(2) Descriptive Statistics

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Jul 16, 2023

Outlines

- Numerical variables
  - Descriptive statistics
  - Plots
- Categorical variables
  - Descriptive statistics
  - Plots

Expected outcomes

- Familiarize with common descriptive statistics and plots for numerical and categorical variables

Numerical variables

Central Tendency

- Mean
- Median
- Mode

$$X = 1, 2, 2, \underline{3, 3, 3}, 4, 4, 5$$

$$\text{Mean} = \bar{X} = \frac{\sum X}{n} = \frac{27}{9} = 3$$

$$\text{Location of median} = \frac{n+1}{2} = \frac{9+1}{2} = \text{5th number}$$

$$Median = 3$$

$$\text{Mode} = \text{most frequent value} = 3$$

Dispersion

- Range
- Inter-quartile range
- Variance
- Standard deviation

$X = \underline{1}, 2, 2, 3, 3, 3, 4, 4, \underline{5}$

$$\text{Range} = \text{Maximum value} - \text{Minimum value} = 5 - 1 = 4$$

X = 1, 2, 2, 3, 3, 3, 4, 4, 5

$$\text{Interquartile range (IQR)} = \text{Quartile 3}\left(Q_3\right) - \text{Quartile 1}\left(Q_1\right)$$

$$\text{Location of } Q_1 = \frac{n+1}{4} = 2.5 = \text{2nd and 3rd numbers} = \left(2,2\right)$$

$$\text{Location of } Q_3 = \frac{3}{4}\left(n+1\right) = 7.5 = \text{7th and 8th numbers} = \left(4,4\right)$$

$$Q_1 = \frac{\left(2+2\right)}{2} = 2 \text{ and } Q_3 = \frac{\left(4+4\right)}{2} = 4$$

