STATISTICS I: DATA EXPLORATION

Basic Statistical Concepts and Simple Linear Regression

Professor:

Dr. Kaushik Jana

Notes by:

Students of Class of 2029

Fall 2025

Indian Statistical Institute, Bangalore

Contents

1 Introduction to Statistics

Definition

Statistics is the science of collecting, classifying, presenting, and analyzing data.

1.1 The Statistical Process

1. **Collecting:** Choosing the question model, sample group, geometric location, etc.
2. **Classifying:** How the data is grouped.
3. **Presenting:** How data is visualized (types of graphs).
4. **Analysis:** Interpreting the data without bias, predetermined results, or hidden motives.

1.2 Descriptive vs. Inferential

- **Descriptive Statistics:** Techniques for interpreting the numbers resulting from the objects analyzed. The philosophy here is that **"The data speaks for itself."** It aims for an unbiased representation of the facts.
- **Inferential Statistics:** Making predictions or inferences about a larger population based on a smaller representative sample.

1.3 Basic Terms

- **Population:** The whole set of individuals or objects to be analyzed.
- **Sample:** A representative subset of the population.
- **Parameter:** A numerical characteristic of an entire population (e.g., Population Mean μ).
- **Statistic:** A numerical characteristic of a sample (e.g., Sample Mean $\bar{x}$).

2 Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. It is the "most typical" value around which most of your data lies.

2.1 Mean (Average)

The mean acts as the **"center of mass"** for all your data points. It is the balance point.

Important conceptual note: The mean is **not** necessarily the point where half the data is above and half is below. Rather, it is the point where the sum of positive deviations equals the sum of negative deviations.

$$\sum(x_i - \bar{x}) = 0$$

Mean Formula

$$\bar{x} = \frac{\sum x_i}{n} \tag{1}$$

Mean for Grouped Data (Step-Deviation Method)

When data is grouped into classes with width h and assumed mean A:

$$\bar{x} = A + h\left(\frac{\sum f_i u_i}{N}\right), \quad \text{where } u_i = \frac{x_i - A}{h} \tag{2}$$

Weakness: The mean is affected significantly by **outliers** (data values that are extremely low or high compared to the rest). When outliers exist, the average is "skewed" and may not represent the "typical" value well. It is best used when data is symmetrical.

2.2 Median

The median is the middle value when data is arranged in ascending or descending order.

Why use Median over Mean? It is used as an alternate measure of central tendency especially when the Mean is not a good representative. This happens when:

- The data has outliers.
- The distribution is skewed (not symmetric).

The Median is better in these cases because it depends only on the **position** of values, not their magnitude. It is robust against extreme values.

2.3 Mode

The most occurring value (the value with the highest frequency/probability).

Empirical Relationship

For moderately skewed distributions, there is an approximate relationship between Mean, Median, and Mode:

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean}) \tag{3}$$

3 Measures of Variability (Dispersion)

Measures of central tendency tell us where the center is, but they do **not** tell us how spread out the data is. Two different datasets can have the same mean but be completely different in how scattered the values are.

There are four main measures: Range, Variance, Standard Deviation, and Interquartile Range.

3.1 Mean Absolute Deviation (MAD)

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n} \tag{4}$$

What it does well: It tells you, on average, how far the data points lie from the Mean. It is the "average error" and is easy to interpret.

The Problem with MAD (Why we need Variance): While MAD is robust, the absolute value function $|x|$ has a sharp corner at zero. This **"breaks algebra"**—you cannot easily perform calculus (differentiation) on it. In advanced statistics, we need functions that are smooth and differentiable to find minimums and maximums easily.

3.2 Standard Deviation (SD) and Variance

To solve the algebraic problem of MAD, we **square** the deviations instead of taking the absolute value.

- **Why Square?** Squaring turns negative errors positive (like absolute value) but creates a smooth parabola that is easy to differentiate.
- **Weighting:** Squaring gives **more weight** to extreme values (outliers). A deviation of 2 becomes 4, but a deviation of 10 becomes 100.

Variance & Standard Deviation

Variance (S^2):

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{(Pop.)} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \text{(Sample)} \tag{5}$$

Standard Deviation (S):

$$\sigma = \sqrt{\text{Variance}} \tag{6}$$

3.3 Other Measures of Dispersion

Coefficient of Variation (CV)

A relative measure of dispersion used to compare consistency between two datasets.

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100 \tag{7}$$

Quartile Deviation (Semi-Interquartile Range)

Also known as the Semi-Interquartile Range (SIQR), this is a robust measure of spread.

$$\text{QD} = \text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{\text{IQR}}{2} \tag{8}$$

Interpretation: It represents the typical deviation of a value from the median.

Variance for Grouped Data (Step-Deviation)

Using $u_i = \frac{x_i - A}{h}$:

$$\text{Variance}(\sigma^2) = h^2 \left[\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N} \right)^2 \right] \tag{9}$$

3.4 Combined Mean and Variance

When we have two sets of data (Group 1 and Group 2) and we want to find the statistics for the combined group:

Combined Formulas

Let Group 1 have size n_1, mean $\bar{x}_1$, variance s_1^2.
Let Group 2 have size n_2, mean $\bar{x}_2$, variance s_2^2.

