What is statistic?

To understand the importance of statistics, we should first understand what is statistics. Well, Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and presenting data. It's all about making sense of numbers and finding patterns in data to make informed decisions. Whether you're looking at sports scores, election polls, or your grades, statistics helps you understand what's going on and predict future outcomes.

Lets understand this by a basic example: Imagine you’re a teacher and you want to understand the performance of your students in a recent math test. You decide to use statistics to analyze the test scores. You find out the average score of the students, then you find out what percentage of students have scored below the average score and decide to frame your teaching materials focused on the students who have scored less than the average score so as to improve their test scores. This way you have used statistics to find out the optimal teaching method for your students.

Think of it as a detective story where numbers are your clues. By analyzing these clues, we can uncover insights that help us understand trends and make predictions."

Is statistics a new concept:

Ancient beginnings: Statistics has its roots in ancient times. Early civilizations like the Egyptians and Babylonians used basic counting and record-keeping for agriculture and taxation.

Birth of modern statistics: The 17th century saw the birth of modern statistics. John Graunt, known as the father of demography, analyzed mortality data in London. In the 19th century, Carl Friedrich Gauss developed the method of least squares, a key technique in regression analysis.

20th century advances: The 20th century brought significant advances in the field of statistics. Pioneers like Ronald Fisher and Karl Pearson developed foundational theories in hypothesis testing and correlation. The advent of computers revolutionized statistics, enabling complex data analysis.

Statistics in digital age: Today, statistics is integral to the digital age. Big data and machine learning rely heavily on statistical methods to turn vast amounts of data into actionable insights, driving innovation across industries

Key Concepts of Statistics

Let's break down some key concepts in statistics starting with data collection

Data collection is the process of gathering information. to analyze and make decisions. It involves various methods to ensure we get accurate and relevant data. Collecting accurate data is essential because it forms the foundation of your analysis. Good data collection leads to reliable insights, while poor data collection can result in misleading conclusions.

There are two main types of data collection methods:

* + **Primary Data Collection**: This involves collecting data directly from the source where you collect data firsthand. Examples include surveys, interviews, and experiments.
  + **Secondary Data Collection**: This uses existing data collected by others where you use existing data. Examples include using reports, historical data, and online databases.

Let’s look at an example. Let's say you're a teacher who wants to know how students feel about a new teaching method. Here’s how you could collect the data:

* **Primary Data Collection**: You could create a survey and ask your students to fill it out, answering questions about their experience with the new method.
* **Secondary Data Collection**: You could look at existing research studies or reports on similar teaching methods to see how students in other schools felt about it.

Sampling Method: Sampling is the process of selecting a subset of individuals from a population to estimate characteristics of the whole population. Sampling methods help us choose a representative group from a larger population, ensuring our data is manageable and meaningful.

Imagine you want to find out the average height of students in a large school. Measuring every student is impractical, so you decide to use a sampling method

Here are some common sampling methods:

**Simple Random Sampling**:

* As the name says its simple & its random. It means that the sample is selected from the population randomly so as every member of the population has an equal chance of being selected.
* Example: Drawing names out of a hat to choose participants for a survey

**Systematic Sampling**:

* Every nth member of the population is selected after a random starting point. First you select a random starting point and start selecting the nth member of the population.
* Example: Surveying every 10th person on a list.

**Stratified Sampling**:

* The population is divided into subgroups (strata) based on a characteristic, and random samples are taken from each stratum.
* Example: Dividing a school into grades and randomly selecting students from each grade.

**Cluster Sampling**:

* The population is divided into clusters, and a few clusters are randomly selected. All members of the chosen clusters are then surveyed.
* Example: Selecting a few classrooms in a school and surveying all students in those classrooms.

**Convenience Sampling**:

* Samples are taken from a group that is easy to access or contact.
* Example: Surveying friends or colleagues because they are readily available.

