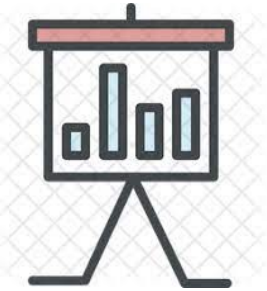
Typical Statistical Testing Procedures

Tushar B. Kute,
http://tusharkute.com



Statistical Testing

- The average business has radically changed over the last decade.
- Whether it's the equipment used at desks or the software used to communicate, very few things look the same as they once were.
- Something else that is completely different is how much data we have at our fingertips. What was once scarce is now a seemingly overwhelming amount of data. But, it's only overwhelming if you don't know how to analyze your business's data to find true and insightful meaning.
- So, how do you go from point A, having a vast amount of data, to point B, being able to accurately interpret that data? It all comes down to using the right methods for statistical analysis, which is how we process and collect samples of data to uncover patterns and trends.

Methods for performing statistical analysis

- Mean
- Standard deviation
- Regression
- Hypothesis Testing, and
- Sample size determination.

Mean

- The first method that's used to perform the statistical analysis is mean, which is more commonly referred to as the average.
- When you're looking to calculate the mean, you add up a list of numbers and then divide that number by the items on the list.
- When this method is used it allows for determining the overall trend of a data set, as well as the ability to obtain a fast and concise view of the data.
- Users of this method also benefit from the simplistic and quick calculation.

Mean

- The statistical mean is coming up with the central point of the data that's being processed. The result is referred to as the mean of the data provided.
- In real life, people typically use mean to in regards to research, academics, and sports.
- Think of how many times a player's batting average is discussed in cricket; that's their mean.

Mean

- To find the mean of your data, you would first add the numbers together, and then divide the sum by how many numbers are within the dataset or list.
- As an example, to find the mean of 6, 18, and 24, you would first add them together.

  6 + 18 + 24 = 48
- Then, divide by how many numbers in the list (3).

  48 / 3 = 16
- The mean is 16.

Problems with Mean

- When using mean is great, it’s not recommended as a standalone statistical analysis method.
- This is because doing so can potentially ruin the complete efforts behind the calculation, seeing as it is also related to the mode (the value that occurs most often) and median (the middle) in some data sets.
- When you’re dealing with a large number of data points with either a high number of outliers (a data point that differs significantly from others) or an inaccurate distribution of data, the mean doesn’t give the most accurate results in statistical analytics for a specific decision.

Standard Deviation

- Standard deviation is a method of statistical analysis that measures the spread of data around the mean.
- When you're dealing with a high standard deviation, this points to data that's spread widely from the mean.
- Similarly, a low deviation shows that most data is in line with the mean and can also be called the expected value of a set.
- Standard deviation is mainly used when you need to determine the dispersion of data points (whether or not they're clustered).

Standard Deviation

- Let’s say you’re a marketer who recently conducted a customer survey.
- Once you get the results of the survey, you’re interested in measuring the reliability of the answers in order to predict if a larger group of customers might have the same answers.
- If a low standard deviation occurs, it would show that the answers can be projected to a larger group of customers.

Standard Deviation

- The formula to calculate the standard deviation is:

  $\sigma 2 = \Sigma(x - \mu)2/n$
- In this formula:
  - The symbol for standard deviation is σ
  - Σ stands for the sum of the data
  - x stands for the value of the dataset
  - μ stands for the mean of the data
  - σ2 stands for the variance
  - n stands for the number of data points in the population

Standard Deviation

- To find the standard deviation:
  - Find the mean of the numbers within the data set
  - For each number within the data set, subtract the mean and square the result (which is this part of the formula $(x - \mu)2$).
  - Find the mean of those squared differences
  - Take the square root of the final answer
- If you used the same three numbers in our mean example, 6, 18, and 24, the standard deviation, or σ, would be 7.4833147735479.

