# Stats Review for Data Science

Notes Prepared by: Uras Demir, Ph.D.

[Work in Progress]

February 10, 2025

# Contents

# 1 Types of Data

Table 1: Types of Data

| **Data Type** | **Explanation** | **Statistics to Compute** | **Examples** |
|---|---|---|---|
| **Binary** | A subtype of categorical data with only two possible values (0 or 1, True/False). | Proportions, logistic regression, chi-square tests. | Pass/Fail, Yes/No, Male/Female. |
| **Categorical (Nominal)** | Data divided into distinct groups or categories without intrinsic order. | Mode, frequencies, chi-square tests. | Eye color, types of fruits, city names. |
| **Ordinal** | Categorical data with a meaningful order, but differences between levels are not measurable. | Median, percentiles, Spearman's rank correlation. | Satisfaction levels (Low/Medium/High), education levels. |
| **Discrete** | Countable quantities, often integers, with a finite or countably infinite range of values. | Counts, frequencies, mode, Poisson regression. | Number of students, cars in a parking lot. |
| **Continuous (Interval)** | Measurable quantities with no true zero. Differences are meaningful, but ratios are not. | Mean, standard deviation, correlation. | Temperature (°C), calendar years. |
| **Continuous (Ratio)** | Measurable quantities with a true zero. Both differences and ratios are meaningful. | Mean, standard deviation, geometric mean. | Weight (kg), height (cm), income ($). |

## 2 Statistical Measures of Location

Table 2: Statistical Measures of Location

| **Measure** | **Explanation** | **Formula** | **Example (Dataset: {1, 1, 3, 5, 5, 15})** | **When to Use** |
|---|---|---|---|---|
| **Mode** | The most frequently occurring value(s) in the dataset. | N/A (find the value with the highest frequency). | Mode = 1, 5 (both appear twice). | Use when identifying the most common category or value is important, such as for nominal data (e.g., survey responses). |
| **Median** | The middle value when the dataset is ordered. If even, it is the average of the two middle values. | Median = $\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$, for $n$ even. | Ordered: {1, 1, 3, 5, 5, 15}. Median = (3 + 5) / 2 = 4. | Use when the dataset has outliers or a skewed distribution, as the median is robust to extreme values. |
| **Weighted Median** | The value where 50% of the cumulative weight lies. | Sort values by weight; find the point where cumulative weight = 50%. | Weights: {1, 1, 1, 1, 1, 1}. Weighted median = 4. | Use when data points have varying importance or weights, and you want a robust central measure. |
| **Mean** | The average of all values in the dataset. | Mean = $\frac{\sum x_i}{n}$. | (1 + 1 + 3 + 5 + 5 + 15) / 6 = 30 / 6 = 5. | Use for normally distributed data where all values are equally important, as the mean is sensitive to outliers. |
| **Weighted Mean** | The average where each value has a weight. | Weighted Mean = $\frac{\sum(x_i \cdot w_i)}{\sum w_i}$. | Values: {1, 1, 3, 5, 5, 15}; Weights: {2, 1, 1, 3, 2, 1}. Weighted Mean = $\frac{(1\cdot2)+(1\cdot1)+(3\cdot1)+(5\cdot3)+(5\cdot2)+(15\cdot1)}{2+1+1+3+2+1}$ $\frac{2+1+3+15+10+15}{10} = 4.6$. | Use when some values contribute more significantly to the average than others, such as in weighted surveys or averages. |
| **Geometric Mean** | The nth root of the product of all values, where $n$ is the total count. | Geometric Mean = $\sqrt[n]{\prod x_i}$. | $\sqrt[6]{1 \cdot 1 \cdot 3 \cdot 5 \cdot 5 \cdot 15} \approx 3.56$. | Use for data involving rates, ratios, or percentages, such as growth rates in finance or population studies. |
| **Trimmed Mean** | The mean after removing a percentage of the smallest and largest values. | Remove $p$% of smallest and largest values; compute mean of remaining. | Remove 1 value from each end (10%). New dataset: {1, 3, 5, 5}. Trimmed Mean = (1 + 3 + 5 + 5) / 4 = 3.5. | Use when mitigating the influence of outliers is important, but you still want to use an average. |

## 3 Descriptive Measures of Variability

Table 3: Descriptive Measures of Variability with Examples

| **Measure** | **Explanation** | **Formula and Example** | **When to Use** |
|---|---|---|---|
| **Variance (MSE)** | Measures variability by averaging squared deviations from the mean. | $\text{Var} = \frac{\sum(x_i - \bar{x})^2}{n-1}.$<br>Dataset: $\{1, 1, 3, 5, 5, 15\}, \bar{x} = 5.$<br>$\text{Var} = \frac{16+16+4+0+0+100}{5} = 22.67.$ | Quantifies variability in normally distributed data. Sensitive to outliers. |
| **Standard Deviation (SD)** | Represents the average distance of values from the mean. | $\text{SD} = \sqrt{\text{Variance}}.$<br>Dataset: Var $= 22.67.$<br>$\text{SD} = \sqrt{22.67} \approx 4.76.$ | Use for data spread in the same unit as the data. Commonly used in normal distributions. |
| **Mean Absolute Deviation (MAD)** | Represents the average absolute differences from the mean. | $\text{MAD} = \frac{\sum \lvert x_i - \bar{x} \rvert}{n}.$<br>Dataset: $\{1, 1, 3, 5, 5, 15\}, \bar{x} = 5.$<br>$\text{MAD} = \frac{4+4+2+0+0+10}{6} = 3.33.$ | Less sensitive to outliers than variance or SD. |
| **Covariance** | Measures the joint variability between two variables. | $\text{Cov}(X, Y) = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n}.$<br>Dataset: $X = \{1, 2, 3, 4, 5\}, Y = \{10, 20, 30, 20, 40\}.$<br>$\text{Cov}(X, Y) = \frac{60}{5} = 12.$ | Used to assess the direction of a relationship. Not standardized. |
| **Range** | The difference between the largest and smallest values. | $\text{Range} = \text{Max}(x) - \text{Min}(x).$<br>Dataset: $\{1, 1, 3, 5, 5, 15\}.$<br>$\text{Range} = 15 - 1 = 14.$ | Quick measure of variability. Highly sensitive to outliers. |
| **Ranks** | The position of each value in the ordered dataset. | Ordered dataset: $\{1, 1, 3, 5, 5, 15\}.$<br>Ranks: $\{1, 2, 3, 4, 5, 6\}.$ | Used for ordinal data or non-parametric statistics like Spearman's correlation. |

Table 3: (continued)

| **Measure** | **Explanation** | **Formula and Example** | **When to Use** |
|---|---|---|---|
| **Percentiles** | The value below which a given percentage of observations fall. | $$P_k = \text{Value at } \frac{k}{100}(n+1).$$ Dataset: $\{1, 1, 3, 5, 5, 15\}$.<br>$P_{25} = 2.75$, value between ranks 2 and 3. | Used for dividing data into parts for analysis (e.g., boxplots). |
| **Interquartile Range (IQR)** | The range between the first quartile ($Q1$) and the third quartile ($Q3$). | $$\text{IQR} = Q3 - Q1.$$ Dataset: $Q1 = 3, Q3 = 5$.<br>$\text{IQR} = 5 - 3 = 2$. | Robust to outliers; useful for skewed datasets. |

# 4 Visualizations of Distribution

Table 4: When to Use Common Data Visualizations and Tools

| **Tool** | **When to Use** | **What It Shows** | **Example** |
|---|---|---|---|
| **Contingency Table** | For categorical data to analyze relationships. | A table of frequencies or proportions for combinations of two categorical variables. | Analyzing the relationship between *gender* (Male/Female) and *purchase decision* (Yes/No). |
| **Frequency Table** | For counting categorical or discrete values. | A table of counts or proportions for each value or category. | Counting the number of students in each *grade level* (Freshman, Sophomore, etc.). |
| **Histogram** | For visualizing numerical data distributions. | Bar plot showing how data is distributed across intervals (bins). | Showing the distribution of *heights* in a population. |
| **Boxplot** | For continuous data to summarize distributions. | Shows median, quartiles, IQR, and outliers for one or more groups. | Comparing test scores for students across different *classes.* |
| **Density Plot** | For smoothed distributions of continuous data. | A curve estimating the probability density function (PDF). | Comparing the distribution of *income levels* for two regions. |
| **Violin Plot** | For continuous data to show both the distribution and summary statistics. | Combines a boxplot (median, quartiles) with a kernel density plot to show data distribution symmetrically. | Comparing the distribution of *reaction times* across different *experimental groups.* |

# 5 Correlation and Correlation Tests

## 5.1 Pearson Correlation

**Assumptions:**

- Both variables are continuous.
- The relationship is linear.
- Variables are normally distributed (or approximately so).
- No significant outliers.

**Formula:**

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- $\text{Cov}(X, Y)$ is the covariance between $X$ and $Y$,
- $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$.

**Example Calculation:**

- Dataset: $\{1, 2, 3, 4, 5\}$, $\{10, 20, 30, 20, 40\}$.
- Mean of $X$: $\bar{X} = 3$, Mean of $Y$: $\bar{Y} = 24$.
- Covariance: $\text{Cov}(X, Y) = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n} = 10$.
- Standard Deviations: $\sigma_X = \sqrt{2}, \sigma_Y = \sqrt{160}$.
- Pearson Correlation: $\rho = \frac{10}{\sqrt{2} \cdot \sqrt{160}} = 0.25$.

**Interpretation:** Weak positive linear correlation ($\rho = 0.25$).

## 5.2 Spearman's Rank Correlation

**Assumptions:**

- Data can be ranked (ordinal or continuous).
- The relationship is monotonic (increasing or decreasing consistently).

**Formula:**

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:

- $d_i$ is the difference between the ranks of $X_i$ and $Y_i$.

**Example Calculation:**

- Dataset: $\{1, 2, 3, 4, 5\}$, $\{10, 20, 30, 20, 40\}$.
- Ranks: $X$ ranks: $\{1, 2, 3, 4, 5\}$, $Y$ ranks: $\{1, 2.5, 4, 2.5, 5\}$.
- Differences ($d_i$): $\{0, -0.5, -1, 1.5, 0\}$, ($d_i^2 = \{0, 0.25, 1, 2.25, 0\}$).
- Spearman Correlation: $\rho_s = 1 - \frac{6 \cdot 3.5}{5(5^2 - 1)} = 0.9$.

**Interpretation:** Strong positive monotonic relationship ($\rho_s = 0.9$).

## 5.3 Kendall's Tau

**Assumptions:**

- Works best for small datasets.
- Data can be ranked.
- The relationship is monotonic.

**Formula:**

$$\tau = \frac{(\text{Number of Concordant Pairs}) - (\text{Number of Discordant Pairs})}{\binom{n}{2}}$$

Where:

- Concordant pairs: If $X_i > X_j$ and $Y_i > Y_j$, or $X_i < X_j$ and $Y_i < Y_j$,
- Discordant pairs: If $X_i > X_j$ and $Y_i < Y_j$, or $X_i < X_j$ and $Y_i > Y_j$,
- $\binom{n}{2} = \frac{n(n-1)}{2}$: Total number of unique pairs.

**Example:**

Dataset: $X = \{1, 2, 3, 4, 5\}$, $Y = \{10, 20, 30, 20, 40\}$.

**Concordant Pairs:**

- Pair 1: $(X_1, Y_1) = (1, 10)$ and $(X_2, Y_2) = (2, 20) \Rightarrow 1 < 2$ and $10 < 20$.
- Pair 2: $(X_1, Y_1) = (1, 10)$ and $(X_3, Y_3) = (3, 30) \Rightarrow 1 < 3$ and $10 < 30$.
- Pair 3: $(X_1, Y_1) = (1, 10)$ and $(X_4, Y_4) = (4, 20) \Rightarrow 1 < 4$ and $10 < 20$.
- Pair 4: $(X_1, Y_1) = (1, 10)$ and $(X_5, Y_5) = (5, 40) \Rightarrow 1 < 5$ and $10 < 40$.
- Pair 5: $(X_2, Y_2) = (2, 20)$ and $(X_3, Y_3) = (3, 30) \Rightarrow 2 < 3$ and $20 < 30$.
- Pair 7: $(X_2, Y_2) = (2, 20)$ and $(X_5, Y_5) = (5, 40) \Rightarrow 2 < 5$ and $20 < 40$.
- Pair 9: $(X_3, Y_3) = (3, 30)$ and $(X_5, Y_5) = (5, 40) \Rightarrow 3 < 5$ and $30 < 40$.
- Pair 10: $(X_4, Y_4) = (4, 20)$ and $(X_5, Y_5) = (5, 40) \Rightarrow 4 < 5$ and $20 < 40$.

**Discordant Pairs:**

- Pair 6: $(X_2, Y_2) = (2, 20)$ and $(X_4, Y_4) = (4, 20) \Rightarrow 2 < 4$ but $20 = 20$ (no increase/decrease in $Y$).
- Pair 8: $(X_3, Y_3) = (3, 30)$ and $(X_4, Y_4) = (4, 20) \Rightarrow 3 < 4$ but $30 > 20$.

