Section A [50 marks]

A1. What is the most important statistical concept you have learned during the statistics part of the module? Explain this concept to someone without a background in statistics, and why you think it is the most important thing you learned.

Ans-There are various important statistical concepts which I have learnt during the statistics part of the module. But from the point of view that a person without a background in statistics can understand I’ll go with numerical description of data, specifically concept of mean. We intentionally or unintentionally use this concept a lot in our daily life without realization. Suppose we are spending the amount between $10 to $15 on the daily basis and we want to summaries our weekly spending but we have for eg:- $12,$14,$10,$15,$13,$11 and $13. We have seven numbers based on each day in a week but we cannot draw any relevant information by looking at all the numbers, so we need a concept that can be a representative on behalf of all the seven number and the most used and common concept we used in this situation is average i.e. sum of weekly expenditure/number of days. This concept easy to understand and used very frequently by all types of personnel’s. There are plenty of examples we can think of in our daily lives where we use this concept. Mean gives the idea about center of our data, and is meant to carry a piece of information from every member of the sample. It represents average value in a dataset.

Ref:--[mean](https://www.statology.org/importance-of-mean/#:~:text=The%20mean%20is%20important%20because,is%20skewed%20or%20contains%20outlies.)

A2. Create a simple data frame of your choice in R with at least 2 variables and 10 observations. Please include your R coding. The data can be real-world or generated data for illustration (but not data from labs). Use any R function from labs to calculate a statistical outcome of interest from the data frame.

R code-

# Create the data frame.

emp.data <- data.frame(

emp\_id = c (1:5),

emp\_name = c("Rick","Dan","Michelle","Ryan","Gary"),

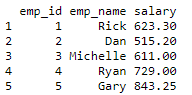
salary = c(623.3,515.2,611.0,729.0,843.25),

stringsAsFactors = FALSE

)

head(emp.data)

Output-



summary(emp.data$salary)

Output:-



A3. Crop rotation is the practice of planting dissimilar crops sequentially in one location to avoid the depletion of nutrients. A researcher conducted an experiment with soybeans: they used crop rotation on 15 plots and left another 15 with no rotation.

• The researcher attaches particular importance to the reliability of the harvest as well as the actual harvest. Set out two descriptive statistics that could be useful to capture reliability.

• Below are the numbers of pounds of soybeans harvested from each plot. Draw a boxplot to explain the data and explore any differences between the two plots. (Do not carry out formal tests or confidence intervals.)

Crop rotation: 37, 42, 44, 46, 41, 39, 29, 31, 37, 41, 42, 45, 38 , 42, 33

No rotation: 29, 35, 34, 27, 31, 36, 41, 40, 37, 29, 31, 36, 24, 23, 22

Ans-

Average and median can be two suitable statistic for drawing the useful information to draw inference about reliability of harvest.

R-code-

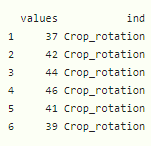
Crop\_rotation<-c(37, 42, 44, 46, 41, 39, 29, 31, 37, 41, 42, 45, 38 , 42, 33)

No\_rotation<-c(29, 35, 34, 27, 31, 36, 41, 40, 37, 29, 31, 36, 24, 23, 22)

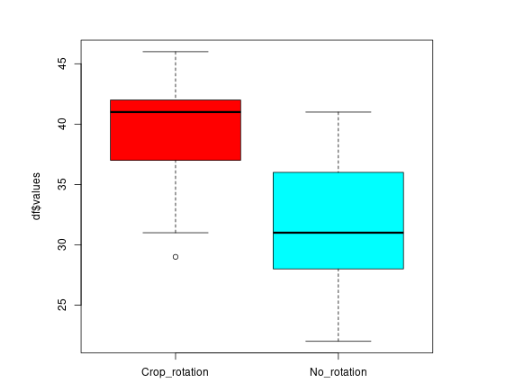
data<-data.frame(Crop\_rotation,No\_rotation)

df <- stack(data)

head(df)



boxplot(df$values~df$ind, col = rainbow(ncol(df)))



In above boxplot we observed that the range of numbers of pounds of soybeans crop rotation data is different from numbers of pounds of soybeans no rotation data. And Median number of pounds of soybeans is significantly different in both cases. Notice that the vertical line inside the box plot for crop rotation which represents the median is much closer to the third quartile than the first quartile, which means the distribution is left-skewed. And this boxplot also showing one outlier while on the other hand no rotation data is seems a bit of positively skewed but has no outlier. One more thing to notice is for no rotation data median is actually pretty close to the minimum of crop rotation data.

