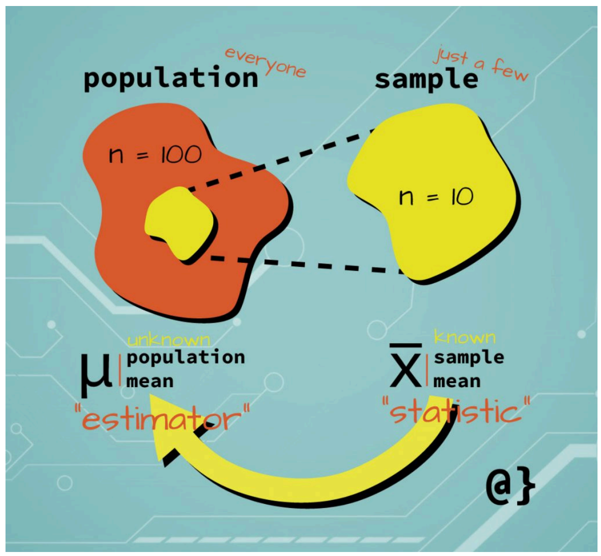
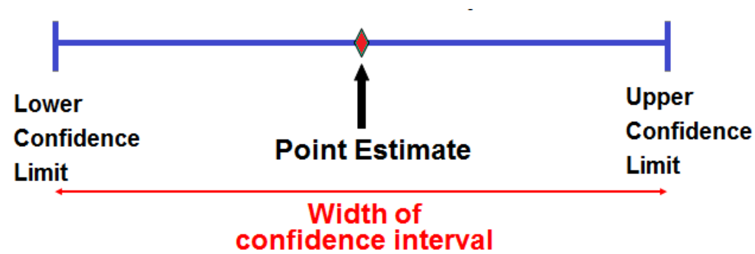
Estimates in Statistics

An estimate is a value or a set of values derived from sample data, used to infer information about an unknown population parameter. Estimation is a fundamental aspect of inferential statistics and involves two main types:

1. **Point Estimate**: A single value computed from sample data to represent a population parameter (e.g., sample mean as an estimate of population mean).
2. **Interval Estimate**: A range of values within which the population parameter is expected to lie, often accompanied by a confidence level (e.g., 95% confidence interval).





Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or draw conclusions about a population based on sample data. It involves formulating and testing assumptions (hypotheses) about a population parameter.

Key Concepts:

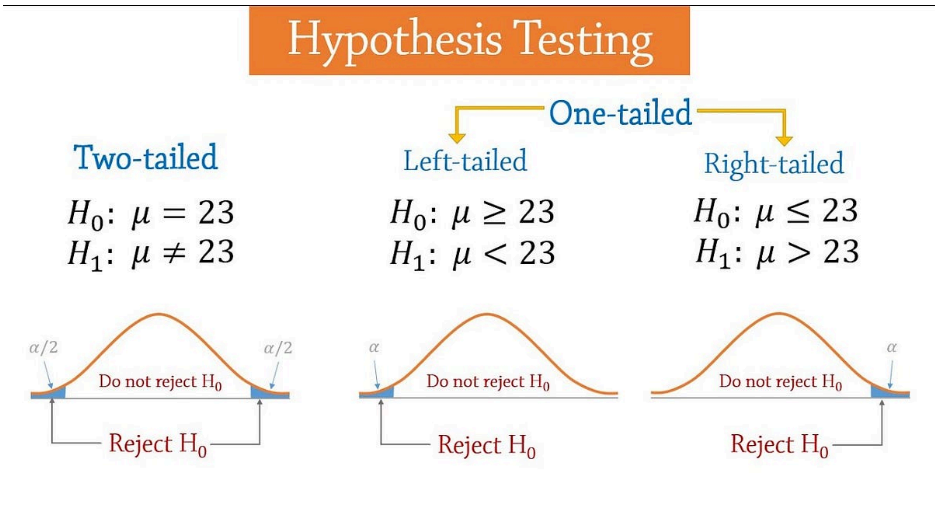
1. **Null Hypothesis (H0)**:
   The default assumption that there is no effect or no difference. It represents the "status quo."
   - Example: H0: $\mu=\mu_0$ (the population mean equals a specified value).
2. **Alternative Hypothesis (Ha)**:
   The claim we seek to support, indicating the presence of an effect or difference. Opposite of null hypothesis.
   - Example: Ha:$\mu\neq\mu_0$ (the population mean is not equal to a specified value).
3. **Test Statistic**:
   A standardized value calculated from sample data to test the hypotheses. Examples include the z-statistic, t-statistic, or F-statistic.
4. **Significance Level (α)**:
   The probability of rejecting the null hypothesis when it is true (Type I error). Common choices are 0.05, 0.01, or 0.10.
5. **P-Value**:
   The probability of observing the test statistic or something more extreme, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against H0.
6. **Critical Value**:
   A threshold that defines the rejection region for the null hypothesis. If the test statistic exceeds the critical value, H0 is rejected.
7. **Decision**:
   - Reject H0: If the test statistic falls in the rejection region (p-value < α).
   - Fail to Reject H0: If there is insufficient evidence to support Ha.

Steps in Hypothesis Testing:

1. **Formulate Hypotheses**:
   - Null hypothesis (H0).
   - Alternative hypothesis (Ha).
2. **Select a Test and Assumptions**: Choose the appropriate statistical test based on data type and assumptions (e.g., normality, sample size).
3. **Set the Significance Level (α)**: Common values: 0.05, 0.01.
4. **Compute the Test Statistics**: Use sample data to calculate the statistic (e.g., z, t).
5. **Determine the P-Value or Compare to Critical Value**:
   - Use the test statistic to find the p-value or compare with the critical value.
6. **Make a Decision**:
   - Reject H0 if p-value < α or test statistic exceeds the critical value.
7. **Draw Conclusions**: Interpret the results in the context of the research question.

