Chapter 1

Introductory Concepts

1.1 Statistical Thinking

With every passing day, we are living in an increasingly data rich world where large amounts of information are being generated on a mass scale in virtually every domain like Business, Economics, Medicine, Finance, Human Resources, Humanities, Environment and the like. The dramatic explosion of fine grained digital data is pervading functions in organizations ranging from marketing to finance, human relations to supply chain management and beyond. At the highest strategic level, it is disrupting entire industries in domains of transportation, media, entertainment, engineering, merchandising, retail banking and financial services. By 2022, it is estimated that for every person on earth, 2 MB of data will be created every second! With each click, swipe, share and like, a wealth of critical information are created. In fact, back in 2017, *The Economist* published a story titled *The world's most valuable resource is no longer oil, but data*[1], which generated such a storm that the phrase "Data is the new oil" eventually gained the status of an adage.

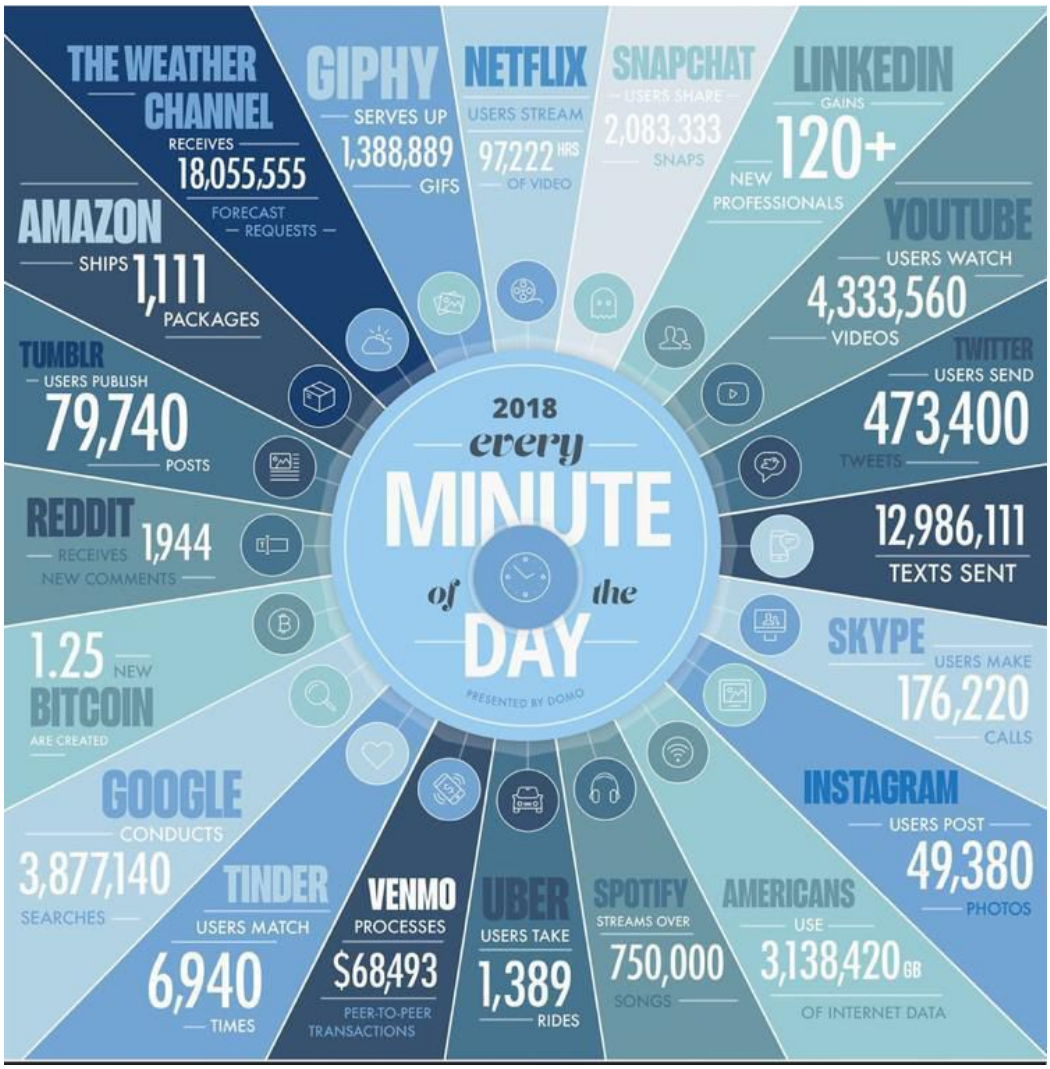
If used appropriately, data can be an incredibly powerful tool to formulate better decisions and implement sound policies. In fact, critical policy decisions made by governments and businesses are increasingly becoming more and more "data-driven" since those rely on evidence and insights obtained from analyzing large datasets. The fact that these decisions usually have far-reaching effects on our day-to-day lives only reinforces the importance of accurate and objective analysis of available data. Having said that, understanding and accurately analyzing such vast amounts of readily available data is interesting but challenging as well. This is precisely where Statistics comes into play.

Statistics or Data Analytics (or Data Science) provides us with the tools and concepts to draw critical insights from large datasets. This is one of those unique fields where demand far outstrips supply. As a result, individuals with sound statistical/analytical skills are (and will be) in high demand now and in the decades to come. In fact, as per the *Forbes* magazine, the job of a Data

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

scientist has consistently ranked as the best job in America for the last four years in a row[2]. The worldwide trend should not be any different.

The following picture depicts the amount of data that was supposed to be created every minute in 2018 and provides a notional overview of the data-deluge that surrounds us. It must have gotten bigger and better by now ! No wonder, we are living in a "data-driven" world !



Due to its ever widening applicability, Statistics is one of the fastest evolving subjects so much so that it is difficult to "define" it in a few words. Having said that, Statistics can be viewed as the "art and science" of deciphering the underlying pattern in a dataset. It enables us to have a "conversation" with the data. In doing so, 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*

[2]https://www.forbes.com/sites/louiscolumbus/2019/01/23/data-scientist-leads-50-best-jobs-in-america-for-2019-according-to-glassdoor/#1e82c1717474

and science of learning from data.

In the next section, we will be introducing an example that will be used as a pivot for our explorations of categorical data analytic techniques in this chapter and beyond. The subject of this example is very "real" in the sense that it touches the lives of each one of us, every passing day. On the other hand, it is multidimensional in nature as it has ramifications in politics, health, economics, education...virtually in every sphere of our lives. Last but not the least, it can be effortlessly positioned as a decision problem which have far reaching implications for all of us. As we will see, the tools and techniques of categorical data analysis can be employed to find an objective, data-driven solution to this realistic problem.

1.2 Motivating Example I: A Prime Minister's Dilemma !

Our prime minister has been facing an acute moral dilemma for quite some time. The question that bothers him is simple to frame but is proving to be notoriously difficult to answer, namely "*When to open the schools ?*" With the trauma of the second wave still vivid in public memory and the threat of an impending third wave looming in the horizon, no wonder that this question is giving sleep less nights to every Indian parent and probably the Prime Minister himself !

The dilemma that this question entails in understandable given the fact that it would take about a year to properly vaccinate the Indian population, even so, with the first dose, which, in the best case scenario, is only 70% effective. In addition, vaccination for children and those below 18 have not started yet. Having said that, prolonged closure of schools and educational institutes has already taken a huge toll on the mental health and well being of students and may even lead to a mental health epidemic in the coming days. The key to a balanced and measured response that will protect

the lives of youngsters while ensuring a continuity in their academic pursuits is something that is eluding the brightest and shrudest minds of the country right from the best scientists to the most seasoned bureaucrats and politicians.

To address this issue in a data-based and scientific manner, suppose the Health and Education ministries decide to obtain some "data-driven" insights from the collective wisdom of the netizens of this country. Accordingly, 18,710 adult Indians, aged 18 years or higher, were surveyed and asked: "*Should schools and educational institutes be reopened before pan-India vaccination is complete provided all necessary precautions (sanitization, social distancing etc) are adhered to* ?". To incorporate this data in the decision making process, the Government convenes an all-party meeting and after a lot of deliberation, decides that re-opening of educational institutes will only be considered if there is strong evidence that the true population proportion of adult Indians who are supportive of re-opening is at least 55%.

Once the data was collected, it was seen that of the 17,230 participants who responded, 9649 answered in the affirmative while the remaining 7581 were opposed to reopening until pan-India vaccination is completed.

Apart from the feedback on the aforementioned question, data was also collected on the following attributes for each respondent: age (in years), gender, marital status (married, unmarried, divorced, separated, widowed), educational qualification (illiterate, medium school, high school, Bachelors, Masters, PhD etc), occupation (unemployed, retired, corporate sector, academics, business etc), gross annual income, political party affiliation (BJP, Congress, AAP etc), number of children, whether any family member or near relative (including themselves) have had Covid-19, whether staying in the vicinity of a containment zone, whether belonging to joint or nuclear family, educational qualification of husband/wife if married, religion (Hindu, Muslim, Christian, Jain, Buddhist, other), ethnicity etc.

In the context of the above example, we will now explore different types of variables which are encountered in any dataset. In fact, a clear understanding of variable nomenclature is the first step towards successful analysis of any data.

1.3 Nomenclature of Variables

As mentioned above, a critical element in any statistical analysis is a proper understanding of the various types of variables that constitute a particular dataset. This is because, the nature and structure of statistical models that are eventually used to quantify the association between the variables usually depend on the types of variables themselves. Following is a brief overview of the major types of variables which are usually encountered in any real-life data analysis problem.

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, whole numbers. **Eg**: Number of Covid-19 patients in your locality, number of years of work experience of a ePGD-ABA participant etc.
- **Continuous** : A quantitative variable is continuous if the possible values belong to an interval. **Eg**: Your cholesterol level, running time of a Bollywood movie etc.

2. **Categorical** : A variable is categorical if each of its observations belong to any one of a set of finite categories. **Eg:** employee satisfaction in a major IT firm (very satisfied, moderately satisfied, neutral, unsatisfied, grossly unsatisfied), choice of transport (bus, metro, auto, taxi, bicyle, walk) etc. Categorical variables can again be of three types:

   - **Binary**: A categorical variable is binary if it has just two categories, generally labelled as "Yes" or "No". **Eg:** whether you support the Indian government's handling of the Covid-19 crisis, whether a participant of the first batch of ePGD-ABA is satisfied with the program etc.
   - **Ordinal**: A categorical variable is ordinal if it has *ordered* categories. **Eg:** Level of education (no education, middle school, high school, undergraduate, graduate), hierarchy in the Indian army (General, lieutenant general, major general,...,lieutenant) etc.
   - **Nominal**: A categorical variable is nominal if it has *unordered* categories. **Eg:** Your favorite color (red, yellow, green, crimson etc), favourite course in the ePGD-ABA program etc.

Q1.1. Identify the nature of the following variables in our motivating example.

	Quantitative		**Categorical**		
Variable	Discrete	Continuous	Binary	Nominal	Ordinal
Feedback					
Age					
Gender					
Marital status					
Education					
Occupation					
Annual income					
Political party					
No of children					
Covid-19 anyone ?					
Containment zone ?					
Joint/nuclear family					
Religion					

Note:

1. Sometimes a continuous variable may be simplified into a categorical one for ease of measurement. For example, cholesterol level, although continuous, may be categorized as {very low, low, normal, high, very high}.

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

1.4 How Corporations use Categorical Data

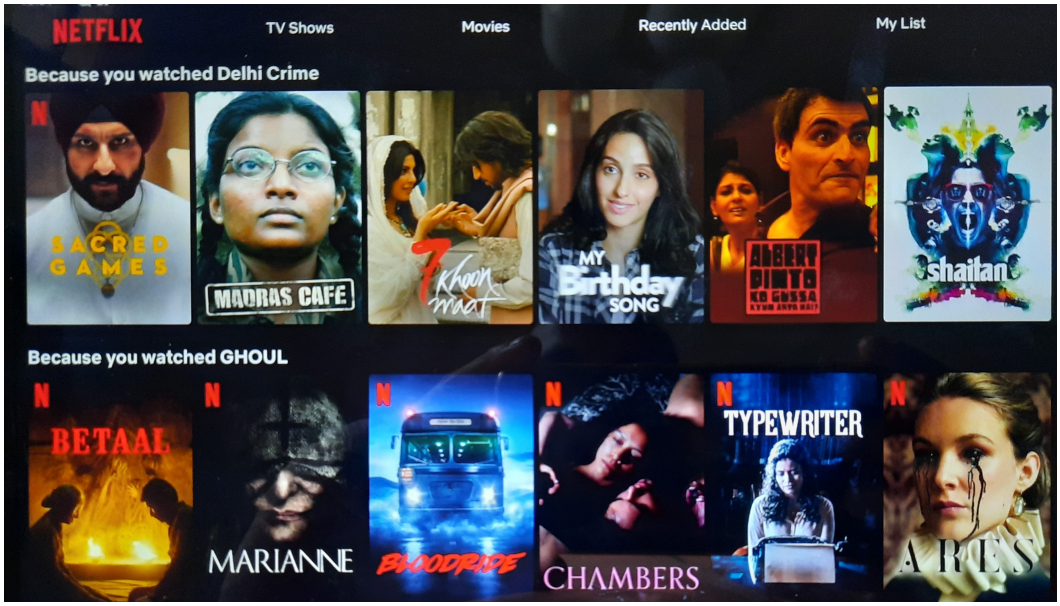
1.4.1 The Curious Case of Netflix

As mentioned in Sec 1, proper harnessing of the power of big data can unleash huge benefits for an organisation, be it public or private. A notable example of that is none other than **Netflix**. From its humble beginning as a DVD rental platform, way back in 1997, it has come a long way to become the defacto destination of viewers looking to binge on the latest movies and TV-shows. Needless to say, a large part of this arduous journey has been fueled by its efficient use of Big data...lots of it.

"*If the Starbucks secret is a smile when you get your latte... ours is that the Web site adapts to the individual's taste.*" - Reed Hastings, CEO of Netflix.

The above quote says it all. With every new membership and every streaming of a video or a web series, Netflix collects lots of valuable information on millions of its viewers like their age,

gender, location, viewing preferences and patterns. By applying statistical concepts and cutting edge machine learning algorithms on such data, it can dive right into the minds of its customers and obtain valuable insights about their viewing patterns. This, in turn, enables them to predict what any particular customer would like to watch next even before he or she has finished watching a movie or a show ! For instance, it has been observed that viewing behaviour depend on day of the week, time of day, the device and sometimes even the location from where the content is being watched. Netflix incorporates insights like these to recommend user-specific movies and TV shows. Needless to say, these are different across customers simply because the viewing patterns and tastes for every customer is unique.



Some of the key metrics that Netflix tracks to predict customer behavioral patterns are:

- Day and time of viewing content.
- Device on which the content was watched.
- Nature of the content.
- Searches on the platform.
- Portions of content that got re-watched.
- Whether content was paused, rewind, or fast forward and if so, the specific portion that was re-viewed.
- User location data.
- When viewing stopped or content changed midway.

- The ratings given by the users
- Browsing and scrolling behavior

Clearly, many of the above variables are categorical in nature (can you identify which ?). No wonder, when sophisticated machine learning algorithms are applied on such granular data of nearly 150 million customers, invaluable critical insights are generated on customer behaviour, which eventually gives Netflix an unbeatable edge that it enjoys in the streaming video market.

1.5 Looking Ahead

In this course, we will discuss various tools and techniques required for the analysis of categorical data. Specifically, we will explore the following three domains:

1. Distributions and inference relating to categorical variables.
2. Quantification of association (dependence or independence) between two or more categorical variables.
3. Modeling and analysing the effect of various explanatory variables on a categorical response variable.

Towards this end, we will use the *Prime Minister's Dilemma* example as our pivot. In the next chapter, we will introduce the twin concepts of probability distribution and statistical inference in the context of categorical variables. As we will see, these concepts lies at the heart of all subsequent analytical techniques.

Chapter 2

Distribution and Inference for Categorical Data

2.1 Probability Distributions for Categorical Data

As we have seen, categorical variables can be broadly classified into two types viz those with two categories i.e Binary variables and those with more than two categories i.e Nominal and Ordinal ones. Accordingly, there are two major types of probability distributions used for modeling such variables, namely the **Binomial** and **Multinomial** distributions. A brief exposition of these two distributions will now be given in the context of the *Prime Minister's Dilemma* example.

2.1.1 Binomial Distribution

It is not hard to imagine that after months of grinding lockdowns, people around the world are desperate for life to return to normal...to what it was before Corona struck. Workers are waiting to return to their workplaces and students of all ages are anxiously waiting to go back to schools and reunite with their friends. Having said that, it also well understood that doing anything in a hurry, without taking proper precaution may have calamitous outcomes.

