Exploratory Data Analysis

Part 1

Exploratory Data Analysis

- *Descriptive statistics* are used to summarize a sample.

Exploratory Data Analysis

- *Descriptive statistics* are used to summarize a sample.

- This topic covers **graphical and numerical summaries** that are the key elements of descriptive statistics.

Exploratory Data Analysis

- *Descriptive statistics* are used to summarize a sample.
- This topic covers **graphical and numerical summaries** that are the key elements of descriptive statistics.
- We shall learn ways of describing the information in our sample.

Exploratory Data Analysis

- *Descriptive statistics* are used to summarize a sample.
- This topic covers **graphical and numerical summaries** that are the key elements of descriptive statistics.
- We shall learn ways of describing the information in our sample.
- These are known as Exploratory Data Analysis (EDA) techniques.

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record
  - number of cigarettes a day (0, 1, 2, etc.)

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record
  - number of cigarettes a day (0, 1, 2, etc.)
  - gender (M, F)

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record
  - number of cigarettes a day (0, 1, 2, etc.)
  - gender (M, F)
  - age

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record
  - number of cigarettes a day (0, 1, 2, etc.)
  - gender (M, F)
  - age
  - lung caner (yes, no)

Variables

Definition 1 (Variable)

A **variable** is any characteristic observed in a study.

- The term variable highlights that the data values vary.
- If we study the relationship of smoking and lung cancer, we should record
  - number of cigarettes a day (0, 1, 2, etc.)
  - gender (M, F)
  - age
  - lung caner (yes, no)
- A variable can be categorical or quantitative.

Categorical and Quantitative Variables

Definition 2 (Categorical and Quantitative Variables)

1. A variable is called **categorical** if each observation belongs to one of a set of categories.

Categorical and Quantitative Variables

Definition 2 (Categorical and Quantitative Variables)

1. A variable is called **categorical** if each observation belongs to one of a set of categories.
2. A variable is called **quantitative** if observations of it take on numerical values that represent different magnitudes of the variable.

Categorical and Quantitative Variables

Definition 2 (Categorical and Quantitative Variables)

1. A variable is called **categorical** if each observation belongs to one of a set of categories.
2. A variable is called **quantitative** if observations of it take on numerical values that represent different magnitudes of the variable.

- Categorical: gender (Male and Female); race (Chinese, Malay and Indian); type of residence (HDB, Condo, EC, Landed house); etc.

Categorical and Quantitative Variables

Definition 2 (Categorical and Quantitative Variables)

1. A variable is called **categorical** if each observation belongs to one of a set of categories.
2. A variable is called **quantitative** if observations of it take on numerical values that represent different magnitudes of the variable.

- Categorical: gender (Male and Female); race (Chinese, Malay and Indian); type of residence (HDB, Condo, EC, Landed house); etc.
- Quantitative: age, height, weight, blood pressure, etc.

Distinguishing Between Quantitative and Categorical Data

- Simply by asking if there is a meaningful distance between any two points in the sample. If such a distance is meaningful, then you have quantitative data.

Distinguishing Between Quantitative and Categorical Data

- Simply by asking if there is a meaningful distance between any two points in the sample. If such a distance is meaningful, then you have quantitative data.

- E.g, Mr Ali has systolic blood pressure of 120 mm Hg, Mr Ben is of 100 mm Hg, the difference of 20 mm Hg means Mr Ali has higher SBP than Mr Ben.

Distinguishing Between Quantitative and Categorical Data

- Simply by asking if there is a meaningful distance between any two points in the sample. If such a distance is meaningful, then you have quantitative data.
- E.g, Mr Ali has systolic blood pressure of 120 mm Hg, Mr Ben is of 100 mm Hg, the difference of 20 mm Hg means Mr Ali has higher SBP than Mr Ben.
- However, it does not make sense to consider the mathematical operation ("Male" - "Female").

Distinguishing Between Quantitative and Categorical Data

- Simply by asking if there is a meaningful distance between any two points in the sample. If such a distance is meaningful, then you have quantitative data.
- E.g, Mr Ali has systolic blood pressure of 120 mm Hg, Mr Ben is of 100 mm Hg, the difference of 20 mm Hg means Mr Ali has higher SBP than Mr Ben.
- However, it does not make sense to consider the mathematical operation ("Male" - "Female").
- **It is important to identify which type of data you have, as it affects the exploration techniques that you can apply**.

Quantitative Variables

A quantitative variable can be discrete or continuous.