**Summary:**

- Total Concordant Pairs: 8.
- Total Discordant Pairs: 2.
- Kendall's Tau:

$$\tau = \frac{\text{Number of Concordant Pairs} - \text{Number of Discordant Pairs}}{\binom{n}{2}}$$

$$\tau = \frac{8 - 2}{10} = 0.6.$$

**Interpretation:** Moderate positive ordinal relationship ($\tau = 0.6$).

## 5.4 Point-Biserial Correlation

**Assumptions:**

- One variable is continuous, and the other is binary (e.g., 0 or 1).

**Formula:**

$$r = \frac{\bar{X}_1 - \bar{X}_0}{s} \cdot \sqrt{\frac{n_1 n_0}{n^2}}$$

Where:

- $\bar{X}_1$: Mean of the continuous variable for group 1 (binary = 1),
- $\bar{X}_0$: Mean of the continuous variable for group 0 (binary = 0),
- $s$: Standard deviation of the continuous variable,
- $n_1, n_0$: Number of observations in group 1 and group 0,
- $n$: Total number of observations ($n = n_1 + n_0$).

**Example:**

- Binary Variable ($Y$): $\{0, 1, 1, 0, 1\}$
- Continuous Variable ($X$): $\{10, 20, 30, 20, 40\}$

**Step 1: Group Statistics**

- Group 1 ($Y = 1$): $X = \{20, 30, 40\}$

$$\bar{X}_1 = \frac{20 + 30 + 40}{3} = 30, \quad n_1 = 3$$

- Group 0 ($Y = 0$): $X = \{10, 20\}$

$$\bar{X}_0 = \frac{10 + 20}{2} = 15, \quad n_0 = 2$$

**Step 2: Compute Standard Deviation of $X$**

$$\bar{X} = \frac{10 + 20 + 30 + 20 + 40}{5} = 24$$

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

$$s = \sqrt{\frac{196 + 16 + 36 + 16 + 256}{5}} = \sqrt{104} \approx 10.2$$

**Step 3: Compute Point-Biserial Correlation**

$$r = \frac{\bar{X}_1 - \bar{X}_0}{s} \cdot \sqrt{\frac{n_1 n_0}{n^2}}$$

$$r = \frac{30 - 15}{10.2} \cdot \sqrt{\frac{3 \cdot 2}{5^2}} = \frac{15}{10.2} \cdot \sqrt{\frac{6}{25}} = 1.47 \cdot 0.49 = 0.72$$

**Interpretation:** There is a strong positive relationship ($r = 0.72$) between the binary variable ($Y$) and the continuous variable ($X$).

## 5.5 Partial Correlation

**When to Use:** Partial correlation is used to measure the strength and direction of the linear relationship between two variables ($X$ and $Y$) while controlling for the influence of a third variable ($Z$). It helps isolate the unique relationship between $X$ and $Y$ by removing the effect of $Z$.

**Assumptions:**

- All variables are continuous.
- Relationships between variables are linear.
- Variables are measured without error.

**Formula:** The partial correlation between $X$ and $Y$ while controlling for $Z$ is:

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ} \cdot \rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}$$

Where:

- $\rho_{XY}$: Correlation between $X$ and $Y$,
- $\rho_{XZ}$: Correlation between $X$ and $Z$,
- $\rho_{YZ}$: Correlation between $Y$ and $Z$.

**Example:** Given the datasets:

$$X = \{1, 2, 3, 4, 5\}, \quad Y = \{10, 20, 30, 20, 40\}, \quad Z = \{5, 15, 25, 20, 35\}$$

**Step 1: Compute Pairwise Correlations** Using a statistical tool or Pearson's correlation formula, we calculate:

- $\rho_{XY} = 0.25$,
- $\rho_{XZ} = 0.5$,
- $\rho_{YZ} = 0.4$.

(Note: These values are chosen to ensure a valid correlation matrix with a non-negative determinant.)

**Step 2: Compute Partial Correlation** Substitute the values into the formula:

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ} \cdot \rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}$$

$$\rho_{XY \cdot Z} = \frac{0.25 - (0.5 \cdot 0.4)}{\sqrt{(1 - 0.5^2)(1 - 0.4^2)}}$$

Simplify step by step:

$$\rho_{XY \cdot Z} = \frac{0.25 - 0.2}{\sqrt{(1 - 0.25)(1 - 0.16)}} = \frac{0.05}{\sqrt{0.75 \cdot 0.84}}$$

$$\rho_{XY \cdot Z} = \frac{0.05}{\sqrt{0.63}} = \frac{0.05}{0.7937} \approx 0.063$$

**Step 3: Interpret the Partial Correlation** The partial correlation coefficient is approximately 0.063. This value suggests a weak positive relationship between $X$ and $Y$ after accounting for the influence of $Z$. The low value indicates that much of the relationship between $X$ and $Y$ is explained by their shared association with $Z$.

# 6 Sampling Distributions

Sampling distributions are the probability distributions of sample statistics (e.g., mean, variance) derived from repeated sampling of a population. This section explains key concepts like sample statistics, sampling distributions, bias, standard error, bootstrap, and confidence intervals.

## 6.1 Sample Statistic

**Explanation:** A sample statistic is a numerical measure computed from a sample, such as the sample mean ($\bar{x}$) or sample variance ($s^2$).

**Formula:** For a sample of size $n$,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

## 6.2 Sampling Distribution

**Explanation:** The sampling distribution is the probability distribution of a sample statistic when drawn from the population multiple times. For large $n$, the sampling distribution of the sample mean approximates a normal distribution (Central Limit Theorem).

**Formula:** The mean and variance of the sampling distribution of the sample mean are:

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n},$$

where $\mu$ and $\sigma^2$ are the population mean and variance.

**Example:** Given a population with $\mu = 10$, $\sigma^2 = 4$, and $n = 16$:

$$\mu_{\bar{x}} = 10, \quad \sigma_{\bar{x}}^2 = \frac{4}{16} = 0.25, \quad \sigma_{\bar{x}} = \sqrt{0.25} = 0.5$$

## 6.3 Bias (Reliability vs. Validity)

**Explanation:** - Bias refers to the systematic error in sample statistics, leading to incorrect estimates of population parameters. - Reliability refers to the consistency of measurements. - Validity refers to the accuracy of measurements.

**Example:** - A biased estimator consistently overestimates or underestimates the true population parameter. - A reliable but invalid estimator might give consistent but inaccurate results.

## 6.4 Standard Error of the Sample

**Explanation:** The standard error (SE) quantifies the variability of a sample statistic, such as the sample mean, across repeated random samples drawn from the population. It indicates how much the sample mean is expected to fluctuate around the population mean if multiple samples are taken. A smaller SE implies more precise estimates of the population parameter.

**Formula:** For the standard error of the sample mean:

$$\text{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

where:

- $\sigma$: The population standard deviation,
- $n$: The sample size.

**Example:** Suppose the population standard deviation is $\sigma = 10$ and the sample size is $n = 25$:

$$\text{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

**Interpretation:** - The sample mean is expected to vary by approximately 2 units from the population mean across repeated samples. - If $n$ increases (e.g., $n = 100$), $\text{SE}_{\bar{x}}$ decreases:

$$\text{SE}_{\bar{x}} = \frac{10}{\sqrt{100}} = 1$$

This demonstrates that larger sample sizes result in more precise estimates (smaller standard error).

**Connection to Sampling Distribution:** - The standard error is the standard deviation of the sampling distribution of the sample mean. - It reflects the spread of the sample means around the true population mean in repeated sampling.

## 6.5 Bootstrap

**Explanation:** Bootstrap is a resampling method to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the original sample. It is particularly useful when the theoretical distribution of a statistic is unknown or when the sample size is small.

**Steps:**

1. Start with an original sample of size $n$, e.g., $\{2, 4, 6, 8, 10\}$.
2. Generate multiple resamples of size $n$ by sampling with replacement, e.g., $\{4, 6, 6, 8, 10\}$, $\{2, 2, 6, 8, 8\}$, etc.
3. Compute the desired statistic (e.g., mean) for each resample.
4. Use the distribution of the resampled statistics to estimate variability (e.g., standard error) or construct confidence intervals.

**Example:** Original sample: $\{2, 4, 6, 8, 10\}$. Bootstrap means:

$$\text{Resample 1: } \{4, 6, 6, 8, 10\} \to \bar{x} = 6.8, \quad \text{Resample 2: } \{2, 2, 6, 8, 8\} \to \bar{x} = 5.2$$

Repeat this process $B = 1000$ times to build the bootstrap distribution of the mean.

—

## 6.6 Jackknifing

**Explanation:** Jackknifing is a resampling method that estimates a statistic's bias or standard error by systematically leaving out one observation at a time from the sample and recomputing the statistic. Unlike bootstrapping, it does not involve random sampling.

**Steps:**

1. Start with an original sample of size $n$, e.g., $\{2, 4, 6, 8, 10\}$.
2. Create $n$ subsets, each leaving out one observation:

$$\text{Subset 1: } \{4, 6, 8, 10\}, \quad \text{Subset 2: } \{2, 6, 8, 10\}, \ldots$$

3. Compute the statistic (e.g., mean) for each subset.
4. Aggregate the results to estimate bias or variability.

**Example:** Original sample: $\{2, 4, 6, 8, 10\}$, with $n = 5$.

$$\text{Subset 1: } \{4, 6, 8, 10\} \to \bar{x}_1 = \frac{4 + 6 + 8 + 10}{4} = 7, \quad \text{Subset 2: } \{2, 6, 8, 10\} \to \bar{x}_2 = \frac{2 + 6 + 8 + 10}{4} = 6.5$$

Repeat for all subsets, then compute:

$$\text{Jackknife Mean: } \bar{x}_{\text{jackknife}} = \frac{\sum_{i=1}^{n} \bar{x}_i}{n} = \frac{7 + 6.5 + \ldots}{5}$$

## 6.7 Confidence Interval (CI)

**Explanation:** A confidence interval provides a range of values within which the population parameter is likely to lie, based on the sample statistic.

**Formula:**

$$\mathrm{CI} = \bar{x} \pm z^* \cdot \mathrm{SE}_{\bar{x}}$$

where $z^*$ is the critical value for a given confidence level, and $\mathrm{SE}_{\bar{x}}$ is the standard error of the sample mean.

**How to Look Up $z^*$:**

1. Choose the Confidence Level $(1 - \alpha)$: For example, a 95% confidence level implies $\alpha = 0.05$.
2. Find the Area in the Standard Normal Table: - Divide $\alpha$ by 2 for the tails: $\alpha/2 = 0.025$ (for 95% CI). - The cumulative probability for the upper tail is $1 - 0.025 = 0.975$.
3. Look Up in the Z-Table: Find the row and column in the Z-table corresponding to 0.975. The value is $z^* = 1.96$.

**Example:** Given $\bar{x} = 6$, $\mathrm{SE}_{\bar{x}} = 2$, and a 95% confidence level ($z^* = 1.96$):

$$\mathrm{CI} = \bar{x} \pm z^* \cdot \mathrm{SE}_{\bar{x}}$$

$$\mathrm{CI} = 6 \pm 1.96 \cdot 2 = 6 \pm 3.92 = [2.08, 9.92]$$

**Interpretation:** There is a 95% probability that the population mean lies within the interval $[2.08, 9.92]$.

**Common $z^*$ Values for Confidence Levels:**

- 90% CI: $z^* = 1.645$
- 95% CI: $z^* = 1.96$
- 99% CI: $z^* = 2.576$

# 7 Normal Distribution

The **Normal Distribution** (or Gaussian distribution) is a continuous probability distribution characterized by the probability density function (PDF), which yields a bell-shaped distribution.

## 7.1 Standard Normal Distribution

A **standard normal distribution** is a special case where the mean and standard deviation are:

$$\mu = 0, \quad \sigma = 1 \tag{1}$$

This standardization allows for easy comparison of different datasets.

## 7.2 Probability Density Function

The **Probability Density Function (PDF)** describes the relative likelihood of a continuous random variable taking on a given value. For a normal distribution, the PDF is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

where:

- $x$ is the value of the random variable,
- $\mu$ is the **mean** (center of the distribution),
- $\sigma$ is the **standard deviation** (spread of the distribution),
- $e$ is Euler's number ($\approx 2.718$),
- $\pi$ is the mathematical constant ($\approx 3.1416$).

**Key Properties of the PDF**

1. The function is always **non-negative**:
2. The total probability sums to 1:
3. The function gives the **relative likelihood** of different values of $X$ but not exact probabilities.

## 7.3 Z-Score (Standardization)

The **Z-score** transforms a normal variable $X$ into a standard normal variable $Z$:

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

- If $Z > 0$: the value is above the mean.
- If $Z < 0$: the value is below the mean.
- If $Z \approx 0$: the value is near the mean.

Z-scores are commonly used in hypothesis testing and normalization.

## 7.4 Q-Q Plot for Normality Check

A **quantile-quantile (Q-Q) plot** is a graphical tool to assess if a dataset follows a normal distribution. It compares the quantiles of the sample data against the quantiles of a theoretical normal distribution.