Taking everything into account, suppose there is a 60% chance that an adult Indian will support the idea of reopening of schools and educational institutes in the current scenario provided all safeguards are strictly adhered to. Let us denote this as p i.e $p = .60$. Let us also assume that the responses are independent across individuals i.e response of a particular individual does not affect that of any other. Let Y_i be a random variable denoting the feedback of the i^{th} respondent $(i = 1, 2, ..., n)$ such that 1: support reopening and 0: does not support reopening. In our motivating example, $n = 17,230$. Then, Y_i is said to have a distribution with probability i.e

$$Y_i \sim \text{Binary}(0.60)$$

Suppose Y be the total number of respondents, out of 17,230, who support the reopening of schools i.e

$Y =$. Then, Y is said to have a distribution with index $n = 17,230$ and parameter $p = 0.60$. In general notations, the probability that Y takes a particular value, say y, is given by

$$P(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \qquad y = 0, 1, 2, ..., n$$

In simple terms, the above expression is the probability that y out of the n respondents support the reopening of schools while the remaining $n - y$ respondents are not supportive of it. The above Binomial distribution has mean and standard deviation .

Q2.1. For our example, find the probability that

1. None of the respondents support the reopening of schools:

2. All of the respondents support the reopening of schools:

3. Exactly half of the respondents support the reopening of schools:

4. At most 55% of the respondents support the reopening of schools:

5. At least 55% of the respondents support the reopening of schools:

The above calculations can be easily performed using `R` as shown below

1. `dbinom(0,17230,.6)` =
2. `dbinom(17230,17230,.6)` =
3. `dbinom(8615,17230,.6)` =
4. `pbinom(  ,17230,.6)` =
5. `1 - pbinom(  ,17230,.6)` =

Following are some properties of the Binomial distribution:

1. The Binomial distribution is at $p =$ i.e when success and failure are equally likely.
2. For fixed n, it gets closer to a distribution as p tends to 0.5.
3. For fixed p, with increasing n, it tends to a normal distribution with mean $\mu =$ and standard deviation $\sigma =$
4. A rule of thumb is that a Binomial distribution can be approximated by a normal distribution if both np and $n(1-p)$ are larger than 10. However this cutoff is larger for values of p close to 0 and 1. For instance, if $p = .10$ or .90, n should be at least 50 for the distribution to achieve a bell-shaped, symmetric shape.

Note: *You can see for yourself how varying the values of n and p results in different shapes of the Binomial distribution using the applet:* ***https://istats.shinyapps.io/BinomialDist/****.*

2.1.2 Multinomial Distribution

Suppose, instead of responding "Yes" or "No", each respondent in the *Prime Minister's Dilemma* survey is asked to select any one of the following categories as their response: *(strongly support, support, neutral, oppose, strongly oppose)*. Let us also assume that the outcome of a particular respondent have no bearing on that of another i.e the independence assumption is valid. Let p_i be the probability that an individual chooses category i as his/her response ($i = 1, 2, ..., 5$). Clearly, $\sum_{i=1}^{5} p_i = 1$. This is known as the Multinomial setup since the response variable Y, can have more than two (in this case, five) possible outcomes. This is a generalization of the Binomial setup discussed in the previous section.

Out $n = 17,230$ respondents, suppose the number of individuals belonging to the five categories are $(y_1, y_2, y_3, y_4, y_5)$ i.e y_1 individuals strongly support reopening, y_2 individuals support reopening, y_3 individuals are neutral about reopening, y_4 individuals oppose reopening and y_5 individuals strongly oppose reopening. In this context, the probability that $Y_1 = y_1$, $Y_2 = y_2$,...,$Y_5 = y_5$ is given by the following Multinomial probability mass function

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_5 = y_5) = \frac{n!}{y_1!y_2!...y_5!} p_1^{y_1} p_2^{y_2} ... p_5^{y_5}$$

Q2.2. Suppose, the probability of an adult Indian choosing the aforementioned response categories are (.07, .14, .1, .41, .28). Then what is the probability that out of the 17,230 respondents, 1250 strongly supports reopening, 2700 supports reopening, 1670 are neutral, 6392 does not support reopening while 5218 strongly oppose such a move ?

The above calculation can be easily performed using `R` as shown below

```
dmultinom(c(1250, 2700, 1670, 6392, 5218),100,c(.07, .14, .1, .41, .28)) =
```

2.2 Inference for Categorical Data

2.2.1 What is Inference ?

Statistical inference comprises of a suite of techniques that enables us to estimate unknown population characteristics, known as **parameters**, from analogous sample estimates, known as **sample statistics**, with an acceptable margin of error.

For instance, in the *Prime Minister's Dilemma* example, the parameter of interest is the true unknown proportion of adult Indians who are supportive of reopening of schools and educational institutions with immediate effect (say, p). However, this will always be unknown since it is impossible to survey each and every adult Indian. Supposing that the sample of 18,710 Indians is a good representation of the population of all adult Indians, we can use the corresponding sample proportion of affirmative responses (say $\hat{p}$), as an estimate of p. By a "good representative" sample, we mean a sample that is like a "mini-population" in accurately reflecting the various characteristics of the population, more so, in the correct proportion but on a much smaller scale. Needless to say, obtaining a good representative sample from the Indian population is no easy task, considering the size and diversity of India.

There are various ways in which inference can be carried out. In this chapter, we will deal with the two most common methods, namely **Hypotheses test** and **Confidence interval**. However, both these methods are based on a particular technique known as the **Maximum Likelihood Estimation (MLE)**. We will now explore this method in the context of the above example.

2.2.2 Maximum Likelihood Estimation

To reiterate the framework of Binomial distribution, let Y_i be the feedback (Yes/No) of the i^{th} respondent in the *Prime Minister's Dilemma* survey, such that

$$Y_i = \begin{cases} 1 \text{ (support reopening)} & \text{with probability } p \\ 0 \text{ (oppose reopening)} & \text{with probability } 1-p \end{cases}$$

where p is the true, unknown population probability that an adult Indian will support reopening of schools with immediate effect. As usual, we assume independence of responses across all respondents.

Let Z denote the total number of respondents out of the $n = 17,230$ sampled subjects, who support reopening i.e $Z =$. Then, under the above assumptions, Z will follow a dis-tribution with indices n and p respectively i.e

$$Z \sim \text{Bin}(n, p)$$
$$\text{i.e } P(Z = z) = \qquad z = 0, 1, 2, ..., n$$

The above function is also known as the *likelihood function* since it expresses the likelihood (or probability) of the observed data, z as a function of the parameter p. For instance, suppose out of the first $m = 100$ subjects selected as part of the above survey, 28 support reopening of schools i.e $z = 28$. Then the likelihood function corresponding to this sample will be

$$l(p|z = 28) =$$

The *maximum likelihood estimate* of p is that value of p which maximize the above function i.e the value of p, for which the probability (or likelihood) of obtaining the above data (i.e 28 successes out of 100 trials) is the highest. Table 2.1 shows the value of the above likelihood function for some randomly chosen values of p while the same is depicted pictorially in Fig 2.1.

p	0	.1	.2	.25	**.28**	.29	.31	.35	.4	.45	.6	.8	1
$l(p\|z)$	0	0	0.014	0.07	**0.089**	0.086	0.071	0.029	0.004	0.0002	0	0	0

TABLE 2.1: Values of $l(p|z = 28)$ for different values of p

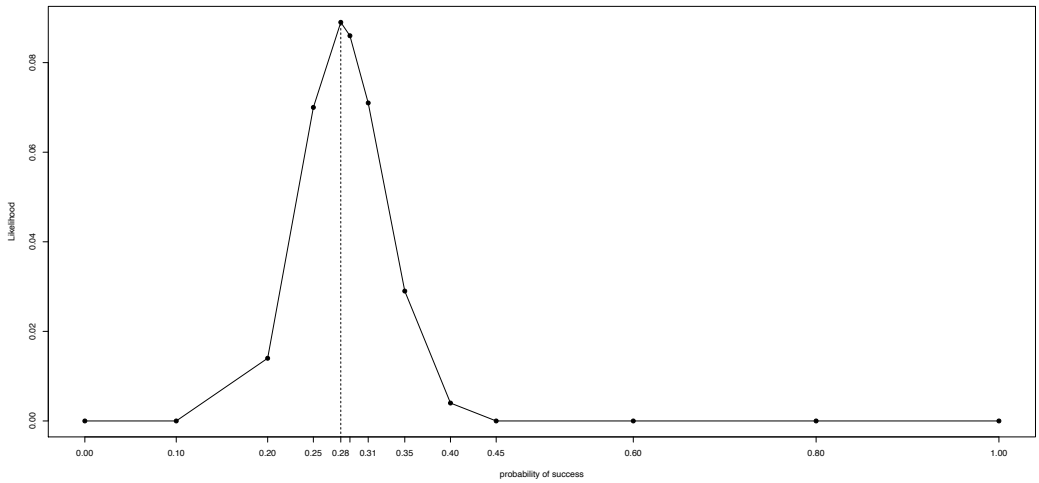


FIGURE 2.1: $l(p|z = 28)$ plotted against p (maximum achieved at $p = .28$)

It is evident, from both Table 2.1 and Fig 2.1 that the likelihood function achieves its highest value at $p = .28$. This value is known as the maximum likelihood estimate of p and is denoted as $\hat{p}$. Similarly, if say, 65 of the 100 respondents support reopening, it can be shown that the corresponding likelihood function

$$l(p|z = 65) = \frac{100!}{65!(100-65)!}p^{65}(1-p)^{100-65}$$

achieves its maximum at $\hat{p} =$ i.e *observing* $z = 65$ *successes out of* $n = 100$ *trials is most likely to occur when* $p = .65$ *than when* p *equals any other value in the range* $[0, 1]$. In general, for a Binomial distribution with indices n and p, the maximum likelihood estimate of p is given by

$$\hat{p} = \frac{\text{Total number of success}}{\text{Number of trials}} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

which can also be viewed as a sample mean. Thus, for the actual survey with 17,230 respondents, the maximum likelihood estimate of p will be

The `R` code for creating Table 2.1 and Fig 2.1 are as follows:

```
x<-28
n<-100
pi<-c(0,.1,.2,.25,.28,.29,.31,.35,.4,.45,.6,.8,1)
l<-rep(0,length(pi))
for (i in 1:length(pi))
{
l[i]<-round(dbinom(x,n,pi[i]),3)
}
plot(pi,l,xlab="probability of success",ylab="Likelihood",pch=19,xaxt='n')
axis(side = 1, at = pi,labels = T)
lines(pi,l,xlab="probability of success",ylab="Likelihood")
lines(c(.28,.28),c(-0.03,.089),lty=2)
```

2.2.3 Hypotheses Test for Population Proportion

Maximum likelihood estimation provides a single value estimate of the parameter. However, these estimates vary from sample to sample and hence comes with a measure of precision vis-a-vis variability (low variability → precision). Naturally, an inferential procedure that incorporates this precision measure will be more realistic. Hypotheses testing is one such inferential procedure in which we decide whether a certain claim (or hypotheses) about the parameter of interest is corroborated by the data obtained from a random sample. The various stages of performing a hypotheses test are as follows:

i) Hypotheses specification:

In the *Prime Minister's Dilemma* study, the Government decides to consider reopening, if there is strong evidence that more than 55% of adult Indians are supportive of it. This is synonymous to determining which of the following two "hypotheses" is best supported by the observed data

$$H_0: \qquad\qquad H_a:$$

where H_0 is the hypotheses which is "tested" against the hypotheses H_a. The value specified in H_0 is called the "null value" of the parameter and is denoted by p_0. Following are some guidelines about framing the null and alternative hypotheses:

1. Hypotheses should be framed before data is collected.
2. Hypotheses should only be expressed in terms of the parameter (say, p) and never in terms of sample statistics ($\hat{p}$).
3. Equality sign ("$=$", "$\geq$" or "$\leq$") should only be in the null while the alternative can only have strict inequality sign ("$>$", "$<$" or "$\neq$").
4. The null and alternative hypotheses, taken together, should encompass the entire parameter space.
5. Decision should only be framed in terms of either "Rejecting the null" or "Failing to reject the null". We can never "Accept" the null.

Q2.3. What would be the null and alternative hypotheses if the research questions were

1. The Government would like to know whether a majority or minority of the adult Indian population are supportive of reopening of schools with immediate effect.

   $$H_0: \qquad\qquad H_a:$$

2. The Government would consider reopening of schools only if the proportion of the adult population who are opposed to it is less than 30%.

   $$H_0: \qquad\qquad H_a:$$

Any hypotheses testing procedure would require us to quantify the evidence that the data provides *against* the null. If that evidence is large enough, we should the null; else we should the null. A test statistic provides us with a measure of this evidence.

ii) The test statistic:

A test statistic is a standardized measure of the "proximity" between the sample proportion, $\hat{p}$, and the null value, p_0. The observed value of the test statistic is given by

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

which follows a standard normal (Z) distribution under the null hypotheses provided $np_0 > 10$ and $n(1-p_0) > 10$. This follows from the theorem.

> Clearly, further is $\hat{p}$ from p_0 (in either direction), is the absolute value of the test statistic, and hence is the evidence against the null leading us to the null. On the other hand, closer $\hat{p}$ is to p_0, will be the absolute value of the test statistic, implying a evidence against the null, leading us to the null.

For our data, $z_{obs} =$

which means that the sample proportion $\hat{p}$ falls standard errors the null value, $p_0 = 0.55$. Now, we need to figure out whether this "distance" is large enough (or small enough) for us to reject (or fail to reject) the null hypotheses. For that, we need to calculate the P-value.

iii) The P-value:

P-value is a measure of the amount of evidence the data/test statistic provides *in support of* the null hypotheses. P-value is a probability measure and hence lies within and . *Smaller the p-value, larger is the evidence the null and hence we should the null and vice versa.*

How to calculate the p-value ? Strictly speaking, *p-value is the probability of obtaining a test statistic value at least as extreme as the one we have obtained if the null hypotheses is true (extremity measured in the direction of the alternative).* Thus, the p-value is the area beyond the observed value of the test statistic, z_{obs}, in the same direction as that of the alternative, under a standard normal curve (since the test statistic follows a standard normal distribution under the null). The following figures illustrate this fact

For our example, we have a alternative. Hence the p-value will be the area in the tail of the standard normal curve beyond $z_{obs} =$. The following figure depicts this.

We can obtain the p-value using the following `R` code.

```
x<-9649
n<-17230
p<-x/n
p0<-.55
z.obs<-(p-p0)/sqrt(p0*(1-p0)/n)
p.value<-1-pnorm(z.obs,0,1)
0.004
```

Interpretation: A p-value of 0.004 implies that, if at most 55% of adult Indians are supportive of reopening of schools, there would be a chance of obtaining 9649 or more affirmative responses in a random, representative sample of 17,230 adult Indians.

Now we have to compare this p-value to a threshold to determine whether it is small enough for us to the null or large enough for us to the null. For that, we need to use the significance level.

iv) Significance level (α):

Significance level is a "threshold", set by the statistician, against which the p-value is compared, to determine whether to reject (or fail to reject) the null hypotheses. It is denoted by α and is usually set at $.01, .05$ and $.1$. Strictly speaking, *it is the probability of rejecting the null hypotheses when it is actually true* and is known as the *Probability of Type I error*.

> We should reject H_0 at a given significance level α, if p-value $< \alpha$ and would fail to reject H_0 if p-value $\geq \alpha$.

Since our p-value of 0.004 is than all the three α values mentioned above, we should the null hypotheses at all the above significance levels i.e at $\alpha = .01, .05$ and $.1$. In other words, the data provide significant support against the null hypotheses at all the above significance levels. This leads us to the following interpretation of significance levels.

> Significance levels tells us how strong the evidence should be for us to reject the null hypotheses. Larger the significance level, is the evidence needed to reject the null and vice versa.

v) Conclusion/Policy decision:

Suppose the Government decides to base its decision on a 5% significance level. At this level, the data clearly provides significant support/evidence that the true proportion of Indian adults who are supportive of reopening of schools and educational institutes (provided all safe guards are strictly adhered to) exceeds the Government approved threshold of 55%. This can serve as a critical "data-driven" component in the decision making process of the Government.

vi) Summary:

The following table provides a bird's eye view of the entire hypotheses testing procedure:

$\hat{p}$ "far" from null value	$\hat{p}$ "close to" null value
↓	↓
test statistic	test statistic
↓	↓
p-value	p-value
↓	↓
evidence against H_0	evidence against H_0
↓	↓
Reject H_0	Fail to reject H_0

Q 2.4. Rework the hypotheses test assuming 9625 affirmative responses out of 17,230 and compare your conclusion with the one above.