Common Tests:

1. **Z-Test**: For population means or proportions with a large sample size or known variance.
2. **T-Test**: For population means with a small sample size or unknown variance.
3. **Chi-Square Test**: For categorical data or goodness-of-fit.
4. **ANOVA**: To compare means across multiple groups.



A factory claims that the average weight of its product is 500 grams. A quality control officer takes a random sample of 25 products and finds an average weight of 495 grams with a standard deviation of 10 grams. Test at the 5% significance level whether the claim is true.

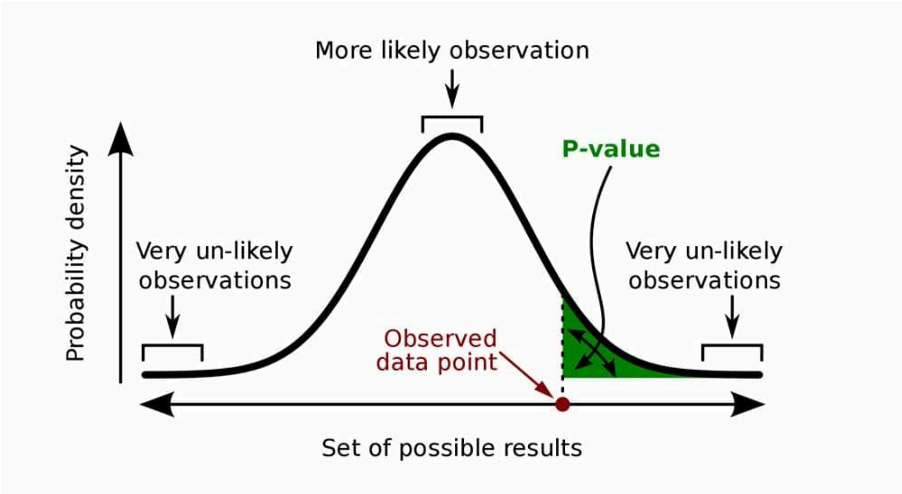
1. **Hypotheses**:
   - $H_0 : \mu = 500$ (The mean weight is 500 grams).
   - $H_a : \mu \neq 500$ (The mean weight is not 500 grams).

P-Value in Hypothesis Testing

The **p-value** (probability value) is a key concept in hypothesis testing. It quantifies the probability of observing a test statistic at least as extreme as the one obtained, assuming the null hypothesis (H0) is true.

Key Points:

1. **Interpretation**:
   - A **small p-value** ($p < \alpha$): Strong evidence against H_0; reject H_0.
   - A **large p-value** ($p \geq \alpha$): Insufficient evidence to reject H_0; fail to reject H_0.
   - Typical significance levels (α) are 0.05, 0.01, or 0.10.
2. **Significance Level** (α):
   - α is the threshold for deciding whether the p-value indicates sufficient evidence to reject H_0.
   - For example, $\alpha = 0.05$ means a 5% risk of rejecting H_0 when it is true.
3. **P-Value and Test Statistic**:
   - The p-value is calculated using the test statistic (e.g., z-statistic, t-statistic) and the sampling distribution under H_0.
   - It represents the area in the tails of the distribution beyond the observed test statistic.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Scenario:

A factory claims the average weight of a product is 500 grams. A sample of 30 products has a mean weight of 495 grams and a standard deviation of 10 grams. Test at α=0.05.

Steps:

1. **Hypotheses**:
   - $H_0 : \mu = 500$
   - $H_a : \mu \neq 500$
2. **Test Statistic**:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{495 - 500}{10/\sqrt{30}} = -2.74$$

3. **P-Value**:
   - For $z = -2.74$, the p-value (from a z-table or software) is approximately 0.0062 (two-tailed).
4. **Decision**:
   - $p = 0.0062 < 0.05$, so reject H_0.
5. **Conclusion**: There is strong evidence to conclude that the mean weight is not 500 grams.

Z-Test

A **z-test** is a statistical test used to determine whether there is a significant difference between the sample statistic (e.g., sample mean or proportion) and a population parameter, or between two sample statistics, assuming the data follows a normal distribution or the sample size is large(n>30).

Assumptions

- Data is approximately normally distributed (or n is large enough for the Central Limit Theorem to apply).
- Population variance is known (or approximated by the sample variance).
- Observations are independent.

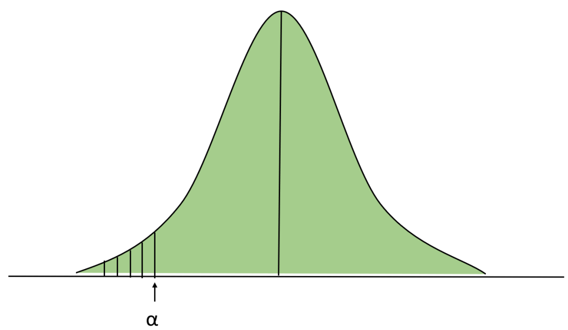
Z Test Statistics Formula

$$\text{Z Test} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Type of Z-test

Left-tailed Test

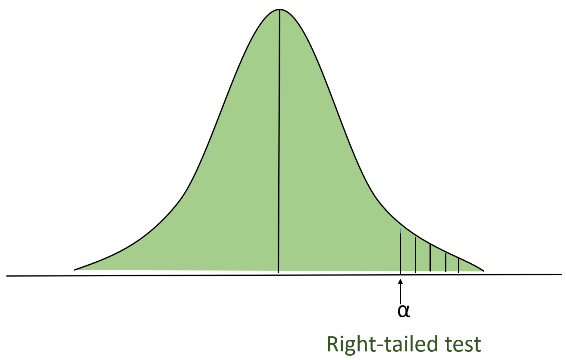
In this test, our region of rejection is located to the extreme left of the distribution.



