gem. But if the evidence isn't convincing enough (more than 5% chance of being wrong), you'll be cautious and might not consider it a gem.

3. **P-value:** The p-value is the probability of obtaining a result as extreme as, or more extreme than, the observed data, under the null hypothesis. In multiple linear regression, each independent variable will have an associated p-value. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so we reject the null hypothesis. For instance, if the p-value for 'Car Age' is 0.01, this provides strong evidence that the effect of 'Car Age' on 'Sale Price' is not zero.

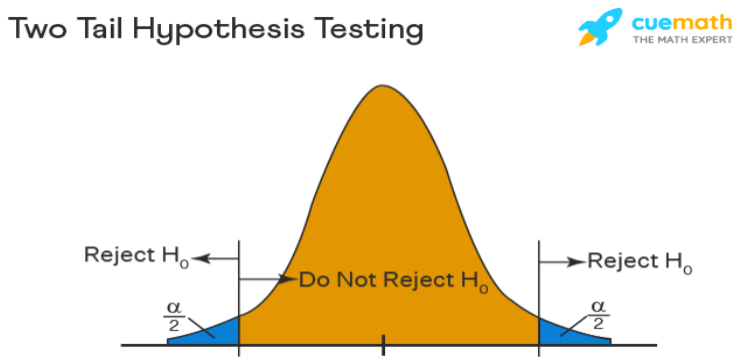
Assuming we have a multiple linear regression model for 'Sale Price' using 'Car Age' and 'Mileage' as independent variables, and we find that 'Car Age' has a coefficient of -500, a 95% confidence interval of [-700, -300], and a p-value of 0.01, we would interpret this as strong evidence that older cars sell for less money, with each additional year of age reducing the sale price by between $300 and $700.

Hypothesis Testing:

Hypothesis Statements:
Ho = Null hypothesis or the default statement (The average time it takes for customers to complete an online purchase is equal to 5 minutes)
HA = Opposite to Ho, e.g (The average time it takes for customers to complete an online purchase is not equal to 5 minutes)

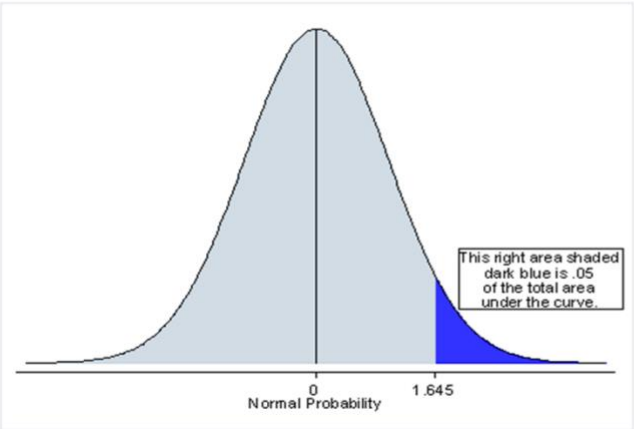
The above-mentioned statements are an example of two-tail test.

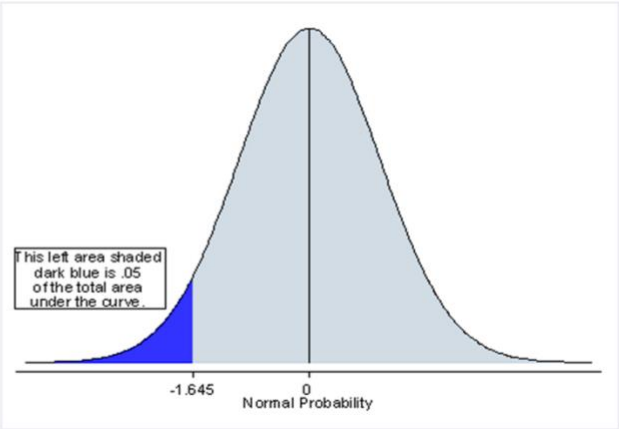
Two Tail Hypothesis Test:
A two-tailed hypothesis test considers both directions of deviation from the null hypothesis (either longer or shorter times).
Ho = 5 and HA != 5



If your z-statistic lies in the blue colored region than you reject the null hypothesis, whereas if it falls in the yellow region, you ***fail*** to reject the null hypothesis.

One Tail Hypothesis Test:
A one-tail hypothesis test, also known as a one-sided hypothesis test, is a type of statistical hypothesis test that focuses on detecting an effect or difference in only one direction. In other words, it tests whether a population parameter is either greater than or less than a specific value, but not both.
Ho = 5 and HA < 5 OR Ho = 5 and HA > 5





If your z-statistic lies in the blue colored region than you reject the null hypothesis, whereas if it falls in the gray region, you fail to reject the null hypothesis.