Note: *A hypotheses test has interesting parallels with a legal battle. In any criminal trial, an under-trial is assumed to be innocent (null hypotheses) until he or she is proven guilty (alternative hypotheses). The job of a judge is to "measure" the evidence against the under-trial and pass a judgement accordingly. If the evidence is strong enough, the under-trial is pronounced guilty (synonymous to rejecting the null). Else, if the evidence is not strong enough, he/she is released (synonymous to failing to reject the null). However, even in the latter case, a judge is not supposed to pronounce the convict "innocent" (just like a statistician is not supposed to "accept the null") ! Having said that, we can think of the following analogies between various aspects of hypotheses test and court room trial :*

Hypotheses test	*Courtroom trial*
Test statistic	
P-value	
Significance level	
Rejection of null	
Failing to reject null	

Q 2.5. Suppose an under-trial is given the death sentence in lower court which is commuted to life-in-prison by a higher court later for the same quantum of evidence. Which of the judges will have a lower significance level ?

2.2.4 Confidence Interval for Population Proportion

From the above hypotheses test, we can only conclude that the data provides significant evidence, at all possible significance levels, that the proportion of Indians who are supportive of re-opening (p) exceeds the government approved threshold of 55%. However, it would be more useful, from a policy perspective, to have an idea of the range of plausible values of p. Confidence intervals enables us to address this issue with a high degree of certainty or confidence.

Every confidence interval comes with a confidence level, denoted as $1-\alpha$, chosen to be either 90%, 95% or 99%. The general form of a $100(1-\alpha)\%$ confidence interval of p is

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{or} \quad \left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

where n is the sample size and $\hat{p}$ is the estimate of p. The above confidence interval is valid provided $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ i.e there are at least 10 "Yes" and 10 "No" responses in your sample. The quantity

$$Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is a measure of precision of $\hat{p}$ as an estimate of p and is known as the
while $Z_{\alpha/2}$ denotes the value of the standard normal variable to the right of which the area under a standard normal curve is $\alpha/2$.

The following table depicts the critical values corresponding t6 different confidence levels.

Confidence level $(1-\alpha)$	Tail area $(\alpha/2)$	Critical value $(Z_{\alpha/2})$
90%	(1-.9)/2 = .05	$Z_{.05} = 1.645$
95%		
99%		

Note 1. *(i) For fixed confidence level, as sample size increases, 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

Q 2.6. Calculate a 95% confidence interval of the true population proportion of adult Indians who are supportive of reopening of schools with immediate effect and interpret the same.

- Assumptions:

- Interval:

- Interpretation:

The `R` code for calculating the above interval is as follows:

```
1.x<-9649
n<-17230
p<-x/n
lower<-p-qnorm(.975,0,1)*sqrt(p*(1-p)/n)
upper<-p+qnorm(.975,0,1)*sqrt(p*(1-p)/n)
2.  install.packages("binom")
library("binom")
prop.test(9649,17230,p=.55,alternative="two-sided",correct=FALSE)
```

Note: *In the above case, the confidence interval contains values greater than 0.55 implying that the true proportion of adult Indians who are supportive of reopening can only be greater than 55% with 95% confidence. This seem to corroborate the conclusion reached from the hypotheses test. However, it may well happen that the 95% confidence interval of p contains the null value even when the null hypotheses is rejected at 5% significance level. This may seem to be counterintuitive but actually it is not. This is because, at a particular significance vis-a-vis confidence level, conclusions from a confidence interval will only match that of a* ***two-sided*** *hypotheses test. However, the conclusions may not match if the alternative is one sided, which is the case here.*

Q 2.7. Verify that a 90% confidence interval will be narrower while a 99% interval will be wider than the above interval.

- 90% interval

```
lower<-p-qnorm(.95,0,1)*sqrt(p*(1-p)/n)
upper<-p+qnorm(.95,0,1)*sqrt(p*(1-p)/n)
```

- 99% interval:

```
lower<-p-qnorm(.995,0,1)*sqrt(p*(1-p)/n)
upper<-p+qnorm(.995,0,1)*sqrt(p*(1-p)/n)
```

2.2.5 Sample Size Determination

Suppose the Government in your state is enthused by the *Prime Minister's Dilemma* survey and would like to carry out a similar survey in the state to gauge public sentiment regarding re-opening of schools and educational institutes. Accordingly it decides that the survey should be such that the parameter is estimated with a precision of .02 with a 95% confidence. What should be the minimum sample size that would ensure the above guidelines ?

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

In the above problem, $Z_{\alpha/2}$ will be . But what about $\hat{p}$? We have the following two options:

1. Use the $\hat{p}$ value obtained in the national survey i.e 0.56. That would result in the following sample size:

2. Use $\hat{p} = 0.5$. That would result in the following sample size:

Note 2. *i)* $p = 0.5$ *will always result in a larger (i.e more conservative) sample size estimate.*
ii) The sample size estimate should always be rounded off to the next higher integer.
iii) The sample size estimate would with higher confidence levels.

2.2.6 Score Confidence Intervals

Let us assume that the true proportion of adult Indians who are supportive of reopening is very close to 0 i.e nearly all Indians prefer reopening of schools only after pan-India vaccination is complete. Suppose out of the first 100 subjects interviewed, none supports reopening. Thus the 95% confidence interval of p will be

$$\hat{p} = 0, \quad 0 \pm 1.96\sqrt{\frac{0 \times (1-0)}{100}} = (0, 0)$$

Obviously this is highly unrealistic. In fact, when the true population proportion is very close to the borderline values i.e 0 or 1, the above confidence interval performs poorly in the sense that the actual coverage probability falls short of the nominal level (say, 95%), unless the sample size n is very large. In situations like this, there is an alternative method of constructing confidence intervals that generates more realistic intervals with better coverage probabilities. These intervals are known as *score confidence intervals* and are based on the duality between hypotheses tests and confidence intervals.

The duality principle states that, at a given significance vis-a-vis confidence level, the results of a *two sided* hypotheses test and confidence interval should match. In other words, the confidence interval should contain *all* values of the parameter p, specified in the null hypotheses, for which the null is *not rejected* i.e for which the two-tailed p-value is $>$ than the chosen significance level. This is synonymous to the absolute value of the test statistic (i.e $|z_{obs}|$) being less that $Z_{\alpha/2}$. Let us illustrate this concept using a toy example.

Suppose, enthused by the *Prime Minister's Dilemma* study, you choose a random sample of 10 of your neighbours and ask them whether primary/montessori schools should remain closed until pan-India vaccination is complete. Suppose 9 out of 10 answered in the affirmative. For this data, the test statistic value for testing the following hypotheses

$$H_0 : p = 0.596 \quad vs \quad H_a : p \neq 0.596$$

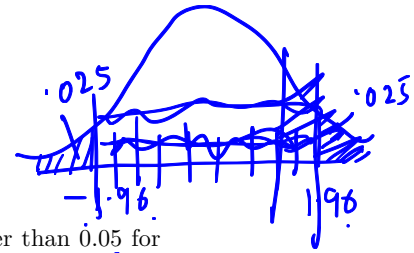
is $z_{obs} = \dfrac{0.9 - 0.596}{\sqrt{\frac{0.596 \times (1-0.596)}{10}}} = 1.96$

Confidence interval $(p_{0.1000}, p_{0.1000})$ → Set of plausible values of p.

Thus, the p-value is the area beyond 1.96 and which is .05. Similarly, for testing

$$H_0 : p = 0.982 \quad vs \quad H_a : p \neq 0.982$$

the test statistic is $z_{obs} = \frac{\cdot 9 - \cdot 982}{\sqrt{\frac{\cdot 982(1-\cdot 982)}{10}}} = -1.96$



which also yields a two-sided p-value of 0.05. In fact, the p-value will be greater than 0.05 for testing any null value lying between 0.596 and 0.982. In other words, we will fail to reject the null hypotheses at $\alpha = 0.05$ for all null values lying in the interval (0.596, 0.982). This implies that at 95% confidence, the interval (0.596, 0.982) contains all the plausible values of the population parameter p. This interval is known as the **Score Confidence Interval** of p.

In general, for a given significance level α and given values of $\hat{p}$ and n, the lower and upper limit of the score confidence interval are solutions to the following equation

$$\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} = Z_{\alpha/2} \qquad (2.1)$$

As mentioned before, score confidence intervals are particularly useful when the sample proportion of success is 0 or 1, as was the case in the above sample for which none of the 100 sampled subjects responded in the affirmative. In that case, the usual confidence interval was (0, 0). However, in this case, the 95% score confidence interval can be shown to be (0, 0.037) which is much more meaningful.

It is not necessary to solve the equation (2.1) to obtain score intervals since it is relatively easy to generate it in R as the following code illustrates:

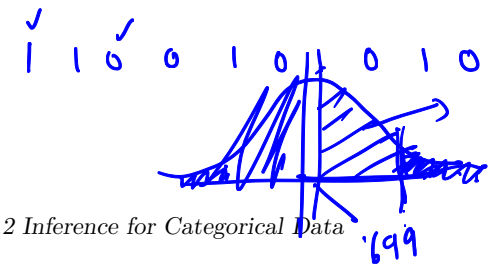
```
1.9 out of 10 success:
install.packages("binom")
library("binom")
binom.confint(9,10,conf.level=0.95, method="wilson")
2.0 out of 100 success:
binom.confint(0,100,conf.level=0.95, method="wilson")
```

2.2.7 Likelihood Ratio Test

As the term suggests, this test uses the ratio of the likelihood function of p, defined in Sec 2.2.2, at two values of p - the null value, p_0 and the maximum likelihood estimate of p say $\hat{p}_{ml}$. Accordingly, the likelihood ratio test statistic is given by

$$G^2_{obs} = 2\log\frac{l(\hat{p}_{ml})}{l(p_0)}$$

which can be shown to follow a Chi-square distribution with 1 degrees of freedom. Since the chi-square distribution is only defined in the positive axis, the p-value will be the right tailed area



beyond G^2_{obs} under the above distribution. Larger the value of $l(\hat{p}_{ml})$ with respect to $l(p_0)$, larger is the test statistic value and hence larger is the evidence against H_0.

In the context of the *Prime Minister's Dilemma* study, let us use the likelihood ratio test for testing

$$H_0 : p \leq 0.55 \quad vs \quad H_a : p > 0.55$$

Thus the null value of p is while the maximum likelihood estimate value of p is

The corresponding values of the likelihood function are

$$l(p_0) = \frac{17230!}{9649!(17230-9649)!}(.55)^{9649}(1-.55)^{17230-9649} = 0.00019$$

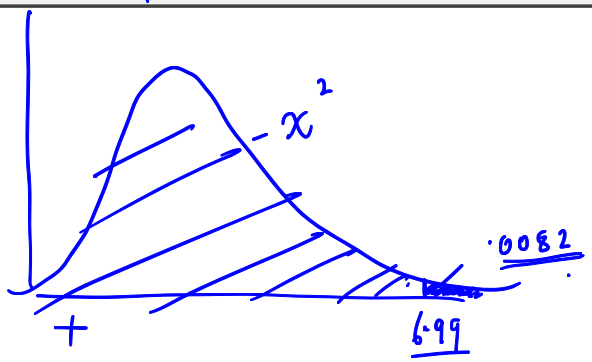
$$l(\hat{p}_{ml}) = \frac{17230!}{9649!(17230-9649)!}(.56)^{9649}(1-.56)^{17230-9649} = 0.0061$$

which yields the test statistic

$$G^2_{obs} = 2\log(.0061/.00019) = 6.99$$

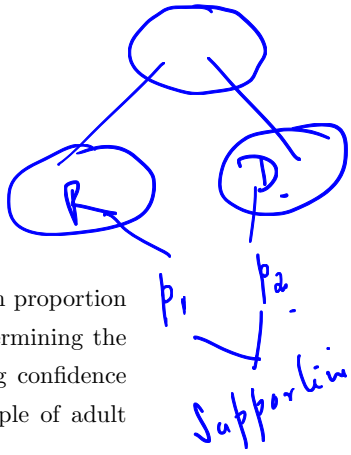
The p-value corresponding to the above test statistic value is found to be 0.0082 which is than all the commonly used significance levels. Hence, at all those significance levels, we the null hypotheses and conclude that the data provides strong evidence that the true proportion of Indians who are supportive of reopening of schools and educational institutions exceed the government mandated threshold of 55% The R codes for performing the likelihood ratio test are as follows:

```
dbinom(9649,17230,.55)
0.00019
dbinom(9649,17230,.56)
0.0061
G2<-2*log(.0061/.00019)
6.99
pvalue<-1-pchisq(G2,1)
pvalue
0.0082
```



Chapter 3

Analyzing Contingency Tables



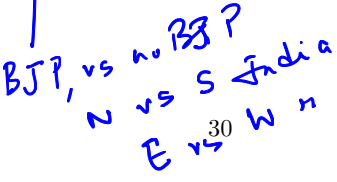
3.1 Introduction

Our discussions in the previous chapter revolved around estimating the true population proportion of Indians who are supportive of reopening of schools (p). This was achieved by determining the plausible values of this proportion either through hypotheses tests or by constructing confidence intervals of various types. In doing so, we treated the population vis-a-vis the sample of adult Indians as one entity.

Suppose the Government wants to determine whether there are any significant differences between different sections of the population in their perception of the above issue. For instance, the Government may be interested in the following research questions:

- *Is there any significant difference between males and females in their "supportiveness" towards reopening of schools vis-a-vis is it plausible that a significantly larger proportion of females are less supportive of reopening of schools compared to males ?*
- *Is is plausible that teachers are less supportive of reopening of schools than individuals in other professions ?*
- *Is it true that individuals who have had a brush with Covid-19 are significantly more cynical about reopening compared to others who do not yet have a similar experience ?*

The underlying similarity between the above questions is that, each involves splitting the population into two segments and comparing the feedback/perception about reopening of schools between those. This can also be viewed as analyzing the strength of association between two variables, namely a grouping variable (say, gender, profession or covid-survivor) and a response variable, namely feedback. In this chapter, we will discuss various inferential techniques for answering the above questions in a data-based manner.



3.2 Contingency Tables

Let us consider the first question mentioned in the previous section. It can also be rephrased as to whether gender has any significant effect on supportiveness towards reopening. Thus gender is the explanatory variable while feedback in the response variable. The following table cross-classifies all the 17,230 respondents according to their gender (male, female) and supportiveness towards reopening (yes, no). These kind of tables are known as 2×2 Contingency tables since both Gender and Feedback have two categories.

	Feedback on reopening		
Gender	Support	Donot support	Total
Females	2631	6121	8752
Males	7018	1460	8478
Total	9649	7581	17,230

3.2.1 Probability Structure

As shown above, a contingency table cross-classifies a sample of subjects into different categories of response and explanatory variables. For instance, in our sample of 17,230 respondents, 6121 females do not support reopening while 7018 males support reopening. Accordingly, we can define the following probability measures for any contingency table:

1. **Joint probability**: This enumerates the probability, p_{ij}, that a randomly chosen adult Indian is classified in the i^{th} row of Gender (G) and j^{th} column of Feedback (F) ($i = 1, 2; j = 1, 2$). This is given by $P(G = i, F = j)$ and it sums to 1 when the sum is taken over all the cells. p_{ij} is estimated by the corresponding cell proportion

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \Rightarrow n_{ij} = n\hat{p}_{ij}$$

where n_{ij} is the number of subjects classified in the i^{th} row of Gender (G) and j^{th} column of Feedback (F) while n is the total sample size. For instance, the estimated joint probability that a "Female" supports "Reopening" will be

$$\hat{p}_{11} = \frac{2631}{17230} = .153 \rightarrow 15.3\%$$ chance that a female will support reopening.

Similarly the estimated probabilities for the other cells are $\hat{p}_{12} = 6121/17230 = .355$ $\hat{p}_{21} = \frac{7018}{17230} = .407$ and $\hat{p}_{22} = \frac{1460}{17230} = .0.084$

2. **Marginal probability**: These are the probabilities that a randomly chosen adult Indian belongs to the i^{th} row or j^{th} column respectively. These are given by p_{i+} and p_{+j}. Clearly,

$$p_{i+} = p_{i1} + p_{i2} \quad \text{and} \quad p_{+j} = p_{1j} + p_{2j}$$

As usual, these are estimated by the corresponding sample proportions of row and column totals. For instance, the sample marginal distribution for feedback on reopening is corresponding to the "Support" and "Do not support" categories.