$$\text{IQR} = Q_3 - Q_1 = 4 - 2 = 2$$

X = 1, 2, 2, 3, 3, 3, 4, 4, 5

$$\text{Sample variance} = s^2$$

$$= \frac{\sum (X - \bar{X})^2}{n-1}$$

$$= \frac{(1-3)^2 + (2-3)^2 + \ldots + (5-3)^2}{9-1}$$

$$= \frac{12}{8} = 1.5$$

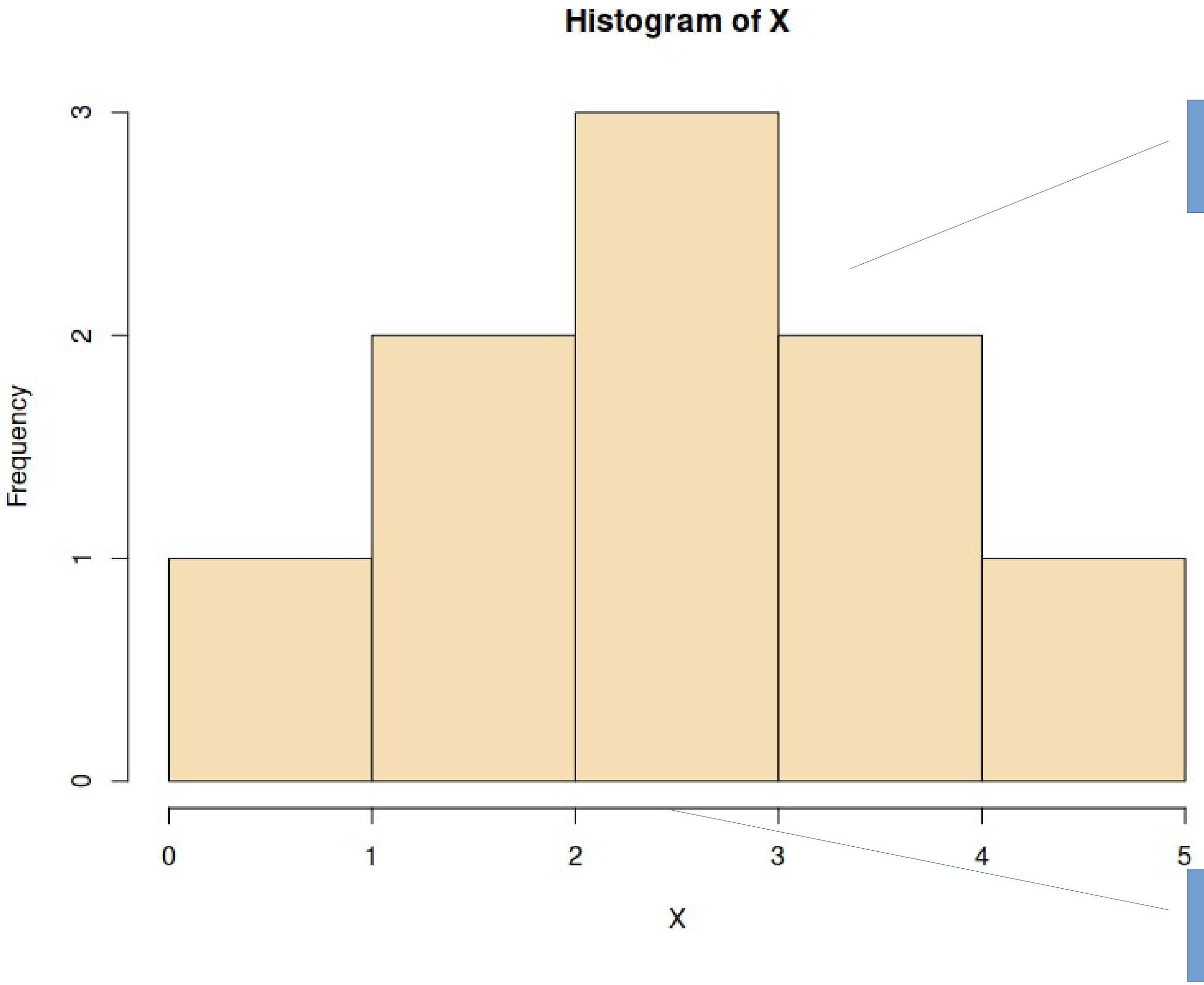
X = 1, 2, 2, 3, 3, 3, 4, 4, 5

$$
\begin{aligned}
\text{Sample standard deviation} &= s \\
&= \sqrt{\text{Sample variance}} \\
&= \sqrt{s^2} \\
&= \sqrt{1.5} = 1.2
\end{aligned}
$$

Plots

- One variable:
  - Histogram
  - Box-and-whisker plot
- Two variables:
  - Scatter plot

