Contents

Chapter 1

Introduction

1.1 What is Statistics ? Why do we need it ?

In everyday life, we always face scenarios (that matters to us) and want to make sense of those - this often boils down to asking ourselves a question and seeking an answer that satisfies it. In order to do that, we have to analyze the available information in an objective and appropriate manner. For example, we may be interested in the following questions

- Does excessive use of social media makes a person more lonely/isolated ?
- Does smoking causes lung cancer ?
- Are astrology predictions better than mere guessing ?
- How likely are you to win a lottery ?
- Does Aspirin reduces the chance of heart attack ?

Statistics is probably the only scientific discipline, that helps us to answer questions like those above in a data-based, scientific manner.

Every statistical study is motivated by a question (like one of the above) that directly relates to reality. In order to answer that question satisfactorily, we have to design a suitable study (we will learn how to do this later on). The study produces data (or information). Statistical techniques are then applied to make sense of the data and the conclusions drawn (by using those statistical methods on the dataset) is (are) the answer(s) to the question we started off with.

Definition: *Statistics* is the art and science of designing studies, analyzing the data those studies produce and drawing objective conclusions based on the analysis. Its ultimate goal is to translate **data** into **knowledge** that would better help us in understanding the world around us. Thus, in a nutshell, **Statistics is the art and science of learning from data.**

1.2 Statistical Enquiry

Every statistical enquiry has four distinct stages viz :

1. **The Question** : Here you should give careful thought to the question that needs to be answered by the study. This is because, *the exact nature of the question influence all the later stages of the study.*
2. **Design** : This deals with formulating a proper and realistic plan or experiment which would generate the data we want and thus would lead us to answer the question we started off with.
3. **Description** : This deals with exploring and summarizing patterns in the raw data obtained from an experiment.
4. **Inference** : This deals with making predictions and decisions based on the above data but in such a way that *the conclusions are applicable to the whole population, not merely those in the dataset.*

Hopefully, these stages will give you a satisfactory answer to the question you started off with !

1.3 Terms to Remember

Following are some important terms/concepts that every statistician should be familiar with.

- **Variable :** It is any characteristic that is observed for subjects/units in a statistical study. As the term suggest, it is something that *varies* across the different subjects.
  Eg : If we want to know the living standard of average Indians, we may select a representative sample of Indians and record their (i) monthly income, (ii) number of dependent family members (iii) monthly expenditure on food and other necessities etc. All these are variables for this study.
- **Observations** : The actual data values we observe for a variable (or a particular subject) are referred to as the observations for that variable (or that particular subject).
  Eg : Suppose the number of dependent family members of a sample of 5 Indians are $(2, 3, 5, 1, 0)$ - these are the observations (or observed values) for this variable.
- **Data set** : In a study, the totality of the values of all the variables for all the selected sample units constitute the dataset corresponding to that particular study.
  Eg : In the above study, the totality of the values of monthly incomes, number of family members etc for all the sampled Indians constitute our data set.
- **Subject** : A unit on which we make observations or measurements on variables i.e on which we collect data. Generally, a subject is a person but can also be a state, country etc. Basically, the data we collect should be characteristics of the subject.

Eg : If you ask a sample of IIMA students whether they support the death penalty or not, the subjects will be ; if we collect data on the Gross Domestic Product (GDP) of different countries, the subjects will be ; if we collect data on the amount (or percentage) of rainfall each Indian state has received so far, the subjects will be

- **Population** : Set of all subjects of interest. A population generally depends on the purpose of the study being undertaken.

  Eg : For the death penalty example above, the population will be all the students currently enrolled at IIMA.

- **Sample** : It is a part of the population on which data is actually collected. Any statistical study uses the sample to learn about the population.

  Eg : In the above example, suppose you asked 20 students about their views on death penalty. So, the sample will be those 20 students.

- **Simple Random Sample** : A sample selected in such a way that every unit in the population has an equal chance of being included in the sample - such a sample is a good representative of the population. Generally, in the statistical field, if not stated otherwise, a sample is always a random sample.

  Eg : In the above example, your sample will be a simple random sample if each of the students currently enrolled at IIMA had an equal chance of being included in your sample (of size 20).

- **Random variable** : It is a numerical measurement of the outcome of a random phenomenon.

  Eg : The feedback (yes/no) of a randomly chosen student in your sample.

- **Parameter** : A parameter is a numerical summary of the population and hence describes the characteristics of a population. A parameter can be calculated only if we have information from the whole population. Generally, a parameter is regarded as an unknown quantity.

  Eg : The true percentage of IIMA students who support death penalty.

- **Statistic** : It is a numerical summary of a sample. A statistic describes a sample as a parameter describes a population. In fact, we estimate a parameter of a population using analogous sample statistics calculated from samples selected from that population.

  Eg : In this case, the statistic will be the percentage of affirmative response in your sample. Thus, if 8 out of the 20 students you interviewed support death penalty, the statistic will be $(8/20) * 100\% = 40\%$.

Eg : Marriage & Personality : An article in BBC future (dated Aug 31, 2017)* dealt with the topic of whether marriage changes our personality and if so, how ? Unfortunately, there has not been much research in this field but empirical evidence are not rare. For example, singles always complain that their married friends are not as much fun as they used to be before marriage ! On the other hand, a lot of married people vouch that marriage brings with it a degree of contentment and happiness that singles can never attain ! Let us see what our class feels about this i.e is it true that the experience of committing to and settling down with another person really does change our personalities (for better and for worse) ?

Identify the :

1. Research question :

2. Variable :

3. Subject :

4. Sample :

5. Sample size :

6. Population :

7. Parameter :

8. Statistic :

*Ref : http://www.bbc.com/future/story/20170831-how-marriage-changes-people-forever.

Chapter 2

Sampling Techniques

2.1 Introduction

As mentioned in Chapter 1, one of the most important component of any statistical investigation is to formulate a realistic and sound plan for data collection. A data set based on a well-formulated study is a good representation of the population and hence statistical inferences based on it would be believable and would help us understand the population better. On the other hand, if a study is not well designed, the results will likely be meaningless and/or misleading. There are different types of statistical studies. However, before going into the details, we need to understand the concept of *association* between variables, a key concept in Statistics.

2.2 Concept of Association

One of the most fundamental aspects of statistical practice is to *analyze* and *interpret* the relationship between different variables in the population. What makes this interesting is that relationships or *association patterns* between variables are often as diverse as the variables themselves. Some variables may have a pretty simple relationship while others may have a much more complicated pattern. Generally speaking, when two variables are associated, one influences the other.[1] Thus, we have the following two types of variables :

1. **Response variable (Y)** : This is the *outcome* or the *dependent* variable.
2. **Explanatory variable (X)** : This is the *independent* variable or the variable which *explains* or is related to the outcome [2].

Eg : (i) There are many factors which influence the ranking of IIMA (or of any institution per se) like salary of graduating students (say, s), placement success (say, p), percentage of international students

[1] Later on we will see that association may not necessarily imply causation. However, we will go by this statement for the time being for the sake of defining response and explanatory variables.

[2] Explanatory variables are also known as *covariates* or *predictors*

(say, i), research output of faculties (say, r) etc. So, ranking is the variable, while constitute the explanatory variables.

(ii) Whether taking Aspirin reduces the chance of heart attacks. Here the explanatory variable is "whether someone takes Aspirin" while the response variable is

Similarly, you can find innumerable examples from day-to-day life where association is at play between two variables. Given this concept, we will now learn about the two main types of statistical studies.

2.3 Statistical Studies

Broadly speaking, there are two major types of statistical studies :

- **Experiment** : Here a researcher conducts an *experiment* by assigning subjects to certain experimental conditions and then observing outcomes on the response variable. The experimental conditions correspond to some assigned values on the explanatory variables and are often called **treatments**.
- **Observational Study** : Here the researcher samples some subjects and just observes values of the response and explanatory variables for them without doing any experiment i.e without assigning the subjects to treatments.

Eg : Nowadays, as cell phones are becoming all-pervasive, a growing concern is whether cell phones have any adverse health effects. Several studies have explored this issue :

1. **Study 1** : An Australian study (Repacholi, 1997) used 200 transgenic mice, specially bred to be susceptible to cancer; 100 mice were exposed an hour every day to the same frequency of microwaves as transmitted from a cell phone. The other 100 mice were not exposed. After 18 months, it was found that, the brain tumor rate for the mice exposed to cell phone radiation was twice as high as the brain tumor rate for the unexposed mice.

   Clearly, this is an ; subjects (mice) were assigned to two different experimental conditions (or treatments) which were categories of the explanatory variable "*whether exposed to radiation ?*". The treatments were compared with respect to the response "*whether developed tumor/cancer ?*" having categories (yes, no).

2. **Study 2** : A German study (Stang *et al.*, 2001) compared 118 patients with a rare form of eye cancer, called uveal melanoma to 475 healthy patients who did not have the eye cancer. The patient's cell phone use was measured using a questionnaire. It was found that the eye cancer patients used cell phones more often, on an average.

   In the above study, the researchers sampled some subjects and merely observed the values of the response and explanatory variables for each of them *without doing anything to them.* So, these are .

So, the main difference between experimental and observational studies is that in the former, it is the researcher who decides on the allocation of treatments to the subjects while in the latter, it is predetermined and the researcher is just an observer.

2.4 Experiment & Observational Studies : A Comparison

In order to understand the advantages and disadvantages of experiment and observational studies, we need to be familiar with what are known as **lurking** or **confounding** variables.

2.4.1 Lurking Variables

Lurking variables are those which are associated with both the response and the covariates (explanatory variables) and in doing so, often gives misleading impression about the response-covariate relationship. For example, the German study (study 2) found an association between cell phone use (covariate) and eye cancer (response) - something that seems a little counterintuitive (why EYE cancer??). In a subsequent study it was found that those who use cell phones a lot tend to use computers a lot too. It is more intuitive that high exposure to computer screens increases the chance of eye cancer. Thus, the higher prevalence of eye cancer for heavier users of cell phone may infact be due to their higher use of computers, NOT due to the high use of cell phones. So, is the lurking variable in this case.

2.4.2 Advantages of Experiments

One big advantage of experimental studies is that it provides a lot more control over lurking variables compared to observational studies. This is because, by randomly allocating subjects to different treatments (*vis-a-vis* different categories of explanatory variables), the different groups will be "balanced" with regard to potentially lurking variables like lifestyle, genetic or health characteristics. Hence, there will be little or no influence of the lurking variables on the results.

For example, for the Australian study, if we randomly allocate the mice to different radiations, the two groups would likely to have similar distributions on health. One group will NOT be significantly healthier than the other; so we can be sure that the higher cancer rate in one of the groups is NOT because its mice have poorer health on average.

Establishing **cause and effect** is central to any scientific discipline. However, it is often not possible to definitively establish cause and effect with observational studies - there is always the possibility that some lurking variable was causing the association (which is actually NOT there). Thus, it is important to remember that **association does not always imply causation**. Since it is much easier to adjust for lurking variables in an experiment than in an observational study, we can study the effect of an explanatory variable on a response variable more accurately with an experiment rather than with an observational study.

2.4.3 Advantages of Observational Studies

The big disadvantage of experimental studies is that, it is not easy and often unrealistic to do experiments specially when dealing with humans. Consider the following experiment : *Select half of the faculty members at IIMA at random and ask them to use cell phone every day, without fail, for at least five hours per day for the next 10 years. Tell the other half not to even touch a cell phone for this time period. At the end of 10 years, analyze whether the former group had a higher proportion of cancer patients.*

Clearly this experimental setup is not only highly unrealistic but unethical too. Secondly, even if you do something like that, it is virtually impossible to enforce it. Lastly, it is absurd to wait for 10 years to get the results of a study. For this reasons, observational studies are almost always preferred, specially when the object of interest/inference are humans. If some experiment has to be done, it is generally done on animals and over a short duration. Fortunately, there are ways to formulate a good observational study which we will learn soon. Another reason for the popularity of observational studies (over experiments) is that, often researchers are not interested in questions related to assessing causality. For example, we may want to gauge the public's opinion on some controversial issues or on how they rate a particular product. In these cases, a observational study is good enough.

2.5 Sampling Techniques

In this course, we will mainly concentrate on observational studies since it is the one that is used more often and specifically in business applications. One of the foundations of virtually any observational study is a good sampling plan i.e selection of a sample that is a good representative of the population. Only then can the inferences drawn from that sample will be applicable to the population as a whole. On the other hand, a sample which does not have proper representation of the population is likely to give us erroneous/invalid results.

In order to select a good, representative sample from the population, we first need to identify the population of interest. This is because, the population depends on the statistical study (or the study question) itself. For example, many news papers collect regular surveys/polls from their readers regarding their views about current happenings. So, here the population will be all the subscribers to that newspaper/s. On the other hand, if your target of inference is the proportion of IIMA students who are vegetarians, the population would be all the students of IIMA.

2.5.1 Sampling Frame and Sampling Design

Once the population is identified, the second step is to compile a list of subjects (or units) in it so that you can sample from it - this list is called the **sampling frame** - *this is applicable when the population is finite (i.e has a fixed size, say "N").*

For example, if you want to know some particular characteristics of registered voters of Gujarat (say, what proportion of them own a vehicle), a possible sampling frame would be the list of all registered voters (of Gujarat) maintained by the Election Commission.

Once you have a sampling frame, you need to decide on a method of drawing samples from it. This method is known as the **sampling design**. The sampling design should be such that the resulting sample is a "good representative" of the population i.e it should reflect all the characteristics or nuances of the population.

For example, if your population is everyone who work/studies in IIMA, a good sample should have a *representative* mixture of students (that too from PGP, PGP-ABM, FPM etc), faculties (from all ranks), staff members, workers (from all departments) etc. However, if for convenience, you only select the diners from the KLMDC canteen on a given day, it may give you misleading information about the population since it will definitely not contain a significant part of the IIMA people (a large chunk of staff members do not go to KLMDC for dinner).

In a bigger canvas, if you want to gauge the national mood about an important issue, then the verdict of an online poll by a leading newspaper/media (say TOI, NDTV etc) will NOT be an accurate indicator since the results will not account for a huge chunk of the population (say those who don't have access to internet or are not subscribers/viewers of the above newspapers).

Thus, the bottom-line is that you have to be very careful in selecting a sample that would accurately reflect (or be a good representative of) the population.

2.5.2 Simple Random Sampling

Based on the discussion above, it should be clear that a sample selected out of convenience will NOT be a good, representative sample. Rather, a sampling design which gives each and every population unit an equal chance of being included in the sample will likely result in a much better sample (in terms of accurately representing the population). Such a sampling design, which is guided by "chance" rather than "convenience" is known as a **simple random sample**.

A **simple random sample** of "n" subjects from a finite population, say of size "N", is one in which each possible sample of that size (i.e n) has the same chance of being selected.

Eg: Suppose a campus club can select two of its officers to attend the club's annual conference in Bangalore. The five officers are President (P), Vice-President (V), Secretary (S), Treasurer (T) and Activity Coordinator (A). So, here the sampling frame will be . The club members want the selection process to be fair - so they decide to select a simple random sample of size . The 5 names are written on a identical slips of paper, placed in a hat, mixed, and then a neutral person blindly selects two slips from the hat - this is the sampling design.

Thus the possible samples of the two officers will be

There are 10 possible samples and the process of blindly (randomly) selecting two slips ensures that each of the above sample has an equal chance of occurring. Thus the chance of selecting any one of the above samples is . Moreover, since each of the 5 officers appears in of the above samples, the chance of each of them going to Bangalore is .

