## Descriptive Statistics

- Describing data with tables and graphs
- Describing center of the data
- Describing variability of the data
- Bivariate descriptive statistics

## Describing data with tables and graphs:

We organize raw data in frequency tables showing classes, frequencies, relative frequencies and cumulative frequencies. Then we draw graphs.

### Example 1: Categorical (nominal) data

Ten people ($n = 10$) were asked about their marital status.

Data: Single, Single, Divorced, Divorced, Widowed,
Single, Married, Married, Widowed, Single

| Class | Frequency | Relative Frequency | Percentage |
|---|---|---|---|
| Divorced | 2 | 2/10 = 0.2 | 20% |
| Married | 2 | 0.2 | 20% |
| Single | 4 | 0.4 | 40% |
| Widowed | 2 | 0.2 | 20% |
| Total | 10 | 1 | 100% |

### Example 2: Categorical (ordinal) data

Respondents in a study are categorized as young (Y), middle-aged (M) and old (O). Data for 20 respondents are given below.

Data: M, M, Y, O, O, M, M, Y, Y, M, Y, M, O, O, M, M, M, Y, O, M

| Class | Frequency | Relative Frequency | Cumulative Frequency |
|---|---|---|---|
| Y | 5 | 5/20 = 0.25 | 5 |
| M | 10 | 0.50 | 15 |
| O | 5 | 0.25 | 20 |
| Total | 20 | 1 | ---- |

**Example 3: Discrete data**

Data were collected from an industrial plant.
$X =$ Number of spots in a metal sheet.
Data: 3, 2, 0, 4, 1, 1, ⋯, 3.

| $X$ | Number of sheets | Relative Frequency | Cumulative Frequency |
|---|---|---|---|
| 0 | 10 | 0.125 | 10 |
| 1 | 27 | 0.338 | 37 |
| 2 | 15 | 0.188 | 52 |
| 3 | 18 | 0.225 | 70 |
| 4 | 10 | 0.125 | 80 |

**Example 4: Continuous data**

$X =$ Duration (minutes) of production halt due to machine failure.
Sample size, $n = 100$.
Data: 49.2, 89.7, 41.3, 50.0, ⋯, 60.9

| $X$ | Frequency | Relative Frequency | Cumulative Frequency |
|---|---|---|---|
| 40-50 | 10 | 0.10 | 10 |
| 50-60 | 20 | 0.20 | 30 |
| 60-70 | 40 | 0.40 | 70 |
| 70-80 | 20 | 0.20 | 90 |
| 80-90 | 10 | 0.10 | 100 |

Here, the class '40-50' means $40 \leq X < 50$ and '50-60' means $50 \leq X < 60$. If data contain the value 50.0, it is counted in '50-60'. That is, the classes are non-overlapping. Another option: '40-50' means $40 < X \leq 50$, and so on.

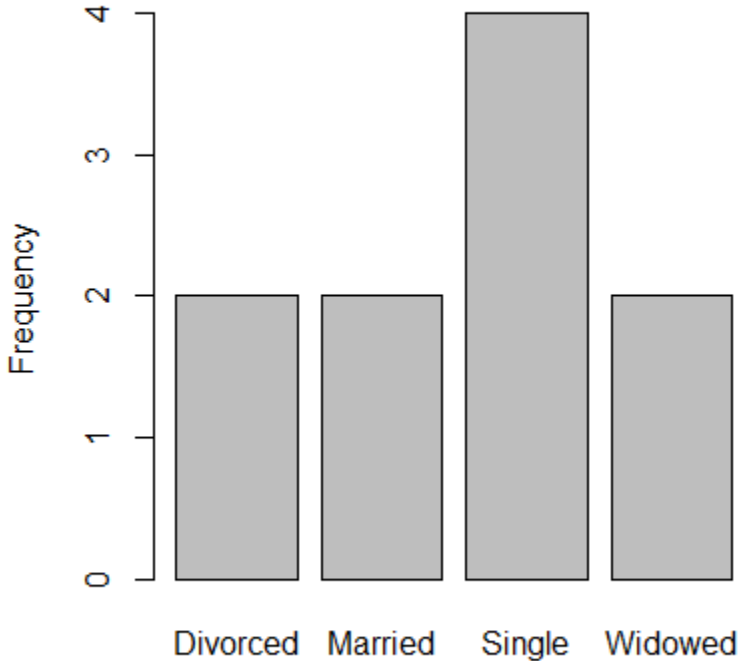
**Remarks** on Frequency distribution:

- The classes should be non-overlapping.
- Helps view how the data are distributed among different classes.