1. Combined Mean ($\bar{x}_c$):

$$\bar{x}_c = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \tag{10}$$

2. Combined Variance (s_c^2):

$$s_c^2 = \frac{n_1(s_1^2 + d_1^2) + n_2(s_2^2 + d_2^2)}{n_1 + n_2} \tag{11}$$

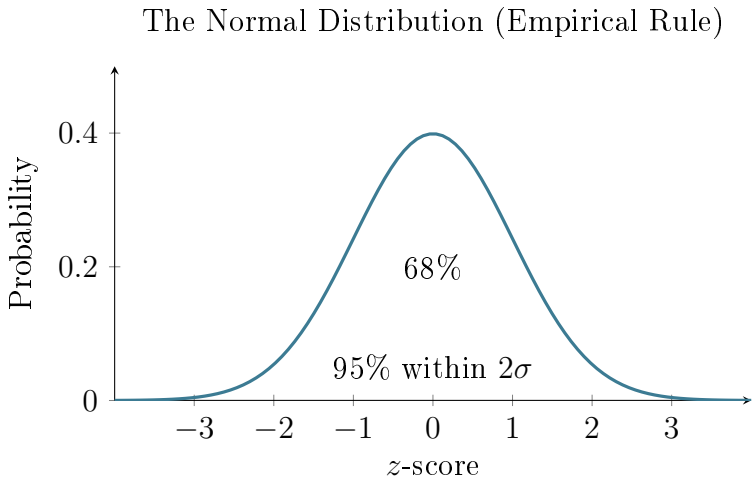
Where $d_1 = \bar{x}_1 - \bar{x}_c$ and $d_2 = \bar{x}_2 - \bar{x}_c$ are the deviations of the group means from the combined mean.

3.5 Distribution Models: Normal vs. General

The Z-Score

A Z-score tells us how many standard deviations a data point (x_i) is from the mean.

$$Z = \frac{x_i - \bar{x}}{\sigma} \tag{12}$$



Chebyshev vs. Empirical Rule

1. The Empirical Rule (Normal Only): Symmetric bell curve.

- $\approx 68\%$ of data within $\bar{x} \pm 1\sigma$
- $\approx 95\%$ of data within $\bar{x} \pm 2\sigma$

2. Chebyshev's Theorem (Any Distribution):
No matter how skewed or irregular the data is, the fraction of data within k standard deviations is at least $1 - \frac{1}{k^2}$.

- For $k = 2$: $1 - 1/4 = 75\%$ (Minimum guaranteed).

3.6 Effect of Linear Transformation

If a random variable X is transformed to Y using the linear equation $Y = aX + b$ (where a and b are constants):

Expectation and Variance Rules

1. Expectation (Mean): The mean is affected by both the scaling factor (a) and the shift (b).

$$E[Y] = aE[X] + b \quad \Longrightarrow \quad \bar{y} = a\bar{x} + b \tag{13}$$

2. Variance: Variance measures spread. Adding a constant (b) shifts the data but does not change the spread. Multiplying by a scales the spread by a^2.

$$Var(Y) = a^2 Var(X) \quad \Longrightarrow \quad \sigma_y^2 = a^2\sigma_x^2 \tag{14}$$

3. Standard Deviation: Standard deviation scales by the absolute value of a.

$$SD(Y) = |a|SD(X) \quad \Longrightarrow \quad \sigma_y = |a|\sigma_x \tag{15}$$

4 Moments, Skewness, and Kurtosis

4.1 Moments

Moments describe the shape of a distribution.

Central Moments

The k-th central moment is the expected value of the deviation from the mean raised to the power of k:

$$\mu_k = E[(X - \mu_x)^k] \tag{16}$$

- $\mu_1 = 0$ (Always)
- $\mu_2 = \sigma^2$ (Variance)
- μ_3 measures Skewness
- μ_4 measures Kurtosis

4.2 Skewness

Skewness measures the asymmetry of the probability distribution.

Measures of Skewness

1. Moment Coefficient of Skewness (γ_1):

$$\gamma_1 = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{\sigma^3} \tag{17}$$

2. Pearson's Coefficient of Skewness (Sk_p):

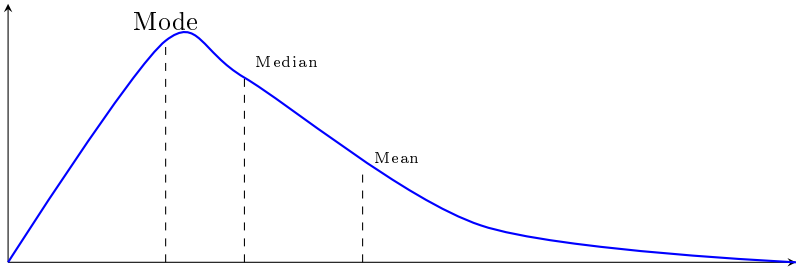
$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{SD}} = \frac{\bar{x} - \text{Mode}}{\sigma} \tag{18}$$

3. Bowley's Coefficient of Skewness (Sk_B): Based on Quartiles:

$$Sk_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \tag{19}$$

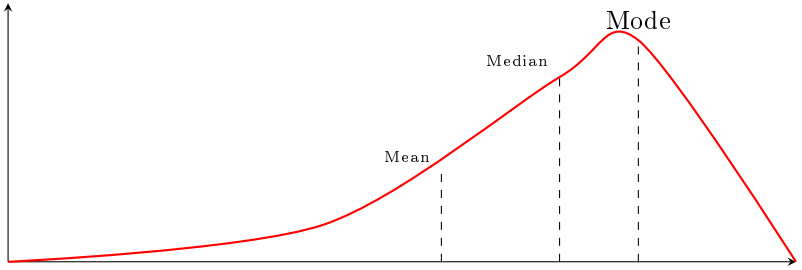
4.3 Visualizing Skewness

Positively Skewed (Right Tail)



Mode < Median < Mean

Negatively Skewed (Left Tail)



Mean < Median < Mode

4.4 Kurtosis

Kurtosis measures the "tailedness" or "peakedness" of the distribution.