Left-tailed Test

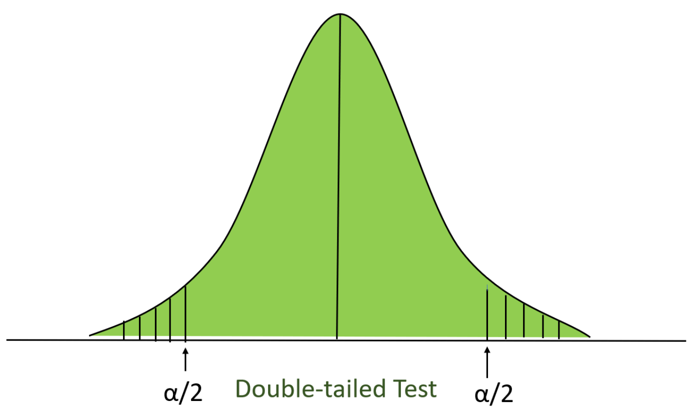
Right-tailed Test

In this test, our region of rejection is located to the extreme right of the distribution. Here our null hypothesis is that the claimed value is less than or equal to the mean population value.



Two-tailed test

In this test, our region of rejection is located to both extremes of the distribution. Here our null hypothesis is that the claimed value is equal to the mean population value.



Two-Tailed Z-Test Example

Scenario:
A company claims that the average life of its LED bulbs is 1000 hours. A sample of 50 bulbs has a mean lifespan of 980 hours with a standard deviation of 30 hours. Test the claim at a 5% significance level ($\alpha=0.05$).

Solution:

1. **State the Hypotheses**:
    - **Null Hypothesis (H0)**: The mean lifespan of the bulbs is 1000 hours (μ=1000).
    - **Alternative Hypothesis (Ha)**: The mean lifespan of the bulbs is not 1000 hours (μ≠1000).

2. **Choose the Significance Level** (α):

- Significance level (α) = 0.05 (5%).
- For a two-tailed test, the critical regions are divided equally between both tails:

$$\alpha_{left} = \alpha_{right} = \frac{\alpha}{2} = 0.025$$

---

3. **Calculate the Test Statistic**: Use the formula for the z-statistic:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where:

- $\bar{x} = 980$: Sample mean
- $\mu = 1000$: Population mean (claimed)
- $\sigma = 30$: Sample standard deviation
- $n = 50$: Sample size

Substitute the values:

$$z = \frac{980 - 1000}{30/\sqrt{50}} = \frac{-20}{30/7.071} = \frac{-20}{4.243} \approx -4.71$$

Locate Critical Values Using the Z-Table:

- **Left Tail** (z_{left}): Find the z-value corresponding to a cumulative probability of 0.025. From the z-table:

$$z_{left} = -1.96$$

- **Right Tail** (z_{right}): Find the z-value corresponding to a cumulative probability of $1 - 0.025 = 0.975$. From the z-table:

$$z_{right} = +1.96$$

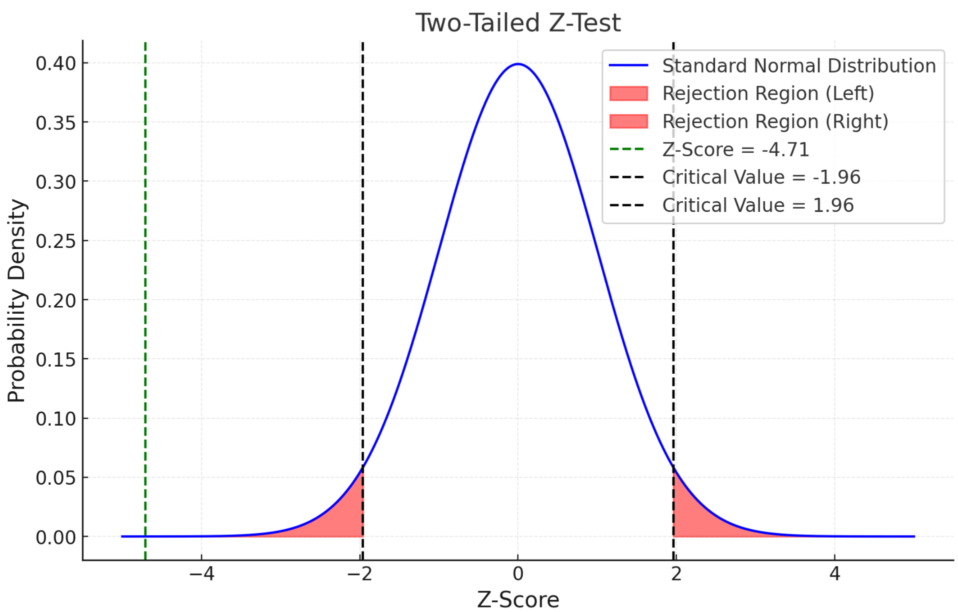
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

5. **Make the Decision**:

    ○ Compare the calculated z=−4.71 to the critical z-values.
    ○ Since z=−4.71<−1.96, the test statistic falls in the rejection region of the left tail.
6. **Conclusion**:
    ○ **Reject H0**.
      There is strong evidence to conclude that the mean lifespan of the bulbs is not 1000 hours.

Visual Representation:

The standard normal distribution, critical regions, and the calculated z-statistic.



The chart above illustrates the two-tailed z-test:

- The **blue curve** represents the standard normal distribution.
- The **red shaded areas** on both ends show the rejection regions (z<−1.96 and z>1.96) for α=0.05.
- The **green dashed line** shows the calculated z=−4.71, which falls within the left rejection region.

One-Tailed Z-Test Example

Scenario:
A pharmaceutical company claims that their new drug reduces blood pressure by at least 10 mmHg on average. A sample of 36 patients shows a mean reduction of 8 mmHg with a

standard deviation of 3 mmHg. Test the claim at a 5% significance level (α=0.05) using a **one-tailed z-test**.