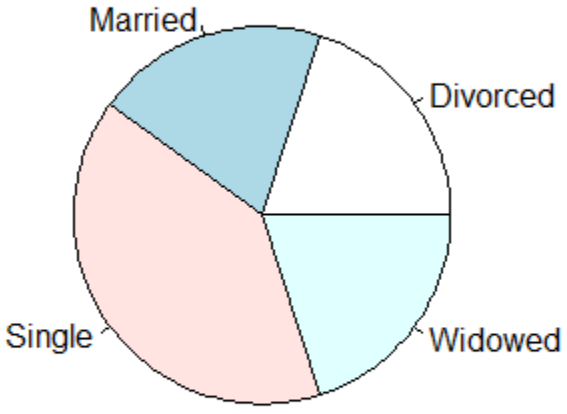
## Graphs

- For **Categorical** variables: Bar chart and Pie chart
- For **Discrete** variables: Bar Chart
- For **Continuous** variables: Histogram

Bar Chart for Example 1 (Marital Status)

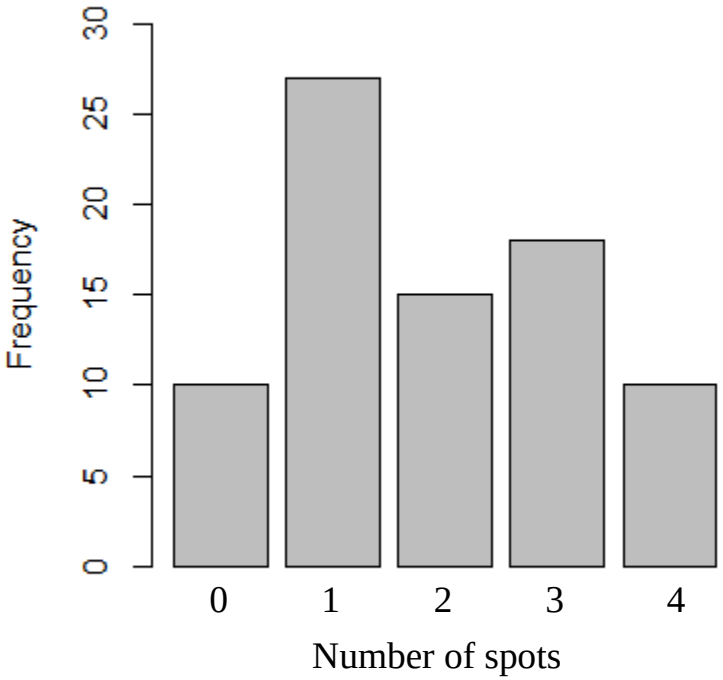



Pie Chart for Example 1 (Marital Status)