Ref- <https://www.statology.org/box-plot-skewness/>

A4. The normal distribution plays an important role in statistics. From our work this semester, explain two reasons why it is a popular distribution to work with. Can you think of a variable where the normal distribution might not be a good distribution to work with?

Ans-

Normal distribution plays a vital role in statistics and there are multiple reasons. Below are two Reasons-

1. many, many situations in the real world can be modelled by a normal distribution, or at least come very close to a normal distribution. In fact, it tends to be the “go-to” distribution, for most purposes. Some examples are the heights of a random population of people, an IQ distribution or the pattern of misses that a shooter makes around a bullseye.
2. If the sample is large enough this sampling distribution will often be approximately normal. This theorem is mainly called as Central limit theorem.

Ref;-

<https://statisticsbyjim.com/basics/normal-distribution/#:~:text=As%20with%20any%20probability%20distribution,values%20for%20many%20natural%20phenomena>.

Since Normal distribution is a symmetric distribution. It will not be suitable in case of skewed Variables such as number of events happening in a particular positive time period, which is example of Poisson distribution. And Poisson distribution is positively skewed.

Ref;- <https://www.statisticshowto.com/probability-and-statistics/non-normal-distributions/#:~:text=Insufficient%20Data%20can%20cause%20a,t%20get%20a%20normal%20distribution>.

A5. • A weather forecaster knows from historical data that the daily probability of rain is 0.3 during April. What is the probability that there are 20 dry daysin April? and at least three wet days in April.

<https://math.stackexchange.com/questions/1112657/what-is-the-probability-of-rain-over-a-number-of-days-given-the-probability-of>

#daily probability of rain is 0.3 during April and not raining is 0.7

#Since there are 30 days in April.

# What is the probability that there are 20 dry days in April?

n=30

x=20

p\_rain=0.3

p\_no\_rain=0.7

# We will use binomial distribution to calculate the probability.

p\_20\_dry\_days <- dbinom(x,n,p\_no\_rain)

p\_20\_dry\_days

#Output - 0.1415617

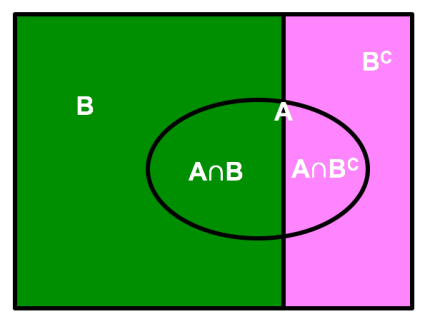
#What is the probability that there are at least three wet days in April?

p\_at\_least\_3\_wet\_days<-1-dbinom(0,n,p\_rain)-dbinom(1,n,p\_rain)-dbinom(2,n,p\_rain)

p\_at\_least\_3\_wet\_days

#Output - 0.9978868

• 70% of people attend their doctor regularly; 30% of those people have no health problems crop up during the following year. Out of the 30% of people who don’t see their doctor regularly, only 5% have no health issues during the following year. What is the probability that a random person will have no health problems in the following year?



So in this scenario B represents people who do see their doctor and Bc represents the people who do not. All the people who do not experience health problems are represented by A, and this is the probability we want to find.

Ref;-

<http://cbseacademic.nic.in/web_material/Manuals/appliedmaths/Chapter9_Probability.pdf>

<https://shotlefttodatascience.com/2018/08/05/law-of-total-probability-worked-examples/>

p(B)=0.7

p(Bc)=0.3

p(B|A)=0.3

P(Bc |A)=0.05

P(A)=P(B)\*P(B|A)+P(Bc)\*P(Bc|A)

=0.7\*0.3+0.3\*0.05

=0.225

A6. • Briefly explain with an example of your choice each of the following statistical terms: a population, a sample, population parameter, a sample statistic

Population data is a whole and complete set. The sample is a subset of the population that is derived using sampling.

A **parameter** is a characteristic of a population. A **statistic** is a characteristic of a sample. Inferential statistics enables you to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.

For example, say you want to know the mean income of the subscribers to a particular magazine—a parameter of a population. You draw a random sample of 100 subscribers and determine that their mean income is $27,500 (a statistic). You conclude that the populations mean income μ is likely to be close to $27,500 as well. This example is one of statistical inference.