Solution:

1. State the Hypotheses:

- **Null Hypothesis (H0)**: The mean reduction in blood pressure is at least 10 mmHg (μ≥10).
- **Alternative Hypothesis (Ha)**: The mean reduction in blood pressure is less than 10 mmHg (μ<10). *(This is a **left-tailed test**.)*

2. Choose the Significance Level (α):

- Significance level (α) = 0.05.
- Since it's a one-tailed test, the rejection region is entirely on the **left side**.

3. Calculate the Test Statistic (Z-Score):

Use the formula for the z-statistic:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Where:

- $\bar{x} = 8$: Sample mean
- $\mu = 10$: Claimed mean
- $\sigma = 3$: Population standard deviation
- $n = 36$: Sample size

Substitute the values:

$$z = \frac{8 - 10}{3/\sqrt{36}} = \frac{-2}{3/6} = \frac{-2}{0.5} = -4.0$$

4. Find the Critical Z-Value:

For a left-tailed test at α=0.05:

- From the z-table, the critical z-value for a cumulative probability of 0.05 is:

zcritical=−1.645

5. Make the Decision:

- **Compare $z_{calculated}$ with $z_{critical}$:**

$$z_{calculated} = -4.0, \quad z_{critical} = -1.645$$

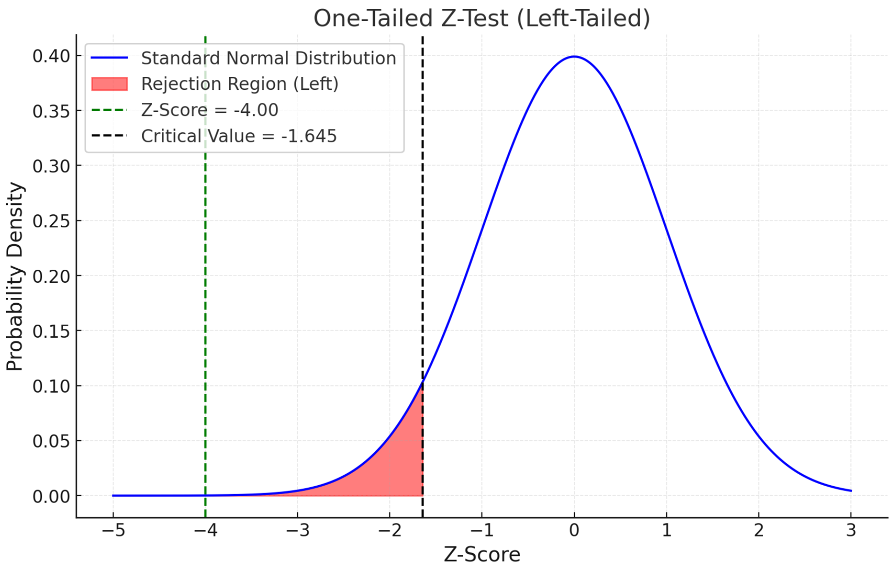
- Since $z_{calculated} < z_{critical}$, the test statistic falls in the **rejection region**.

6. Conclusion:

- **Reject H0**.
  There is strong evidence to conclude that the drug reduces blood pressure by less than 10 mmHg on average.

Visualization:

The rejection region and the calculated z-statistic for this left-tailed test.



The chart above illustrates the one-tailed z-test:

- The **blue curve** represents the standard normal distribution.
- The **red shaded region** shows the rejection region ($z < -1.645$) for a left-tailed test at $\alpha = 0.05$.
- The **green dashed line** indicates the calculated z-score (-4.0), which falls within the rejection region.

T-Test: Overview

A **t-test** is a statistical test used to determine whether there is a significant difference between the means of one or more groups, particularly when the sample size is small (n<30) or when the population standard deviation is unknown.

Types of T-Tests

1. **One-Sample T-Test**:
    - Compares the mean of a single sample to a known or hypothesized population mean.
    - Example: Testing if the average height of students in a class differs from the national average height.
2. **Independent Two-Sample T-Test**:
    - Compares the means of two independent groups.
    - Example: Comparing test scores between two different schools.
3. **Paired (Dependent) T-Test**:
    - Compares the means of two related groups (e.g., before and after a treatment).
    - Example: Measuring weight loss before and after a fitness program for the same individuals.

When to Use a T-Test?

- Data is approximately normally distributed.
- The scale of measurement is continuous (e.g., height, weight, test scores).
- The sample size is small (n<30).
- Population standard deviation is unknown.

T-Test Formula

The formula for the t-statistic depends on the type of t-test. For a one-sample t-test:

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

Where:

- $\bar{x}$: Sample mean
- μ: Population mean (hypothesized mean)
- s: Sample standard deviation
- n: Sample size

Degrees of Freedom (df)

Degrees of Freedom (df) refers to the number of independent values in a dataset that are free to vary while calculating a statistic, such as the mean or variance. It plays a critical role in various statistical tests, particularly in determining critical values from distributions like the t-distribution or chi-distribution.

In general, degrees of freedom account for the constraints imposed by using sample data. For example:

- If we compute the sample mean, one degree of freedom is "used up" because all values must sum up to match the mean.
- As a result, degrees of freedom are typically calculated as the sample size minus the number of estimated parameters.

Formula for Degrees of Freedom

1. **One-Sample T-Test**:

$$df = n - 1$$

- n: Sample size.

2. **Independent Two-Sample T-Test**:

$$df = n_1 + n_2 - 2$$

- n_1, n_2: Sample sizes of the two groups.

3. **Paired T-Test**:

$$df = n - 1$$

- n: Number of pairs.

4. χ^2-**Test**:

$$df = (r - 1)(c - 1)$$

- r: Number of rows.
- c: Number of columns.

Example: One-Sample T-Test

Scenario:
A teacher claims that the average score in her class is 75. A random sample of 10 students gives the following scores:
70,68,75,80,74,72,77,78,69,7370, 68, 75, 80, 74, 72, 77, 78, 69, 73,70,68,75,80,74,72,77,78,69,73

Test the claim at a 5% significance level.