For large populations, there are better, more automated ways of selecting random samples. One of the popular method is using random number tables or computerised random number generators as follows :

- Number the subjects/units in the sampling frame.
- Generate random numbers of the same length/digits as the above numbers (this can easily be done through freely available softwares like `www.random.org`).
- Select those units from the sampling frame whose numbers were generated in the above step.
- Stop the process once you have reached the desired sample size.

Eg: State agriculture officials regularly visit different districts to check the productivity of crop fields. However, it is not possible for the auditors to examine all the crop fields (since there are so many of those in a particular district). So, they select a random sample of the fields and check their productivity.

Suppose a particular district has 70 large crop fields and the auditor want to randomly select 10 of those. Thus, the sampling frame consists of the 70 accounts $\{01, 02, ..., 70\}$. The auditor generates 2 digits at a time from a random number generator until he/she has 10 unique two digit numbers from 01 to 70. Any number from 71 to 99 are rejected and also if any number appears twice, it is ignored (since it does not make sense to select the same field more than once). Suppose the selected numbers are $\{05, 09, 15, 23, 44, 46, 52, 59, 63, 69\}$ - the official should then check the productivity of the fields with the above numbers.

The Income Tax department is said to use a similar method to randomly select those tax return accounts that it should audit.

Note : *The above selection procedure is known as **sampling with replacement** (SWR) since a particular unit can be selected more than once. However, if any unit can be selected/included in the sample only once, it is known as **sampling without replacement** (SWOR). Although SWR is a valid way of selecting a simple random sample, it is SWOR that is used most often. In fact, unless explicitly stated, a simple random sample implies a sample selected without replacement.*

2.6 Sample Surveys

Once a sample is selected, there are different ways in which one can collect data from the sample units. Some of the commonly used methods are **personal interview**, **telephonic interview** and **self-administered questionnaire**. In majority of the cases though, these data are results of **sample surveys**. As the term suggests, it is based on selecting a representative sample from the population and collecting data on those. Clearly, sample surveys are

2.6.1 Potential Biases in Sample Surveys

One of the main issues with sample surveys is that, often the responses from the sample tend to favor some parts of the population over others. Then the results of the sample are not representative of the population and are said to be **biased**. For example, for the population of adults in your home town, results of an opinion survey may be biased in a liberal direction if you sample only educators or biased in the conservative direction if you sample only business owners. Bias can occur in a sample due to various reasons as follows :

1. **Sampling Bias** : As the term suggests, this kind of bias results from a flaw in the sampling method, most likely if the sample is **non-random** (like the one mentioned above). Another way it can occur is due to **under-coverage** - having a sampling frame that lacks representation from parts of the population. Responses by those not in the sampling frame might be quite different from those in it thus leading to misleading conclusions about the population.

   Eg : A telephone survey (*vis-a-vis* the sampling frame on which it is based) will not reach homeless people or prison inmates; incidentally, these groups of people may have very different views about life in general. Similarly, online surveys by newspapers suffer from serious under-coverage.

2. **Non-response bias**: This kind of bias results when some of the sampled subjects cannot be reached or refuse to participate. In fact, the subjects who are willing to participate may be different from the overall sample in some way, perhaps having strong views about the survey issues. The subjects who do participate may not respond to some questions, resulting in non-response bias due to **missing data**.

   Nearly all major surveys suffer from some non response biases ranging from 20 – 30% (Eurobarometer in UK) to 6 – 7% (Current Population Surveys in US).

3. **Response bias** : This kind of bias results from the actual responses. The responses of subjects may differ based on the particular manner the interviewer asks questions; subjects can often lie because they think that their responses may be socially unacceptable. In fact the wording of questions can greatly affect the responses - it is always preferable to word questions in a direct, clear and understandable manner and avoid, wordy and confusing questions.

   Eg: A Roper poll (Newsweek, July 25, 1994) asked a sample of adult Americans : “Does it seem possible or does it seem impossible to you that the Nazi extermination of Jews never happened ?” 22% said that it was possible the Holocaust never happened !! The Roper organization later admitted that the question was worded in a confusing manner. When they asked instead : “Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened ?”, only 1% said that it is possible that it never happened !!

2.7 Observational Studies

There are different ways of sampling which can result in a well-designed observational study/survey. Some of the commonly used ones are :

2.7.1 Cluster Sampling

Often a sampling frame consisting of ALL the population units is hard to come by. In that case, a population can be divided into a large number of clusters and a simple random sample of a pre-specified number of clusters can be selected. The **cluster random sample** will consist of all the units in those clusters.

Eg : Suppose you like to sample about 1% of the families in your city. You can use a map to label and number the city blocks (which are the clusters) and can select a simple random sample of 1% of the blocks. Now you can select each and every family in those blocks - those will be your observations. The sectors of Chandigarh in Punjab or the blocks of Salt Lake City in Kolkata may be taken as good examples of clusters.

2.7.2 Stratified Sampling

Here, a population is divided into separate groups or stratas and a simple random sample is selected from each stratum.

Eg : Suppose you want to estimate the mean number of hours a week that IIMA students spend in the library and also how it compares between PGP I, PGP II, PGPX and FPM. You can easily identify all these groups of students using the registration records - those will be your strata. If you want a sample size of, say 40, you can select a simple random sample of size 10 from each strata (or you can take it to be proportional to the size of the respective student bodies).

Historically, part of the old city of Ahmedabad is neatly divided into "pols", mainly along religious lines (Jain, Hindu, Muslim, Buddhist etc). Suppose you want to compare the average income/wealth of the various pols. Then you can treat each pol as a and select a random sample of households from each. For each such sample of households, you can calculate the mean income/wealth.

Stratified sampling ensures that you have adequate representations from each strata you want to compare. However, you should have access to the sampling frame and the strata into which each subject belongs. *The main difference between cluster and stratified sampling is that in the former the within-cluster variability is large while the between-cluster variability is small; however in stratified sampling, the within-strata variability is small but the between-strata variability is large.*

Allocation Schemes

Naturally, an important issue with stratified sampling is to formulate a way to decide on the sample size to be chosen from each strata. Although various allocation schemes have been formulated for this purpose, the most commonly used one is **proportional allocation** where the sample size (n_h) allocated to stratum h is given by

$$n_h = n\left(\frac{N_h}{N}\right)$$

Where N is the population size while N_h is the size of the h^{th} stratum.

Eg. Suppose a survey carried out by ICICI bank across all its branches revealed the following information about the age and rank of their executives.

Rank	Mean age	SD of age	Strata size	Sample size
Asst Manager	23	3.5	1500	
Manager	27	3.0	1200	
Senior Manager	31	2.8	850	
AVP	35	4.0	400	
VP	41	3.7	200	

Suppose, for a particular survey, you want to select 300 employees in total. Then what would be the sample sizes for each stratum under proportional allocation ? (complete the last column in the above table)

Eg : Marriage & Personality : Continuing with the same example as in Chapter 1, it is often stated by psychologists that the two most important marital traits/skills are (i) **self-control** i.e having the ability to bite your tongue for the long term sake of marriage and (ii) **forgiveness** i.e ability to look past the errors of your partner and move on.

To test whether marriage hones these skills, a team of Dutch psychologists at Tilburg university recruited 199 newlywed couples (who were current and past students of the university) and measured the above traits three months after their marriage and each year for the next four years thereafter. They asked questions like "*Don't you think it is best to just forgive and forget when your partner wrongs you ?*" or "*Don't you think that it is best to just overlook minor digressions on the part of your partner for the longterm sake of marriage ?*". They received complete feedback from 176 of the 199 couples. Based on the results, it was apparent that participants became progressively better in both the above skills with time. Identify whether (and if so, why) the following may be of concern :

1. Sampling bias :

2. Undercoverage :

3. Response bias :

4. Non-response bias :

What kind of study is this (experiment/observational) ? Justify.

Eg : Stress & Ageing : It is well known that stress can adversely affect our health and mental wellbeing. However, not much research has been done to analyze how exactly this happens at the cellular level. Researchers at the University of California, San Francisco wanted to probe this issue. Accordingly, they took blood samples from 58 mothers, 19 of whom were non-stressed while the rest 39 were "stressed-out" with varying levels of stress. For each of them, the length of Telomeres and the levels of Telomerase were measured (these are indicators of ageing with shorter Telomeres and lower levels of Telomerase indicative of accelerated ageing). In doing so, they found that higher levels of stress was associated with shorter Telomeres and lower levels of Telomerase. In fact, the most stressed-out women in the group had Telomeres/Telomerase levels that translated into an extra decade or so of ageing compared to those who were least stressed. Thus, they concluded that feeling stressed does not only damage our health it literally ages us.

What kind of study is this (experiment/observational) ? Justify.

Identify whether (and if so, why) the following may be of concern :

1. Sampling bias :
2. Undercoverage :
3. Response bias :
4. Non-response bias :

References

1. Repacholi, M. H. (1997), "Radio frequency field exposure and cancer", *Environ. Health Prospect*, **105** : 1565-1568.
2. Stang, A., et al. (2001), "The possible role of radio frequency radiation in the developement of uveal melanoma", *Epidemiology*, **12(1)** : 7-12.
3. http://www.bbc.com/future/story/20170831-how-marriage-changes-people-forever.
4. http://www.bbc.com/future/story/20140701-can-meditation-delay-ageing.

Chapter 3

Exploratory Analysis

3.1 Introduction

Any statistical experiment/survey usually generates a lot of raw data. In order to make sense of it, we need to somehow summarize it using some numerical and/or graphical measures. The procedure of doing this is broadly known as **Descriptive Statistics**. In this chapter we will deal with the two main ways of summarizing data i.e **numerical** and **graphical** summaries. However, prior to that, we need to understand the different types of variables as explained below.

3.1.1 Types of Data/Variable

A variable can be of different types (resulting in different types of data and hence the methods to analyze those) as follows :

1. **Quantitative** : A variable is quantitative if the observations on it takes numerical values that represent different magnitudes. **Eg :** Number of friends you have, your height, weight etc. Quantitative variables can be of two types :

   - **Discrete** : A quantitative variable is discrete if the possible values belong to a set of distinct numbers like 1, 2, 3,... **Eg** : Number of pets you have.
   - **Continuous** : A quantitative variable is continuous if the possible values belong to an interval. **Eg** : Your exact height since it may literally be any number between, say, 4ft and 7ft.

2. **Categorical** : A variable is categorical if each of its observations belong to any one of a set of categories. **Eg :** Your religious affiliation : {no religion, Hindu, Muslim, Buddhist, Jain, Christian etc}. Categorical variables can be of two types :

   - **Ordinal** : A categorical variable is ordinal if it has *ordered* categories.
     Eg : Level of education having categories {no education, middle school, high school, undergraduate, graduate}.

- **Nominal** : A categorical variable is nominal if it has *unordered* categories. **Eg** : Your favorite color may be any of {red, yellow, green, crimson etc}.

1. *Sometimes a continuous variable may be simplified into a categorical one for ease of measurement. For example, length of hair, although continuous, may be categorized as {very short, short, medium, long, very long}.*

2. *Not all variables those are numbers are quantitative. For example, Section numbers of courses, zip codes, passport, Aadhar card numbers for example do not measure the magnitude of anything although they have numerical values. In fact, these are just convenient numerical labels used for identification.*

3.2 Graphical Summaries

One of the best ways to summarize (and understand) a large chunk of raw data is to represent it pictorially. As you know, a picture is worth a thousand words !

3.2.1 Graphs for Categorical Variables

The two most common graphical summaries of a categorical variable are (i) **Bar graphs** and (ii) **Pie charts**.

1. **Bar graph** : It displays a vertical bar for each category, the height of the bar representing the percentage of observations in that category.

2. **Pie chart** : It is a circle with a "slice of pie" for each category, the size of the slice representing the percentage of observations in that category.

Eg : The following table shows the different sources of electricity in India and the corresponding percent uses of each source.

Source	Coal	Hydropower	Natural Gas	Nuclear	Petroleum	Other	Total
Percentage	51	6	16	21	3	3	100

The corresponding bar graph and pie chart are shown in Figs 3.1 and 3.2. Clearly, sources of electricity with higher percent use have higher bars/larger slices of pie. However, bar graphs are superior to pie charts since it is easier to distinguish between categories in the former specially when two categories are very close (in terms of percentage/frequency).

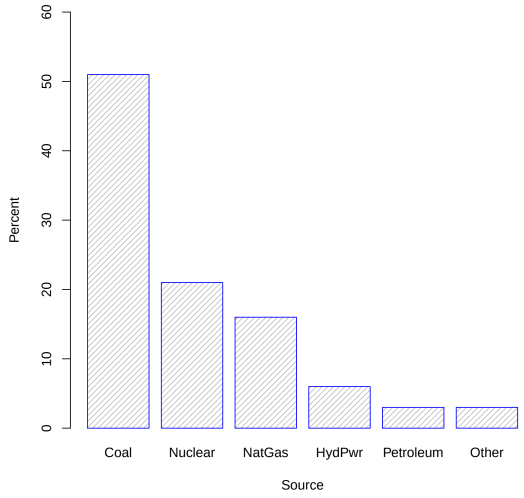


FIGURE 3.1: Bar graph of electric usage

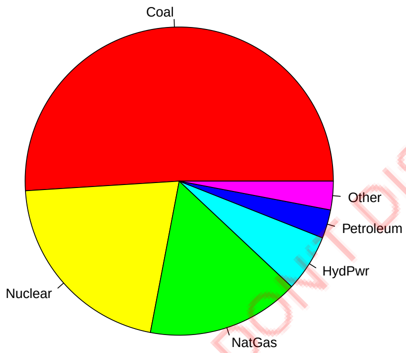


FIGURE 3.2: Pie chart of electric usage

3.2.2 Graphs for Quantitative Variables

The most common graphical summary of quantitative variables is **Histogram**. However, before going on to histograms, we need to know about **frequency tables**. For large data-sets, it is often useful to look at the possible values (of a variable) and count/list the number of occurrence of each. For **categorical variables**, this is just the number of values in each category. We can then divide the number in each category by the total number (in all the categories combined) to get a proportion (or percentage) for each category.

Eg : The following table shows the number/frequency of shark attacks in some (shark infested) states of the U.S and also in some other countries.

Region	Frequency	Proportion	Percentage
Florida	365	$365/701 = .52$	52%
Hawaii	60	.086	8.6%
California	40	.057	5.7%
Australia	94	.134	13.4%
Brazil	66	.094	9.4%
South Africa	76	.108	10.8%
Total	701	1.00	100

Here   is the categorical variable with   categories. For example, out of 701 total shark attacks, 365 i.e   were in Florida.

For **quantitative variables**, a frequency table usually divides the possible values into a set of intervals and displays the number of observations (or frequencies) in each interval. A **Histogram** is a graph which use bars to represent the frequencies (or proportions) of the possible outcomes of a quantitative variable. For discrete variable, a histogram usually has one bar for each possible value. For a particular value, the height of the bar is proportional to the frequency or relative frequency of that value. For continuous variables or discrete variable with a large number of unique values, it is convenient to segregate the values into different intervals and have a separate bar for each.

