Introduction to Statistical Methods in Political Science

Lecture 3: Summarizing Data II - Graphs

Ignacio Urbina

Shape of a Distribution

Distribution of a Variable

A variable's distribution is a description (function) that shows its possible values and the frequency with which they occur.
By examining a variable's distribution, we learn how likely is a given value relative to others.

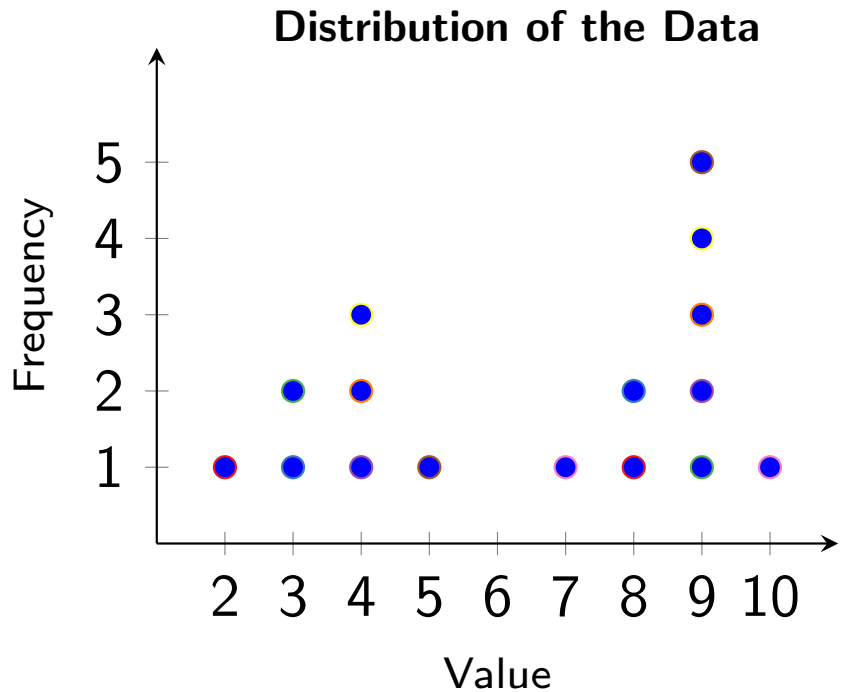
Dot Plot Example with Discrete Numerical Data

Sample:

$$\{9,\ 10,\ 4,\ 3,\ 8,\ 9,\ 3,\ 9,\ 5,\ 2,\ 4,\ 4,\ 8,\ 9,\ 9,\ 7\}$$

Frequency Table:

Value	Frequency
2	1
3	2
4	3
5	1
6	0
7	1
8	2
9	5
10	1



Identifying Distribution Shapes

Key Distribution Shapes:

- **Symmetric:** Equal spread on both sides.
- **Right Skewed:** Tail on the right (positive skewness).
- **Left Skewed:** Tail on the left (negative skewness).

Key Modalities:

- **Unimodal:** Single peak.
- **Bimodal:** Two distinct peaks.
- **Multimodal:** More than two peaks.
- **Uniform:** Flat, no peaks.

Symmetry and Skewness

Symmetric Distributions:

- The distribution looks the same on both sides of the mean.

Skewed Distributions:

- **Right skewed:** Long tail on the right.
- **Left skewed:** Long tail on the left.

The direction of skewness indicates where most of the data points lie relative to the tail.

Skewness and the Mean-Median Relationship

- In a **right-skewed distribution**, the mean is typically greater than the median due to the influence of the long tail.
- In a **left-skewed distribution**, the mean is typically less than the median for similar reasons.
- In a **symmetric distribution**, the mean and median are roughly equal.

Summary:

- The relative position of the mean and median can indicate the skewness of a distribution.

Unimodal, Bimodal, Multimodal, and Uniform Distributions

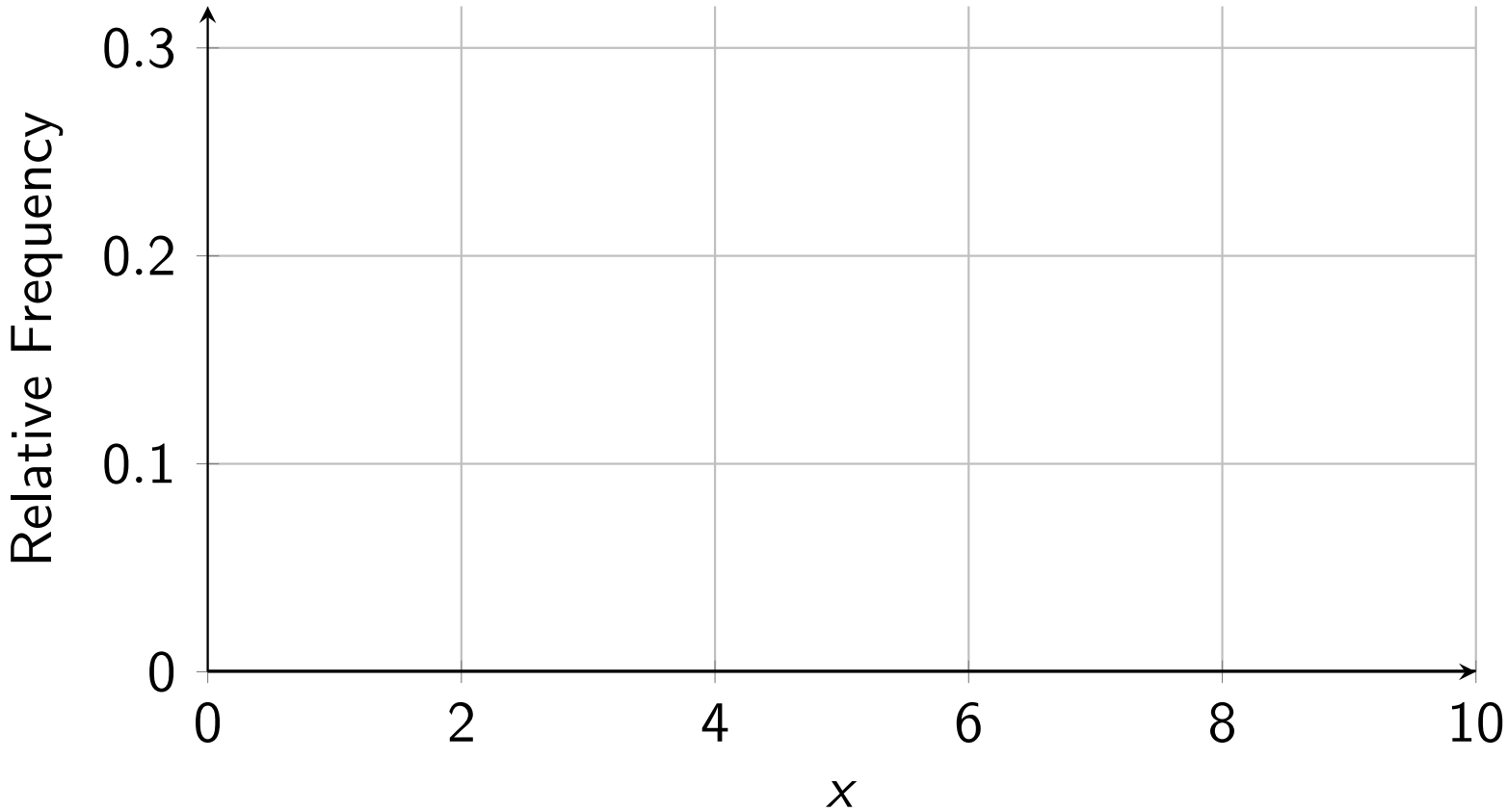
- **Unimodal:** A distribution with one clear peak or mode.
- **Bimodal:** A distribution with two distinct peaks.
- **Multimodal:** More than two peaks, indicating multiple clusters or groups.
- **Uniform:** No peaks; all values have roughly the same frequency.

These characteristics help describe the overall shape of the data and can indicate the presence of subpopulations.