3. **Conditional probability**: As the term suggest, these are the probabilities that a randomly chosen adult Indian will belong to a particular category of Feedback given his/her gender. The population conditional probability of belonging to the j^{th} category of the response given the i^{th} category of the explanatory variable is denoted by $p_{j|i}$ and is estimated by the corresponding sample proportion of individuals. For instance, given that a subject is male/female, the conditional proportions that he/she will not support/will support reopening are

These are estimates of the corresponding conditional probabilities $p_{ns|m}$ and $p_{s|f}$ respectively.

3.3 Independence of Two Categorical Variables

A very important concept in Statistics is that of independence between a response and an explanatory variable. In the context of our example, Feedback will be statistically independent of Gender, if the population conditional distribution of each category of Feedback is identical for Males and Females. This can happen if the joint probability of any cell is equal to the product of the corresponding marginal probabilities i.e if

$$P(\text{Gender} = i, \text{Feedback} = j) = p_{ij} = p_{i+} \times p_{+j} \quad \text{for} \quad i = 1, 2;\ j = 1, 2$$

i.e the marginal probabilities determines the joint probabilities. For instance, under the assumption of independence, the counts corresponding to cell (1, 1) i.e (Females, Support) will be

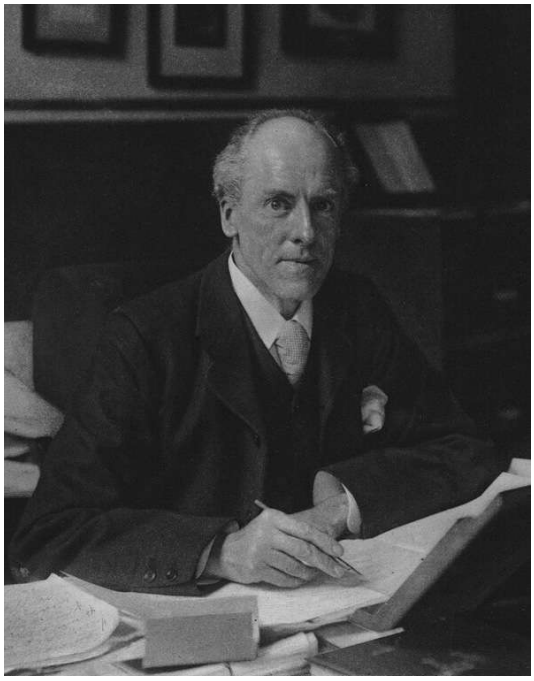
while the conditional probability of supporting reopening given any gender is In fact this is same as the marginal probability of supporting reopening.

Q 3.1 Determine the cell counts corresponding to the different categories of Feedback and Gender under the assumption of independence between the two

	Feedback on reopening		
Gender	Support	Donot support	Total
Females	4901.22	3850.78	8752
Males	4747.78	3730.22	8478
Total	9649	7581	17,230

3.4 Chi-square Test of Independence

Based on the above discussion, it should be intuitively clear that one way of testing for independence between two variables is to measure the "distance" between the observed cell counts in a contingency table and the cell counts which are expected under the assumption of independence i.e the ones calculated above for instance. If this distance is large, that would indicate ~~large~~ strong evidence against independence while a relatively small distance measure would indicate weak evidence of independence between the variables. This logic forms the foundation of the well known **Pearson's Chi square test** for independence, named after the famous British statistician Karl Pearson who proposed it in 1900.



This test has the following stages:

1. **Assumptions**
   - The sample of 18,710 adult Indians should be a random and representative sample from the population - *satisfied.*
   - Expected counts in each cell should be at least 5 - *satisfied.*
2. **Hypotheses**
   - H_0: Feedback is statistically independent of gender.
   - H_a: " " not independent of gender.

3. The test statistic is given by

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} and e_{ij} are respectively the observed and expected cell counts (under independence) in the $(i,j)^{th}$ cell of the contingency table. The expected cell count in the $(i,j)^{th}$ cell is given by

$$e_{ij} = \frac{n_{i+} \times n_{+j}}{n}$$

where n_{i+} and n_{+j} are the i^{th} row total and j^{th} column total while n is the sample size.

The above test statistic follows a Chi-square (χ^2) distribution with $(r-1)(c-1)$ degrees of freedom for a contingency table with r rows and c columns i.e r categories of the explanatory variable and c categories of the response variable. For the Gender-Feedback example, the degrees of freedom of the above test statistic will be since there are two rows and columns.

Intuition: *If independence holds, the observed and expected counts will tend to be for each cell leading to a value of the above test statistic. On the other hand, if the variables are not independent, at least some of the cell counts will tend to far apart from their expected counterparts, resulting in a value of the above test statistic.*

The following table depicts the observed (expected) cell counts corresponding to the Gender-Feedback contingency table

	Feedback on reopening		
Gender	Support	Do not support	Total
Females	2631 ( )	6121 ( )	8752
Males	7018 ( )	1460 ( )	8748
Total	9649	7581	17,230

Accordingly, the test statistic can be found to be 4855 which has degrees of freedom.

4. **P-value:** Since the χ^2 distribution is only defined in the positive axes, p-value will be the tailed area above the observed value of the test statistic i.e under a χ^2 distribution with 1 degrees of freedom. Needless to say, this area and hence the p-value is nearly 0.

5. **Conclusion:** Since the p-value is smaller than all the commonly used α values $(.01, .05, .1)$, we the null hypotheses of independence and conclude that there is strong evidence of association between gender and feedback on reopening of schools. In other words, males and females have significantly different views about reopening of schools and educational institutions before with immediate effect.

The R codes for obtaining the above test statistic is as follows:

```
feedback.gender<-matrix(c(2631,7018,6121,1460),ncol=2,byrow=TRUE)
chisq.test(feedback.gender)
Pearson's Chi-squared test with Yates' continuity correction
data:  feedback.gender
X-squared = 4855, df = 1, p-value < 2.2e-16
```

3.5 Comparison of Proportions

For the *Prime Minister's Dilemma* example, the Chi-square test indicated that there is a strong evidence of association between gender and feedback on reopening. However, it did not provide any information regarding the exact strength of association between gender and feedback. This is a major shortcoming of the Chi-square test. In this section, we will highlight some specific measures of the strength of association between two categorical variables.

3.5.1 Difference of Proportions

As the term suggests, this is the difference in the conditional probabilities for a particular category of response taken over the categories of the explanatory variable. The following table depicts the conditional proportion values corresponding to all the cells of the gender-feedback contingency table:

	Feedback on reopening ?		
Gender	Support	Do not support	Total
Females	0.3	0.7	8752
Males	0.828	0.172	8748
Total	9649	7581	17,230

Thus the difference of proportion between males and females who are supportive of reopening is 0.828 − 0.3 = 0.528 i.e 52.8%. more males are supportive of reopening compared to females. Following are some properties of difference of proportions:

1. It lies between −1 and 1 .
2. It is 0 when two variables are independent.
3. Closer the value is to −1 or +1 , stronger is the association between the two variables.

3.5.2 Relative Risk

Here, instead of taking the difference, we take the ratio of the conditional proportions for a particular response category over the categories of the explanatory variables. For our example, the relative

risk of opposition towards reopening between females and males will be

$$RR = \hat{p}_{ns|f} / \hat{p}_{ns|m} = .7/.18 = 3.89 \approx 4$$

which implies that the sample proportion of females who are opposed to reopening is 3.89 times the corresponding proportion of males. Similarly, the relative risk of support towards reopening between males and females is $\frac{.828}{.3} = 2.76$ implying that the sample proportion of males who are supportive of reopening is 2.76 times the corresponding proportion of females. Following are some properties of relative risk:

1. The relative risk can be any non-negative number.
2. Two variables are independent if $\hat{p}_{s|m} = \hat{p}_{s|f}$ which corresponds to a relative risk of 1
3. Further the value of relative risk is from 1, stronger is the strength of association between the two variables.
4. Two values of the relative risk such that one is the reciprocal of the other represent the same strength of association but in opposite direction.

Note: *The difference of proportion measure is more meaningful when both the sample/population proportion values are closer to the middle of the range than when they are closer to the borderline i.e 0 or 1. In the latter case, relative risk is a much more relevant and meaningful measure in terms of quantifying the true strength of association between the response and predictors.*

As an illustration of the above comment, consider the following two situations - suppose the proportion of males and females who are supportive of reopening are i) .010 and .001 and ii) .510 and .501 respectively. In both the cases, the difference of proportions is .009 . However, it is not difficult to see that the difference is much more striking in the first instance since in that case, 10 times more females were averse to reopening than males. This is evident in the relative risk measure

$$RR_1 = .01/.001 = 10$$

Interestingly, the relative risk for the second instance is

$$RR_2 = .510/.501 = 1.02$$

a value that is nearly 10 times less than that for the first instance regardless of the same difference in proportion measure for the two cases ! => RR is a more robust measure of association b/c Compared to difference.

3.5.3 Odds ratio

The odds ratio is probably the most well known measure of the strength of association between two categorical variables. Just like difference of proportions and relative risk, it is also a function of the conditional proportion values. Before delving into odds ratios, we need to have a clear understanding of the odds.

What is the odds ? For success probability π, the odds of success is given by

$$\text{Odds} = \frac{\pi}{1-\pi}$$

So, if the success probability is, say $\pi = .8$, the odds of success will be $.8/.2 = 4$ i.e we expect to see 4 success for every one failure. Similarly, an odds of 1/4 would imply that a failure is 4 times as likely as a success. Clearly, the success probability can be expressed in terms of odds as

$$\pi = \text{Odds}/(1+\text{Odds})$$

So, what is the odds ratio ? As the term suggests, it is the ratio of the odds of a particular category of the response taken across the two categories of the explanatory variable. For instance, for our contingency table,

- The odds that females will support reopening (a success) is $O_1 = .3/.7$
- The odds that males will support reopening is $O_2 = .828/.172$

This implies an odds ratio of

$$OR^S_{f/m} = \frac{.3/.7}{.828/.172} = .089 \qquad OR^S_{m/f} = \frac{1}{.089} = 11$$

i.e odds that females will be supportive of reopening of schools is .09 times the corresponding odds that males will be supportive of reopening. In other words, the odds that a female will be opposed to reopening is ≈ 11 times the corresponding odds that a male will be opposed to reopening. Following are some properties of Odds ratio:

1. The odds ratio can be any non-negative number
2. Two variables are independent if Odds$_f$/Odds$_m$ i.e when the odds ratio is 1
3. Higher the value of odds ratio (from 1), stronger is the strength of association between two variables.
4. When the order of the rows is reversed, the new value of the odds ratio is inverse of the original one but the strength of association remains the same.
5. Odds of a "No" response is always the inverse of the odds of a "Yes" response.
6. The odds ratio remains unchanged if the orientation of the table reverses i.e if the rows becomes the columns and the columns becomes the rows.

3.5.4 Inference for Odds ratios

Often it is important to have a range of plausible values for the odds ratio, θ. This is obtained indirectly by first constructing a confidence interval of the *log-odds*, $(\log\theta)$ and then exponentiating the two limits. It can be shown that a $100(1-\alpha)\%$ confidence interval of $\log\theta$ is given by

$$\log\hat{\theta} \pm z_{\alpha/2} se(\log\hat{\theta})$$

iii) " " " M supporting reopening is at least 10·52 times and atmost 12·09 times the odds of F supporting reopening

where $se(\log\theta)$ can be shown to be

$$se(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

where n_{ij} is the observed cell count for the i^{th} row and j^{th} column for the contingency table $(i = 1, 2, j = 1, 2)$. For our example, the 95% confidence interval of the $\log\theta$ of supporting reopening will be

$\log(\cdot 09) \pm 1\cdot 96 \times \sqrt{\frac{1}{2631} + \frac{1}{6121} + \frac{1}{7018} + \frac{1}{1460}}$ $= (-2\cdot 49, -2\cdot 35)$

$(1\cdot 23, 1\cdot 37)$

Thus, the 95% confidence interval of the corresponding odds ratio will be

$(e^{-2\cdot 49}, e^{-2\cdot 35}) = (\cdot 083, \cdot 095)$

Interpretation: i) Since the interval does not contain 1, there is significant evidence of association betn Gender & Supportiveness towards reopening
ii) With 95% conf, we can say that the odds of F supporting reopening is atleast ·083 times and atmost ·095 times the odds of males supporting reopening

The following R codes can be used to evaluate the odds ratio and the corresponding confidence intervals for the *Prime Minister's Dilemma* example:

```
install.packages("epitools")
library("epitools")
OR<-oddsratio(c(2631,7018,6121,1460), method="wald",conf=.95,correct=FALSE)
OR
```

3.5.5 Odds Ratio and Relative Risk

The relation between odds ratio and relative risk is simple as well as important. Supposing that the conditional proportions of supporting reopening for females and males are $\hat{p}_{s|f}$ and $\hat{p}_{s|m}$ respectively, the odds ratio will be

$OR = \frac{\hat{p}_{s|f}/(1-\hat{p}_{s|f})}{\hat{p}_{s|m}/(1-\hat{p}_{s|m})}$

$\hat{p}_{s|f}$ $\hat{p}_{s|m}$

$\frac{\hat{p}_{s|f}}{\hat{p}_{s|m}} \times \frac{1-\hat{p}_{s|m}}{1-\hat{p}_{s|f}}$

which can be rewritten as

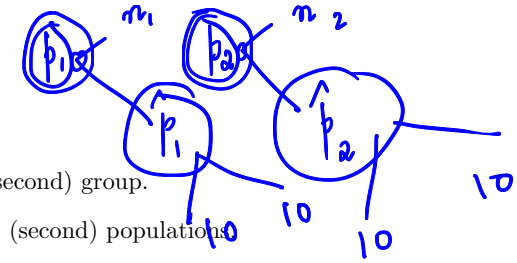
Relative risk$\times$ $\frac{1-\hat{p}_{s|m}}{1-\hat{p}_{s|f}}$

Clearly, the odds ratio and relative risk will have similar values if $\hat{p}_{s|m} = \hat{p}_{s|f}$.

3.6 Inferential Techniques

Suppose, you report the aforementioned measures of association to the Cabinet Committee who are tasked with taking the final decision on reopening of schools. Suppose, of all the members, the Prime minister summons you and ask "*Good that you have done all this. It seems the data depicts a strong association between gender and feedback and also that females are lot less supportive of reopening compared to males. However, are you sure that the same pattern will be visible in the population of all adult Indians ?*"

Well, the Prime minister was dead on. After all, the Chi square test indicated that gender and feedback was associated in the population but did not provide any indication of i) the actual strength or significance of association nor ii) the direction of association. On the other hand, the difference of proportions, relative risks and odds ratios indicated strengths of association but only in the observed sample. So, the obvious "missing link" is to figure out the strength, direction and significance of association between gender and supportiveness (or opposition) towards reopening in the population of all adult Indians. This can be determined using significance tests and confidence intervals for the difference of proportions (of males and females) who are supportive/opposed towards reopening. These techniques will be elaborated below.



3.6.1 Hypothesis Tests

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* Normal *under the CLT.*
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 : $(p_1 - p_2) = 0$, $p_1 - p_2 \leq 0$, $p_1 - p_2 \geq 0$ } Rule of hypotheses.
   - H_a : $p_1 - p_2 \neq 0$, $p_1 - p_2 > 0$, $p_1 - p_2 < 0$ }

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

$\hat{p}$ = 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 Standard error 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 p-value $< \alpha$ and would fail to reject H_0 otherwise.

Prime Minister's Dilemma: Now let us apply the above technique on the *Prime Minister's Dilemma* example to test whether there is a significant difference in the true proportion of males and females who are supportive of reopening.

1. **Assumptions** : Here the assumptions are satisfied because (i) the sample of 18,710 Indians was a random and representative sample and ii) the number of people of each gender who supported and opposed reopening of schools are much higher than 10.

2. **Hypotheses** :

   - H_0 : $p_1 - p_2 = 0$
   - H_a : $p_1 - p_2 \neq 0$

   $p_1 = p_f$, $p_2 = p_m$

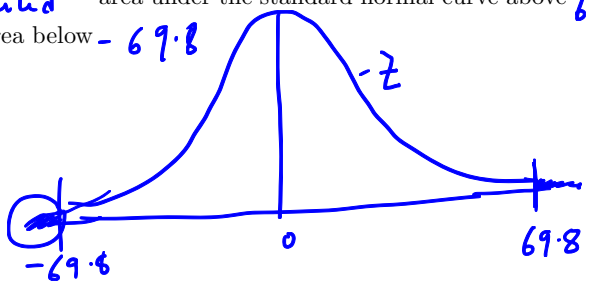
   $\hat{p}_1 = .3$ $\hat{p}_2 = .828$

3. **Test statistic** : The total number of supporters in the two samples taken together is 9649 and the total sample size is 17230 . Hence, the pooled sample estimate of supporters of reopening will be $9649/17230 = .56$. Hence the test statistic will be

$$Z = \frac{.3 - .828 - 0}{\sqrt{.56(1-.56)\left(\frac{1}{8752} + \frac{1}{8478}\right)}} = -69.8$$

Thus, the $\hat{p}_1 - \hat{p}_2$ falls about 69.8 (estimated) standard errors below the null value of 0.