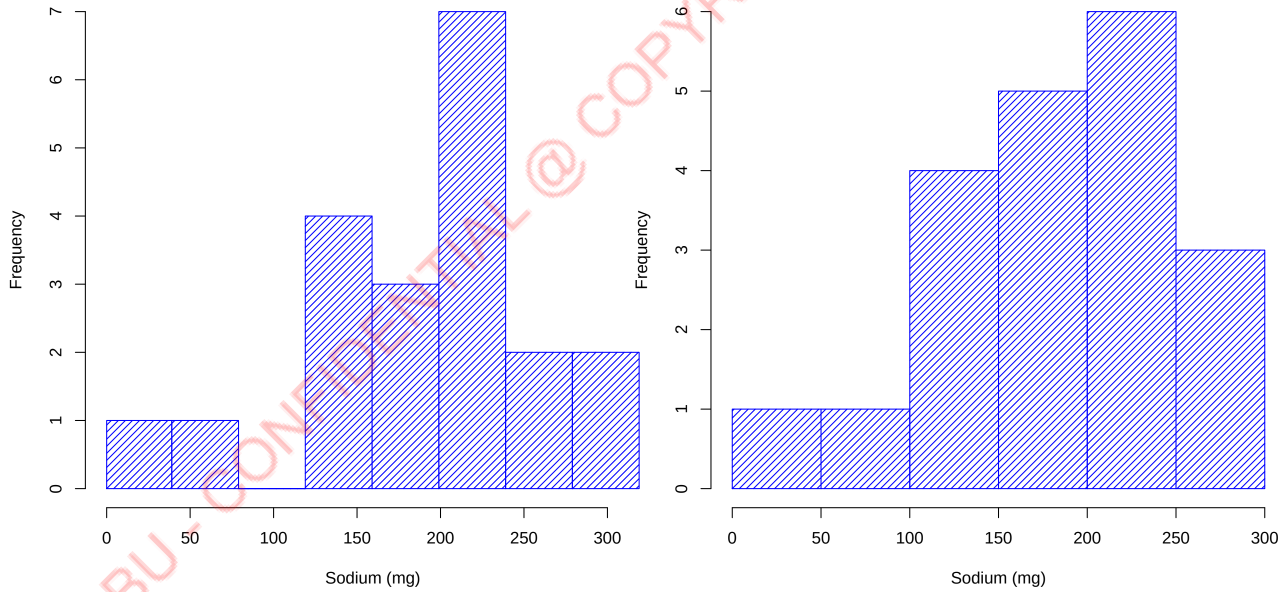
Eg : Nutritional labels on packaged foods provide information on the amount of different minerals contained in one serving of that food. The following table lists 20 popular brands of cereals and the amounts of sodium and sugar contained in a single serving (about 3/4 cup) as listed on the label.

Cereal	Sodium (mg)	Sugar
Frosted mini wheats	0	7
Raisin bran	210	12
All bran	260	5
Apple Jacks	125	14
Capt Crunch	220	12
Cheerios	290	1
Cinnamon Toast Crunch	210	13
Crackling Oat Bran	140	10
Crispix	220	3
Frosted Flakes	200	11
Froot Loops	125	13
Grape Nuts	170	3
Honey Nut Cheerios	250	10
Life	150	6
Oatmeal Raisin Crisp	170	10
Honey Smacks	70	15
Special K	230	3
Wheaties	200	3
Corn flakes	290	2
Honeycomb	180	11

For the above data, there are many possible values for the amount of Sodium (0-319). So, it is best to form intervals before drawing the histogram. For example, the frequency 4 in the interval 120-159 implies that 4 cereals had Sodium content between 120 and 159 mg. The table in the next page shows the intervals and the corresponding frequencies for the cereal data.

Interval	Frequency	Proportion	Percentage
0-39	1	0.05	5.0
40-79	1	0.05	5.0
80-119	0	0.0	0.0
120-159	4	0.20	20.0
160-199	3	0.15	15.0
200-239	7	0.35	35
240-279	2	0.10	10
280-319	2	0.10	10
Total	20	1.00	100.0

Two possible histograms for the cereal data are shown below with two different interval structures (first histogram : above interval).



3.2.3 Graphs for Time Varying Variables

Often we come across variables which are measured over time, for example, our blood pressure when measured every month for one year, the monthly temperature of a place measured for the last two years etc. It is helpful to plot the observations on these variables over time to understand whether any trend (generally long term) is present. These plots are known as **Time plots**. Fig 3.3 depicts a time plot showing the variation of the annual average temperature (average of the daily temperature for a year) of Central Park in New York City between 1901 to 2000.

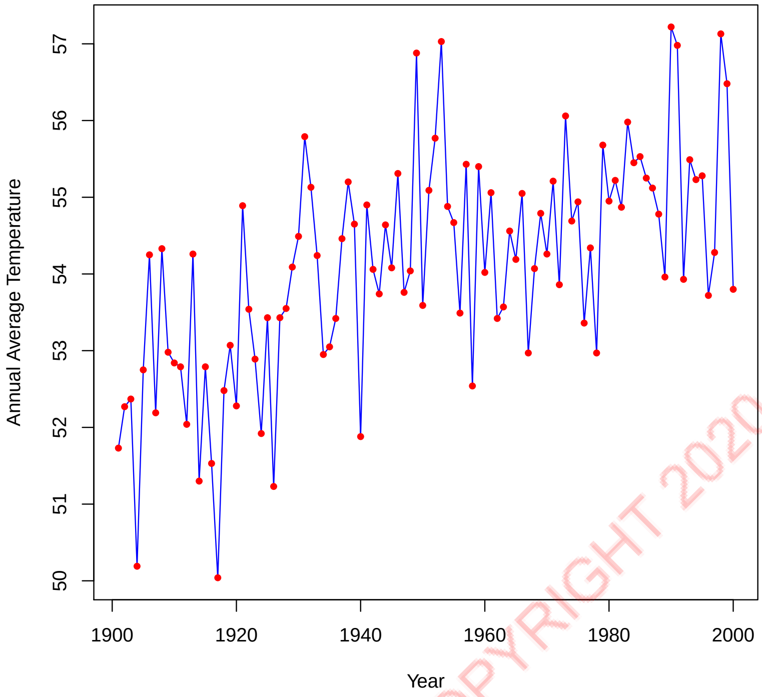


FIGURE 3.3: Time plot of average annual temperature of Central Park, NY

Although, the observations have considerable fluctuations, there is a clear increasing trend over the 100 year period...a tell-tale sign of global warming!

3.3 Data Distribution and Shape

The **distribution** of a variable is explained by the **values** that the variable takes and the **frequency** of occurrence of each values. So, the frequency table and the histogram can give us an idea of the distribution of a variable. The most important aspect of a distribution is its shape. Some of the common shapes we may encounter are :

1. **Unimodal** : Distribution with one mode.

2. **Bimodal** : Distribution with two modes.

3. **Multimodal** : Distribution with more than two modes.

4. **Symmetric** : Two sides of a distribution about a central point mirror images of each other.
   Eg : IQ, height, weight.

5. **Left skewed** : Distribution whose left tail is longer than the right tail.
   Eg : Life span (large majority of people live at least 65 years but some die at a young age); distribution of scores in an easy PS exam.

6. **Right skewed** : Distribution whose right tail is longer than the left tail.
   Eg : Distribution of scores in a difficult PS exam, amount of donation received by a temple (large majority donates a standard amount but few wealthy individuals may donate a huge amount).

3.4 Numerical Summaries

Generally graphical summaries are the first step forward in analyzing a statistical data since they sort of provides a “bird’s eye view” of the overall data pattern and often indicates the next step in the analytical procedure. This next step is often calculating the **numerical summaries** which are basically some values reflecting some important characteristics of the data set.

The numerical summaries are of three types viz (i) **Measures of center** (ii) **Measures of spread** and (iii) **Measures of position.**

3.4.1 Measures of Center

There are basically two types of measures of center as follows :

1. **Mean** : The mean (or average) of a variable is just the sum of its observations divided by the number of observations. However, if you have a set of values, say $\{x_1, x_2, ..., x_n\}$ with frequencies $\{f_1, f_2, ..., f_n\}$, then the mean will be

$$m = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} \tag{3.1}$$

2. **Median** : The median of a variable is the midpoint of its observations when ordered in increasing or decreasing order.

   Note : When the number of observations (say, n) is odd, median is exactly the middle observation (i.e the $(n+1)/2^{th}$ observation). However, when the number of observations (n) is even, median is the average of the two middle observations (i.e the average of the $n/2^{th}$ and $(n/2)+1^{th}$ observation).

Eg: For the cereal data set, the observations (amount of sugar) are {1, 2, 3, 3, 3, 3, 5, 6, 7, 10, 10, 10, 11, 11, 12, 12, 13, 13, 14, 15 }. Here the mean will be

To calculate the median, we first arrange the observations in increasing order (as is done above) and calculate the mean of the two middle most observation. So, the median will be

Some important points to note :

- Median does not depend on the actual values of the observations while mean does. This is precisely why mean is highly influenced by **outliers** (observations falling way above or below the rest of the data) while the median is not.

  Eg : The following table depicts the per capita CO_2 emissions for some countries

Country	CO_2 emission
China	2.3
India	1.1
United States	19.7
Indonesia	1.2
Brazil	1.8
Russia	9.8
Pakistan	0.7
Bangladesh	0.2

Do you spot an outlier/s ?

The mean $C0_2$ emissions is while the median is

Conclusion : The mean is so high compared to the median due to the outliers which virtually drag the mean towards themselves. However, the median only depends on the relative position of the observations and hence is not affected (thus remains low). *So, we say that median is resistant to extreme observations while the mean is not.*

For example, if the $C0_2$ emission of US would have been the same as that of Russia (9.8), the mean would have decreased to ; however, the median would have remained the same at 1.5. Similarly, if the $C0_2$ emission of the U.S increased to, say, 25, the mean would have shot up to but the median would still remain the same - this is because, this change will not affect the relative positions of the observations.

- Suppose Brazil's $C0_2$ emission is 1.0 instead of 1.8. Then the new ordering would be $\{0.2, 0.7, 1, 1.1, 1.2, 2.3, 9.8, 19.7\}$. So, the new median would be . In the old data, Brazil was between Indonesia and China; however in the new data, it is between India and Pakistan. Thus, it's relative position has changed, resulting in the change in the median.

- The shape of a distribution influences whether the mean should be larger or smaller than the median. Often an extremely large value in the right(left) hand tail (of the data distribution) may drag the mean towards the right (left) so much that it may fall above (below) the median. Thus, we have

  – Symmetric distribution ⇒ mean = median.

  – Right-skewed distribution ⇒

  – Left-skewed distribution ⇒

  In a nut shell, the mean is always drawn towards the longer tail of the distribution.

- However, mean also have some advantages over the median that we will learn later - so, the use of mean or median would depend on the data we have.

3.4.2 Measures of Spread

Measures of center only tells us about the average or middle-most value of the distribution. Another key feature of a distribution is the amount of spread of the observations. This is important because two variables may have the same mean and median but very different spread of the individual observations. There are two commonly used measures of spread viz

1. **Range** : It is simply the difference between the largest and smallest values of a variable. **Eg**: For the CO_2 emission data, the maximum value is 19.7 and the minimum is 0.2. So, the range is . Unfortunately, range is also sensitive to the present of outliers. For example, if the CO_2 emission for U.S was, say, 25, the range would have shot up to

2. **Standard Deviation** : This is the most popular measure of spread since it uses all the observations in the data. Basically, it computes the deviation of each observation from the mean and then combines all the deviations into a single number which reflects the overall spread in the data.

   Let there be n observations and x_i be the i^{th} observation and $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$ be the mean. Then, the standard deviation is given by

$$\begin{aligned} s &= \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \end{aligned}$$

   Eg : For the cereal data, the mean (of Sodium) is $\bar{x} = 185.5$. Honey Nut Cheerios has sodium content 250 mg; so its deviation (from the mean) will be . Similarly, the deviation of Honey Smacks (Sodium content 70 mg) will be . Once we have the deviations of all the cereals, the standard deviation will be

$$\sqrt{\frac{(0 - 185.5)^2 + (70 - 185.5)^2 + \ldots + (290 - 185.5)^2}{19}} = 71.25 \qquad (3.2)$$

   This sort of means that the typical difference between an observation from the mean is 71.25.

Note

- The square of the standard deviation (s) is known as the **variance** i.e

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- The sum of all the deviations is 0 since $\sum_{i=1}^{n}(x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$.

- Clearly, greater the spread of the data, larger will be s.

- $s = 0$ only when all the observations have the same value (i.e same as the mean). **Eg**: if all the Cherrios had Sodium content 100 mg, the mean would have been 0, hence each deviation (about the mean) would have been 0, resulting in a 0 standard deviation.

- Since s uses the mean, it is also sensitive to outliers; this is intuitive because an outlier would have large deviation (and hence very large squared deviation) and hence increases the standard deviation.

3.4.3 Measures of Position

As the term suggests, these are some ‘positional landmarks’ of a distribution in that they divide the distribution into some well defined areas (in terms of the proportion of observations that fall above and/or below these values).

For example, the **median** is also a measure of position since it specifies a location such that *half of the data falls above it and the other half below it.* In fact, the median is a special case of a general set of measures of position called the **percentiles**.

Definition : The p^{th} percentile is a value such that $p\%$ of the observations fall below it.

Eg : Suppose a student scores 1200 (out of 1600) in the GRE and is told that his score is at the 90^{th} percentile. This implies that 90% of those who took the exam scored between the minimum score and 1200 i.e only 10% of the scores were higher than your friend's.

Thus, it is clear that the median is the 50^{th} percentile. Infact, the three popular percentiles are :

- **First quartile ($Q1$)** : Lowest 25% of the observations fall below it i.e $p =$
- **Second quartile ($Q2$)/median**: 50% of the observations fall below it i.e $p =$
- **Third quartile ($Q3$)** : 75% of the observations fall below it (or highest 25% of the observations fall above it) i.e $p =$

Thus, the quartiles split the distribution into 4 distinct parts, each containing 25% of the observations.

Finding the Quartiles

1. Arrange the data in increasing order.
2. Identify the median ($Q2$).
3. Consider the observations *below* the median. The first quartile ($Q1$) will be the median of these observations.
4. Consider the observations *above* the median. The third quartile ($Q3$) will be the median of these observations.

Eg. Let us find the quartiles of the sodium values in the 20 cereals. The Sodium values, arranged in increasing order are : $\{0, 70, 125, 125, 140, 150, 170, 170, 180, 200, 200,$
$210, 210, 220, 220, 230, 250, 260, 290, 290\}$. Thus, we have

- Median =
- $Q1 =$
- $Q3 =$

Note : Quartiles also indicates the shape of the distribution. As a rule of thumb, if the distance between $Q1$ and $Q2$ is larger (smaller) than that between $Q2$ and $Q3$, the distribution is left (right) skewed.

For the Cereal data, $Q_2 - Q_1 =$ while $Q_3 - Q_2 =$. Hence the distribution of the Sodium values is . The quartiles are also used to define a measure of spread known as the **Inter Quartile Range (IQR)** which is the distance between the third and the first quartiles i.e IQR = Q3 - Q1 i.e it is range for the middle 50% of the data.

Eg : For the Sodium data, IQR =

As with range and standard deviation, IQR increases with the spread of the data. However, compared to the range and the standard deviation, the IQR is much more resistant to outliers. This is because, it is based on the quartiles, which in turn are resistant to outliers (as they depend only on the relative positions of observations). So, for highly skewed distributions (or distributions with outliers), it is better to use the IQR rather than range or standard deviation.