Data Types

Next, let's talk about data types. In statistics, data can be classified into different types based on their characteristics. Understanding these data types is crucial for selecting the appropriate statistical methods for analysis.

Here are the main data types

Quantitative: We have quantitative data, which includes numbers. Represents numerical values and can be measured or counted and can be either discrete or continuous.

Discrete data are countable items like the number of students in a class, while continuous data are measurable quantities like height or weight."

Qualitative: On the other hand, qualitative data includes non-numerical information like colors or names. This can be nominal, which means no specific order like Eye color (e.g., blue, green, brown), or ordinal, Categories with a meaningful order but no fixed intervals between them like Movie ratings (e.g., poor, fair, good, excellent)

Importance of data types in statistics: By categorizing your data correctly, you can choose the right statistical methods to analyze it and draw meaningful conclusions

Descriptive Statistics:

Descriptive statistics involves summarizing and organizing data so that it can be easily understood. This type of statistics is used to describe the basic features of the data in a study. Imagine you have the tests scores of a class. Descriptive statistics can be used to summarize and organize the data in the following way:

Measure of Central Tendency

Mean:

The mean gives us the average which is essentially the sum of all data values divided by the number of values. Following our example if we add all the test scores of our class and divide it by the total number of students we will get the average score.

The median shows the middle value when data values are ordered from least to greatest. Example: Consider the test scores: 76, 78, 84, 85, 88, 89, 90, 90, 92, 90

Median = (88 + 89) / 2 = 88.5

The mode tells us the most frequent value. The value that occurs most frequently in the data set. Example: Consider the test scores: 76, 78, 84, 85, 88, 89, 90, 90, 92, 90

Mode = 90

The range is a measure of dispersion. The range is the difference between the highest and lowest values. Example: Consider the test scores: 76, 78, 84, 85, 88, 89, 90, 90, 92, 90

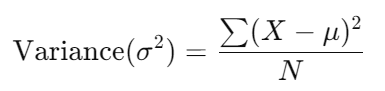
Range = 92-76= 16

The variance: The variance is a measure of dispersion. Variance measures the average squared deviation of each data point from the mean.

Calculation:

1. Find the mean (average) of the data set.
2. Subtract the mean from each data point and square the result.
3. Find the average of these squared differences.

Formula:



where XXX is each individual data point, μ\muμ is the mean of the data, and NNN is the number of data points.

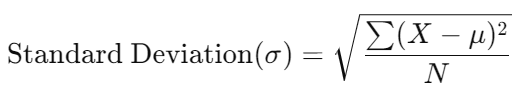
**Example**: Consider the data set: 2, 4, 6, 8, 10

* Mean = (2 + 4 + 6 + 8 + 10) / 5 = 6
* Squared differences from the mean: (2-6)², (4-6)², (6-6)², (8-6)², (10-6)² = 16, 4, 0, 4, 16
* Variance = (16 + 4 + 0 + 4 + 16) / 5 = 8

standard deviation to measure the spread of our data. Standard deviation is the square root of the variance. It provides a measure of the average distance from the mean.

Calculation:

1. Calculate the variance.
2. Take the square root of the variance.

Formula: 

Example: Continuing from the variance example:

* Variance = 8
* Standard Deviation = √8 ≈ 2.83

Inferential Statistics: Inferential statistics helps us make conclusions or predictions about a larger group (population) based on a smaller group (sample). Imagine you have a big jar of candies, and you want to know the average number of red candies in the jar without counting every single candy. Instead, you take a handful of candies (a sample) and use that information to make an educated guess about the entire jar. That’s what inferential statistics is all about.

Population vs. sample: The entire group you’re interested in or the entire group for which you are drawing inference for. In this case ass the candies in the jar. Sample is a smaller group selected from the population which helps us draw inference for the population. In this case the handful of candies.

