#UpSkillWithKalpesh

Day 06

# Data Science Unlocked

From Zero to Data Hero

## Intermediate Statistics for Data Science

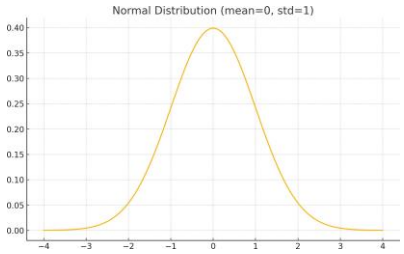
Kalpesh Pathade
@DataSimplified

# Intermediate Statistics(CLT, Hypothesis Tests, etc.)

## I. Distributions

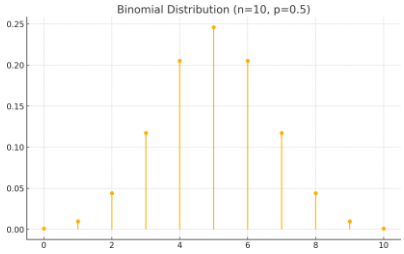
### Normal Distribution

- **Definition:** A continuous probability distribution that is symmetric around its mean. It is also known as the Gaussian distribution.
- **Properties:**
  - Bell-shaped curve.
  - Mean › Median › Mode.
  - Completely defined by its mean (μ) and standard deviation (σ).
  - 68-95-99.7 rule (empirical rule): About 68% of data falls within 1σ, 95% within 2σ, and 99.7% within 3σ.
- **Applications in Data Science:**
  - Used in hypothesis testing and confidence intervals.
  - Commonly applied in predictive modeling and machine learning algorithms.
  - Assumes normality in linear regression and other statistical models.
- **Chart:**



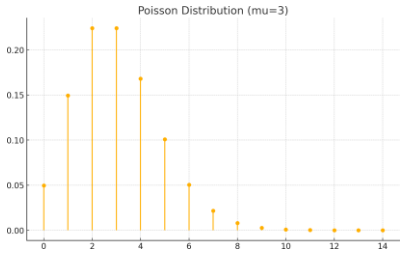

### Binomial Distribution

- **Definition:** A discrete probability distribution of the number of successes in a fixed number of independent Bernoulli trials.
- **Properties:**
  - Two possible outcomes: Success or Failure.
  - Defined by parameters n (number of trials) and p (probability of success).
  - Mean = np, Variance = np(1-p).
- **Applications in Data Science:**
  - Used in classification problems, especially in logistic regression.
  - Modeling binary outcomes such as click-through rates or churn prediction.
  - Applied in A/B testing to compare conversion rates.
- **Chart:**



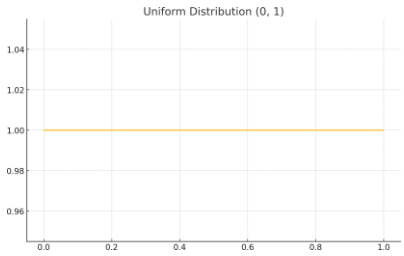

## Poisson Distribution

- **Definition:** A discrete distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space.
- **Properties:**
  - Defined by the rate parameter λ (lambda).
  - Mean › Variance = λ.
  - Suitable for rare events.
- **Applications in Data Science:**
  - Used to model count data, such as the number of occurrences of an event.
  - Applied in anomaly detection, for example, identifying unusual traffic spikes.
  - Common in natural language processing for modeling word frequencies.
- **Chart:**



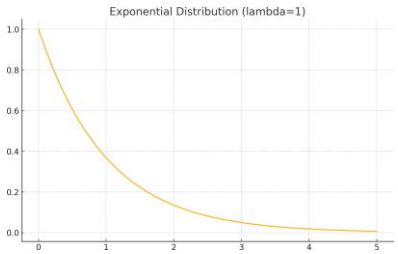

## Uniform Distribution

- **Definition:** A type of probability distribution in which all outcomes are equally likely.
- **Properties:**
  - Continuous uniform distribution is defined by lower and upper bounds (a, b).
  - Mean = (a + b) / 2.
  - Variance = (b - a)^2 / 12.
- **Applications in Data Science:**
  - Used in simulations and random sampling.
  - Basis for generating random numbers in algorithms.
  - Helps in initializing weights in machine learning models.
- **Chart:**




### Exponential Distribution

- **Definition:** A continuous probability distribution that describes the time between events in a Poisson process.
- **Properties:**
  - Defined by the rate parameter λ.
  - Mean › 1/λ, Variance › 1/λ^2.
  - Memoryless property: The probability of an event occurring in the next interval is independent of previous intervals.
- **Applications in Data Science:**
  - Used in survival analysis and reliability engineering.
  - Applied in modeling time-to-event data, such as customer churn or system failures.
  - Relevant in queuing theory for predicting wait times.
- **Chart:**




## II. Central Limit Theorem (CLT)

### Definition:

The Central Limit Theorem (CLT) is a fundamental statistical principle that explains how the distribution of sample means becomes approximately normal as the sample size increases, even if the original data is not normally distributed. This remarkable property allows statisticians to make inferences about population parameters using the normal distribution, which simplifies analysis and interpretation.