Kurtosis (β_2)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \tag{20}$$

4.5 Quantiles

4.5.1 Calculation for Discrete Data (Ungrouped)

Before calculating quartiles, data must be ordered from smallest to largest.

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

To find the p-th percentile (e.g., $p = 0.25$ for Q_1), calculate the position index $k = (n+1)p$.

- If k is an integer, the value is $x_{(k)}$.
- If k is not an integer (e.g., 2.75), interpolate between $x_{(k)}$ and $x_{(k+1)}$. Hence the value is $x_m + (x_{m+1} - x_m)\{k\}$ where $\{k\}$ is the fractional part of k and m is the integer part of k.

4.5.2 Quantiles for Grouped Data

To calculate any Quantile (Q) for grouped data (where N is total frequency, f is frequency of the class, h is class width, and r is the rank/fraction):

General Quantile Formula

$$Q = L + \left(\frac{N \cdot r - (cf)_{prev}}{f} \right) \times h \tag{21}$$

Where:

- L: Lower limit of the quantile class.
- $(cf)_{prev}$: Cumulative frequency of the class *preceding* the quantile class.
- r: The fractional position (e.g., for Median $r = 0.5$, for Q_1, $r = 0.25$).

4.6 Box and Whisker Plot

A graphical representation of the distribution of data using the "Five-Number Summary":

1. Minimum Value (L)
2. First Quartile (Q_1)
3. Median (Q_2)
4. Third Quartile (Q_3)
5. Maximum Value (H)

It allows for visual identification of outliers and symmetry.

5 Correlation Analysis (Bivariate Data)

Bivariate Data

Data comprising two variables (e.g., X and Y) is called bivariate data. We often use a **Scatter Diagram** to visualize the relationship between the Independent Variable (X) and the Dependent Variable (Y).

5.1 Types of Correlation

- **Positive Correlation:** As X increases, Y tends to increase. (e.g., Hours worked vs. Money earned).
- **Negative Correlation:** As X increases, Y tends to decrease. (e.g., Money spent vs. Money saved).
- **No Correlation:** No discernible pattern between X and Y.

5.2 Correlation Coefficient (r)

The Pearson Product-Moment Correlation Coefficient measures the strength and direction of the linear relationship.

Pearson's Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{22}$$

Properties:

- $-1 \leq r \leq +1$
- $r = +1$: Perfect Positive Correlation.
- $r = -1$: Perfect Negative Correlation.
- $r = 0$: No Linear Correlation.

Effect of Change of Origin and Scale

Correlation is independent of the change of origin and scale. If two new variables U and V are defined as:

$$U = \frac{X - a}{b}, \quad V = \frac{Y - c}{d}$$

Then the correlation coefficient between U and V (r_{uv}) is related to r_{xy} by:

$$r_{uv} = \frac{bd}{|bd|} r_{xy}$$

This means:

- $|r_{uv}| = |r_{xy}|$ (Magnitude is unchanged).
- The sign changes only if b and d have opposite signs.

6 Simple Linear Regression (SLR)

6.1 The Model

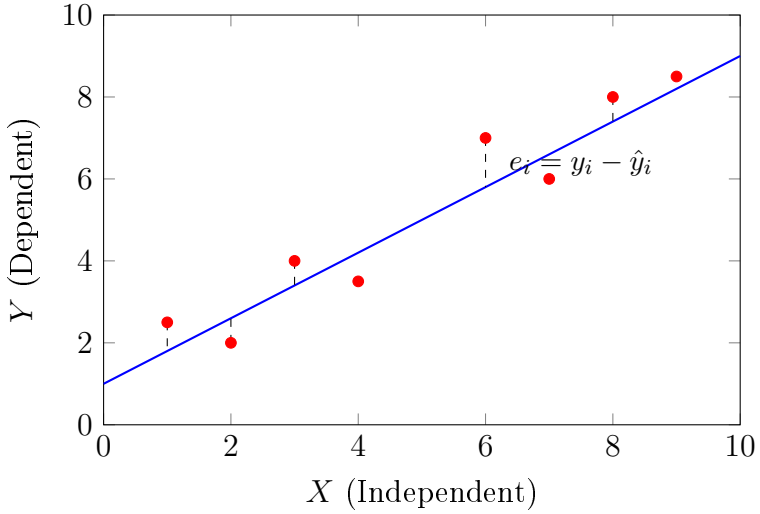
The goal of linear regression is to find a linear relationship between a set of data.

- **Population model:** $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- **Estimated line:** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Here, Y is the output variable (Dependent) and X is the input variable (Independent). ϵ is the unobservable error component.