If the data is normally distributed, the points should align along a diagonal line.

# 8 Student's t-distribution

The **Student's t-distribution** is a continuous probability distribution used in statistical inference, particularly when estimating the mean of a normally distributed population with small sample sizes.

## 8.1 Difference from the Normal Distribution

The t-distribution is similar to the normal distribution but has heavier tails, meaning it gives more probability to extreme values. This accounts for additional uncertainty due to small sample sizes.

- When the sample size ($n$) is large, the t-distribution approximates the standard normal distribution.
- When $n$ is small, the t-distribution has wider tails to reflect increased variability.

## 8.2 Degrees of Freedom (df)

The **degrees of freedom** ($\nu$) represent the number of independent pieces of information available for estimating variability. It is typically calculated as:

$$\nu = n - 1 \tag{4}$$

where $n$ is the sample size.

- A smaller $\nu$ leads to a distribution with heavier tails.
- As $\nu \to \infty$, the t-distribution converges to the standard normal distribution.

## 8.3 t-Score

The t-score is used in hypothesis testing and confidence interval estimation. It is given by:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{5}$$

where:

- $\bar{X}$ is the sample mean,
- $\mu$ is the population mean (hypothesized value),
- $s$ is the sample standard deviation,
- $n$ is the sample size.

The t-score follows a t-distribution with $\nu = n - 1$ degrees of freedom.

## 8.4 Confidence Interval (CI) Calculation

The t-distribution is used to calculate confidence intervals when the population standard deviation ($\sigma$) is unknown and must be estimated from the sample.

A two-sided confidence interval for the population mean is given by:

$$\bar{X} \pm t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} \tag{6}$$

where:

- $t_{\alpha/2,\nu}$ is the critical t-value from the t-table at significance level $\alpha$,
- $s/\sqrt{n}$ is the standard error of the mean.

# 9 Binomial Distribution

## 9.1 Definition and Properties

The **Binomial Distribution** models the number of successes in a fixed number of independent trials, where each trial has two possible outcomes: **success** or **failure**. It is widely used in probability and statistics for discrete random variables.

The probability mass function (PMF) of a binomially distributed random variable $X$ is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{7}$$

where:

- $n$ is the number of trials,
- $k$ is the number of successes,
- $p$ is the probability of success in a single trial,
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

## 9.2 Binomial Trials and Success Probability

A **binomial trial** (or Bernoulli trial) is a single experiment where:

- There are only two possible outcomes: success (1) or failure (0).
- The probability of success, $p$, remains constant across trials.
- The trials are **independent**, meaning the outcome of one does not affect another.

If we repeat a binomial trial $n$ times, the number of successes follows a **binomial distribution**.

## 9.3 Mean and Variance

For a binomially distributed random variable $X \sim \mathrm{Bin}(n, p)$, the expected value (mean) and variance are given by:

$$E(X) = np \tag{8}$$

$$\mathrm{Var}(X) = np(1-p) \tag{9}$$

These properties show that as $n$ increases, the distribution's spread depends on $p$.

## 9.4 Normal Approximation to the Binomial

For large $n$, the binomial distribution can be approximated by a normal distribution using the Central Limit Theorem. If $n$ is large and $p$ is not too close to 0 or 1, then:

$$X \approx \mathcal{N}(np, np(1-p)) \tag{10}$$

This normal approximation is useful when working with large datasets, as calculating binomial probabilities directly can be computationally intensive.

# 10 Poisson Distribution

## 10.1 Definition and Properties

The **Poisson distribution** is a discrete probability distribution that models the number of events occurring within a fixed interval of time or space, assuming that the events occur independently and at a constant average rate.

The probability mass function (PMF) of a Poisson-distributed random variable $X$ is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots \tag{11}$$

where:

- $k$ is the number of occurrences,
- $\lambda$ is the expected number of occurrences in the given interval (mean rate),
- $e$ is Euler's number ($\approx 2.718$).

## 10.2 Poisson Process and Rate Parameter

A **Poisson process** describes a sequence of events occurring randomly over time or space, characterized by:

- Events occurring **independently** of each other.
- A **constant rate** $\lambda$, meaning the probability of an event occurring is proportional to the interval size.
- At most **one event per infinitesimally small interval**.

The parameter $\lambda$ is both the mean and variance of the Poisson distribution:

$$E(X) = \lambda, \quad \mathrm{Var}(X) = \lambda. \tag{12}$$

# 11 Hypothesis Testing

## 11.1 Introduction

**Hypothesis testing** is a statistical method used to evaluate whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis. It is widely applied in scientific research, business analytics, and A/B testing to assess whether an observed effect is statistically significant or occurred by chance.

## 11.2 Key Concepts in Hypothesis Testing

- **Null Hypothesis ($H_0$)**: The assumption that there is no effect, no difference, or no relationship between groups or variables.
- **Alternative Hypothesis ($H_1$)**: The hypothesis suggesting that an effect, difference, or relationship exists.
- **Significance Level ($\alpha$)**: The probability of rejecting $H_0$ when it is actually true. A common threshold is 0.05, meaning a 5% risk of a false positive.
- **P-Value**: The probability of obtaining results as extreme as (or more extreme than) the observed data, assuming $H_0$ is true. A p-value below $\alpha$ typically leads to rejecting $H_0$.
- **Type I Error**: Occurs when $H_0$ is wrongly rejected (false positive). The probability of this occurring is $\alpha$.
- **Type II Error**: Occurs when $H_0$ is not rejected even though $H_1$ is true (false negative). The probability of this occurring is $\beta$.
- **Statistical Power ($1 - \beta$)**: The probability of correctly rejecting $H_0$ when $H_1$ is true. A power of at least 0.8 (80%) is often desired.
- **Effect Size**: Measures the magnitude of the difference between groups. A small effect size requires a larger sample size to detect reliably.
- **Sample Size**: The number of observations included in the test. Larger samples reduce variability and increase statistical power.

## 11.3 Understanding Type I and Type II Errors

When conducting hypothesis tests, two types of errors may occur:

- **Type I Error ($\alpha$)**: Incorrectly rejecting a true $H_0$ (false positive). Lowering $\alpha$ (e.g., from 0.05 to 0.01) reduces this risk but increases the chance of a Type II error.
- **Type II Error ($\beta$)**: Failing to reject $H_0$ when $H_1$ is actually true (false negative). This often happens when the sample size is too small or the effect size is weak.

Statistical power is the complement of $\beta$:

$$\text{Power} = 1 - \beta. \tag{13}$$

Higher power increases the ability to detect true effects and is improved by larger sample sizes or stronger effect sizes.

## 11.4 One-Tailed vs. Two-Tailed Tests

- **One-Tailed Test**: Tests for a difference in a specific direction (e.g., whether Method A is better than Method B).

$$H_1 : \mu_A > \mu_B \quad \text{or} \quad H_1 : \mu_A < \mu_B.$$

- **Two-Tailed Test**: Tests for any significant difference in either direction (e.g., whether Method A and Method B perform differently).

$$H_1 : \mu_A \neq \mu_B.$$

A one-tailed test has greater power to detect an effect in the expected direction, while a two-tailed test is more conservative and considers deviations in both directions.

## 11.5 One-Way vs. Two-Way Tests

- **One-Way Test**: Compares the means of two independent groups. Example: Testing whether a new drug improves recovery rates compared to a placebo.
- **Two-Way Test**: Examines the effects of two independent variables and their interaction. Example: Testing whether both drug type and dosage affect recovery rates.

## 11.6 Steps in Hypothesis Testing

1. **Define the Hypotheses:** Formulate $H_0$ and $H_1$.
2. **Choose the Significance Level ($\alpha$)**: Typically 0.05 or 0.01.
3. **Select the Statistical Test:** Depends on data type and assumptions (e.g., Z-test, t-test, chi-square test).
4. **Compute the Test Statistic**: A measure of deviation from $H_0$.
5. **Compare to Critical Value or Compute P-Value**: Determine whether the result is statistically significant.
6. **Make a Decision**: Reject $H_0$ if the p-value is below $\alpha$, otherwise fail to reject $H_0$.

## 11.7 Determining Statistical Power and Sample Size

**Power analysis** is used to calculate the required sample size for a given effect size and significance level. The key factors influencing power are:

- **Effect Size**: Larger effects are easier to detect.
- **Sample Size**: A larger sample reduces variability and increases power.
- **Significance Level ($\alpha$)**: A stricter $\alpha$ (e.g., 0.01) reduces false positives but requires a larger sample to maintain power.
- **Variability**: Higher variance increases the required sample size.

The formula for estimating the minimum sample size needed for a test with power $1 - \beta$ is:

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\Delta^2}. \tag{14}$$

where:

- $Z_\alpha$ is the critical value for the chosen significance level.
- $Z_\beta$ corresponds to the desired power (e.g., 80% power corresponds to $Z_\beta = 0.84$).
- $\sigma$ is the standard deviation.
- $\Delta$ is the expected difference between groups.

# 12 A/B Testing using Two-Proportion Z Test

## 12.1 Introduction

A/B testing is a statistical method used to compare two versions of a treatment (e.g., a webpage, advertisement, or drug) to determine which performs better. It is widely used in marketing, product optimization, and clinical trials.

In an A/B test:

- **Group A** (control group) receives the original version.
- **Group B** (treatment group) receives the modified version.
- The outcome of interest (e.g., click-through rate, conversion rate) is compared between the two groups.

## 12.2 Hypothesis Setup

A/B testing is typically framed as a hypothesis test:

- **Null Hypothesis** ($H_0$): There is no difference between A and B.
- **Alternative Hypothesis** ($H_1$): There is a significant difference between A and B.

Mathematically, let $p_A$ and $p_B$ be the success probabilities for groups A and B:

$$H_0 : p_A = p_B. \tag{15}$$

The choice of alternative hypothesis depends on whether a two-tailed test or a one-tailed test is appropriate.

- **Two-Tailed Test:**
  $$H_1 : p_A \neq p_B$$
  Used when testing for any significant difference in either direction.
- **One-Tailed Test (Greater):**
  $$H_1 : p_A < p_B$$
  Used when testing if B performs significantly better than A.
- **One-Tailed Test (Lesser):**
  $$H_1 : p_A > p_B$$
  Used when testing if B performs significantly worse than A.

## 12.3 Example: Testing Click-Through Rate (CTR)

Suppose a company wants to test whether changing a button color on a webpage increases the click-through rate (CTR).

- Version A (Control): Blue button.
- Version B (Treatment): Red button.

The company randomly assigns 10,000 visitors:

- 5,000 see the blue button ($n_A = 5000$), with 250 clicks ($X_A = 250$).
- 5,000 see the red button ($n_B = 5000$), with 300 clicks ($X_B = 300$).

The observed click-through rates (CTR) are:

$$\hat{p}_A = \frac{250}{5000} = 0.05, \quad \hat{p}_B = \frac{300}{5000} = 0.06. \tag{16}$$

## 12.4 Statistical Test: Two-Proportion Z-Test

Since we are comparing two proportions, we use a two-proportion Z-test:

$$Z = \frac{(\hat{p}_A - \hat{p}_B)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{17}$$

where:

- $\hat{p} = \frac{X_A + X_B}{n_A + n_B}$ is the pooled proportion.
- $n_A, n_B$ are sample sizes.

For our example:

$$\hat{p} = \frac{250 + 300}{5000 + 5000} = 0.055. \tag{18}$$

Computing the Z-score:

$$Z = \frac{(0.05 - 0.06)}{\sqrt{0.055(1 - 0.055)\left(\frac{1}{5000} + \frac{1}{5000}\right)}}. \tag{19}$$

## 12.5 Interpreting Results: One-Tailed vs. Two-Tailed Tests

Once the $Z$-value is computed, we compare it to critical values from the standard normal distribution.

- Two-Tailed Test: If $|Z|$ is greater than the critical value at $\alpha/2$ (e.g., $\pm1.96$ for $\alpha = 0.05$), we reject $H_0$ and conclude a significant difference.
- One-Tailed Test: If $Z$ is greater than the critical value for $\alpha$ (e.g., 1.645 for $\alpha = 0.05$), we reject $H_0$ in favor of the alternative hypothesis.

## 12.6 Choosing Between One-Tailed and Two-Tailed Tests

- Use a one-tailed test when you have a clear directional hypothesis.
- Use a two-tailed test when any difference, in either direction, is important.

## 12.7 Practical Considerations

- **Sample Size:** A/B tests require sufficient sample sizes to detect meaningful effects.
- **Multiple Testing:** Running many A/B tests increases the risk of false positives (Type I error).
- **Effect Size:** Even if a difference is statistically significant, it must be practically meaningful.

# 13 Permutation Test

## 13.1 Introduction

A **permutation test** is a non-parametric statistical method used to assess whether two groups differ significantly. Unlike traditional hypothesis tests that rely on assumptions about normality, the permutation test makes minimal assumptions and is particularly useful for small sample sizes or non-normal data.

The test is based on randomly shuffling the observed data to generate a null distribution, then comparing the observed test statistic to this distribution.