### Key Concepts:

- **Sampling Distribution:** This is the probability distribution of a statistic (like the mean) obtained from a large number of samples drawn from the same population.
- **Sample Mean (μ̂):** The average of the values in a sample, which estimates the population mean.
- **Independence:** The sampled observations must not influence each other.
- **Identically Distributed:** All sampled observations must come from the same underlying distribution with the same probability.

### Mathematical Expression:

Consider X1,X2,...,Xn as independent and identically distributed random variables with mean μμ and standard deviation σ. The standardized sample mean is calculated as:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

As the sample size nn approaches infinity, the distribution of Z approaches the standard normal distribution N(0,1), regardless of the original data distribution.

## Conditions for CLT:

1. **Sample Size:** Generally, a sample size of  is considered sufficient for the CLT to hold. However, if the population distribution is heavily skewed, a larger sample size might be necessary.

   n>30
2. **Independence:** The sampled data points must be independent of each other. This condition can be violated in time series data where observations are sequentially dependent.
3. **Finite Variance:** The population from which the samples are drawn must have a finite variance. CLT does not apply to distributions with infinite variance, like the Cauchy distribution.
4. **Random Sampling:** Samples must be collected through a process that ensures each member of the population has an equal chance of being included.

## Why is CLT Important in Data Science?

- **Hypothesis Testing:** Many statistical tests assume normality of the sampling distribution of the test statistic. CLT justifies this assumption even when the data itself isn't normal.
- **Confidence Intervals:** CLT allows the construction of confidence intervals for population parameters by using the normal distribution.
- **Predictive Modeling:** Machine learning algorithms, such as linear regression and logistic regression, assume normally distributed errors. CLT supports this assumption for large datasets.
- **Bootstrapping:** This resampling technique relies on CLT to approximate the distribution of a statistic, helping estimate confidence intervals and standard errors.
- **Error Analysis:** CLT aids in understanding the distribution of prediction errors, crucial for refining models and improving accuracy.

## Real-World Examples:

1. **A/B Testing in Marketing:** When testing different versions of a webpage or advertisement, the conversion rates are sampled. As the sample size grows, the average conversion rates tend to follow a normal distribution, allowing for reliable statistical comparisons.
2. **Quality Control in Manufacturing:** By sampling batches of products and calculating the average dimensions or weights, manufacturers can use the normal distribution to detect defects and maintain quality standards.
3. **Financial Modeling:** Daily stock returns, when aggregated over a period (like monthly or yearly), tend to form a normal distribution. This helps in risk assessment, portfolio optimization, and forecasting.
4. **Medical Research:** In clinical trials, the average effects of treatments across different patient samples approximate a normal distribution, aiding in the evaluation of new drugs or procedures.

## Visualization of CLT:

Imagine repeatedly taking random samples from a skewed distribution (like income, which often has a long right tail). Each sample may have a different mean. If you plot the distribution of these sample means, you'll notice that as the sample size increases, the shape of this distribution becomes more and more bell-shaped, resembling a normal distribution.

- **Small Sample Sizes:** The distribution of sample means may still reflect the skewness of the population.
- **Large Sample Sizes:** The distribution smooths out and approaches a normal distribution, even if the original population was skewed or non-normal.

## Intuitive Understanding:

Consider rolling a fair die. The outcome of a single roll follows a uniform distribution (each number from 1 to 6 is equally likely). However, if you roll the die 30 times and take the average, and then repeat this process many times, the distribution of these averages will form a

bell curve. This transformation from a uniform distribution to a normal distribution through repeated sampling is a direct consequence of the Central Limit Theorem.

The Central Limit Theorem is the backbone of statistical inference and data science. It provides the theoretical foundation for many analytical techniques, allowing data scientists to make valid conclusions from sample data and apply probabilistic models to real-world problems with confidence.

## III. Sampling Techniques

### 1. Random Sampling

- **Definition:** Random sampling is a technique where every member of the population has an equal chance of being selected. This ensures that the sample represents the population without bias.
- **How It Works:** Use random number generators or lottery methods to select participants.
- **Advantages:**
  - Reduces selection bias.
  - Simple to implement.
  - Provides a good representation of the population when the sample size is large enough.
- **Disadvantages:**
  - May not account for specific subgroups within the population.
  - Requires a complete list of the population, which may not always be available.
- **Applications in Data Science:**
  - Used in creating training and test datasets for machine learning.
  - Applied in surveys and polls to generalize results to the broader population.