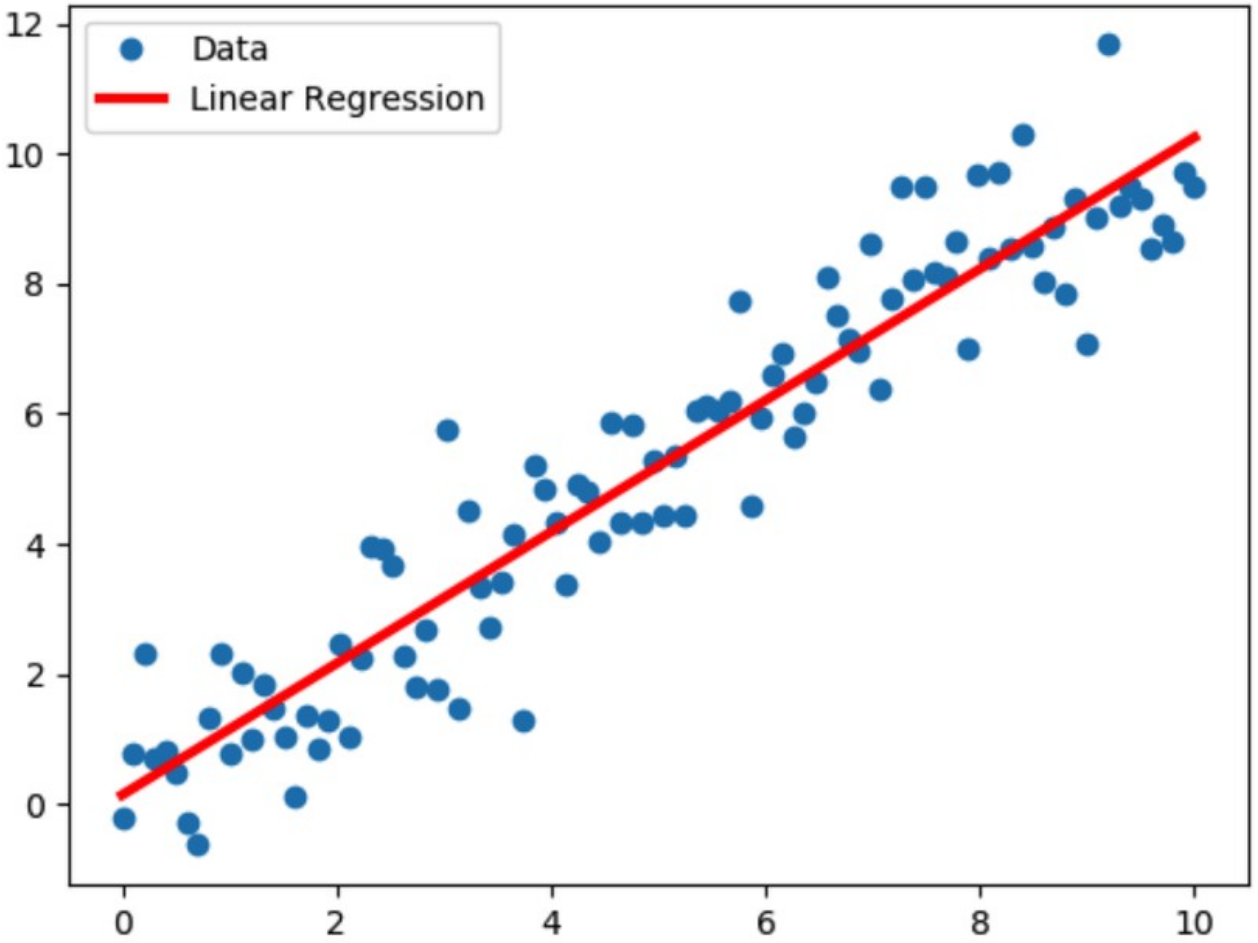
Standard Deviation – Problems

- On a similar note to the downside of using mean, the standard deviation can be misleading when used as the only method in your statistical analysis.
- As an example, if the data you're working with has too many outliers or a strange pattern like a non-normal curve, then standard deviation won't provide the necessary information to make an informed decision.

Regression

- When it comes to statistics, regression is the relationship between a dependent variable (the data you're looking to measure) and an independent variable (the data used to predict the dependent variable).
- It can also be explained by how one variable affects another, or changes in a variable that trigger changes in another, essentially cause and effect.
- It implies that the outcome is dependent on one or more variables.

Regression



Regression

- The line used in regression analysis graphs and charts signify whether the relationships between the variables are strong or weak, in addition to showing trends over a specific amount of time.
- These studies are used in statistical analysis to make predictions and forecast trends.
- For example, you may use regression to predict how a specific product or service may sell to your customers. Or, here at G2, we use regression to predict how our organic traffic will look 6 months from now.