Definition 3 (Discrete and Continuous Variables)

A quantitative variable is **discrete** if its possible values form a set of separate numbers such as 0, 1, 2, 3, ...
A quantitative variable is **continuous** if its possible values form an interval.

Discrete Variables

- Discrete variables are usually countable.

Discrete Variables

- Discrete variables are usually countable.

- Examples: number of pets in a home, the number of children that a couple has, the number of dengue cases in a GRC, etc.

Discrete Variables

- Discrete variables are usually countable.
- Examples: number of pets in a home, the number of children that a couple has, the number of dengue cases in a GRC, etc.
- It is impossible for the number of pets in a home is not a whole number, like 1.7.

Continuous Variables

- Continuous variables have a continuum of infinitely many possible values.

Continuous Variables

- Continuous variables have a continuum of infinitely many possible values.

- Examples: age, height, weight, systolic blood pressure, IQ, income ($), time taken for an event, etc.

Continuous Variables

- Continuous variables have a continuum of infinitely many possible values.
- Examples: age, height, weight, systolic blood pressure, IQ, income ($), time taken for an event, etc.
- It is possible for age of a person be a whole number like 42 (42 years) or a number like 36.42 (36 years and 5 months).

Categorical Variables

A categorical variable has categories which can be nominal or ordinal.

Definition 4 (Nominal and Ordinal Variables)

A categorical variable is **ordinal** if the observations can be ordered, but do not have specific quantitative values.
A categorical variable is **nominal** if the observations can be classified into categories, but the categories have no specific ordering.

Categorical Variables

- Nominal categorical variables or *Nominal random variables*: These consist of readings that cannot be ordered, such as: gender, race, pregnancy status, etc.

Categorical Variables

- Nominal categorical variables or *Nominal random variables*: These consist of readings that cannot be ordered, such as: gender, race, pregnancy status, etc.

- Ordinal categorical variables or *Ordinal random variables*: consist of readings that can be ordered, such as: ratings (good, normal and bad), final position of a competition (first , second, third and no prize), etc.

Can Age be a Categorical Variable?

- Variable age (in a dataset) with values as age of participants is a quantitative, continuous variable.

Can Age be a Categorical Variable?

- Variable age (in a dataset) with values as age of participants is a quantitative, continuous variable.

- Order the values of age, divide into 3 ranges: 25-45, 46-65, more than 65. Count the number of participants in each range.

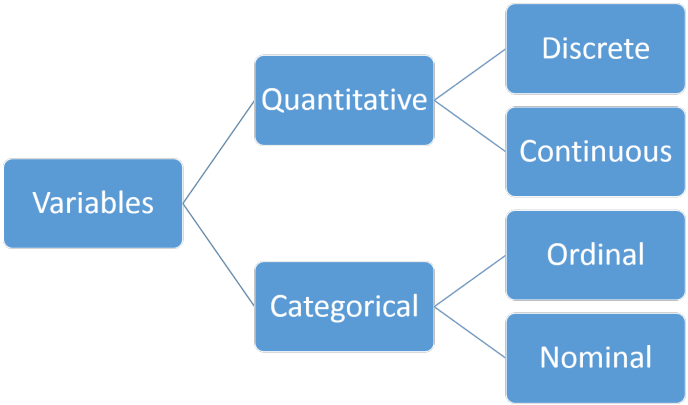
Can Age be a Categorical Variable?

- Variable age (in a dataset) with values as age of participants is a quantitative, continuous variable.
- Order the values of age, divide into 3 ranges: 25-45, 46-65, more than 65. Count the number of participants in each range.
- With the step of "categorizing" above, variable age can be treated as a categorical variable, which has 3 categories ordered.

Can Age be a Categorical Variable?

- Variable age (in a dataset) with values as age of participants is a quantitative, continuous variable.
- Order the values of age, divide into 3 ranges: 25-45, 46-65, more than 65. Count the number of participants in each range.
- With the step of "categorizing" above, variable age can be treated as a categorical variable, which has 3 categories ordered.
- This step can be done for many quantitative variables.

Overview of Types of Variable



1 Introduction

2 Types of Variables

3 Single Variable Exploration
- Frequency Tables
- Graphical Summaries

Summarizing Data with Tables

- Usually, the first step in summarizing data is to **look at the possible values**, and count how often each one occurs.

Summarizing Data with Tables

- Usually, the first step in summarizing data is to **look at the possible values**, and count how often each one occurs.

- For categorical variables, the number of times each category turns up is counted and displayed in a table.

