Statistical Methods I

B.Stat. 1st Year

What is Statistics?
Statistics is the study of the collection, organization, analysis and interpretation of data.

Objective: To make decisions in the presence of uncertainty based on information contained in the data.

Some Statistical Problems:

- A market analyst wants to know the effectiveness of a new diet.
- A pharmaceutical company wants to know if a new drug is superior to already existing drugs, or possible side effects.
- A fisherman wants to know the total amount of fishes in a lake.

Reference books:

- F.E. Croxton and D.J. Cowden: Applied General Statistics.
- A.M. Goon, M.K. Gupta and B. Dasgupta. Fundamentals of Statistics, Vol. 1.
- D. Freedman, R. Pisani and R. Purves: Statistics.
- M.G. Kendall and A. Stuart: The Advanced Theory of Statistics, Vol. 1 and 2.
- J.F. Kenney and E.S. Keeping: Mathematics of Statistics.
- G.W. Snedecor and W.G. Cochran: Statistical Methods.
- M. Tanner: An Investigation for a Course in Statistics.
- J.M. Tanur (ed.): Statistics: A Guide to the Unknown.
- J.W. Tukey: Exploratory Data Analysis.
- W.A. Wallis and H.V. Roberts: Statistics: A New Approach.
- G.U. Yule and M.G. Kendall: An Introduction to the Theory of Statistics.

There are two branches of statistics:

1. **Descriptive Statistics:** Quantitatively describing the main features of data, or the quantitative description itself; *e.g.,* average or range of data, representation of data via plots and tables.
2. **Inferential Statistics:** Drawing conclusions from data; *e.g.,* drug A is better than drug B.

Some Definitions:

- **Population:** A population is a collection of all units of interest; *e.g.,* all fishes in a lake.
- **Variable:** A measurable property or attribute associated with each unit of a population; *e.g.,* weight and hight of a fish.
- **Random Variable:** A variable whose value is subject to variations due to chance; *e.g.,* weight of a fish caught randomly from a lake.
- **Parameter:** A common characteristic of the population that is being studied; *e.g.,* average weight of all fishes in a lake.

  Enumerating big population is difficult. We generally get "estimates" of the population parameters based on an randomly chosen subset of population (called a sample).
- **Sample:** A subset of a population that is actually observed; *e.g.,* 20 fishes which are caught in a particular day.
- **Observation:** Values of all variables for an individual unit in the sample; *e.g.,* weight 2.5 kg, height 30 cm.
- **Statistic:** (not to be confused with Statistics) Characteristic or measure obtained from a sample; *e.g.,* average weight of fishes which are caught.

There are two types of variables:

1. **Qualitative Variables:** Variables which assume non-numerical values.
   (a) **Nominal Levels:** Categories without rank and order; *e.g.,* gender, mother tongue.
   (b) **Ordinal Levels:** Categories that can be ranked; *e.g.,* opinion for reservation – agree, neutral, disagree.
2. **Quantitative Variables:** Variables which assume numerical values.
   (a) **Discrete:** *e.g.,* number of children, number of accident.
   (b) **Continuous:** *e.g.,* age, income, weight and height.

Sometimes quantitative variables are categorized so that the response is presented in terms of a qualitative variables, although the original variable is essentially quantitative. For example, incomes are categorized as – high income, middle income and low income.

1 Data Collection: Randomized Study vs. Observational Study

Suppose we want to study the possible effect of a treatment over two groups of subjects – group I and group II. In *observational study* we draw inference, where the assignment of subjects into group I or group II is outside the control of the investigator. But in *randomized study* each subject is randomly assigned to group I or group II. Consider the following examples:

1. We want to determine whether an experimental drug is useful in reducing blood pressure.
2. We want to determine whether males have higher rate of cardiac related problem than females.