Ref-<https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics>

• We are interested in independent and identically distributed (iid) random variables when sampling. What do you understand by an iid sample and can you think of a sample not meeting this criterion (other than the family IQ example in lectures)?

 definition of an IID statistics is that random variables X1, X2, . . . , Xn are IID if they share the same [probability distribution](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/) and are [independent events](https://www.statisticshowto.com/probability-and-statistics/dependent-events-independent/#or). Sharing the same probability distribution means that if you plotted all of the [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)together, they would resemble some kind of distribution.

 imagine the sequence x where each element x\_i is either one higher or one lower than the preceding element, with a 50-50 chance as to which of these happens. Then one possible sequence is 1,2,3,2,3,4,3,2. It should be clear that there are some permutations of this sequence that are not equiprobable: in particular, sequences starting 1,4,... have probability zero. You can instead consider pairs of the form x\_i | x\_i-1 to be iid if you wish.

Ref-

<https://www.statisticshowto.com/iid-statistics/>

A7. A clinical researcher knows from experience that average body temperature for a particular group they are studying is approximately normally distributed with μ = 98.2 degrees Fahrenheit and population standard deviation σ = 0.73 deg Fahrenheit. Use R to simulate 30 values from this distribution (call it “sample1”) and then simulate 1,000 values from this distribution (call it “sample2”) assuming normality. Plot histograms and derive the summary statistics for sample1 and sample2. Can you explain your findings? Why is random sampling important in statistics?

mu=98.2

sigma=0.73

sample1<-rnorm(30,mu,sigma)

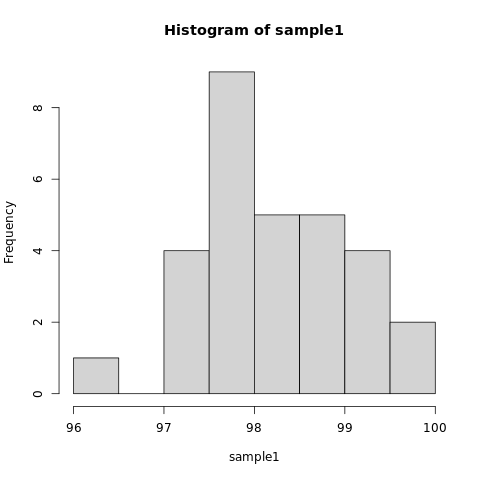
summary(sample1)

hist(sample1)

output-

Min. 1st Qu. Median Mean 3rd Qu. Max.

96.42 97.64 98.10 98.21 98.69 99.85



sample2<-rnorm(1000,mu,sigma)

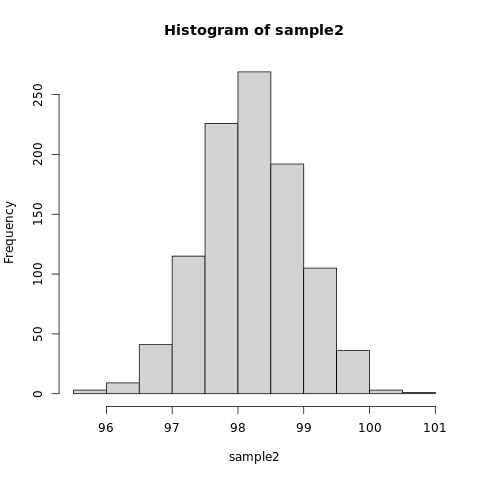
summary(sample2)

Output

Min. 1st Qu. Median Mean 3rd Qu. Max.

95.87 97.74 98.21 98.20 98.71 100.76

hist(sample2)



Means from both samples are approximately same. But the histogram for sample 2 looks more normal than histogram of sample1 because size of sample is much larger in sample2.

Randomly selecting the members of a sample is important because it helps prevent bias in your results.

A8.

• The 95% Confidence Interval for an unknown population mean is calculated using a sample size of 64 and is presented as (273, 663). Determine the sample mean and the population standard deviation.

To calculate the sample mean and population standard deviation form 95% ciwe will use the below formula to calculate the population standard deviation first and then using this we will calculate mean.

https://handbook-5-1.cochrane.org/chapter_7/image014.gif

Hence population standard deviation is 795.92.Now using below formula we will find out sample mean.

x̄ ± z\* σ / (√n),

z\* σ / (√n)=1.96\*795.92/8=195

now we can use any of upper of lower boundry value from confidence interval to calculate sample mean and sample mean will be 468.

• Two students, A and B, from different countries take a similar psychometric test. The tests have a different marking scheme and past experience shows that the marks are approximately normally distributed. Below are the student marks, country mean and country standard deviation of marks. Which student did best in their country and why?

Z-score for student in country A is 0.214 where as Z-score for student in country B is 0.3. Since z-score is greater for student in country B so student in country B did best in their country because their score is relatively far from their country mean compare to student in country A.

