Statistics Notes

Some general terms in statistics are:

1. **Population**: The entire group of individuals or instances about whom we hope to learn.
2. **Sample**: A subset of the population selected for observation and analysis.
3. **Variable**: Any characteristic, number, or quantity that can be measured or quantified. Variables can be classified as:
   * **Independent Variable**: The variable that is manipulated or controlled in an experiment.
   * **Dependent Variable**: The variable that is observed and measured to determine the effect of the independent variable.
4. **Data**: Information collected for analysis. Data can be:
   * **Quantitative Data**: Numerical data that can be measured.
   * **Qualitative Data**: Descriptive data that can be observed but not measured.
5. **Mean**: The average of a set of numbers, calculated by summing all the values and dividing by the number of values.
6. **Median**: The middle value in a set of numbers arranged in ascending or descending order.
7. **Mode**: The value that appears most frequently in a data set.
8. **Range**: The difference between the highest and lowest values in a data set.
9. **Variance**: A measure of how much the values in a data set differ from the mean. It is the average of the squared differences from the mean.
10. **Standard Deviation**: A measure of the amount of variation or dispersion in a set of values, calculated as the square root of the variance.
11. **Probability**: The likelihood or chance of an event occurring, ranging from 0 (impossible) to 1 (certain).
12. **Normal Distribution**: A bell-shaped probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence.
13. **p-value**: The probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true.
14. **Confidence Interval**: A range of values, derived from a sample, that is likely to contain the value of an unknown population parameter.
15. **Hypothesis Testing**: A method of making decisions using data, whether to reject the null hypothesis or not. It involves:
    * **Null Hypothesis (H0)**: A statement that there is no effect or no difference.
    * **Alternative Hypothesis (H1 or Ha)**: A statement that there is an effect or a difference.
16. **Correlation**: A measure of the strength and direction of a linear relationship between two variables.
17. **Regression**: A statistical method used to determine the relationship between a dependent variable and one or more independent variables.
18. **ANOVA (Analysis of Variance)**: A statistical technique used to compare the means of three or more samples to see if at least one mean is different from the others.
19. **Chi-Square Test**: A statistical test used to determine if there is a significant association between categorical variables.
20. **Bias**: Systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others.
21. **Outlier**: An observation point that is distant from other observations, which can distort statistical analyses.

Statistics play a crucial role in researching academic topics by providing tools and methods to collect, analyze, interpret, and present data. Here are some ways statistics are used in academic research:

1. **Designing Experiments and Studies**:
   * **Sample Selection**: Determining the appropriate sample size and sampling method to ensure the sample is representative of the population.
   * **Randomization**: Ensuring subjects are randomly assigned to different groups to reduce bias.
2. **Data Collection and Analysis**:
   * **Descriptive Statistics**: Summarizing and describing the main features of a dataset. This includes measures such as mean, median, mode, range, and standard deviation.
   * **Inferential Statistics**: Making inferences and predictions about a population based on a sample of data. This involves hypothesis testing, confidence intervals, and significance testing.
3. **Testing Hypotheses**:
   * **Null and Alternative Hypotheses**: Formulating and testing hypotheses to determine if there is a significant effect or relationship in the data.
   * **p-values**: Assessing the strength of the evidence against the null hypothesis.
   * **Confidence Intervals**: Estimating the range within which a population parameter lies with a certain level of confidence.
4. **Modeling Relationships**:
   * **Correlation and Regression Analysis**: Understanding the relationships between variables. Correlation measures the strength and direction of a relationship, while regression models the relationship and makes predictions.
   * **Multivariate Analysis**: Analyzing more than two variables simultaneously to understand complex relationships.
5. **Evaluating Reliability and Validity**:
   * **Reliability**: Ensuring that the results are consistent and reproducible.
   * **Validity**: Ensuring that the study measures what it is intended to measure.
6. **Interpreting Results**:
   * **Data Visualization**: Using graphs, charts, and plots to present data in a clear and understandable way.
   * **Statistical Significance**: Determining whether the results are likely due to chance or represent a true effect.
7. **Drawing Conclusions**:
   * **Generalization**: Extending the findings from the sample to the larger population.
   * **Policy and Decision Making**: Using statistical evidence to inform decisions, develop policies, and guide further research.