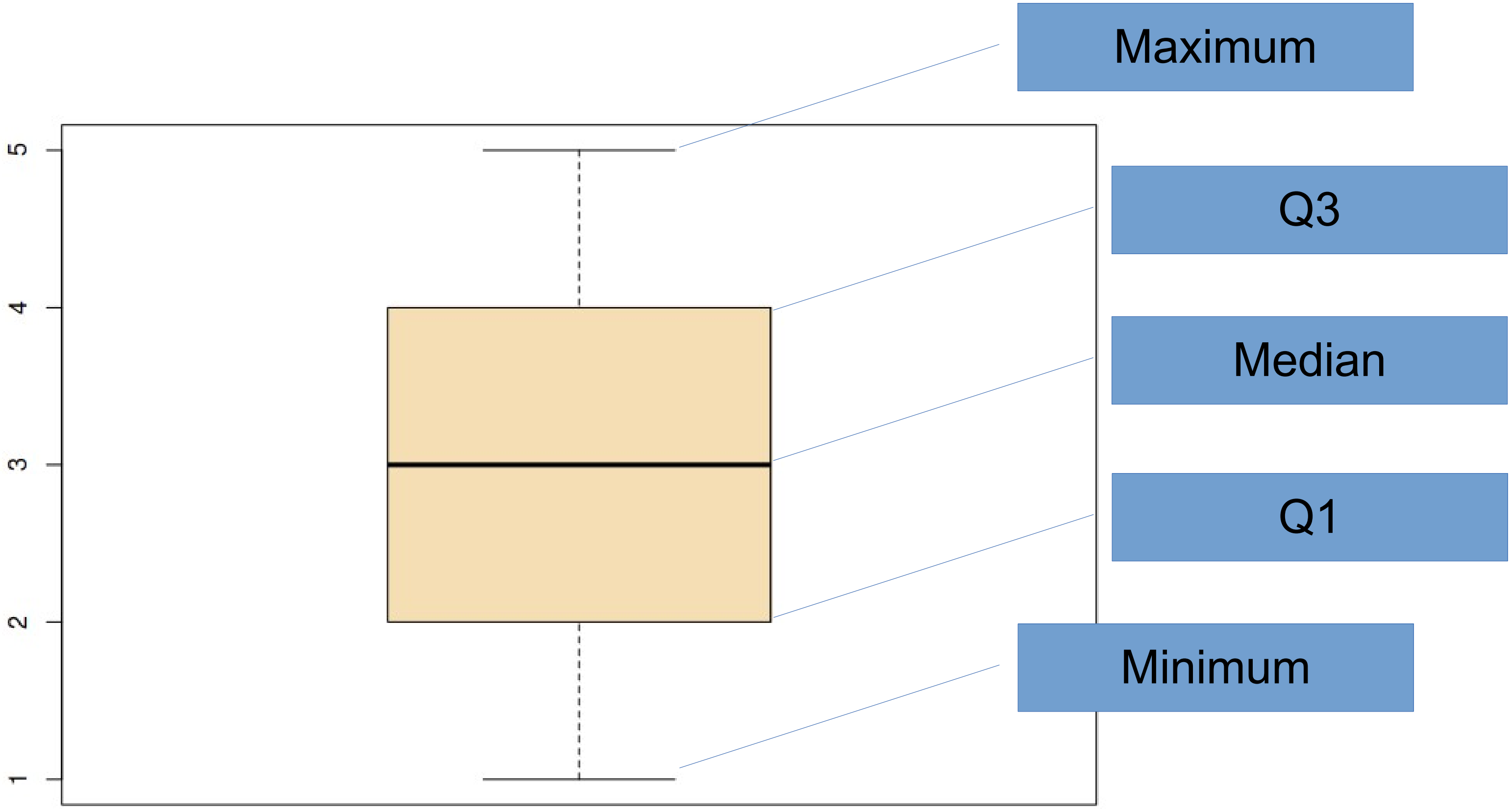
Plots: Histogram



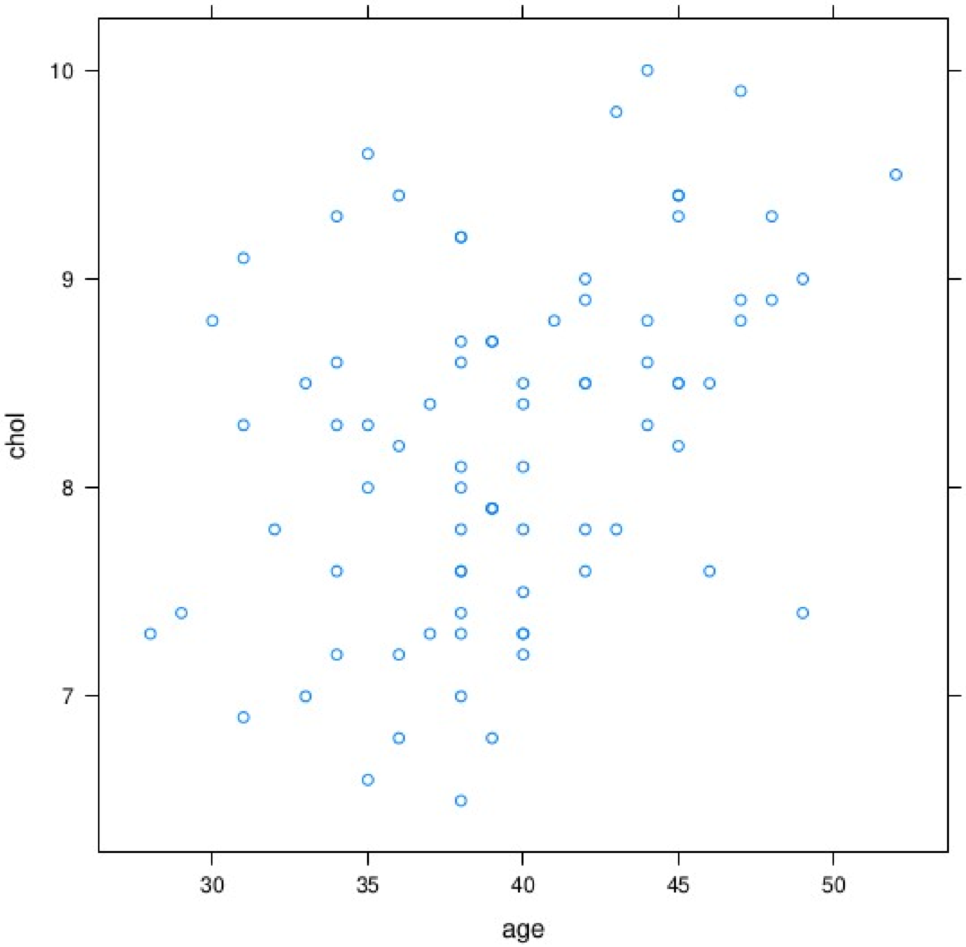
Normally distributed data

Mean, Median, Mode = 3

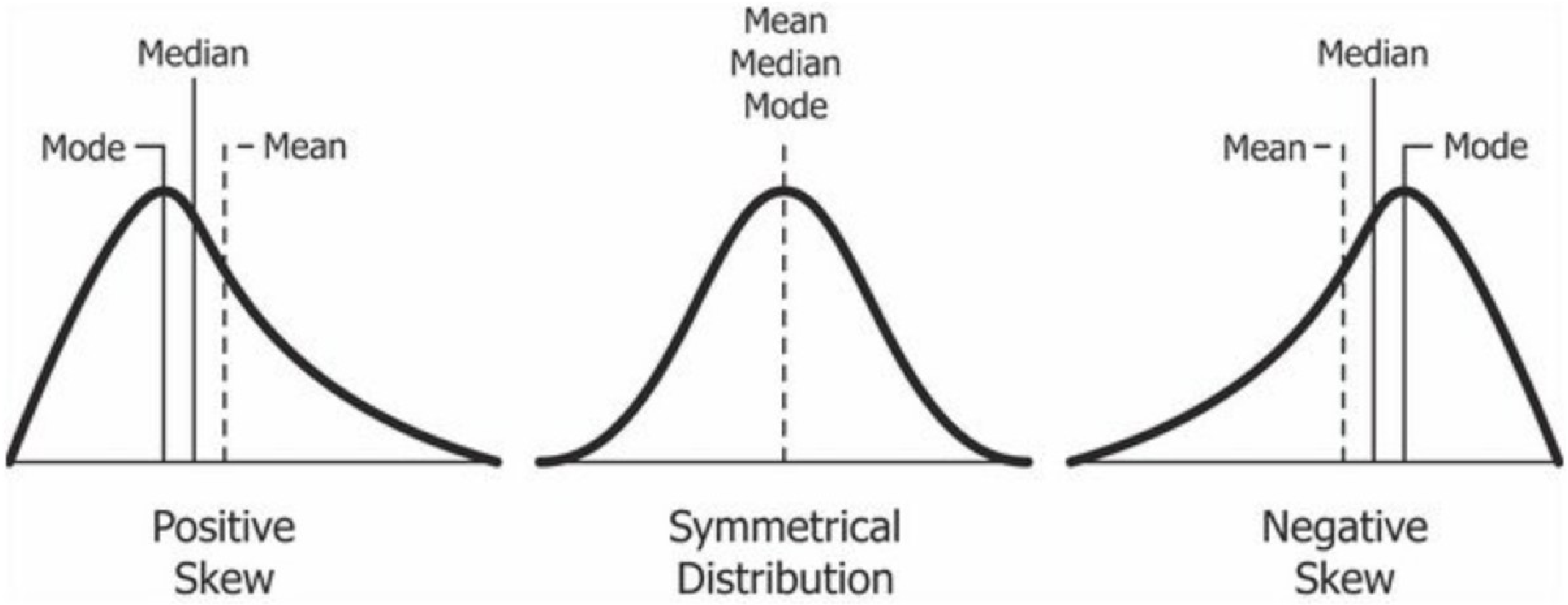
Plots: Boxplot



Plots: Scatter Plot



Skewness



Skewed to the right

Normal distribution

Skewed to the left

Source: Diva Jain, https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Implication

- When data is not normally distributed – use median (IQR) in place of mean (SD)

Categorical variables

Count and proportion

$$\text{Count} = n\, per\, category$$

$$\text{Proportion} = p = \frac{n\,\text{per category}}{n}$$

$$\text{Percentage} = \frac{n\,\text{per category}}{n} \times 100\,\% = p \times 100\,\%$$