Ref:- <https://www.statology.org/comparing-z-scores/>

The psychometric test in country B is also used as part of a company’s graduate recruitment process. The company wants to set a cut off mark so only the top 8% of applicants pass the test. What should the mark be set at? [

P(z>-zB)=0.08

P(z>0.3)=0.08

(x-80)/10=1.4051

X=94.051

Ref:-

<https://statistics.laerd.com/statistical-guides/standard-score-3.php>

Section B.

B1. This question is based on the UN’s Human Development Index (HDI) statistical dataset (for more background see http://hdr.undp.org/en/content/ human- development-index-hdi). The file is called hdi.xlsx and is on Blackboard. life is lifeexpectancy in years, school is average schooling in years and income is $ nationalincome per capita for 189 countries. You need to first import the data into R.

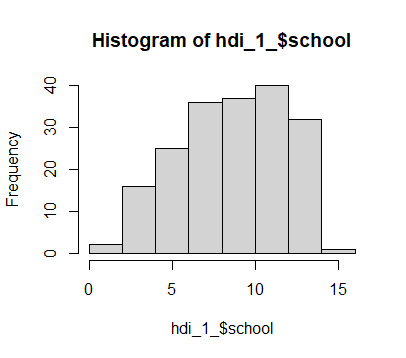
We first imported the data in R using readxl library using below commands.

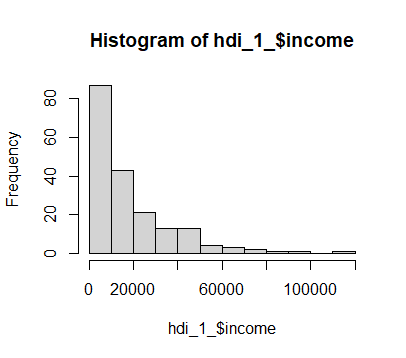
library(readxl)

hdi\_1\_ <- read\_excel("C:/Users/lenovo/Downloads/hdi (1).xlsx")

• Plot histograms of life expectancy, schooling and income. Comment on their shapes.







Histogram of life expectancy and schooling negatively skewed in shape where as the histogram of income is positively skewed.

• Are there any “outliers” in any of the three variables? Does this surprise you?

To check the outliers we plotted the boxplot of all three using the below code in R.

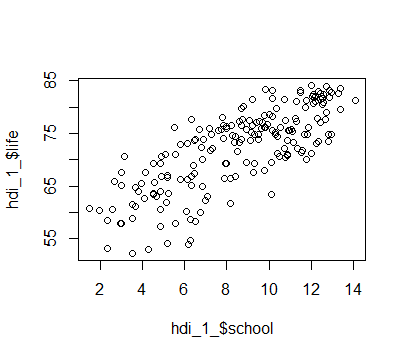
and for income there are some outliers present but that does not surprise me since there are many people which earn in a certain range but there are less people which have comparatively high income than the average.

• Explain the 75th percentile of schooling and 15th percentile of life expectancy to someone with no background in statistics.

75 % percentile of schooling means that the person has better schooling than 75 percent of the other people in the data.

Same goes with 15% percentile of life expectancy which means that the person has higher life expectancy than 15% people in data.

• A researcher is interested in whether there is a relationship between life expectancy and average levels of schooling. Use one diagram to explore a possible relationship and comment on this.



Above is the scatter plot for life expectancy and average levels of schooling and it shows a positive linear trend which means the people having higher level of education have higher life expectancy.

• Can you conclude that changes in schooling cause changes in life expectancy?

Yes from the above scatter plot we can conclude because higher level of schooling brings awareness and opportunity of earning high and affording a healthy lifestyle.

#R code

library(readxl)

hdi\_1\_ <- read\_excel("C:/Users/lenovo/Downloads/hdi (1).xlsx")

hist(hdi\_1\_$life)

hist(hdi\_1\_$school)

hist(hdi\_1\_$income)

boxplot(hdi\_1\_$life)

boxplot(hdi\_1\_$school)

boxplot(hdi\_1\_$income)

quantile(hdi\_1\_$school,probs = 0.75)

quantile(hdi\_1\_$life,probs=0.15)

plot(hdi\_1\_$life~hdi\_1\_$school)

B2.

• A lecturer wants to take a sample of 40 students from a class of 160 to complete a survey on the module. In the past they have tried two ways: asking the first 40 students who turn up to lectures or emailing a survey to the entire class. The lecturer’s friend, a statistics lecturer, suggests asking each student in lecture to toss a coin twice and select the students who get two heads. Which approach would you prefer and why, or does it actually matter?