6.2 Visualizing Residuals (The "Vertical" Distance)

In OLS (Ordinary Least Squares), we minimize the **vertical distance** between the observed point and the line. **Why Vertical?** In standard regression, we assume the input X is known or fixed (e.g., time, dosage), so there is no error in X. All the measurement error is assumed to be in the response variable Y. Therefore, we minimize the error in the Y direction.



Types of Regression

1. **Vertical Offsets (OLS):** Minimizes $\sum(y_i - \hat{y}_i)^2$. Assumes error is only in Y.
2. **Horizontal Offsets:** Minimizes $\sum(x_i - \hat{x}_i)^2$. Assumes error is only in X.
3. **Perpendicular Offsets:** Minimizes Euclidean distance. Assumes error in both X and Y.

7 Derivation of OLS Estimators

7.1 Mathematical Lemmas (Tools for Derivation)

Before deriving, recall these summation rules:

- $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$.
- $S_{xx} = \sum(x_i - \bar{x})^2 = \sum(x_i - \bar{x})x_i$ (because sum with constant $\bar{x}$ is 0).

7.2 Minimization Calculus

We minimize SSE: $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$.

Step-by-Step Minimization

Step 1: Partial Derivative w.r.t β_0

$$\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0 \\
&\implies -2\left(\sum y_i - n\beta_0 - \beta_1 \sum x_i\right) = 0 \\
&\implies n\beta_0 = \sum y_i - \beta_1 \sum x_i \\
&\implies \beta_0 = \bar{y} - \beta_1 \bar{x}
\end{aligned}$$

Step 2: Partial Derivative w.r.t β_1

$$\begin{aligned}
\frac{\partial S}{\partial \beta_1} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0 \\
&\implies \sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0
\end{aligned}$$

Step 3: Solve System Substitute $\beta_0 = \bar{y} - \beta_1 \bar{x}$ into Eq 2:

$$\begin{aligned}
\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i - \beta_1 \sum x_i^2 &= 0 \\
\sum x_i y_i - n\bar{x}\bar{y} &= \beta_1 \left(\sum x_i^2 - n\bar{x}^2\right) \\
\hat{\beta}_1 &= \frac{\sum x_i y_i - \frac{1}{n}\sum y_i \sum x_i}{\sum x_i^2 - \frac{1}{n}\sum x_i^2}
\end{aligned}$$

Using $S_{xx} = \sum(x_i - \bar{x})^2$ and $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

8 Properties of OLS Estimators

8.1 Setup: The Linear Constant k_i

To simplify derivations for Expectation and Variance, we define a constant k_i dependent only on x values:

$$k_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}} \tag{23}$$

Properties of k_i

1. $\sum k_i = \frac{\sum(x_i - \bar{x})}{S_{xx}} = 0$
2. $\sum k_i x_i = \frac{\sum(x_i - \bar{x})x_i}{S_{xx}} = \frac{S_{xx}}{S_{xx}} = 1$
3. $\sum k_i^2 = \sum \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 = \frac{1}{S_{xx}^2} \sum(x_i - \bar{x})^2 = \frac{1}{S_{xx}}$

Using this, the slope estimator becomes a linear combination of Y_i:

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i$$

8.2 Assumptions (Gauss-Markov)

1. $E[\epsilon_i] = 0$ (Zero conditional mean).
2. $Var(\epsilon_i) = \sigma^2$ (Homoscedasticity).
3. $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ (No autocorrelation).

8.3 Unbiasedness (Expectations)

Expectation of Slope $\hat{\beta}_1$

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum k_i Y_i\right] = \sum k_i E[Y_i] \\ &= \sum k_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \underbrace{\sum k_i}_{0} + \beta_1 \underbrace{\sum k_i x_i}_{1} \\ &= \beta_1 \end{aligned}$$

Expectation of Intercept $\hat{\beta}_0$

Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$:

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{y}] - \bar{x}E[\hat{\beta}_1] \\ &= (\beta_0 + \beta_1\bar{x}) - \bar{x}(\beta_1) = \beta_0 \end{aligned}$$

8.4 Derivation of Variances

Variance of Slope $\hat{\beta}_1$

Using $\hat{\beta}_1 = \sum k_i Y_i$:

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\sum k_i Y_i\right) \\ &= \sum k_i^2 Var(Y_i) \quad \text{(Since } Y_i \text{ are independent)} \\ &= \sigma^2 \sum k_i^2 \\ &= \sigma^2 \left(\frac{1}{S_{xx}}\right) \quad \text{(From Property 3)} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Variance of Intercept $\hat{\beta}_0$

Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$:

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}\underbrace{Cov(\bar{y}, \hat{\beta}_1)}_{0} \\ &= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \frac{\sigma^2 \sum x_i^2}{nS_{xx}} \end{aligned}$$

Side Note: Derivation of Covariance of $\bar{y}$ and $\hat{\beta}_1$

To justify the step above, recall that $\hat{\beta}_1 = \sum k_i Y_i$ and $\bar{y} = \frac{1}{n}\sum Y_i$.