For example, if the maximum Sodium content was 1000 (instead of 290), both the range and standard deviation would have increased but the IQR would have remained the same.

3.4.4 Five Number Summary and Box-plot

The five number summaries of a dataset/distributions are the minimum value, first quartile, median, third quartile and the maximum value. These are the basis of a graphical display known as **Box-plot**. For the Sodium data, the box-plot is given below

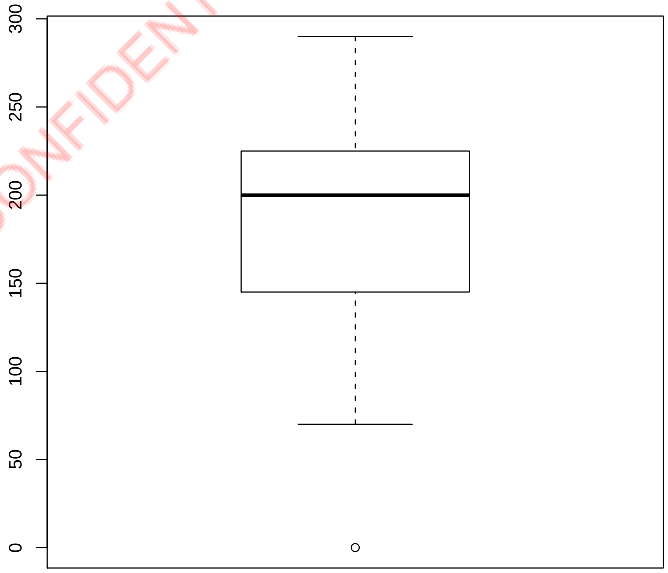


FIGURE 3.4: Boxplot of Sodium content in Cereals

A Box-plot has the following features :

- The **box** of a box-plot goes from Q1 to Q3 i.e it contains the central 50% of the distribution.
- A line inside the box marks the median.
- Lines extending from the box in either direction encompasses the rest of the data except for potential outliers. These lines are called **whiskers**.
- Outliers are shown separately, generally using the "*" symbol.

For the Box-plot of the Cereal data, the box extends from to i.e it contains the central 50% of the observations. The only outlier (Frosted Mini Wheats) is shown as the dot at the extreme bottom. Lastly, the two **whiskers** extend to and which are the minimum and maximum values except the outliers.

Note : A boxplot can also indicate whether the data is skewed i.e the side with the larger part of the box and the longer whisker usually has the skew in that direction.

3.4.5 The Mode

Mode is the value with the largest frequency i.e the value that occurs most frequently in a distribution. Usually, it is used with regard to a categorical variable to denote the category that has the highest frequency. For quantitative variables, it is generally used for discrete variables taking a small number of possible values.

Eg : (i) For the "Electricity source" data, is the mode since it corresponds to the highest percentage/frequency.

(ii) For the "Shark attack" data, is the mode.

(iii) For the "Sodium content" example, the interval (200-239) or (200-250) corresponds to the mode since the highest proportions of cereals has Sodium content belonging to this interval.

Note :

- The Mode need not be near or at the center of the distribution - so it not really a measure of center; nor does it tells us anything about the spread of the distribution. So, it is usually classified as a measure of position.
- For perfectly symmetric distributions, **Mean** = **Median** = **Mode**.

3.5 R codes

Following R codes were used to perform the analysis in this chapter. Necessary explanations are provided alongside the codes. (R can be downloaded from `https://www.r-project.org/`)

Important : *Create a folder, say `SDA` in your desktop. Transfer all the datafiles therein. Open the folder, right-click on it and go to `Properties` and copy the `Location` indicator, say `D:Desktop`. Now you need to add the folder name to it and change the backslashes to forward slashes and use this path name in the codes below, as I have done. Remember, your pathname will be unique and may be different from `D:/Desktop/SDA`.*

1. Setting up the working directory :

   ```
   setwd("D:/Desktop/SDA")
   getwd()
   ```

2. Importing the `Electricity` data, Sec 3.2.1:

   ```
   energy<-read.csv("D:/Desktop/SDA/Energy usage.csv",header=TRUE)
   attach(energy)
   Source<-energy$Source
   percent.use<-energy$percent.use.
   ```

3. Bargraph and pie-chart (Figs. 3.1 and 3.2):

   ```
   barplot(c(51,21,16,6,3,3),names.arg=c("Coal","Nuclear","NatGas","HydPwr","Petroleum","Other"),
   ylab="Percent",col="blue")
   barplot(percent.use,names.arg=c("Coal","Nuclear","NatGas","HydPwr","Petroleum","Other"),
   ylab="Percent",col="blue")
   pie(c(51,21,16,6,3,3),labels=c("Coal","Nuclear","NatGas","HydPwr","Petroleum","Other"))
   pie(percent.use,labels=c("Coal","HydPwr","NatGas","Nuclear","Petroleum","Other"))
   ```

   To know more about a particular function you can type ”?name of the function()”. For example `?barplot()` or `?pie()` will take you to the manuals of those functions.

4. Saving the above pictures as a `.pdf` file in the working directory :

   ```
   pdf("energy-barchart.pdf")
   barplot(percent.use,names.arg=c("Coal","Nuclear","NatGas","HydPwr","Petroleum","Other"),
   ylab="Percent",col="blue")
   dev.off()
   pdf("energy-piechart.pdf")
   pie(c(51,21,16,6,3,3),labels=c("Coal","Nuclear","NatGas","HydPwr","Petroleum","Other"))
   dev.off().
   ```

5. Importing the `Cereal` data :

   ```
   cereal<-read.csv("D:/Desktop/SDA/Cereal.csv",header=TRUE)
   attach(cereal).
   ```

6. Histograms of `Sodium`, Sec 3.2.2 :

   ```
   hist(sodium,col="blue",xlab="Sodium(mg)",ylab="Frequency",main="Histogram of sodium
   content in cereals").
   ```

7. Importing the `Temperature` data :

   ```
   temp<-read.csv("D:/Desktop/SDA/Temperature.csv",header=TRUE)
   attach(temp).
   ```

8. Time plot, Fig 3.3 :

   ```
   plot(YEAR,TEMP,col="red",pch=20,xlab="Year",ylab="Annual Average Temperature")
   lines(YEAR,TEMP,col="blue")
   ```

   The argument `pch` determines the texture of the points. Use `pch = 1,2,3...` and observe the textures !

9. Obtaining fitted values and residuals for sampled subjects (Sec 10.6) :

   ```
   fitted.mm<-fitted(fit.mm)
   resid.mm<-residuals(fit.mm).
   ```

10. Mean and Median values of `Sugar` values, Sec 3.4:

    ```
    mean(sugar)
    median(sugar).
    ```

11. Importing the `CO2 emissions` data:

    ```
    emission<-read.csv("D:/Desktop/SDA/CO2 emission.csv",header=TRUE)
    attach(emission).
    ```

12. Mean and median of CO2:

    ```
    mean(CO2.emission)
    median(CO2.emission).
    ```

13. Modifying CO2 values of USA and Brazil and resulting mean/medians:

    ```
    CO2.emission[3]<-9.8 (changes CO2 of USA to 9.8)
    mean(CO2.emission)
    median(CO2.emission)
    CO2.emission[3]<-25 (changes CO2 of USA to 25)
    mean(CO2.emission)
    median(CO2.emission)
    CO2.emission[3]<-19.7 (changes CO2 of USA to 19.7)
    CO2.emission[5]<-1 (changes CO2 of Brazil to 1)
    mean(CO2.emission)
    median(CO2.emission).
    ```

14. Calculating the `Range` of CO2 emission, Sec 3.4.2:

    ```
    CO2.emission[5]<-1.8 (changes CO2 of Brazil back to 1.8)
    ```

```
range<-max(CO2.emission)-min(CO2.emission).
```

15. Standard deviation of `Sodium` values:

    ```
    sd(sodium).
    ```

16. Median, Q1, Q3 and IQR of Sodium values, Sec 3.4.3 :

    ```
    Quant.sodium<-quantile(sodium)
    Q1<-Quant.sodium[2]
    Q2<-Quant.sodium[3]
    Q3<-Quant.sodium[4]
    IQR<-Q3-Q1
    Quant.sodium
    Q1
    Q2
    Q3
    IQR.
    ```

    The `quantile()` function generates the 0%, 25%, 50%, 75% and 100% quantiles of the sodium values.

17. Box-plot of Sodium values:

    ```
    boxplot(sodium).
    ```

18. *R does not have an inbuilt function for calculating mode. However, once you have a frequency distribution of any variable, it will be obvious what the mode is.*

Chapter 4

Basic Concepts of Probability

4.1 Introduction

In everyday life, we often make decisions in the face of uncertain outcomes. Should you invest money in a stock market ? Should you start a new business ? Should you buy an expensive lottery ticket ? Should you carry an umbrella to work today ? We always have to deal with random phenomenon like these and weigh the possible outcomes of such phenomenon before taking a decision.

Probability is the tool that helps us to quantify uncertainty and thus can be used to measure the chances of all possible outcomes of a random phenomenon. No wonder, it is one of the foundations of Statistics.

Eg : Rolling a dice or flipping a coin are the two most common examples of random phenomenon; the random outcomes being the different faces of a die i.e $\{1, 2, 3, 4, 5, 6\}$ and the "head" and "tail" of a coin.

The performance of a random phenomenon is called a **random experiment or trial**. So, when you toss a coin or roll a dice, you are performing a random experiment or trial (provided that the coin/die is fair) since you are not sure which face of the coin or die will occur.

4.2 Long-run Behavior of Random Outcomes

If you perform an experiment or trial a small number of times, the outcomes may seem to follow a particular pattern. This can seem to be counterintuitive since you expect the outcomes to be totally random.

Eg : If you toss a fair coin 10 times, you may get a "head" 5 times in a row, and maybe a total of 8 times, although you expected it to happen 5 times, more so in a random fashion. However, this is not surprising because unpredictability is the essence of randomness. In fact, if we perform a random experiment a very large number of times, the number (and proportion) of any particular outcome would "settle down" and would get closer and closer towards the number that we would expect.

Eg : If we let a computer perform a coin tossing experiment a million times, the number of heads/tails would be very close (and maybe equal) to 1/2 a million, which is what we would expect. This "long-run behavior" of random outcomes forms the basis of the definition of probability.

Probability of a particular outcome for a random phenomenon is the proportion of times that outcome would occur in a long sequence of random experiments or trials.

Eg : (i) When we say that a roll of dice has outcome 6 with probability , we mean that the proportion of times a 6 would occur in a long run of observations is . This would also be the probability for each of the other possible outcomes $\{1, 2, 3, 4, 5\}$.

(ii) When the weather channel says that the chance of precipitation today is 20% i.e 0.20, it means that in a large number of days with atmospheric conditions like those today, precipitation would take place in 20% of the days.

Note : Probability refers to the "long run" because we can't accurately access a probability with a small number of trials. If you randomly sample 10 people and all of them happen to be right handed, you cannot conclude the probability of being right-handed is 1.0 in the population. It would take a much larger sample of people to accurately predict the proportion of people in the population who are right handed.

Independent Trials : Different trials of a random phenomenon are independent if the outcome of any one trial is not affected by the outcome of another trial.

Eg : Whether or not you get a "head" or a "tail" in this toss of a coin should have absolutely no bearing on the outcome of the next toss.

Clearly, a probability should be between 0 and 1; however, nearly all probability values lies between 0 and 1. In fact, the probability that the sun will rise tomorrow is NOT 1, but 0.999999....!

4.3 Sample Spaces and Events

The first step in finding the probabilities is to list all the possible outcomes of a random phenomena. The set of all possible outcomes of a random phenomenon is called the **sample space.**

Eg : (i) If you roll a die once, the sample space will be

(ii) If you toss a coin once, the sample space will be ; if you toss it twice, the sample space will be

(iii) Suppose I give a surprise pop quiz containing 3 multiple choice questions. Each question has 5 options and your answer can be either correct (C) or incorrect (I). The sample space of a student's response can be obtained by drawing the following tree diagram :

So, the student's performance will have possible outcomes given by

Often we need to specify a specific group of outcomes in the sample space, known as the **event**. Events are generally denoted by A, B, C etc.
Eg : Subset of outcomes for which student answers all 3 questions correctly :

4.4 Finding Probabilities of Events

Each outcome in the sample space has a probability attached to it; so does each event. The probabilities of outcomes in the sample space must follow the following two rules

- The probability of each individual outcome is between 0 and 1.
- The sum total of all individual probabilities is 1 i.e the probability of the sample space is 1.

In order to find the probability of a particular event, A :

- Find the probability of each individual outcome in the sample space.
- Add the probabilities of each outcome that the event contains i.e

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Number of outcomes in the sample space}}$$

The above definition assumes that the outcomes of the random phenomenon are **equally likely.**
Eg : Suppose a researcher is conducting a randomized experiment to compare a herbal remedy and placebo for treating common cold. The response variable are the cold's severity and duration. There are 4 volunteers, two men (Anuj and Rishi) and two women (Swati and Prabha). Two of these volunteers will be randomly chosen to receive the herbal remedy and the rest two will receive the placebo.
Q1. Identify the possible samples to receive the herbal remedy.

This is the sample space for randomly choosing 2 out of 4 people.
Q2. For each possible sample, what is the probability that it is chosen ?
Since each possible sample is equally likely, each of them has probability of of being chosen.
Clearly the total probability of the sample space is
Q3. What is the probability of the event that a man-woman pair will receive the herbal remedy ?
The event in which a man-woman pair will be selected consists of the outcomes

Since each outcome has probability , the probability of this event is.

4.4.1 Different Types of Events

1. **Complement of an event:** For an event A, its complement, A^c consists of all outcomes in the sample space NOT in A i.e

$$P(A^c) = 1 - P(A)$$

Eg : Let A = event that it will rain tomorrow; so A^c = event that it will not rain tomorrow. Suppose, $P(A) = 0.30$, then $P(A^c) =$

2. **Disjoint events** : Two events, A and B are said to be disjoint if they do not share any common outcomes.

Eg : For the pop quiz example, the event that the student answers exactly one question correctly is while the event that he/she answers exactly two questions correctly is . These are disjoint events since they have no outcomes in common.

Clearly, an event (A) and its complement (A^c) are disjoint events.

3. **Intersection of two events :** The intersection of two events A and B consists of outcomes that are in both A and B.

Eg : For the pop quiz example, let

A : event that a student answers first question correctly =

B : event that a student answers two questions correctly =

The intersection of A and B is . Thus, P(A and B) =

Note : In the special case when A and B are **independent**,

$$P(A \text{ and } B) = P(A) \times P(B)$$

Eg : Suppose a basket ball player shoots two free throws - he has a 80% chance of making each of them. Assuming that the two throws are independent, the probability that he will make both of them is

4. **Union of two events :** The union of two events, A and B consists of outcomes that are in A *or* B. In probability, this means "in A or in B or in both A and B".

Thus,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are disjoint, P(A and B) = , hence P(A or B) = .

Eg : Consider a family with two children. The sample space of the possible genders are

Thus,
A : first child is a girl =
B : second child is a girl =
(A and B) : both are girls =

Assuming that the possible outcomes in the sample space are equally likely, we have, $P(A) =$, $P(B) =$ while P(A and B) = .

Thus, (A or B) : event that first child is a girl or second child is a girl or both = at least one child is a girl. Then, P(A or B) = P(A) + P(B) - P(A and B) =

Note : Events are rarely independent. So, please give careful thought about the situation at hand before assuming independence.