In the first example we conduct the following steps:

- Collect two independent random samples of individuals of size n_1 and n_2, respectively, from the population of interest.
- Administer the experimental drug to group I (all n_1 individuals), and keep group II (all n_2 individuals) in control state.
- Take blood pressure reduction reading for each individuals.

Here in the first example the treatment under study (drug) is randomly assigned, so it is a randomized study. But in the second example the treatments cannot be randomly assigned, so it is called observational study. Sometimes randomized study is not possible due to various reasons, such as lack of power or an ethical cause.

1.1 How large is "large" enough?

Larger sample sizes generally lead to increased precision and power when we make inference about unknown parameters. In practice the sample size is determined based on the desired confidence level or maximum tolerable margin of error. The confidence level tells us how sure we can be about the information provided by the sample. It is expressed in the form of a percentage, and indicates the reliability of a given estimate or statistic. The standard for most studies the confidence interval is 95%. This means we can be 95% certain that the sample size reflects the total population. In other words, we can tolerate 5% error in prediction.

A manager of a store claims that 10% of his male employees are regularly late for duty, whereas 50% of his female employees are regularly late. It seems that there is a huge difference between male and female. But the store has 40 male employees and only 2 female employees. Is the sample size large enough to make this claim?

2 Descriptive statistics

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data. It aims to summarize a data set to describe patterns and general trends. Statistics can be viewed as a means of finding order and meaning in apparent chaos. At the end of the data collection phase, what we get is a bunch of numbers with no apparent order or meaning. The first phase of data analysis involves the placing of some order on that chaos. Typically the data are reduced down to one or two descriptive summaries like the mean and standard deviation or correlation, or by visualization of the data through various graphical procedures like histograms, frequency distributions, and scatter plots.

2.1 Data Visualization

2.1.1 Frequency Distribution

Frequency distribution is a tabular representation organized to show the number of observations of each possible outcome or within a given interval.

Example 1. These are the numbers of newspapers sold at a local shop over the last 75 days (row by row):

Table 1: News paper sell data.

22	24	25	23	23	23	23	24	24	25	22	23	22	25	23
22	24	25	23	23	23	23	23	23	22	24	24	25	24	22
24	23	23	24	24	24	25	24	21	23	23	24	23	25	23
24	25	22	25	23	23	23	23	24	23	24	26	22	24	25
23	24	23	23	22	23	22	24	25	22	23	22	24	25	22

The minimum and the maximum number of newspapers sold are 21 and 26 respectively. Table 2 represents the frequency distribution of news paper sell data[1].

It is also possible to group the values. Here they are grouped in three intervals, however, Table 3 shows that the groups are too small for this example. See Section 2.1.5 for a better example.

[1] V should be replaced by //// and a diagonal cross line.

Table 2: Frequency distribution of news paper sell[1].

Papers Sold	Tally Mark	Frequency	Relative Frequency
21	/	1	0.0133
22	V V ///	13	0.1733
23	V V V V V ///	28	0.3733
24	V V V V	20	0.2667
25	V V //	12	0.1600
26	/	1	0.0133
Total	–	75	1

Table 3: Grouped frequency distribution of news paper sell.

Class Limit	Class Boundary	Freq.	Relative Frequency	Cumulative Frequency
21–22	20.5–22.5	14	0.1867	14
23–24	22.5–24.5	48	0.6400	62
25–26	24.5–26.5	13	0.1733	75
Total	–	75	1	–

For the grouped frequency distribution of quantitative variables the following rules should be maintained.

- The number of cases should not be too large or too small (relative to the sample size).
- The lower boundary of the first class must be smaller than the smallest observation. The upper boundary of the last class must be larger than the largest observation.
- The classes should be non-overlapping.
- The classes should be exhaustive over the range of the data.
- The number of classes should increase with the sample size.

Procedure for constructing a grouped frequency distribution:

1. Decide on the number of classes you want (genderally 5 to 20 classes).
2. Calculate the class width.
   Class width = Range / number of classes,
   where Range = maximum value – minimum value.
   Round up the class width to get a convenient number.