Some concepts involved in statistics include:

1. **Descriptive Statistics**:
   * **Mean**: The average of a set of numbers.
   * **Median**: The middle value in a set of numbers.
   * **Mode**: The most frequently occurring value in a set of numbers.
   * **Range**: The difference between the highest and lowest values.
   * **Standard Deviation**: A measure of the amount of variation or dispersion in a set of values.
2. **Inferential Statistics**:
   * **Population vs. Sample**: The entire group you're interested in vs. a subset of that group.
   * **Hypothesis Testing**: Procedure to test if a hypothesis about a population parameter is true.
   * **Confidence Intervals**: A range of values that is likely to contain a population parameter with a certain level of confidence.
   * **p-value**: The probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis is true.
3. **Probability**:
   * **Basic Probability**: The likelihood of an event occurring.
   * **Conditional Probability**: The probability of an event occurring given that another event has occurred.
   * **Independent and Dependent Events**: Whether the occurrence of one event affects the probability of another.
4. **Distributions**:
   * **Normal Distribution**: A bell-shaped distribution that is symmetrical about the mean.
   * **Binomial Distribution**: The distribution of the number of successes in a fixed number of trials.
   * **Poisson Distribution**: The distribution of the number of events occurring in a fixed interval of time or space.
5. **Correlation and Regression**:
   * **Correlation**: Measures the strength and direction of a relationship between two variables.
   * **Linear Regression**: A method for modeling the relationship between a dependent variable and one or more independent variables.
6. **Data Collection and Sampling**:
   * **Random Sampling**: Every member of the population has an equal chance of being selected.
   * **Stratified Sampling**: The population is divided into strata and a random sample is taken from each stratum.
   * **Bias**: Systematic error introduced into sampling or testing.
7. **Data Visualization**:
   * **Histograms**: Used to represent the frequency distribution of a set of data.
   * **Box Plots**: Displays the distribution of data based on a five-number summary.
   * **Scatter Plots**: Shows the relationship between two variables.

**Notes from Statistics for Absolute Beginners**

Every bit of data has a story to tell. But, left isolated, these parcels of information rest dormant and underutilized.

While primary methods of statistical analysis date back to at least the 5th Century BC, it wasn’t until the 18th Century AD that these and newly evolved methods coalesced into a distinctive sub-field of mathematics and probability known today as statistics.

Derived from the Latin stem "sta," meaning "to stand, set down, make or be firm," the field of statistics was initially limited to policy discussions and the condition of the state.

A large number of people study statistics as part of their education, but very few know how to apply these methods past examination day despite an inherent curiosity and interest in measuring things and especially performance.

There are four major categories of statistical measures used to describe data:

* Measures of frequency analyze how often each value appears in a dataset, such as counting Democrat and Republican voters in a sample population.
* Measures of Central Tendency describe the "middle" values of a dataset, including median and mode.
* Measures of Spread show how varied the data is, e.g. standard deviation.
* Measures of Position show where a value stands in a dataset.

As a popular branch of mathematics, statistics involves studying how data is collected, organized, analyzed, interpreted, and presented. The goal of statistics is to determine the meaning of the data and whether variations, if any, are meaningful or due merely to chance.

Descriptive statistics helps to organize data and provides a summary of data features numerically and/or graphically. Typical methods of descriptive statistics include the mode (most common value), mean (average value), standard deviation (variance), and quartiles. ﻿In general, descriptive statistics helps give you a better sense of your data and may be used in advance of inferential methods.

In inferential statistics, there isn’t the luxury of a full population. The analysis is instead subject to the nuances of probability theory, dealing with random phenomena, and inferring what is likely based on what is already known to be true.

As a critical distinction from inferential statistics, descriptive statistical analysis applies to scenarios where all values in the dataset are known.

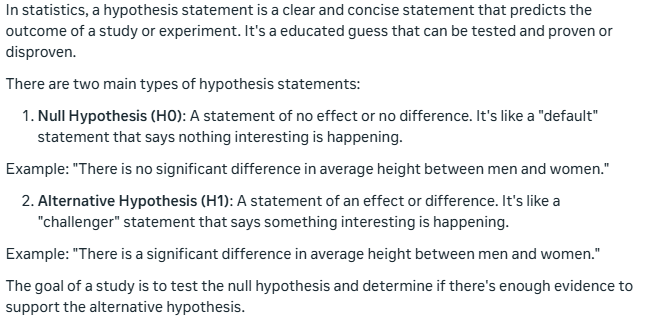
A sample is a representative subset of a larger population, used to make inferences about the whole population.

Although we can attempt to collect a sample dataset representative of the population, some margin of error is expected. In inferential statistics, we account for this margin of error with a statistical measure of prediction sureness called confidence.

Confidence measures how closely sample results match the population's true value, expressed as a percentage (0-100%). A higher percentage (e.g. 95%) indicates greater confidence that the results accurately represent the population, while 0% indicates no confidence.

It’s impossible to have a confidence level of 100%. The only way to prove the results are 100% accurate is by analyzing the entire population, which would render the study descriptive rather than inferential.

A crucial part of inferential statistics is the hypothesis test, in which you evaluate two mutually exclusive statements to determine which statement is correct given the data presented.