Step-by-Step Solution

1. **State the Hypotheses**:
   - Null Hypothesis (H_0): The mean score is 75 ($\mu = 75$).
   - Alternative Hypothesis (H_a): The mean score is not 75 ($\mu \neq 75$).
2. **Calculate the Sample Mean ($\bar{x}$) and Standard Deviation (s)**:

$$\bar{x} = \frac{\text{Sum of scores}}{\text{Number of scores}}$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

3. **Compute the Test Statistic (t)**: Use the t-test formula:

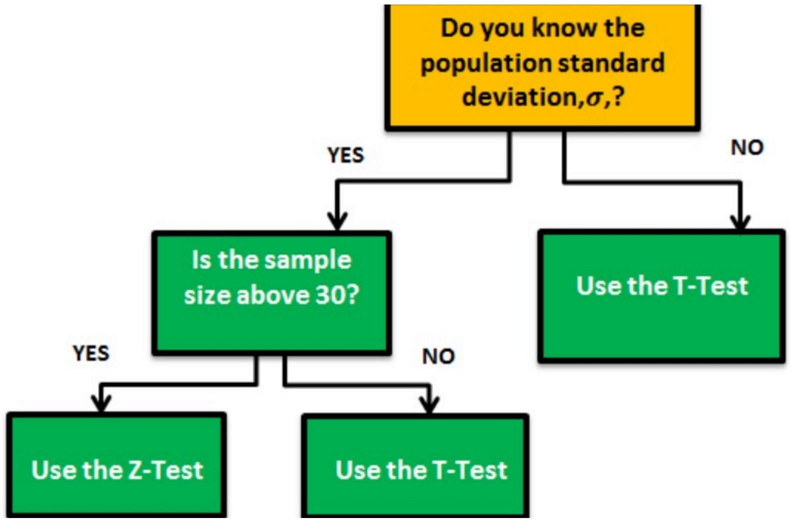
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

4. **Find the Critical T-Value**:
   - Degrees of freedom (df) = $n - 1$ = 9.
   - Use a t-table or software to find the critical t-value for a two-tailed test with $\alpha = 0.05$.
5. **Decision Rule**:
   - If $|t| > t_{critical}$, reject H_0.
   - Otherwise, fail to reject H_0.
6. **Conclusion**: Compare the calculated t-value with the critical t-value to make a decision.

t-test table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Z-test vs T-test



Type 1 and Type 2 Errors

In hypothesis testing, errors can occur when making decisions about rejecting or not rejecting the null hypothesis (H0). These errors are classified as **Type 1** and **Type 2** errors.

Type 1 Error (α):

- Occurs when the **null hypothesis (H0) is true**, but we incorrectly reject it.
- It is also called a **false positive**.
- The probability of making a Type 1 error is represented by the **significance level (α)**, often set to 0.05 or 5%.

Example:

- A medical test concludes that a patient has a disease (rejects H0) when they are actually healthy (H0 is true).

Consequences:

- In critical applications (e.g., medicine, justice systems), a Type 1 error can lead to severe consequences like unnecessary treatments or false accusations.

Type 2 Error (β):

- Occurs when the **null hypothesis (H0) is false**, but we fail to reject it.
- It is also called a **false negative**.
- The probability of making a Type 2 error is represented by β, and the **power of a test** is $1-\beta$.

Example:

- A medical test concludes that a patient does not have a disease (fails to reject H0) when they actually have it (H0 is false).

Consequences:

- Missing a genuine effect or failing to detect a true positive outcome (e.g., missing a diagnosis).

Balancing Type 1 and Type 2 Errors

- **Reducing α** decreases the likelihood of a Type 1 error but increases the chance of a Type 2 error (and vice versa).
- Larger sample sizes can reduce both errors by increasing the test's power.

Comparison

	Null Hypothesis is True	Null Hypothesis is False
Reject H_0	Type 1 Error (α)	Correct Decision
Fail to Reject H_0	Correct Decision	Type 2 Error (β)

Bayes' Theorem

Bayes' Theorem provides a mathematical framework for updating probabilities based on new evidence.

The Formula

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where:

- $P(A|B)$: **Posterior probability** - Probability of A given B (updated probability after considering B).
- $P(B|A)$: **Likelihood** - Probability of B given A.
- $P(A)$: **Prior probability** - Initial probability of A.
- $P(B)$: **Evidence** - Overall probability of B.

Example: Using Bayes' Theorem in Machine Learning

Scenario:
Imagine you're building a spam filter for emails. You want to predict whether an email is **spam** (S) or **not spam** (N) based on the presence of certain words (features).

Setup

- **Features**:
  - $x_1 = \text{"offer"}$
  - $x_2 = \text{"win"}$
  - $x_3 = \text{"urgent"}$
- **Goal**:
  - Predict whether an email is spam (S) or not spam (N) based on these features.
- **Given Probabilities**:
  - $P(S) = 0.2$ (20% of emails are spam).
  - $P(N) = 0.8$ (80% of emails are not spam).
  - Likelihoods:
    - $P(x_1|S) = 0.8$, $P(x_1|N) = 0.1$
    - $P(x_2|S) = 0.7$, $P(x_2|N) = 0.2$
    - $P(x_3|S) = 0.9$, $P(x_3|N) = 0.3$

Question

If an email contains the words "offer," "win," and "urgent," what is the probability that it is spam (P(S∣x1,x2,x3)?