$$\begin{aligned}
Cov(\bar{y}, \hat{\beta}_1) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n} k_j Y_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} k_i Var(Y_i) \quad \text{(Cross-terms vanish due to independence)} \\
&= \frac{1}{n}\sum_{i=1}^{n} k_i \sigma^2 \\
&= \frac{\sigma^2}{n}\underbrace{\sum_{i=1}^{n} k_i}_{0} \quad \text{(Since } \sum(x_i - \bar{x}) = 0\text{)} \\
&= 0
\end{aligned}$$

8.5 Covariance of Estimates

Covariance

$$\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) \\
&= \underbrace{Cov(\bar{y}, \hat{\beta}_1)}_{0} - \bar{x}Cov(\hat{\beta}_1, \hat{\beta}_1) \\
&= -\bar{x}Var(\hat{\beta}_1) \\
&= \frac{-\bar{x}\sigma^2}{S_{xx}}
\end{aligned}$$

8.6 Relationship between Regression and Correlation

Relation between Regression coefficients of Y on X and X on Y

Let b_{yx} be the slope of the regression of Y on X, and b_{xy} be the slope of X on Y.

$$b_{yx} = \frac{Cov(X,Y)}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$$
$$b_{xy} = \frac{Cov(X,Y)}{\sigma_y^2} = r\frac{\sigma_x}{\sigma_y}$$

Multiplying the two coefficients:

$$b_{yx} \cdot b_{xy} = \left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = r^2$$

$$\implies r^2 = b_{yx}b_{xy} \implies |r| = \sqrt{b_{yx}b_{xy}}$$

Determining the Sign: Since standard deviations σ_x and σ_y are always positive, the sign of the regression coefficients depends entirely on r.

- If $r > 0$, both b_{yx} and b_{xy} are positive.
- If $r < 0$, both b_{yx} and b_{xy} are negative.

Thus, r must have the same sign as the regression coefficients:

$$r = \text{sign}(b_{yx})\sqrt{b_{yx}b_{xy}}$$

8.7 Unbiased Estimator of Error Variance (σ^2)

We define the estimator for the error variance as the Mean Square Error (MSE):

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} \tag{24}$$

Proof

We know the ANOVA identity: $SST = SSR + SSE$, so $SSE = SST - SSR$.

1. **Expectation of Total Sum of Squares (SST):**
   $SST = \sum(Y_i - \bar{Y})^2$. Using the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$:

   $$Y_i - \bar{Y} = \beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})$$

   Squaring and summing (cross-terms vanish in expectation):

   $$E[SST] = \beta_1^2 S_{xx} + E\left[\sum(\epsilon_i - \bar{\epsilon})^2\right]$$

   Since ϵ_i are i.i.d with variance σ^2, $\sum(\epsilon_i - \bar{\epsilon})^2 \sim \sigma^2\chi^2_{n-1}$.

   $$E[SST] = \beta_1^2 S_{xx} + (n-1)\sigma^2$$

2. **Expectation of Regression Sum of Squares (SSR):**
   $SSR = \hat{\beta}_1^2 S_{xx}$.

   $$E[SSR] = S_{xx}E[\hat{\beta}_1^2] = S_{xx}\left(Var(\hat{\beta}_1) + (E[\hat{\beta}_1])^2\right)$$

   Substituting $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ and $E[\hat{\beta}_1] = \beta_1$:

   $$E[SSR] = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx}$$

3. **Expectation of Error Sum of Squares (SSE):**

   $$E[SSE] = E[SST] - E[SSR]$$
   $$E[SSE] = \left[(n-1)\sigma^2 + \beta_1^2 S_{xx}\right] - \left[\sigma^2 + \beta_1^2 S_{xx}\right]$$
   $$E[SSE] = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

Therefore:

$$E[\hat{\sigma}^2] = E\left[\frac{SSE}{n-2}\right] = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

This proves that $\hat{\sigma}^2$ is an unbiased estimator.

Intuition: Why do we divide by $n-2$?

You might notice that for the standard Sample Variance s^2, we divide by $n-1$. In Regression, we divide by $n-2$. Why?

1. Residuals vs. True Errors: We cannot observe the true random errors ϵ_i. We only see the residuals $e_i = Y_i - \hat{Y}_i$. Because the regression line is calculated specifically to *minimize* these residuals for this specific sample, the e_i's are slightly smaller than the true ϵ_i's.

2. Degrees of Freedom (The Cost of Estimation): To calculate the residuals, we first had to estimate two parameters: the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$.

- Every parameter we estimate "costs" us one degree of freedom.
- We started with n independent data points.
- We used 2 degrees of freedom to fix the line.
- We are left with $n-2$ independent pieces of information to estimate the variance.

If we divided by n, we would underestimate the true variance. Dividing by $n-2$ corrects for this bias.