## 13.2 When to Use a Permutation Test

- When normality assumptions of parametric tests (e.g., t-test) do not hold.
- When sample sizes are small, making standard tests unreliable.
- When analyzing experimental or observational data without clear distributional assumptions.

## 13.3 Steps in a Permutation Test

1. **Define the Hypotheses:**
   - $H_0$ (Null Hypothesis): The two groups come from the same distribution.
   - $H_1$ (Alternative Hypothesis): The two groups have different distributions.
2. **Compute the Observed Test Statistic:** Calculate a metric such as the difference in means, medians, or another relevant statistic.
3. **Shuffle (Permute) the Data:** Randomly reassign the observed values between the two groups multiple times.
4. **Recalculate the Test Statistic:** Compute the statistic for each shuffled dataset to form the null distribution.
5. **Compare the Observed Statistic to the Null Distribution:** Compute a p-value based on how extreme the observed statistic is compared to the permuted distribution.

## 13.4 Example: Testing a New Drug vs. Placebo

A researcher tests whether a new drug improves recovery times compared to a placebo. The recovery times (in days) for each group are:

- **Drug Group**: $[5, 7, 6, 4, 5]$
- **Placebo Group**: $[8, 9, 6, 10, 7]$

**Step 1: Compute the Observed Statistic**
The observed difference in mean recovery times is:

$$\bar{X}_{\text{drug}} - \bar{X}_{\text{placebo}} = (5.4 - 8) = -2.6. \tag{20}$$

**Step 2: Generate the Null Distribution**
The values are pooled and randomly reassigned to two groups multiple times (e.g., 10,000 permutations). The mean difference is recalculated for each permutation, forming the null distribution.

**Step 3: Compute the p-value**
The p-value is the proportion of permuted differences as extreme as or more extreme than the observed difference. If $p < 0.05$, the researcher rejects $H_0$ and concludes that the drug significantly affects recovery time.

# 14 t-Test

## 14.1 Introduction

A **t-test** is a statistical method used to determine whether there is a significant difference between the means of one or two groups. It is widely used in hypothesis testing, particularly when the sample size is small and the population variance is unknown.

There are four main types of t-tests:

- **One-Sample t-Test**: Compares the mean of a single sample to a known population mean.
- **Independent (Two-Sample) t-Test (Student's t-Test)**: Compares the means of two independent groups.
- **Welch's t-Test**: A modification of the independent t-test that does not assume equal variances.
- **Paired t-Test**: Compares means from the same subjects under two different conditions (e.g., before and after treatment).

## 14.2 Assumptions of the t-Test

- The data should be approximately **normally distributed**, especially for small samples.
- The samples should be **independent** (except for the paired t-test).
- For the standard two-sample t-test, the **variances of the two groups should be equal** (Welch's t-test relaxes this assumption).

# 15 One-Sample t-Test

## 15.1 Definition

A **one-sample t-test** is used to determine whether the mean of a sample differs significantly from a known population mean.

## 15.2 Hypotheses

- $H_0 : \mu = \mu_0$ (The sample mean equals the population mean).
- $H_1 : \mu \neq \mu_0$ (The sample mean is different from the population mean).

## 15.3 Formula

The test statistic is calculated as:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \tag{21}$$

where:

- $\bar{X}$ is the sample mean,
- $\mu_0$ is the population mean,
- $s$ is the sample standard deviation,
- $n$ is the sample size.

## 15.4 Example: Average Coffee Consumption

A coffee shop owner believes that customers drink an average of $\mu_0 = 3.2$ cups of coffee per day. A sample of 25 customers reports:

- $\bar{X} = 3.5$ cups
- $s = 0.8$ cups

The t-value is:

$$t = \frac{3.5 - 3.2}{0.8/\sqrt{25}} = \frac{0.3}{0.16} = 1.875. \tag{22}$$

The calculated $t$-value is compared to the critical $t$-value for $n - 1 = 24$ degrees of freedom.
If $|t| > t_{\alpha/2}$, the owner concludes that customers drink a different amount than expected.

# 16 Independent (Two-Sample) t-Test (Student's t-Test)

## 16.1 Definition

A **two-sample t-test** compares the means of two independent groups to determine if they are significantly different.

## 16.2 Hypotheses

- $H_0 : \mu_1 = \mu_2$ (The two groups have the same mean).
- $H_1 : \mu_1 \neq \mu_2$ (The two groups have different means).

## 16.3 Formula

The test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{23}$$

## 16.4 Example: Effect of a New Teaching Method

A school tests whether a new teaching method improves student scores:

- **Traditional Method**: $n_1 = 30$, $\bar{X}_1 = 75$, $s_1 = 10$.
- **New Method**: $n_2 = 30$, $\bar{X}_2 = 80$, $s_2 = 12$.

The t-value is:

$$t = \frac{75 - 80}{\sqrt{\frac{10^2}{30} + \frac{12^2}{30}}} = \frac{-5}{\sqrt{3.33 + 4.8}} = \frac{-5}{2.67} = -1.87. \tag{24}$$

# 17 Welch's t-Test

## 17.1 Definition

**Welch's t-test** is a variation of the two-sample t-test that does not assume equal variances. It is more reliable when sample sizes and variances are unequal.

## 17.2 Formula

The t-value is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{25}$$

where:

- $\bar{X}_1, \bar{X}_2$ are the sample means,
- $s_1, s_2$ are the sample standard deviations,
- $n_1, n_2$ are the sample sizes.

Since Welch's t-test does not assume equal variances, the degrees of freedom $(df)$ are approximated using:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}. \tag{26}$$

## 17.3 Example: Salaries in Different Industries

A company compares employee salaries between two industries:

- **Industry A**: $n_1 = 40$, $\bar{X}_1 = 60,000$, $s_1 = 15,000$.
- **Industry B**: $n_2 = 25$, $\bar{X}_2 = 55,000$, $s_2 = 20,000$.

**Step 1: Compute the t-Statistic**

$$\begin{aligned} t &= \frac{60,000 - 55,000}{\sqrt{\frac{(15,000)^2}{40} + \frac{(20,000)^2}{25}}} \\ &= 1.08. \end{aligned}$$

**Step 2: Compute the Degrees of Freedom**

$$\begin{aligned} df &= \frac{\left(\frac{225,000,000}{40} + \frac{400,000,000}{25}\right)^2}{\frac{\left(\frac{225,000,000}{40}\right)^2}{39} + \frac{\left(\frac{400,000,000}{25}\right)^2}{24}} \\ &= 27.67. \end{aligned}$$

**Step 3: Interpretation**

Using a t-table, we compare $t = 1.08$ with the critical value for $df = 27.67$ at $\alpha = 0.05$. Since $|t|$ is less than the critical value ( 2.05 for two-tailed test), we fail to reject $H_0$, meaning there is no significant salary difference.

# 18 Paired t-Test

## 18.1 Definition

A **paired t-test** compares the means of the same subjects measured under two conditions.

## 18.2 Formula

$$t = \frac{\bar{D}}{s_D/\sqrt{n}} \tag{27}$$

where:

- $\bar{D}$ is the mean of the paired differences,
- $s_D$ is the standard deviation of the differences,
- $n$ is the number of pairs.

## 18.3 Example: Effect of a Workout Program

A fitness trainer tests if a 6-week workout program reduces resting heart rate. The resting heart rates of 10 participants are recorded before and after the program:

| **Participant** | **Before (bpm)** | **After (bpm)** |
|---|---|---|
| 1 | 72 | 68 |
| 2 | 75 | 70 |
| 3 | 78 | 73 |
| 4 | 80 | 74 |
| 5 | 76 | 72 |
| 6 | 74 | 70 |
| 7 | 79 | 75 |
| 8 | 77 | 71 |
| 9 | 75 | 70 |
| 10 | 78 | 72 |

Table 5: Heart rate measurements before and after training

**Step 1: Compute Differences and Mean Difference**

- Differences: $D = \{4, 5, 5, 6, 4, 4, 4, 6, 5, 6\}$.
- Mean difference:

$$\bar{D} = \frac{4+5+5+6+4+4+4+6+5+6}{10} = \frac{49}{10} = 4.9. \tag{28}$$

**Step 2: Compute Standard Deviation of Differences**

- Squared differences from mean:

$$s_D^2 = \frac{\sum (D_i - \bar{D})^2}{n-1}. \tag{29}$$

- Individual deviations: $(-0.9, 0.1, 0.1, 1.1, -0.9, -0.9, -0.9, 1.1, 0.1, 1.1)$.
- Squared deviations: $(0.81, 0.01, 0.01, 1.21, 0.81, 0.81, 0.81, 1.21, 0.01, 1.21)$.
- Sum of squared deviations: 6.99.

- Standard deviation:

$$s_D = \sqrt{\frac{6.99}{9}} = \sqrt{0.776} = 0.88. \tag{30}$$

**Step 3: Compute t-Statistic**

$$t = \frac{4.9}{0.88/\sqrt{10}} = \frac{4.9}{0.278} = 17.63. \tag{31}$$

**Step 4: Interpretation**

Using a t-table, we compare $t = 17.63$ to the critical value for $df = 9$ at $\alpha = 0.05$ ( 2.26 for a two-tailed test). Since $t$ is much larger, we reject $H_0$, concluding that the workout significantly reduces heart rate.

# 19 Analysis of Variance (ANOVA)

**Analysis of Variance (ANOVA)** is a statistical method used to compare the means of multiple groups to determine whether at least one of them significantly differs from the others. It is an extension of the t-test for comparing more than two groups. Unlike multiple pairwise comparisons, which involve testing each pair of groups separately, ANOVA provides a single omnibus test to determine if any significant difference exists across all groups.

## 19.1 Key Concepts in ANOVA

- Pairwise Comparison Pairwise comparison tests whether the means of two groups are significantly different. While useful, conducting multiple t-tests increases the risk of Type I error (false positives). ANOVA overcomes this by providing a single overall test, reducing the probability of false findings.
- Omnibus Test The omnibus test in ANOVA assesses whether at least one group mean significantly differs from the others. However, it does not indicate which groups are different. If the ANOVA test is significant, post-hoc tests (e.g., Tukey's HSD) are used for pairwise comparisons.
- Decomposition of Variance ANOVA partitions the total variability in the data into two components:
  - **Between-Group Variance**: The variability due to differences between group means.
  - **Within-Group Variance (Error Variance)**: The variability due to differences within each group.
- Sum of Squares (SS) ANOVA uses **sum of squares** (SS) to measure variance:
  - **Total Sum of Squares ($\mathbf{SS_T}$)**: Measures overall variability in the data.
  - **Between-Group Sum of Squares ($\mathbf{SS_B}$)**: Measures variability due to differences between group means.
  - **Within-Group Sum of Squares ($\mathbf{SS_W}$)**: Measures variability within each group.

  These components follow:

  $$SS_T = SS_B + SS_W. \tag{32}$$

- F-Statistic The test statistic for ANOVA is the F-statistic, defined as:

  $$F = \frac{\text{Between-Group Variance}}{\text{Within-Group Variance}} = \frac{MS_B}{MS_W}. \tag{33}$$

  where:

  - $MS_B = SS_B/df_B$ is the mean square between groups.
  - $MS_W = SS_W/df_W$ is the mean square within groups.
  - $df_B = k - 1$ (degrees of freedom for between-group variance, where $k$ is the number of groups).
  - $df_W = N - k$ (degrees of freedom for within-group variance, where $N$ is the total number of observations).

  If the F-statistic is significantly large, it suggests at least one group mean is different from the others.

| **Group** | **Scores** | **Mean** ($\bar{X}$) | **Variance** ($s^2$) |
|---|---|---|---|
| Method A | 75, 78, 82, 85, 79 | 79.8 | 13.7 |
| Method B | 70, 72, 68, 74, 71 | 71.0 | 6.5 |
| Method C | 85, 88, 92, 90, 86 | 88.2 | 8.7 |

Table 6: Exam Scores by Teaching Method

### 19.2 Example: Examining Exam Scores Across Three Teaching Methods

A researcher examines whether three different teaching methods lead to different exam scores. They randomly assign 15 students into three groups:

**Step 1: Compute Sum of Squares**

$$SS_T = \sum (X_i - \bar{X}_T)^2 = 544.4. \tag{34}$$

$$SS_B = \sum n(\bar{X}_i - \bar{X}_T)^2 = 475.1. \tag{35}$$

$$SS_W = SS_T - SS_B = 544.4 - 475.1 = 69.3. \tag{36}$$

**Step 2: Compute Mean Squares and F-Statistic**

- $df_B = k - 1 = 3 - 1 = 2.$
- $df_W = N - k = 15 - 3 = 12.$
- $MS_B = SS_B/df_B = 475.1/2 = 237.55.$
- $MS_W = SS_W/df_W = 69.3/12 = 5.78.$

$$F = \frac{MS_B}{MS_W} = \frac{237.55}{5.78} = 41.1. \tag{37}$$

**Step 3: Interpretation**

Comparing $F = 41.1$ to the critical F-value from an F-table ($F_{2,12} \approx 3.89$ at $\alpha = 0.05$), we see that $41.1 > 3.89$. Thus, we reject $H_0$, concluding that at least one teaching method significantly affects exam scores.