Solution Using Bayes' Theorem

$$P(S|x_1, x_2, x_3) = \frac{P(x_1, x_2, x_3|S) \cdot P(S)}{P(x_1, x_2, x_3)}$$

1. Compute the Likelihood ($P(x_1, x_2, x_3|S)$):

Assuming the features are independent (naive assumption):

$$P(x_1, x_2, x_3|S) = P(x_1|S) \cdot P(x_2|S) \cdot P(x_3|S)$$

$$P(x_1, x_2, x_3|S) = 0.8 \cdot 0.7 \cdot 0.9 = 0.504$$

2. Compute the Prior Probability ($P(S)$):

$$P(S) = 0.2$$

3. Compute the Evidence ($P(x_1, x_2, x_3)$):

$$P(x_1, x_2, x_3) = P(x_1, x_2, x_3|S) \cdot P(S) + P(x_1, x_2, x_3|N) \cdot P(N)$$

First, compute $P(x_1, x_2, x_3|N)$:

$$P(x_1, x_2, x_3|N) = P(x_1|N) \cdot P(x_2|N) \cdot P(x_3|N)$$

$$P(x_1, x_2, x_3|N) = 0.1 \cdot 0.2 \cdot 0.3 = 0.006$$

Now, compute $P(x_1, x_2, x_3)$:

$$P(x_1, x_2, x_3) = (0.504 \cdot 0.2) + (0.006 \cdot 0.8) = 0.1008 + 0.0048 = 0.1056$$

4. Compute the Posterior Probability ($P(S|x_1, x_2, x_3)$):

$$P(S|x_1, x_2, x_3) = \frac{P(x_1, x_2, x_3|S) \cdot P(S)}{P(x_1, x_2, x_3)}$$

$$P(S|x_1, x_2, x_3) = \frac{0.504 \cdot 0.2}{0.1056} = \frac{0.1008}{0.1056} \approx 0.954$$

Conclusion

The probability that the email is spam, given the presence of the words "offer," "win," and "urgent," is approximately **95.4%**.

Confidence Interval

A **Confidence Interval** provides a range of values within which we expect the true population parameter to lie with a certain level of confidence.

Formula for Confidence Interval:

$$\mathrm{CI} = \hat{x} \pm \mathrm{MoE}$$

Where:

- $\hat{x}$: Sample statistic (e.g., sample mean or sample proportion).
- **MoE**: Margin of Error.
- $\pm$: Indicates the range around the estimate.

Interpretation: A 95% confidence interval means that if we took 100 different samples and calculated a CI for each, approximately 95 of those intervals would contain the true population parameter.

Margin of Error

The **Margin of Error** quantifies the maximum expected difference between the sample statistic and the true population parameter due to sampling variability.

Formula for Margin of Error:

$$\mathrm{MoE} = z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Where:

- z^*: Critical value from the standard normal distribution (e.g., 1.96 for 95% confidence).
- σ: Population standard deviation (if unknown, use the sample standard deviation).
- n: Sample size.

Steps to Calculate CI and MoE

1. **Choose Confidence Level**:
   - Common levels: 90%, 95%, 99%.
   - Determine the critical value (z^*) from a z-table or t-table.
2. **Compute Standard Error (SE)**:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

3. **Calculate Margin of Error**:

$$\mathrm{MoE} = z^* \cdot \mathrm{SE}$$

4. **Determine the Confidence Interval**:

$$\mathrm{CI} = \hat{x} \pm \mathrm{MoE}$$

Example

Scenario: A researcher wants to estimate the average height of students in a school.

- Sample mean (x^) = 170 cm.
- Sample standard deviation (s) = 10 cm.
- Sample size (n) = 25.
- Confidence level = 95%.

Step 1: Find the critical value (z^*): For a 95% confidence level, $z^* = 1.96$.

Step 2: Compute Standard Error (SE):

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Step 3: Calculate Margin of Error (MoE):

$$\text{MoE} = z^* \cdot \text{SE} = 1.96 \cdot 2 = 3.92$$

Step 4: Determine the Confidence Interval:

$$\text{CI} = \hat{x} \pm \text{MoE} = 170 \pm 3.92$$

$$\text{CI} = (166.08, 173.92)$$

Interpretation: We are 95% confident that the true average height of students lies between **166.08 cm and 173.92 cm**.

Key Points

1. **Wider Intervals**:
    - Lower confidence level (e.g., 90%) → Narrower interval.
    - Higher confidence level (e.g., 99%) → Wider interval.
2. **Larger Sample Sizes**:
    - Reduce the margin of error, making the confidence interval narrower.

Chi-Square Test

The **Chi-Square Test** is a statistical method used to determine whether there is a significant association between categorical variables or whether observed data fits an expected distribution. It is a non-parametric test, meaning it does not assume the data follows a normal distribution.

Types of Chi-Square Tests

1. **Chi-Square Test of Independence**:
    - Used to test if two categorical variables are independent of each other.
2. **Chi-Square Goodness-of-Fit Test**:
    - Used to test if the observed data fits a specific theoretical or expected distribution.

Formula

For both types of tests, the test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i: Observed frequency in each category.
- E_i: Expected frequency in each category.
- χ^2: Chi-square statistic.

Chi-Square Goodness-of-Fit Test Example

Scenario:

A candy company claims that their bag of candies contains the following proportions of colors:

- Red: 30%
- Blue: 20%
- Green: 20%
- Yellow: 15%
- Orange: 15%

A customer randomly selects 100 candies from a bag and counts the colors:

Color	Observed Frequency (O)
Red	40
Blue	25
Green	20
Yellow	10
Orange	5

Test whether the observed distribution matches the company's claim at a significance level of α=0.05.

Step 1: State the Hypotheses

- **Null Hypothesis (H0)**: The observed frequencies match the expected proportions.

- **Alternative Hypothesis (H1)**: The observed frequencies do not match the expected proportions.

Step 2: Calculate Expected Frequencies

The expected frequency (E) for each category is calculated as:

$E=p \cdot N$

Where:

- p: Proportion of each color (given in the claim).
- N: Total number of candies sampled (100 in this case).