Various sampling methods:

**Simple Random Sampling**:

* As the name says its simple & its random. It means that the sample is selected from the population randomly so as every member of the population has an equal chance of being selected.
* Example: Drawing names out of a hat to choose participants for a survey

**Systematic Sampling**:

* Every nth member of the population is selected after a random starting point. First you select a random starting point and start selecting the nth member of the population.
* Example: Surveying every 10th person on a list.

**Stratified Sampling**:

* The population is divided into subgroups (strata) based on a characteristic, and random samples are taken from each stratum.
* Example: Dividing a school into grades and randomly selecting students from each grade.

**Cluster Sampling**:

* The population is divided into clusters, and a few clusters are randomly selected. All members of the chosen clusters are then surveyed.
* Example: Selecting a few classrooms in a school and surveying all students in those classrooms.

**Convenience Sampling**:

* Samples are taken from a group that is easy to access or contact.
* Example: Surveying friends or colleagues because they are readily available.

The central limit theorem, which helps us understand the distribution of sample means. The Central Limit Theorem (CLT) is a fundamental concept in statistics that describes how the distribution of sample means approaches a normal distribution, regardless of the shape of the population distribution, as the sample size becomes large.

The Central Limit Theorem tells us that the distribution of sample means will tend to be normal, given a sufficiently large sample size, regardless of the population's distribution shape. This is crucial because it enables the use of normal probability methods in many practical situations, even when the population distribution is unknown

Hypothesis Testing: A method to test if a certain claim about a population is true or not

**Null Hypothesis (H₀)**: The default assumption (e.g., there is no difference between two groups). For example lets suppose you want to test if the average height is 5.6 feet. The null hypothesis(H0) is that the average height is 5.6 feet.

**Alternative Hypothesis (H₁)**: The opposite of the null (e.g., there is a difference). So for the same assumption the alternate hypothesis (H1) will be The average height is not 5.6 feet.

Hypothesis testing helps us determine if our findings are statistically significant. We start with a null hypothesis (H0) and an alternate hypothesis (H1). We then calculate the p-value to decide whether to accept or reject the null hypothesis. Common tests include the Z-test, T-test, Chi-square test, and ANOVA.

Before we move on to our common tests of hypothesis lets learn about the Type I and Type II errors and P-Value.

**Type I and II Errors**: Type I is rejecting H0 when it is true; Type II is failing to reject H0 when it is false.

**Type I Error**: False positive, like thinking a medicine works when it doesn’t. For example lets suppose that you are a farmer and you are using a new fertilizer for your crops and you conduct a test to find out that weather the new fertilizer works or not and the Type I error in this case would be concluding the new fertilizer works when it doesn’t.

**Type II Error**: False negative, like thinking a medicine doesn’t work when it does. Concluding the new fertilizer doesn’t work when it does.

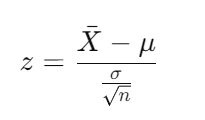
**P-value**: The probability of obtaining the observed results, or more extreme, if H0 is true. If this is low, it means your results are unlikely to have happened by chance, so you can reject the null hypothesis. If the probability of getting our observed increase in yield by chance is very low, we reject the null hypothesis

**What is a P-Value?**

The p-value is a measure used in statistical hypothesis testing to determine the strength of the evidence against the null hypothesis (H₀). It represents the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true.

* **Small p-value (≤ 0.05)**: Strong evidence against the null hypothesis, leading to its rejection.
* **Large p-value (> 0.05)**: Weak evidence against the null hypothesis, failing to reject it.

