stats101

alexander

May 29, 2025

Statistics is the study of how to gather, summarize, and draw conclusions from data, often in presence of uncertainty.

A key distinction in statistics is between inference (making conclusions about a larger population) and description (summarizing and describing data). Inferential statistics involves making educated guesses based on sample data, while descriptive statistics provide summaries and descriptions of data without any attempt to make inferences.

data: a collection of values, characteristics, or observations that can be analyzed and interpreted.
raw data: unprocessed and unstructured data collected from various sources.

types:

- quantitative: numerical data that can be measured, such as heights or weights
- qualitative: non-numerical data that cannot be measured, such as text or categorical labels
- categorical: data organized into categories or groups, such as colors or brands

characteristics:

- scalability: the ability to increase or decrease the size of the dataset without affecting its structure
- completeness: the presence or absence of all relevant data points in the dataset
- consistency: the uniformity and accuracy of the data within the dataset

sources:

- primary: original data collected directly from sources, such as surveys or experiments

- secondary: existing data that is reused or repurposed, such as publicly available datasets or literature reviews
- external: data sourced from external organizations or systems, such as social media platforms or government databases

quality:

- validity: the degree to which the data accurately represents the real world
- reliability: the consistency and accuracy of the data over time
- accuracy: the closeness of a measured value to its true value

Central tendency is a statistical measure that describes the "center" or typical value of a dataset. It provides a single value that best represents the middle or avaerage of a set of data. measures of central tendency:

- mean: the average value of a dataset
- median: the middle value of a dataset when sorted in ascending or descending order
- mode: the most frequiently occurring value in a dataset (if there is one)

Variability refers to the amount of scatter or dispersion present in a dataset. It measures how much individual data poiunt deviate from the central tendency. In other words, it describes the spread or range of values within the dataset. measures of variability:

- range: the difference between the largest and smallest values in a dataset
- variance: the average of the squared differences from the mean
- standard deviation: a measure of the spread or dispersion of a dataset
- skewness: a measure of the asymmetry of the distribution, with positive skew indicating a longer tail to the right and negative skew indicating a longer tail to the left.
- kurtosis: a measure of the "tailedness" or "peakedness" of the distribution

inferences:

- descriptive inference: making statements about the caracteristics of a population based on a sample, such as estimating a population mean or proportion
- analytical inference: drawing conclusions about the relationships between variables within a population

procedures:

- hypothesis testing: formulating hypotheses and testing them against data to determine if they are supported or rejected
- confidence intervals: constructing intervals around sample estimates that have a specific level of confidence for containing the true population parameter
- regression analysis: modeling relationships between variables within a population

assumptions:

- independence: observations are independent are not influenced by each other
- normality: the distribution of residuals or errors follows a normal distribution
- homoscedasticity: the variance of residuals is constant across all levels of predictor variables

limitations:

- sampling error: errors that occur due to the fact that only a sample is used instead of the entire population
- biased sampling: samples are not representative of the population or have some inherent flaws.
- limited generalizability: inferences may be restricted to the specific context and conditions of the study.