4. **P-value** : Since our alternative is $<$ and our test statistic value is -69.8 the p-value will be the two-tailed area under the standard normal curve above 69.8 and below -69.8 which is double the area below -69.8



So, the p-value is 0

5. **Conclusion** : At all significance levels, the data set points towards a Sig. difference between the proportion of males & females who are supportive of reopening of schools until pan-India Vaccination is complete.

3.6.2 Confidence Interval

A confidence interval will yield a set of plausible values of the difference in the proportion of males and females who are supportive of reopening in the Indian population, with a prespecified degree of confidence. Following are the steps for evaluating such a confidence interval for our example.

1. **Notation**: Same as that for hypotheses test.

2. **Assumptions**: Same as that for hypotheses test.

3. **Structure** : As for one proportion, the confidence interval for $(p_1 - p_2)$ is obtained by adding and subtracting a Margin of error 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

$$\hat{se}(\hat{p_1} - \hat{p_2}) = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\sqrt{V(X_1+X_2)}$

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

$z^*_{\alpha/2}$

resulting in the confidence interval

$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}$

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.

Prime Minister's Dilemma: The following table depicts the conditional proportion values in the four cells corresponding to the data on 17,230 sampled subjects in our example:

	Support reopening ?	
Gender	Yes	No
Female	·3	·7
Male	·828	·172

1. **Assumptions**: Satisfied, as explained in the confidence interval section.
2. **Structure** : Here

   $\hat{p}_1 =$ ·3 $\quad \hat{p}_2 =$ ·828 $\quad , \hat{se}(\hat{p}_1 - \hat{p}_2) =$ $\sqrt{\frac{·3 \times ·7}{8752} + \frac{·828 \times ·172}{8478}}$ = ·0064

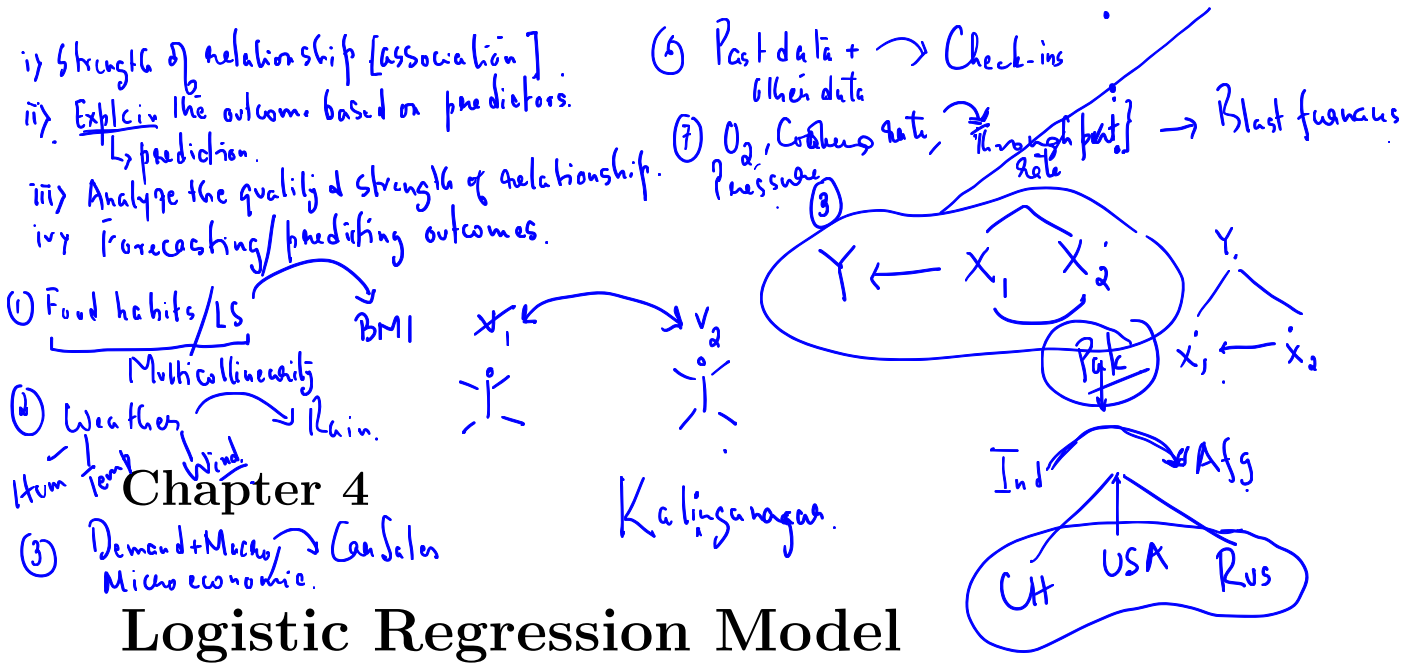
   Hence the required 95% confidence interval of $p_1 - p_2$ will be

   ·3 − ·828 ± 1·96 × ·0064 = (−·54, −·51)

3. **Interpretation**: We are 95% confident that the population proportion of females supporting reopening is at least 51% less and atmost 54% less than the corresponding proportion of males. Since the interval does not contain 0, the data provides significant evidence (at 95%) of association between gender & supportiveness towards reopening.
4. **Conclusion**: Thus there is a significant evidence towards differential perception about reopening between the genders.

The following R codes can be used to compute the confidence interval of the difference of proportions for the *Prime Minister's Dilemma* example:

```
test<-prop.test(c(2631, 7018),c(8752, 8748),conf.level=.95,correct=FALSE)
test
```



Chapter 4

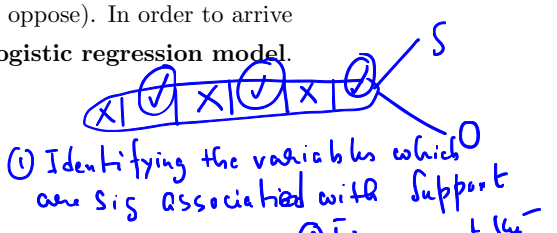
Logistic Regression Model

4.1 Motivation

The statistical techniques discussed in Chapters 1 and 2 were used to address the following two aspects of the *Nation's Dilemma* study in a data-based, scientific manner :

- Determination of the plausible values of the population proportion of adult Indians who are supportive of reopening of schools and educational institutes before the availability of a vaccine for Corona virus.
- Figuring out whether a particular attribute/characteristic of Indians (say, gender, occupation etc) significant affects their perception about reopening.

Well, so far so good. However, suppose the Prime minister would like to know whether a set of characteristics of individuals, say gender, occupation, age and number of children, taken together, have a collective effect on their perception about reopening (support or oppose). In order to arrive at a data-based, scientific answer to this question, we need to use the **Logistic regression model**. This is the theme of the current chapter.



4.2 What is Regression ?

Regression analysis is one of the most important branch of Statistics in that it helps us to analyse and quantify how different variables relate to each other and in doing so, aids in predicting the value/s of one of the variable from given values of the others. Due to its wide applicability in diverse disciplines, regression analysis has become one of the most (or probably the most) applicable technique of Statistical science.

As in the *Nation's Dilemma* study, variables can be of different types (say, discrete, continuous, categorical etc) and so are the regression models used to analyse the association between them. As a result, various types of regression models have (and are) been developed, from the simple to the

highly sophisticated ones, in order to precisely estimate the association pattern between variables and to generate accurate predictions of the response variable. These predictions can then be used by Governments and Businesses in framing policies and taking decisions which can have far reaching implications in our lives. Thus, in an increasingly data-centric world that we inhabit, the importance of regression analysis cannot be over-stated.

As mentioned above, the specific type of a regression model is usually dictated by the nature of variables to be modelled in the first place. More specifically, it is the nature of the response (or outcome) variable that determines the structure of the regression model. The most commonly used regression model, namely the **Linear regression model** is used when the response variable is Continuous . For example, linear regression model can be used to analyse how *market share of an FMCG product is influenced by factors like product category, price, amount spent in advertising the product, Gross Nielsen rating points etc.*

However, in many real life scenarios, we may be interested in analyzing how a set of variables impact a binary response variable, as is the case for the *Nation's Dilemma* study (in which we are interested in analyzing who different attributes of individuals influence their opinion about reopening of schools). As will be illustrated in the upcoming sections, linear regression models cannot be used in scenarios like these. The model that will help us analyze these kind of association patterns is known as the **Logistic regression model**.

4.3 Logistic Regression Models

4.3.1 Pervasiveness

As mentioned above, logistic regression models are used to analyse and quantify the association between a set of predictors (covariates) and a response variable that is binary in nature i.e has two possible outcomes, viz. success and failure. These kind of models have wide applicability across diverse disciplines as illustrated by the following examples:

Eg 1 [Marketing]: Whether a consumer would switch to Patanjali cosmetic products (yes/no) may depend on various factors like his/her location, age, income, education, monthly expenditure, expenditure on marketing the product etc.

Eg 2 [Medicine/Biology]: We may want to analyse whether the occurance of respiratory problems (yes/no) depends on factors/predictors like cigarette smoking behavior, gender, age and various other physiological factors of a group of patients.

Eg 3 [Finance]: Analysing whether and how payment of credit card bills on time (yes/no) depend on the size of the bill, annual income, occupation, mortgage and debt obligations and percentage of bills paid on time in the past of a group of subscribers etc.

Eg 4 [Political Science]: Analyzing how various factors like age, education, gender and income influence whether someone votes for a particular political party or not.

Eg 5 [Strategic Management]: Whether an Indian company/conglomerate would internationalize (i.e open offices/production units abroad) will depend on various factors like domain, yearly

revenue, employee size, whether listed or not, brand value at home, market share etc.

4.3.2 Why not Linear ?

In all the above examples, the response variable have two categories viz. success and failure, each of which comes with a specific probability. In the context of our research question, the response for the i^{th} individual can be expressed as

$$Y_i \begin{cases} 1 \rightarrow \text{success with } \pi_i \\ 0 \rightarrow \text{failure with } (1-\pi_i) \end{cases}$$

i.e Y_i follows a Bernoulli(π_i) distribution with probability of success $\pi_i = P(Y_i = 1)$ reflecting the probability that the i^{th} individual in the population is supportive of reopening. Thus,

$$E(Y_i) = 1 \times \pi_i + 0 \times (1-\pi_i) = \pi_i \quad \text{--- (i)}$$

Now, suppose we frame our model just like the simple linear regression model i.e

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $Y_i = 0, 1$ and $\epsilon_i \sim N(0, \sigma^2)$. Taking expectations of both sides, we have

$$E(Y_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) = \beta_0 + \beta_1 X_i = \pi_i$$

$$\text{(i) + (ii)} \Rightarrow \pi_i = \beta_0 + \beta_1 X_i$$

However, there are major problems with the above model viz.

1. **Constraints on the response function :** Since $0 \leq \pi_i = E(Y) \leq 1$, by the above equation, $\beta_0 + \beta_1 X$ should also lie within [0, 1] which may not always be true. For instance, suppose, we fit the above model to the *Nation's Dilemma* dataset with just one predictor, say age and obtain the following fitted model

$$\hat{\pi}_i = 1.456 - .25 x_i$$

where $\hat{\pi}_i$ is the predicted probability that an individual of age x_i will support reopening. Accordingly, the predicted probability that an individual aged, 75 years will be supportive of reopening will be

$$\hat{\pi}(75) = 1.456 - .25 \times 75 = -17.29$$

which is absurd since probability should always lie between 0 and 1.

2. **Non-normal errors** : In the above framework, $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$ whose possible values are

$$\epsilon_i \begin{cases} 1 - \beta_0 - \beta_1 X_i \\ 0 - \beta_0 - \beta_1 X_i \end{cases}$$

which clearly the normal distributional assumption for ϵ_i.

3. **Non-constant errors** : Since $Y_i \sim \text{Ber}(\pi_i)$, $\text{Var}(Y_i) = \pi_i(1-\pi_i)$ implying $\text{Var}(\epsilon_i) = (\beta_0+\beta_1 x_i)(1-\beta_0-\beta_1 x_i) \neq \sigma^2$

   Thus ϵ_i is a function of the predictor, which violates the constant variance assumption of linear regression models.

Based on the above discussion, it should be clear that the linear regression framework does not work binary response variables. Hence we have to opt for an alternative modeling framework - this is known as the **logistic regression model**.

4.3.3 Logistic Regression Function

It is quite clear from the above section that, instead of the linear function $(\beta_0 + \beta_1 x)$, we need a function of X, that also goes from 0 to 1. One such function is the logistic distribution function given by

$$F(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Let $\pi(x) = P(Y = 1|X = x) = E(Y|X = x)$. Then, the logistic regression model is given by

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Alternatively, it can also be written as

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x. \quad (4.1)$$

The transformation

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

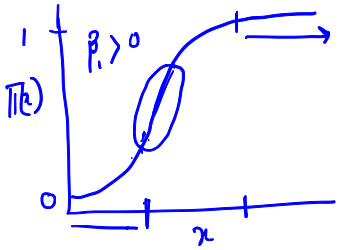
is known as the **logit transformation** of π and the ratio $\pi/(1-\pi)$ is known as the **odds**.

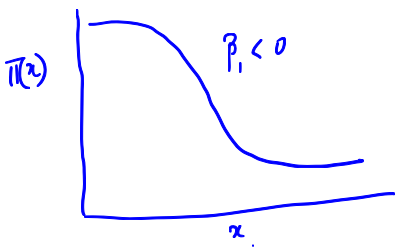
4.3.4 Characteristics

Some of the main properties of the logistic regression function are :

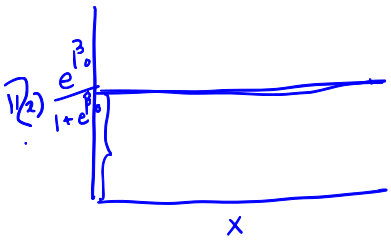
1. The mean response function $0 \leq E(Y|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \leq 1$ and it approaches these limits asymptotically.

2. The sign of β_1 determines whether $\pi(x)$ is increasing or decreasing with x. For $\beta_1 > 0$, the mean response function is monotonically increasing while for $\beta_1 < 0$, it is monotonically decreasing. The following figures depict the same.

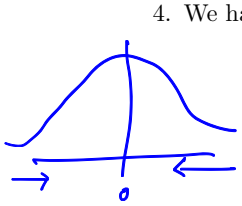




3. For $\beta_1 = 0$, Y is independent of X and the response function is a horizontal line.



4. We have



$$P(Y_i = 0) = 1 - P(Y_i = 1) = 1 - \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{\exp(0)\exp(-\beta_0 - \beta_1 X_i)}{1 + \exp(-\beta_0 - \beta_1 X_i)}$$

Thus, if we switch the 1's and the 0's in the response (Y), the signs of the regression coefficients are reversed. This is the symmetry property of the logistic response function.

5. It is important to note that unlike the linear regression model, β_1 is not the slope of the logistic mean function. This is because, the rate of change in $\pi(x)$ per unit x is not a constant but is a function of x, given by $\beta_1\pi(x)[1 - \pi(x)]$.

6. For $x = -\beta_0/\beta_1$

$$\pi(x) = \frac{\exp\left(\beta_0 + \beta_1\left(-\beta_0/\beta_1\right)\right)}{1 + \exp(\cdots)} = \frac{e^0}{1 + e^0} = \frac{1}{2} = .5$$

This is the point of the steepest slope and is called the **median effective level**. In toxicology studies, it is called LD_{50} (LD : lethal dose) i.e the dose with a 50% chance of a lethal result.

4.3.5 Interpretation of Regression Coefficients

As mentioned above, β_1 is no longer the slope of the logistic regression function. So, does β_1 have any interpretation after all ? The following discussion will make it clear.

$$X = x \rightarrow \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x \quad (i)$$

$$X = x + k \rightarrow \log\left(\frac{\pi(x+k)}{1 - \pi(x+k)}\right) = \beta_0 + \beta_1(x + k) \quad (ii)$$

$$(ii) - (i) = \log\left[\frac{\pi(x+k)/(1 - \pi(x+k))}{\pi(x)/(1 - \pi(x))}\right] = k\beta_1 \Rightarrow \log\left[\frac{\text{odds}(x+k)}{\text{odds}(x)}\right] = k\beta_1$$

$$\Rightarrow \text{odds}(x + k) = e^{k\beta_1}\,\text{odds}(x)$$

Thus, for k unit increase in x, the odds increase multiplicatively by $e^{k\beta_1}$. This is how β_1 is interpreted for a logistic regression model.