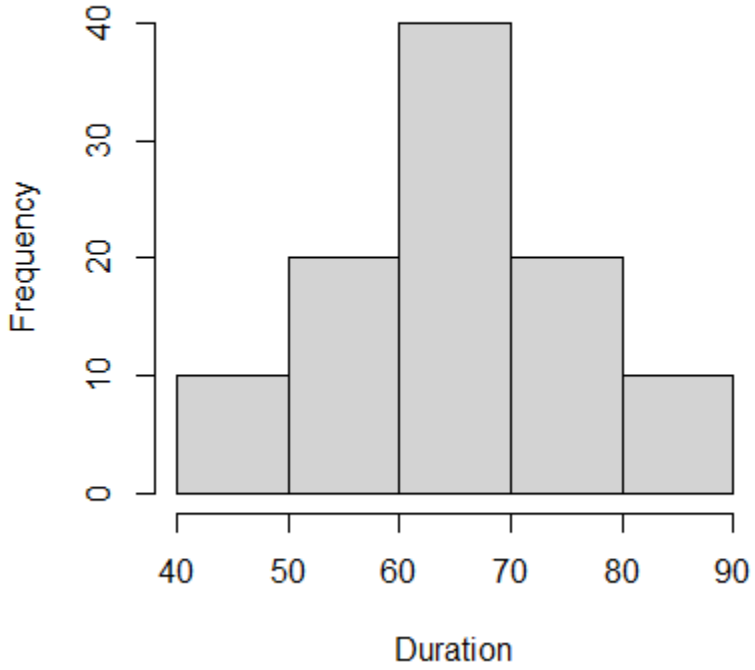


Bar chart and pie chart can also be drawn for Example 2 (Ordinal data)

Bar chart for Example 3 (Discrete)




Histogram for Example 4 (Continuous)

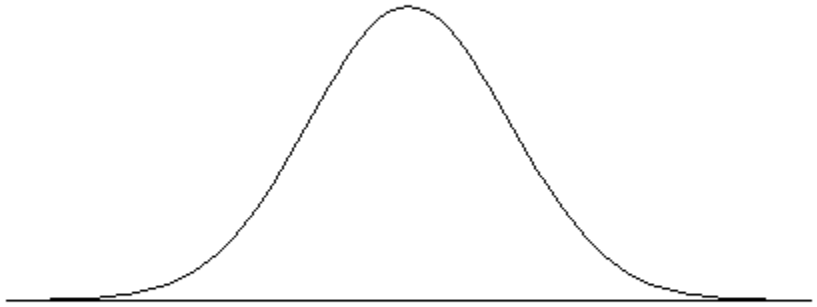


**Difference between bar chart and histogram**

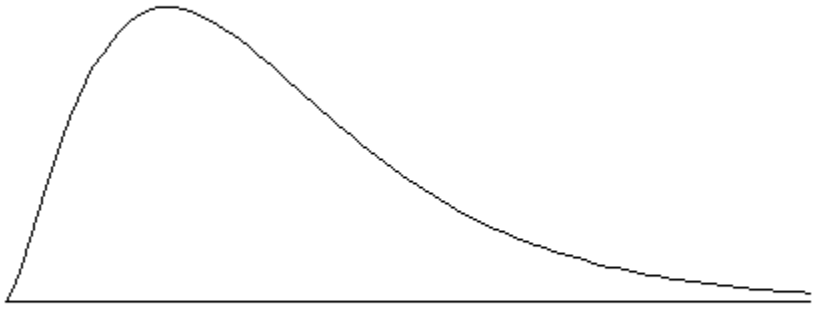
1. In a bar chart, bars are separate. In a histogram, bars are attached to each other.
2. In a bar chart, height of bar represents frequency. In a histogram, area of bar represents frequency.

**Shape of a distribution**

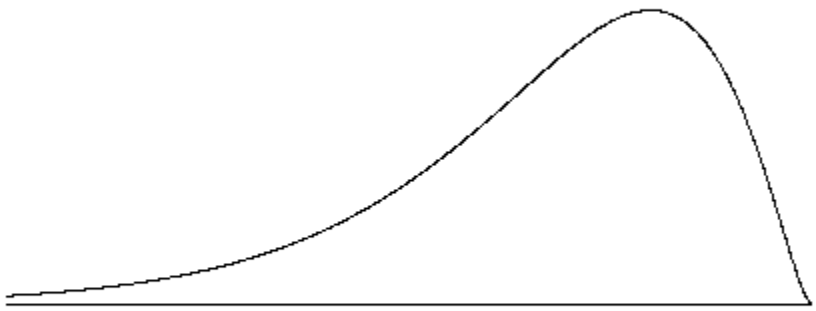
A histogram can have different shapes:



Symmetric



Positively skewed (rightward skewed)



Negatively skewed (leftward skewed)