MSIT 431: PROBABILITY AND STATISTICAL METHODS

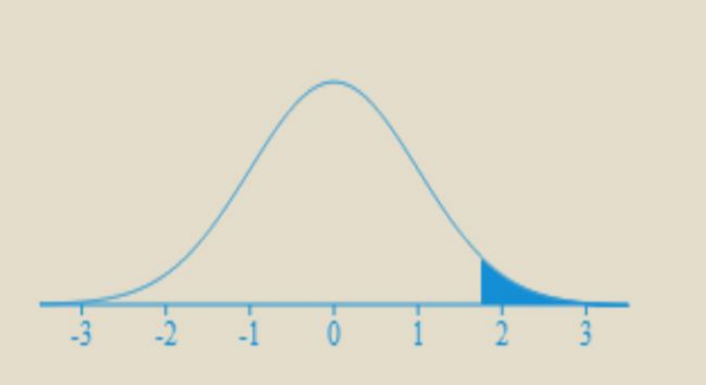
Submitted By : Ruma Anand
Student ID : 3074142

Ques 1. Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

Answer 1.

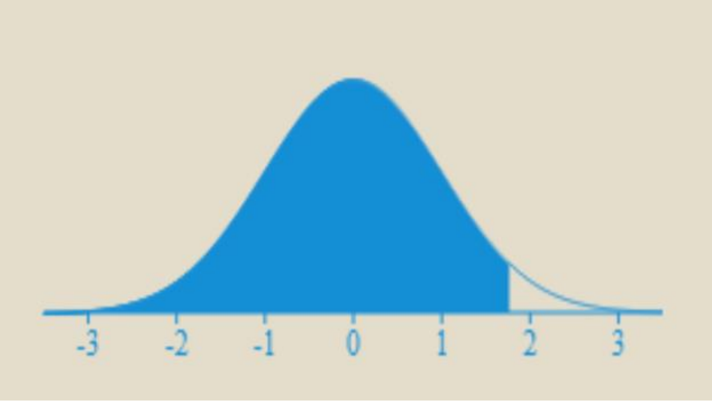
(a) Z>1.75

From Table A, standard normal probability for Z=1.75 is .9599
Proportion of observation from a standard normal distribution=1-0.9599=0.0401 or 4%



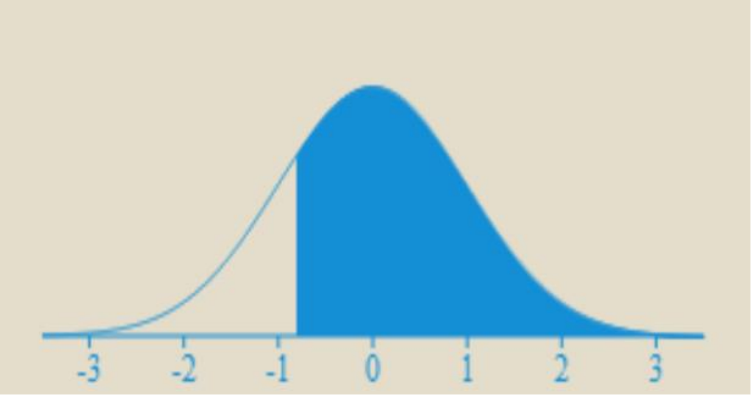
(b) Z<1.75

From Table A, standard normal probability for Z=1.75 is .9599
Proportion of observation from a standard normal distribution 0.9599 or 96%



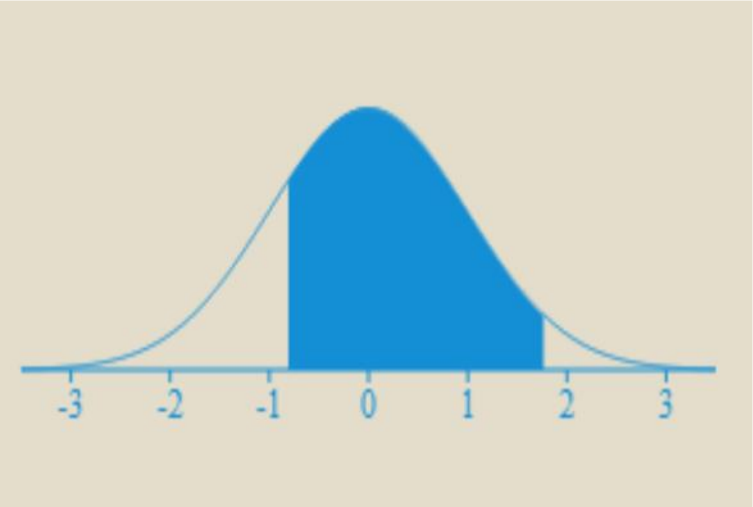
(c) Z>-0.80

From Table A, standard normal probability for Z=0.80 is .2119
Proportion of observation from a standard normal distribution 1-.2119 =.7881