4.3.6 Age and Feedback

Let us illustrate the concepts learnt above in analyzing the effect of Age on whether an individual supports reopening or not. Let Y denote the feedback and X the age of a respondent such that

$$Y = \begin{cases} 1 & \text{if supports} \\ 0 & \text{o.w.} \end{cases}$$

On fitting the logistic model to the data, suppose we obtain the following fitted model

$$\hat{\pi}(x) = \frac{\exp(-1.13 - .02x)}{1 + \exp(-1.13 - .02x)}$$

where $\pi(x) = P(Y = 1|X = x)$.

- Since $\beta_1 < 0$, $\pi(x)$ ↓ with x i.e probability of being supportive of reopening with age.
- An individual will have a 50% chance of supporting reopening if his/her age is $-\beta_0/\beta_1 = -(-1.13)/(-.02) = 56.5$.
- For one year increase in age, the estimated odds of being supportive of reopening decreases by $e^{-.02} = .98$ i.e by 2%. So, for 10 years increase in age, the estimated odds of being supportive of reopening will ↓ by $e^{-10 \times .02} = e^{-.2}$ i.e a person will be 20% less likely to be supportive of reopening if his/her age increases by 10 years.

change in multiples of .98

For 10 yrs ↓ in age, estimated odds of supporting reopening increases multiplicatively by $1/e^{-.2} = e^{.2} = 1.2 = 22\%$.

4.3.7 Multivariate Logistic Models

The simple, one-predictor logistic regression model can be easily extended by including multiple predictors, say $\boldsymbol{X} = (X_1, X_2, ..., X_p)$. Thus, we have

$$\text{logit}(\pi(\boldsymbol{x})) = \log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = \boldsymbol{x}'\boldsymbol{\beta} \quad (4.2)$$

where $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$. Alternatively, we can also state

$$\begin{aligned} & Y_i \sim \text{Ber}(\pi_i) \\ \text{where} \quad & E(Y_i) = \pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \end{aligned} \quad (4.3)$$

Note 3. *As for multiple linear regression model, the predictors can be continuous, categorical and can contain polynomial and interaction terms. Thus, the multiple logistic model has wide applicability in real life scenarios.*

4.3.8 Revisiting Nation's Dilemma

Let us incorporate three more predictors (on top of age) in the *Nation's Dilemma* example viz. gender (1: male, 0: female), occupation (1: teaching, 0: otherwise) and number of children, in order to better predict the probability of supportiveness towards reopening. Suppose the corresponding

coefficients for the intercept and the predictors are $(-2.40, -.08, .12, -.18, .10)$. So, the fitted logistic regression model will be

$$\text{logit}\left(\hat{\pi}(x)\right) = \log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -2.4 - .08\,\text{Age} + .12\,\text{Gen} - .18\,\text{Occp} + .10\,\text{Child}.$$

Based on the coefficients, we have the following interpretations regarding the effect of the predictors:

- The estimated probability of supporting reopening with age controlling for Gen, Occp & Child. older the individual, more likely he/she will not support reopening controlling for the other predictors.
- Males are more likely to be supportive of reopening compared to Females controlling for age, occupation and number of children. In fact, the odds that a male will be supportive of reopening is $e^{.12} = 1.13$ times the odds that a female will be supportive of reopening controlling for
- Non-teachers are more likely to be supportive of reopening compared to teachers controlling for age, gender and number of children. In fact, the odds that a teacher will be supportive of reopening is $e^{-.18} = .835$ times the odds that a non-teacher will be supportive of reopening controlling for
  $\Rightarrow$ The estimated odds that non-teachers will be supportive is $1/e^{-.18} = e^{.18} = 1.20$ times the odds that teachers will be supportive.
- The estimated probability of supporting reopening with the number of children one has controlling for the other predictors. In fact, the odds that a person having 2 children will be supportive of reopening is $e^{.1} = 1.10$ times the odds that a single child parent will be supportive of reopening controlling for Age, Gender & Occup.

  $1/.835$ ↑ 1.20

  $X_1 \to X_1 + k$

  3 vs 1 $\to e^{2\times.1}$ 2.

Predictor

The estimated probability that a female teacher, aged 43 years having two children will be supportive of reopening will be

$$\hat{\pi}(43, F, T, 2) = \frac{\exp(-2.40 - .08\times 43 + .12\times 0 - .18\times 1 + .10\times 2)}{1 + \text{''}}$$

(F: 0, T: 1)

4.4 Parameter Inference

In addition to interpreting the effect of predictors, it is of paramount importance to test whether a certain predictor (or a set of predictors) have a significant influence on the probability of supporting reopening. We now provide a brief exposition on how to carry out such inferential procedures for a logistic regression model in the context of the *Nation's Dilemma* study.

The following R output depicts the parameter estimates and estimated standard errors for the above study:

Predictor	Estimate	Standard error
Intercept	-2.40	0.12
Age	-0.08	0.01
Gender	0.12	0.005
Occupation	-0.18	0.014
Number of children	0.10	0.08

4.4.1 Hypotheses Test

Suppose we want to test whether *gender* has significant influence on supportiveness controlling for the other predictors. So, the null and alternative hypotheses will be

$$H_0: \beta_2 = 0 \qquad H_a: \beta_2 \neq 0$$

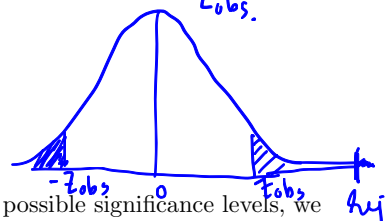
The test statistic for testing the above hypotheses is known as the **Wald test statistic**, named after the celebrated statistician Prof. Abraham Wlad. It is given by

$$Z_{obs} = \frac{\hat{\beta}_2 - 0}{\hat{se}(\hat{\beta}_2)} = \frac{.12 - 0}{.005} = 24$$

$$\frac{\hat{p} - p_0}{\hat{se}(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and it follows the $Z \sim N(0,1)$ distribution under H_0.

Since the alternative hypotheses is two-sided, the p-value will be the two tailed area above Z_{obs} and below $-Z_{obs}$ i.e double the area below $-Z_{obs}$ which can be assumed to be 0 for all practical purposes.



Since the p-value is less than all possible significance levels, we reject H_0 at all the commonly used significance levels and conclude that gender has significant influence on supportiveness controlling for the other predictors. In fact, since the gender coefficient is ~~negative~~ positive, Males are more likely to Support reopening compared to ~~males~~ Females. Significantly

4.4.2 Confidence Interval

The 95% confidence interval for the coefficient of gender i.e β_2 will be

$$\hat{\beta}_2 \pm Z_{.025} \times \hat{se}(\hat{\beta}_2) = .12 \pm 1.96 \times .005 = (.11, .13)$$

$\beta_2 \to \exp(\beta_2)$ $\frac{d\beta_2}{d\exp(\beta_2)}$

k $k-1$ $y = x^3$ $\frac{dx}{dy}$ $x_j \to \beta_3$

and that for the odds ratio $\exp(\beta_2)$ will be approximately

Thus, we are 95% confident that controlling for the other predictors, the odds that a male will support reopening is at least 1.12 and at most 1.14 times the odds that a female will do the same. No wonder that the conclusion corroborates the one we had drived from the hypotheses test. Similar tests can also be done for the other regression coefficients as well.

4.4.3 Likelihood Ratio Test

Often, we may want to test whether a subset of the predictors have any significant association with the response/probability of success. So, this is analogous to the Partial-F test for multiple linear regression and is based on comparing the full and reduced models. Following is a stepwise procedure of how it is done :

1. The *full model* contains all the p predictors, $(X_1, ..., X_p)$. Let $\hat{\boldsymbol{\beta}}_f$ be the vector of estimated regression coefficients for this model and $L(\hat{\boldsymbol{\beta}}_f)$ be the log-likelihood function evaluated at these estimates.
2. Suppose we want to test whether a subset, say $(\beta_{q+1}, ..., \beta_p)$ is significant or not i.e our hypotheses will be :
3. So, the *reduced model* contain the predictors $(X_1, ..., X_q)$ and let the corresponding estimated coefficients be $\hat{\boldsymbol{\beta}}_r$ while $L(\hat{\boldsymbol{\beta}}_r)$ be the log-likelihood function evaluated at these estimates.
4. It can be shown that $L(\hat{\boldsymbol{\beta}}_r) \leq L(\hat{\boldsymbol{\beta}}_f)$ since the likelihood function will always achieve its maximum at the maximum likelihood estimates.
5. The *likelihood ratio* test statistic is given by

   which follows a chi-square distribution with $p - q$ degrees of freedom i.e
6. **Decision rule** : If $G^2 > \chi^2_{\alpha, p-q}$, we reject H_0 at significance level α; and fail to reject H_0 otherwise.

Eg: Let us illustrate the likelihood ratio test for checking whether number of children can be dropped from our model when gender, occupation and age are already there.

The full model containing all these predictors has a log likelihood value of $L_f = -50.527$ while the reduced model containing the variables (G, O, Age.) has the log likelihood value of $L_r = -53.102$. Thus, the likelihood ratio test statistic is

From the table, we have $\chi^2_{.05}(1) = 3.84$ which is less than G^2. Hence, we reject the null hypotheses and conclude that number of children cannot be dropped from the model given the other predictors. Similar tests for other predictors and intercation/polynomial terms can be carried out as well.

4.5 Model Selection

As always, it is important to select the correct predictors and/or the relevant functional form of the predictors in the logistic regression model. For that purpose, three criterions are generally used : (i) **Akaike information criterion (AIC)**, (ii) **Bayesian information criterion (BIC)** and (iii) **Deviance information criterion (DIC)**. The modified forms of these criterions are as follows (here p denotes the number of predictors):

- $AIC_p = -2log_eL(\hat{\boldsymbol{\beta}}) + 2(p+1)$
- $BIC_p = -2log_eL(\hat{\boldsymbol{\beta}}) + (p+1)log(n)$
- $DIC = -2log_eL(\hat{\boldsymbol{\beta}})$

The AIC and BIC penalize the likelihood for large number of parameters while the DIC does not. As always, lower values of these criterion will indicate a better model.

Apart from these, automated model selection procedures like **best subsets** or **stepwise model selection** can be adapted to identify good subset models and eventually the optimal model for a given dataset/scenario. The procedures (and considerations) are very similar to those for multiple linear regressions. However, for stepwise model selection, addition or deletion of a predictor depends on the Wald (Z) statistic (not the t-statistic as for the multiple linear model).

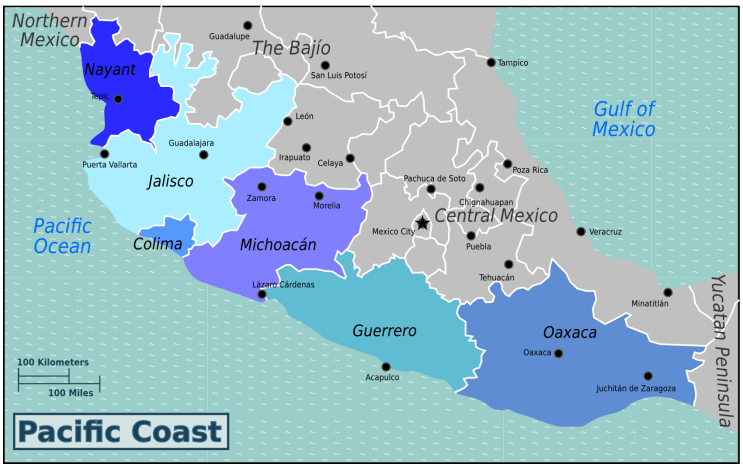
4.6 Case Study: Disease Outbreak

We will now implement the aforementioned model selection techniques on a data set relating the outbreak of a particular disease, with the sole aim of identifying the optimal predictive model. We will denote this as the *Disease Outbreak Study*.

4.6.1 Context

Dengue is a mosquito borne viral disease that commonly occurs in tropical and sub-tropical regions. According to the Centers for Disease Control and Prevention, Dengue is prevalent in more than 100 countries around the world and affects around 400 million people annually of whom around 22,000 people succomb to it. Accurate identification of the patterns and risk factors of Dengue is very important in planning for proper interventions. This also aids in the prediction of the trends of the disease and identification of specific attributes that makes certain type of individuals more prone to the disease. In this case study, we will leverage the tools of logistic regression analysis to identify

the patterns and risk factors of Dengue, using dataset related to an outbreak that occurred in the pacific coast of Mexico during the late 1980's. In doing so, we would endeavour to determine an optimal predictive model that can be used to ascertain the likelihood of a person contracting this disease based on certain attributes specific to him/her. Last but not the least, the study and the ensuing analysis assumes special relevance in the midst of the current Covid-19 outbreak that has devastated the world in general and India in particular.



4.6.2 Data Description

The dataset for this study was obtained from a random sample of 196 subjects selected from two sectors of a Mexican city. The response variable was Disease status (Y, coded as 1 if dengue, 0 if not) while the explanatory variables were Age (X_1), Socio-economic status (X_2 and X_3) and Sector (X_4). Socio-economic status (SES) has three categories, namely Upper, Middle and Lower, which are coded using the two dummy variables, X_2 and X_3, such that

$$X_2 = \begin{cases} 1 & \text{if Middle} \\ 0 & \text{otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 & \text{if Lower} \\ 0 & \text{otherwise} \end{cases}$$

implying that the Lower SES category would correspond to . Lastly, Sector is coded as

$$X_4 = \begin{cases} 0 & \text{if Sector 2} \\ 1 & \text{if Sector 1} \end{cases}$$

The reason behind the aforementioned codings are that the prevalence of Dengue was expected to be lowest in the Upper SES and in Sector 1. There was a fourth explanatory variable, namely Savings account status which was not considered in the analysis since it had negligible impact on disease status. A brief description and nomenclature of the variables is given in Tab 1 below:

Table 1 : Brief description of the data

S.No	Variable	Type of variable	Description
1	Identification no.	Discrete	Provides information on other variables for a single person
2	Age	Discrete	Age of a person (in Years)
3	Socio-Economic Status	Categorical	• Upper class • Middle Class • Lower class
4	Sector	Categorical	Indicates sector within the city; • Sector_1 • Sector_2
5	Savings account status	Binary	• No Savings Account = 0 • Person Has Savings Account = 1
6	Disease Status	Binary	Indicates the presence/absence of the disease • Absence = 0 • Presence =1

4.6.3 Model Fitting

The primary purpose of the study is to assess the strength of association between each of the aforementioned predictors and the likelihood of disease and in doing so, identification of the optimal predictive model. Towards that end, the dataset was split equally into training and testing components, each consisting of 98 individuals selected randomly from the main data. The training data will be utilized to identify/train the optimal model while the testing data will be used to evaluate the predictive capability of that model and any other competitive models.

1. Model with all predictors

At first, we fit the model with all the 3 predictors on the training dataset i.e

$$\text{logit}(\pi(\boldsymbol{x})) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Middle} + \beta_3 \text{Lower} + \beta_4 \text{Sector2} \tag{4.4}$$

resulting in the following estimates table

Predictor	Estimate	Standard error
Intercept	-2.31	0.64
Age	0.03	0.013
Middle	0.41	0.6
Lower	-0.31	0.60
Sector2	1.57	0.50

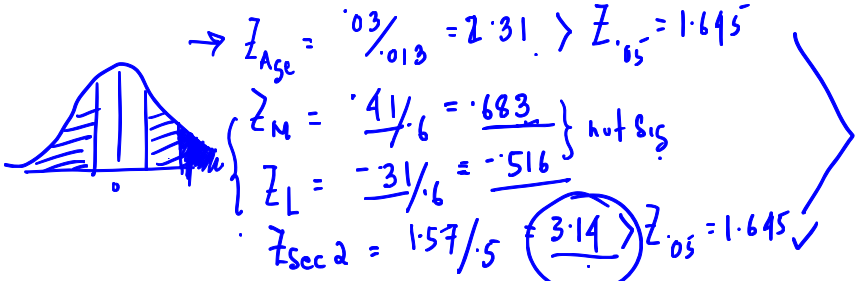
Thus, the fitted logistic regression model is given by

$$\hat{\text{logit}}(\hat{\pi}(x)) = -2.31 + 0.03\,\text{Age} + 0.41\,\text{Middle} - 0.31\,\text{Lower} + 1.57\,\text{Sector2}$$

The R codes for fitting the above model are as follows.

```
epidemic<-read.table("C:/Users/IIMA/Desktop/CDA/Epidemic.txt",header=TRUE)
epidemic.tr<-epidemic[1:98,]
install.packages("car")
library(car)
age.tr<-epidemic.tr$Age
sec.tr<-recode(epidemic.tr$Sec,"1=0;2=1")
sesmid.tr<-recode(epidemic.tr$SES,"2=1;c(1,3)=0")
seslow.tr<-recode(epidemic.tr$SES,"3=1;c(1,2)=0")
dis.tr<-epidemic.tr$Dis
Training<-data.frame(dis.tr,age.tr,sesmid.tr,seslow.tr,sec.tr)
fit1.tr<-glm(dis.tr~age.tr+sesmid.tr+seslow.tr+sec.tr,family=
binomial(link="logit"),data=Training)
summary(fit1.tr)
```

Do it: Can you identify which of the predictors have a significant effect on disease status ?