### 2. Stratified Sampling

- **Definition:** Stratified sampling involves dividing the population into distinct subgroups or "strata" based on specific characteristics, then randomly sampling from each stratum.
- **How It Works:** Identify key subgroups (e.g., age, gender, income level), then perform random sampling within each subgroup proportionally.
- **Advantages:**
  - Ensures representation of all key subgroups.
  - Increases precision and accuracy of results.
- **Disadvantages:**
  - More complex to organize and analyze.
  - Requires detailed knowledge of population characteristics.
- **Applications in Data Science:**
  - Used in predictive modeling to ensure all classes are represented equally, especially in imbalanced datasets.
  - Applied in market research to understand behaviors across different demographic groups.

### 3. Systematic Sampling

- **Definition:** Systematic sampling selects every kth member of the population from a randomly chosen starting point.
- **How It Works:** Determine the sampling interval k (e.g., every 10th person) and start from a randomly chosen position within the first interval.
- **Advantages:**
  - Simple and quick to implement.

- Ensures even coverage of the population.

- **Disadvantages:**
  - Can introduce bias if there is a hidden pattern in the population list that coincides with the sampling interval.
  - Not suitable for populations with periodic patterns.
- **Applications in Data Science:**
  - Used in time-series analysis for regular interval sampling.
  - Applied in quality control processes in manufacturing.

---

### 4. Sampling Distributions

- **Definition:** A sampling distribution is the probability distribution of a given statistic (like the mean or proportion) based on a random sample.
- **Key Concepts:**
  - **Sample Mean Distribution:** The distribution of sample means from multiple samples of the same size.
  - **Standard Error:** The standard deviation of the sampling distribution, which decreases as the sample size increases.
- **Importance:**
  - Provides the foundation for inferential statistics, enabling hypothesis testing and confidence interval estimation.
  - Essential for understanding variability and reliability of sample statistics.
- **Applications in Data Science:**
  - Used to assess the variability of model performance metrics (e.g., accuracy, precision).
  - Applied in bootstrapping techniques for estimating sampling distributions when theoretical distributions are unknown.

---

### 5. Law of Large Numbers (LLN)

- **Definition:** The Law of Large Numbers states that as the size of a sample increases, the sample mean will get closer to the population mean.
- **Types:**
  - **Weak Law of Large Numbers:** The sample mean converges in probability towards the population mean.
  - **Strong Law of Large Numbers:** The sample mean almost surely converges to the population mean.
- **Importance:**
  - Ensures that results become more reliable as sample size increases.
  - Justifies the use of large datasets in making accurate predictions and inferences.
- **Applications in Data Science:**
  - Used in Monte Carlo simulations where repeated random sampling approximates complex probabilities.
  - Supports the stability of machine learning models as more data is provided.
  - Helps in validating models by comparing long-run averages of predictions to actual outcomes.

---

## Ⅳ. Point Estimation and Confidence Intervals

### 1. Estimators and Properties of Good Estimators

- **Definition:** An estimator is a statistic used to infer the value of an unknown population parameter. For example, the sample mean ($\hat{\mu}$) is an estimator of the population mean ($\mu$).
- **Properties of Good Estimators:**
  - **Unbiasedness:** An estimator is unbiased if its expected value equals the true population parameter.
  - **Consistency:** As the sample size increases, the estimator converges to the true parameter value.
  - **Efficiency:** Among unbiased estimators, the one with the smallest variance is considered efficient.

- **Sufficiency:** An estimator is sufficient if it uses all the information in the data relevant to estimating the parameter.

- **Applications in Data Science:**
  - Used in parameter estimation for statistical models.
  - Applied in evaluating machine learning model coefficients.

---

### 2. Confidence Interval for Mean and Proportion

- **Definition:** A confidence interval (CI) is a range of values, derived from the sample, that is likely to contain the true population parameter with a specified level of confidence (e.g., 95%).

  **Confidence Interval for Mean (when population standard deviation is known):**

$$CI = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

  - $\bar{X}$ is the sample mean
  - Z is the Z-score corresponding to the confidence level
  - σ is the population standard deviation
  - n is the sample size

  **Confidence Interval for Proportion:**

$$\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  Where:

  - $\hat{p}$ is the sample proportion
  - Z is the Z-score corresponding to the confidence level
  - n is the sample size

- **Applications in Data Science:**
  - Used in A/B testing to estimate the difference in conversion rates.
  - Applied in model evaluation to assess the reliability of performance metrics.

---

### 3. Margin of Error

- **Definition:** The margin of error (MOE) quantifies the amount of random sampling error in a survey's results. It defines the range within which the true population parameter is expected to fall.
- **Formula:**

$$MOE = Z \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

Where  is the Z-score for the desired confidence level,  is the standard deviation, and  is the sample size.

- **Importance:**
  - A smaller margin of error indicates more precise estimates.
  - The margin of error decreases as the sample size increases.
- **Applications in Data Science:**
  - Helps in interpreting the results of surveys and experiments.
  - Used to quantify uncertainty in predictive models and forecasts.