Eg : Suppose the pop-quiz has only two questions and the probabilities of the events are : P(II) = 0.26, P(IC) = 0.11, P(CI) = 0.05 and P(CC) = 0.58. Let

A = {first question correct} = { } i.e P(A) =

B = {second question correct} = { } i.e P(B) =

while P(A and B) = . However, $P(A) \times P(B) =$

This is because, A and B are NOT independent. (Responses to different questions on a quiz are typically not independent. Students who gets the first question right are more likely to get the second question correct than those gets the first question wrong).

4.5 Conditional Probability

Often, we need to find the probability of a particular event conditional on the outcome of another event. These kind of probabilities are known as **conditional probabilities**. Most commonly, these are used to find the probabilities of a category of one variable (for instance, a person having a disease) given the outcome of another variable (the person having tested positive for that disease).

Definition : For events A and B, the conditional probability of A, given that B has occurred is

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

where $P(A \bigcap B)$ is the same as P(A and B).

Eg : April 15 is tax day in the U.S - the deadline for filing federal income tax forms. The main factor in the amount of tax owed is a taxpayer's income level. The following is a contingency table that cross-tabulates the 80.2 million long-form federal returns received in 2010 by the taxpayer's income level and whether the tax form was audited.

	Audited		
Income level	Yes	No	**Total**
Under \$ 25,000	90	14010	**14100**
\$ 25,000-\$ 49,999	71	30629	**30700**
\$ 50,000-\$ 99,999	69	24631	**24700**
\$ 100,000 or more	80	10620	**10700**
Total	**310**	**79890**	**80200**

Let us look at the following questions based on this table :

- What is the probability that a randomly selected tax-payer will be audited ?

- What is the probability that a randomly selected tax-payer will belong to the highest income bracket (i.e $\geq 100,000$) ?

- What is the probability that a randomly selected tax-payer will belong to the highest income bracket *and* will be audited ?

  Let A = {audited = Yes} and B = {income ≥ 100,000}. So, the above event will be the of A and B i.e . So, it's probability will be

- Given that a taxpayer has income ≥ 100,000, what is the probability that he/she will be audited ?
  Given B, the probability of A is the proportion of A and B cases out of the B cases. This will be

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} =$$

4.5.1 Multiplication Rule

We have seen that if A and B are independent events, P(A and B) = P(A)×P(B). However, the definition of conditional probability provides a more general formula for P(A and B) that holds *regardless of whether A and B are independent or not.* Thus, rewriting the conditional probability formulas given above, we have

$$\text{P(A and B)} = \text{P(B)}\times P(A|B)$$

Eg : In the 2006 Wimbledon tournament, Roger Federer faulted on his first serve 44% of the time. Given that he faulted on his first serve, he faulted on his second serve only 2% of the time. Assuming that these are typical of his serving performance, what is the probability that he makes a double fault ?

4.5.2 Concept of Independence

Two events A and B are independent if the probability that one occurs is not affected by the occurrence of the other event. This can be expressed more formally using conditional probabilities as follows :

Events A and B are independent if $P(A|B) = P(A)$, or equivalently, if $P(B|A) = P(B)$. If one of these holds, the other does too and either of these implies P(A and B)=P(A)×P(B).

Eg: For the children example done before, A = {first child is a girl} and B = {second child is a girl}. Then P(A)=P(B)=1/2 while P(A and B) = 1/4. Applying the definition of conditional probability, we have

Thus, $P(B|A) = P(B)$ and hence A and B are independent events.

Eg : Following is a contingency table showing the results of a diagnostic blood test (POS=positive; NEG = negative) to whether or not a woman's fetus is affected with Down syndrome. The study is based on the results from 5282 women aged 35 or over.

	Blood test		
Disease status	POS	NEG	**Total**
Affected (A)	48	6	54
Not affected (NA)	1307	3921	5228
Total	1355	3927	5282

Let us try to answer the following questions :

- What is the probability that a woman will have a positive test result i.e P(POS)?
- What is the probability that a woman has Down syndrome i.e P(A) ?
- What is the probability that a woman has Down syndrome and tests positive i.e P(A and POS) ?
- Are the events "POS" and "A" independent or dependent ?

4.6 Probability Distributions

The possible outcomes of a random phenomenon and the corresponding probabilities are summarized using what are known as **probability distributions** as we will learn now. Before delving into probability distributions, we need to understand the concept of random variables.

4.6.1 Random Variables

The characteristics we measure for different subjects or units in a sample are called **variables**. For example, our height, weight, scores obtained in an exam etc. Generally, the values of a variable are the result of a random phenomenon like selecting a random sample from a population or performing a randomized experiment. In those cases, the variable is known as **random variable.** Thus, a **random variable** is a numerical measurement of the outcome of a random phenomenon.

Eg : The number of heads in 3 flips of a coin is a random variable. We denote a random variable using big caps (say, X) and its values with small caps (say, x). Since each possible outcome of a random phenomenon (i.e each possible values of a random variable) has a probability attached to it, we use the **probability distribution** of a random variable to specify its values and the corresponding probabilities.

4.6.2 Discrete Random Variable

A discrete random variable, say X, takes distinct values, say, $\{0, 1, 2, ...\}$. The probability distribution of a discrete random variable assigns a probability, say $P(x)$ to each possible value, x such that

- For each $x, 0 \leq P(x) \leq 1$.
- The sum total of all probabilities over all x values is 1 i.e $\sum_x P(x) = 1$.

Eg : The following table depicts the probability distribution of the number of children in middle class Indian families for 2016.

Number of children	0	1	2	3	4	5	6 or more
Probability	0.13	0.38	0.32	0.13	0.03	0.01	0.00

Here the discrete variable is with values

Q1. Are the properties of probability distribution satisfied for the above table ?

Q2. What is the probability that a middle class Indian family will have at least 3 children ?

As before, we can use the **mean** and **standard deviation** (or **variance**) to describe the center and spread of a probability distribution. We obtain the **mean**(μ) by multiplying each possible value of the random variable by the corresponding probability and summing over all possible values i.e

$$\mu = \sum_x xP(x)$$

The mean of the probability distribution of the number of children is

This value reflects the average number of children per Indian family.

The mean (μ) is also known as the expected value of X since it reflects the average value of X in a long run of observations. It is also a **weighted average** since every value x is weighed by its corresponding probability $P(x)$ i.e values that are more likely receive greater weight and vice versa.

The **Standard Deviation** (σ) of a probability distribution measures the spread of a probability distribution. Larger values of σ implies greater spread and vice versa. It is given by

$$\sigma = \sqrt{\sum x^2 P(x) - (\sum xP(x))^2}$$

For the above example, the standard deviation will be

This value reflects the typical deviation of the number of children for a particular middle class household from the average number of children across all household.

4.6.3 Continuous Random Variable

Continuous variables are those whose values fall within an interval. For example, your height, playing time of a music etc. Probability distributions of continuous random variables assign probabilities to any interval of the possible values a random variable can take. In fact, if the intervals are very narrow, a smooth curve can be used to approximate the probability distribution.

- The probability that the random variable takes values in a particular interval is between 0 and 1 and is equal to the area under the curve above that interval.
- The interval containing all possible values has probability 1 - so, the total area under the curve equals 1.

Eg : A 2017 report by a Bangalore-based NGO suggested that 65% of IT workers have a commuting time to work more than 45 minutes. The following graph shows the probability distribution of commuting times

The area under the total curve is and the area under the curve corresponding to values greater than 45 is .

Chapter 5

Sampling Distribution

5.1 Introduction

Statistical inference is all about using a sample to predict characteristics of a population. The sample characteristics are summarized using the **sample statistics** while the characteristics of a population are summarized by the population **parameters**. *Thus, statistical inference boils down to estimating a population parameter using analogous sample statistic.*

Eg 1. The traditionally low percentage of female students in IIMA has been a topic of heated debate for quite some time. You can estimate the percentage of female students in the current batch (i.e PGP, PGPX, FPM, FDP, AFP combined) using the corresponding percentage (of females) in a random sample of students you may select from the current batch. In this case, the *parameter* is the true (unknown) proportion of females in the population of all students currently enrolled in IIMA while the *statistic* is the sample proportion of female students in your sample.

Two most commonly used statistics are the **sample mean** ($\bar{x}$) and **sample proportion** ($\hat{p}$). The corresponding parameters are the **population mean** (μ) and **population proportion** (p).

Clearly, in order for the inferential procedure to be "good", the sample statistic should be pretty close to the unknown population parameter. In order to ensure that, we have to study the **sampling distribution** of sample statistics.

Since sample statistics are based on samples, they will have different values for different samples. Thus, a statistic is also a random variable and will have a probability distribution that will assign probabilities to the possible values it can take for the different samples. *The probability distribution of a sample statistic is called a sampling distribution.*

Eg 1. contd Suppose the true proportion of females in the current batch of students is $p = 0.21$. Suppose you select all possible random samples of 20 students - each of those samples will yield a value of the sample proportion (of females in that sample). If you construct a histogram of those values, what you will get is precisely the sampling distribution of the sample proportion (of females).

Eg 2. IAS Officers A few of years ago, a female IAS officer was suspended by the UP government for her actions against the sand mafias. Suppose you want to estimate the average tenure of an IAS officer (at a particular posting). For that purpose, suppose you select random samples of 20 IAS officers for the pool of all IAS officers currently in service. For each sample, you calculate the average tenure of an officer. If you continue doing this for a large number of samples and draw the histogram (of the sample average tenure value of an IAS officer), what you will have is precisely the sampling distribution of the sample mean tenure of IAS officers[1].

Having said all the above, in reality, it is impossible to collect all possible samples and calculate the sample statistic value for each. *In fact, in a real-life study, we can only collect one sample.* However, the theory of sampling distribution will tell us how much a statistic would vary from sample to sample and will help us to predict how close a statistic will be to the parameter it estimates.

5.1.1 Necessary tools

Before delving into some examples, let us briefly discuss a couple of important concepts which are essential for a proper understanding of sampling distribution.

1. **Empirical rule** : Suppose the mean and standard deviation of a sample is $\bar{x}$ and s respectively[2] If it can be assumed that the sample comes from a distribution that is approximately bell-shaped and symmetric, then

   - of the observations would fall within 1 standard deviation of the mean i.e within $(\bar{x} - s, \bar{x} + s)$.
   - of the observations would fall within 2 standard deviation of the mean i.e within $(\bar{x} - 2s, \bar{x} + 2s)$.
   - of the observations would fall within 3 standard deviation of the mean i.e within $(\bar{x} - 3s, \bar{x} + 3s)$.

   This is important to know because a lot of variables observed in reality have approximately

[1] Here the population is the set of ALL IAS officers currently on duty; the parameter is the true average tenure of an IAS officer which is unknown (since we cannot survey each and every IAS officer). In fact, a recent Harvard study has concluded that the average tenure of an IAS officer is only 16 months !

[2] s is known as the sample standard deviation and is given by $\sqrt{\frac{1}{n-1}(x_i - \bar{x})^2}$, $x_i, i = 1, ..., n$ being the sample units and n the sample size.

bell shaped distributions.

2. **Central Limit Theorem (CLT)** : This is a very well known result in Statistics which basically says that as the sample size increases (i.e as we take larger and larger samples), the sampling distribution of the sample statistic (mean or proportion) tends to a **Normal distribution**. In fact this holds true even if the population (from which the sample is drawn) is moderately skewed or discrete. The only restriction is that the mean and standard deviation of the population distribution should exist.

   For all practical purposes, $n \geq 30$ is good enough to ensure the normality of the sample mean $(\bar{X})$ if the population distribution is not too skewed. Needless to say, larger the sample size, better will be the approximation.

Now, we will discuss the sampling distribution of sample means and proportions in some detail.

5.2 Sampling Distribution of the Sample Mean

Since means or averages are so ubiquitous in Statistics, it is useful to learn about their sampling distribution i.e how sample means vary from sample to sample and how close they will be to the population mean in repeated sampling from the population. However, in reality, it is impossible to collect many samples and obtain the distribution of the sample means manually. The following result provide us with an automated way of achieving the same even when we collect only one sample.

Suppose you draw samples of size n from a population with mean μ and standard deviation σ and calculate the sample mean $(\bar{x})$ for each. Then the means will have a sampling distribution which will be centered around the true population mean μ and will have a standard deviation (or standard error) $\sigma/\sqrt{n}$. In fact, if $n \geq 30$, then this distribution will be approximately with mean μ and standard error $\sigma/\sqrt{n}$.

In the light of this result, let us consider the following example :

Eg 3. Rambhai's income : The sales of food and drink in Rambhai's stall (located adjacent to the boundary wall of IIM, a little to the left of the mail gate) vary from day to day. The daily sales figures fluctuate with mean $\mu =$ Rs 900 and standard deviation $\sigma =$ Rs 300. Suppose Rambhai wants to calculate the mean daily sales for the week to check how he is doing.

Q1. What would the mean daily sale figures for the week center around ?

Q2. How much variability would you expect in the mean daily sales figures for the week ? Interpret.

Thus, if Rambhai were to observe the mean daily sales for several weeks, those will center around with a standard deviation .

Q3. Suppose Rambhai now wants to look at the monthly sales. What will be the sampling distribution ? Will his mean daily sales for the month vary more or less than the mean daily sales for the week ?

His mean daily sales for the month will be centered around with standard error . Thus, the mean daily sales for the month will tend to vary than the mean daily sales for the week and thus be closer to the true mean sales of Rs 900. *In fact, since $n = 30$, CLT holds and the sampling distribution of the mean daily sales for the month is approximately with mean and standard error.*

Clearly there is variability in the mean daily sales from month-to-month than from week-to-week than there is from day-to-day in the daily sales.

Q4. What is the probability that the mean daily sales of the month will be between 800 and 1000 Rupees ?

5.3 Sampling Distribution of the Sample Proportion

If $\hat{p}$ is the sample proportion for a random sample of size n drawn from a population with proportion p, then $\hat{p}$ has mean and standard deviation

Note : (i) If n is sufficiently large such that both np and $n(1-p)$ are at least 10, then this sampling distribution is approximately normal due to CLT.
(ii) The standard deviation of a sampling distribution is known as the **standard error**. So, the standard error of the sampling distribution of $\hat{p}$ is

Eg 4. Internship : Suppose out of all first year students enrolled in the top business schools across India, about 55% went abroad for summer internship last year. Suppose you randomly select a business school and it turns out to be XLRI which has about 350 students enrolled in the first year.

a) What is the sampling distribution of the proportion of females in your sample ?

b) What is the probability that at least 50% of the 350 XLRI students will go abroad for internship this year ?

c) What is the probability that at most 70% of the 350 XLRI students will go abroad for internship this year ?

d) What is the probability that between 50% and 70% of the 350 XLRI students will go abroad for internship this year ?

Note :

(i) Clearly, the standard error will decrease as you your sample size. For example, if you select a larger B school which has 500 students, the standard error of $\hat{p}$ will be .

Thus, smaller the standard error, closer will be the sample proportion to the population proportion.
(ii) If the sampling distribution is approximately normal, we can use the **Empirical rule**. For example, in the above example, nearly all the sample proportions (corresponding to all possible samples/B-schools of size 350) will lie between

Chapter 6

Confidence Intervals

6.1 Introduction

The process of making decisions and predictions about one or more **population parameters** using the corresponding **sample statistics** (obtained from a randomly selected representative sample from the population) is called Statistical Inference. Broadly, there are three ways of making statistical inference as follows :

1. **Point Estimation** : Here we put forward a single estimate (usually the sample statistic obtained from a random sample) for the population parameter.
   Eg : The proportion of vegetarians in a random sample of 50 IIMA students can be a point estimate of the corresponding proportion in the population of all IIMA students.

