تكليف بحث المصطلحات

مادة تنقيب بيانات

منال السلال االقادري
cs_4_AM

1. Skewness

Definition:

is a statistical measure that quantifies the asymmetry of a probability distribution. It measures the deviation or departure from symmetry in a dataset.

Calculation:

Skewness is typically calculated using the third standardized moment of a distribution. The formula to calculate skewness is:

Skewness = (٣ * (Mean - Median)) / Standard Deviation

Interpretation:

Skewness provides information about the shape of the distribution's tail and indicates whether the data is skewed to the left or right.

1. Positive Skewness:

   If the skewness value is positive, it indicates that the distribution has a longer tail on the right side (right-skewed) or more positive outliers. In a positively skewed distribution, the mean is typically greater than the median, and the tail extends towards higher values.

2. Negative Skewness:

   If the skewness value is negative, it indicates that the distribution has a longer tail on the left side (left-skewed) or more negative outliers. In a negatively skewed distribution, the mean is typically less than the median, and the tail extends towards lower values.

3. Zero Skewness:

   If the skewness value is zero, it means the data is perfectly symmetrical. In a symmetric distribution, the mean and median are equal, and the data is evenly distributed around the center.

Applications:

Skewness is commonly used in various fields for different purposes:

1. Data Analysis: Skewness helps in understanding the shape of a distribution and detecting departures from normality. It is useful in exploratory data analysis to identify potential outliers and understand the distributional characteristics of a dataset.

2. Finance: Skewness is used to assess the risk and return characteristics of financial assets. Skewed returns can impact investment strategies, portfolio construction, and risk management decisions.

3. Economics: Skewness is employed in economic research to analyze income distributions, wealth distributions, and other economic variables. It provides insights into inequality and can help in policy-making decisions.

4. Risk Management: Skewness is used in risk models to capture the asymmetry and tail behavior of asset returns. It is particularly relevant in financial risk analysis, such as Value-at-Risk (VaR) estimation.

5. Portfolio Optimization: Skewness is considered in portfolio optimization models to account for the non-normality of asset returns. It helps in constructing efficient portfolios that consider both risk and return characteristics.

Limitations:

It's important to note some limitations of skewness:

1. Skewness is sensitive to outliers: Skewness can be influenced by extreme values or outliers in the dataset. Outliers can disproportionately impact the skewness measure, leading to misinterpretation.

2. Skewness does not capture all aspects of distribution: Skewness provides information about the asymmetry of a distribution but does not convey details about other characteristics such as kurtosis (peakedness) or specific shape patterns.

3. Interpretation depends on context: The interpretation of skewness values may vary depending on the context and the specific dataset. It is crucial to consider the domain knowledge and the underlying data generating process.

In summary, skewness is a statistical measure that quantifies the asymmetry of a probability distribution. It helps in understanding the shape of a distribution, detecting departures from normality, and assessing the risk and return characteristics of data. However, it has limitations and should be interpreted carefully in conjunction with other statistical measures and domain knowledge.

Let's consider a dataset of example scores for a class of students :

٧٨، ٨٢، ٨٥، ٨٨، ٩٠، ٩٢، ٩٥، ٩٨، ١٠٠

To calculate the skewness, we'll follow these steps:

Step ١: Calculate the mean:

Mean = (٧٨ + ٨٢ + ٨٥ + ٨٨ + ٩٠ + ٩٢ + ٩٥ + ٩٨ + ١٠٠) / ٩ = ٨٩,٣٣

Step ٢: Calculate the median:

The dataset has an odd number of values, so the median is the middle value, which is ٩٠.

Step ٣: Calculate the standard deviation:

To calculate the standard deviation, we first need to calculate the variance. The variance is the average of the squared differences from the mean.