Summarizing Data with Tables

- Usually, the first step in summarizing data is to **look at the possible values**, and count how often each one occurs.
- For categorical variables, the number of times each category turns up is counted and displayed in a table.
- The category with the highest frequency is the **modal category**.

Frequency Tables

1. A **frequency table** is a listing of possible values, together with the frequency of each value.

Frequency Tables

1. A **frequency table** is a listing of possible values, together with the frequency of each value.

2. The **proportion** of observations in a certain category is the count of observations in that category divided by the total number of observations.

Frequency Tables

1. A **frequency table** is a listing of possible values, together with the frequency of each value.

2. The **proportion** of observations in a certain category is the count of observations in that category divided by the total number of observations.

3. The **percentage** is the proportion multiplied by 100.

Frequency Tables

1. A **frequency table** is a listing of possible values, together with the frequency of each value.

2. The **proportion** of observations in a certain category is the count of observations in that category divided by the total number of observations.

3. The **percentage** is the proportion multiplied by 100.

4. Proportions and percentages are also known as **relative frequencies**.

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.
- Description of the variables in this study:

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.
- Description of the variables in this study:

  Age: year

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.
- Description of the variables in this study:

  Age: year

  Smoke: 1 = yes and 0 = no

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.
- Description of the variables in this study:

  Age: year

  Smoke: 1 = yes and 0 = no

  Cancer: 1 = yes and 0 = no

Lung Cancer Data

- A study to investigate on the effect of smoking on the incidence of lung cancer.
- 17 lung cancer patients were randomly sampled; and 17 healthy people were randomly sampled.
- Description of the variables in this study:

  Age: year

  Smoke: 1 = yes and 0 = no

  Cancer: 1 = yes and 0 = no

  Gender: 1 = male and 0 = female

Frequency Tables for Categorical Variable

Table summary of Gender in frequency count; in proportion and in percentage. The modal category is "Male".

```
gender
Female   Male
    16     18
```

```
gender
   Female      Male
0.4705882 0.5294118
```

```
gender
  Female     Male
47.05882 52.94118
```

Frequency Tables for Categorical Variable

Table summary of Gender in frequency count; in proportion and in percentage. The modal category is "Male".

```
gender
Female   Male
    16     18
```

```
gender
   Female      Male
0.4705882 0.5294118
```

```
gender
  Female     Male
47.05882 52.94118
```

The table summary of Cancer

```
cancer
 No Yes
 17  17
```

```
cancer
 No Yes
0.5 0.5
```

```
cancer
 No Yes
 50  50
```

Frequency Table for Quantitative Variable

- Consider variable Age.
- Categorize age into 3 ranges: 25-45, 46-65 and older than 65. The modal category is $(46, 65)$.

```
age
       25-45        46-65 more than 65
          10          14              10

age
       25-45        46-65 more than 65
   0.2941176    0.4117647    0.2941176

age
       25-45        46-65 more than 65
    29.41176     41.17647     29.41176
```

What to Say

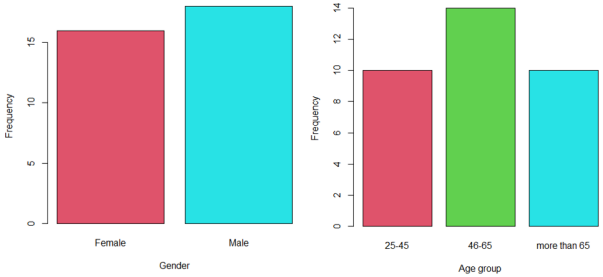
When you are asked to summarize a frequency table, be sure to mention:

- The modal category.
- The proportion or percentage for modal category.

Bar Plots

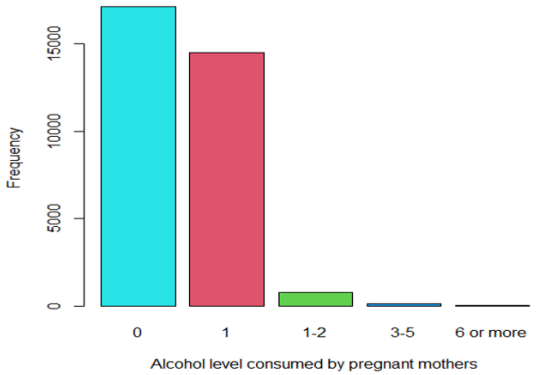
- Bar plot is a common way to display a **single categorical variable**.
- It consists of a vertical bar for each category, with the height proportional to the frequency of that category.