2. **Interval Estimation** : Here we form an interval containing the most plausible values of the population parameter and within which the parameter is believed to lie with a certain confidence. Unlike point estimates, interval estimates give us an idea of the precision of our estimates.
   Eg: The interval (5.1, 5.9) may be a 95% confidence interval of the true average height of an Indian adult.

3. **Hypotheses Tests** : This is an inferential procedure that yields a decision on whether a claim about the value of a parameter (framed in terms of hypotheses) is supported by data observed from a random sample.

An important part of any confidence interval is the **Confidence level -** it is the level of confidence with which the interval actually contains the true parameter value. It is usually chosen to be very close to 1 with $0.90, 0.95$ and 0.99 being the commonly used values and is denoted by $1-\alpha$, where α can be $0.1, 0.05$ or 0.01 respectively for the above confidence levels.

Thus, if (a, b) is a 95% confidence interval of a parameter, say θ, then we can be 95% confident that (a, b) contains the true unknown value of θ in the population.

What this statement really implies is that, if we go on collecting a large number of random samples from the population and form a 95% confidence interval from each of those, then, in the long run, about 95% of those intervals would contain the true population parameter and the rest 5% will not.

Confidence intervals generally have the form :

where the margin of error measures the accuracy with which the sample statistic estimates the unknown population parameter.

Now, we will separately discuss confidence intervals for population means and proportions.

6.2 Confidence Interval for Population Proportion

- Unknown parameter :
- Sample statistic :
- Standard error (se) of $\hat{p}$:
- Estimated se of $\hat{p}$:
- Margin of error :

where $Z_{\alpha/2}$ is that value of the standard normal variable above which the area under a standard normal curve is $\alpha/2$. So, for a 90%, 95% and 99% confidence interval, the corresponding $Z_{\alpha/2}$ will be and - these values are 1.645, 1.96 and 2.58 respectively.

Thus the resulting $100(1-\alpha)\%$ confidence interval of p will be

However, the above confidence interval is valid only under the following assumptions

1. Random sample of observations from a distribution.
2. # success ≥ 10; # failures ≥ 10

Note 1. *(i) For fixed confidence level, as sample size increases, standard error and hence the margin of error . So, the confidence interval gets*
(ii) For fixed sample size, as confidence level increases, the Z-score ; hence the margin of error . Thus the confidence interval gets

Eg : NDTV randomly selected 1000 final year students across different management schools in India and asked them about their career choices. 4% said they want to take the plunge and start their own companies even if that meant giving up lucrative job offers from established MNCs. Find a 99% confidence interval of the true population proportion of management students in India who want to work their start-ups.

Conclusion : We can be 99% confident that between and of all final year management students in Indian B-schools want to work on their start-ups right after graduating.

(i) A 95% confidence interval of p will be :

(ii) A 90% confidence interval of p will be :

Thus, comparing the above three confidence intervals, it is obvious that the intervals get with increasing confidence level.

6.3 Confidence Interval for Population Mean

- Unknown parameter :
- Sample statistic :
- Standard error (se) of $\bar{X}$:
- Estimated se of $\bar{X}$: where S is the sample standard deviation.
- Margin of error :

Note 2. *A t-distribution is similar to the standard normal distribution but with thicker tails and it approaches the standard normal distribution as the degrees of freedom increases. Here, $t_{\alpha/2,n-1}$ is that value of the t-variate such that the area above it under a t curve with $n-1$ degrees of freedom is*

The resulting $100(1-\alpha)\%$ confidence interval of μ will be

The above confidence interval is valid under the following assumptions :

1. Random sample
2. Population distribution approximately normal

Following is a section of a t-table :

	Confidence Level		
	90%	95%	99%
	Right tail probability		
df	$t_{.05}$	$t_{.025}$	$t_{.005}$
1	6.314	12.706	63.656
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169

Eg : A recent survey asked 899 randomly selected college students “On an average day, about how many hours do you watch TV ?” The sample mean was 2.865 and standard deviation was 2.617. Find a 95% confidence interval of the population mean number of hours per day college students watch TV.

Thus, we are 95% confident that the average number of hours an Indian college student watch TV per day is between

(i) 90% confidence interval of μ :

(ii) 95% confidence interval of μ :

6.4 Sample Size Determination

In many real-life studies, we often know the level of precision and need to determine the sample size that will yield that precision. This is doable because precision is measured by the margin of error of a confidence interval which in turn is a function of the sample size. Now we will discuss how this is done for the case of proportion.

The sample size (n) for which a confidence interval of a population proportion has margin of error m is

As usual, the Z-score depends on the confidence level (i.e Z = 1.96 for a 95% confidence interval). The value of $\hat{p}$ can either be determined from past experience/studies or can be assumed to be 0.5.

Eg : A study by a social sciences institute concluded that 19% of university students in India do not want to go abroad for jobs even if they are presented with the opportunity. To estimate this proportion in the IIMs with precision 0.05 with a 95% confidence interval, what should be your sample size if

a) You have absolutely no idea of the proportion in the IIMs.

b) You use the social science study as your guideline

Now, what will be the sample size for a

- 90% confidence interval :

- 99% confidence interval :

Note 3. *i)* $p = 0.5$ *will always result in a larger (i.e more conservative) sample size estimate. ii) Always round the sample size estimate to the next higher integer.*

Chapter 7

Testing of Hypotheses

7.1 Introduction

Significance tests (or Tests of Hypotheses) is an integral part of statistical inference along side point and interval estimation. Any significance test procedure has five (5) distinct steps viz.

1. **Making assumptions** : simple random sampling etc.

2. **Constructing hypotheses** : Each significance test is composed of two hypotheses

   - **Null hypotheses (H_0)** : It is a statement that specifies a particular value for the parameter (in our case, p or μ) which is pre-determined from *experience and/or prior belief*. **Eg :** You believe that on an average an IIMA student spends 15 hours per week discussing/solving cases outside of class i.e the null hypotheses will be

   - **Alternative hypotheses (H_a)** : It states that the population parameter takes values in some alternative parameter space (than what is stated by the null). It may be **one** or **two sided**.
     Eg : Suppose on the contrary, your friend believes that an IIMA student spends more than 15 hours per week solving cases i.e the alternative hypotheses will be
     The above alternative is a one-sided one. However, two-sided alternatives are also possible.

   Here μ is the population mean number of hours an IIMA student spends solving cases per week outside of class. Similar one/two-sided hypotheses can be framed for the population proportion p.

3. **Determining the test statistic** : A test statistic measures how close the point estimate of the population parameter is to the null hypotheses value (of the parameter). This "closeness" is measured in terms of the **standard error** of the point estimate. Thus, the test statistic is given by

4. **P-values** : p-values represent the amount of evidence against the null hypotheses based on the available data. *Smaller the p-value, stronger is the evidence against the null hypotheses and vice versa.* The smallness of the p-value is measured with respect to the significance level (α).

5. **Drawing conclusion** : In order to come to a definite conclusion (about rejecting or not rejecting the null hypotheses), we compare the p-value with the **significance level** (denoted by α). The significance level is usually set at 0.05, 0.1 or 0.01 *and it is chosen by the statistician.* We would reject H_0 at a given significance level α, if

   and fail to reject H_0 otherwise (i.e if p-value α).

Now we will separately discuss significance tests for population proportion (p) and mean (μ).

7.2 Significance Tests for Population Proportion

Let us go through the steps sequentially :

1. **Assumptions** :
   - Simple random sample.
   - Sample size (n) should be large enough such that $np_0 \geq 10, n(1 - p_0) \geq 10$.

2. **Hypotheses** :
   - **Null** : $p = p_0$
   - **Alternative** :

   where p_0 is known as the *null value* of p.

3. **Test statistic**:

4. **P-value** : The p-value would depend on the *direction of the alternative* as follows :
   - If $H_a : p > p_0$, p-value will be the **right** tailed area **above** the observed value of the test statistic (Z_{obs}) under the standard normal curve.

- If $H_a : p < p_0$, p-value will be the tailed area **below** the observed value of the test statistic under the standard normal curve.

- If $H_a : p \neq p_0$, p-value will be the tailed area beyond the observed value of the test statistic under the standard normal curve. Since the normal curve is symmetric, it can also be calculated as twice the one-tailed area above (or below) the observed value of the test statistic.

5. **Drawing conclusion** : We will reject H_0 if and fail to reject H_0 otherwise.

Eg 1: **Female managers** : Traditionally, the percentage of managers who are female in the Indian corporate sector has been pretty low, about 18%. The HRD ministry wants to know whether the percentage has improved during recent times. Accordingly, a random sample of 100 managers were chosen and 25 of them were females. Perform an appropriate test of hypotheses for the above problem.

Let p be the unknown population proportion of female managers. Let us go through the steps one by one.

1. **Assumptions** :
   - Since the 100 managers were chosen randomly, the random sampling assumption is preserved.
   - Here the null value is $p_0 =$. So $np_0 =$ and $n(1 - p_0) =$.

   Thus, all our assumptions are satisfied and we can proceed with the test.

2. **Hypotheses** :
   - **Null** :

- **Alternative** :

3. **Test statistic**:

   This implies that the sample proportion (0.25) falls standard errors the null value 0.18. Now, we will check whether this is good enough evidence to reject H_0.

4. **P-value** : The observed value of our test statistic is . Since the alternative is , the p-value will be the tailed area above under the Z curve i.e

   From the normal tables, the p-value will be

5. **Conclusion** : Let us choose a significance level of 0.05. Since , we H_0 at $\alpha = 0.05$. Thus, there is significant evidence (at $\alpha = 0.05$) that the proportion of female managers in the Indian corporate sector has increased in recent times.

7.3 Significance Tests for Population Mean

As for proportions, significance tests for means also have five distinct steps viz.

1. **Assumptions** :
   - Simple random sample.
   - Population distribution approximately
2. **Hypotheses** :
   - **Null** : $\mu = \mu_0$
   - **Alternative** :

   where μ_0 is the *null value* of μ.

3. **Test statistic**:

where $\bar{X}$ and S are respectively the point estimates of μ and σ and t_{n-1} denotes a t distribution with $(n-1)$ degrees of freedom. As for confidence intervals, our test statistic follows a t distribution under H_0 since we replace σ by S.

4. **P-value** : The p-value would depend on the *direction of the alternative* as follows :

   - If $H_a : \mu > \mu_0$, p-value will be the **right** tailed area **above** the observed value of the test statistic under a t_{n-1} curve.

   - If $H_a : \mu < \mu_0$, p-value will be the **left** tailed area **below** the observed value of the test statistic under a t_{n-1} curve.

   - If $H_a : \mu \neq \mu_0$, p-value will be the **two** tailed area beyond the observed value of the test statistic under a t_{n-1} curve. Since the t distribution is symmetric, it can also be calculated as twice the one-tailed area above (or below) the observed value of the test statistic.

5. **Drawing conclusion** : We will reject H_0 if and fail to reject H_0 otherwise.

Eg 2. Based on past records, it is generally believed that on an average, a typical IIMA student spends about 25 hours in the Vikram Sarabhai library per week. Recently, the library has been shifted to a new location in KLMDC which is further away from the dorms. As a result, the administration feels that students may be spending less time in the library. Accordingly a random sample of 41 students were selected and the average number of hours they spend in the library came out to be 16.78 with a standard deviation of 5.17. Carry out an appropriate test of hypotheses for the above problem to test whether the shifting of the library has adversely impacted the study time in the population of all students. (You may assume that study times/week approximately follow a normal distribution in the population).

1. **Assumptions** :
   - 41 students were selected randomly - so this is a simple random sample.
   - Population distribution of study times is Normal
2. **Hypotheses** :
   - **Null** :
   - **Alternative** :

   where μ is the population mean number of hours/week a student spends in the library.
3. **Test statistic**:

   which follows a t distribution with df under H_0. This implies that the sample mean ( ) falls standard errors the null value of 25. Now, we will check whether this is good enough evidence to reject H_0.
4. **P-value** : Since the alternative is <, the p-value will be the tailed area below under the t curve with df i.e

   Following is a small part of the t table :

	Area to the right					
df	.100	.050	.025	.010	.005	.001
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261

5. **Drawing conclusion** :

7.4 Errors in Hypotheses Tests

Two types of errors may result from drawing conclusions from hypotheses tests :

- **Type I error** : We commit a type I error when we mistakenly **reject** H_0 when it is **true**. In that case,

  $$\text{P(type I error)} = \text{significance level } (\alpha)$$

  Thus, we control the probability of type I error by our choice of the significance level. In practice, $\alpha = 0.05$ is most common.

  Eg 2 : Suppose for the library example, you fix the significance level to be 0.05. Then, the probability that we will conclude $H_a : \mu < 25$ when in fact the average study time has increased is

  In other words, the probability of taking a correct decision i.e $H_0 : \mu = 25$ is

  As α decreases, we are to reject H_0 i.e P(type I error) goes down.

- **Type II error** : We commit a type II error when we **fail to reject** H_0 even when it is **false**. It is usually denoted by β.

  When H_0 is false, we would want the probability of rejecting it to be as as possible. The probability of rejecting H_0 when it is false is called the **power** of the test. It is given by :

  $$\text{Power} = 1 \text{ - P(type II error)} = 1\text{- } \beta$$

  Obviously, higher the power, is our test.

  Eg 3 : Suppose for the female managers example, the true proportion of managers who are females has increased significantly from 0.18. However, we still conclude the null hypotheses (i.e $H_0 : p = 0.18$) with probability 0.01. Then the power of our test will be

Chapter 8

Comparison of Two Populations

8.1 Introduction

In the previous lectures, we learned how to construct confidence intervals of and test hypothesis corresponding to **one** population. Now, we shall extend the above method to **two** groups or populations i.e we will learn how to construct confidence intervals and hypothesis tests on the **difference of proportions** corresponding to two populations. By doing this, we can compare the characteristics of the subjects belonging to these two groups.
Eg. We want to compare the proportions of male and female in the India who believe in miracles. Here the two groups/populations are respectively the populations of all and in India while the characteristic we want to compare (between the two populations) is

The two groups mentioned above can be compared with regard to a categorical or quantitative outcome. For **categorical** outcomes, we compare population **proportions** and for **quantitative** outcomes, we compare population **means**. Here we will only deal with techniques for comparing population **proportions** across two groups.

8.2 Comparing Two Population Proportions

8.2.1 Confidence Interval

1. **Notation** :

   - $p_1(p_2)$: population proportion of success in the first(second) group.
   - $n_1(n_2)$: sizes of random samples drawn from the first (second) populations.
   - $\hat{p_1}(\hat{p_2})$: corresponding sample proportions in the first (second) group.

2. **Assumptions** :

   - Independent random samples from the two groups.

- Large enough sample sizes so that in each sample there are at least 10 success and 10 failures.

These will ensure that the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ is approximately under the CLT.