Color	p	$E = p \cdot 100$
Red	0.30	$0.30 \cdot 100 = 30$
Blue	0.20	$0.20 \cdot 100 = 20$
Green	0.20	$0.20 \cdot 100 = 20$
Yellow	0.15	$0.15 \cdot 100 = 15$
Orange	0.15	$0.15 \cdot 100 = 15$

Step 3: Compute the Chi-Square Statistic

Use the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

For each color:

- Red:

$$\frac{(40-30)^2}{30} = \frac{100}{30} \approx 3.33$$

- Blue:

$$\frac{(25-20)^2}{20} = \frac{25}{20} = 1.25$$

- Green:

$$\frac{(20-20)^2}{20} = \frac{0}{20} = 0.0$$

- Yellow:

$$\frac{(10-15)^2}{15} = \frac{25}{15} \approx 1.67$$

- Orange:

$$\frac{(5-15)^2}{15} = \frac{100}{15} \approx 6.67$$

Summing these values:

χ2=3.33+1.25+0.0+1.67+6.67=12.92 = 3.33 + 1.25 + 0.0 + 1.67 + 6.67 = 12.92

χ2=3.33+1.25+0.0+1.67+6.67=12.92

Step 4: Find the Degrees of Freedom

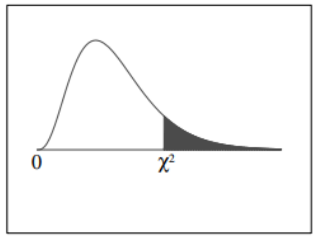
df=Number of Categories−1=5−1=4 = 5 - 1 = 4

Step 5: Determine the Critical Value

From the Chi-Square distribution table:

- For α=0.05 and df=4, the critical value is **9.488**.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490

Step 6: Make the Decision

- **Test Statistic**: χ2=12.92
- **Critical Value**: 9.488

Since 12.92>9.488, we **reject the null hypothesis** (H0).

Step 7: Conclusion

The observed distribution of candy colors **does not match** the company's claimed proportions.

ANOVA (Analysis of Variance)

ANOVA is a statistical technique used to determine whether there are significant differences between the means of three or more independent groups. It extends the t-test to multiple groups and helps answer the question: **"Do the group means significantly differ?"**

Types of ANOVA

1. **One-Way ANOVA**
   Compares means of three or more groups based on one independent variable.
   Example: Comparing test scores among students in three different teaching methods.
2. **Two-Way ANOVA**
   Compares means based on two independent variables and their interaction.
   Example: Examining the effects of diet and exercise on weight loss.
3. **Repeated Measures ANOVA**
   Used when the same subjects are measured under different conditions or over time.
   Example: Testing the effect of a drug on blood pressure at different time points.
4. **MANOVA (Multivariate ANOVA)**
   Extends ANOVA to include multiple dependent variables.
   Example: Testing the impact of a training program on both productivity and satisfaction.
5. ### Factorial ANOVA
   a. **Purpose**: A generalization of two-way ANOVA to include more than two factors.
   b. **Example**: Investigating the effects of diet, exercise, and sleep patterns on weight loss.
   c. **Key Feature**: Multiple factors with interactions among them.

Key Assumptions of ANOVA

1. **Independence**
   Observations are independent of each other.
2. **Normality**
   Data in each group should be approximately normally distributed.
3. **Homogeneity of Variance**
   The variance among the groups should be roughly equal (tested using Levene's Test).

---

1. Factors

A **factor** is an independent variable that is hypothesized to influence the dependent variable (response variable). It represents the categorical variable being tested.

- **Example**: In an experiment comparing the effect of different teaching methods on student performance, the teaching method is the **factor**.

2. Levels

Levels are the specific categories, conditions, or values of a factor.

- **Example**: For the teaching method factor, the levels might be:
    - Traditional Lecture
    - Online Course
    - Hybrid Approach

Key Points

- **Single Factor ANOVA**: Has one factor with multiple levels.
    - Example: Comparing crop yields using 3 fertilizers (factor: fertilizer, levels: A, B, C).
- **Two-Factor ANOVA**: Has two factors, each with multiple levels.
    - Example: Examining the effects of diet (levels: vegetarian, keto, paleo) and exercise type (levels: yoga, cardio, weights) on weight loss.

Key Terms in ANOVA

- **Null Hypothesis (H_0)**: All group means are equal ($\mu_1 = \mu_2 = \mu_3 \ldots$).
- **Alternative Hypothesis (H_a)**: At least one group mean is different.
- **F-statistic**: Ratio of variance between groups to variance within groups.

$$F = \frac{\text{Between-Group Variance}}{\text{Within-Group Variance}}$$

- **Degrees of Freedom (df)**:
    - Between groups: $k - 1$, where k is the number of groups.
    - Within groups: $N - k$, where N is the total number of observations.

Steps in Performing ANOVA

1. **State Hypotheses**
   Null Hypothesis (H_0): Group means are equal.
   Alternative Hypothesis (H_a): At least one group mean is different.
2. **Check Assumptions**
   - Independence
   - Normality
   - Homogeneity of Variance
3. **Calculate ANOVA Table**
   - Sum of Squares (SS): Measures variability.
     - $SS_{Between}$: Variance due to differences between group means.
     - SS_{Within}: Variance within groups.
   - Mean Square (MS): Average variance ($MS = SS/df$).
   - F-Statistic: Ratio of mean squares.
4. **Determine p-value**
   Compare the F-statistic with critical values or use software to find the p-value.
5. **Make a Decision**
   If $p \leq \alpha$ (e.g., 0.05), reject H_0.

ANOVA Table Structure

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS) = SS/df	F-ratio (MS_between / MS_within)
Between Groups	$SS_{Between}$	$k-1$	$MS_{Between}$	F
Within Groups	SS_{Within}	$N-k$	MS_{Within}	
Total	SS_{Total}	$N-1$		

Applications of ANOVA

- Comparing the effectiveness of different treatments.

- Evaluating marketing strategies across different regions.
- Studying behavioral patterns in psychology.