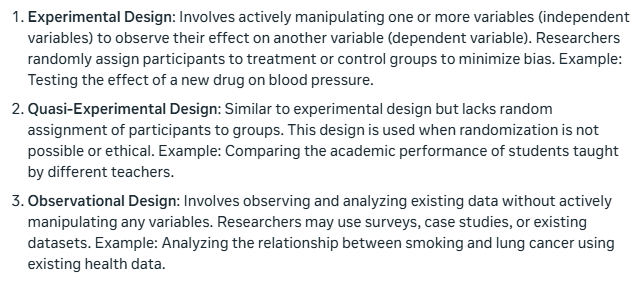
It’s important to note that the term “null” does not mean “invalid” or associated with the value zero but rather a hypothesis that the researcher attempts or wishes to “nullify.”

A clear hypothesis tests only one relationship and avoids conjunctions such as “and,” “nor” and “or.” According to California State University Bakersfield, a good hypothesis should include an “if” and “then” statement, such as: If [I study statistics] then [my employment opportunities increase].

A dependent variable represents what you’re attempting to predict, and the independent variable is the variable that supposedly impacts the outcome of the dependent variable.

A good hypothesis statement should be testable through an experiment, controlled test, or observation.

Statistical-based research design can be separated into three categories: experimental, quasi-experimental, and observational.

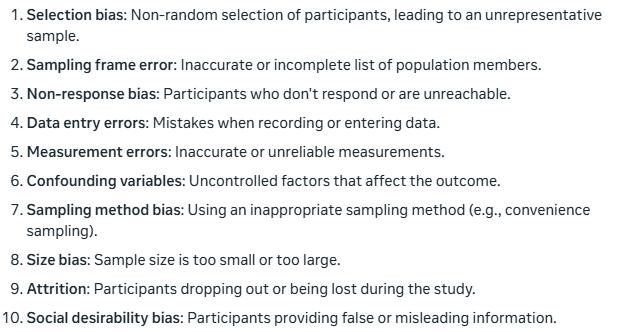


For most large-scale studies, it’s prohibitively expensive or impossible to survey and analyze an entire population of items, people, objects, etc. Instead, statisticians typically gather a sample, which is a subset of data that is representative of the full population.

A random sample is a subset of individuals or data points selected from a larger population, allowing us to make unbiased and precise inferences about the population, and enabling statistical analyses and generalizations.

For sample data to be reliable, each individual sample (item, person, object, etc.) must have an equal probability of selection and measurement.

Some factors that can contaminate a sample are:



When surveys are used to try and collect sample data, it is important that the survey topic remain unknown to the participants, so as not to affect the answers they provide.

The first tip to remember is that more relevant sample data is usually better than less. While more data allows you to cover more potential combinations and generally leads to more accurate predictions, there is no hard-and-fast rule on how much data is needed to perform statistical analysis.

Probability is the likelihood of something happening and is typically expressed as a number with a decimal value called a floating-point number. The closer the floating-point number is to 1.0, the more likely the hypothesis is true. For example: if one probability results in 0.2, and one other result is .99 - the .99 is more likely to happen since it is closer to 1.0 than .02.

Probability is a measure of how likely something is to happen, usually shown as a percentage or decimal. Odds, on the other hand, compare the number of times something happens to the number of times it doesn't happen. In other words, probability looks at the chance of something happening, while odds look at the ratio of successes to failures.

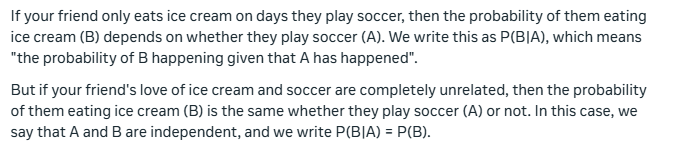
Correlation shows how two things are connected and tend to change together. When one thing changes, the other thing tends to change too.

Causation is when one thing (the cause) directly affects another thing (the effect). In other words, changing the cause leads to a change in the effect. For example, if you turn on a light switch (cause), the light turns on (effect).

Correlation and causation are commonly confused and for obvious reasons; when two variables vary together, they can appear connected. But the observed correlation could be caused by a third and previously unconsidered variable, known as a lurking variable or confounding variable.

The curse of dimensionality means that as data gets more complex, it becomes harder to work with and understand. Bear in mind, though, that the curse of dimensionality tends to affect machine learning and data mining analysis more than traditional hypothesis testing due to the high number of variables placed under consideration.

In statistics, we’re interested in relationships between variables and whether their apparent covariance is meaningful.



Bayes' Theorem is a mathematical formula that helps us update our belief in a hypothesis (H) based on new evidence (E). It's a way to revise our probability estimate of something being true, given new information.

Here's the simple formula:

P(H|E) = P(E|H) × P(H) / P(E)

Let's break it down:

* P(H|E) is the updated probability of the hypothesis (H) being true, given the new evidence (E)
* P(E|H) is the probability of the evidence (E) occurring, assuming the hypothesis (H) is true
* P(H) is our initial belief in the hypothesis (H) being true
* P(E) is the probability of the evidence (E) occurring, regardless of the hypothesis

In simple terms, Bayes' Theorem helps us answer questions like:

* Given new evidence, how likely is it that our hypothesis is true?
* How should we update our belief in a hypothesis based on new data?