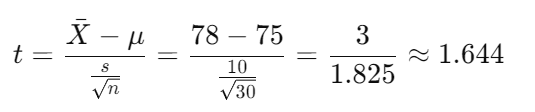
**Steps to Calculate the P-Value:**

1. **State the Hypotheses**:
   * Null Hypothesis (H₀): Assumes no effect or no difference.
   * Alternative Hypothesis (H₁): Assumes some effect or difference.
2. **Choose the Appropriate Test**:
   * Select a statistical test based on the data type and sample size (e.g., t-test, z-test, chi-square test).
3. **Calculate the Test Statistic**:
   * Compute the test statistic using sample data. The test statistic measures how far the sample data diverge from the null hypothesis.
   * Example for a z-test: 
4. **Find the P-Value**:
   * Use the test statistic to determine the p-value from the appropriate distribution (e.g., z-distribution for z-test, t-distribution for t-test).
   * This can be done using statistical tables or software.

**Example Calculation:**

Imagine you want to test if a new teaching method has changed the average score of students compared to the previous average score of 75. You collect a sample of 30 students with an average score of 78 and a standard deviation of 10.

1. **Hypotheses**:
   * H0:μ=75 (the average score has not changed)
   * H1:μ≠75 (the average score has changed)
2. **Choose the Test**:
   * Use a t-test since the sample size is relatively small and the population standard deviation is unknown.
3. **Calculate the Test Statistic**:



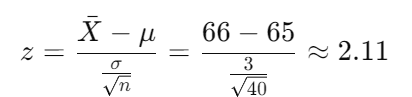
1. **Find the P-Value**:
   * Look up the t-statistic (1.644) in the t-distribution table with 29 degrees of freedom (n-1) or use software.
   * Assuming a two-tailed test, the p-value might be around 0.11.

Since the p-value (0.11) is greater than 0.05, you fail to reject the null hypothesis. This means there isn’t strong enough evidence to say the new teaching method has changed the average score.

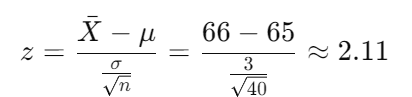
Common hypothesis testing process: Hypothesis testing is a statistical method used to make decisions about a population based on sample data. There are tests to evaluate our hypothesis testing and some common hypothesis tests: Z-test, T-test, Chi-Square test, and ANOVA. Lets understand these one by one.

**1. Z-Test**

**Purpose**: Used to determine whether there is a significant difference between the sample mean and the population mean when the population variance is known and the sample size is large (n > 30).

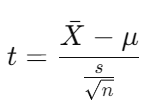
The formula for the test is 

**Example**:

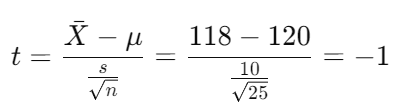
* **Scenario**: You want to test if the average height of a group of students is different from the known population average height of 65 inches.
* **Hypotheses**:
  + Null Hypothesis (H₀): The average height is 65 inches.
  + Alternative Hypothesis (H₁): The average height is not 65 inches.
* **Sample Data**: Average height of the sample is 66 inches, standard deviation is 3 inches, and sample size is 40.
* **Test Statistic Calculation**: 
* **Conclusion**: Compare the z-value to a standard normal distribution. If the z-value is beyond the critical value (e.g., 1.96 for a 5% significance level), reject the null hypothesis.

**2. T-Test**

**Purpose**: Used to determine whether there is a significant difference between the sample mean and the population mean when the population variance is unknown and/or the sample size is small (n ≤ 30). The formula is



**Example**:

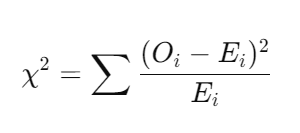
* **Scenario**: You want to test if a new drug affects blood pressure differently from the average blood pressure of 120 mmHg.
* **Hypotheses**:
  + Null Hypothesis (H₀): The average blood pressure is 120 mmHg.
  + Alternative Hypothesis (H₁): The average blood pressure is not 120 mmHg.
* **Sample Data**: Average blood pressure of the sample is 118 mmHg, standard deviation is 10 mmHg, and sample size is 25.
* **Test Statistic Calculation**: 
* **Conclusion**: Compare the t-value to the critical value from the t-distribution table with n−1n-1n−1 degrees of freedom. If the t-value is beyond the critical value, reject the null hypothesis.