Right-Skewed Distribution Example

Right-Skewed Distribution:

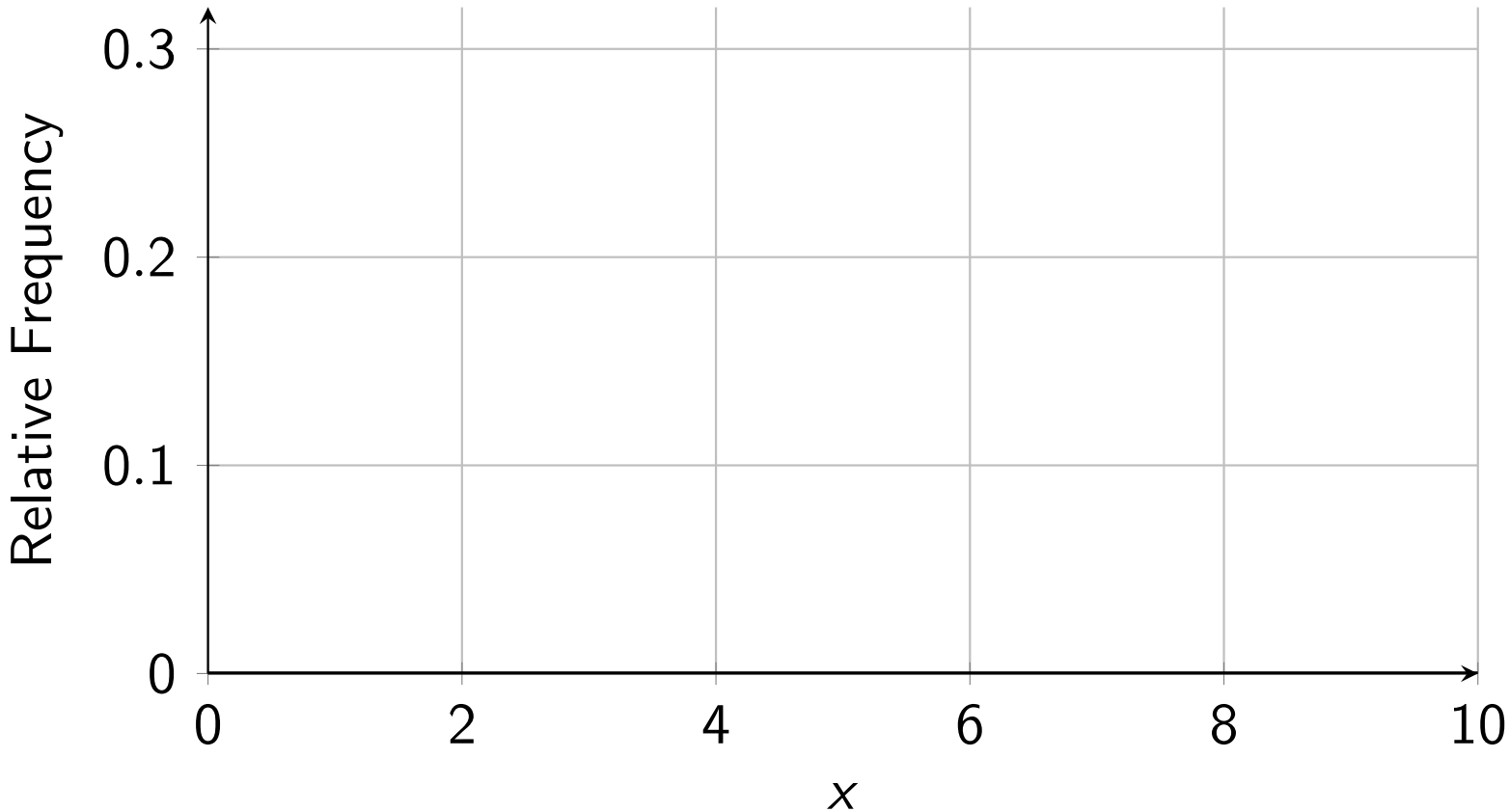
- The tail extends to the right, meaning more data is concentrated on the left.



Left-Skewed Distribution Example

Left-Skewed Distribution:

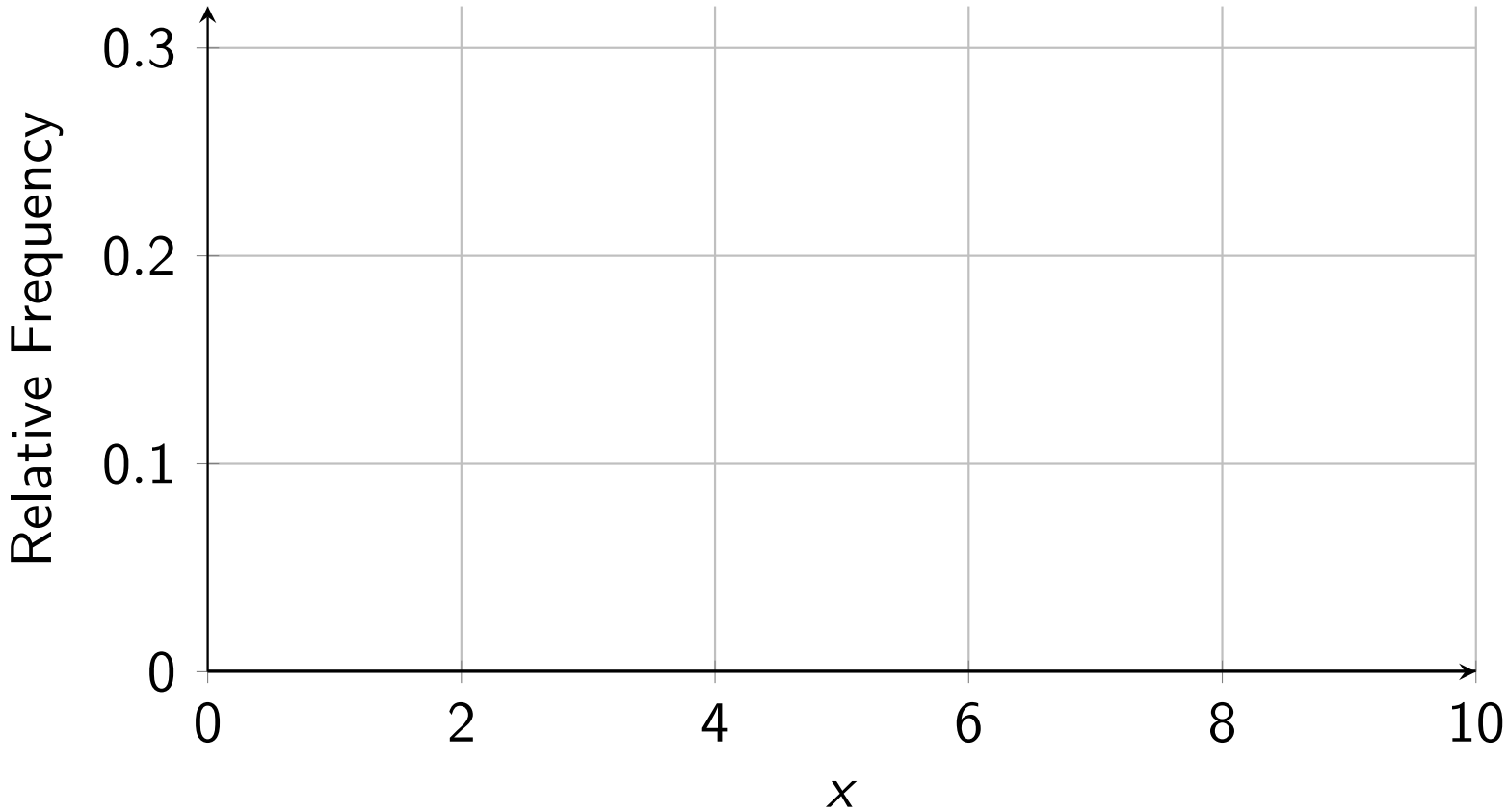
- The tail extends to the left, meaning more data is concentrated on the right.