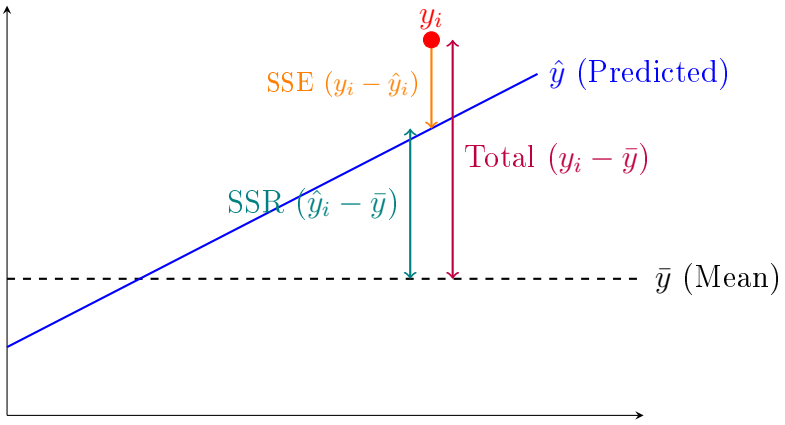
9 Decomposition of Variability (ANOVA)

ANOVA Identity

SST (Total Variation) = **SSR** (Explained by Model) + **SSE** (Unexplained Error)

9.1 Decomposition of Deviations

Visualizing the components for a single data point (x_i, y_i):



Derivation of ANOVA Decomposition

$$\begin{aligned}\sum(y_i - \bar{y})^2 &= \sum[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 + 2\sum \underbrace{(y_i - \hat{y}_i)}_{e_i}(\hat{y}_i - \bar{y})\end{aligned}$$

The cross-term $2\sum e_i(\hat{y}_i - \bar{y}) = \sum e_i\beta_1(x_i - \bar{x})$ vanishes because $\sum e_i = 0$ and

$\sum e_i x_i = 0$ (from Normal Equations).

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

$$\therefore SST = SSE + SSR$$

10 Goodness of Fit (R^2)

10.1 Coefficient of Determination (R^2)

Formula

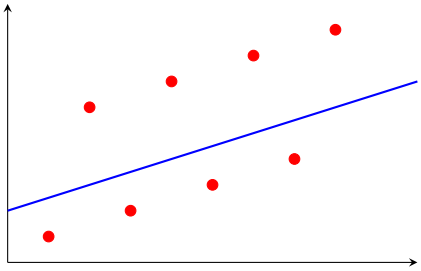
R^2 represents the proportion of variation in Y that is explained by the variation in X.

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} \leq 1 \tag{25}$$

Range: $0 \leq R^2 \leq 1$.

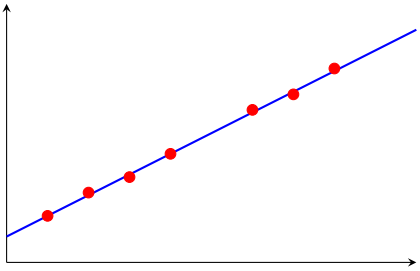
10.2 Visualizing R^2 and SSE

Low R^2 (High SSE)



High variation in errors

High R^2 (Low SSE)



Low variation in errors

10.3 Degrees of Freedom (df)

Degrees of Freedom

The number of independent values that can vary in an analysis without breaking any constraints.

$$df = n - k - 1 \tag{26}$$

Where:

- n: Number of observations.
- k: Number of explanatory variables (predictors).
- -1: Represents the intercept estimation.

For Simple Linear Regression ($k = 1$), $df = n - 2$.

11 Model Assumptions and Diagnostics

For the Simple Linear Regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we rely on specific assumptions about the error terms (ϵ) to ensure our estimates and predictions are valid.

11.1 Assumption 1: Normality of Residuals

The Normality Assumption

We assume the random errors follow a Normal Distribution centered at zero:

$$\epsilon_i \sim N(0, \sigma^2) \tag{27}$$

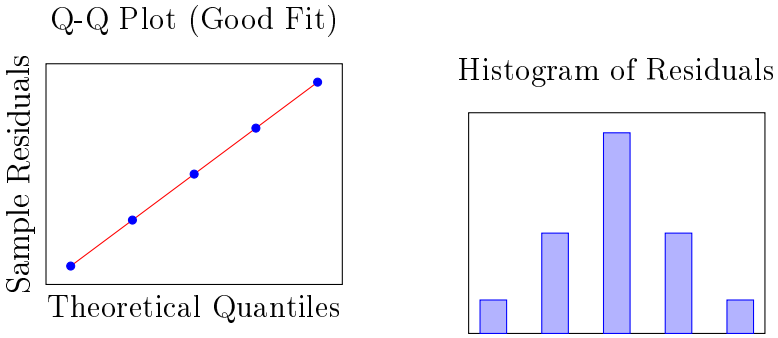
11.1.1 Why is this important?

This assumption is strictly required for **Statistical Inference**, not for calculating the line itself.

- **Estimation:** You can calculate the OLS line (slope and intercept) regardless of the distribution of errors.
- **Inference:** To calculate **Confidence Intervals** or **p-values** (e.g., testing $H_0 : \beta_1 = 0$), we use the t-distribution. The t-distribution derives from the assumption that errors are Normal. If errors are non-normal, p-values may be invalid (unless n is large, per CLT).

11.1.2 Graphical Diagnostics

- **Q-Q Plot (Quantile-Quantile):** Plots "Theoretical Normal Quantiles" vs. "Sample Residuals".
- **Histogram:** Checks for a symmetric bell curve.



11.1.3 Remedial Measures

If normality fails, try **transforming the response variable** (e.g., $\log(Y)$ or $\sqrt{Y}$). This often "squeezes" skewed data into a normal shape.