**3. Chi-Square Test**

**Purpose**: Used to determine whether there is a significant association between categorical variables.

**Example**:

* **Scenario**: You want to test if there is an association between gender (male, female) and preference for a type of movie (action, comedy).
* **Hypotheses**:
  + Null Hypothesis (H₀): There is no association between gender and movie preference.
  + Alternative Hypothesis (H₁): There is an association between gender and movie preference.
* **Sample Data**: A contingency table showing the frequency of preferences.
* **Test Statistic Calculation**:



where OiO\_iOi​ is the observed frequency, and EiE\_iEi​ is the expected frequency.

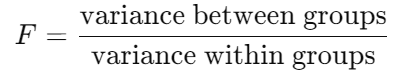
* **Conclusion**: Compare the chi-square value to the critical value from the chi-square distribution table with (r−1)(c−1)(r-1)(c-1)(r−1)(c−1) degrees of freedom. If the chi-square value is beyond the critical value, reject the null hypothesis.

**4. ANOVA (Analysis of Variance)**

**Purpose**: Used to determine whether there are significant differences between the means of three or more groups.

**Example**:

* **Scenario**: You want to test if different types of diets (A, B, C) lead to different weight loss.
* **Hypotheses**:
  + Null Hypothesis (H₀): All diets result in the same average weight loss.
  + Alternative Hypothesis (H₁): At least one diet results in a different average weight loss.
* **Sample Data**: Weight loss data from individuals on each diet.
* **Test Statistic Calculation**:
  + Calculate the F-statistic using the ratio of between-group variance to within-group variance.

​

* **Conclusion**: Compare the F-value to the critical value from the F-distribution table with appropriate degrees of freedom. If the F-value is beyond the critical value, reject the null hypothesis.

**Summary**

* **Z-Test**: Used for large samples and known population variance.
* **T-Test**: Used for small samples and unknown population variance.
* **Chi-Square Test**: Used for categorical data to test for association.
* **ANOVA**: Used to compare means across three or more groups.

Each test helps determine whether observed data significantly deviate from what we would expect under the null hypothesis.

Regression Analysis"\*

Regression analysis is a powerful statistical method used to examine the relationship between two or more variables. It's widely used in data science for prediction, forecasting, and determining causal relationships. Here are the key types of regression analysis that a data scientist should be familiar with:

* Simple Linear Regression: Examines the relationship between two variables: one dependent variable (Y) and one independent variable (X). *Example*: Predicting house prices based on square footage.

**Equation**: 

* Y: Dependent variable
* X: Independent variable
* β0​: Intercept
* β1​: Slope
* ϵ: Error term

Multiple Linear Regression: Examines the relationship between one dependent variable and two or more independent variables. *Example*: Predicting house prices based on square footage, number of bedrooms, and age of the house.

**Equation**

* Y: Dependent variable
* X1,X2,…,Xn​: Independent variables
* β0​: Intercept
* β1,β2,…,βn​: Slopes
* ϵ: Error term

Logistic Regression: Used when the dependent variable is categorical (often binary). Estimates the probability that a given input point belongs to a certain class. *Example*: Predicting whether a student will pass or fail an exam based on study hours and attendance.

**Equation**: 

* p: Probability of the event occurring
* 1−p1: Probability of the event not occurring

There are few more regression:

**3. Polynomial Regression**

* A form of regression where the relationship between the independent variable and the dependent variable is modeled as an nth degree polynomial.
* Used for datasets where the relationship between variables is non-linear.
* *Example*: Predicting the growth rate of a plant based on the amount of fertilizer used.
  + **Equation**: 
  + Y: Dependent variable
  + X: Independent variable
  + X2,Xn: Polynomial terms