Count and proportion

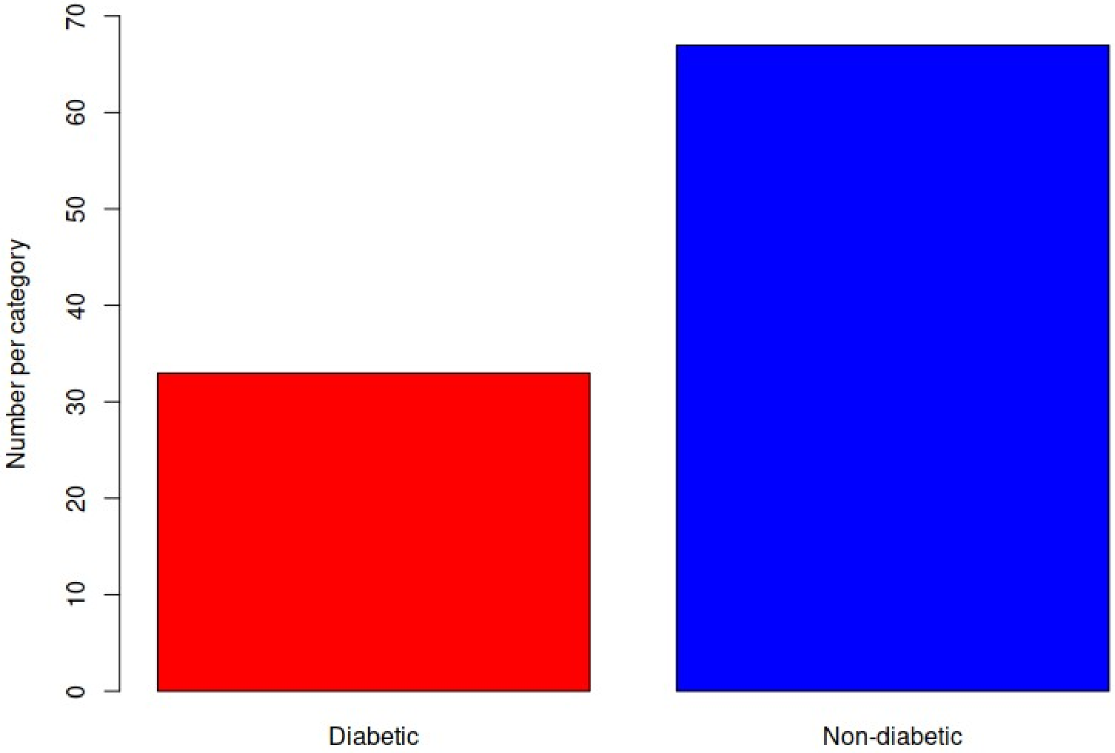
Variable	**Category**	**n**	**p**	**%**
Gender	***Male***	40	0.4	40.0%
	Female	60	0.6	60.0%
Diabetic	***Yes***	33	0.33	33.0%
	No	67	0.67	67.0%