I will prefer the approach suggested by lecturer’s friend and statistics lecturer because a sample is representative of whole population and It should be chosen randomly and in a way that it should capture most of characteristics of the population. So there should not be any bias in choosing sample. In Past two ways bias can come but in the way suggested by the statistics lecturer is a random way and will exclude bias and sample collected from this way will be purely random sample and will represent the population in better way.

• 10% of your emails are spam emails. Your spam filter catches spam 90% of the time. Your spam filter misidentifies non-spam as spam 3% of the time. How might you use probability to understand what percent of the emails sitting in your spam folder are genuinely spam emails? (Hint: Bayes’ Law might be useful; also think about false positives).

Let A denote the event that an email is detected as spam and B denote the event that an email is spam.

Given that 10% of the emails are spam, i.e., P(B)=0.1

Thus P(B′)=1−P(B)=0.9

Spam filter can detect 90% of spam emails.

That is P(A|B)=0.90.

And the probability for a false positive (a non-spam email detected as spam) is 3%.

That is P(A|B′)=0.03

We need to find the probability that the email is spam given that it is detected as spam.

Using Bayes' Theorem, required probability is

P(B|A)= P(A|B)P(B)

P(A)

= P(A|B)P(B)

P(A|B)P(B)+P(A|B′)P(B′)

= 0.769

77% percent of the emails approximately sitting in spam folder are genuinely spam emails.

Ref:-

<https://novelanswer.com/it-is-estimated-that-50-of-emails-are-spam-emails-some-software-has-been-applied-to-filter-these-spam-emails-before-they-reach-your-inbox/>

• Use either the letters in your surname, e.g. ZACK to explain the difference between “permutations” and “combinations.” Give a real-world example where the distinction between permutations and combinations is crucial.

A permutation is a method of arranging all the members in order. The combination is selection of elements from a collection.

So in surname Zack the order matters if the order changes then it won’t be Zack so If we find the a method of arranging letters Z,A,C,K in order to make the surname of 4 letters using these 4 letters as Zack. That will be permutation.

And to make the surname of 4 letters using these 4 letters where order does not matter will be combination.

combination locks: they are actually ‘permutation’ locks because it matters what order the numbers are in; that’s the whole point of them. and the distinction is crucial another example is dna sequences.

B3.

A researcher is interested in the average age people first develop symptoms of a particular condition they are investigating. They record the ages of first developing symptoms from 30 random people (below) who have been involved in a recent largetrial on managing the condition.

25, 19, 60, 76, 52, 42, 56, 99, 61, 42, 70, 23, 23, 19, 18, 54, 72, 18, 42, 33, 42, 19, 41, 24, 67, 70, 23, 21, 30, 22

• Calculate 95% and 99% confidence intervals for the average age of firstdeveloping symptoms. Carefully explain the two confidence intervals and why they differ.

We used R to calculate the confidence intervals using the given R code.95% confidence interval is (34.20271, 49.99729) and 99% confidence interval is (31.7212, 52.4788).955 confidence interval means that we are 95% confident that population mean lies in this interval and 99% confidence interval means we are 99% confident that population mean contains in this interval.

#R code-

age<-c(25, 19, 60, 76, 52, 42, 56, 99, 61, 42, 70, 23, 23, 19, 18, 54, 72,

18, 42, 33, 42, 19, 41, 24, 67, 70, 23, 21, 30, 22)

xbar<-mean(age)

n<-length(age)

stdev<-sd(age)

## 95 percent confidence interval so tails are .925

error <- qnorm(0.975)\*stdev/sqrt(n)

lower\_bound <- xbar - error

lower\_bound

upper\_bound <- xbar + error

upper\_bound

error1<- qnorm(0.995)\*stdev/sqrt(n)

lower\_bound1 <- xbar - error1

lower\_bound1

upper\_bound1<- xbar + error1

upper\_bound1

Ref:-

<http://www.stat.ucla.edu/~rgould/110as02/bsci>

• Previous well-established research shows that the average age of first developing symptoms was around 30. Are your findings consistent with this and why?

From our findings it shows that our 95% and 99% confidence intervals do not indicate this since 30 does not belong to either of confidence intervals.

• Do you run the risk of being wrong in ii) and why?

Yes Since the small size is small we cannot fully support our findings.

• A colleague has asked you for a smaller confidence interval from those in i). How might this be achieved?

By reducing the confidence there will be increase in 1- (confidence interval value /100).

Suppose 80% CI. This means alpha = .20 We can get z(alpha/2) = z(0.10)

from R: > qnorm(.10) or qnorm(.90)

and put all the values in below formula.

Xbar +- z(alpha/2) \* sigma/sqrt(n)