Z-Test

Used when sample size is greater than 30 and the population standard deviation is known

$Z = (\bar{X}-\mu)/(s/\sqrt{n})$

Where $\bar{X}$ = sample avg

μ = population avg

s = population standard deviation

n = no of samples

T-Test

Used when sample size is less than 30 and the population standard deviation is not known

$Z = (\bar{X}-\mu)/s_x \sqrt{n})$

Where s_x = sample standard deviation

Chi-Square

A chi-square test is a statistical hypothesis test used to determine whether there is a significant association or relationship between two ***categorical*** variables.

E.g., Testing whether there is an association between gender (male/female) and smoking status (smoker/non-smoker).

$X^2 = (O-E)^2/ E$

Where O = observed value

E = expected value

ANOVA

ANOVA is a statistical technique that examines the variability between groups and within groups to assess whether there are significant differences in the means of these groups. It does this by comparing the variation between groups to the variation within groups, and it calculates a test statistic (F-statistic) to determine if these differences are statistically significant.
E.g., Comparing the effectiveness of different drug treatments on patients from multiple hospitals. Analyzing the impact of various diets or exercise routines on weight loss across different age groups.

Example of Two-Tailed Hypothesis Test:

Problem Statement

We want to determine whether the average daily homework time for high school students is different from 60 minutes, with a significance level (α) of 0.05 or confidence interval of 95%

Step 1: Hypothesis:

Ho = The average daily homework time for high school students is equal to 60 minutes.

HA = The average daily homework time for high school students is different from 60 minutes.

Step 2: Data Collection:

Data = [55, 60, 58, 62, 61, 59, 63, 56, 64, 57]

Step 3: Calculate Mean And Standard Deviation:

Sample Mean ($\bar{x}$) = 55 + 60 + 58 + 62 + 61 + 59 + 63 + 56 + 64 + 57 / 10 = 60.5 minutes

Sample Standard Deviation (s): 3.22 minutes

Step 4: Calculate Critical Values:

With the help of z-statistic table or an online calculator you can calculate critical values for α/2 = 0.025 which would turn out to be +1.96 and -1.96 (the values of blue regions)

Step 5: Calculate the Z-statistic:

$Z = (\bar{X}-\mu)/(s/\sqrt{n})$ -> (60.5 – 60)/3.22/ √10

Z = 0.30

Conclusion:

The calculated test statistic (0.30) falls between -2.26 and +2.26 thus we fail to reject the null hypothesis (H0).

P-Value:

Instead of doing all these calculations, statisticians have calculated a P-value parameter for us which tells us the probability that our z-value falls in the rejected region.

So, calculate the P-Value and if:

P value < α -> Reject Ho

P value > α -> Fail to reject Ho

Difference between Correlation and Causality

Correlation and causality are two fundamental concepts in data analysis and interpretation, but they convey different types of relationships.

Correlation refers to a statistical relationship between two variables, which could be positive (both variables increase or decrease together), negative (one variable increases when the other decreases), or zero (no relationship). For instance, in our dataset, we might observe a negative correlation between 'Age of Car' and 'Sale Price', suggesting that as the age of the car increases, the sale price tends to decrease.

However, correlation doesn't imply causation. It only shows that two variables move together, but it doesn't indicate whether one variable causes the other to change. Two variables could be correlated because they are both caused by a third variable, or even due to random chance.

Causality, on the other hand, implies a cause-and-effect relationship between two variables. It suggests that a change in one variable (the cause) directly results in a change in another variable (the effect). Establishing causality often requires experimental or longitudinal data and more rigorous statistical or experimental methods.

In our example, while we observe correlations (e.g., between car age and sale price), we cannot definitively establish causality based on the available data. We might suspect that a car's age causes it to have a lower sale price due to factors like wear and tear and outdated features, but our dataset doesn't definitively prove this. Other factors that we haven't considered might also be at play.

To establish causality, we would need to control for all other factors, which is often difficult in practice. For example, we might need to compare the sale prices of two identical cars, with one being older than the other, while ensuring that factors like mileage, previous owners, and other

variables are the same. Such an experimental setup is challenging to achieve, especially in observational studies like this.