3. Choose a number for the lower limit of the first class and construct all class limits.

4. Tally the frequency for each class.

Some Definitions:

- **Class Boundary:** Closing the gap between one class to the next class. The class limits should have the same decimal value as the data, but the class boundaries have an additional place value and end with a 5.

  e.g., if data are whole numbers

  lower class boundary = lower class limit – 0.5,

  Upper class boundary = upper class limit + 0.5;

  e.g., if data are one decimal place

  lower class boundary = lower class limit – 0.05,

  Upper class boundary = upper class limit + 0.05;

  e.g., if data are two decimal places

  lower class boundary = lower class limit – 0.005,

  Upper class boundary = upper class limit + 0.005.

- **Class Mark:** The midpoint of each class.

  Class Mark = (lower class limit + upper class limit) / 2.

- **Cumulative Frequency:** The sum of the frequencies accumulated up to the upper boundary of a class. This is also called less than type cumulative frequency. The greater than type cumulative frequency sums up the frequencies greater than the lower boundary of a class.

- **Relative Frequency:** The frequency of each class divided by the total number of sample.

  Relative frequency = frequency / total sample size.

R code

```
> x=c(22, 24, 25, ... 22, 25, 22)
> table(x)
x
21  22  23  24  25  26
 1  13  28  20  12   1

> table(cut(x,c(20,22,24,26)))

(20,22] (22,24] (24,26]
     14      48      13
```

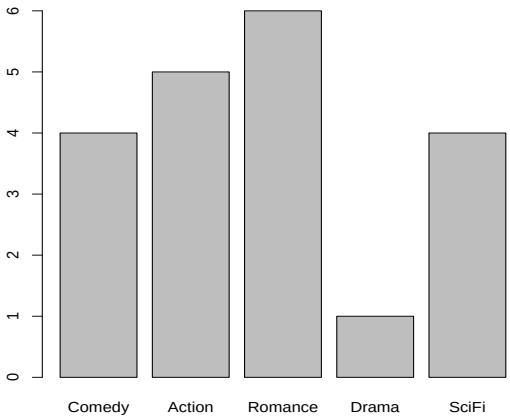

Figure 1: Bar diagram showing preference of different types of movies.

2.1.2 Bar Diagram:

A special graph that uses bars to show relative sizes of data. The bars can be plotted vertically or horizontally. One axis of the graph shows the specific categories being compared, and the other axis represents the values or levels of the variable.

Example 2. Imagine you just did a survey among your friends to find which kind of movies they like most. Here are the results:

Table 4: Data on favorite movie.

Comedy	Action	Romance	Drama	SciFi	Total
4	5	6	1	4	20

You may present your data on a bar graph as in Figure 1. It is a really good way to show relative sizes. It is easy to see, at a glance, which types of movies are most liked, and which are least liked. You can use bar graphs to show the relative sizes of many things, such as what type of car people have, how many customers a shop has on different days and so on.

R code

```
> x=c(4, 5, 6, 1, 4)
> barplot(x) # basic command
> barplot(x,names.arg=c('Comedy', 'Action','Romance','Drama','SciFi'))
```

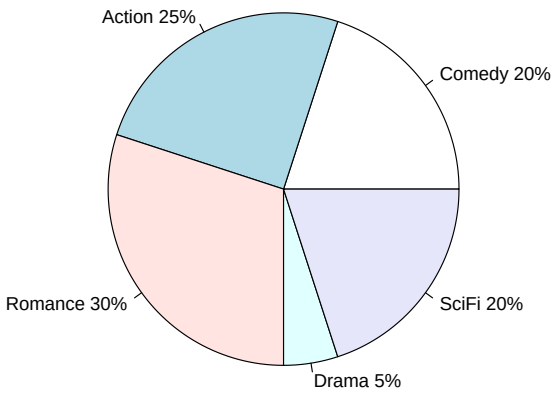

Figure 2: Pie chart showing relative preference of different types of movies.