Limitations

- Sensitive to violations of assumptions.
- Cannot indicate which specific groups differ without post-hoc tests.
- Only tests for differences in means, not other characteristics.

Example Scenario

A researcher wants to test whether three different fertilizers (A, B, and C) lead to different crop yields. The crop yield (in kg) for each fertilizer type is recorded from five plots.

Fertilizer	Crop Yields (kg)
A	20, 22, 23, 21, 20
B	30, 32, 31, 29, 30
C	25, 27, 26, 24, 26

Step 1: State the Hypotheses

- H0: The mean crop yields for all three fertilizers are the same. (μA=μB=μC)
- Ha: At least one mean is different.

Step 2: Calculate the F-statistic

1. Compute Group Means and Overall Mean

$$\text{Mean for A } (\bar{X}_A) = \frac{20+22+23+21+20}{5} = 21.2$$

$$\text{Mean for B } (\bar{X}_B) = \frac{30+32+31+29+30}{5} = 30.4$$

$$\text{Mean for C } (\bar{X}_C) = \frac{25+27+26+24+26}{5} = 25.6$$

$$\text{Overall Mean } (\bar{X}) = \frac{20+22+23+21+20+30+32+31+29+30+25+27+26+24+26}{15} = 25.73$$

2. Compute the Sum of Squares (SS)

1. **Between-Group Sum of Squares ($SS_{Between}$):**

$$SS_{Between} = n \sum (\bar{X}_i - \bar{X})^2$$

Where $n = 5$ (samples per group).

$$SS_{Between} = 5 \left[(21.2 - 25.73)^2 + (30.4 - 25.73)^2 + (25.6 - 25.73)^2 \right]$$

$$SS_{Between} = 5 \left[20.7361 + 21.9529 + 0.0169 \right] = 214.035$$

2. **Within-Group Sum of Squares (SS_{Within}):**

$$SS_{Within} = \sum \sum (X_{ij} - \bar{X}_i)^2$$

For group A:

$$SS_A = (20 - 21.2)^2 + (22 - 21.2)^2 + (23 - 21.2)^2 + (21 - 21.2)^2 + (20 - 21.2)^2 = 5.2$$

For group B:

$$SS_B = (30 - 30.4)^2 + (32 - 30.4)^2 + (31 - 30.4)^2 + (29 - 30.4)^2 + (30 - 30.4)^2 = 6.8$$

For group C:

$$SS_C = (25 - 25.6)^2 + (27 - 25.6)^2 + (26 - 25.6)^2 + (24 - 25.6)^2 + (26 - 25.6)^2 = 7.2$$

$$SS_{Within} = SS_A + SS_B + SS_C = 5.2 + 6.8 + 7.2 = 19.2$$

3. **Total Sum of Squares (SS_{Total}):**

$$SS_{Total} = SS_{Between} + SS_{Within} = 214.035 + 19.2 = 233.235$$

---

3. Compute Mean Squares (MS)

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}}, \quad df_{Between} = k - 1 = 3 - 1 = 2$$

$$MS_{Between} = \frac{214.035}{2} = 107.0175$$

$$MS_{Within} = \frac{SS_{Within}}{df_{Within}}, \quad df_{Within} = N - k = 15 - 3 = 12$$

$$MS_{Within} = \frac{19.2}{12} = 1.6$$

---

4. Compute F-Statistic

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{107.0175}{1.6} = 66.89$$

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS) (SS/df)	F-ratio (MS_between / MS_within)
Between Groups	$SS_{Between} = 214.035$	$df_{Between} = 2$	$MS_{Between} = 107.0175$	$F = 66.89$
Within Groups	$SS_{Within} = 19.2$	$df_{Within} = 12$	$MS_{Within} = 1.6$	
Total	$SS_{Total} = 233.235$	$df_{Total} = 14$		

Step 3: Determine the Critical Value or p-value

Using an F-distribution table or software:

- $df_{Between} = 2$, $df_{Within} = 12$, and $\alpha = 0.05$.

The critical F-value at $F(2, 12, 0.05) \approx 3.89$.

Since $F = 66.89$ is much larger than 3.89, we reject H_0.

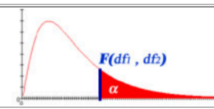
Step 4: Conclusion

There is a statistically significant difference in crop yields among the three fertilizers (p<0.05).

F Distribution Tables

The F distribution is a right-skewed distribution used most commonly in Analysis of Variance. When referencing the F distribution, the **numerator degrees of freedom are always given fi**
freedom changes the distribution (e.g., $F_{(10,12)}$ does not equal $F_{(12,10)}$). For the four F tables below, the rows represent denominator degrees of freedom and the columns represent numerato
given in the name of the table. For example, to determine the .05 critical value for an F distribution with 10 and 12 degrees of freedom, look in the 10 column (numerator) and 12 row (deno
$_{10,\ 12)}$ = 2.7534. You can use the interactive F-Distribution Applet to obtain more accurate measures.

F Table for α = 0.10



\	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61.22034	61.74029	62.00205	62.26497	62.52905	62.79428	63.06064	63.32812
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9.42471	9.44131	9.44962	9.45793	9.46624	9.47456	9.48289	9.49122
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5.20031	5.18448	5.17636	5.16811	5.15972	5.15119	5.14251	5.13370
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3.87036	3.84434	3.83099	3.81742	3.80361	3.78957	3.77527	3.76073
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3.23801	3.20665	3.19052	3.17408	3.15732	3.14023	3.12279	3.10500
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2.87122	2.83634	2.81834	2.79996	2.78117	2.76195	2.74229	2.72216
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2.63223	2.59473	2.57533	2.55546	2.53510	2.51422	2.49279	2.47079
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196	2.46422	2.42464	2.40410	2.38302	2.36136	2.33910	2.31618	2.29257
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888	2.33962	2.29832	2.27683	2.25472	2.23196	2.20849	2.18427	2.15923