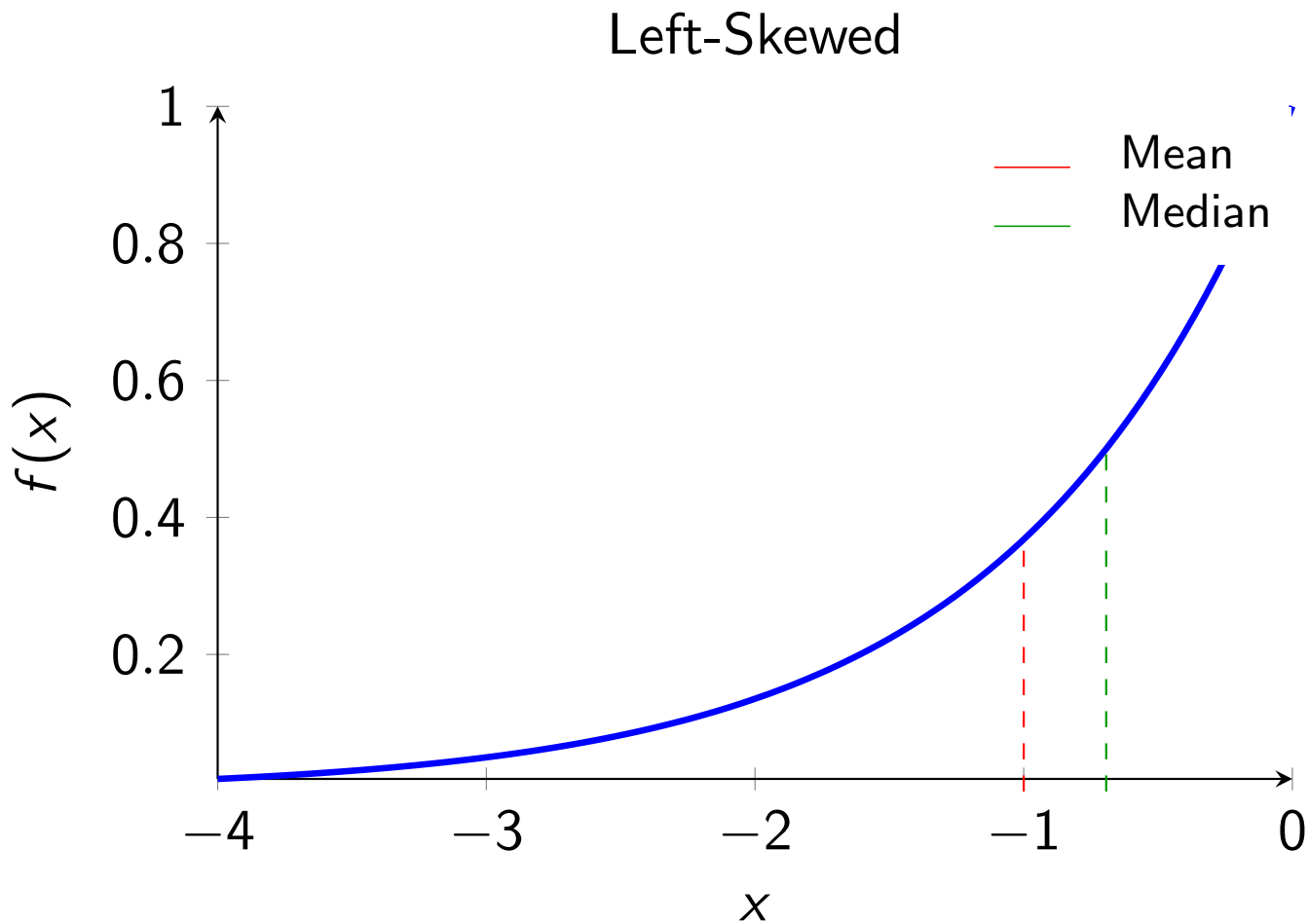
Symmetric Distribution Example

Symmetric Distribution:

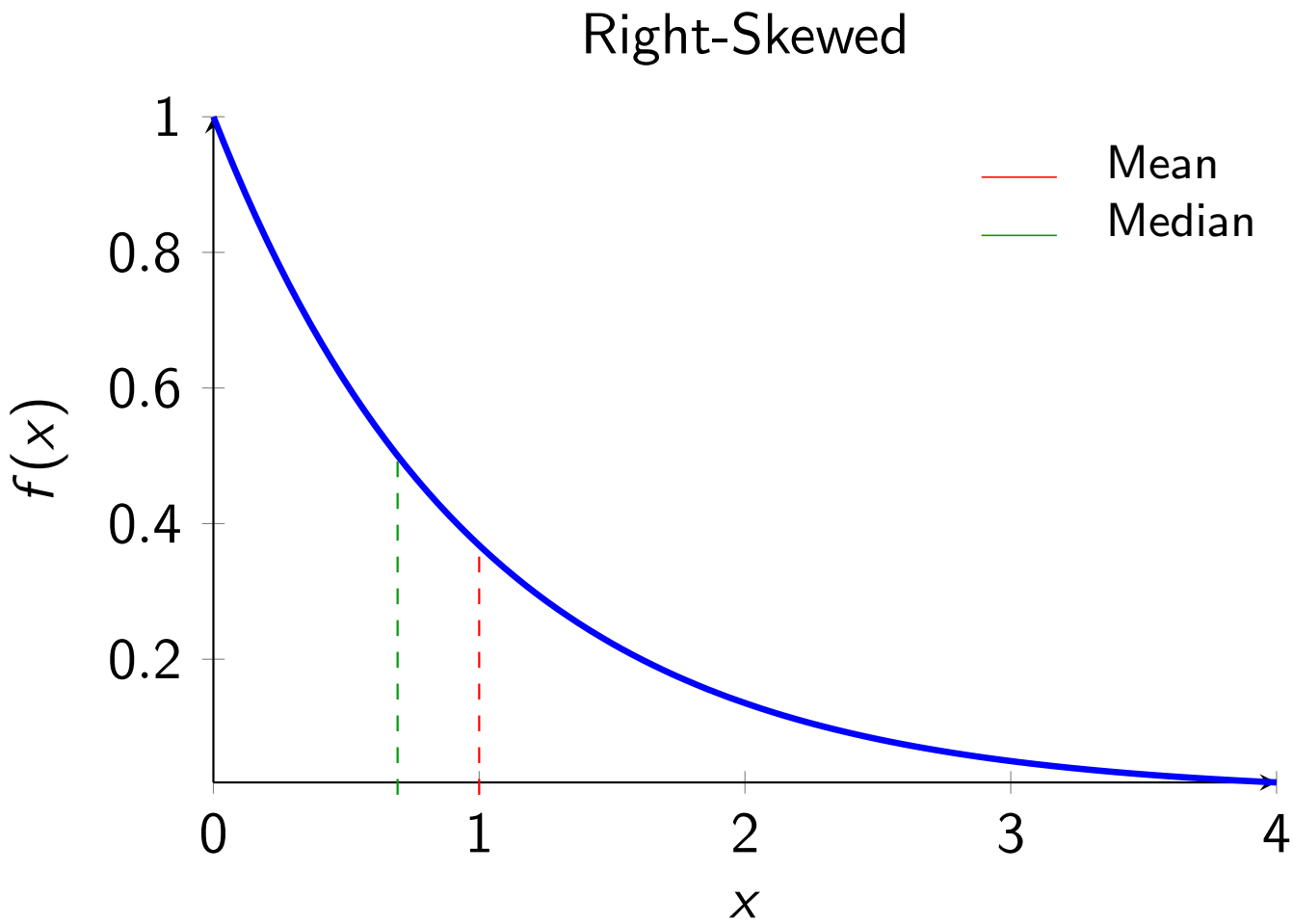
- A symmetric distribution has equal spread on both sides of the mean.



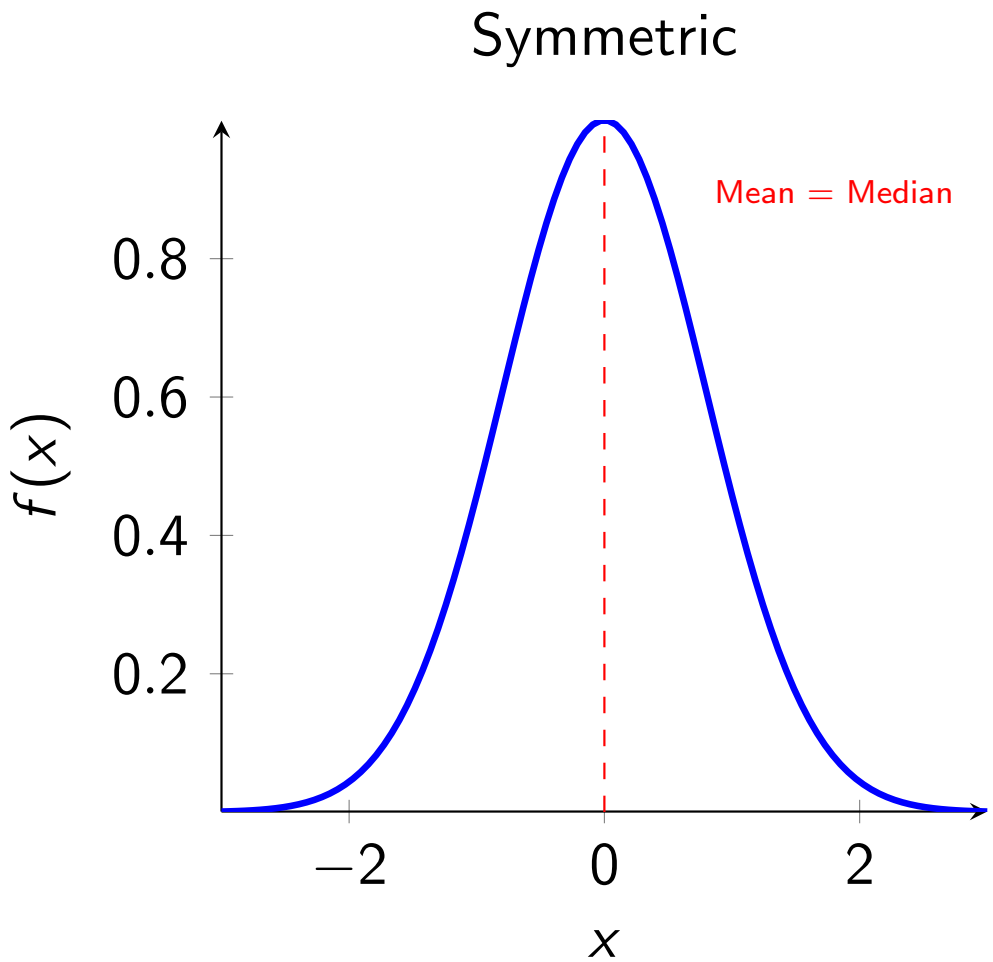
Left-Skewed Distribution



Right-Skewed Distribution



Symmetric Distribution



A symmetric distribution is a distribution in which the shape on the left and right sides of the center are mirror images of each other, meaning that the frequencies change at the same rate and direction as one moves away from the center in both directions.