Variance = ((٧٨ - ٨٩,٣٣)^٢ + (٨٢ - ٨٩,٣٣)^٢ + (٨٥ - ٨٩,٣٣)^٢ + (٨٨ - ٩٠) + ٢^(٨٩,٣٣ - ٩٢) + ٢^(٨٩,٣٣ - ٩٥) + ٢^(٨٩,٣٣ - ٩٨) + ٢^(٨٩,٣٣ - ١٠٠)) / ٩

= (١٣٥,١١ + ١٢,٤٤ + ١٦,٨٩ + ١,٧٨ + ٠,١١ + ٥,٧٨ + ٣١,٥٦ + ٦٦,٤٤ + ٦٧,١١) / ٩

= ٢١,٥٦

Standard Deviation = sqrt(Variance) = sqrt(٢١,٥٦) ≈ ٤,٦٤

Step ٤: Calculate the skewness:

Skewness = (٣ * (Mean - Median)) / Standard Deviation

= (٣ * (٨٩,٣٣ - ٩٠)) / ٤,٦٤

= -٠,٣٩

In this example, the skewness value is approximately -٠,٣٩, Since the skewness value is negative, it indicates that the distribution is left-skewed, meaning it has a longer tail on the left side.

2. Variance

is a statistical measure that quantifies the spread or dispersion of a dataset. It measures how far each data point in the set is from the mean and provides an indication of the variability or scatter of the data points.

Definition:

Variance is the average of the squared differences between each data point and the mean of the dataset. It is calculated using the following formula:

Variance =)∑)xi - μ)²) / N

Where:

xi - represents each data point in the dataset

μ - represents the mean of the dataset

N - represents the total number of data points

Interpretation:

Variance provides information about the dispersion or spread of the data points around the mean. A higher variance indicates that the data points are more spread out, while a lower variance suggests that the data points are closer to the mean.

Key Points:

1. Squaring of Differences: Variance involves squaring the differences between each data point and the mean. This is done to avoid cancellation of positive and negative deviations and to focus on the magnitude of the differences
2. Units: Variance is expressed in squared units of the original data. For example, if the original data represents measurements in meters, the variance will be in square meters.

3. Non-Negative Value: Variance is always a non-negative value since it is the sum of squared differences. It is zero when all data points are the same (no variability), and it increases as the data points become more spread out.

4. Influence of Outliers: Variance is sensitive to outliers or extreme values in the dataset. Outliers can have a significant impact on the variance calculation, as their squared differences from the mean can be large.

Applications:

Variance is widely used in various fields for different purposes:

1. Descriptive Statistics: Variance is a fundamental measure used in .descriptive statistics to summarize the variability of a dataset

2. Risk and Volatility: In finance and investment, variance is used to measure the risk and volatility of financial assets or portfolios. Higher .variance indicates higher risk and potential for larger price fluctuations

3. Quality Control: Variance is used in quality control to assess the variation in manufacturing processes. It helps in identifying whether the process is consistent or whether there is excessive variability in the .output

4. Experimental Design: Variance is used in experimental design to analyze the variability and assess the significance of differences .between treatment groups or factors

5. Machine Learning: Variance is used in various machine learning algorithms to measure the importance or relevance of features in a .dataset. It helps in feature selection and model building

Limitations:

It's important to consider some limitations of variance:

1. Sensitive to Scale: Variance is affected by the scale or units of measurement of the data. For example, if the data is measured in different units, such as weight and height, the variance may be dominated by the larger-scale variable.

2. Not Robust to Outliers: The presence of outliers can significantly impact the variance calculation, making it less robust to extreme values.

3. Difficult to Interpret: Variance is expressed in squared units, which can make it difficult to interpret in the original scale of the data. Taking the square root of variance gives the standard deviation, which is in the original units and is often preferred for interpretation.

In summary, variance is a statistical measure that quantifies the spread or dispersion of a dataset. It provides insights into the variability or scatter of data points around the mean. Variance is used in descriptive statistics, risk assessment, quality control, experimental design, and machine learning. However, it has limitations and should be interpreted carefully, considering the scale of the data and the presence of outliers.

variance is calculated in a dataset:

Let's consider a dataset of exam scores for a class of students:

٨٥، ٩٠، ٩٢، ٨٨، ٩٥

To calculate the variance, we'll follow these steps:

Step ١: Calculate the mean:

Mean = (٨٥ + ٩٠ + ٩٢ + ٨٨ + ٩٥) / ٥ = ٩٠

Step ٢: Calculate the squared differences from the mean:

Squared difference for each data point = (Data point - Mean)^٢

Squared differences:

٢٥ = ٢^(٩٠ - ٨٥)

٠ = ٢^(٩٠ - ٩٠)

٤ = ٢^(٩٠ - ٩٢)

٤ = ٢^(٩٠ - ٨٨)

٢٥ = ٢^(٩٠ - ٩٥)

Step ٣: Calculate the sum of squared differences:

Sum of squared differences = ٢٥ + ٠ + ٤ + ٤ + ٢٥ = ٥٨

Step ٤: Calculate the variance:

Variance = Sum of squared differences / Number of data points

٥ / ٥٨ =

١١,٦ =

In this example, the variance of the dataset is ١١,٦, The variance represents the average squared difference between each data point and the mean. It indicates the spread or dispersion of the data points around the mean. A higher variance suggests greater variability or scatter in the dataset.

specific range of values that is considered "high" or "low" for the variance:

There is no specific range of values that universally defines what is considered "high" or "low" variance. The interpretation of whether a variance value is high or low depends on the context and the nature of the data being analyzed.

we can make some general observations :

1. Comparing Variances: To assess whether a variance value is high or low, it is often helpful to compare it with the variances of other datasets of similar nature or within the same domain. By comparing variances, you can get a relative understanding of the spread of the data.

2. Magnitude Relative to Mean: The magnitude of the variance can be considered in relation to the mean of the dataset. If the variance is relatively small compared to the mean, it suggests that the data points are tightly clustered around the mean, indicating low variability. Conversely, if the variance is relatively large compared to the mean, it suggests a wider spread of data points, indicating high variability.

3. Data Characteristics: The interpretation of what constitutes a high or low variance also depends on the characteristics of the data being analyzed. For example, in some domains, such as financial markets or stock returns, higher variance values may be expected due to the inherent volatility of the data. On the other hand, in controlled experiments or quality control processes, low variance values may be desirable to indicate consistency and precision.

In summary, the determination of what is considered a high or low variance value is subjective and context-dependent. It is often helpful to compare variances with other datasets or consider the magnitude relative to the mean and the specific characteristics of the data domain.

some techniques for comparing variances between different datasets:

There are several techniques you can use to compare variances between different datasets. Here are a few commonly employed methods:

1. Coefficient of Variation (CV): The coefficient of variation is a standardized measure that compares the variability (standard deviation or variance) of datasets relative to their means. It is calculated as the ratio of the standard deviation to the mean, multiplied by ١٠٠ to express it as a percentage. By comparing the CV values of different datasets, you can assess and compare their relative variability. A lower CV suggests less relative variability, while a higher CV indicates greater relative variability.

2. F-Test: The F-test is a statistical test that compares the variances of two datasets to determine if they are significantly different. It calculates the F-statistic by dividing the larger variance by the smaller variance. The F-statistic is then compared to a critical value based on the degrees of freedom to assess the significance of the difference in variances. If the calculated F-statistic exceeds the critical value, it suggests that the variances are significantly different.

3. Box Plots: Box plots provide a visual representation of the distribution of data and can be used to compare variances. By comparing the widths of the boxes (interquartile ranges) or the lengths of the whiskers, you can get an idea of the variability and compare the spread of different datasets. Wider boxes or longer whiskers indicate greater variability.

4. Levene's Test: Levene's test is a statistical test used to compare the equality of variances between multiple groups or datasets. It calculates a test statistic based on the absolute deviations of the data from their group means and assesses the significance of the differences in variances. If the p-value of Levene's test is below a pre-defined significance level (e.g., ٠٫٠٥)، it suggests that the variances are significantly different between the groups.

5. Confidence Intervals: Confidence intervals can also be used to compare variances. By calculating the confidence intervals for the variances of different datasets, you can assess whether the intervals overlap or not. Non-overlapping confidence intervals indicate a significant difference in the variances, while overlapping intervals suggest no significant difference.

It's important to note that the choice of technique depends on the specific context and nature of the data. The suitability of a particular method will vary based on the type of data, sample size, and the research question at hand. It's always advisable to consult with a statistician or refer to appropriate statistical textbooks or references when selecting and applying these techniques.