2.1.3 Pie Chart

A special chart that uses "pie slices" to show relative sizes of data. It illustrates numerical proportion of different categories under study.

Let us consider the data about the favorite movies given in Table 4. The percentage of different types of favorite movies are calculated below:

Table 5: Data on favorite movie.

Comedy	Action	Romance	Drama	SciFi	Total
4	5	6	1	4	20
20%	25%	30%	5%	20%	100%

In this case we can use pie chart to show percentage clearly (see Figure 2).

R code

```
> x=c(4, 5, 6, 1, 4)
> pie(x) # basic command
> pie(x,labels=c('Comedy', 'Action','Romance','Drama','SciFi'))
```

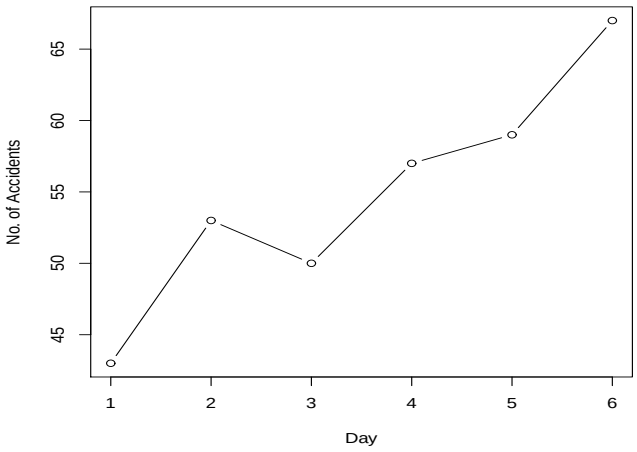

Figure 3: Line graph showing the number of accidents in West Bengal in 6 days.

```
> # displays both name and percentage
> prct <- round(x/sum(x)*100) # percentage
> lbls = c('Comedy', 'Action','Romance','Drama','SciFi')
> lbls <- paste(lbls, prct) # add percents to labels
> lbls <- paste(lbls,"%",sep="")   # add % to labels
> pie(x,labels = lbls)
```

2.1.4 Line Graph

A line chart or line graph displays information as a series of data points connected by straight line segments. It shows information that is connected in some way (such as change over time). A line chart is often used to visualize a trend in data over intervals of time, so the line is generally drawn chronologically.

Example 3. Table 6 shows number of accidents in West Bengal, recorded in 6 consecutive days. We can show the trend of accident using the line graph in Figure 3.

R code

```
> x=c(43, 53, 50, 57, 59, 67)
> plot(x, type='b') # basic command
> plot(x, type='b', xlab='Day', ylab='No. of Accidents')
```

Table 6: Data showing number of accidents in West Bengal.

Day	No. of Accidents
1	43
2	53
3	50
4	57
5	59
6	67
Total	329

2.1.5 Histogram

A histogram is a graphical representation, similar to a bar chart, that organizes a group of data points into user specified ranges. Histogram presents bars on the X-axis with the base being the class interval and the height being propotional to the class frequency. Sometimes we standardize the histogram so that the area under bars is one.

Histograms are commonly used in statistics to demonstrate how many of a certain type of variable occurs within a specific range. For example, in a census on the demography of a city we may use a histogram to know how many people are there between the ages of 0 and 10, 11 and 20, 21 and 30, 31 and 40, 41 and 50 etc.

Example 4. Suppose we had a survey to check the weight of fish in a lake, and we caught forty fish. The data are given below in Table 7.

Table 7: Weight (in kg.) of fish.