Histograms

Introduction to Histograms

What is a Histogram?

- A histogram is a graphical representation of the distribution of numerical data.
- It uses bins (or intervals) to group data points and shows the frequency of data points in each bin.
- The height of each bar represents the count or frequency of data within that bin.

Step 1: Understanding Absolute Frequency

Absolute Frequency:

- The absolute frequency of an interval is the number of data points that fall within that interval.
- It's simply a count of occurrences of data points in each interval.

Example:

- Consider the data: $\{3, 5, 7, 9, 11, 5, 7, 9, 3, 5\}$
- The absolute frequency of the interval $[3, 6]$ is 5 (as four data points fall in this interval: 3, 3, 5, 5, 5).

Step 2: Choosing Bins (Intervals)

What are Bins?

- Bins (or intervals) divide the entire range of values into equal-sized chunks.
- Each bin contains a specific range of values, and data points are placed into the bin they fall within.

General Definition:

- *First* bin: $[\text{min}, \text{min} + \text{bin\_size})$
- Bin k: $[\text{min} + (k-1) \cdot \text{bin\_size}, \text{min} + (k) \cdot \text{bin\_size})$
- *Last* bin: $[\text{max} - \text{bin\_size}, \text{max}]$
  - It's normal to close the last bin on the right to capture the maximum.

Step 3: Assigning Data to Bins

Assigning Data to Bins:

- Once bins are defined, we assign each data point to the appropriate bin.
- This process results in an absolute frequency count for each bin.

Example:

- Suppose a discrete variable that goes from 1 to 15.
- Suppose we have 3 bins: [1, 6), [6, 11), [11, 15].
- If the data is $\{3, 7, 9, 12, 3, 5\}$, then:
  - Bin [1, 6): 3 data points $\{3, 3, 5\}$
  - Bin [6, 11): 2 data points $\{7, 9\}$
  - Bin [11, 15]: 1 data point $\{12\}$

Step 4: Constructing the Histogram

Constructing the Histogram:

- Now that we have the counts (absolute frequencies) for each bin, we can construct the histogram.
- The x-axis represents the bin intervals, and the y-axis represents the count of data points in each bin.
- Draw a bar for each bin where the height corresponds to the count (absolute frequency).

Example: Data Table

Step-by-Step Example: Consider the following data set:

3	5	7	9	11	3	4	6	8	10
12	15	1	2	14	5	7	9	11	13

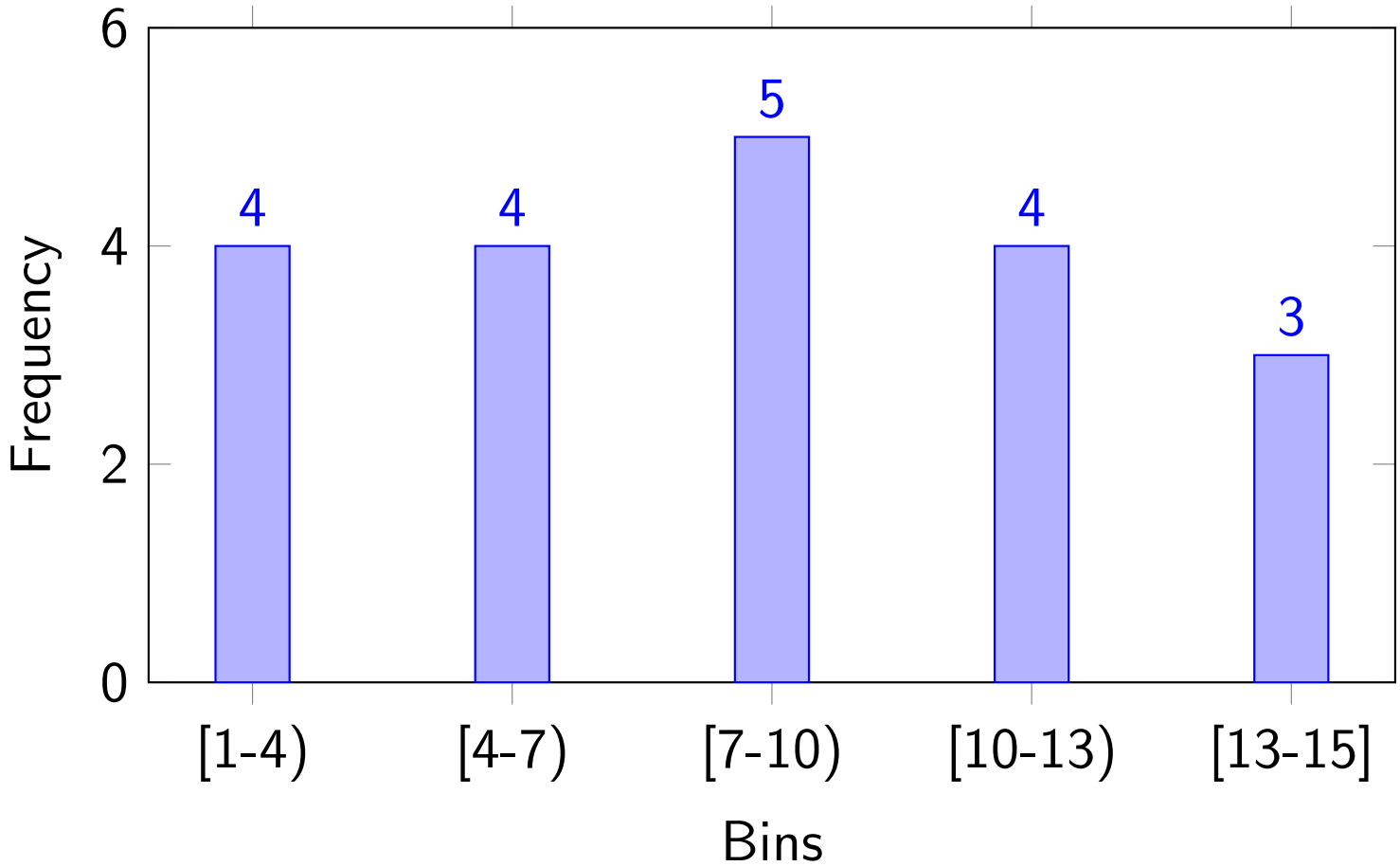
We will construct a histogram with 5 bins.

Example: Histogram with 5 Bins

Step-by-Step Example:

- Range of data: 1 to 15
- Bin size $= \frac{15-1}{5} = 2.8$ (rounded to 3)
- Bins:
  - $[1, 4)$: 4 data points $\{1, 2, 3, 3\}$
  - $[4, 7)$: 4 data points $\{4, 5, 5, 6\}$
  - $[7, 10)$: 5 data points $\{7, 7, 8, 9, 9\}$
  - $[10, 13)$: 4 data points $\{10, 11, 11, 12\}$
  - $[13, 15]$: 3 data points $\{13, 14, 15\}$

Example: Histogram with 5 Bins



Bar Plots

When Do We Use Bar Plots?

Bar Plots:

- Bar plots are used to display the frequency or proportion of categorical data.
- They are especially useful for comparing different categories.
- Each bar represents a category, and the height of the bar represents the value (frequency, proportion, etc.).

Use Cases:

- Visualizing survey responses.
- Comparing the frequency of different groups (e.g., gender, age groups).

What is a Frequency Distribution?

Frequency Distribution:

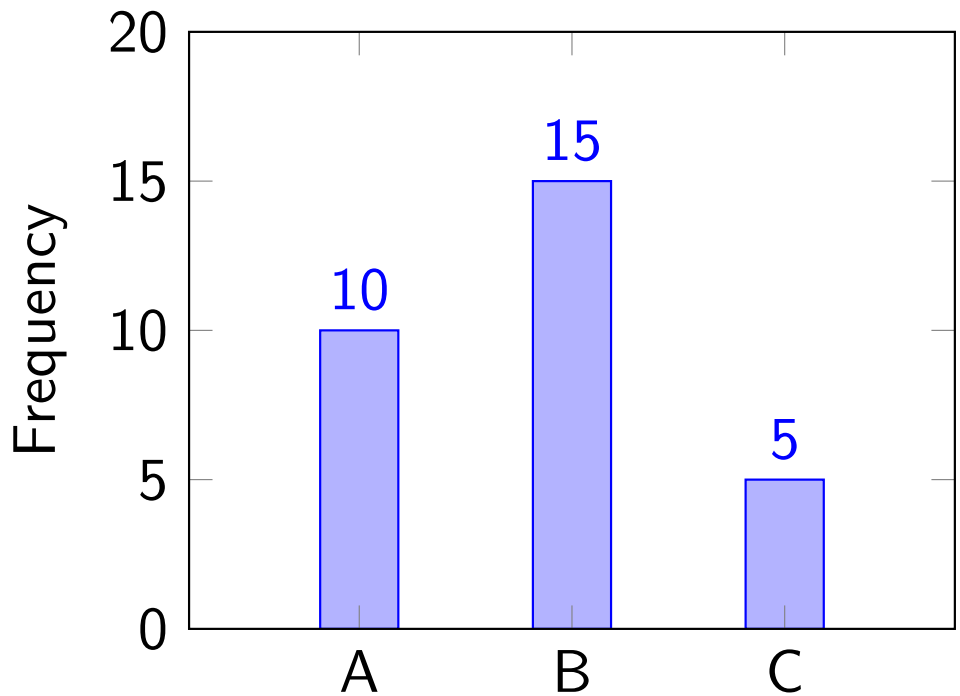
- A frequency distribution is a table that shows the frequency (or count) of each value or category.
- It can be visualized using a bar plot where each category corresponds to a bar.