---

## V. Hypothesis Testing

## 1. Introduction to Hypothesis Testing

- **Definition:** Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data. It evaluates two competing statements to determine which is better supported by the data.
- **Purpose:** To assess whether there is enough evidence in a sample to infer that a certain condition holds true for the entire population.

## 2. Null Hypothesis ($H_0$)

- **Definition:** The null hypothesis is a statement that there is no effect, no difference, or no relationship between variables. It represents the default or status quo assumption.
- **Notation:** Typically denoted as .**$H_0$**
- **Examples:**
  - In testing a new drug: **$H_0$** : The new drug has no effect compared to the placebo.
  - In quality control: **$H_0$**: The mean weight of a product batch is equal to the specified standard.
- **Role in Testing:** The null hypothesis is tested directly. If the data provides strong enough evidence against it, we reject **$H_0$**

## 3. Alternative Hypothesis ($H_1$ or $H_a$)

- **Definition:** The alternative hypothesis is a statement that contradicts the null hypothesis. It suggests that there is an effect, a difference, or a relationship.
- **Notation:** Typically denoted as **$H_1$ or $H_a$** .
- **Types of Alternative Hypotheses:**
  - **One-Tailed Test:** Tests if a parameter is either greater than or less than a certain value.
    - Example: **$H_a$** : The new drug is more effective than the placebo.
  - **Two-Tailed Test:** Tests if a parameter is simply different from a certain value (could be either higher or lower).
    - Example: **$H_a$**: The mean weight of the product batch is not equal to the specified standard.
- **Role in Testing:** If the null hypothesis is rejected, the alternative hypothesis is accepted as the more plausible explanation.

## 4. Steps in Hypothesis Testing

1. **State the Hypotheses:** Define **$H_0$ and $H_a$** .
2. **Choose a Significance Level (α):** Common levels are 0.05, 0.01, or 0.10, representing the probability of rejecting **$H_0$** when it is actually true.
3. **Select the Appropriate Test:** Depending on the data type and hypothesis (e.g., t-test, z-test, chi-square test).
4. **Calculate the Test Statistic:** Use sample data to compute a value that will help in decision-making.
5. **Determine the p-Value or Critical Value:** The p-value indicates the probability of obtaining the observed results if **$H_0$** is true.
6. **Make a Decision:**
   - If **p-value ≤α**, reject **$H_0$**.
   - If **p-value >α** do not reject **$H_0$** .
7. **Interpret the Results:** Explain what the decision means in the context of the original research question.

## 5. Applications in Data Science

- **A/B Testing:** Used to compare two versions of a product or feature (e.g., website design, marketing emails) to see if changes improve performance.
- **Model Evaluation:** Hypothesis tests can assess if model improvements are statistically significant.
- **Experiment Design:** Helps in validating the effectiveness of new algorithms, treatments, or interventions.
- **Quality Control:** Monitors production processes to ensure products meet specifications.

### 6. Real-Life Example

**Scenario:** A company wants to test if a new website layout increases user engagement compared to the current layout.

1. **Null Hypothesis ($H_0$):** The new website layout has no effect on user engagement compared to the current layout.
2. **Alternative Hypothesis (Ha):** The new website layout increases user engagement.
3. **Significance Level (α):** 0.05 (5%)
4. **Data Collection:** The company randomly splits users into two groups: one sees the old layout (control group) and the other sees the new layout (treatment group).
5. **Test Selection:** A t-test is conducted to compare the average time spent on the website by both groups.
6. **Results:**
   - If the p-value is 0.03 (which is less than 0.05), the company rejects $H_0$ and concludes that the new layout significantly increases user engagement.
   - If the p-value is 0.07 (which is greater than 0.05), the company does not reject $H_0$, meaning there isn't enough evidence to conclude the new layout has an effect.

---

## VI. Type I and Type II Errors

### 1. Introduction to Errors in Hypothesis Testing

- **Definition:** In hypothesis testing, errors occur when incorrect conclusions are drawn from sample data. These errors are categorized into two types: Type I and Type II errors.
- **Importance:** Understanding these errors is crucial for interpreting the results of statistical tests accurately and minimizing incorrect decisions.

---

### 2. Type I Error (False Positive)

- **Definition:** A Type I error occurs when the null hypothesis ($H_0$) is rejected when it is actually true. This means we conclude that there is an effect or difference when none exists.
- **Notation:** The probability of committing a Type I error is denoted by α, also known as the significance level.
- **Example:**
  - In a medical trial, concluding that a new drug is effective when it actually has no effect.
  - In a courtroom, convicting an innocent person.
- **Consequences:**
  - Can lead to the adoption of ineffective treatments or interventions.
  - In business, it may result in unnecessary changes based on false findings.
- **How to Control:**
  - Set a lower significance level (e.g., $\alpha=0.01$ instead of $\alpha=0.05$ ).
  - Use more rigorous testing methods and larger sample sizes.