Regression

- The regression formula that's used to see how data could look in the future is:

  $Y = a + b(x)$

- In this formula:
  - A refers to the y-intercept, the value of y when x = 0
  - X is the dependent variable
  - Y is the independent variable
  - B refers to the slope, or rise over run

Regression - Problems

- One disadvantage of using regression as part of your statistical analysis is that regression isn't very distinctive, meaning that although the outliers on a scatter plot (or regression analysis graph) are important, so are the reasons as to why they're outliers.
- This reason could be anything from an error in analysis to data being inappropriately scaled.
- A data point that is marked as an outlier can represent many things, such as your highest selling product. The regression line entices you to ignore these outliers and only see the trends in data.

Hypothesis Testing

- In statistical analysis, hypothesis testing, also known as “T Testing”, is a key to testing the two sets of random variables within the data set.
- This method is all about testing if a certain argument or conclusion is true for the data set. It allows for comparing the data against various hypotheses and assumptions.
- It can also assist in forecasting how decisions made could affect the business.

Hypothesis Testing

- In statistics, a hypothesis test determines some quantity under a given assumption.
- The result of the test interprets whether the assumption holds or whether the assumption has been violated.
- This assumption is referred to as the null hypothesis, or hypothesis 0.
- Any other hypothesis that would be in violation of hypothesis 0 is called the first hypothesis, or hypothesis 1.

Hypothesis Testing

- The results of a statistical hypothesis test need to be interpreted to make a specific claim, which is referred to as the p-value.
- Let's say what you're looking to determine has a 50% chance of being correct.
- The formula for this hypothesis test is:

  H0: P = 0.5

  H1: P ≠ 0.5

Hypothesis Testing – Problems

- Hypothesis testing can sometimes be clouded and skewed by common errors, like the placebo effect.
- This occurs when statistical analysts conducting the test falsely expect a certain result and then see that result, no matter the circumstances.
- There's also the likelihood of being skewed by the Hawthorne effect, otherwise known as the observer effect.
- This happens when participants being analyzed skew the results because they know they're being studied.

Sample Size Determination

- When it comes to analyzing data for statistical analysis, sometimes the dataset is simply too large, making it difficult to collect accurate data for each element of the dataset.
- When this is the case, most go the route of analyzing a sample size, or smaller size, of data, which is called sample size determination.

Sample Size Determination

- To do this correctly, you'll need to determine the right size of the sample to be accurate. If the sample size is too small, you won't have valid results at the end of your analysis.
- To come to this conclusion, you'll use one of the many data sampling methods.
- You could do this by sending out a survey to your customers, and then use the simple random sampling method to choose the customer data to be analyzed at random.

Sample Size Determination

- On the other hand, a sample size that is too large can result in wasted time and money.
- To determine the sample size, you may examine aspects like cost, time, or the convenience of collecting data.

Sample Size Determination

- Unlike the other four statistical analysis methods, there isn't one hard-and-fast formula to use to find the sample size.
- However, there are some general tips to keep in mind when determining a sample size:
  - When considering a smaller sample size, conduct a census
  - Use a sample size from a study similar to your own. For this, you may want to consider taking a look at academic databases to search for a similar study
  - If you're conducting a generic study, there may be a table that already exists that you can use to your advantage….

Sample Size Determination

- Continued...
  - Use a sample size calculator
  - Just because there isn’t one specific formula doesn’t mean you won’t be able to find a formula that works.
    - There are many you could use, and it depends on what you know or don't know about the purposed sample.
    - Some that you may consider using are Slovin’s formula and Cochran’s formula

Sample Size Determination - Problems

- As you analyze a new and untested variable of data within this method, you'll need to rely on certain assumptions.
- Doing so could result in a completely inaccurate assumption. If this error occurs during this statistical analysis method, it can negatively affect the rest of your data analysis.
- These errors are called sampling errors and are measured by a confidence interval.
- For instance, if you state that your results are at a 90% confidence level, it means if you were to perform the same analysis again and again, 90% of the time your results will be the same.

Which method to choose?

- No matter which method of statistical analysis you choose, make sure to take special note of each potential downside, as well as their unique formula.
- Of course, there's no gold standard or right or wrong method to use.
- It's going to depend on the type of data you've collected, as well as the insights you're looking to have as an end result.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com