Example:

Category	Frequency
A	10
B	15
C	5

Basic Bar Plot Structure

- The x-axis represents categories (e.g., different groups).
- The y-axis represents the frequency, proportion, or value corresponding to each category.
- Bars can be vertical or horizontal.



What is Cross-Tabulation?

Cross-Tabulation (Crosstab):

- Cross-tabulation is a method used to analyze the relationship between two categorical variables.
- It creates a matrix (or table) that shows the frequency distribution of the variables across their categories.

Example:

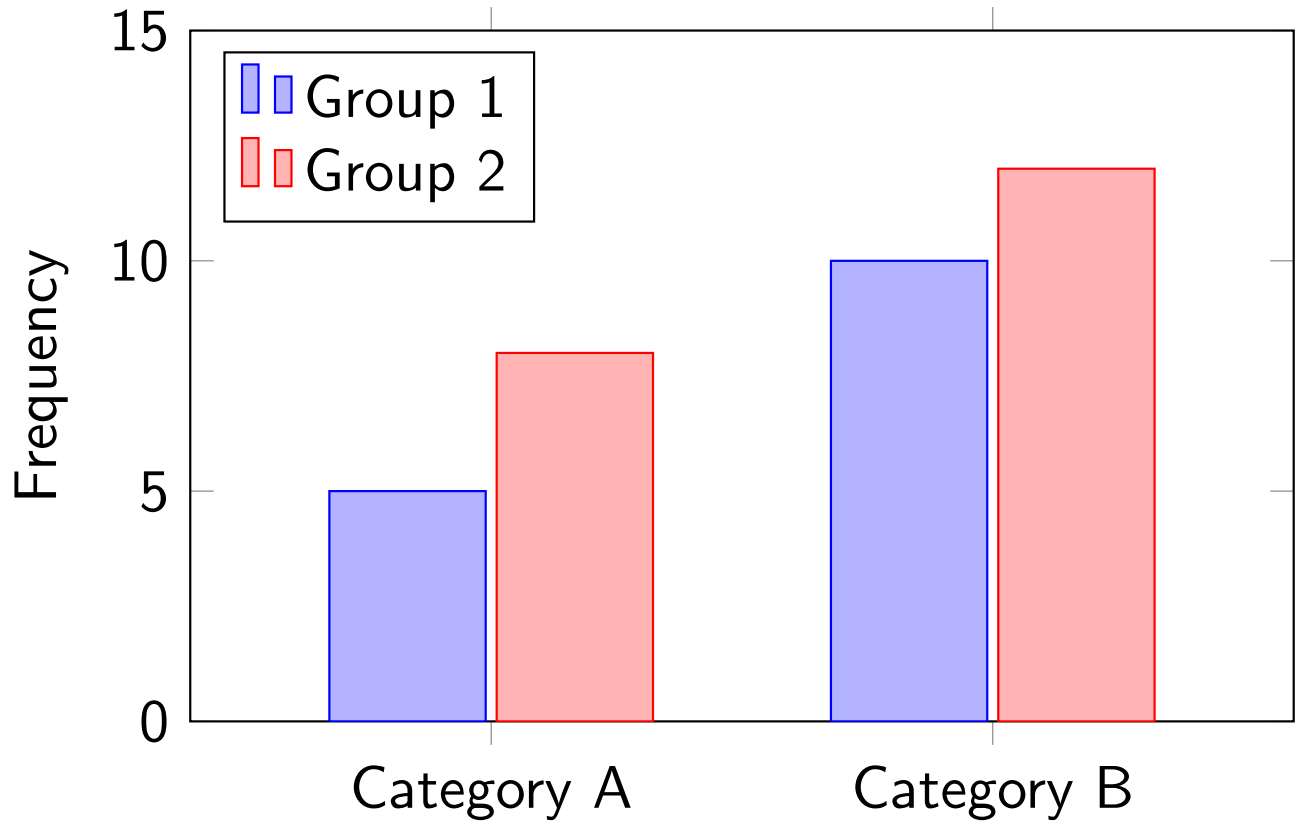
	Group 1	**Group 2**
Category A	5	8
Category B	10	12

Using Bar Plots for Cross-Tabulation

Bar Plots for Two Variables:

- Bar plots can also represent cross-tabulation by plotting grouped bars.
- Each group (e.g., Group 1, Group 2) has its own bar for each category.

Example:

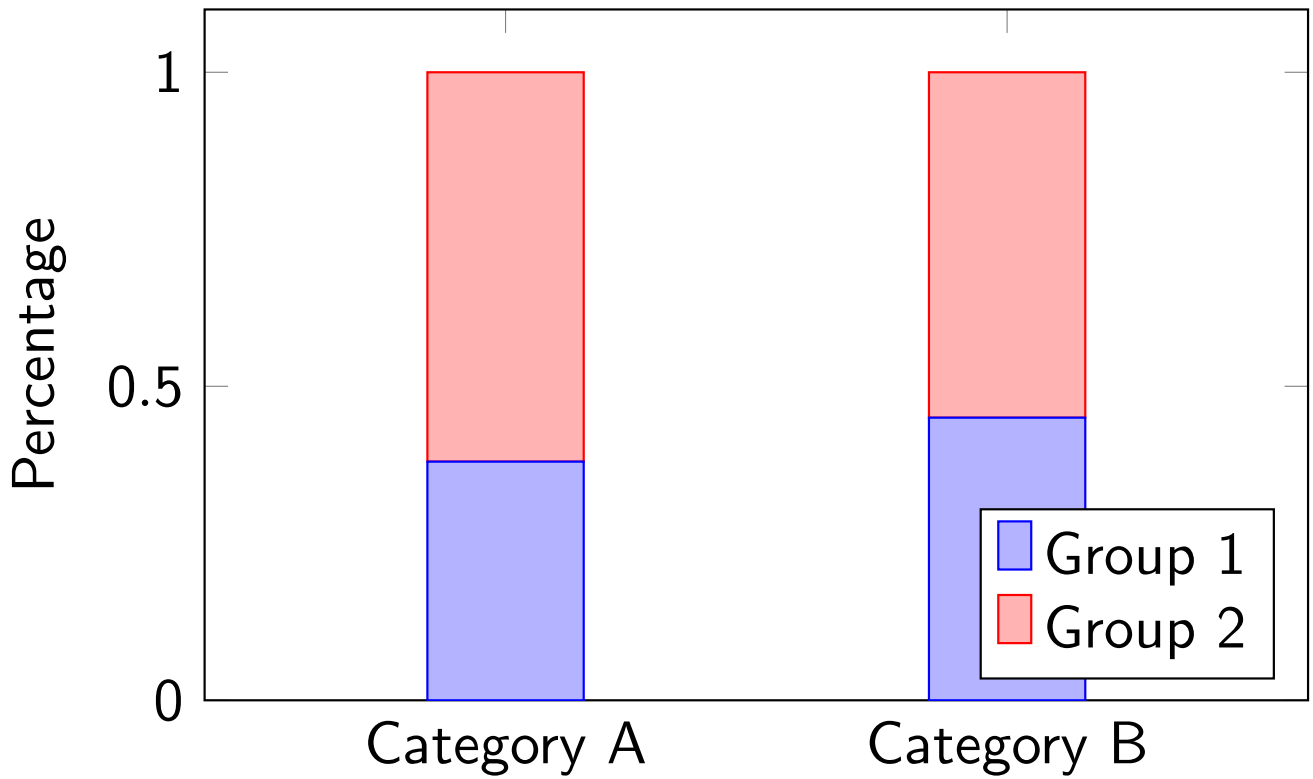


Using Bar Plots for Cross-Tabulation

Stacked Bar Plots for Two Variables:

- Bar plots can also be *stacked* within each category.
- We often do this when we represent percentages in the y-axis.