### 19.3 Post-Hoc Analysis: Tukey's HSD

Since ANOVA indicates a significant difference, we perform **Tukey's Honest Significant Difference (HSD)** test to determine which groups differ.

$$HSD = q \times \sqrt{\frac{MS_W}{n}}. \tag{38}$$

For $q = 3.77$ (from Tukey's table at $\alpha = 0.05$):

$$HSD = 3.77 \times \sqrt{\frac{5.78}{5}} = 3.77 \times 1.07 = 4.03. \tag{39}$$

Comparing mean differences:

- $|79.8 - 71.0| = 8.8$ (Significant)
- $|79.8 - 88.2| = 8.4$ (Significant)
- $|71.0 - 88.2| = 17.2$ (Significant)

Since all differences exceed 4.03, all groups significantly differ from each other.

# 20 Two-Way Analysis of Variance (ANOVA)

**Two-Way ANOVA** is a statistical method used to analyze the effects of two independent categorical variables (factors) on a continuous dependent variable. It extends one-way ANOVA by evaluating the individual effects of each factor (main effects) and their combined effect (interaction effect).

## 20.1 Key Concepts in Two-Way ANOVA

- **Main Effects**: The independent effect of each factor on the dependent variable.
- **Interaction Effect**: Whether the effect of one factor depends on the level of the other factor.
- **Decomposition of Variance**: Total variability is partitioned into three components:
  - **Factor A Variance**: Variability due to differences in levels of factor A.
  - **Factor B Variance**: Variability due to differences in levels of factor B.
  - **Interaction Variance**: Variability due to the combined influence of both factors.
  - **Error Variance**: Variability within groups not explained by factors.

**Sum of Squares (SS) Components**:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_W. \tag{40}$$

Where:

- $SS_T$ = Total sum of squares
- $SS_A$ = Sum of squares for factor A
- $SS_B$ = Sum of squares for factor B
- $SS_{AB}$ = Sum of squares for the interaction
- $SS_W$ = Within-group sum of squares (error variance)

**F-Statistics for Each Effect:**

$$F_A = \frac{MS_A}{MS_W}, \quad F_B = \frac{MS_B}{MS_W}, \quad F_{AB} = \frac{MS_{AB}}{MS_W}. \tag{41}$$

Where:

- $MS_A = SS_A/df_A$ is the mean square for factor A.
- $MS_B = SS_B/df_B$ is the mean square for factor B.
- $MS_{AB} = SS_{AB}/df_{AB}$ is the mean square for the interaction.
- $MS_W = SS_W/df_W$ is the mean square error.

## 20.2 Example: Examining Exam Scores Based on Teaching Method and Study Environment

A researcher investigates whether students' exam scores are influenced by **Teaching Method (A, B, C)** and **Study Environment (Quiet, Noisy)**. Each student is randomly assigned to a combination of teaching method and study environment.

**Step 1: Compute Sum of Squares**

$$SS_T = \sum (X_i - \bar{X}_T)^2. \tag{42}$$

| **Teaching Method** | **Study Environment** | **Scores** | **Mean ($\bar{X}$)** |
|---|---|---|---|
| Method A | Quiet | 75, 78, 82 | 78.3 |
| Method A | Noisy | 68, 72, 71 | 70.3 |
| Method B | Quiet | 80, 82, 85 | 82.3 |
| Method B | Noisy | 73, 74, 76 | 74.3 |
| Method C | Quiet | 85, 88, 90 | 87.7 |
| Method C | Noisy | 77, 79, 81 | 79.0 |

Table 7: Exam Scores by Teaching Method and Study Environment

$$SS_A = \sum n_A(\bar{X}_A - \bar{X}_T)^2. \tag{43}$$

$$SS_B = \sum n_B(\bar{X}_B - \bar{X}_T)^2. \tag{44}$$

$$SS_{AB} = \sum n_{AB}(\bar{X}_{AB} - \bar{X}_T)^2. \tag{45}$$

$$SS_W = SS_T - (SS_A + SS_B + SS_{AB}). \tag{46}$$

**Step 2: Compute Mean Squares and F-Statistics**

- $df_A = a - 1 = 3 - 1 = 2$ (for Teaching Method).
- $df_B = b - 1 = 2 - 1 = 1$ (for Study Environment).
- $df_{AB} = (a - 1)(b - 1) = (3 - 1)(2 - 1) = 2$ (for Interaction).
- $df_W = N - ab = 18 - 6 = 12$ (for Within-Group variance).

$$MS_A = SS_A/df_A, \quad MS_B = SS_B/df_B, \quad MS_{AB} = SS_{AB}/df_{AB}, \quad MS_W = SS_W/df_W. \tag{47}$$

$$F_A = \frac{MS_A}{MS_W}, \quad F_B = \frac{MS_B}{MS_W}, \quad F_{AB} = \frac{MS_{AB}}{MS_W}. \tag{48}$$

**Step 3: Interpretation** - If $F_A$ is significant, teaching method affects scores. - If $F_B$ is significant, study environment affects scores. - If $F_{AB}$ is significant, teaching method and study environment interact.

## 20.3 Post-Hoc Analysis: Tukey's HSD

If the ANOVA test shows a significant main effect, post-hoc analysis determines which groups differ.

$$HSD = q \times \sqrt{\frac{MS_W}{n}}. \tag{49}$$

For $q = 3.77$ (from Tukey's table at $\alpha = 0.05$):

$$HSD = 3.77 \times \sqrt{\frac{MS_W}{n}}. \tag{50}$$

Comparing mean differences:

- $|78.3 - 70.3| = 8.0$ (Check if $> HSD$)
- $|82.3 - 74.3| = 8.0$ (Check if $> HSD$)
- $|87.7 - 79.0| = 8.7$ (Check if $> HSD$)

**Conclusion:** If $F_{AB}$ is significant, the effect of teaching method depends on the study environment.

# 21 Chi-Square Test

**Chi-Square Test** is a non-parametric statistical method used to examine the relationship between categorical variables. It determines whether there is a significant association between two categorical variables in a contingency table.

## 21.1 Key Concepts in Chi-Square Test

- **Categorical Variables**: The test applies to data divided into discrete categories (e.g., gender, preference, education level).
- **Observed vs. Expected Frequencies**: The test compares the actual data (observed frequencies) with the frequencies expected under the assumption of independence.
- **Independence**: The null hypothesis states that the two categorical variables are independent, meaning changes in one variable do not affect the other.

**Types of Chi-Square Tests**:

- **Chi-Square Test for Independence**: Determines if two categorical variables are associated.
- **Chi-Square Goodness-of-Fit Test**: Evaluates if a sample distribution matches an expected theoretical distribution.

## 21.2 Chi-Square Test for Independence

This test assesses whether two categorical variables are related by comparing observed and expected frequencies.

**Hypotheses**:

- $H_0$ (Null Hypothesis): The two categorical variables are independent.
- $H_a$ (Alternative Hypothesis): The two categorical variables are dependent (associated).

**Formula:** The chi-square test statistic is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{51}$$

where:

- $O_{ij}$ = Observed frequency in cell $i, j$.
- $E_{ij}$ = Expected frequency in cell $i, j$, calculated as:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}. \tag{52}$$

**Degrees of Freedom:**

$$df = (r-1)(c-1), \tag{53}$$

where $r$ is the number of rows and $c$ is the number of columns.

| Gender | Product A | Product B | Total |
|---|---|---|---|
| Male | 30 | 20 | 50 |
| Female | 40 | 30 | 70 |
| **Total** | 70 | 50 | 120 |

Table 8: Contingency Table of Gender and Product Preference

### 21.3 Example: Examining the Relationship Between Gender and Product Preference

A company surveys customers on their preferred product (A or B) and records responses by gender.

**Step 1: Compute Expected Frequencies**

- $E_{Male,A} = \frac{(50\times70)}{120} = 29.2.$
- $E_{Male,B} = \frac{(50\times50)}{120} = 20.8.$
- $E_{Female,A} = \frac{(70\times70)}{120} = 40.8.$
- $E_{Female,B} = \frac{(70\times50)}{120} = 29.2.$

**Step 2: Compute the Chi-Square Statistic**

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{54}$$

$$\chi^2 = \frac{(30-29.2)^2}{29.2} + \frac{(20-20.8)^2}{20.8} + \frac{(40-40.8)^2}{40.8} + \frac{(30-29.2)^2}{29.2}. \tag{55}$$

$$\chi^2 = \frac{0.64}{29.2} + \frac{0.64}{20.8} + \frac{0.64}{40.8} + \frac{0.64}{29.2} = 0.0219 + 0.0308 + 0.0157 + 0.0219 = 0.0903. \tag{56}$$

**Step 3: Determine Significance**

- Degrees of freedom: $df = (2-1)(2-1) = 1$.
- Critical value at $\alpha = 0.05$ from the chi-square table: 3.84.
- Since $0.0903 < 3.84$, we fail to reject $H_0$ (no significant association).

### 21.4 Chi-Square Goodness-of-Fit Test

This test determines whether observed categorical data matches a theoretical expected distribution.

**Hypotheses**:

- $H_0$: The observed distribution matches the expected distribution.
- $H_a$: The observed distribution differs from the expected distribution.

**Example: Testing Survey Responses**

A company predicts that customers are equally likely to prefer three different flavors ($\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$). They collect responses:

**Chi-Square Calculation:**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}. \tag{57}$$

$$\chi^2 = \frac{(50-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50}. \tag{58}$$

$$\chi^2 = 0 + \frac{100}{50} + \frac{100}{50} = 0 + 2 + 2 = 4. \tag{59}$$

**Step 3: Compare with Critical Value**

| **Flavor** | **Observed** | **Expected** |
|---|---|---|
| Chocolate | 50 | $150 \times \frac{1}{3} = 50$ |
| Vanilla | 40 | $150 \times \frac{1}{3} = 50$ |
| Strawberry | 60 | $150 \times \frac{1}{3} = 50$ |
| **Total** | 150 | 150 |

Table 9: Observed vs. Expected Preferences

- $df = k - 1 = 3 - 1 = 2$.
- Critical value from chi-square table ($\alpha = 0.05$) is 5.99.
- Since $4 < 5.99$, we fail to reject $H_0$ (data is consistent with the expected distribution).

**Conclusion:** Chi-square tests help determine relationships between categorical variables (independence test) and whether distributions follow expected patterns (goodness-of-fit test).

# 22 Simple Linear Regression

**Simple Linear Regression** is a statistical method used to model the relationship between a dependent variable ($Y$) and a single independent variable ($X$). The goal is to find the best-fitting line that predicts $Y$ based on $X$.

## 22.1 Key Concepts in Simple Linear Regression

- **Dependent Variable** ($Y$): The outcome we want to predict.
- **Independent Variable** ($X$): The predictor variable.
- **Regression Line**: The best-fit line that minimizes the differences between observed and predicted values.
- **Residuals**: The differences between observed and predicted values.

## 22.2 Regression Equation

The equation of a simple linear regression model is:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{60}$$

where:

- $Y$ = Dependent variable (response).
- $X$ = Independent variable (predictor).
- $\beta_0$ = Intercept (value of $Y$ when $X = 0$).
- $\beta_1$ = Slope (change in $Y$ for a one-unit change in $X$).
- $\epsilon$ = Error term (captures random variability).

## 22.3 Estimating Parameters

The slope ($\beta_1$) and intercept ($\beta_0$) are estimated using the least squares method, which minimizes the sum of squared residuals:

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{61}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{62}$$

## 22.4 Example: Predicting Exam Scores Based on Study Hours

A researcher collects data on students' study hours ($X$) and their exam scores ($Y$).

| **Study Hours** ($X$) | **Exam Score** ($Y$) |
|---|---|
| 2 | 50 |
| 3 | 55 |
| 5 | 65 |
| 7 | 75 |
| 9 | 85 |

Table 10: Study Hours vs. Exam Scores

**Step 1: Compute Means**

$$\bar{X} = \frac{2+3+5+7+9}{5} = 5.2, \quad \bar{Y} = \frac{50+55+65+75+85}{5} = 66$$

**Step 2: Compute Slope** ($\beta_1$)

$$\begin{aligned}\sum(X_i-\bar{X})(Y_i-\bar{Y}) &= (2-5.2)(50-66)+(3-5.2)(55-66)+(5-5.2)(65-66)+(7-5.2)(75-66)+(9-5.2)(85-66)\\ &= (-3.2)(-16)+(-2.2)(-11)+(-0.2)(-1)+(1.8)(9)+(3.8)(19)\\ &= 51.2+24.2+0.2+16.2+72.2 = 164\end{aligned}$$

$$\begin{aligned}\sum(X_i-\bar{X})^2 &= (2-5.2)^2+(3-5.2)^2+(5-5.2)^2+(7-5.2)^2+(9-5.2)^2\\ &= (-3.2)^2+(-2.2)^2+(-0.2)^2+(1.8)^2+(3.8)^2\\ &= 10.24+4.84+0.04+3.24+14.44 = 32.8\end{aligned}$$

$$\beta_1 = \frac{164}{32.8} = 5$$

**Step 3: Compute Intercept** ($\beta_0$)

$$\beta_0 = \bar{Y} - \beta_1\bar{X} = 66 - (5 \times 5.2) = 66 - 26 = 40$$

**Final Regression Equation:**

$$Y = 40 + 5X \tag{63}$$

## 22.5 Interpreting the Results

- **Intercept** ($\beta_0 = 40$): If a student studies for 0 hours, their expected score is 40.
- **Slope** ($\beta_1 = 5$): Each additional study hour increases the expected score by 5 points.