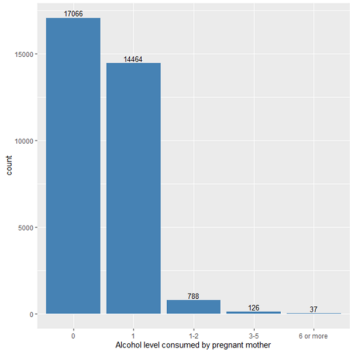
Bar plots for Gender and Age in Example 1.



Bar Plots

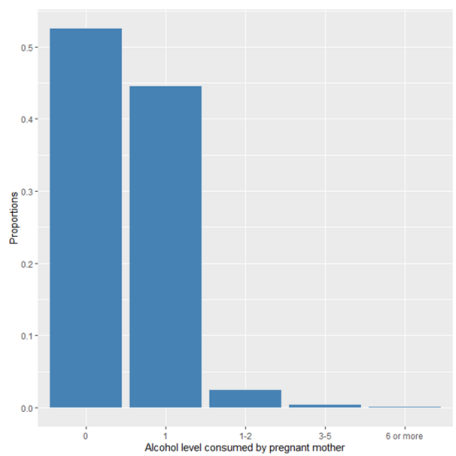
Variable: **level of alcohol** consumed by pregnant mothers (higher level means consuming more alcohol)

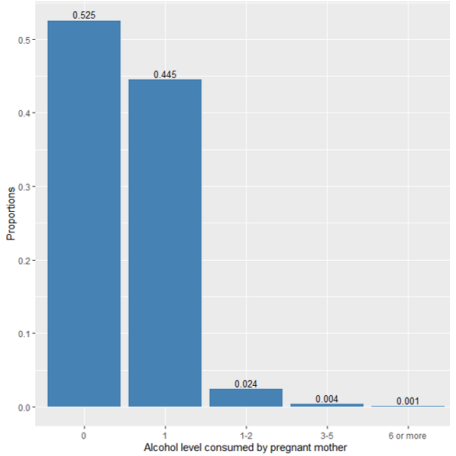




The modal category is "0" (no alcohol) and followed by "1". Gradually less pregnant mothers consume high level of alcohol.

Bar Plots


Proportions
0.5
0.4
0.3
0.2
0.1
0.0
0
1
1-2
3-5
6 or more
Alcohol level consumed by pregnant mother


0.525
0.445
0.024
0.004
0.001
Proportions
0.5
0.4
0.3
0.2
0.1
0.0
0
1
1-2
3-5
6 or more
Alcohol level consumed by pregnant mother

What to Say

When you are asked to summarize a bar plot, mention the same points as for a frequency table. In addition,

- Mention if there are groups of categories with high/low proportions. Such a pattern is usually easier to see visually.
- If there is an ordering to the categories, mention if there is any apparent trend in proportions.

Histograms

Definition 5 (Histogram)

A **histogram** is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.

Histograms

Definition 5 (Histogram)

A **histogram** is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.

Histograms

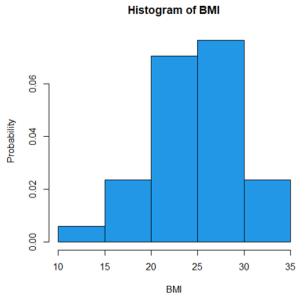
Definition 5 (Histogram)

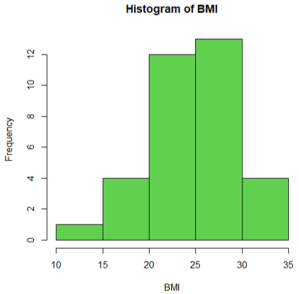
A **histogram** is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.

1. Divide the range of the data into intervals of equal width.
2. Count the number of observations that fall within each interval, forming a frequency table.
3. Label the intervals on the x-axis, and draw a bar over each interval, that has height equal to its frequency or relative frequency.

Histograms for a BMI Variable

Body mass index (BMI) of 34 people.

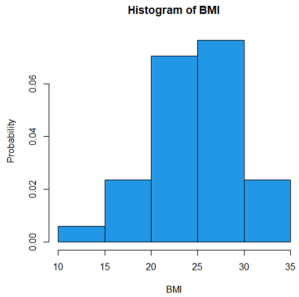


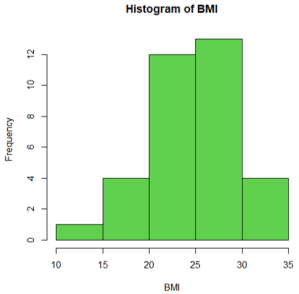


- From right histogram: one person has BMI from 10 to 15; 4 people have BMI 15 to 20; etc.