Example:



Box Plots

Introduction to Box Plots

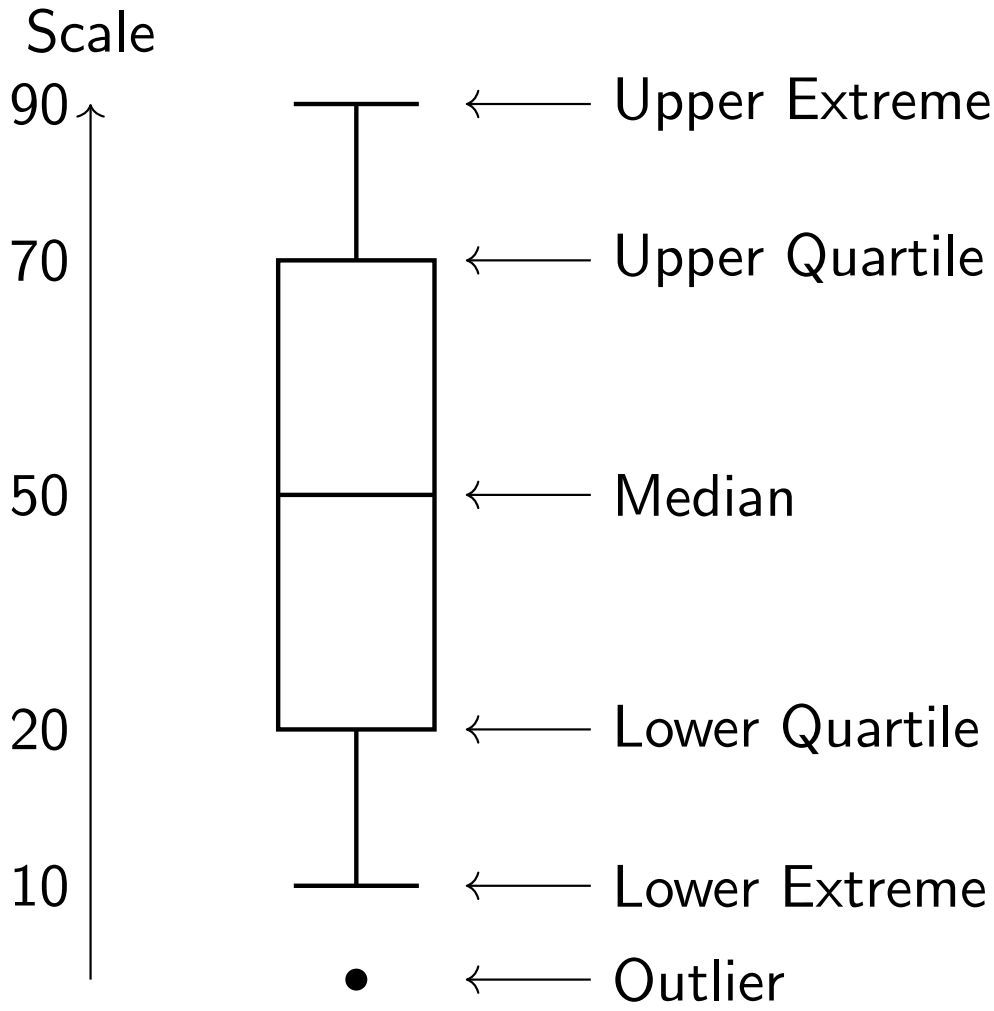
- A box plot (or box-and-whisker plot) displays the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable.
- The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the inter-quartile range.
- Box plots give a clear summary of data distribution and variability and are particularly useful for highlighting outliers and for comparing distributions across groups.

Constructing a Box Plot

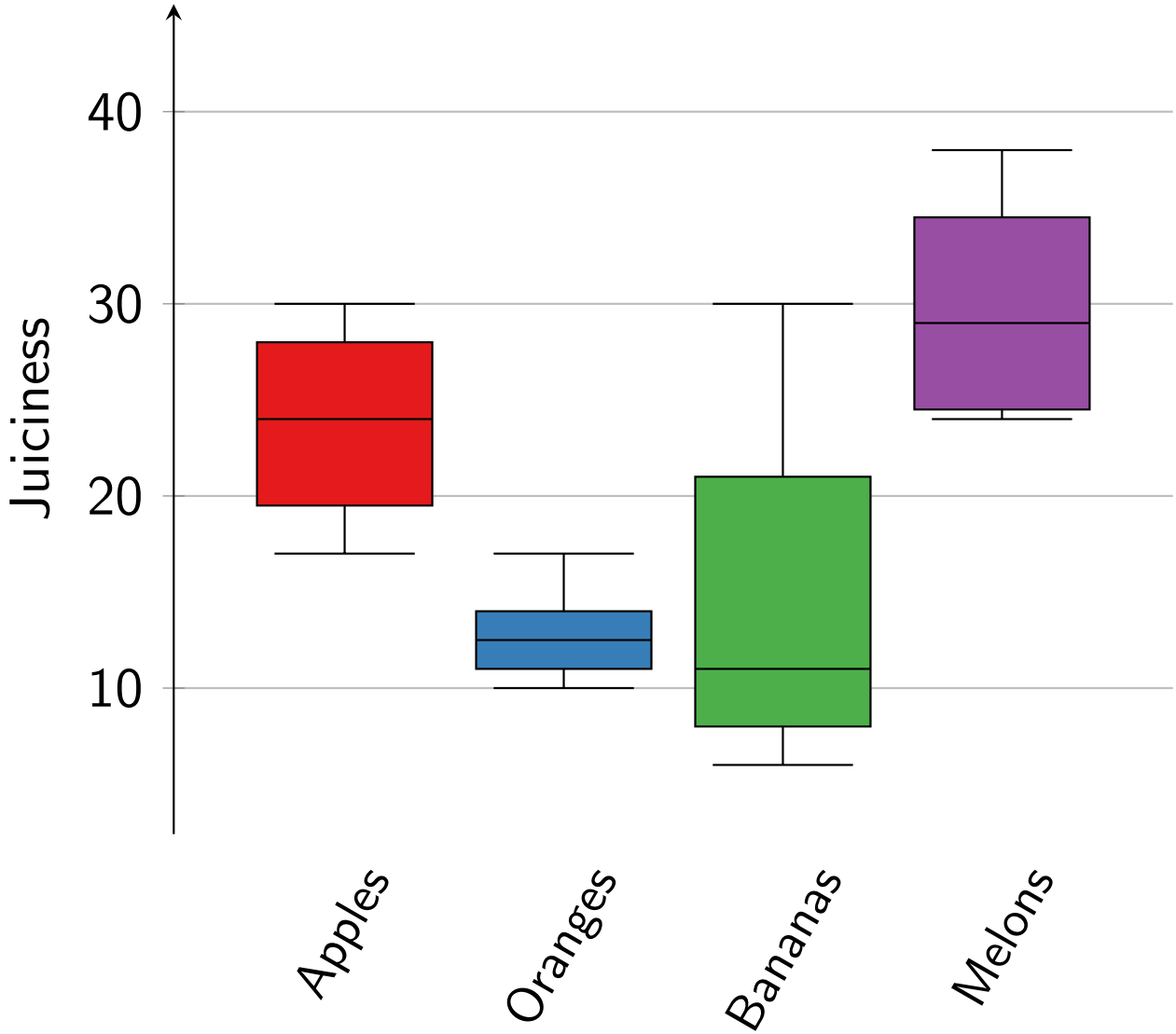
Steps to Construct a Box Plot:

1. Calculate the first (Q1) and third quartiles (Q3).
2. Find the interquartile range (IQR = Q3 - Q1).
3. Determine the "whiskers" which are typically set at 1.5 * IQR above Q3 and below Q1. Data points outside this range are considered outliers.
4. The median (Q2) is marked by a line inside the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side.

Structure of a Box Plot



Example: Box Plot with Multiple Categories



Dot Charts

Introduction to Dot Charts

What is a Dot Chart?

- A dot chart is a statistical chart consisting of data points plotted on a simple scale.
- It is used to compare frequency, count, or any measure across different categories.
- Similar to bar charts but dots are used instead of bars, making it less cluttered.

Why Use Dot Charts?

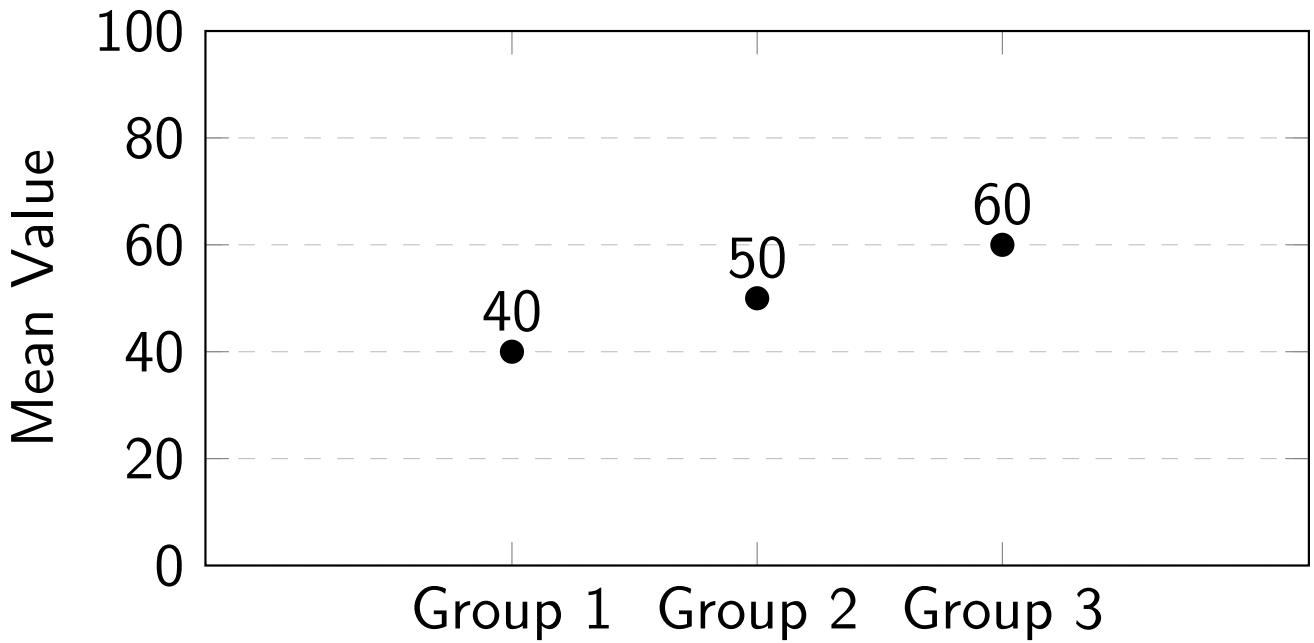
- **Clarity:** Provides a clear and precise representation of data points.
- **Comparison:** Facilitates easy comparison of multiple groups.
- **Space-efficient:** More effective in space usage than bar graphs.