Cross-tabulation table

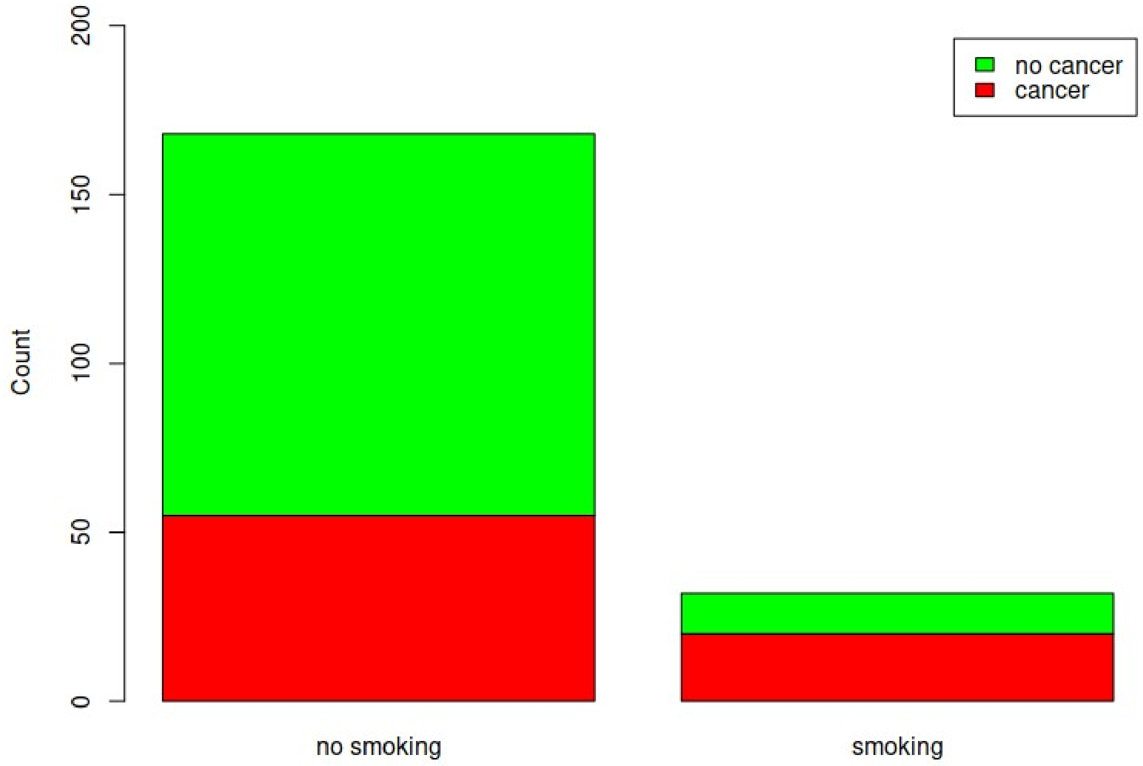
- Between two categorical variables

Smoking	**Lung Cancer**	
	Yes	***No***
Yes	20 (62.5%)	12 (37.5%)
No	55 (32.7%)	113 (67.3%)