Uses

Variance has various practical uses across different fields. Here are some common applications of variance:

1. Descriptive Statistics: Variance is a fundamental measure used in descriptive statistics to quantify the variability or spread of data points around the mean. It provides insights into the dispersion of the data and helps summarize the dataset.

2. Quality Control: Variance is used in quality control to assess the variation in manufacturing processes. By calculating the variance of a specific process parameter, one can determine whether the process is consistent and within acceptable limits. A lower variance indicates less variability and better quality control.

3. Risk Assessment: In finance and investment, variance is widely used to measure risk. In portfolio management, the variance of asset returns is calculated to analyze the volatility or price fluctuations of investments. Higher variance indicates higher risk, and investors often consider it when making investment decisions.

4. Experimental Design: Variance is utilized in experimental design to analyze the variability and assess the significance of differences between treatment groups or factors. By comparing the variances, researchers can determine if the observed differences are statistically significant or occur due to random chance.

5. Machine Learning: Variance is employed in various machine learning algorithms and techniques. For example, in feature selection, variance is used as a criterion to measure the importance or relevance of features in a dataset. Features with low variance may be considered less informative and can be excluded from the model.

6. Process Optimization: Variance analysis is commonly used in process optimization to identify sources of variation and reduce them. By analyzing the variance components in a process, practitioners can identify areas for improvement and implement strategies to minimize variability, leading to enhanced efficiency and quality.

7. Comparison of Data Sets: Variance is often used to compare the spread or variability of different datasets. By comparing the variances, one can assess which dataset has greater or lesser variability. This is useful in various fields, such as comparing the performance of different groups, analyzing the variability of measurements, or assessing the stability of processes over time.

8. Sensitivity Analysis: In sensitivity analysis, variance is employed to assess the sensitivity of model outputs to variations in input parameters. By analyzing the variance of model outputs for different input scenarios, one can identify the most influential input factors and their effect on the overall system.

These are just a few examples of how variance is applied in different domains. Variance is a versatile statistical measure that aids in understanding the dispersion and variability of data, enabling informed decision-making and analysis in various fields.

3. Deviation

is a statistical concept that measures the dispersion or spread of data points around a central value, typically the mean. Deviation provides information about the variability or distance of individual data points from the central tendency. There are two common types of deviation: absolute deviation and standard deviation.

1. Absolute Deviation

The absolute deviation measures the absolute difference between each data point and the central value (e.g., mean or median). Here's how to calculate the absolute deviation for a dataset:

Step ١: Calculate the central value (mean or median)

Step ٢: For each data point, calculate the absolute difference between the data point and the central value

Step ٣: Sum the absolute differences

Step ٤: Divide the sum by the number of data points to calculate the average absolute deviation.

2. Standard Deviation

The standard deviation is a widely used measure of deviation that takes into account the squared differences between each data point and the central value. It provides a single value that represents the overall dispersion of the data. Here's how to calculate the standard deviation for a dataset:

Step ١: Calculate the central value (mean or median)

Step ٢: For each data point, calculate the squared difference between the data point and the central value

Step ٣: Sum the squared differences

Step ٤: Divide the sum by the number of data points minus one (for a sample) or by the number of data points (for a population)

Step ٥: Take the square root of the result to obtain the standard deviation

The standard deviation is widely used because it has several desirable properties, including:

- It considers all data points in the calculation, providing a comprehensive measure of dispersion.
- It is expressed in the same units as the original data.
- It is sensitive to outliers, giving them more weight in the calculation compared to the absolute deviation.
- It is commonly used in various statistical analyses and modeling techniques.

The concept of deviation is closely related to variance, which is the squared value of the standard deviation. Variance and standard deviation are both measures of dispersion, with the standard deviation being the more commonly used measure due to its additional interpretability.

In summary, deviation is a statistical concept that quantifies the dispersion or spread of data points around a central value. It provides valuable information about the variability of data and is used in various statistical analyses, decision-making processes, and modeling techniques.

Uses

Deviation, specifically standard deviation, has numerous practical uses in various fields. some common applications of deviation:

1. Descriptive Statistics: Deviation, particularly standard deviation, is widely used in descriptive statistics to provide a summary measure of the dispersion or spread of a dataset. It helps to understand the variability of the data points around the mean or median, providing insights into the distribution of the data.