8.76	7.98	8.53	6.53	10.78	7.84	7.41	6.80	9.48	8.35
9.82	6.08	7.21	11.18	7.14	6.82	6.81	6.10	7.31	11.58
6.17	5.87	10.12	9.08	5.32	9.50	8.01	8.27	8.21	7.29
7.58	10.37	8.97	11.46	8.36	11.48	9.83	8.37	7.64	5.87

If you examine the data above closely, you will see that there are a lot of mostly different numbers. It would not make much sense to make a bar graph showing the frequency of each individual mass, since most of the bars would just be one unit high, and there would be a lot of bars. Instead we group the data and the histogram is plotted in Figure 4. Grouping means to count how many fishes are there in different weight categories. For example, how many fishes are there between 6 and 7 kilograms? The answer is seven. They are 6.53, 6.80, 6.08, 6.82, 6.81, 6.10 and 6.17 kg. The grouped frequencies are given in Table 8 and the histogram is presented in Figure 4

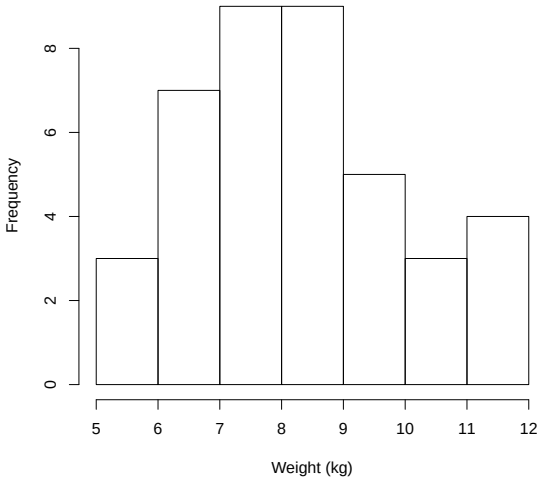

Figure 4: Histogram showing weight of fish.

R code

```
> x = c(8.76, 7.98 ... 7.64, 5.87)
> hist(x)
> hist(x, main = '',xlab='Weight (kg)')
```

2.1.6 Frequency Polygon

A graph that displays data by using lines that connect points plotted for the frequencies at the class marks. They serve the same purpose as histograms, but are especially helpful for comparing different sets of data.

Steps to draw a frequency polygon:

1. Choose suitable class intervals as we do in case of grouped frequency distribution.
2. Calculate frequencies for each class.
3. Prepare a graph where X axis represents the value of the variable and Y axis represents the frequency. Place a point in each class mark at the height corresponding to its frequency.
4. Connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X axis on both sides.

Table 8: Grouped frequency distribution of the weight of fishes.

Class Interval	Class Mark	Freq.	Relative Frequency	Cum. Freq. (< type)	Cum. Freq. (> type)
5–6	5.5	3	0.075	3	40
6–7	6.5	7	0.175	10	37
7–8	7.5	9	0.225	19	30
8–9	8.5	9	0.225	28	21
9–10	9.5	5	0.125	33	12
10–11	10.5	3	0.075	36	7
11–12	11.5	4	0.100	40	4
Total	–	40	1	–	–

R code

```
> # read data from file and convert to a matrix
> data=as.matrix(read.table('hist_data.txt'))
> min_value = min(data)
> max_value = max(data)
> x_value = c(min_value, 5.5:11.5, max_value)
> # convert the frequency table to a vector
> freq = as.vector(table(cut(x,5:12)))
> y_value = c(0, freq, 0) # both ends meet at X axis
> plot(x_value, y_value, type='b', xlab='Mass (kg)', ylab='Frequency')
```

2.1.7 Ogive

A line graph that represents the cumulative frequencies for the classes in a frequency distribution. Ogive can be used to determine how many data values lie above or below a particular value in a data set. The less than type cumulative frequency is calculated from a frequency table, by adding each frequency to the total of the frequencies of all data values before it in the data set. The last value for the cumulative frequency will always be equal to the total number of data values, since all frequencies will already have been added to the previous total.

Steps to draw a ogive for the less than type cumulative frequency:

1. Calculate the less than type cumulative frequencies.
2. Place a point in each upper class limit at the height corresponding to its (less than type) cumulative frequency.
3. Connect the points. Join the graph with the X at the lower limit of the first class.