Plots: Bar Chart



Plots: Stacked Bar Chart



Descriptive in Jourrnal

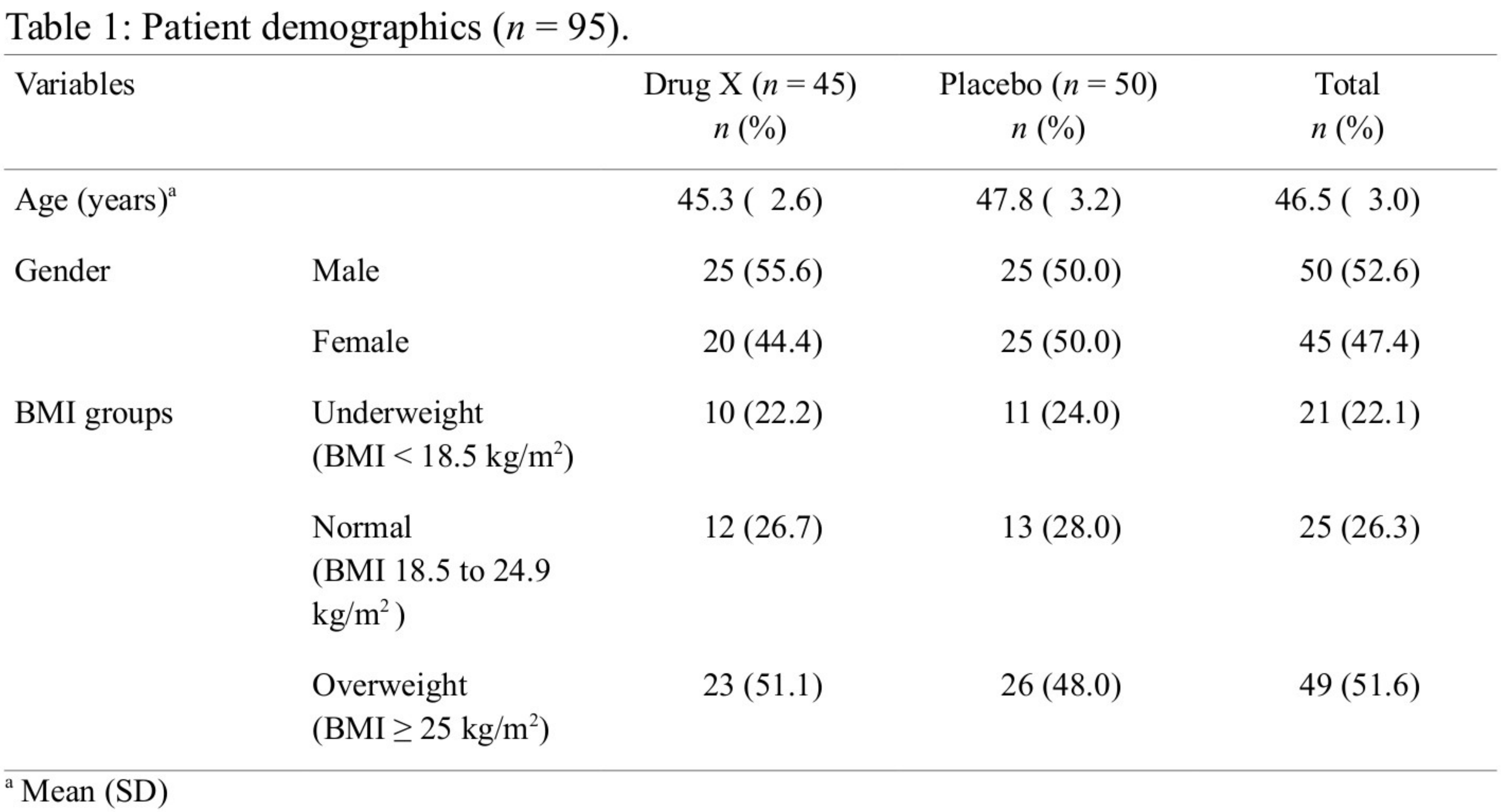


Table 1: Patient demographics (*n* = 95).

Variables		Drug X (*n* = 45)	Placebo (*n* = 50)	Total
		n (%)	*n* (%)	*n* (%)
Age (years)[a]		45.3 (2.6)	47.8 (3.2)	46.5 (3.0)
Gender	Male	25 (55.6)	25 (50.0)	50 (52.6)
	Female	20 (44.4)	25 (50.0)	45 (47.4)
BMI groups	Underweight (BMI < 18.5 kg/m^2)	10 (22.2)	11 (24.0)	21 (22.1)
	Normal (BMI 18.5 to 24.9 kg/m^2)	12 (26.7)	13 (28.0)	25 (26.3)
	Overweight (BMI ≥ 25 kg/m^2)	23 (51.1)	26 (48.0)	49 (51.6)

[a] Mean (SD)

Quiz

- For numerical variable:
  - – List measures of central tendency
  - – List measures of dispersion
  - – Describe suitable plots
  - – Describe "skewness" in relation to mean and median
- For categorical variable:
  - – Describe suitable statistics
  - – Describe suitable data presentation

Thank You

Plots: Histogram Extra

Raw SBP data, n = 300

```
127.2 110.1 114.3 122.2 117.5 122.6 105.1 117.8 121.4 110.9 133.7
121.2 124.5 117.1 116.2 118.7 123.3 111.9 130.9 106.3 123.3 119.4
131.9 111.5 122.5 117.8 117.6 123.9 120.9 118.1 121.8 116.2 126.7
121.4 126.7 126.2 117.7 119.2 118.8 121.3 117.7 115.3 130.8 117.9
131.5 116.0 114.2 117.9 123.6 120.7 118.9 117.1 108.0 124.6 117.2
118.0 127.3 115.3 123.2 119.0 124.3 110.2 130.9 131.1 102.9 113.7
124.0 122.8 115.9 121.7 124.9 115.7 111.2 121.6 110.3 122.4 119.4
122.4 104.2 123.8 110.6 115.3 114.8 120.8 115.2 118.6 129.9 120.9
119.0 127.5 129.6 110.7 124.6 134.3 113.3 115.3 118.3 119.9 137.1
119.0 102.9 115.7 110.8 107.3 113.2 117.2 127.3 117.3 122.6 114.2
122.7 113.2 123.9 113.7 106.5 116.9 127.6 118.2 105.9 114.6 119.4
121.4 117.9 125.4 117.7 115.0 122.4 124.0 122.2 109.6 130.0 126.9
117.8 123.9 131.4 124.1 130.7 127.5 112.0 105.8 122.3 124.2 117.4
128.0 114.6 122.4 118.0 109.8 117.2 122.6 112.0 110.3 115.7 131.6
131.2 126.0 126.2 115.9 123.6 121.6 129.9 121.6 120.1 114.3 128.6
132.0 114.5 131.1 132.5 113.6 125.9 123.5 102.9 132.1 109.5 110.6
117.1 112.4 113.2 117.4 117.8 113.0 129.8 126.6 132.7 118.5 109.0
110.2 129.5 136.3 109.4 117.6 119.2 120.1 127.2 126.7 128.9 125.9
121.6 122.0 133.8 111.5 115.8 120.2 115.6 125.7 121.6 135.0 110.0
125.7 103.6 129.3 121.5 120.8 123.0 117.5 122.9 122.0 129.3 132.9
123.3 115.8 118.1 126.6 117.9 123.1 122.5 122.3 118.1 121.4 110.3
108.3 117.8 120.8 122.6 108.6 121.2 129.0 124.5 127.2 116.5 106.9
120.7 117.1 136.7 127.9 125.5 116.4 119.4 111.7 123.9 121.5 119.3
116.1 115.8 120.4 116.5 109.1 112.1 125.1 126.4 126.5 130.8 124.4
128.3 128.2 116.3 114.4 113.3 109.9 119.7 124.6 110.1 114.7 122.0
119.1 112.0 121.2 122.4 113.8 124.2 109.7 137.5 124.1 102.5 131.3
125.9 132.0 119.8 120.3 114.4 111.6 119.5 114.3 121.1 120.5 117.0
121.9 113.0 114.3
```

