2.3 Tabular Presentations of Data

he sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

Data source: https://www.kaggle.com/varimp/a-mostly-tidyverse-tour-of-the-titanic

Here's a quick summary of our variables:

Variable Name	Description
PassengerID	Passenger ID (just a row number, so obviously not useful for prediction)
Survived	Survived (1) or died (0)
Pclass	Passenger class (first, second or third)
Name	Passenger name
Gender	Passenger Gender
Age	Passenger age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare
Cabin	Cabin
Embarked	Port of embarkation (S = Southampton, C = Cherbourg, Q = Queenstown)

Raw Data

- Raw data are collected data that have not been organized numerically
- Eg: Passenger age

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Moran, Mr. James	male	NA

Raw data: Age

```
##  [1] 22 38 26 35 35 NA 54  2 27 14  4 58 20 39 14 55  2 NA 31 NA 35 34 15 28  8
## [26] 38 NA 19 NA NA 40 NA NA 66 28 42 NA 21 18 14
```

An array

- An array is an arrangement of raw numerical data in ascending or descending order of magnitude.
- Eg: Passenger age

```
##  [1]  2  2  4  8 14 14 14 15 18 19 20 21 22 26 27 28 28 31 34 35 35 35 38 38 39
## [26] 40 42 54 55 58 66
```

Frequency Table (Frequency Distributions)

- A frequency table (frequency distribution) is a listing of the values a variable takes in a data set, along with how often (frequently) each value occurs
- frequency can be recorded as a
  - **frequency or count:** the number of times a value occurs, or
  - **percentage frequency:** the percentage of times a value occurs

- Percentage frequency can be calculated as,

$$Percentagefrequency = \frac{a}{b} \times 100\%$$

- The objective of constructing a frequency table are as follows
  - to organize the data in a meaningful manner
  - to determine the nature or shape of the distribution
  - to draw charts and graphs for the presentation of data
  - to facilitate computational procedures for measures of average and spread
  - to make comparisons between different data sets
- There are two basic types of frequency tables
  1. Simple frequency tables (Ungrouped frequency distribution)
  2. Grouped frequency distribution

2.3.0.1 Simple frequency table (Ungrouped frequency distribution)

- Each possible value or category is taken as a class
- More suitable for
  - Qualitative variables
  - Discrete variables
- Sometimes construct for continuous variables when there is a small number of possible values between the minimum and maximum.

Examples:

CASE I:

Example 1

The native countries of 56 students from a certain education institute are as follows:

```
##  [1] "SL" "BD" "SL" "SL" "SL" "SL" "IN" "SL" "SL" "SL" "BD" "SL" "SL" "SL" "IN"
## [16] "SL" "SL" "BD" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "MD" "SL" "SL"
## [31] "SL" "SL" "SL" "SL" "PK" "MD" "PK" "SL" "SL" "SL" "SL" "SL" "PK" "MD" "SL"
## [46] "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "SL" "MD" "MD"
```

BD- Bangladesh, IN-India, MD-Maldives, PK-Pakistan, SL- Sri Lanka

Construct a frequency table

Native Country	Count	Percentage (%)
Bangladesh	3	5.357
India	2	3.571
Maldives	5	8.929
Pakistan	3	5.357
Sri Lanka	43	76.786
Total	56	100.000

CASE II:

Example 2

The grades of 30 students for Statistics are as follows:

```
##  [1] "B" "C" "B" "D" "B" "C" "C" "A" "B" "C" "C" "B" "E" "B" "B" "D" "D" "F" "B"
## [20] "D" "D" "A" "B" "A" "B" "C" "E" "A" "A"
```

Construct a frequency table

Grade	Count	Percentage (%)
A	5	17.241
B	10	34.483
C	6	20.690
D	5	17.241
E	2	6.897
F	1	3.448
Total	29	100.000

CASE III:

Example 3

The number of family members of a sample of undergraduates of Batch 19 are as follows:

```
##  [1] 7 5 3 4 5 4 3 6 4 4 5 2 7 4 5 6 4 4 3 5
```

Construct a frequency table

Number of family members	Count	Percentage (%)
2	1	5
3	3	15
4	7	35
5	5	25
6	2	10
7	2	10

CASE IV:

Example 4

The ages (in years) of a sample of undergraduates of Batch 19 are as follows:

```
##  [1] 21 22 22 23 22 24 24 23 21 22 23 22 22 23 21 21 22 23 22 23
```

Construct a frequency table

Age (years)	Count	Percentage (%)
21	4	20
22	8	40
23	6	30
24	2	10

2.3.1 Grouped frequency distribution

- A grouped frequency distribution (table) is obtained by constructing classes (or intervals) for the data and then listing the corresponding number of values in each interval.
- Suitable for quantitative variables with large number of possible values in the range of data.
- Note that when items have been grouped in this way, their individual values are lost.
- When studying about frequency distributions it is very important to know the meaning of the following terms

i Class intervals

- In a frequency distribution the total range of the observations are divided into a number of classes. Those are called *class intervals*
- Eg: Class intervals: 10-14, 15-19, 20-24, . . . , 40-44

ii Class limits

- Class limits are the smallest and largest piece of data value that can fall into a given class.
- In the class interval 10-14, the end numbers, 10 and 14, are called class limits
- The smaller number (10) is the *lower class limit*
- The larger number (14) is the *upper class limit*

iii Class boundaries

- Class boundaries are obtained by adding the upper limit of one class interval to the lower limit of the next-higher class interval and dividing by 2.
- Class boundaries are also called **True class limits**

- Class boundaries **should not** *coincide with actual observations*

Class interval	Class boundaries
10 - 14	9.5 – 14.5
15 - 19	14.5 – 19.5
20 - 24	19.5 – 24.5
25 - 29	24.5 – 29.5
30 - 34	29.5 – 34.5
35 - 39	34.5 – 39.5
40 - 44	39.5 – 44.5

iv The size or width of a class interval

- The size or width of a class interval is the difference between the *lower and upper class boundaries*
- It is also referred to as the *class width, class size, or class length*
- Eg: The class width for the class 10-14 is = 14.5-9.5 = 5

v The class mark (Midpoint of the class)

- Midpoint of the class
- Also called as *class midpoint*
- Midpoint of the class = $\frac{\text{Lower limit+Upper limit}}{2}$

or

- Midpoint of the class = $\frac{\text{Lower boundary+Upper boundary}}{2}$

vi Open class intervals

- A class interval that, at least theoretically, has either no upper class limit or no lower class limit indicated is called an *open class interval*

- For example, referring to age groups of individuals, the class interval “65 year and over” is an open class interval

Rules and Practices for constructing grouped frequency tables

- Every data value should be in an interval
- The intervals should be mutually exclusive
- The classes of the distribution must be arrayed in size order.

- The number of classes not less than 5 or not greater than 15 is recommended.
- The following formula is often used to determine the number of classes: If n is the number of observations, then

$$\text{Number of classes} = \sqrt{n}$$

$$\text{Width of the class interval} = \frac{Range}{\sqrt{n}} = \frac{Min - Max}{\sqrt{n}}$$

- Data should be represented within classes having limits which the data can attain
- Classes should be continuous
- By convention, the beginning of the interval is given the appropriate exact value, rather than the end.
  Eg: intervals of 0-49, 50-99,100-149 would be preferred over the intervals 1-50, 51-100, 101-150 etc.
- The number f observations falling into each category or class interval (class frequency) can be easily found using *tally marks*.

Examples:

In a grouped frequency distribution, class intervals can be constructed in different ways

Example 1

Class interval	Number of students
10 - 14	4
15 - 19	5
20 - 24	11
25 - 29	9
30 - 34	6
35 - 39	3
40 - 44	2

2.3.2 Two-way frequency table

- Cross tabulation, Cross classification table, Contingency table, Two-way table
- Display the relationship between two or more qualitative variables (categorical variables (nominal or ordinal))

Two-way frequency table

Survived	First	Second	Third
died	80	97	372
Survived	136	87	119

Column %

Survived	First	Second	Third
died	0.37	0.53	0.76
Survived	0.63	0.47	0.24

Row %

Survived	First	Second	Third
died	0.15	0.18	0.68
Survived	0.40	0.25	0.35

Total %

Survived	First	Second	Third
died	0.09	0.11	0.42
Survived	0.15	0.10	0.13