The ogive for the greater than type cumulative frequency is calculated in a similar way.

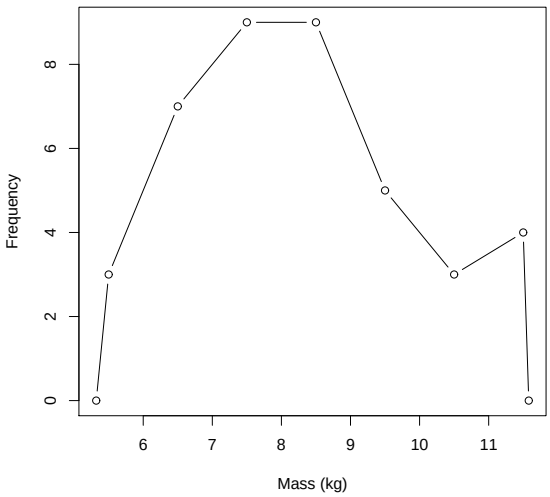

Figure 5: Frequency polygon showing weight of fish.

2.1.8 Scatter Plots

A scatter plot is a graph of plotted points that shows the relationship between two sets of data. It uses Cartesian (X, Y) coordinates to display values of two variables for a set of data. The data set is displayed as a collection of points, each having the values of one variable plotted on X axis and the values of the other variable plotted on Y axis.

The relationship between two variables is called "correlation" (mathematical definition will be given later). The closer the data points with respect to a straight line, the higher the correlation between the two variables, or the stronger the relationship. If the pattern of dots slopes from lower left to upper right, it suggests a positive correlation between two variables. Similarly if the pattern of dots slopes from upper left to lower right, it suggests a negative correlation. There might not have any notable association, i.e, if the plotted points do not indicate any trends whatsoever, that means the two variables are uncorrelated.

A scatter plot is also used when a variable is under the control of the experimenter which is called independent variable. It is plotted along X axis. The measured or dependent variable is customarily plotted along Y axis. If no dependent variable exists (*e.g.,* height and weight), either type of variable can be plotted on either axis.

Example 5. Suppose we have height and weight data for eleven students in a class. The observations are given in the following table.

Table 9: Data showing weight of students.

Height (cm)	170	165	172	166	178	177	167	170	175	173	169
Weight (kg)	66	62	67	59	74	77	66	70	75	72	70

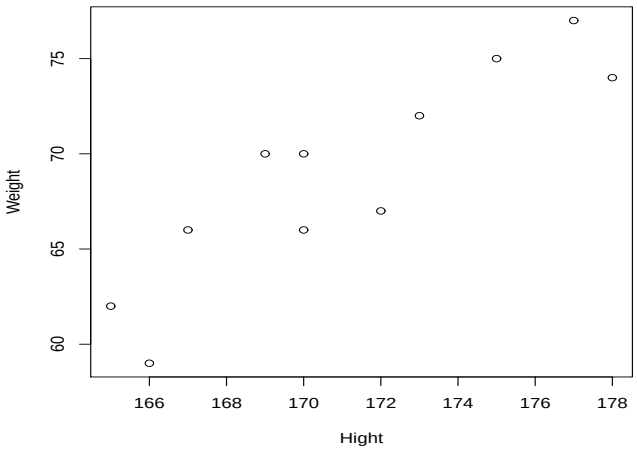

Figure 6: Scatter plot showing height and weight 11 students.

The data are plotted in Figure 6. Here the scatter plot shows a positive correlation between height and weight of the students.

R code

```
> x = c(170, 165, 172, 166, 178, 177, 167, 170, 175, 173, 169)
> y = c(66, 62, 67, 59, 74, 77, 66, 70, 75, 72, 70)
> plot(x,y) # basic command
> plot(x,y, xlab='Hight', ylab='Weight')
```