## 22.6 Goodness of Fit: $R^2$

The coefficient of determination ($R^2$) measures how well the model explains variability in $Y$:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residuals}}{SS_{Total}}. \tag{64}$$

**Key Interpretations**:

- $R^2 = 1$: Perfect fit.
- $R^2 = 0$: Model explains no variability in $Y$.
- Higher $R^2$ values indicate better model performance.

## 22.7 Significance Testing: t-Test for $\beta_1$

To determine if $X$ significantly predicts $Y$, we test:

- $H_0 : \beta_1 = 0$ (No relationship).
- $H_a : \beta_1 \neq 0$ (Significant relationship).

The t-statistic is:

$$t = \frac{\beta_1}{SE_{\beta_1}}, \tag{65}$$

where $SE_{\beta_1}$ is the standard error of the slope.

**Decision Rule:** - Compare $|t|$ to the critical value from the t-distribution. - If $p < \alpha$ (e.g., 0.05), reject $H_0$ (significant relationship).

## 22.8 Conclusion

Simple linear regression provides a powerful way to model relationships between two variables, estimate trends, and make predictions. However, it assumes:

- Linearity between $X$ and $Y$.
- Homoscedasticity (constant variance of errors).
- Normally distributed residuals.
- No strong multicollinearity (only applies to multiple regression).

When assumptions hold, regression models provide interpretable, useful insights for decision-making.

# 23 Multiple Linear Regression

**Multiple Linear Regression (MLR)** is an extension of simple linear regression that models the relationship between a dependent variable ($Y$) and multiple independent variables ($X_1, X_2, \dots, X_p$). It is used when more than one predictor variable is necessary to explain variations in the dependent variable.

## 23.1 Regression Equation

The general form of a multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{66}$$

where:

- $Y$ = Dependent variable (response).
- $X_1, X_2, ..., X_p$ = Independent variables (predictors).
- $\beta_0$ = Intercept (value of $Y$ when all $X$'s are zero).
- $\beta_1, \beta_2, ..., \beta_p$ = Regression coefficients (effect of each $X$ on $Y$).
- $\epsilon$ = Error term capturing random variability.

## 23.2 Estimating Parameters

The regression coefficients ($\beta$) are estimated using the **Ordinary Least Squares (OLS)** method by minimizing the sum of squared residuals:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{67}$$

where:

- $\mathbf{X}$ = Matrix of independent variables (including a column of ones for the intercept).
- $\mathbf{Y}$ = Vector of observed values.
- $\mathbf{b}$ = Vector of estimated regression coefficients.

## 23.3 Goodness of Fit: $R^2$ and Adjusted $R^2$

**Coefficient of Determination ($R^2$)**:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residuals}}{SS_{Total}}. \tag{68}$$

- $R^2$ measures the proportion of variance in $Y$ explained by the model.
- $R^2 = 1$ indicates perfect fit, $R^2 = 0$ means no explanatory power.

**Adjusted** $R^2$ corrects for the number of predictors:

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n-1)}{n-p-1}\right). \tag{69}$$

- Penalizes adding unnecessary predictors.
- Helps prevent overfitting.

| **Size (sq ft) $X_1$** | **Bedrooms $X_2$** | **Price ($Y$ in 1000s)** |
|---|---|---|
| 1500 | 3 | 300 |
| 1800 | 4 | 360 |
| 2100 | 3 | 420 |
| 2500 | 5 | 500 |
| 2800 | 4 | 550 |

Table 11: House Prices Dataset

## 23.4 Example: Predicting House Prices Based on Size and Bedrooms

A researcher models house prices ($Y$) based on house size ($X_1$*insquarefeet)andthenumberofbedrooms*($X_2$).

**Regression Model:**

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{70}$$

Assume the estimated coefficients are:

$$\hat{Y} = 50 + 0.18X_1 + 10X_2. \tag{71}$$

**Interpretation:**

- Intercept $\beta_0 = 50$: Predicted price when $X_1 = 0$ and $X_2 = 0$ (not practically meaningful).
- $\beta_1 = 0.18$: Each additional square foot increases price by 0.18 (or $180 per sq ft).
- $\beta_2 = 10$: Each additional bedroom increases price by $10,000.

## 23.5 Hypothesis Testing for Coefficients

Each coefficient is tested using:

$$t = \frac{\beta_j}{SE_{\beta_j}} \tag{72}$$

where $SE_{\beta_j}$ is the standard error of $\beta_j$.

**Decision Rule:** - If $p < \alpha$ (e.g., 0.05), reject $H_0$ (significant predictor). - If $p > \alpha$, fail to reject $H_0$ (not significant).

## 23.6 Conclusion

Multiple linear regression provides a flexible approach to modeling relationships between multiple predictors and an outcome variable. However, it is crucial to:

- Interpret coefficients carefully.
- Validate assumptions using diagnostic tests.
- Avoid overfitting by selecting relevant predictors.

When assumptions hold, the OLS estimator remains BLUE, ensuring reliable and unbiased predictions.

# 24 Interpreting Regression Coefficients for Different Data Types

Regression coefficients represent the estimated effect of an independent variable on the dependent variable while holding all other variables constant. The interpretation of coefficients varies depending on the type of independent variable.

## 24.1 Binary Variables (0/1, Dummy Variables)

**Example: Predicting Salary Based on Gender**

$$\text{Salary} = \beta_0 + \beta_1 \text{Gender} \tag{73}$$

where:

- Gender = 1 if Male, 0 if Female.
- $\beta_1$ represents the average difference in salary between males and females.

**Interpretation:** If $\beta_1 = 5000$, then males earn \$5,000 more on average than females, assuming all else is constant.

## 24.2 Categorical (Nominal) Variables with Dummy Coding

Categorical variables with $k$ levels require $k - 1$ dummy variables.

**Example: Predicting Test Scores Based on Education Level**

$$\text{Score} = \beta_0 + \beta_1 D_1 + \beta_2 D_2 \tag{74}$$

where:

- $D_1 = 1$ if Bachelor's, 0 otherwise.
- $D_2 = 1$ if Master's, 0 otherwise.
- The reference group (PhD) is not included.

**Interpretation:** - $\beta_1$ represents the difference in test scores between Bachelor's and PhD. - $\beta_2$ represents the difference in test scores between Master's and PhD.

## 24.3 Ordinal Variables

Ordinal variables maintain order but do not have equal spacing. They can be:

- Treated as continuous (if approximately equidistant).
- Coded as dummy variables.

**Example: Predicting Job Satisfaction Based on Work Stress (Low, Medium, High)**

$$\text{Satisfaction} = \beta_0 + \beta_1 D_{\text{Medium}} + \beta_2 D_{\text{High}} \tag{75}$$

**Interpretation:** - If $\beta_1 = -2$, medium stress workers report 2 points lower satisfaction than low-stress workers. - If $\beta_2 = -5$, high-stress workers report 5 points lower satisfaction than low-stress workers.

## 24.4 Discrete Variables (Count Data)

**Example: Predicting Customer Purchases Based on Number of Visits**

$$\text{Purchases} = \beta_0 + \beta_1 \text{Visits} \tag{76}$$

**Interpretation:** If $\beta_1 = 0.3$, each additional store visit increases expected purchases by 0.3.

For Poisson regression (log transformation):

$$\log(\text{Purchases}) = \beta_0 + \beta_1 \text{Visits} \tag{77}$$

Here, $e^{\beta_1}$ represents the multiplicative change in purchases per additional visit.

## 24.5 Continuous Variables (Interval and Ratio)

**Example: Predicting Blood Pressure Based on Age**

$$\text{BP} = \beta_0 + \beta_1 \text{Age} \tag{78}$$

**Interpretation:** If $\beta_1 = 1.2$, every one-year increase in age is associated with a 1.2 mmHg increase in blood pressure.

## 24.6 Interaction Terms

When the effect of one variable depends on another variable, an interaction term is included.

**Example: Predicting Salary Based on Education and Experience**

$$\text{Salary} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Experience} + \beta_3 (\text{Education} \times \text{Experience}) \tag{79}$$

**Interpretation:** - $\beta_1$ represents the effect of education when experience is zero. - $\beta_2$ represents the effect of experience when education is zero. - $\beta_3$ represents how education modifies the effect of experience.

## 24.7 Comprehensive Example: Interpreting Regression Coefficients in a Realistic Model

To demonstrate coefficient interpretation across different variable types, consider a multiple regression model predicting annual salary ($Y) based on the following factors:

- $X_1$: Years of experience (continuous, ratio).
- $X_2$: Education level (categorical, ordinal: Bachelor's, Master's, PhD).
- $X_3$: Gender (binary: 1 for Male, 0 for Female).
- $X_4$: Industry (categorical, nominal: Tech, Finance, Education with Education as the reference).
- $X_5$: An interaction term between experience and education.

The estimated regression equation is:

$$\text{Salary} = 35,000 + 2,500X_1 + 8,000D_{\text{Master's}} + 15,000D_{\text{PhD}} + 5,000X_3 + 12,000D_{\text{Tech}} + 10,000D_{\text{Finance}} + 500(X_1 \times D_{\text{PhD}}) + \epsilon \tag{80}$$

**Interpretation of Coefficients:**

- **Intercept (35,000)**: The baseline predicted salary for a female ($X_3 = 0$) with a Bachelor's degree ($D_{\text{Master's}} = 0$, $D_{\text{PhD}} = 0$) working in the Education industry ($D_{\text{Tech}} = 0$, $D_{\text{Finance}} = 0$) and zero years of experience ($X_1 = 0$).
- **Experience ($X_1 = 2,500$)**: Each additional year of experience increases salary by $2,500, assuming no interaction effect.
- **Education Level:**
  - $\beta_2 = 8,000$ (Master's Degree): Holding all else constant, individuals with a Master's degree earn $8,000 more than those with a Bachelor's.
  - $\beta_3 = 15,000$ (PhD): Holding all else constant, individuals with a PhD earn $15,000 more than those with a Bachelor's.
- **Gender ($X_3 = 5,000$)**: Males ($X_3 = 1$) earn $5,000 more than females ($X_3 = 0$), holding all other variables constant.

- **Industry Type:**
  - $\beta_5 = 12,000$ (Tech Industry): Working in Tech increases salary by $12,000 compared to Education.
  - $\beta_6 = 10,000$ (Finance Industry): Working in Finance increases salary by $10,000 compared to Education.
- **Interaction Effect ($\beta_7 = 500$):** The interaction term modifies the effect of experience for PhD holders. A PhD holder earns an additional $500 per year of experience, meaning their experience-based salary increase is:
$$(2,500 + 500) = 3,000 \text{ per year.} \tag{81}$$

**Example Calculation:** Consider a male employee ($X_3 = 1$) with a PhD ($D_{\text{PhD}} = 1$, $D_{\text{Master's}} = 0$), working in Tech ($D_{\text{Tech}} = 1$, $D_{\text{Finance}} = 0$), with 10 years of experience ($X_1 = 10$).

$$\begin{aligned}
\text{Predicted Salary} &= 35,000 + (2,500 \times 10) + (15,000 \times 1) + (5,000 \times 1) + (12,000 \times 1) + (500 \times 10 \times 1) \\
&= 35,000 + 25,000 + 15,000 + 5,000 + 12,000 + 5,000 \\
&= 97,000.
\end{aligned}$$

**Final Interpretation:**

- The base salary (without experience or additional factors) is $35,000.
- This employee gains $2,500 per year of experience plus an extra $500 due to holding a PhD.
- His PhD degree gives him an additional $15,000.
- Being male adds $5,000 to his salary.
- Working in Tech increases his salary by $12,000.
- The final predicted salary is $97,000.

This example illustrates how different types of variables (binary, categorical, discrete, and continuous) and interaction terms influence salary prediction in multiple linear regression.

## 24.8 Conclusion

Interpreting regression coefficients correctly depends on:

- The type of independent variable.
- Whether it is transformed (log, interaction).
- The reference category (for categorical variables).

Proper interpretation ensures meaningful insights from regression analysis.

# 25 Robust Standard Errors

**Robust Standard Errors (RSEs)** are adjustments to the standard errors of regression coefficients that provide valid statistical inference when classical assumptions (e.g., homoscedasticity and independence) are violated. They are particularly useful in the presence of heteroscedasticity or autocorrelation.

## 25.1 Why Use Robust Standard Errors?

In Ordinary Least Squares (OLS) regression, standard errors are computed under the assumption of homoscedastic residuals (constant variance):

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{82}$$

When heteroscedasticity is present, standard errors are incorrectly estimated, leading to unreliable hypothesis testing. Robust standard errors correct for this issue by adjusting the variance-covariance matrix.