3. **Structure** : As for one proportion, the confidence interval for $(p_1 - p_2)$ is obtained by adding and subtracting a to its point estimate $(\hat{p}_1 - \hat{p}_2)$. As before, the margin of error is the product of the Z-score and the estimated standard error of $(\hat{p}_1 - \hat{p}_2)$ given by

$$\hat{se}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

So, for a $100(1 - \alpha)\%$ confidence interval, the margin of error will be

resulting in the confidence interval

4. **Observations :**

- If the confidence interval contains only positive (negative) values, we can conclude (with the appropriate confidence) that
- If the confidence interval contains 0, we conclude (with the appropriate confidence) that p_1 and p_2 are not significantly different.

Eg 1. Let us go back to the "belief-in-miracles" example. Suppose 523 males and 498 females were randomly sampled and each of them were asked the question "Do you believe in miracles?". The following table summarizes the observations. We want to compare the population proportion of males and females who believe in miracles using a 95% confidence interval.

	Belief in miracles		
Gender	Yes	No	Total
Male	225	298	523
Female	276	222	498

(*The above table is called a 2 x 2 contingency table. It cross-classifies the observations corresponding to the response and explanatory variables*).

1. **Assumptions** : Here the assumptions are satisfied because (i) the samples of males and females were chosen randomly and ii) the number of successes (belief in miracles) and failures (no belief in miracles) in each sample (male and female) are much higher than 10.

2. **Structure** : Here

$$\hat{p_1} = \qquad \hat{p_2} = \qquad , \hat{se}(\hat{p_1} - \hat{p_2}) =$$

Hence the required 95% confidence interval of $p_1 - p_2$ will be

3. **Conclusion** : Since the above confidence interval only contains negative numbers, we can be 95% confident that the population proportion of males who believes in miracles is between to than the population proportion of who females believe in miracles.

8.2.2 Hypothesis Tests

Significance test is another avenue through which the comparison of two groups can be carried out. As for confidence intervals, the general methodology for significance tests is perfectly analogous for the one and two sample case.

1. **Notation** : same as for confidence intervals.
2. **Assumptions** : same as for confidence intervals.
3. **Hypotheses** : Similar to the case for single proportion, we have to formulate two hypothesis for $(p_1 - p_2)$: a **null** hypothesis based on our experience and an **alternative** one challenging our belief. These can be expressed as:
   - $H_0 : p_1 - p_2 = 0$
   - H_a :
4. **Test statistic** : The test statistic is given by

$$Z = \frac{\hat{p_1} - \hat{p_2} - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

   where $\hat{p}$ is the pooled estimate of p i.e the common population proportions of success given by

$$\hat{p} = \text{total \# of success in the two samples/total sample size}$$

   The above standard error is obtained from the estimated standard error of $\hat{p_1} - \hat{p_2}$ by replacing both $\hat{p_1}$ and $\hat{p_2}$ by $\hat{p}$. Analogous to the one proportion case, the above test statistic would measure the number of that separate $\hat{p_1} - \hat{p_2}$ from the null value 0.
5. **P-values :** As in the one proportion case, the p-value will be the one (for one sided alternative) or two (for two sided alternative) tailed probability of values even more extreme than the observed test statistic value.

6. **Rejection rule** : As before, we would reject H_0 at significance level α if and would fail to reject H_0 otherwise.

Eg 2. Let us go back to the "belief in miracles" example. We want to test whether there is any difference in the population proportion of males and females who believe in miracles.

1. **Assumptions** : All the assumptions have been satisfied.
2. **Hypotheses** :
   - H_0 :
   - H_a :
3. **Test statistic** : The total number of believers in the two samples taken together is and the total sample size is . Hence, the pooled sample estimate of believers will be . Hence the test statistic will be

   Thus, the $\hat{p_1} - \hat{p_2}$ falls about (estimated) standard errors below the null value of 0.
4. **P-value** : Since our alternative is and our test statistic value is , the p-value will be the area under the standard normal curve above and below which is double the area below

   So, the p-value is
5. **Conclusion** : Since our p-value is less than all possible significance levels, we H_0 and conclude that there is strong evidence that the proportion of male and female believers in miracles are different in the population. In fact, since the test statistic is negative, the population proportion of male believers should be than the population proportions of female believers.

   Observe that the conclusions we have drawn from significance tests and confidence intervals match perfectly.

Chapter 9

Analysis of Variance (ANOVA)

9.1 Introduction

In Chapter 7, we have learnt how to compare two groups (or populations) with respect to a categorical response variable. However, in reality, we may have to compare a characteristic across multiple (i.e more than 2) groups. The statistical technique through which this is accomplished is known as **Analysis of Variance** or **ANOVA**.

Eg 1 : A recent news article in the *Times of India* reported that in the current placement season, the average starting (domestic) salaries of graduating PGP students at IIMA have increased by a whooping 28%. Suppose, in view of this news, you want to compare the true (population) average starting salaries across the top five IIMs (IIMA, IIMB, IIMC, IIML and IIMK). Thus, we are comparing groups with respect to a response variable (salaries).

In ANOVA, the categorical explanatory variables identifying the groups are called **factors** while the individual categories being the **levels**. So, in the above example, there is one factor (IIMs) with levels (or categories) : . When there is only one factor, the technique is called **one-way ANOVA**, which will be the topic of this chapter.

9.2 One-way ANOVA

ANOVA is basically a hypotheses testing problem. As with any hypotheses test, it is based on certain assumptions as follows :

- Simple random samples from each group (or population).
- Response should follow an approximate in the population. However, this is mainly important for small samples ($n < 30$).

As mentioned above, suppose we want to compare a quantitative response variable across g groups. This is tantamount to comparing the population means of the response for the g groups (say

$\mu_1, \mu_2, ..., \mu_g$). Thus, the null and alternative hypotheses will respectively be

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = ... = \mu_g = \mu \\ H_a &: \end{aligned}$$

In order to test the above hypotheses i.e whether there is any significant differences in the population means, the following test statistic is used

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}} = \frac{\sigma_b^2}{\sigma_w^2} \tag{9.1}$$

Observations :

- Higher the variability between groups relative to the variability within groups, will be the evidence against the null and the test statistic (i.e F) value will be Thus, we will H_0 and would conclude that the means cannot be assumed to equal.
- If the variability between groups is similar to the variability within groups (i.e $F \sim$), we would H_0 and conclude that the means ($\mu_1, \mu_2, ..., \mu_g$) are not significantly different from each other.

Since the population variances between and within groups are generally unknown, we will not know the values of σ_b^2 and σ_w^2 in (1). So, we replace those by their estimates **Mean Squares Between (MSB)** and **Mean Squares Within (MSW)** (also known as **Mean Squares Error (MSE)**). Thus, the test statistic becomes

$$F = \frac{MSB}{MSW} \tag{9.2}$$

Now, MSB and MSE can again be represented as the ratios of the corresponding **Sums of Squares** and the degrees of freedom;

$$MSB = \frac{SSB}{g-1} \qquad\qquad MSW = \frac{SSW}{N-g}$$

Here, SSB is known the **Sums of Squares Between** while SSE is the **Sums of Squares of Errors** with degrees of freedom $g-1$ and $N-g$ respectively, N being the total combined sample size from all the groups. Thus, replacing these values in (2), the test statistic becomes

$$F = \frac{SSB/g-1}{MSW/N-g} \tag{9.3}$$

which follows a F distribution with $(g-1, N-g)$ degrees of freedom.

P-values : Since the F distribution is only defined on the positive axis, p-value will always be the tailed area above the observed F value. We can use the F table at the back of the book for that.

Decision rule : As always, we will reject H_0 is p-value and would fail to reject H_0 otherwise.

The output of a ANOVA procedure is summarized in an ANOVA which has the following form :

Source	Df	Sum of squares	Mean squares	F
Between groups	g-1	SSB	MSB = SSB/g-1	F = MSB/MSE
Within groups (Error)	N-g	SSE	MSE=SSE/N-g	
Total	N-1	SST		

Here SST is the Sum of Squares Total and is equal to $SSB + SSE$.

Eg 2 : Suppose a researcher want to compare 3 diet regimens (Low Fat, Low Cal and Low Carb) by the amount of weight loss they induce among subjects. Accordingly, she randomizes 10 subjects into each of these regimens and measure their weights before and after (they take the diet) and take the difference of the same. It can be assumed that weight loss has a normal distribution in the population.

So, the explanatory variable (or factor) is which has categories i.e while the response is . The assumptions are valid because

- The researcher randomizes the subjects into the 3 regimens.
- Weight loss has a normal distribution in the population.

We want to test whether the mean weight loss induced (in the population) by these 3 diets are the same or not. Thus, the null and alternative hypotheses will respectively be

H_0 : H_a :

Here, $n_1 = n_2 = n_3 =$, $N =$ and $g =$
Moreover, it can be shown that $SSB = 122.1$ and $SST = 182.85$. Thus,

- Df of SSB =
- Df of SSE =
- Df of SST =
- SSE =
- MSB =
- MSE =
- F =
- Df of F =

Based on the above values, the ANOVA table will be :

Source	Df	Sum of squares	Mean squares	F
Between groups				
Within groups (Error)				
Total				

P-value : From the F table at the back of the book, we have, for df (2, 27), the right tailed area above 5.49 is .01. Hence the area to the right of 27.13 will be i.e our p-value will be

Decision : Since the p-value is approximately 0, we would H_0 at all the commonly used significance levels i.e $\alpha = .01, .05, .1$.

Conclusion : At $\alpha = .01, .05, .1$, We have strong evidence to believe that the mean weight loss induced by the 3 diet regimens are not the same (or the mean weight loss induced by at least one diet regimen is significantly different from the rest).

Notations:

Let Y_{ij} be the value of the response variable for j^{th} observation in the i^{th} group, $i = 1, 2, ..., g$; $j = 1, 2, ..., n_i$. Then,

- Sample mean for the i^{th} group : $\bar{Y}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, $i = 1, 2, ..., g$.
- Overall sample mean : $\bar{Y} = \dfrac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} Y_{ij}$.
- Total sample size in all the groups combined : $N = \sum_{i=1}^{g} n_i$.
- Sum of squares between (SSB) : $\sum_{i=1}^{g} n_i(\bar{Y}_i - \bar{Y})^2$.
- Sum of squares within (or error, SSE) : $\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$.
- Sum of squares total (SST) : $\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$

Executive Program in Advanced Business Analytics

Problem set for Foundational Materials

These problems are based on the foundational materials. These will not be graded but you should work these out to gain a good understanding of the concepts

Q1. [**Surgical strike]** In the context of the Pulwama attack and the ensuing air strike in Balakot (Pakistan), suppose the next Prime minister of India wants to take a decisive action against Pakistan based terror groups, specially Jaish-e-Mohammad, with the sole aim of neutralizing it completely. Accordingly, the Cabinet Committee on Security, headed by the PM, decides that this "action" will be a top secret mission that will be designed in line with the now-famous "Operation Neptune Spear" carried out by the US navy seals in 2011 to liquidate Osama Bin Laden in Abbotabad, Pakistan. The Indian mission would entail a crack team of highly trained personnel from various units of Indian defence forces to sneak into Pakistan, destroy the terror training camps of the Jaish-e-Mohammad and liquidate, or better still, capture and bring back their leader Maulana Masood Azhar to face trial in Indian soil.

Accordingly, it is decided that the following eight (8) segments of the Indian defence and intelligence forces will take part in the mission – *Air force, Garud commandos, MARCOS, Para commandos, Ghatak, COBRA, RAW* and *NSG*. Being mindful of the fact that even a single error will lead to far reaching consequences for the sub-continent, our Prime minister decides that a final go-ahead will only be given if at least 90% of the members of the above-mentioned strike groups (taken together) are certain about success in this mission. Accordingly, the Chief secretary of the PM takes up the responsibility of doing this survey and selects the MARCOS commando force and asks each and every commando "*In view of success of the US special forces in capturing and killing Osama-Bin-Laden, can't you do the same for Masood Azhar and in the process teach Pakistan a lesson and elevate the prestige of your homeland ?*" Of the 180 commandos who were asked, 35 did not participate and of the 145 who did, 132 answered in the affirmative.

A) For this study, identify the

i) Research Question :

ii) Parameter:

iii) Subject:

iv) Sample:

v) Sample size:

vi) Statistic:

vii) The population of this study is (*circle the correct one*)

a) Garud b) Indian defence force c) MARCOS d) the 8 strike groups

e) Indians f) Indian army g) Indian Navy f) Indian air force

B) What kind of sampling did the Chief secretary did ? (*circle the correct one/s*)

a) stratified b) simple random c) convenience d) cluster e) volunteer

C) Are there any flaws in the above sampling procedure ? If so, what should have been a better or more correct way of sampling. Justify your answer.

D) Based on the feedback received, the Chief secretary promptly reports to the Prime Minister that 91% of the strike force personnel are confident of success in the proposed mission. Is this conclusion justified and/or should the PM decide to give the go-ahead based on this report ? Justify.

E) Identify *whether/how/why* the following may be of concern:

a) Sampling design / Undercoverage:

b) Response bias:

c) Non-response bias:

F) Unassured by the previous survey, the Prime minister wants to do a second and final survey. Accordingly his office employs a trained statistician who decides to collect a sample of size 150 from the eight units. Supposing that the total number of personnel in all these units taken together is 5000 and there are 300 personnel in the Para commandos, how many Para commandos should the statistician ideally select in his sample ? Justify with reasons.

G) In addition to the above survey, the PMO also takes high level feedback from spy agencies of Israel and America about the likelihood of success of these kind of operations. It does so because, America and Israel are the only two nations who have successfully conducted high-risk covert operations on foreign soil (capturing Osama-Bin-Laden and Adolf Eichmann respectively). Based on their feedback, the PMO concludes that there is a 75% chance of success for the proposed mission. Based on this information, answer the following questions :

a) What will be the mean and standard error of the sampling distribution of the sample proportion of affirmative responses that will be obtained from the sample in part F) ?

Mean:

Standard error:

b) Can you comment on the shape of the above distribution ? Justify.

c) What is the probability that at least 70% of the 150 personnel surveyed in F) will be optimistic about this operation ?

d) What is the probability that between 70% and 80% of the 300 strike group personnel surveyed in F) will be optimistic about this operation ?

e) Suppose the RAW chief advises the PM that he/she should only give a go-ahead if, with probability 80%, at least 70% of the strike group personnel are optimistic about this mission, the logic being that in case the mission fails, the PM can fall back on this statistic as a face-saver. How many personnel should now be sampled to ensure this ?

H) A critical factor in covert operations like the one above is the time it takes to finish the operation. Needless to say, operation planners always try to ensure that this time is minimised. For covert operations like the one proposed, 5 hours (300 mins) is usually taken as the mean time for completion with a standard deviation of 60 mins.

Three crack teams has been created for this operation and their training has commenced. The training entails simulating near-identical scenarios and replicating the proposed operation in those scenarios. In doing so, it is expected that the crack team will gain confidence and expertise and the chance of error will be minimized vis-à-vis chance of success will be maximised. In the first round, the crack teams conduct 12 simulated operations in total in the Garo hills of North east. For each operation, the completion time is noted by the planners.

Suppose the actual completion time (in mins) of the 12 simulated operations are : 330, 312, 295, 321, 289, 359, 341, 308, 376, 314, 301 and 408. Based on this information, answer the following questions:

a) Find the mean, median, first and third quartiles (Q1 and Q3) of the completion times. Interpret the median, Q1 and Q3 respectively.

i) Mean:

ii) Median:

Interpretation:

iii) Q1:

Interpretation:

iv) Q3:

Interpretation:

b) Based on the above measures, the distribution of the completion times is

i) Left skewed ii) Symmetric iii) Right skewed iv) Need more information

c) What will be the mean and standard error of the sampling distribution of the average completion time ? Can you attribute a shape to this distribution ? Justify.