11.2 Assumption 2: Homoscedasticity (Constant Variance)

Definition

Homoscedasticity (Homo = same, scedastic = scatter) means the variance of the errors is constant for all levels of X.

$$Var(\epsilon_i|X_i) = \sigma^2 \quad \text{(Constant)} \tag{28}$$

If the variance changes (e.g., variance increases as X increases), it is called **Heteroscedasticity**.

11.2.1 Consequences of Heteroscedasticity

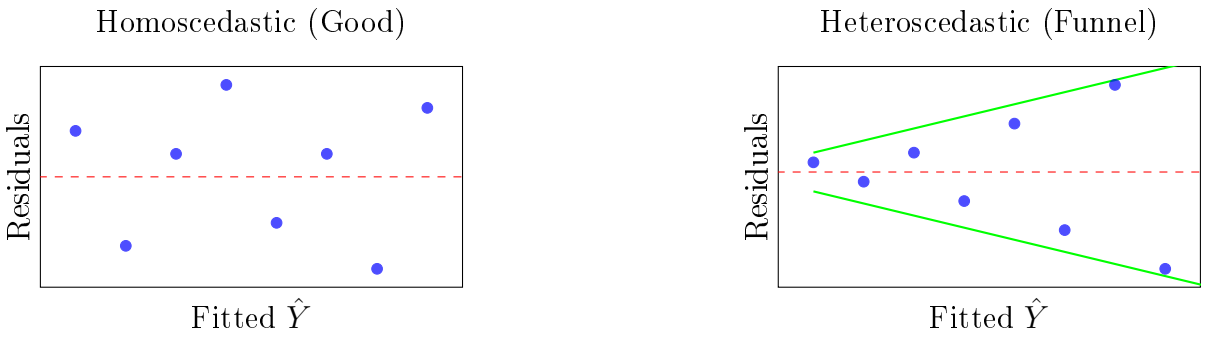
If the variance is **not** constant:

1. **Unbiasedness holds:** OLS estimators $\hat{\beta}$ remain unbiased (correct on average).
2. **Efficiency fails:** OLS is no longer the Best Linear Unbiased Estimator (BLUE). Weighted Least Squares (WLS) would be more precise.
3. **Inference fails:** Standard error formulas $SE(\hat{\beta})$ become incorrect, making t-tests and confidence intervals invalid.

11.2.2 Graphical Check: Residuals vs. Fitted Plot

We plot the Residuals (e_i) against the Predicted Values ($\hat{Y}_i$).

- **Homoscedastic (Ideal):** Random rectangular cloud.
- **Heteroscedastic (Problematic):** Funnel or Megaphone shape.



12 Sampling Theory

In statistical analysis, we often cannot observe every individual in a group. We must rely on sampling. In this context, we strictly distinguish between the **Population** (Capital Letters) and the **Sample** (Lowercase Letters).

12.1 Notation and Definitions

Population vs. Sample Parameters

1. **The Population (Fixed Constants):** Size N. Units $U = \{Y_1, Y_2, \ldots, Y_N\}$.
   - **Population Mean ($\overline{Y}$):** $\frac{1}{N}\sum_{i=1}^{N} Y_i$
   - **Population Variance (S^2):** In sampling theory, we often use $N-1$ as the denominator for the population variance to simplify later derivations:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

2. **The Sample (Random Variables):** Size n. Units $s = \{y_1, y_2, \ldots, y_n\}$.
   - **Sample Mean ($\bar{y}$):** $\frac{1}{n}\sum_{i=1}^{n} y_i$ (Estimator for $\overline{Y}$)
   - **Sample Variance (s^2):** $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ (Estimator for S^2)

12.2 Sampling Schemes

- **SRSWOR (Simple Random Sampling Without Replacement):** A method where n units are selected from N such that every distinct sample of size n has an equal probability of selection.

$$P(s) = \frac{1}{\binom{N}{n}}$$

  Key feature: No unit can appear twice. This provides more information than measuring the same unit multiple times.

- **SRSWR (Simple Random Sampling With Replacement):** Units are drawn one by one and returned to the population. A unit can appear multiple times.

$$P(s) = \frac{1}{N^n}$$

12.3 Properties of the Sample Mean ($\bar{y}$)

We use the sample mean $\bar{y}$ to estimate the population mean $\overline{Y}$. A good estimator should be **Unbiased** and have minimal **Variance**.

Proof: Sample Mean is Unbiased Estimator of Population Mean

Let a_i be an indicator variable for the i-th unit in the population:

$$a_i = \begin{cases} 1 & \text{if unit } Y_i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

In SRSWOR, the probability of any unit being selected is $P(a_i = 1) = \frac{n}{N}$. Therefore, $E[a_i] = \frac{n}{N}$.
We can rewrite the sample mean as a sum over the *population*:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i Y_i$$

Taking the expectation:

$$E[\bar{y}] = \frac{1}{n}\sum_{i=1}^{N} Y_i E[a_i] = \frac{1}{n}\sum_{i=1}^{N} Y_i \left(\frac{n}{N}\right)$$

$$E[\bar{y}] = \frac{1}{N}\sum_{i=1}^{N} Y_i = \overline{Y}$$

Variance of Sample Mean & FPC

The variance of the sample mean depends on the sample size and the population variance.

$$Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = (1-f)\frac{S^2}{n} \tag{29}$$

The Finite Population Correction (FPC): The term $(1-f)$ where $f = \frac{n}{N}$ is the sampling fraction.