---

### 3. Type II Error (False Negative)

- **Definition:** A Type II error occurs when the null hypothesis ($H_0$) is not rejected when it is actually false. This means we fail to detect an effect or difference that actually exists.
- **Notation:** The probability of committing a Type II error is denoted by β.
- **Example:**
  - In a medical trial, concluding that a new drug is ineffective when it actually works.
  - In a courtroom, acquitting a guilty person.
- **Consequences:**
  - Can result in missing out on beneficial treatments or innovations.

- In business, it may lead to ignoring valuable opportunities for improvement.

- **How to Control:**
  - Increase the sample size to improve the power of the test.
  - Use a higher significance level (e.g., $\alpha=0.1$ instead of $\alpha=0.05$ ).

### 4. Relationship Between Type I and Type II Errors

- **Trade-Off:** Reducing the probability of a Type I error increases the risk of a Type II error, and vice versa. Balancing these risks depends on the context and consequences of the errors.
- **Power of a Test:** The power of a test ($1-\beta 1$) is the probability of correctly rejecting a false null hypothesis. Increasing power reduces the likelihood of a Type II error.

### 5. Real-Life Examples

1. **Medical Testing:**
   - **Type I Error:** A test incorrectly indicates a patient has a disease (false positive).
   - **Type II Error:** A test fails to detect a disease that the patient actually has (false negative).
2. **Legal System:**
   - **Type I Error:** Convicting an innocent person.
   - **Type II Error:** Acquitting a guilty person.
3. **Business Decisions:**
   - **Type I Error:** Implementing a new marketing strategy based on a test result that falsely indicates it will increase sales.
   - **Type II Error:** Ignoring a potentially successful marketing strategy because the test failed to detect its effectiveness.

## VII. P-Values

### 1. Introduction to P-Values

- **Definition:** A p-value is the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis ($\mathbf{H_0}$) is true. It quantifies the strength of evidence against the null hypothesis.
- **Importance:** P-values help determine the statistical significance of results in hypothesis testing. A smaller p-value indicates stronger evidence against .

### 2. Interpreting P-Values

- **Threshold for Significance:** The significance level ($\alpha$) is a pre-defined threshold against which the p-value is compared.
  - **If p-value $\leq \alpha$** Reject $\mathbf{H_0}$ (the results are statistically significant).
  - **If p-value $>\alpha$:** Do not reject $\mathbf{H_0}$ (the results are not statistically significant).
- **Common Significance Levels:**
  - $\alpha=0.05(5\%)$: Most commonly used threshold.
  - $\alpha=0.01$ (1%): Indicates stronger evidence required to reject $\mathbf{H_0}$.
  - $\alpha=0.10$ (10%): Used in exploratory research or less strict contexts.

### 3. Misconceptions About P-Values

- **P-value is NOT the probability that $\mathbf{H_0}$ is true:** It only measures the probability of observing the data if is true.
- **A low p-value does NOT prove $\mathbf{H_1}$ is true:** It only suggests that $\mathbf{H_0}$ is unlikely given the observed data.
- **P-value does NOT measure the size of an effect:** Statistical significance does not imply practical significance.

### 4. Calculating P-Values

- **Test Statistic:** The p-value is derived from the test statistic (e.g., z-score, t-score) based on the sample data.

- **Using Statistical Software:** Tools like Python (SciPy), R, and SPSS can compute p-values automatically.
- **One-Tailed vs. Two-Tailed Tests:**
  - **One-Tailed Test:** P-value represents the probability of an extreme result in one direction.
  - **Two-Tailed Test:** P-value represents the probability of extreme results in both directions.

---

### 5. Real-Life Examples

1. **Medical Research:**
   - A study tests whether a new drug lowers blood pressure more effectively than an existing drug.
   - **Null Hypothesis ($H_0$):** There is no difference in blood pressure reduction between the two drugs.
   - **Result:** The p-value is 0.03.
     - **Interpretation:** Since $0.03 < 0.05$, the null hypothesis is rejected, suggesting the new drug is more effective.
2. **Business A/B Testing:**
   - A company tests two versions of a website to see which drives more sales.
   - **Null Hypothesis ($H_0$):** Both versions generate the same sales.
   - **Result:** The p-value is 0.08.
     - **Interpretation:** Since $0.08 > 0.05$, the null hypothesis is not rejected, indicating no statistically significant difference.
3. **Environmental Studies:**
   - Researchers test if a new policy reduces pollution levels.
   - **Null Hypothesis ($H_0$):** The policy has no effect on pollution.
   - **Result:** The p-value is 0.001.
     - **Interpretation:** Since $0.001 < 0.01$, the null hypothesis is rejected, providing strong evidence that the policy reduces pollution.

---

# IX .Z-Test and T-Test (One-Sample, Two-Sample)

### 1. Introduction to Z-Test and T-Test

- **Definition:** Z-tests and t-tests are statistical methods used to determine whether there is a significant difference between sample data and a population or between two samples.
- **Purpose:** Both tests are used to test hypotheses about population parameters (mean or proportion) based on sample data.
- **Difference:**
  - **Z-Test:** Used when the population variance is known or when the sample size is large (typically $n>30$ ).
  - **T-Test:** Used when the population variance is unknown and the sample size is small (typically $n \leq 30$).