Comparing Means Across Groups Using Dot Charts

How to Use Dot Charts for Comparing Means

- Dot charts are excellent for displaying the mean (or any central tendency) of different groups.
- Each dot represents the mean of a group, aligned along a single axis.
- The position of each dot on the scale directly reflects the value of the mean, making comparisons intuitive.

Example:



Scatter Plot

Introduction to Scatter Plots

What is a Scatter Plot?

- A scatter plot is a type of data visualization that shows the relationship between two numerical variables.
- Each point on the plot represents a pair of values: one on the x-axis and one on the y-axis.
- Scatter plots are useful for identifying correlations, trends, and outliers in data.

Applications of Scatter Plots

- Visualizing correlations between variables.
- Spotting clusters and patterns.
- Detecting outliers and anomalies.

How to Create a Scatter Plot

Steps to Create a Scatter Plot:

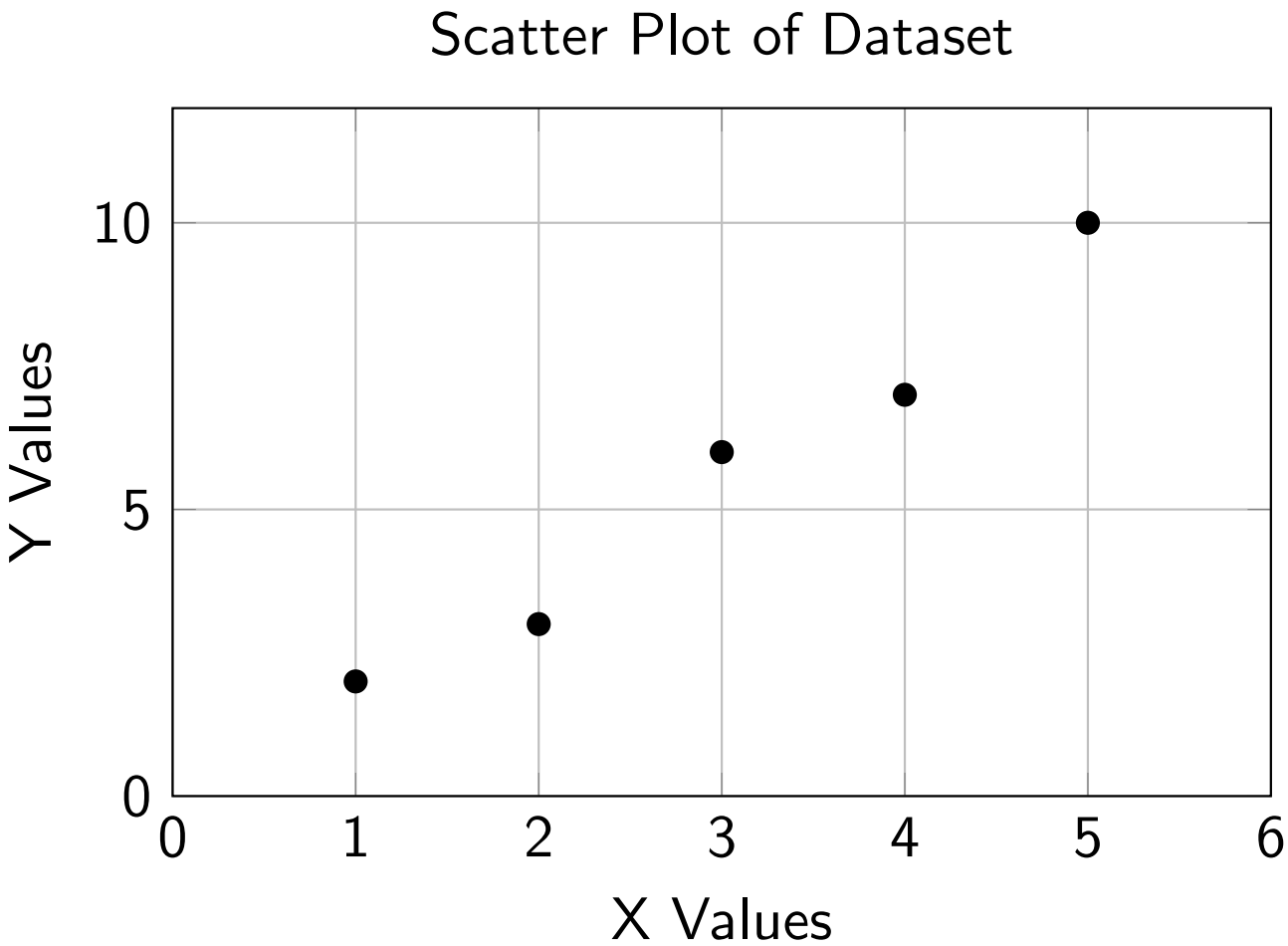
- **Step 1:** Gather your data points. You need two variables, one for the x-axis and one for the y-axis.
- **Step 2:** Plot each pair of values on a coordinate system.
- **Step 3:** Optionally, add labels, grid lines, and color coding to enhance interpretation.

Example:

- Dataset: (1,2), (2,3), (3,6), (4,7), (5,10)
- The scatter plot will show these points in a linear relationship.

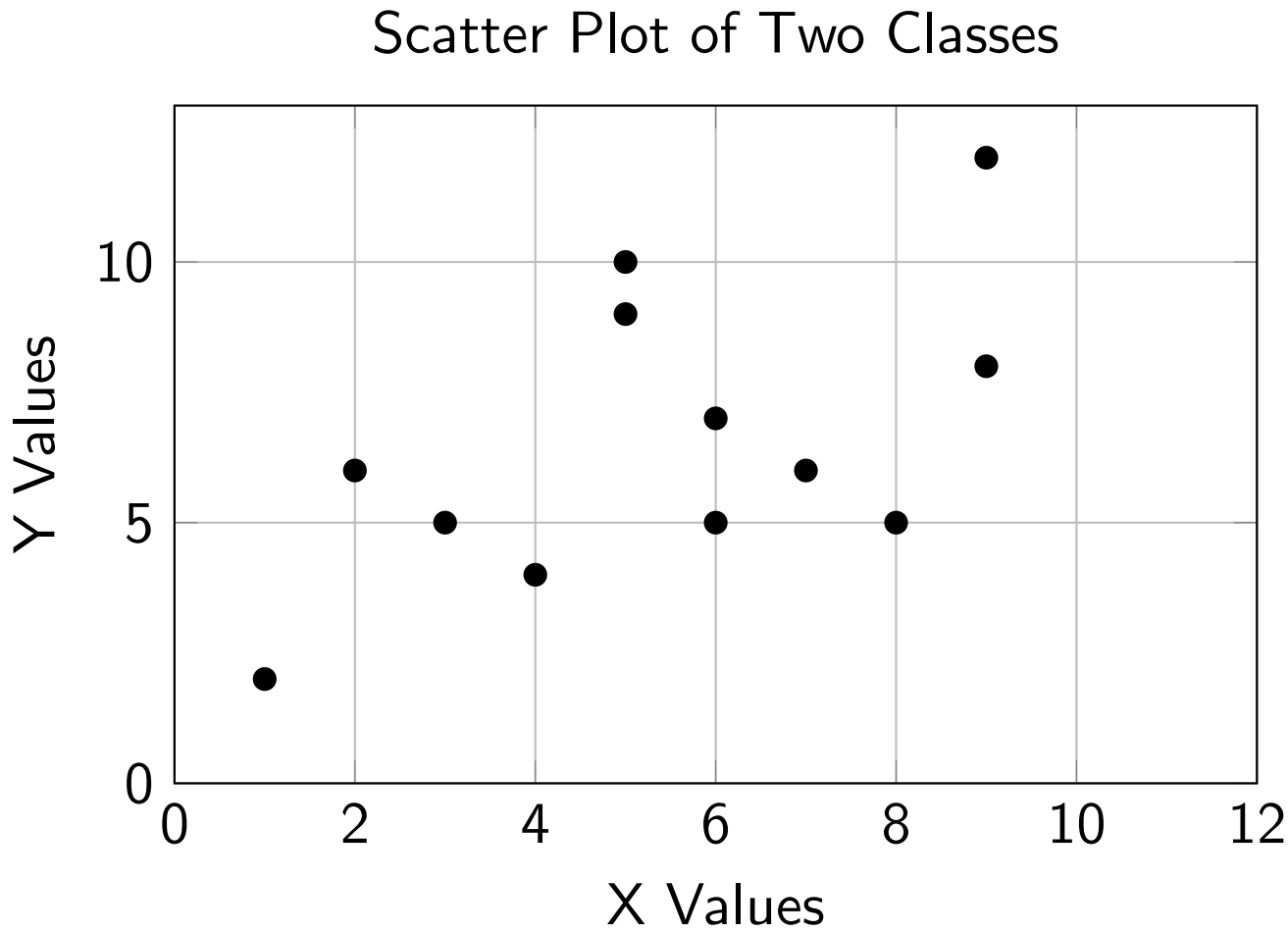
Scatter Plot Example (Basic)