Interpretation: Let us interpret the impact of Age and Sector on Disease status:

1. Controlling for Socio-economic status and Sector location, Age has a positive association with Disease status. Specifically, the estimated odds of contracting Dengue increase in multiples of $e^{.03}=1.03$ for every one year increase in age. In other words, everything else remaining constant, older people have a higher odds of contracting Dengue.
2. Controlling for Age & SES, the estimated odds of contracting Dengue for a person staying in Sector 2 is nearly $e^{1.57}=4.8$ times that of someone residing in Sector 1. In other words, Sector 2 residents are at a much higher risk of contracting Dengue, controlling for the other predictors.

2. Variable selection

The three variables, namely Age, Socio-economic status and Sector were included in the primary model since those are considered *apriori* important. Now, we will perform likelihood ratio (LR) tests to determine the significance of each of these predictors *given* the others. We will do so by dropping each predictor in turn and assessing the impact that has on the model per se.

1. **Age:** Once Age was dropped, the reduced model, say R consisted of Sector & SES. The log-likelihood of R can be shown to be –53.102. On the other hand, the log-likelihood of the

full model, say F, consisting of all the 3 predictors, is –50.527. Thus, the likelihood ratio test statistic for testing the following hypotheses

$$H_0: \beta_1 = 0 \qquad H_a: \beta_1 \neq 0$$

would be

$$G^2_{obs} = 2\left[L(F) - L(R)\right] = 2\left[-50.527 + 53.102\right] = 5.15$$

which follows a χ^2_1 distribution under the null hypotheses. At $\alpha = .05$, it can be shown that $\chi^2_{.05,1} = 3.84$ implying that the p-value for the above test statistic would be $< .05$. Thus, we would reject H_0 implying that given Sector and Socio-economic status, Age cannot be removed from the model.

The R codes for performing the above LR test is given below:

```
fit2.tr<-glm(dis.tr~sesmid.tr+seslow.tr+sec.tr,family=binomial(link=
"logit"),data=Training)
anova(fit1.tr,fit2.tr,test="LRT")
```

2. **SES:** The likelihood ratio test for SES would involve testing for the following hypotheses

$$H_0: \beta_2 = \beta_3 = 0 \qquad H_a: \text{at least one of } \beta_2, \beta_3 \neq 0$$

On performing the LR test, we obtain a p-value of 0.55 implying that, given Age and Sector, Socio-economic status can be removed from the model i.e it is not significantly impactful on the likelihood of Dengue in the presence of other predictors.

The R codes for performing the above LR test is given below:

```
fit3.tr<-glm(dis.tr~age.tr+sec.tr,family=binomial(link="logit"),data=
Training)
anova(fit1.tr,fit3.tr,test="LRT")
```

3. **Sector:** The LR test to verify whether Sector can be dropped from a model containing Age and Socio-economic status yielded a p-value of 0.001 implying that Sector cannot be dropped i.e that it has significant association with disease status controlling for the other variables.

The R codes for performing the above LR test is given below:

```
fit4.tr<-glm(dis.tr~age.tr+sesmid.tr+seslow.tr,family=binomial(link=
"logit"),data=Training)
anova(fit1.tr,fit4.tr,test="LRT")
```

3. Multicollinearity check

Although the likelihood ratio tests are indicative of the redundancy of Socio-economic status given Age and Sector, it was retained keeping in mind its *apriori* importance. This decision was also corroborated by the fact that the estimated regression coefficients and the corresponding standard errors of Age and Sector were *not* sensitive to the presence or absence of Socio-economic status implying low multicollinearity between these variables.

The R codes for performing the sensitivity analysis is given below:

```
summary(fit1.tr)
summary(fit3.tr)
```

4. Interaction effects

In addition to testing for the main effects, it is also important to verify whether the variables interact with each other in impacting the disease status. We will now test for the significance of any two-factor interaction terms *in addition* to the main effects.

The interaction or Full model (F) is given by

$$\text{logit}(\pi(\boldsymbol{x})) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1\text{Age} + \beta_2\text{Middle} + \beta_3\text{Lower} + \beta_4\text{Sector2} + \beta_5\text{Age} \times \text{Middle} + \beta_6\text{Age} \times \text{Lower} + \beta_7\text{Age} \times \text{Sector2} + \beta_8 \text{Mid} \times \text{Sec2} + \beta_9 \text{Low} \times \text{Sect2}$$

Testing for the significance of the interaction effects is synonymous to testing for the following hypotheses

$$H_0: \beta_5 = \cdots = \beta_9 = 0 \qquad H_a: \text{at least one of } \beta_5 \cdots \beta_9 \neq 0$$

Thus, the reduced model, R will be

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Middle} + \beta_3 \text{Low} + \beta_4 \text{Sec 2}$$

The log-likelihoods corresponding to the Full and Reduced models are given by $L(F) = -46.998$ and $L(R) = -50.527$ implying that the LR test statistic is

$$G^2_{obs} = 2\left[L(F) - L(R)\right] = 2\left[-46.998 + 50.527\right] = 7 < 11.07$$

which follows a Chi-square distribution with 5 degrees of freedom under the null. It can be shown that at $\alpha = .05$ $\chi^2_{.05,5} = 11.07$ which is $> G^2_{obs}$ than G^2_{obs}. Thus, we fail to rej H_0 at $\alpha = .05$ and conclude that, given the main effect terms, two factor interaction effects are insignificant. In other words, the training data does not provide significant evidence that the variables interact with one another in impacting the disease status.

The R codes for performing the LR test for interaction effects is given below:

```
fit5.tr<-glm(dis.tr~age.tr+sesmid.tr+seslow.tr+sec.tr+age.tr*sesmid.tr+
age.tr*seslow.tr+age.tr*sec.tr+sesmid.tr*sec.tr+seslow.tr*sec.tr,family=binomial
(link="logit"),data=Training)
anova(fit1.tr,fit5.tr,test="LRT")
```

5. Best subsets procedure

We will now be applying the best subsets procedure, mentioned in Sec 4.5, in an attempt to identify the optimal subset of predictors for the Disease Outbreak study. In doing so, we will be basing our judgement on the values of AIC, BIC and DIC measures for each of the candidate model corresponding to a particular predictor combination. Since there are 4 predictors, there will be 16 possible models. The table in the next page depicts the above measures for each of these models. (The presence or absence of a particular predictor in a given model is denoted by 1 or 0 in the respective predictor column corresponding to that model.)

Model	Age	SES(Middle)	SES(Lower)	Sector	AIC	BIC	DIC
1	0	0	0	0	124.32	126.9	122.32
2	1	0	0	0	118.91	124.08	114.91
3	0	1	0	0	124.88	130.05	120.88
4	0	0	1	0	122.23	127.4	118.23
5	0	0	0	1	111.53	116.7	107.53
6	1	1	0	0	119.11	126.86	113.11
7	1	0	1	0	117.97	125.72	111.97
8	1	0	0	1	108.26	116.01	102.26
9	0	1	1	0	124.08	131.84	118.08
10	0	1	0	1	112.88	120.64	106.88
11	0	0	1	1	112.37	120.13	106.37
12	1	1	1	0	119.5	129.84	111.5
13	1	1	0	1	109.31	119.65	101.31
14	1	0	1	1	109.52	119.86	101.52
15	0	1	1	1	114.20	124.54	106.20
16	1	1	1	1	111.05	123.98	101.05

The R codes for evaluating the model selection criterion for the full model is given below. Criterions for the other models can be evaluated similarly :

```
summary(fit1.tr) ##AIC and DIC are given below the estimates table##
BIC(fit1.tr) ##BIC criterion##
```

Based on the model comparison measures, the following models can be shortlisted in order of preference (or "goodness").

	AIC Criterion		BIC Criterion	
Rank	Predictors	AIC	Predictors	BIC
1	Age, Sector	108.26	Age, Sector	116.01
2	Age, Middle, Sector	109.31	Sector	116.7
3	Age, Lower, Sector	109.52	Age, Middle, Sector	119.65
4	Age, Middle, Lower, Sector	111.05	Age, Lower, Sector	119.86

Thus, both the AIC and BIC criterion has identified the model with Age and Sector as the optimal model as far as the training data is concerned. In order to reconfirm this finding, we perform a stepwise model selection routine next.

6. Stepwise model selection

As the term suggests, in this procedure, inclusion of variables is carried out in a stepwise manner. At first, the "best" predictor is included, followed by the next best provided the p-values are less than .05. The process terminates as soon as the p-values of a newly added predictor exceeds .05.

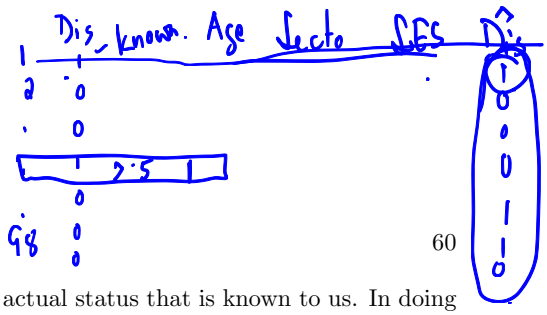
On performing the stepwise procedure for the Disease outbreak study, it is observed that the Sector variable is added first, its p-value being 0. In the next step, age is added, its p-value being 0.026. At this stage, the process terminates since any additional variable has p-values higher than .05. The following table has the results:

	Predictor	Estimate	Standard error	Test statistic	P-value
Step 1	Intercept	-3.33	0.76	18.99	0
	Sector	1.74	0.47	13.59	0
Step 2	Intercept	-4.00	0.87	21.06	0
	Sector	1.67	0.49	11.79	0
	Age	0.03	0.013	4.95	0.026

Thus, the stepwise procedure identifies the model with Age and Sector as the optimal model, thus corroborating the results of the best subsets procedure.

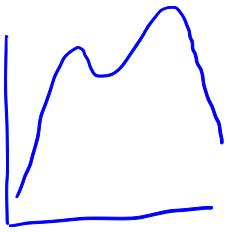
4.6.4 Predictive Ability

One of the best metrics for identifying the optimal logistic regression model is its predictive ability i.e the precision with which the model can accurately predict the response (1 or 0) for a new dataset. For the Disease outbreak study, we will use the testing dataset to assess the prediction accuracy of the model with Age, Socio-economic status and Sector as predictors. In other words, we will use this model to predict the disease status of the 98 subjects included in the testing data and in doing



so, would compare the predicted disease status with the actual status that is known to us. In doing so, we will have a fair idea of the prediction accuracy and the generalizability of the above model.

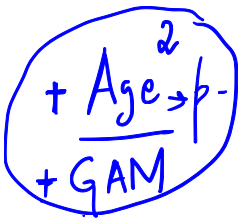
For the purpose of prediction, it is imperative to determine the cutoff point for the predicted probability, $\hat{\pi}$, above which a response would be deemed a "success" and below which it would be considered a "failure". For the purpose of our example, we will use the most common threshold, namely 0.5 i.e our prediction rule will be



$$\hat{Y}_h = \begin{cases} 1 & \text{if } \hat{\pi}_h > 0.5 \\ 0 & \text{if } \hat{\pi}_h \leq 0.5 \end{cases}$$

On implementing the above prediction rule on the testing data, we have the following 2×2 table:

	Predicted disease state		
True disease state	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	47	20	67
Y = 1	8	23	31
Total	55	43	98



Thus, out of 67 diseased subjects, the number of correct predictions was 47 while the same for healthy subjects was 23 . In other words, the *sensitivity* and *specificity* of the procedure are

$$P(\hat{Y} = 1|Y = 1) = 23/31 = 74.2\%. \quad P(\hat{Y} = 1|Y = 0) = 20/67 = 29.9\%.$$

Overall, this implies a prediction error rate of $28/98 = 28.6\%$.

One way of verifying the predictive ability of the model is by comparing the prediction error rates corresponding to the testing and training datasets. If the rates are close, that would imply that the model is robust or generalizable as a predictive model. Towards that end, we use our model to predict the disease status for the training data as well. The following table has the results

	Predicted disease state		
True disease state	$\hat{Y} = 0$	$\hat{Y} = 1$	Total
Y = 0	50	17	67
Y = 1	9	22	31
Total	59	39	98

Thus, the sensitivity and specificity for the training data are $22/31 = 73\%$ and $17/67 = 26\%$ respectively while the overall prediction error rate is $26/98 = 26\%$. Since the prediction error rates corresponding to the training and testing data are pretty close, we can conclude the superiority of the model with Age, Sector and Socio-economic status in terms of its predictive ability. In other words, we can use this model on other datasets in order to predict the disease status of a subject for given values of his/her predictors.

Chapter 5

Multicategory Logit Models

5.1 Motivation

In the previous chapter, we have explored ways of assessing the impact of various attributes on the likelihood of supporting or opposing the reopening of educational institutions through a logistic regression model. However, categorizing supportiveness into just two categories, viz. "support" or "oppose" may eclipse finer differences in attitudes among subjects regarding this issue. For instance, some individuals may strongly oppose reopening while some may have mild opposition. Not able to differentiate these attitudes (or assuming them to be similar) may be unrealistic and hence result in unintended oversimplifications.

Along the same lines, it might be more realistic to reclassify supportiveness into the following categories: (strongly support,).
Since the response variable has five categories, a binary logistic model will not be adequate for modelling it. In fact, the distribution of the counts of individuals belonging to the aforementioned categories will be as opposed to Binomial which formed the basis for the logistic regression model. In order to account for the multicategory (> 2) nature of supportiveness and to assess the effect of various attributes on it, we need to use a class of models known as **Multicategory Logit Models** which, as the term suggests, is an extension of binary logistic models to the case of multicategory outcomes.

In the context of the "Nation's Dilemma" example, a multicategory logit model will enable us to assess whether attributes such as gender, occupation, age and number of children etc have a collective effect on the degree of supportiveness towards reopening where supportiveness is categorized into 5 categories as shown above. The remainder of this chapter is devoted to an exploration of these models.

5.2 Applicability

Just like linear and logistic regression models, multicategory logit models have wide applicability across diverse disciplines as illustrated in the following examples:

Eg 1. [**Education**]: Educational achievement of students in a foreign university, categorised as (high distinction, distinction, credit, pass and fail) may depend on various factors like final exam score, case study score, team presentation score, tutorial grade, total weighted score, student gender, highest degree of the instructor (Phd or not), immigrant status of student (whether native or foreigner), instructor experience, number of classes missed, GMAT score etc.

Eg 2. [**Health**]: Satisfaction about the Government's handling of the Coronavirus crisis, categorised as (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied) may depend on various factors like age, gender, educational qualification, occupation, gross annual income, political party affiliation (BJP, Congress, AAP etc), religion, ethnicity etc.

Eg 3. [**Finance**]: The decision regarding a credit card application (accept, reject or put on hold) may depend on various factors of the applicant like age, income, gender, number of defaults, number of years owning a credit card, whether salaried or not etc.

Eg 4. [**Travel/Tourism**]: The ratings of hosts/properties on Airbnb (excellent, very good, good, average and poor) may depend on various factors like locality, per night charges, amenities, type of establishment, number of reviews, review score ratings, cancellation policy, years of operation, number of bedrooms etc.

Eg 5. [**Neuroscience/Spirituality**]: The cortical thickness in the hippocamus of the brain (higher thickness → better memory & learning abilities) categorized as (very thick, thick, normal, thin, very thin) may depend on age, gender, number of hours spent each day in meditation, number of years of practising meditation, type of meditation practised (Vippasanna, Zen, Om, transcendental, mindfulness, loving-kindness), occupation, nature of home/work environment etc.

The nature of categorization of the response variable influence the type of multicategory model that will be used in a particular situation. For instance, in examples 1, 2, 4 & 5, the categories have a natural ordering (i.e are ordinal) while for example 3, the categories are not ordered i.e are nominal. When the response categories are ordered, **Cumulative Logit Models** (also known as **Proportional Odds Models**) are used while for nominal responses, **Baseline Category Logit Models** are used. In the following sections, we will be exploring each of these models in the context of a real life example.