---

### 2. Z-Test

### 2.1 One-Sample Z-Test

- **Purpose:** To determine whether the sample mean is significantly different from a known population mean.
- **Formula:**

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\bar{X}$ : Sample mean

- μ = Population mean
- σ = Population standard deviation
- n › Sample size
- **Example:**
  - A factory claims that the average weight of its product is 500g. A sample of 40 products shows an average weight of 495g with a known standard deviation of 10g. Is there a significant difference?

### 2.2 Two-Sample Z-Test

- **Purpose:** To compare the means of two independent samples to determine if they come from populations with the same mean.

Formula

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- $\bar{X}_1, \bar{X}_2$ › Sample means
- σ1,σ2 › Population standard deviations n1,n2›
- Sample sizes
- D0› Hypothesized difference between population means (often 0)
- **Example:**
  - Comparing the average test scores of two different schools with large sample sizes and known population variances.

---

## 3. T-Test

### 3.1 One-Sample T-Test

- **Purpose:** To determine whether the sample mean is significantly different from a known population mean when the population variance is unknown.
- **Formula:**

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Where:

- $\bar{X}$ › Sample mean
- μ = Population mean
- s › Sample standard deviation n
- › Sample size
- **Example:**
  - Testing if the average height of students in a class differs from the national average height when the population variance is unknown.

### 3.2 Two-Sample T-Test

- **Purpose:** To compare the means of two independent samples when population variances are unknown.
- **Types:**
  - **Independent Two-Sample T-Test:** Compares means from two different groups.
  - **Paired T-Test:** Compares means from the same group at different times.
  - Formula

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{X}$ › Sample mean
- s1,s2 › Sample standard deviations
- n1,n2 › Sample sizes
- **Example:**
  - Comparing the effectiveness of two different teaching methods by analyzing student test scores from two independent groups.

### 4. When to Use Z-Test vs. T-Test

- **Use Z-Test when:**
  - The population standard deviation is known.
  - The sample size is large ().

    n>30
- **Use T-Test when:**
  - The population standard deviation is unknown.
  - The sample size is small ().

    n≤30

### 5. Real-Life Examples

1. **Quality Control (Z-Test):**
   - A company claims that its light bulbs last 1,000 hours on average. A sample of 50 bulbs shows an average life of 990 hours with a known standard deviation of 20 hours. A one-sample z-test can determine if there's a significant difference.
2. **Medical Research (T-Test):**
   - Researchers test whether a new medication lowers blood pressure more effectively than an existing medication. Two groups of patients are tested, and a two-sample t-test is used to compare the average blood pressure reductions.
3. **A/B Testing (T-Test):**
   - A company tests two versions of a website to see which one results in more sales. A two-sample t-test is used to compare the average sales between the two groups.

## X. Chi-Square Test for Independence

### 1. Introduction to Chi-Square Test for Independence

- **Definition:** The Chi-Square Test for Independence is a statistical test used to determine if there is a significant association between two categorical variables.
- **Purpose:** It assesses whether the distribution of sample categorical data matches an expected distribution under the assumption of independence.

### 2. When to Use the Chi-Square Test for Independence

- When you have two categorical variables from a single population.
- To test if the variables are independent or associated.
- Example scenarios:
  - Testing if gender is related to voting preference.
  - Checking if product type influences customer satisfaction levels.

## 3. Assumptions of the Chi-Square Test

- The data is in the form of frequencies or counts of cases.
- The observations are independent.
- The expected frequency in each cell of the contingency table should be at least 5.

## 4. How the Chi-Square Test Works

- **Step 1:** Create a contingency table displaying the frequency distribution of the variables.
- **Step 2:** Calculate the expected frequencies for each cell assuming the null hypothesis of independence.
- **Step 3:** Use the Chi-Square formula to compare observed and expected frequencies.

**Chi-Square Formula:**

$$X2 = X^{2} = \sum \frac{(O-E)^2}{E}$$

Where:

- O › Observed frequency E
- › Expected frequency

## 5. Hypotheses

- **Null Hypothesis (H0):** There is no association between the two categorical variables (they are independent).
- **Alternative Hypothesis (Ha):** There is an association between the two categorical variables (they are not independent).

## 6. Interpreting the Results

- Compare the calculated $\chi^2$ statistic to the critical value from the Chi-Square distribution table (based on degrees of freedom and significance level).
- **If $\chi^2$ > critical value** or **p-value < α**, reject the null hypothesis.
- **If $\chi^2$ \leq** critical value or **p-value ≥α**, do not reject the null hypothesis.

## 7. Real-Life Example

**Scenario:** A researcher wants to examine if there is an association between gender (male, female) and preference for a new product (like, dislike).