Histograms for a BMI Variable

Body mass index (BMI) of 34 people.





- From right histogram: one person has BMI from 10 to 15; 4 people have BMI 15 to 20; etc.
- Left histogram was created using probability (or density).

What to Say

What do we look for in a histogram?

- The overall pattern. Data cluster together, or there is a gap such that one or more observations deviate from the rest. Any suspected outliers?

What to Say

What do we look for in a histogram?

- The overall pattern. Data cluster together, or there is a gap such that one or more observations deviate from the rest. Any suspected outliers?

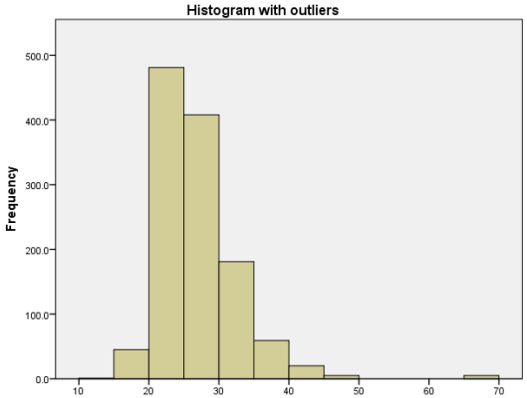
- Do the data have a single mound? This is known as a unimodal distribution. Data with two or more are known as bimodal or multimodal distribution.

What to Say

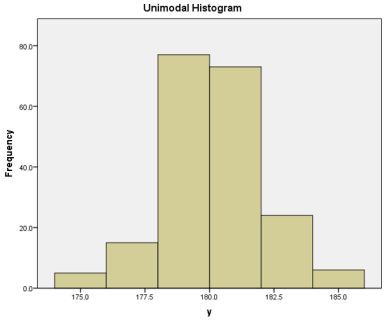
What do we look for in a histogram?

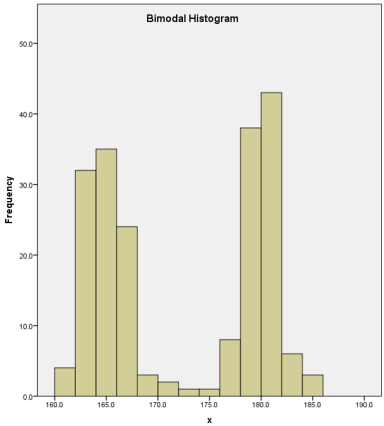
- The overall pattern. Data cluster together, or there is a gap such that one or more observations deviate from the rest. Any suspected outliers?
- Do the data have a single mound? This is known as a unimodal distribution. Data with two or more are known as bimodal or multimodal distribution.
- Is the distribution symmetric or skewed?

A Histogram With Outliers



Unimodal and Bimodal Histograms





Skewness

Definition 6 (Skew)

- To **skew** is to pull in one direction.

Skewness

Definition 6 (Skew)

- To **skew** is to pull in one direction.
- A distribution is **skewed to the left** if the left tail is longer than the right tail.

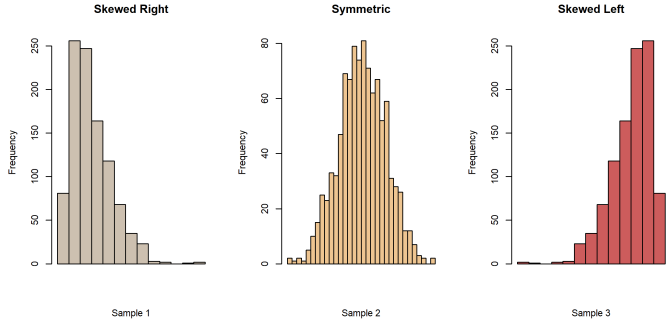
Skewness

Definition 6 (Skew)

- To **skew** is to pull in one direction.
- A distribution is **skewed to the left** if the left tail is longer than the right tail.
- A distribution is **skewed to the right** if the right tail is longer than the left tail.

Note: When describing the tail of a histogram or a distribution, some terminologies are used with same meaning: longer = thicker = heavier while shorter = thinner = lighter.

Skewed Data



Some examples: Income is typically right-skewed; IQ is typically symmetric; Life-span is typically left-skewed, etc.

Summary

We have learnt

- Variable and types of variables
- How to explore a single variable using frequency table, bar plot (for categorical variables) and histogram (for quantitative variables).
- Points to note when analyzing the outputs mentioned above.

THANK YOU!