## 25.2 Types of Robust Standard Errors

### 25.2.1 White-Huber Robust Standard Errors

White's (or Huber-White) standard errors correct for heteroscedasticity in cross-sectional data. The robust variance-covariance matrix is:

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i^T\right)(\mathbf{X}^T\mathbf{X})^{-1}. \tag{83}$$

**Interpretation:** - Ensures valid statistical inference when heteroscedasticity is present. - Does not assume homoscedastic errors.

### 25.2.2 Clustered Standard Errors

Clustered standard errors adjust for within-group correlation, commonly used in panel data or hierarchical datasets where residuals may be correlated within clusters (e.g., individuals in the same company).

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{c=1}^{C} \mathbf{X}_c^T \hat{e}_c \hat{e}_c^T \mathbf{X}_c\right)(\mathbf{X}^T\mathbf{X})^{-1}. \tag{84}$$

**Interpretation:** - Accounts for correlation within clusters. - Recommended for repeated observations on the same unit (e.g., firms, states).

### 25.2.3 Heteroscedasticity and Autocorrelation Consistent (HAC) Standard Errors

Also known as Newey-West standard errors, these are used when errors exhibit both heteroscedasticity and autocorrelation (common in time-series data).

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{t=1}^{T} \hat{e}_t^2 \mathbf{x}_t \mathbf{x}_t^T + \sum_{j=1}^{L} w_j \sum_{t=j+1}^{T} \hat{e}_t \hat{e}_{t-j} \mathbf{x}_t \mathbf{x}_{t-j}^T\right)(\mathbf{X}^T\mathbf{X})^{-1}. \tag{85}$$

where $w_j$ are kernel weights.

**Interpretation:** - Corrects for heteroscedasticity and serial correlation in time-series data. - Commonly used in financial and macroeconomic studies.

## 25.3 When to Use Robust Standard Errors

- **Use White-Huber standard errors** when residual variance is not constant across observations.
- **Use clustered standard errors** when data are grouped (e.g., firm-level, state-level).
- **Use Newey-West standard errors** when errors exhibit autocorrelation, typically in time-series data.

## 25.4 Example: Impact of Education and Experience on Salary

A researcher estimates the effect of education and experience on salary:

$$\text{Salary} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Experience} + \epsilon. \tag{86}$$

After running an OLS regression, they suspect heteroscedasticity. To ensure valid inference, they compute robust standard errors:

| Variable | OLS Standard Error | Robust Standard Error |
|---|---|---|
| Education | 1.5 | 2.2 |
| Experience | 0.8 | 1.3 |

Table 12: Comparison of Standard Errors

**Interpretation:** - The robust standard errors are larger than the OLS standard errors, suggesting heteroscedasticity was present. - Using robust standard errors avoids overconfident conclusions.

## 25.5 Conclusion

Robust standard errors are essential when OLS assumptions are violated. Choosing the right type depends on the data structure:

- White-Huber for cross-sectional heteroscedasticity.
- Clustered for grouped data.
- HAC (Newey-West) for time-series autocorrelation.

By correctly specifying robust standard errors, researchers ensure valid hypothesis testing and improve model reliability.

# 26 Model Fit, Diagnostics, and Selection

Assessing the validity and reliability of a regression model involves evaluating goodness-of-fit, checking diagnostic assumptions, and using model selection criteria. This section covers essential methods for evaluating regression models.

## 26.1 Residual Analysis

Residuals measure the difference between the observed and predicted values:

$$e_i = Y_i - \hat{Y}_i. \tag{87}$$

**Why Residuals Matter:**

- Residuals should be randomly distributed with a mean of zero.
- Systematic patterns indicate misspecification (e.g., omitted variables, incorrect functional form).
- Large residuals may suggest influential observations or outliers.

## 26.2 Goodness-of-Fit Metrics

### 26.2.1 R-Squared ($R^2$) and Adjusted $R^2$

**Definition:** $R^2$ measures the proportion of variance in $Y$ explained by the independent variables.

$$R^2 = 1 - \frac{SS_R}{SS_T}, \tag{88}$$

where:

- $SS_R = \sum(Y_i - \hat{Y}_i)^2$ (Residual Sum of Squares).
- $SS_T = \sum(Y_i - \bar{Y})^2$ (Total Sum of Squares).

**Adjusted $R^2$:** Adjusted $R^2$ accounts for the number of predictors, preventing overestimation of fit:

$$R^2_{\text{adj}} = 1 - \left( \frac{(1 - R^2)(n - 1)}{n - p - 1} \right), \tag{89}$$

where $p$ is the number of predictors and $n$ is the sample size.

**Interpretation:** - $R^2 = 0.8$ means the model explains 80% of the variance in $Y$. - Adjusted $R^2$ is preferred when comparing models with different numbers of predictors.

### 26.2.2 Pseudo-$R^2$ for Generalized Linear Models (GLMs)

For non-linear models (e.g., logistic, Poisson regression), traditional $R^2$ is not applicable. Instead, Pseudo-$R^2$ provides an alternative measure:

- **McFadden's Pseudo-$R^2$**:
$$R^2_{\text{McF}} = 1 - \frac{\log L_{\text{model}}}{\log L_{\text{null}}}. \tag{90}$$
- **Cox-Snell Pseudo-$R^2$**:
$$R^2_{\text{CS}} = 1 - \exp\left( \frac{2(\log L_{\text{null}} - \log L_{\text{model}})}{n} \right). \tag{91}$$

**Interpretation:**

- Used in logistic and Poisson regression models.
- Higher values indicate better fit, but interpretation differs from traditional $R^2$.

## 26.3 Model Significance and Error Metrics

### 26.3.1 F-Test for Overall Model Significance

**Definition:** The F-test checks whether at least one predictor significantly explains $Y$.

$$F = \frac{(SS_T - SS_R)/p}{SS_R/(n-p-1)}. \tag{92}$$

**Hypotheses:**

- $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ (No predictor is significant).
- $H_a$ : At least one $\beta_j \neq 0$.

**Interpretation:**

- A significant F-test ($p < 0.05$) suggests the model is useful.
- The test does not indicate which predictors are significant—this requires individual t-tests.

### 26.3.2 Root Mean Squared Error (RMSE)

**Definition:** RMSE measures the average prediction error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}. \tag{93}$$

**Interpretation:**

- Lower RMSE indicates better predictive accuracy.
- Best used for comparing models on the same scale.

### 26.3.3 Residual Standard Error (RSE)

**Definition:** RSE estimates the standard deviation of residuals.

$$RSE = \sqrt{\frac{SS_R}{n-p-1}}. \tag{94}$$

**Interpretation:**

- Measures how much actual values deviate from model predictions.
- Smaller RSE suggests better model fit.

## 26.4 Model Selection Criteria

### 26.4.1 Akaike Information Criterion (AIC)

**Definition:** AIC balances model fit and complexity.

$$AIC = -2\log L + 2p. \tag{95}$$

**Interpretation:**

- Lower AIC suggests a better model.
- Penalizes excessive parameters to avoid overfitting.

### 26.4.2 Bayesian Information Criterion (BIC)

**Definition:** BIC introduces a stronger penalty for model complexity.

$$BIC = -2 \log L + p \log n. \tag{96}$$

**Interpretation:**

- Similar to AIC but penalizes large models more heavily.
- Lower BIC indicates a better model.

### 26.4.3 Log-Likelihood

**Definition:** Log-likelihood measures how well the model explains the observed data.

$$\log L = \sum_{i=1}^{n} \log P(Y_i | X_i). \tag{97}$$

**Interpretation:**

- Higher log-likelihood values indicate better fit.
- Used in likelihood ratio tests for nested models.

## 26.5 Conclusion

Model evaluation requires multiple criteria:

- Goodness-of-fit: Use $R^2$, Adjusted $R^2$, or Pseudo $R^2$.
- Significance tests: Use the F-test to determine if predictors matter.
- Error metrics: RMSE and RSE measure prediction accuracy.
- Model selection: AIC, BIC, and log-likelihood help compare competing models.

A well-fitted model should explain a significant portion of variance while maintaining simplicity and generalizability. When diagnostics indicate violations of assumptions, applying corrective measures such as transformations, robust regression, or regularization (e.g., Ridge or Lasso) can improve model reliability.

# 27 Lasso and Ridge Regression

Regularization techniques such as Lasso and Ridge Regression are used in linear models to prevent overfitting by adding a penalty term to the loss function. These methods are especially useful when dealing with high-dimensional datasets where traditional Ordinary Least Squares (OLS) regression may suffer from multicollinearity or model complexity.

## 27.1 Lasso Regression (Least Absolute Shrinkage and Selection Operator)

### 27.1.1 Definition and Objective

Lasso regression is a form of penalized regression that performs both variable selection and shrinkage by adding an $L_1$-norm penalty to the sum of squared residuals.

**Objective Function:**

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{98}$$

where:

- $\sum (Y_i - X_i\beta)^2$ is the residual sum of squares (RSS).
- $\lambda \sum |\beta_j|$ is the $L_1$-norm penalty.
- $\lambda$ controls the degree of regularization.
- If $\lambda = 0$, the model reduces to standard OLS regression.
- As $\lambda$ increases, some coefficients $\beta_j$ shrink to exactly zero, effectively performing variable selection.

### 27.1.2 Interpretation of Coefficients

- Unlike OLS, Lasso sets some coefficients to exactly zero, removing irrelevant predictors. - Helps in identifying the most important variables in the model. - Reduces overfitting by selecting only significant predictors.

### 27.1.3 Example: Predicting House Prices

Suppose we are modeling house prices using features such as square footage, number of bedrooms, location, and age of the house:

$$\text{Price} = \beta_0 + \beta_1\text{Size} + \beta_2\text{Bedrooms} + \beta_3\text{Location} + \beta_4\text{Age} + \epsilon. \tag{99}$$

Applying Lasso regression: - If $\beta_4 = 0$, the model determines that house age does not contribute significantly to predicting price. - If $\beta_2$ is shrunk but nonzero, the number of bedrooms is somewhat informative but less critical.

### 27.1.4 When to Use Lasso Regression

- When there are many predictors, and feature selection is needed.
- When you expect that some variables have no effect (sparse solutions).
- When multicollinearity exists (it selects only one correlated predictor).

—

## 27.2 Ridge Regression (Tikhonov Regularization)

### 27.2.1 Definition and Objective

Ridge regression is another form of penalized regression that shrinks coefficients but does not set them to zero. Instead, it adds an $L_2$-norm penalty to the loss function, ensuring all coefficients remain small.

**Objective Function:**

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{100}$$

where:

- $\sum (Y_i - X_i\beta)^2$ is the residual sum of squares (RSS).
- $\lambda \sum \beta_j^2$ is the $L_2$-norm penalty.
- Unlike Lasso, Ridge does not force coefficients to zero but instead shrinks them toward zero.

### 27.2.2 Interpretation of Coefficients

- Unlike Lasso, Ridge retains all variables but reduces their influence by shrinking coefficients. - Useful when all predictors contribute to the response variable. - Prevents overfitting by reducing the model's complexity.

### 27.2.3 Example: Predicting House Prices

Consider the same house price model:

$$\text{Price} = \beta_0 + \beta_1\text{Size} + \beta_2\text{Bedrooms} + \beta_3\text{Location} + \beta_4\text{Age} + \epsilon. \tag{101}$$

Applying Ridge regression: - All coefficients are shrunk but remain nonzero. - If $\beta_4$ is small but nonzero, age still has some influence, albeit limited.

### 27.2.4 When to Use Ridge Regression

- When all predictors are expected to contribute.
- When multicollinearity exists (reduces variance in correlated predictors).
- When feature selection is not needed but regularization is required.

—

## 27.3 Comparison: Lasso vs. Ridge Regression

| **Aspect** | **Lasso Regression** | **Ridge Regression** |
|---|---|---|
| **Penalty Term** | $L_1$-norm ($\sum \lvert\beta_j\rvert$) | $L_2$-norm ($\sum \beta_j^2$) |
| **Feature Selection** | Yes (sets some coefficients to 0) | No (shrinks coefficients but keeps all) |
| **Effect on Multicollinearity** | Selects one variable among correlated ones | Shrinks all correlated predictors |
| **When to Use** | When sparsity is expected | When all features contribute |

Table 13: Comparison of Lasso and Ridge Regression

## 27.4 Elastic Net: Combining Lasso and Ridge

Elastic Net combines the benefits of both Lasso and Ridge regression:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2. \tag{102}$$

**Benefits:**

- Performs feature selection (like Lasso) but retains correlated predictors (like Ridge).
- Useful when predictors exhibit high collinearity.

## 27.5 Conclusion

Both Lasso and Ridge regression address overfitting and improve model generalization:

- Use Lasso when you expect only a few important predictors (sparse models).
- Use Ridge when all predictors are expected to contribute.
- Use Elastic Net when you need both regularization and feature selection.

Choosing the right method depends on the dataset's structure, the presence of multicollinearity, and the need for feature selection.