**4. Ridge Regression (L2 Regularization)**

* Addresses multicollinearity (independent variables are highly correlated) by adding a penalty to the size of coefficients.
* Shrinks coefficients towards zero, but not exactly zero, which can help with overfitting.
* *Example*: Used in situations where you have many predictors and need to prevent overfitting.
  + **Equation**: 
  + λ: Regularization parameter

**5. Lasso Regression (L1 Regularization)**

* Similar to Ridge Regression but uses L1 regularization, which can shrink some coefficients to zero, performing feature selection.
* *Example*: Useful when you have a large number of features and want to identify the most important ones.
  + **Equation**: 
  + λ: Regularization parameter

**6. Elastic Net Regression**

* Combines L1 and L2 regularization.
* Useful when there are multiple features that are correlated with each other.
* *Example*: Predicting customer churn where multiple factors (usage metrics, demographics, etc.) are considered.
  + **Equation**: 
  + λ1,λ2​: Regularization parameters

**7. Poisson Regression**

* Used for modeling count data and rates.
* The dependent variable is a count (non-negative integer) or rate (count per unit time/space).
* *Example*: Predicting the number of calls received by a call center in an hour.
* **Equation**: 
  + λ: Expected count

**8. Quantile Regression**

* Estimates the conditional quantile (e.g., median or other percentiles) of the response variable.
* Useful for understanding the impact of variables at different points in the distribution.
* *Example*: Analyzing how different factors affect the median house price.
* **Equation**: Similar to linear regression but focuses on quantiles rather than the mean.

**Why These Regression Types Matter**

* **Prediction Accuracy**: Different regression techniques can improve the accuracy of your predictions depending on the nature of the data and the problem.
* **Interpretability**: Some regression models (like linear regression) are easier to interpret, which is important for understanding the relationships between variables.
* **Feature Selection**: Techniques like Lasso Regression help in identifying the most important features, reducing dimensionality, and improving model performance.
* **Handling Different Data Types**: Different regression methods are suited to different types of dependent variables (e.g., logistic regression for binary outcomes, Poisson regression for count data).

Applications in Data Science

Statistics is crucial in data science for exploratory data analysis, making predictions, building predictive models, and decision making. By applying these statistical methods, data scientists can uncover insights that drive business and scientific advancements.

**Confidence Intervals**

* A range of values within which the true parameter is expected to fall with a certain level of confidence (e.g., 95%). This gives a range that we’re fairly sure the true value lies within. For example, if we say we’re 95% confident that the average height students in a class is between 5.5 and 6 feet, it means that in 95 out of 100 cases, the true average height would be in this range.

They measure the average height of their sample and calculate a range (confidence interval) where they believe the true average height of all students falls. For instance, they might say they are 95% confident that the true average height is between 4.8 and 5.2 feet.

If we say we are 95% confident that the average yield of corn is between 80 and 100 bushels per acre, it means that if we repeated this study many times, 95% of the time, the true average yield would be within this range

**Applications in Data Science**

**1. Exploratory Data Analysis (EDA)**

* Using descriptive statistics and visualization techniques to understand the data. Looking at your data with graphs and summary statistics to understand what’s going on. Creating graphs and summaries to understand patterns in crop yields and animal productivity.

**2. Probability Models for Prediction**

* Applying probability distributions to model uncertainties and make predictions (e.g., Bayesian methods). Using probability to make predictions. For example, weather forecasts use past data and probability models to predict rain. Using past weather data to predict future rainfall, helping us plan irrigation.

**3. Statistical Inference for Decision Making**

* Using hypothesis testing and confidence intervals to make informed decisions based on data. Making decisions based on data analysis. For instance, a company might use hypothesis testing to decide if a new product is better than an old one. Deciding whether to use a new type of seed based on testing a sample and analyzing the results.

**4. Regression for Predictive Modeling**

* Building predictive models to forecast outcomes and understand relationships between variables. Creating models that can predict future outcomes. For example, predicting house prices based on size, location, and other factors. Building models to predict future harvests based on current planting conditions and historical data.