Example of a Scatter Plot:



Scatter Plot Example with Two Classes

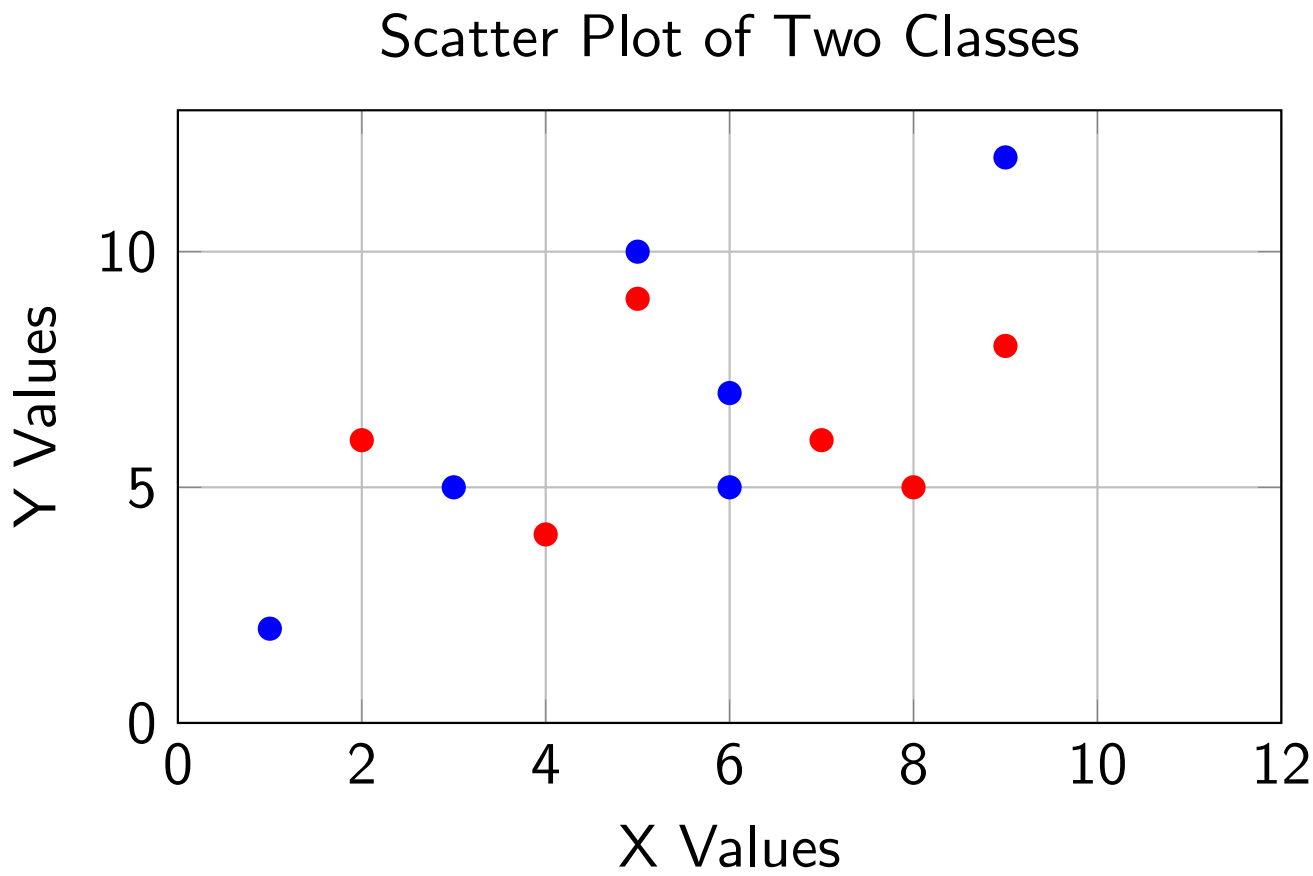
Imagine our dataset has two classes (Class A and Class B).



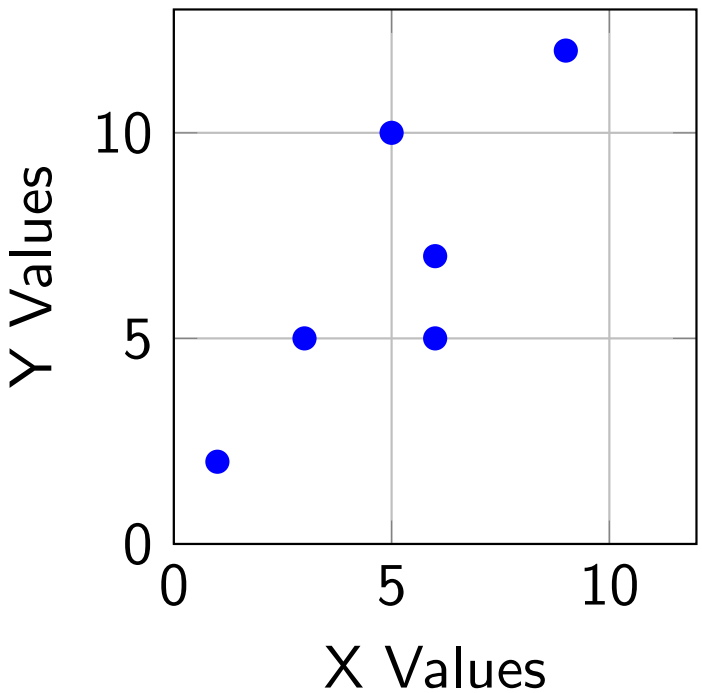
Scatter Plot Example with Two Classes

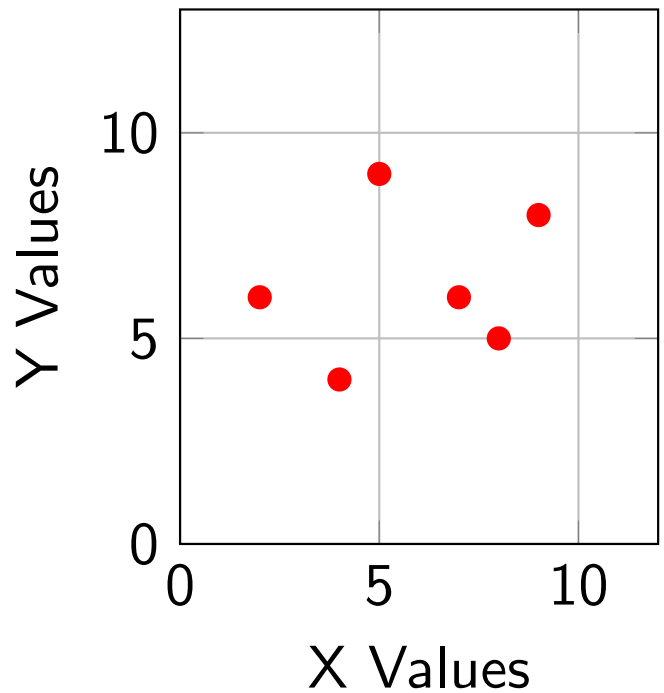
We have two different classes (Class A and Class B), each represented by different colors:

- Blue: Class A
- Red: Class B



Side-by-Side Comparison





Summary and Takeaways

Summary: Distributions and Plots

Key Takeaways:

- **Shape of Distribution:** Understand skewness (right, left) and mode (unimodal, bimodal, multimodal).
- **Histogram:** Useful for visualizing the distribution of intervals or bins of data.
- **Bar Plot:** Best for categorical data comparisons.
- **Box Plot:** Ideal for displaying the distribution's quartiles and identifying outliers.
- **Dot Chart:** Great for comparing multiple categories without the clutter of bars.
- **Scatter Plot:** Essential for identifying relationships and trends between two variables.