5.3 Cumulative Logit Model

5.3.1 Motivating Example

In order to illustrate the modeling of ordinal response variables, we will be using a dataset related to educational achievement of undergraduate management students at a major Australian university. Interest is in identifying attributes which have a significant effect in determining educational

achievement, which is categorized as (high distinction, distinction, credit, pass and fail). Some of the predictors which are considered in this study are final exam score, case study score, team presentation score, tutorial grade, total weighted score, student gender, highest degree of the instructor (Phd or not), immigrant status of student (whether native or foreigner), instructor experience, number of classes missed, entry score etc. Data was collected on a random and representative sample of 990 undergraduate students enrolled in the business school of the above university.

In order to answer the aforementioned question vis-a-vis figure out the predictors those have a significant association with educational achievement, we need to use a Cumulative Logit model, as explained in the following sections.

5.3.2 Structure

Let the population probabilities corresponding to the five categories of educational achievement are $(\pi_1, ..., \pi_5)$ such that $\sum_{i=1}^{5} \pi_i = 1$. The cumulative probability corresponding to response category j is given by

$$P(Y \le j) = \pi_1 + \cdots + \pi_j \qquad , j = 1, 2, ..., 5$$

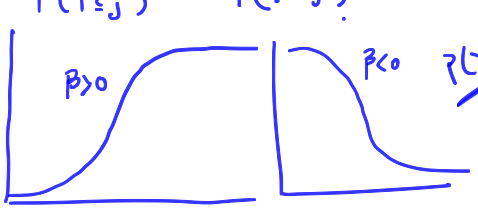
For instance, if $j = 2$, the above cumulative probability corresponds to the likelihood of obtaining at least a distinction. As the term suggests, a cumulative logit model links the logit of the above cumulative probabilities with a linear function of the predictors. In the most general form, it is given by

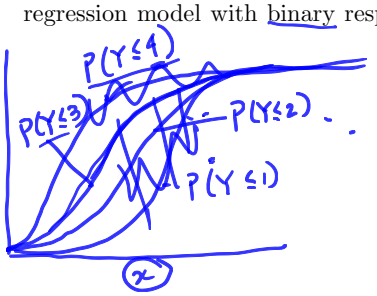
$$\text{logit}[P(Y \le j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \qquad j = 1, ..., 4.$$

where

$$\text{logit}[P(Y \le j)] = \log\left[\frac{P(Y \le j)}{P(Y > j)}\right]$$

In the above model, β_p quantifies the effect of the p^{th} predictor x_p, on the log odds of a response in category j or below, controlling for the remaining predictors. (In the context of this particular problem, this is synonymous to a log odds of attaining an educational achievement corresponding to that of category j or *higher*). One critical assumption underlying this model is that the effect of a particular predictor, say x_p, remains the same for all the 4 cumulative logits. This is reflected in the following figure which depicts each of the four cumulative probabilities plotted as functions of a predictor x, (in a single predictor setup). Clearly, the curves have similar shapes which is due to the constant effect of the predictor x. It is important to note that the curve corresponding to $P(Y \le j)$ is the same as that for a logistic regression model with binary response outcomes and $P(Y > j)$ respectively.





5.3.3 Interpretation

Suppose we would like to interpret the regression coefficient β_p vis-a-vis assess the effect of the predictor x_p on the log odds of response belonging to category j or lower, controlling for the other predictors. Towards this end, let us consider two values of X_p, say $X_p = x_p$ and $X_p = x_p + k$. For these two values, the corresponding models will be

$$\text{logit}[P(Y \leq j|x_p)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_p x_p, \quad j = 1, ..., 4. \tag{5.1}$$

$$\text{logit}[P(Y \leq j|x_p + k)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_p (x_p + k), \quad j = 1, ..., 4. \tag{5.2}$$

Subtracting (5.1) from (5.2), we have

$$\text{logit}[P(Y \leq j|x_p + k)] - \text{logit}[P(Y \leq j|x_p)] = k\beta_p.$$

$$= \log\left[\frac{P(Y \leq j|x_p + k)}{P(Y > j|x_p + k)}\right] - \log\left[\frac{P(Y \leq j|x_p)}{P(Y > j|x_p)}\right] = k\beta_p.$$

$$= \left[\frac{P(Y \leq j|x_p + k)/P(Y > j|x_p + k)}{P(Y \leq j|x_p)/P(Y > j|x_p)}\right] = e^{k\beta_p}$$

The above expression represents the *cumulative odds ratio* of belonging to category j or below corresponding to $X_p = x_p$ and $X_p = x_p + k$. The fact that the above expression holds true for any category $j = 1, 2, ..., 4$, is known as the *proportional odds* property. In fact, cumulative logit models are also referred to as *proportional odds model* because of this characteristic. In the special case of $k = 1$, this implies that for 1 unit increase in x_p, the cumulative odds of belonging to any particular category j or lower increases in multiples of e^{β_p} controlling for the remaining predictors.

5.3.4 Educational Achievement

We will now illustrate the applicability of cumulative logit model in the context of the educational achievement data. As mentioned before, this data consists of observations collected on various attributes of nearly 1000 undergraduate students enrolled in the business school of a major Australian university. The research question relates to identifying student-specific attributes that significantly affects in-class educational achievement. Predictors those are considered for this study are as follows:

- Gender of instructor (1: male, 0: female).
- Educational achievement of instructor (1: Phd, 0: non-Phd).
- Student age.

- Number of tutorials missed during a semester.
- Entry score.
- Background degree (0: BCom, 1: Non-BCom)
- Whether student is indigenous or not (0: non-indigenous, 1: indigenous).
- Whether student is Australian born or not (1: Australia born, 0: otherwise).
- Whether student is of Chinese origin (1: Chinese, 0: otherwise).

Before fitting the model to this data, it is important to note that none of the students whose observations were collected obtained a "High distinction".

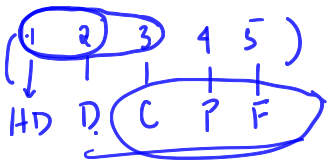
The R code for fitting the cumulative logit model to the above data is given below:

```
fit<-polr(Result~Tutor.Gender+...+ Chinese, data=learning, Hess=TRUE)
summary(fit)
ctable <- coef(summary(fit))
p <- round(pnorm(abs(ctable[, "t value"]), lower.tail = FALSE)*2,digits=3)
ctable <- cbind(ctable, "p value" = p)
ci <- confint.default(fit)
exp(cbind(OR = coef(fit), ci))
```

The corresponding output is depicted in the following table:

Variables	Estimate	Standard error	z-value	p-value	OR	2.5%	97.5%
Tutor.Gender	-0.25	0.135	-1.86	0.063	0.778	0.596	1.014
Tutor.Phd	0.60	0.148	4.045	0.00			
Student.Age	-0.187	0.0574	-3.25	0.001			
Tutorial.missed	-0.647	0.060	-10.74	0.00	0.523	0.465	0.589
Entry.score	0.114	0.020	5.719	0.00	1.121	1.078	1.166
Indigenous	-2.219	1.368	-1.622	0.105	0.109	0.007	1.587
Degree.name	0.489	0.167	2.988	0.003	1.631	1.183	2.248
Austr.born	0.706	0.170	4.16	0.00	2.026	1.452	2.825
Chinese	-0.967	0.198	-4.88	0.00	0.380	0.258	0.560
Fail—Pass	2.860	2.367	1.207	0.227			
Pass—Credit	6.55	2.36	2.77	0.006			
Credit—Distinction	9.454	2.378	3.97	0.000			

Based on the above output it is clear that all the predictors except Indigenous are significant at $\alpha = .1$. In fact, all predictors except Indigenous and Tutor Gender are significant even at $\alpha = .01$.



Hence, each of these predictors have a significant influence on the likelihood of a student receiving a certain higher grade (say. Distinction, Credit or Pass) relative to the corresponding lower grades.

5.3.5 Parameter Interpretation

We will now briefly discuss the interpretation of parameter estimates in light of the discussion in Sec 4.3.3. Specifically, we will do so for the `Tutor.Phd`, `Student Age` and `Tutorial.missed` variables. We have specifically chosen these three predictors since, taken together, they encompass all the three common types of variables encountered in real life applications, namely binary, continuous and discrete.

1. **Tutor.Phd**: For students who were taught by a tutor having a PhD, the cumulative odds of obtaining a higher grade (say, Distinction, Credit or Pass) rather than Failing is $e^{.60} = 1.82$ times that of students who were taught by a tutor not having PhD, controlling for the other variables. In other words, tutors having PhD are more effective in helping students achieve a higher grade compared to tutors not having PhD.

   The 95% confidence interval of the regression coefficient for Tutor.Phd is given by

   $$.60 \pm 1.96 \times .148 = (.31, .89)$$

   Thus the 95% interval of the cumulative odds ratio will be

   $$(e^{.31}, e^{.89}) = (1.36, 2.43)$$

   Since the above interval does not contain 1, `Tutor.Phd` has a significant impact on student achievement. In fact, the cumulative odds that students taught by tutors having PhD will obtain better grades is at least 1.36 times and at most 2.43 times the corresponding odds for students who are taught by tutors without a PhD, controlling for other relevant variables.

   Policy impact: The aforementioned finding can induce university authorities to have a relook at their hiring policies and only consider recruiting tutors who have a Phd.

2. **Student Age**: For every one year increase in student age, the cumulative odds of obtaining a higher grade (say, Distinction, Credit or Pass) rather than Failing changes multiplicatively by a factor of $e^{-.187} = .83$ controlling for the other variables. In other words, for every one year decrease in student age, the odds of obtaining a higher grade (Distinction, Credit or Pass) rather than Failing increases multiplicatively by a factor of $1/.83 = 1.20$ (i.e by 20%) controlling for the other variables. In a nutshell, younger students are significantly more likely to obtain a higher grade (Distinction, Credit or Pass) rather than Failing at all significance levels.

The 95% confidence interval of the regression coefficient of Student Age is given by

$$-1.87 \pm 1.96 \times .057 = (-.30, -.07)$$

Thus the 95% interval of the cumulative odds ratio will be

$$(.741, .929)$$

Since the above interval does not contain 1, `Student Age` has a significant impact on student achievement. In fact, the cumulative odds that students who are younger by a year will obtain better grade is at least $1/.929 = 1.076$ times and at most $1/.741 = 1.35$ the corresponding odds for students who are older by a year, controlling for other relevant variables.

3. **Tutorial missed**: For every additional tutorial that is missed, the cumulative odds of obtaining a higher grade (say, Distinction, Credit or Pass) rather than Failing changes multiplicatively by a factor of $e^{-.697} = .53$ controlling for the other variables. In other words, for every additional tutorial that is not-missed (i.e attended), the cumulative odds of obtaining a higher grade (Distinction, Credit or Pass) rather than Failing increases multiplicatively by a factor of (i.e by 92%), controlling for the other variables. In a nutshell, for every additional tutorial attended, the likelihood of obtaining a higher grade (Distinction, Credit or Pass) rather than Failing significantly improves (nearly doubles).

   The 95% confidence interval of the regression coefficient of Tutorial missed is given by

$$(-.76, -.53)$$

   Thus the 95% interval of the cumulative odds ratio will be

$$(.465, .589)$$

   Since the above interval does not contain 1, `Tutorial missed` has a significant impact on student achievement. In fact, for every additional tutorial attended, the cumulative odds of obtaining a better grade is at least $1/.589 = 1.7$ times and at most $1/.465 = 2.15$ times the corresponding odds for students who fail to attend that tutorial, controlling for the other variables.

Interpretation for the effect of the remaining variables can be carried out in a similar fashion.

5.4 Baseline Category Logit Model

5.4.1 Motivation

Although ordered categories are more prevalent in the Management literature, there can instances where a categorical variable has no apparent ordering. Some examples are given below:

Eg 1. [**Placement@IIMA**]: Suppose a graduating student in the PGP program of IIM Ahmedabad can be placed in any one of the following cohorts [Banking, Consulting, FMCG, IT, NGO]. The final placement may depend on various attributes such as work experience, grade point, gender, extracurriculars, age, previous institute etc. It may be of interest to determine which of the above attributes have a significant influence on the final placement.

Eg 2. [**Spirituality**]: It may be of interest to analyze whether gender and race affects our belief about life after death, categorized as [Yes, Undecided, No].

Eg 3. [**Politics**]: Whether a Bengali will vote for the Trinamool Congress in the upcoming assembly election of 2021 [Yes, No, Undecided] can depend on various factors like age, gender, income, place of residence, employment status etc.

In the above examples, the categorization of the response variables, namely
, and are do not follow any particular ordering i.e are nominal in nature. In order to quantify the impact of various predictors on the response, **Baseline Category Logit** models are used. In the following section, we give a brief overview of these models in the context of the "Belief in Afterlife" example.

5.4.2 Structure

As the term suggests, Baseline Category Logit models treats a prespecified category as the baseline category and pairs each of the other categories with respect to it. For c categories, the baseline category logits are given by

$$\log\left(\frac{\pi_j}{\pi_c}\right), \qquad j = 1, 2, ..., c-1.$$

The above expression can be viewed as the log odds of a response belonging to category j provided that there are only two categories possible, namely category j and category c. The baseline category logit model with p predictors is given by

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + ... + \beta_{jp}x_p, \qquad j = 1, 2, ..., c-1.$$

Controlling for the other variables, the effect of a predictor, say x_k, is dependent on the particular category paired with the baseline. One nice aspect of these models is that it is straightforward to determine the estimated response probabilities for any given category. For instance, the estimated probability for category j is given by

$$\hat{\pi}_j = \frac{e^{\hat{\alpha_j}+\hat{\beta_{j1}}x_1+\hat{\beta_{j2}}x_2+...+\hat{\beta_{jp}}x_p}}{\sum_{h=1}^{c} e^{\hat{\alpha_h}+\hat{\beta_{h1}}x_1+\hat{\beta_{h2}}x_2+...+\hat{\beta_{hp}}x_p}}, \qquad j = 1, 2, ..., c-1.$$

where $(\hat{\alpha_j}, \hat{\beta_{j1}}, ..., \hat{\beta_{jp}})$ are the estimated model parameters obtained by fitting the above model to a particular dataset.

5.4.3 Belief in Afterlife

We now illustrate some of the features of baseline category logit models in the context of a real dataset related to a survey about Belief in Afterlife. The dataset is represented in the following contingency table:

		Belief in Afterlife		
Race	**Gender**	Yes	Undecided	No
White	Female	371	49	74
	Male	250	45	71
Black	Female	64	9	15
	Male	25	5	13

As mentioned before, the response variable, namely "Belief in Afterlife" is categorized into three distinct categories, [Yes, Undecided, No]. It is of interest to determine whether Race (1: white, 0: black) and Gender (1: male, 0: female) have a significant effect on the response.

Treating "No" as the baseline category, the baseline category logit model is given by

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \qquad j = 1, 2.$$

Here β^G and β^R are the coefficients corresponding to "Race" and "Gender" respectively. Specifically, β_1^G represents the log odds ratio between Gender and categories 1 and 2 (i.e "Yes" and "No") of Belief in Afterlife conditional on Race.

5.4.4 Results and Interpretation

The R code for fitting the model to the above dataset is given below:

```
Afterlife<-read.table(''http://www.stat.ufl.edu/~aa/cat/data/Afterlife.dat'',
header=TRUE)
fit<-vglm(cbind(yes,undecided,no)~ gender+race,family=multinomial,
data=Afterlife)
summary(fit)
```

The corresponding output is depicted in the following page

As far as interpretation of parameter estimates are concerned, let us consider "Gender:1". Since $\hat{\beta}_1^G$ = , the estimated odds of Males believing in afterlife (versus not believing) is times that of females controlling for . In other words, the estimated odds of Females believing in afterlife (versus not believing) is times that of males.

Variables	Estimate	Standard error	z-value	p-value	OR	2.5%	97.5%
Intercept:1	1.302	0.226	5.747	0			
Intercept:2	-0.653	0.340	-1.918	0.0551			
Gender:1	-0.419	0.171	-2.44	0.0145	0.658	0.470	0.919
Gender:2	-0.105	0.246	-0.426	0.67			
Race:1	0.342	0.237					
Race:2	0.271	0.354					

As far as the effect of Race is concerned, since $\hat{\beta}_1^R$ = ·342 , the estimated odds of Whites believing in afterlife (versus not believing) is $e^{·342} = 1·41$ times that for Blacks controlling for Gender . In other words, whites are 41% more likely to believe in afterlife compared to blacks.

$\hat{\beta}_2^R = ·271 \rightarrow e^{·271} = 1·31 \rightarrow$ Estimated odds of whites being Undecided vs not believing is 1·31 times " " " Blacks " "