Tabulate

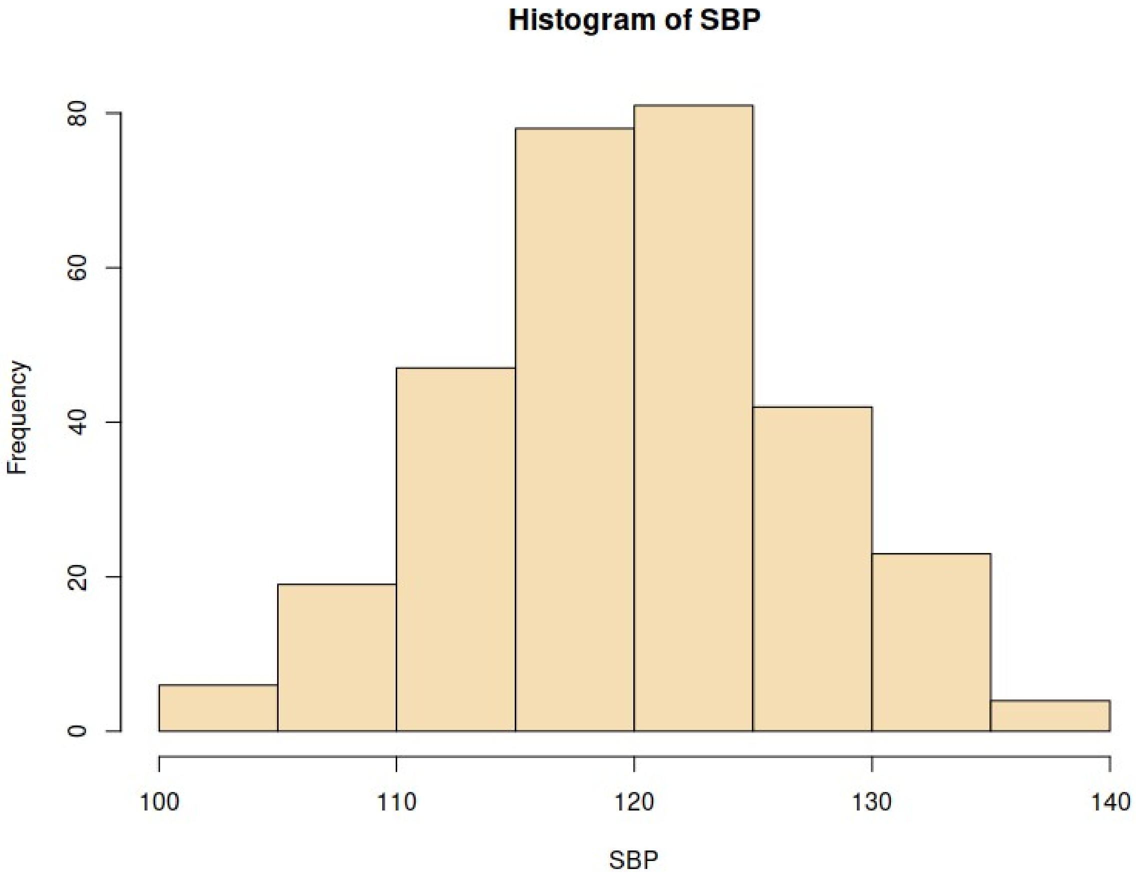
Group	SBP	Frequency
1	[-Inf,105)	6
2	[105,110)	19
3	[110,115)	47
4	[115,120)	78
5	[120,125)	81
6	[125,130)	42
7	[130,135)	23
8	[135,140)	4

Plots: Histogram Extra

Tabulate

Group	SBP	Frequency
1	[-Inf,105)	6
2	[105,110)	19
3	[110,115)	47
4	[115,120)	78
5	[120,125)	81
6	[125,130)	42
7	[130,135)	23
8	[135,140)	4

Plot


Histogram of SBP
Frequency
SBP