# 28 Hierarchical Linear Models (HLM)

Hierarchical Linear Models (HLM), also known as Mixed Effects Models, account for grouped data structures where observations are nested within higher-level units. These models capture both within-group and between-group variation, correcting for the dependency among observations.

## 28.1 Why Use Hierarchical Models?

Standard regression models assume independent observations, but many real-world datasets involve hierarchical structures, such as:

- Students nested within schools.
- Employees nested within companies.
- Repeated measures on the same individuals over time.

Ignoring this structure can lead to biased standard errors and incorrect inferences. HLM accounts for:

- **Group-level effects**: Different groups may have distinct intercepts and slopes.
- **Within-group vs. between-group variation**: Separates variations at different levels.
- **Unobserved heterogeneity**: Controls for latent factors at the group level.

## 28.2 Fixed Effects Model (FE)

Fixed Effects (FE) models control for unobserved group-specific characteristics by including a separate intercept for each group. The assumption is that these characteristics are correlated with the independent variables.

**Model Equation:**

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \gamma_t + \epsilon_{it} \tag{103}$$

where:

- $\alpha_i$ represents the group-specific fixed effect (absorbing time-invariant differences).
- $\gamma_t$ represents time-specific fixed effects (controlling for shocks affecting all groups).
- $\beta_1$ captures the within-group effect of $X$ on $Y$.
- $\epsilon_{it}$ is the idiosyncratic error term.

**Interpretation:**

- Estimates how changes in $X$ affect $Y$ **within the same group over time**.
- Eliminates between-group variation, making cross-group comparisons impossible.
- Time-invariant variables (e.g., industry type) cannot be estimated because they are absorbed into $\alpha_i$.

## 28.3 Random Effects Model (RE)

Random Effects (RE) models assume that group-specific effects ($\alpha_i$) are randomly distributed and uncorrelated with the independent variables.

**Model Equation:**

$$Y_{it} = \gamma_0 + \beta_1 X_{it} + u_i + \gamma_t + \epsilon_{it} \tag{104}$$

where:

- $u_i \sim N(0, \sigma_u^2)$ represents unobserved group-level effects.

- $\gamma_t$ represents time-specific effects.
- $\beta_1$ captures both within-group and between-group variation.

**Interpretation:**

- Estimates both within-group and between-group effects.
- Assumes that $u_i$ is uncorrelated with $X_{it}$; if this assumption is violated, the model is biased.
- Allows estimation of time-invariant variables.

## 28.4 Fixed vs. Random Effects: The Hausman Test

The Hausman Test helps determine whether a Fixed Effects or Random Effects model is appropriate.

**Hypotheses:**

$$H_0 : \mathbb{E}[u_i|X_{it}] = 0 \quad \text{(Random Effects is valid)} \tag{105}$$

$$H_a : \mathbb{E}[u_i|X_{it}] \neq 0 \quad \text{(Fixed Effects is required)} \tag{106}$$

**Decision Rule:**

- If $p < 0.05$, reject $H_0$ → Use Fixed Effects.
- If $p > 0.05$, fail to reject $H_0$ → Random Effects can be used.

## 28.5 Diagnostics for Panel Data Models

### 28.5.1 Pesaran CD Test (Cross-Sectional Dependence)

Residuals across different groups may be correlated due to shared external factors. The Pesaran CD test detects this.

**Test Statistic:**

$$CD = \frac{1}{N(N-1)} \sum_i \sum_{j>i} \hat{\rho}_{ij}. \tag{107}$$

**Decision Rule:**

- If $CD$ is significantly different from 0, cross-sectional dependence exists.
- Solutions include spatial models or Driscoll-Kraay standard errors.

### 28.5.2 Breusch-Pagan LM Test (Random Effects vs. OLS)

Determines whether random effects are needed.

**Hypotheses:**

$$H_0 : \sigma_u^2 = 0 \quad \text{(No random effects)} \tag{108}$$

$$H_a : \sigma_u^2 > 0 \quad \text{(Use Random Effects)} \tag{109}$$

### 28.5.3 Intraclass Correlation Coefficient (ICC)

Measures how much of the total variance is explained by group-level differences:

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}. \tag{110}$$

**Interpretation:**

- If $ICC > 0.05$, substantial between-group variation → Use hierarchical models.
- If $ICC \approx 0$, little group-level variation → OLS may be sufficient.

| **Aspect** | **Fixed Effects (FE)** | **Random Effects (RE)** |
|---|---|---|
| Unobserved Group Effects | Controlled for, treated as fixed | Modeled as random |
| Estimation | Uses only within-group variation | Uses both within- and between-group variation |
| Time-Invariant Variables | Cannot be estimated | Can be estimated |
| Assumption on Group Effects | Correlated with $X_{it}$ | Uncorrelated with $X_{it}$ |
| Efficiency | Less efficient, loses degrees of freedom | More efficient if assumptions hold |

Table 14: Comparison of Fixed Effects and Random Effects Models

## 28.6 Comparison of Fixed Effects and Random Effects

## 28.7 Conclusion

- Use Fixed Effects when unobserved group characteristics are correlated with independent variables.
- Use Random Effects when group differences are assumed to be random and uncorrelated with the independent variables.
- Perform the Hausman Test to determine the correct model empirically.

Hierarchical models improve inference by properly accounting for structured data dependencies, ensuring robust and reliable statistical analysis.

# 29 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) extend traditional Linear Models (LMs) to allow for non-normal response variables. Unlike ordinary least squares (OLS) regression, which assumes normally distributed errors and a linear relationship between predictors and the dependent variable, GLMs enable modeling with various distributions and link functions.

## 29.1 Why Use GLMs? How Are They Different from Linear Models?

Traditional Linear Models (LMs) assume:

- The dependent variable $Y$ is continuous and normally distributed.
- A linear relationship between the predictors $X$ and the mean of $Y$.
- Homoscedasticity (constant variance of errors).
- Additive effects of predictors.

However, in many real-world scenarios, these assumptions do not hold. Examples include:

- Binary outcomes (e.g., disease presence: Yes/No).
- Count data (e.g., number of customer complaints).
- Skewed or bounded outcomes (e.g., proportions or survival times).

GLMs solve these issues by:

- Allowing the dependent variable $Y$ to follow different probability distributions (e.g., binomial, Poisson, gamma).
- Transforming the relationship between $X$ and $Y$ via a **link function**.
- Using Maximum Likelihood Estimation (MLE) instead of OLS to estimate coefficients.

## 29.2 General Structure of a GLM

A Generalized Linear Model consists of three components:

- **Random Component:** Specifies the distribution of the response variable $Y$.
- **Systematic Component:** A linear predictor $\eta$, which is a function of independent variables $X$.
- **Link Function:** Transforms the expected value of $Y$ to fit the linear predictor.

Mathematically, a GLM is represented as:

$$g(\mathbb{E}[Y]) = \eta = X\beta, \tag{111}$$

where:

- $g(\cdot)$ is the **link function**.
- $\mathbb{E}[Y]$ is the expected value (mean) of $Y$.
- $X\beta$ is the linear predictor.

| **Characteristic** | **Linear Models (LMs)** vs. **GLMs** |
|---|---|
| Distribution of $Y$ | LMs assume normality; GLMs allow various distributions |
| Relationship with $X$ | LMs assume a linear relationship; GLMs use link functions |
| Estimation Method | LMs use OLS; GLMs use Maximum Likelihood Estimation (MLE) |
| Error Structure | LMs assume constant variance (homoscedasticity); GLMs allow non-constant variance |

Table 15: Comparison of LMs and GLMs

## 29.3 Comparison: GLMs vs. Linear Models

## 29.4 When to Use GLMs

GLMs should be used when:

- The dependent variable is **binary**, **count-based**, or **skewed**.
- Variance increases with the mean (heteroscedasticity).
- The relationship between predictors and outcome is non-linear.
- Probabilities or rates need to be modeled (e.g., logistic regression for classification).

## 29.5 Conclusion

Generalized Linear Models extend traditional regression by accommodating different types of response variables. The choice of model depends on the nature of $Y$:

- Use **Linear Models** when $Y$ is continuous and normally distributed.
- Use **Logistic Regression** for binary outcomes.
- Use **Poisson Regression** for count data.
- Use **Gamma Regression** for positive, skewed data.

GLMs provide a flexible framework for modeling diverse data types, ensuring more accurate and reliable statistical analysis.

# 30 Logistic Regression

## 30.1 Introduction

Logistic regression is a statistical method for modeling binary outcomes ($Y \in \{0, 1\}$) based on predictor variables. It transforms the response variable into probabilities using the logistic (sigmoid) function.

## 30.2 Mathematical Formulation

Unlike linear regression, logistic regression models the probability that $Y = 1$ using the logit function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}} \tag{112}$$

Taking the log-odds transformation:

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p. \tag{113}$$

where:

- $P(Y = 1)$ is the probability of success.
- $\beta_0$ is the intercept.
- $\beta_1, \beta_2, ..., \beta_p$ are regression coefficients.
- $X_1, X_2, ..., X_p$ are predictor variables.

## 30.3 Interpretation of Coefficients

Each coefficient $\beta_j$ represents the log-odds change in $Y$ per unit change in $X_j$. The exponentiation of coefficients provides the odds ratio:

$$\text{Odds Ratio} = e^{\beta_j}. \tag{114}$$

**Example:** Suppose a logistic regression model predicts whether a customer buys a product ($Y = 1$) based on advertising spending ($X$):

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = -2 + 0.05X. \tag{115}$$

- $e^{0.05} \approx 1.051$ means that for each additional dollar spent, the odds of purchase increase by 5.1

## 30.4 Model Fitting: Maximum Likelihood Estimation

Unlike OLS in linear regression, logistic regression estimates coefficients using Maximum Likelihood Estimation (MLE) by maximizing the likelihood function:

$$L(\beta) = \prod_{i=1}^{n} P(Y_i|X_i)^{Y_i}(1 - P(Y_i|X_i))^{(1-Y_i)}. \tag{116}$$

The log-likelihood is:

$$\log L(\beta) = \sum_{i=1}^{n} Y_i \log P(Y_i|X_i) + (1 - Y_i)\log(1 - P(Y_i|X_i)). \tag{117}$$

MLE finds the coefficients $\beta$ that maximize $\log L(\beta)$.

## 30.5 Model Performance Metrics

Since logistic regression does not use $R^2$, alternative measures assess model fit:

### 30.5.1 Pseudo-$R^2$

- McFadden's $R^2$:

$$R^2_{\text{McF}} = 1 - \frac{\log L_{\text{model}}}{\log L_{\text{null}}}. \tag{118}$$

  - Higher values indicate better fit, but not directly comparable to traditional $R^2$.

- Likelihood Ratio Test (LRT):

$$\chi^2 = -2(\log L_{\text{null}} - \log L_{\text{model}}). \tag{119}$$

  - A significant $p$-value suggests at least one predictor is useful.

### 30.5.2 Classification Accuracy

- Confusion Matrix:

| | **Predicted 0** | **Predicted 1** |
|---|---|---|
| **Actual 0** | True Negative (TN) | False Positive (FP) |
| **Actual 1** | False Negative (FN) | True Positive (TP) |

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{120}$$

- Precision (Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{121}$$

  - Measures how many predicted positives are actual positives.

- Recall (Sensitivity, True Positive Rate):

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{122}$$

  - Measures how many actual positives were correctly identified.

- F1-Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{123}$$

  - Balances precision and recall.

## 30.6 Diagnostics and Assumptions

### 30.6.1 Multicollinearity (VIF)

Variance Inflation Factor (VIF) detects multicollinearity:

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{124}$$

- If $VIF > 5$, multicollinearity may be problematic.

### 30.6.2 Linearity in Log-Odds

The relationship between predictors and the log-odds should be linear. Box-Tidwell Test checks this assumption.

### 30.6.3 Hosmer-Lemeshow Test

Tests goodness-of-fit:

$$H_0 : \text{Model fits well.} \tag{125}$$

- A large $p$-value suggests a good fit.

## 30.7 ROC Curve and AUC

### 30.7.1 Receiver Operating Characteristic (ROC) Curve

The ROC curve plots:
- True Positive Rate (Recall) vs. False Positive Rate.

### 30.7.2 Area Under the Curve (AUC)

Measures overall model discrimination:

$$AUC = \int_0^1 \text{TPR}(x) dx. \tag{126}$$

**Interpretation:**

- $AUC = 0.5 \rightarrow$ No better than random guessing.
- $AUC > 0.7 \rightarrow$ Acceptable model.
- $AUC > 0.9 \rightarrow$ Excellent model.

## 30.8 Conclusion

Logistic regression is a powerful tool for binary classification. Model performance should be evaluated using:

- Goodness-of-fit: Pseudo-$R^2$, likelihood ratio test.
- Classification accuracy: Precision, recall, F1-score.
- Model diagnostics: VIF, Hosmer-Lemeshow test.
- Predictive ability: ROC-AUC.

Logistic regression serves as a foundation for advanced classification models such as regularized logistic regression (Lasso/Ridge), decision trees, and neural networks.