- As $n \to N$, the variance approaches 0 (a census has no sampling error).
- If N is very large ($N \to \infty$) or $f < 0.05$, the FPC ≈ 1, and we use the standard formula S^2/n.

Estimator for Population Total

To estimate the Population Total ($Y_{total} = N\overline{Y}$), we use the estimator:

$$\hat{Y}_{total} = N \times \bar{y} \tag{30}$$

Variance of Total Estimator:

$$Var(\hat{Y}_{total}) = N^2 Var(\bar{y}) = N^2(1-f)\frac{S^2}{n} \tag{31}$$

12.4 Types of Survey Errors

- **Sampling Error:** Occurs because we only observe a part of the population. This is natural and can be reduced by increasing sample size n.
- **Non-Sampling Error:** Occurs due to human mistakes, faulty methods, or defects in data collection (e.g., Response bias, Non-response error, Measurement errors). These **cannot** be fixed simply by increasing n.

13 Categorical Data Analysis

When analyzing two categorical variables (e.g., "Yes/No", "Treatment/Control"), we cannot use means or variances. Instead, we analyze frequencies using **Contingency Tables**.

13.1 The 2×2 Contingency Table

Standard Notation

		Outcome (Y)		
		Success (Yes)	Failure (No)	Total
Group (X)	Group A	a	b	$a+b$
	Group B	c	d	$c+d$
	Total	$a+c$	$b+d$	n

13.2 Probability vs. Odds

A common mistake is confusing Probability with Odds. They measure the likelihood of an event differently.

Definitions

1. **Probability (P):** The proportion of successes out of the *Total*.

$$P(\text{Success}|\text{Group A}) = \frac{a}{a+b} \tag{32}$$

2. **Odds (O):** The ratio of Successes to Failures.

$$\text{Odds}(\text{Group A}) = \frac{a}{b} = \frac{P}{1-P} \tag{33}$$

13.3 The Odds Ratio (OR)

The Odds Ratio measures the strength of association between two categorical variables. It compares the odds of an outcome occurring in one group versus another.

Odds Ratio Formula

$$OR = \frac{\text{Odds}(\text{Group A})}{\text{Odds}(\text{Group B})} = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{34}$$

Interpretation of OR

- **OR = 1**: **Independence**. The exposure does not affect the outcome.
- **OR > 1**: **Positive Association**. Group A has higher odds of the outcome than Group B.
- **OR < 1**: **Negative Association**. Group A has lower odds of the outcome than Group B.

13.4 Testing for Independence (Expected Counts)

To determine if a relationship is statistically significant (Chi-Square Test), we first calculate what the counts *should* look like if there were absolutely no relationship between the variables.

Expected Count Formula

For any cell in the table (row i, column j):

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}} \tag{35}$$

Chi-Square Test Statistic

To measure the discrepancy between Observed (O_{ij}) and Expected (E_{ij}) counts:

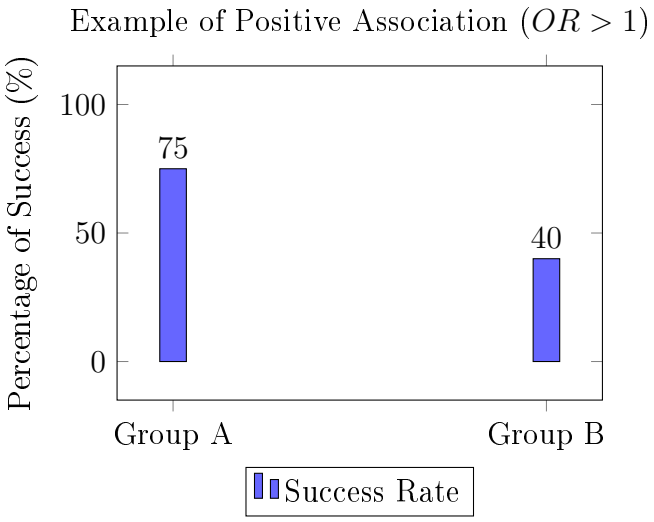
$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{36}$$

The degrees of freedom for an $r \times c$ table are given by:

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

13.5 Visualizing Association

If there is an association, the proportions (probabilities) of the outcome will differ between groups.



14 The Empirical Cumulative Distribution Function (ECDF)

Order Statistics

Before calculating ECDFs, data must be ordered from smallest to largest. If we have a sample $x_1, x_2, \ldots, x_n$, the **order statistics** are denoted as:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

where $x_{(1)}$ is the minimum and $x_{(n)}$ is the maximum.

14.1 Definition of ECDF

The ECDF, denoted as $F_n(t)$, is a step function that jumps up by $1/n$ at each of the n data points. It represents the proportion of observations that are less than or equal to t.

ECDF Formula

$$F_n(t) = \frac{\text{number of elements in sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t) \tag{37}$$

Key Properties

- It is 0 for values smaller than the minimum ($t < x_{(1)}$).
- It is 1 for values larger than or equal to the maximum ($t \geq x_{(n)}$).
- It is a **right-continuous** non-decreasing step function.

14.2 Visualizing ECDF

The graph below shows the step function for a sample. Note how the steps occur exactly at the data points, and the height of the step corresponds to the frequency.