1. **Data Collection:** A survey is conducted, and the responses are tabulated as follows:

| | Like | Dislike | Total |
|---|---|---|---|
| **Male** | 30 | 20 | 50 |
| **Female** | 40 | 10 | 50 |
| **Total** | 70 | 30 | 100 |

1. **Expected Frequencies:**
   - For males who like the product:
     $E = \frac{50 \times 70}{100} = 35$
   - For females who dislike the product:
     $E = \frac{50 \times 30}{100} = 15$
2. **Chi-Square Calculation:**
   - Apply the Chi-Square formula to all cells.
3. **Interpretation:**

- If the p-value is less than 0.05, the researcher concludes there is a significant association between gender and product preference.

# XI. ANOVA (Analysis of Variance)

## 1. Introduction to ANOVA

- **Definition:** ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more groups to determine if at least one group mean is significantly different from the others.
- **Purpose:** It helps in understanding whether variations between group means are due to actual differences or random chance.

## 2. When to Use ANOVA

- When comparing means across three or more groups.
- When the dependent variable is continuous, and the independent variable is categorical.
- Example scenarios:
  - Comparing the average test scores of students from different teaching methods.
  - Analyzing the effectiveness of different marketing campaigns on sales.

## 3. Assumptions of ANOVA

- The samples are independent.
- The data within each group is normally distributed.
- Homogeneity of variances: The variance among the groups should be approximately equal.

## 4. How ANOVA Works

- **Step 1:** Calculate the overall mean of all groups.
- **Step 2:** Compute the variance between the group means (Between-Group Variance).
- **Step 3:** Compute the variance within each group (Within-Group Variance).
- **Step 4:** Use the ANOVA formula to compare these variances.

**ANOVA Formula:**

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}}$$

Where:

- **Variance Between Groups:** Measures how much the group means differ from the overall mean.
- **Variance Within Groups:** Measures the variability within each individual group.

## 5. Hypotheses

- **Null Hypothesis (H0):** All group means are equal (no significant difference).
- **Alternative Hypothesis (Ha):** At least one group mean is different.

## 6. Types of ANOVA

- **One-Way ANOVA:** Compares means of three or more independent groups based on one factor.
  - Example: Comparing the average scores of students from three different schools.
- **Two-Way ANOVA:** Examines the effect of two factors on a dependent variable, including their interaction.
  - Example: Studying the impact of different diets and exercise programs on weight loss.

- **Repeated Measures ANOVA:** Used when the same subjects are measured multiple times under different conditions.
  - Example: Measuring blood pressure of patients at different time intervals after medication.

### 7. Interpreting ANOVA Results

- The **F-statistic** is calculated and compared to a critical value from the F-distribution table.
- **If** F **> critical value** or **p-value < α\alpha**, reject the null hypothesis.
- **If** F**\leq** critical value or **p-value ≥α\geq \alpha**, do not reject the null hypothesis.
- If the null hypothesis is rejected, post-hoc tests (like Tukey's HSD) are used to identify which specific groups differ.

### 8. Real-Life Example

**Scenario:** A researcher wants to evaluate the effectiveness of three different diets (Diet A, Diet B, and Diet C) on weight loss.

1. **Data Collection:** 30 participants are randomly assigned to one of the three diets, and their weight loss is measured after 8 weeks.
2. **Hypotheses:**
   - **Null Hypothesis (H0):** All three diets result in the same average weight loss.
   - **Alternative Hypothesis (Ha):** At least one diet leads to a different average weight loss.
3. **ANOVA Calculation:**
   - Calculate the F-statistic based on the between-group and within-group variances.
4. **Interpretation:**
   - If the p-value is less than 0.05, the researcher concludes that at least one diet has a significantly different effect on weight loss. Further analysis (post-hoc tests) will determine which diets differ.

## XII. Power of a Test

### 1. Introduction to Power of a Test

- **Definition:** The power of a statistical test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (Ha) is true. In other words, it measures a test's ability to detect an actual effect when there is one.
- **Importance:** High power reduces the risk of committing a Type II error (false negative), ensuring that meaningful differences or effects are not overlooked.

### 2. Key Concepts Related to Power

- **Significance Level (α\alpha):** The probability of committing a Type I error (false positive). Commonly set at 0.05.
- **Effect Size:** The magnitude of the difference or relationship being tested. Larger effect sizes increase the power of the test.
- **Sample Size (n):** Larger samples provide more reliable estimates and increase power.
- **Variance:** Lower variability within the data increases power.

### 3. Factors Affecting the Power of a Test

1. **Sample Size:** Increasing the sample size reduces the standard error, making it easier to detect true effects.
2. **Effect Size:** Larger differences between groups are easier to detect, increasing the power.
3. **Significance Level (α):** A higher significance level (e.g., 0.10 instead of 0.05) increases power but also increases the risk of a Type I error.
4. **Variance in Data:** Less variability within groups makes it easier to detect differences, increasing power.