2. Risk Assessment and Finance: In finance and investment, deviation is crucial for measuring risk. Standard deviation is commonly used as a measure of volatility or price fluctuations in asset returns. It helps investors assess the potential variability or risk associated with an investment and make informed decisions.

3. Quality Control and Process Improvement: Deviation is employed in quality control to assess the consistency and stability of manufacturing processes. By monitoring the standard deviation of key process parameters, practitioners can identify variations and take corrective actions to improve product quality and reduce defects.

4. Performance Evaluation: Deviation is utilized in performance evaluation and benchmarking. For example, in sports analytics, standard deviation can be used to measure the consistency or variability of player performance over a season or across different seasons. It helps identify outliers and assess the overall performance of individuals or teams.

5. Experimental Design and Research: Deviation plays a role in experimental design and research. By analyzing the standard deviation of data from different treatment groups or conditions, researchers can assess the variability within groups and determine the statistical significance of observed differences. It helps in drawing reliable conclusions and making valid comparisons.

6. Process Control and Monitoring: Deviation is used in process control to monitor the performance of systems or processes over time. By continuously measuring the standard deviation of process variables, deviations from expected values can be identified, allowing for timely adjustments and maintenance.

7. Quality Assurance and Six Sigma: Deviation is a key concept in quality assurance methodologies such as Six Sigma. By analyzing the standard deviation of product characteristics or process outputs, practitioners can evaluate process capability, identify sources of variation, and implement strategies for quality improvement.

8. Machine Learning and Data Analysis: Deviation is employed in various machine learning algorithms and statistical analysis techniques. For instance, it is used in anomaly detection to identify data points that deviate significantly from the expected patterns. Deviation measures also play a role in feature selection and feature importance ranking in machine learning models.

These are just a few examples of how deviation, particularly standard deviation, is used in different fields. Deviation provides valuable insights into the variability and spread of data, enabling better decision-making, risk assessment, process control, and quality improvement in diverse domains.

how standard deviation is calculated:

The standard deviation is a measure of dispersion that quantifies the spread or variability of a dataset. It is calculated using the following steps

1. Calculate the Mean: First, calculate the mean (average) of the dataset. Add up all the values in the dataset and divide the sum by the total number of data points. Let's denote the mean as μ.
2. Calculate the Difference: For each data point, calculate the difference between the value and the mean (x - μ) for each data point in the dataset.
3. Square the Differences: Square each difference obtained in the previous step [(x - μ)^٢] for each data point.
4. Sum the Squares: Add up all the squared differences.
5. Divide by the Sample Size: If you are working with a sample (a subset of a larger population), divide the sum of squared differences by (n - ١), where n represents the sample size. This is known as Bessel's correction and is used to provide an unbiased estimate of the population standard deviation.
6. Calculate the Square Root: Take the square root of the result obtained in the previous step. This gives you the standard deviation.

Mathematically, the formula for calculating the standard deviation (σ) is:

σ = √])Σ(x - μ)^٢) / n-١]

Where:

σ - represents the standard deviation.

Σ - denotes the sum of the squared differences.

(x - μ) - represents the difference between each data point and the mean .

(n-١) - is the sample size minus one, used for sample standard deviation calculation.

The standard deviation provides a measure of the average amount of deviation or dispersion of data points from the mean. A smaller standard deviation indicates that the data points are closely clustered around the mean, while a larger standard deviation suggests greater variability or spread in the .data

specific range of values that is considered a "good" or "bad" standard deviation:

There is no specific range of values that can universally be considered as "good" or "bad" for standard deviation. The interpretation of whether a particular standard deviation is "good" or "bad" depends on the context and the specific application.

In some cases, a smaller standard deviation may be desirable, indicating that the data points are closely clustered around the mean. This can suggest a higher level of precision, consistency, or stability. For example, in quality control, a smaller standard deviation may indicate that a manufacturing process is producing consistent and reliable output.