Advanced Concepts

1. **Probability Distributions:** Let's consider that the 'Sale Price' of our cars follows a normal distribution with a mean (μ) of $50k and a standard deviation (σ) of $10k. According to the properties of a normal distribution, we can say that about 68% of all car sale prices fall within one standard deviation from the mean, i.e., between $40k and $60k. Similarly, about 95% fall within two standard deviations, i.e., between $30k and $70k.
2. **Central Limit Theorem (CLT):** Suppose we start taking samples of 30 car sale prices and calculate the mean sale price of these samples. According to the CLT, if we do this many times, the distribution of these sample means will approximate a normal distribution, regardless of the shape of the original distribution of all car sale prices. So even if the sale prices are skewed right, the distribution of sample means of sale prices will still be approximately normal if our sample size is large enough. This is crucial for making statistical inferences about the mean sale price using techniques that assume normality.
3. **Bayesian Statistics:** Consider we have prior knowledge that the average sale price of a sedan car model is usually around $35k, but we then record new data from our current inventory where the average price appears to be $40k. A Bayesian approach would allow us to update our prior belief about the average sale price given this new data, leading to a posterior belief that might suggest that the average price is likely somewhere between $35k and $40k.
4. **Confidence Intervals:** Suppose we have a sample mean ($\bar{x}$) sale price of $50k with a standard deviation (s) of $10k for a sample of 30 cars. We could calculate a 95% confidence interval for the population mean sale price using the formula $\bar{x} \pm 1.96*(s/\sqrt{n})$, which in this case gives us $50k ± $3.58k, or about $46.42k to $53.58k. We interpret this as: we're 95% confident that the true average sale price of all cars falls within this range.
5. **Hypothesis Testing:** Hypothesis testing involves stating a null hypothesis (H0, a statement of no effect or no difference) and an alternative hypothesis (Ha, what we suspect might be true). Suppose we believe cars with one previous owner sell for a higher price on average than cars with two previous owners.

We would set this up as: H0: $\mu1 - \mu2 = 0$ (The average sale price is the same for cars with one or two Previous owners) Ha: $\mu1 - \mu2 > 0$ (Cars with one previous owner have a higher average sale price)

If we then collect data and calculate the average sale prices for cars with one and two previous owners, we can use a t-test to determine the probability of observing such a difference in averages if the null hypothesis were true. If this probability, the p-value, is less than a predetermined significance level (commonly 0.05), we would reject the null hypothesis in favor of the alternative.

For example, if cars with one previous owner have an average sale price of $55k and cars with two previous owners have an average sale price of $50k, and we get a p-value of 0.03 from our t-test, we would reject the null hypothesis and conclude that cars with one previous owner sell for a higher price on average, with a 95% confidence level.

Probability Distributions

Probability distributions can provide a lot of insight into your data and inform your decision-making process. Let's discuss a few common types:

1. **Normal Distribution:** As mentioned earlier, many natural phenomena follow a normal (or Gaussian) distribution, including many aspects of business and finance. If 'Sale Price' follows a normal distribution with a mean of $50k and standard deviation of $10k, for example, we know that most cars are sold close to the average price, with fewer cars sold at higher or lower prices. This information can be used to set expectations for future sales or to identify outliers (cars that are sold at unusually high or low prices).
2. **Binomial Distribution:** If we categorize cars as either 'Luxury' or 'Non-Luxury', the number of luxury cars sold could follow a binomial distribution. This distribution tells us the probability of selling a certain number of luxury cars, given the total number of cars sold and the probability of any one car being a luxury car. For example, if 30% of cars are luxury cars, and we sell 100 cars, the distribution can tell us how likely it is we sell exactly 20 luxury cars, or at least 50, and so on.
3. **Poisson Distribution:** If we're looking at the number of cars sold per day, this could follow a Poisson distribution if the sales are relatively rare and independent. For example, if we sell on average 2 cars per day, the Poisson distribution can tell us the probability of selling exactly 0, 1, 2, 3, etc., cars on any given day.

To calculate and interpret these distributions, a data scientist would first need to understand the underlying assumptions and properties of each distribution. They'd also need to be comfortable with the formulas for calculating probabilities and expectations under each distribution, or be familiar with software that can do these calculations.

In terms of decision-making, understanding your data's distribution can help you model and predict future outcomes. For instance, if you know car sales follow a Poisson distribution with an average of 2 cars sold per day, you can predict the probability of selling a certain number of cars in the future, which can help with planning and inventory management.

On the other hand, if you're trying to increase the sale of luxury cars and know that these sales follow a binomial distribution, you can model how changes in the underlying probability (e.g., through a targeted marketing campaign) might affect total sales.

In the case of a normal distribution of 'Sale Price', you can identify when a car is sold at an unusually high or low price (which might prompt further investigation) and make predictions about future sales.

In each case, understanding the distribution of your data enables you to make more informed decisions and predictions.

1. **Multivariate Analysis:** In a dataset, let's say we have two independent variables: Car's
2. **Age (X1) and Mileage (X2), and two dependent variables**: Sale Price (Y1) and Time on Market (Y2). A multivariate regression model may look like Y1 = a1 + b1X1 + c1X2 and Y2 =

a2 + b2X1 + c2X2. Here, 'a' is the intercept, and 'b' and 'c' are the coefficients for Age and Mileage respectively. These coefficients measure the change in Y1 and Y2 for a one-unit change in X.