### 4. Calculating Power

- **Power Analysis:** A statistical technique used to determine the required sample size to achieve a desired power level (typically 80% or 0.80).
- **Software Tools:** Power can be calculated using software like R, Python (Statsmodels), or G*Power.

**Power Formula (simplified for basic tests):**

$$\text{Power} = 1 - \beta$$

Where βis the probability of a Type II error.

---

### 5. Interpreting Power

- **High Power (>0.80):** Indicates a strong likelihood of detecting an effect if it exists.
- **Low Power (<0.80):** Increases the risk of missing a true effect, leading to inconclusive results.
- **Balancing Power and Error Rates:** While increasing power is desirable, it must be balanced against the risk of Type I errors and practical constraints like sample size and cost.

---

### 6. Real-Life Example

**Scenario:** A pharmaceutical company is testing a new drug to lower blood pressure. They want to ensure that their study can detect a meaningful reduction in blood pressure if the drug is effective.

1. **Hypotheses:**
   - **Null Hypothesis (H0):** The drug has no effect on blood pressure.
   - **Alternative Hypothesis (Ha):** The drug lowers blood pressure.
2. **Power Analysis:**
   - The researchers conduct a power analysis and determine that a sample size of 100 participants per group is needed to achieve 80% power at a significance level of 0.05.
3. **Outcome:**
   - If the study detects a significant reduction in blood pressure, the researchers can be confident that the effect is real.
   - If no significant effect is found, they can be reasonably confident that the drug likely has no effect, given the high power of the test.

---

## XIII. Python Codes

```
import numpy as np
from scipy import stats
from statsmodels.stats.power import TTestIndPower

# 1. Z-Test (One-Sample)
def z_test_one_sample(data, population_mean, population_std):
    sample_mean = np.mean(data)
    n = len(data)
    z_statistic = (sample_mean - population_mean) / (population_std / np.sqrt(n))
    p_value = stats.norm.sf(abs(z_statistic)) * 2  # two-tailed test
    return z_statistic, p_value

# 2. Z-Test (Two-Sample)
def z_test_two_sample(data1, data2, pop_std1, pop_std2):
    mean1, mean2 = np.mean(data1), np.mean(data2)
    n1, n2 = len(data1), len(data2)
    z_statistic = (mean1 - mean2) / np.sqrt((pop_std1**2 / n1) + (pop_std2**2 / n2))
    p_value = stats.norm.sf(abs(z_statistic)) * 2  # two-tailed test
    return z_statistic, p_value

# 3. T-Test (One-Sample)
```

```
def t_test_one_sample(data, population_mean):
    t_statistic, p_value = stats.ttest_1samp(data, population_mean)
    return t_statistic, p_value

# 4. T-Test (Two-Sample Independent)
def t_test_two_sample(data1, data2):
    t_statistic, p_value = stats.ttest_ind(data1, data2)
    return t_statistic, p_value

# 5. T-Test (Paired)
def t_test_paired(data1, data2):
    t_statistic, p_value = stats.ttest_rel(data1, data2)
    return t_statistic, p_value

# 6. Chi-Square Test for Independence
def chi_square_test(contingency_table):
    chi2_statistic, p_value, dof, expected = stats.chi2_contingency(contingency_table)
    return chi2_statistic, p_value, dof, expected

# 7. ANOVA (One-Way)
def anova_one_way(*groups):
    f_statistic, p_value = stats.f_oneway(*groups)
    return f_statistic, p_value

# 8. Power of a Test (T-Test Independent)
def power_of_test_ttest(sample_size, effect_size, alpha=0.05):
    power_analysis = TTestIndPower()
    power = power_analysis.power(effect_size=effect_size, nobs1=sample_size, alpha=alpha)
    return power

# Example data for demonstration
np.random.seed(0)
data1 = np.random.normal(100, 10, 30)
data2 = np.random.normal(105, 10, 30)
contingency_table = np.array([[30, 20], [40, 10]])

# Perform tests
z_one_sample_result = z_test_one_sample(data1, population_mean=100, population_std=10)
z_two_sample_result = z_test_two_sample(data1, data2, pop_std1=10, pop_std2=10)
t_one_sample_result  =  t_test_one_sample(data1,  population_mean=100)
t_two_sample_result = t_test_two_sample(data1, data2)
t_paired_result = t_test_paired(data1, data2)
chi_square_result = chi_square_test(contingency_table)
anova_result = anova_one_way(data1, data2)
power_result  =  power_of_test_ttest(sample_size=30,  effect_size=0.5)

# Compile results
results = {
    'Z-Test  (One-Sample)':  z_one_sample_result,
    'Z-Test  (Two-Sample)':  z_two_sample_result,
    'T-Test  (One-Sample)': t_one_sample_result,
    'T-Test (Two-Sample Independent)': t_two_sample_result,
    'T-Test (Paired)': t_paired_result,
    'Chi-Square Test': chi_square_result,
    'ANOVA (One-Way)': anova_result,
    'Power of T-Test': power_result
}

results
```