On the other hand, a larger standard deviation may be more acceptable or even expected in certain situations. For instance, in financial markets, higher volatility is often associated with higher potential returns. In this case, a larger standard deviation may reflect the inherent risk or variability of an investment.

The appropriateness of a standard deviation value also depends on the specific domain and the characteristics of the dataset being analyzed. What is considered a "good" or "bad" standard deviation can vary significantly depending on the field of study, the nature of the data, and the specific research or business goals.

It is often more meaningful to compare the standard deviation to other relevant benchmarks or reference points. This can include comparing the standard deviation of a dataset to historical values, industry standards, or established norms.

example how to calculate standard deviation:

Let's consider the following dataset of exam scores for a class of students:

٨٠، ٨٥، ٩٠، ٧٥، ٩٥

To calculate the standard deviation, follow these steps:

Step ١: Calculate the Mean

Add up all the values in the dataset and divide by the total number of (data points (in this case, ٥

Mean (μ) = (٨٠ + ٨٥ + ٩٠ + ٧٥ + ٩٥) / ٥ = ٤٢٥ / ٥ = ٨٥

Step ٢: Calculate the Difference

For each data point, subtract the mean from the value:

٥- = ٨٥ - ٨٠

٠ = ٨٥ - ٨٥

٥ = ٨٥ - ٩٠

١٠- = ٨٥ - ٧٥

١٠ = ٨٥ - ٩٥

Step ٣: Square the Differences

Square each difference obtained in the previous step:

٢٥ = ٢٨(٥-)

٠ = ٢٨٠

٢٥ = ٢٨٥

١٠٠ = ٢٨(١٠-)

١٠٠ = ٢٨١٠

Step ٤: Sum the Squares

Add up all the squared differences:

٢٥٠ = ١٠٠ + ١٠٠ + ٢٥ + ٠ + ٢٥

Step ٥: Divide by the Sample Size

Since we are working with a sample, divide the sum of squared differences by (n - ١)، where n is the sample size (٥ in this case):

٦٢,٥ = ٤ / ٢٥٠ = (١ - ٥) / ٢٥٠

Step ٦: Calculate the Square Root

Take the square root of the result obtained in the previous step:

٧,٩١ ≈ √٦٢,٥

Therefore, the standard deviation of the exam scores in this dataset is approximately ٧,٩١.

example how to calculate standard . Variance:

To calculate the variance, you can follow these steps:

1. Calculate the Mean: Calculate the mean (average) of the dataset by adding up all the values and dividing the sum by the total number of data points.
2. Calculate the Difference: For each data point, subtract the mean calculated in the previous step from the value of that data point.
3. Square the Differences: Square each difference obtained in the previous step.
4. Sum the Squares: Add up all the squared differences obtained in the previous step.
5. Divide by the Sample Size: If you are working with a sample (a subset of a larger population), divide the sum of squared differences by (n - ١)، where n represents the sample size. This is known as Bessel's correction and is used to provide an unbiased estimate of the population variance. If you have the entire population, divide by the total number of data points (n) without subtracting ١.

Let's calculate the variance using the following dataset of exam scores for a class of students:

٨٠، ٨٥، ٩٠، ٧٥، ٩٥

Step ١: Calculate the Mean

Mean (μ) = (٨٠ + ٨٥ + ٩٠ + ٧٥ + ٩٥) / ٥ = ٤٢٥ / ٥ = ٨٥

Step ٢: Calculate the Difference

٥- = ٨٥ - ٨٠

٠ = ٨٥ - ٨٥

٥ = ٨٥ - ٩٠

٧٥ - ٨٥ = -١٠

٩٥ - ٨٥ = ١٠

Step ٣: Square the Differences

(-٥)^٢ = ٢٥

٨٠^٢ = ٠

٨٥^٢ = ٢٥

(-١٠)^٢ = ١٠٠

١٠^٢ = ١٠٠

Step ٤: Sum the Squares

٢٥ + ٠ + ٢٥ + ١٠٠ + ١٠٠ = ٢٥٠

Step ٥: Divide by the Sample Size

(٥ - ١) / ٢٥٠ = ٤ / ٢٥٠ = ٦٢,٥

Therefore, the variance of the exam scores in this dataset is ٦٢,٥.