3. **Time Series Analysis:** If car sales are 200, 220, 230, and 240 in four consecutive months, a simple time series model might be Sales = a + b*Month, where 'b' captures the monthly trend in sales.
4. **Survival Analysis:** Suppose five cars are sold in 10, 20, 30, 40, and 50 days respectively. Survival analysis might reveal that 80% of cars are sold by day 40, giving the survival function S(t) = e^(-t/λ), where λ is the mean time until sale.
5. **Machine Learning:** A simple decision tree might split cars based on age, then mileage, predicting different sale prices in each final group (or 'leaf'). More complex methods like random forests or neural networks involve more intricate computations.
6. **Principal Component Analysis (PCA):** Suppose three variables—Horsepower, Torque, and Engine Size—are all highly correlated. PCA might combine them into a single 'Engine Performance' factor that explains most of the variation in these variables.
7. **Factor Analysis:** Suppose we have survey data on car performance in several categories—Comfort, Speed, and Fuel Efficiency. Factor analysis might reveal these categories all load heavily on a latent 'Car Quality' factor, with loadings (correlations with the factor) of, say, 0.8, 0.7, and 0.9 respectively.
8. **Cluster Analysis:** Suppose we measure cars on two features—Fuel Efficiency and
9. **Power—and find two clear groups:** one with high efficiency and low power, and one with low efficiency and high power. These could be classified as 'economy' and 'performance' cars respectively.
10. **Design of Experiments:** Suppose we implement a new sales strategy on 50 randomly selected cars and use the old strategy on another 50. If the average sale price for the new strategy is $30,000 with a standard deviation of $5,000, and for the old strategy, it's $27,000 with a standard deviation of $4,000, we could conduct a t-test to see if the new strategy significantly increases sale prices.

In actual practice, these computations are done with the help of statistical software, as they usually involve larger datasets and more complex formulas. The interpretations of these formulas and the decisions based on these interpretations are where the real skills of data scientists come into play.

Advanced statistics concepts on Banking Case:

Suppose our dataset has the following observations (displayed as Age, Annual Income, Credit Score, Loan Amount, Interest Rate, Loan Term, Defaulted):

25	50000	700	20000	5	5	0
40	80000	750	35000	4	10	0
35	60000	650	15000	6	4	1
50	90000	730	40000	4.5	8	0
30	70000	710	25000	5.5	6	0
45	85000	740	30000	4.2	9	0
40	65000	680	20000	5.8	7	1
55	95000	780	45000	3.8	11	0
33	52000	690	22000	5.2	5	1
38	75000	720	28000	5	7	0
50	88000	770	38000	4.3	9	0
45	82000	760	32000	4.6	8	0
52	92000	740	40000	4.1	10	0
30	56000	700	20000	5.4	5	0
37	63000	680	24000	5.7	6	1
41	80000	760	36000	4.8	8	0
46	86000	730	30000	4.3	9	0
54	95000	780	44000	3.9	11	0
32	55000	690	21000	5.1	5	1
39	77000	710	27000	4.9	7	0

Now let's try to explain these concepts with some imaginary numbers:

1. **Multivariate Analysis:** Let's say we find the logistic regression equation predicting
2. **Defaulted from Age, Annual Income, and Credit Score to be:** P(Defaulted=1) = 1 / (1 + e^-(3 + 0.02Age - 0.00001Annual_Income - 0.01*Credit Score)). We could then plug in the values for a specific customer to predict their probability of default.
3. **Central Limit Theorem (CLT):** Let's assume we take many samples of 10 customers' Credit Scores. The average Credit Score across these samples will follow a normal distribution due to the CLT, regardless of the shape of the original distribution.
4. **Bayesian Statistics:** Suppose we have a prior belief that the mean Credit Score of customers who defaulted is 650 with a standard deviation of 20. After observing the credit scores of several defaulters, we can use Bayes' rule to update our belief and calculate a new posterior mean and standard deviation.
5. **Survival Analysis:** Here, the survival function could estimate the probability of a customer not defaulting after a certain number of years.
6. **Time Series Analysis:** Suppose we record the number of defaults each year. We could identify patterns, like an increase in defaults during economic recessions.

7. **Principal Component Analysis (PCA):** For example, we could find that the first principal component (PC1) is a weighted combination of Age, Annual Income, and Credit Score that captures the most variance in our dataset.

Design of Experiments: Suppose we randomly assign a new scoring system to half of the new customers and compare the default rates. If the p-value is less than 0.05, we may conclude the new system has a significant effect on default rates.

While these examples are oversimplified and the numbers are imaginary, they demonstrate how these concepts could be applied. In a real-world scenario, a data scientist would perform these analyses using statistical software, which would handle the complex calculations and assumptions checks.