Mean :

Standard error:

Shape:

In the final stage, the crack teams conduct 30 simulated operations in total at the Hemis National Park in Leh since conditions out there mimics those in the POK where the actual operation would take place.

a) Comment on the sampling distribution of the average completion time.

Taking into account the nature of terrain, forest cover, distance from LOC and other relevant factors, Indian intelligence agencies conclude that, to maximise the chances of success and minimise the chance of casualties on the Indian side, the proposed operation must be completed within 280 mins at the maximum.

b) Find the probability that the average completion time of the 30 simulated operations at the Hemis National Park will be at most 280 mins.

c) Find the probability that the average completion time of the 30 simulated operations will be more than 300 mins.

Q2. [**Trump & Iran]** In the aftermath of the U.S withdrawal from the Iran nuclear deal (JCPOA) in 2017, there has been discussions in different quarters as to whether U.S will launch military strikes on Iranian nuclear facilities. However, doing so is fraught with risks since Iran may immediately retaliate against Israel while China and Russia may support Iran against the U.S, thus deepening the crisis. As a result, events may soon go out of control and can lead to an all out war, not only between U.S and Iran but involving major military powers like China and Russia – a perfect recipe for a Third world war !

CNN ILLUSTRATION/GETTY IMAGES

However, the current U.S president, being very sensitive to how media (and the U.S public) portrays him and his presidency, decides to take a call (about military strikes against Iran) only if he is certain that he has the support of the U.S public. In fact, he decides that he will go ahead with the plan only if at least 70% of Americans support it. Accordingly, a major news channel affiliated with the Republican party is tasked with carrying out a survey to gauge the opinion of the U.S adult population on the aforementioned topic. The news channel floats the survey through its Facebook page and official webpage. It also uses Whatsapp to survey all its employees across its different offices in the U.S and overseas. The survey question was : "*In view of Iran's continuing violation of JCPOA guidelines and its increasingly hostile rhetoric against the U.S and Iran ("Death to Israel" and "Death to U.S" for example), don't you think this is high time for U.S to launch tactical military strikes against Iranian* *nuclear facilities in order to neutralize any impending threat from Iran once and for all ?*" A total of 15,584 responded to the survey and of them about 88% voted in favour of military strikes against Iran.

i) For this study, identify the

a) Population:

b) Parameter:

c) Subject:

d) Sample:

e) Statistic:

ii) Suppose the new channel concludes "Based on our nation-wide survey, it is apparent that 88% of adult Americans support military strikes against Iran". Is this justified ?

iii) Suppose before the survey results are officially presented to the U.S president, the news channel hires you as the lead statistician and seeks your advice as to the validity of the results. What would be your feedback ? Is there anything that you can do, as a statistician, that may lead to a more unbiased result (and in the process save mankind from an impending Third world war) ? Elaborate. (*The news channel will permit you to do another survey, if you may*).

iv) Suppose the news channel asks you to find a 95% confidence interval based on the survey results. Calculate the interval and interpret the same in the context of the problem.

a) Interval calculation **:**

b) Interpretation :

v) Suppose you decide to do a new survey to gauge the opinion of the American public. What will be the size of your sample if you want a precision of .02 with a confidence level of .95 ? You can take the survey result of the news channel as a benchmark.

vi) Suppose you do a new survey on a random and representative sample of U.S adults. The sample size is the one you have obtained above in v). Surprisingly, 51% of your sample support the idea of military strikes against Iran. Perform a hypotheses test to verify whether it is presumable that 70% of Americans (Mr. Trump's cut-off) support military strikes or whether the actual proportion is less than that.

a) State the null and alternative hypotheses for this problem.

H_0 : H_a :

b) Evaluate the test statistic and state its distribution under the null hypotheses.

Test statistic :

Null distribution :

c) Evaluate the p-value (*draw picture if necessary*).

d) Based on the p-value you obtain in c), you will

i) Reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.
ii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.01$ but not at $\alpha = 0.1$.
iii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.1$ but not at $\alpha = 0.01$.
iv) Reject H_0 at $\alpha = 0.1$ but not at $\alpha = 0.01$ and $\alpha = 0.05$.
v) Fail to reject H_0 at all the above α values.
vi) Reject H_0 at all the above α values.

e) State your conclusion in the context of the problem.

Q3. [Traffic @ Mt. Everest] A recent news that has created worldwide sensation concerns the mad rush to climb Mt. Everest in the ongoing climbing season. The story originated from the following picture taken by mountaineer *Nirmal Purja* which clearly shows a long queue of climbers waiting to ascent the summit of Mt. Everest a few days ago. In fact, this deadly "traffic jam" has already taken the lives of 11 mountaineers making this season one of the deadliest on record. A majority of these deaths happened because of sheer exhaustion and/or lack of oxygen resulting from the long hours of wait in the queue.

Since Edmund Hillary and Tenzing Norgay ascended Mt. Everest, it has been assumed that on an average, it takes about 60 mins to reach the summit from Camp 4 (the final camp). However, based on recent events, like the one above, the Nepalese government wants to verify whether, on an average, it is taking significantly longer to reach the summit nowadays. Accordingly, they select a random sample of 51 climbers who have ascended Mt. Everest in the recent past, including those who have did it in the current season. The average ascent-time they got from this sample was 100 mins with a standard deviation of 45 mins. It can be assumed that ascent-time has a normal distribution in the population.

<u>Part A</u>:

a) Explaining notations, state the null and alternative hypotheses

H_0 : H_a :

Notation :

b) The value of the test statistic is __________and it follows a____________ distribution with ___________degrees of freedom under the null hypotheses.

c) The p-value corresponding to the above test statistic is__________________ (*Draw picture if necessary*)

d) Based on the p-value you calculated above, you will

i) Reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.

ii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.01$ but not at $\alpha = 0.1$.

iii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.1$ but not at $\alpha = 0.01$.

iv) Reject H_0 at $\alpha = 0.1$ but not at $\alpha = 0.01$ and $\alpha = 0.05$.

v) Fail to reject H_0 at all the above α values.

vi) Reject H_0 at all the above α values.

e) The Nepalese government has decided that if, at 5% significance level, it is found that the ascent-time has overshot the accepted limit of 60 mins, it will put a cap on the yearly number of permits that is issued to climbers. Based on the answer in d) above, will the government implement the cap ? Justify.

f) Suppose, for the above problem, you fix the level of significance at 0.05. Then, the probability that you would conclude that the average ascent-time has not significantly increased when in fact that is the case (i.e it has not increased) is ______________________.

g) Suppose with probability 0.15 you conclude that the average ascent-time has not significantly increased when in fact it has increased significantly. Then the power of your test is___________

h) Evaluate the 99% confidence interval of the population average ascent time. Interpret the same in the context of the problem.

Interval:

Interpretation:

Part B:

[The Mountaineers] A related issue that is emerging out of this scenario is that many of the "climbers" are actually not trained mountaineers in the first place. In fact, they are novices who completely rely on the sherpas to ascend Mt. Everest just for the sake of it. It is a predominant feeling in the mountaineering community that only experienced mountaineers/trekkers should be allowed to climb Mt. Everest or similar peaks. That would lead to a more safer and less crowded scenario thus negating the deadly queues as above. On the other hand, doing so would mean a decline in the permits being sold, which in turn may dent the coffers of the Nepal government which heavily relies on the revenues obtained from selling these permits.

In view of this dilemma, suppose the Nepal government asks you to calculate a 95% confidence interval of "*p*", the true proportion of mountaineers who feel that permits should only be given to experienced mountaineers. To help you decide on the sample size, you are told that the resulting confidence interval should have a precision of .07.

a) How many mountaineers should you sample for this purpose ? Mention any assumption you make in the process.

b) Suppose, you collect a sample of size obtained above and 70% of the sampled mountaineers are of the opinion that permits should only be given to experienced trekkers/mountaineers. Based on this, calculate a 95% confidence interval of "*p*" and interpret the same in the context of the problem.

Interval:

Interpretation:

c) The 99% confidence interval of the true proportion of mountaineers who feel that permits should not be restricted to experience mountaineers is (*circle the correct alternative*)

i) Wider than the one obtained in b) above.
ii) Narrower than the one obtained in b) above.
iii) Of same length as the one obtained in b) above.
iv) Need more information to answer.

d) Suppose the Nepalese government decides that it would revisit its policy of issuance of permits if the proportion of mountaineers who support the assertion (of giving permits to experienced mountaineers) is significantly higher than 50%. If they rely on your confidence interval in b), what would be their decision ? Justify.

Q4. [Mohamed Salah]. Mohamed Salah is currently the new “blue-eyed boy” of world football. Originally from Egypt but currently playing for Liverpool, his on-field skills have put him in the same league as Lionel Messi and Christiano Ronaldo. Salah has scored 44 goals for the club season so far and many believe that this is just the beginning of a long and illustrious career ahead.

Suppose, for a particular match, the probabilities that Salah scores 0, 1, 2 and 3 goals are respectively 0.1, 0.5, 0.3 and 0.1 respectively. On the other hand, his earnings (per match) are tied with the number of goals he scores such that he gets USD 2 million (the base fee) if he does not score any goal while he earns USD 4, 7 and 10 million for scoring 1, 2 and 3 goals respectively. Based on this information, answer the following questions :

i) Find the distribution of Salah's earnings per match along with the mean and standard deviation of the same. *(You can express the answers in millions).*

a) Distribution :

b) Mean :

c) Standard deviation :

ii) Suppose Salah plays five matches in a row. Find the mean and standard error of the average earnings per match. Can its sampling distribution be approximated by a Normal distribution ? Justify.

iii) As mentioned before, Salah has played 44 matches in the current club season. Specify the sampling distribution of his average earnings per match for a club season like the current one with proper reasoning.

iv) What is the probability that Salah's average per match earning in a particular club season, consisting of 44 matches will be between USD 4 million and 6 million ?

v) Suppose you select a random sample of 20 football players who play for all the different English and Spanish Premier League clubs. Their average earning during a club season is USD 13 million with a standard deviation of USD 2.3 million. Find a 95% confidence interval of the true population average earning of a club football player and interpret the same. State any assumption that you need to make. *(You can state your answer in millions).*

Q5. [**H4 visas**] Currently a lot of discussion is taking place in the US administration regarding overhauling the Immigration system in line with Donald Trump's "*Buy American, Hire American*" philosophy. One aspect of this discussion revolves around instituting a merit-based immigration system which may lead to restricting the issuance of H1B visas which has enabled Indian IT sector workers to temporarily work in the US on onsite projects and as a pathway to permanent residency as well.

In addition to bringing in strict regulations to limit the misuse of H1B visas, President Trump has also proposed scrapping the H4 visas which provided legal work authorisation to spouses (husband or wife) of immigrants having H1B visas. As per US government data, about 90% of H4 visas are used by Indian women who are highly qualified. Hence, scrapping it would be devastating for these people (and their families) since they would lose the right to work in the US although being qualified and suited to do so.

Traditionally, Democrats have been more liberal than Republicans regarding immigration. However, terminating the H4 visas would require bipartisan support in the US Congress (i.e support from both Democrats and Republicans). In this context, a Washington based think tank would like to gauge the opinion of Democrat and Republican lawmakers on this issue, specifically whether Republicans are more supportive of this proposal than Democrats. Accordingly they asked a random sample of 50 Republican and 50 Democrat lawmakers as to whether they would support the termination of H4 visas. 39 Republicans and 18 Democrats responded in the affirmative. Suppose the true proportion of Republican and Democrat lawmakers who support the termination of H4 visas are p_1 and p_2 respectively.

i) What would be a reasonable set of hypotheses for the above problem ?

H_0 : H_a :

ii) Calculate the test statistic and state its distribution under the null hypotheses.

Test statistic :

Null distribution :

iii) Calculate the p-value of your test *(draw picture if necessary)*.

iv) Based on the p-value calculated above, your conclusion will be to

i) Reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.
ii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.01$ but not at $\alpha = 0.1$.
iii) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.1$ but not at $\alpha = 0.01$.
iv) Reject H_0 at $\alpha = 0.1$ but not at $\alpha = 0.01$ and $\alpha = 0.05$.
v) Fail to reject H_0 at all the above α values.
vi) Reject H_0 at all the above α values.

v) Write down your conclusion in the context of the problem so that even President Trump can understand it !.

vi) Calculate the 99% confidence interval of the difference in the population proportion of Republican and Democrat lawmakers who support the termination of H4 visas. Interpret your interval.

Construction :

Interpretation :

vii) Suppose President Trump, in his usual style, tweets "*I do not like the interval. It is SO WIDE*". What should you do to achieve a shorter and more precise interval ?

i) Decrease the confidence level and also the sample sizes of Republican and Democrats.

ii) Increase the confidence level and the sample sizes of Republican and Democrats.

iii) Increase the confidence level but decrease the sample sizes of Republican and Democrats..

iv) Decrease the confidence level but increase the sample sizes of Republican and Democrats..

v) Not possible to shorten the interval and please Mr. Trump; hence I will resign.

Q6. [14 marks] [Politician's wealth]. Before the next general elections, the election commission would like to get an idea of the personal wealth of politicians (Ministers, MPs, MLAs) belonging to the major political parties of India. Specifically, they would like to verify whether belonging to a particular political party has any association with the wealth of politicians.

Accordingly, they select a random sample of 11 politicians (a mix of ministers, MPs and MLAs) from *each of* BJP, Congress, Trinamool Congress, Aam Aadmi Party and Communist Party of

India. For each of the selected politicians, they collect data on the current market values of all movable and immovable assets including bank balance and values of investment. They cross check the same with the declared assests of each of them. Let μ_i denote the population average wealth of a politician (in lakhs of Rupees) belonging to the i^{th} political party.

Based on the observed data, the Election Commission obtains the following values :

Between Sum of Squares = 2054.67
Total Sum of Squares = 9752.32

i) State the null and the alternative hypotheses.

H_0 :

H_a :

ii) Complete the following ANOVA table.

Source	Degrees of freedom	Sum of Squares	Mean Squares	F statistic

iii) Are the assumptions satisfied for performing the ANOVA test ? Justify.

iv) The value of your test statistic is ___________________ and its degrees of freedom is____________

v) What is the p-value for your test statistic ? (*Draw picture if necessary*).

vi) Based on the p-value, you will

a) Reject H_0 at $\alpha = 0.01$ but not at $\alpha = 0.05$.

b) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.01$ but not at $\alpha = 0.1$.

c) Reject H_0 at $\alpha = 0.05$ and $\alpha = 0.1$ but not at $\alpha = 0.01$.

d) Reject H_0 at $\alpha = 0.1$ but not at $\alpha = 0.01$ and $\alpha = 0.05$.

e) Fail to reject H_0 at all the above α values.

f) Reject H_0 at all the above α values.

vii) Based on your decision above, what can you conclude at $\alpha = 0.05$?

a) Being a member of a particular political party has no bearing on the personal wealth.
b) At least two of the political parties differ significantly from the other three with regard to the mean wealth of its politicians.
c) At least one of the political parties differ significantly from the other four with regard to the mean wealth of its politicians.
d) All the political parties significantly differ from each other with respect to the mean wealth of its politicians.
e) We cannot conclude anything definite regarding the mean wealth of politicians from this data.