(d) -0.80<Z<1.75

From Table A, standard normal probability for Z=1.75 is .9599 and for Z=0.80 is .2119
Proportion of observation from a standard normal distribution .9599-.2119=.748 or 74%



Ques 2. The Wechsler Adult Intelligence Scale (WAIS) is the most common IQ test. The scale of scores is set separately for each age group. And the scores are approximately Normal with mean 100 and standard deviation 15. People with WAIS scores below 70 are considered developmentally disabled when, for example, applying for social security disability benefits. What percent of adults are developmentally disabled by this criterion?

Answer2.

Mean =100

Standard deviation = 15

Z=(X-μ)/σ

Z=(70-100)/15 = -2

According to Table A standard normal probability for Z= -2 is 0.0228. Hence, 2.3% of adults are developmentally disabled.

Ques 3. Textbook Exercise 2.5 High click counts on Twitter. A study was done to identify variables that might produce high click counts on Twitter. You and nine of your friends collect data on all of your tweets for a week, You record the number of click counts, the time of day, the day of week, the sex if the person posting the tweet, and the length of the tweet.

(a) What are the cases for this study?

Number of tweets are the cases for the study.

(b) Classify each of the variables as categorical or quantitative.
Number of click counts : Quantitative
Time of day : Categorical/Quantitative
The day of week : Categorical
Sex of the person posting the tweet : Categorical
Length of the tweet : Quantitative

(c) **Classify each of the variables as explanatory, response, or neither. Explain your answers**

A response variable measures an outcome of a study. An explanatory variable explains or causes changes in the response variable

Number of click counts : *Response variable* since it measures an outcome of a study.

Time of day : *Potentially explanatory* as it can be used to explain changes in response variable.

The day of week : *Explanatory* - Number of click counts may vary based on the day of week selected.

Sex of the person posting the tweet : *Explanatory.* Click count can be summarized by gender of the person posting the tweet.

Length of the tweet : *Explanatory.* Length of the tweet could explain why certain long tweets don't get as many clicks as shorter ones do.

Ques 4. Text book exercise 2.34 Internet use and babies

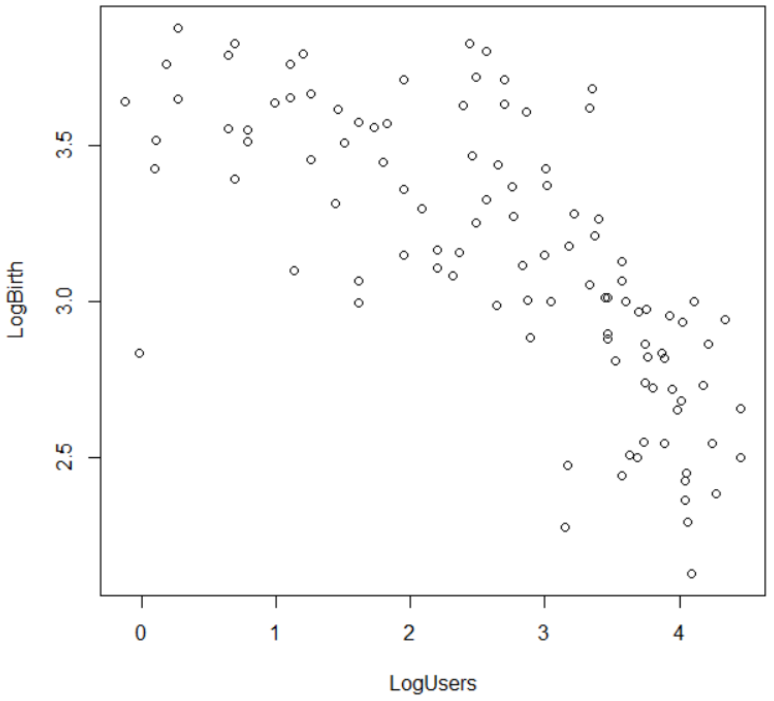
(a) Describe the relationship between two variables.
Relationship has non-linear form and negative curved relationship

(b) A friend looks at the graph and concludes that using the internet will decrease the number of babies born.Explain why the association seen in the scatterplot does not provide a reason to draw this conclusion.
Internet may not always affect birthrate. Although they have a negative relationship, there is no guarantee that one should be the cause for the other. Also, there are multiple value of y for a single value of x. For eg. X=20 has 4 different values within the range (10-40).

Ques 5. Text book exercise 2.59

(a) Make a plot of the data similar to Figure 2.13 and report the correlation :

```
plot(LogUsers,LogBirth)
```



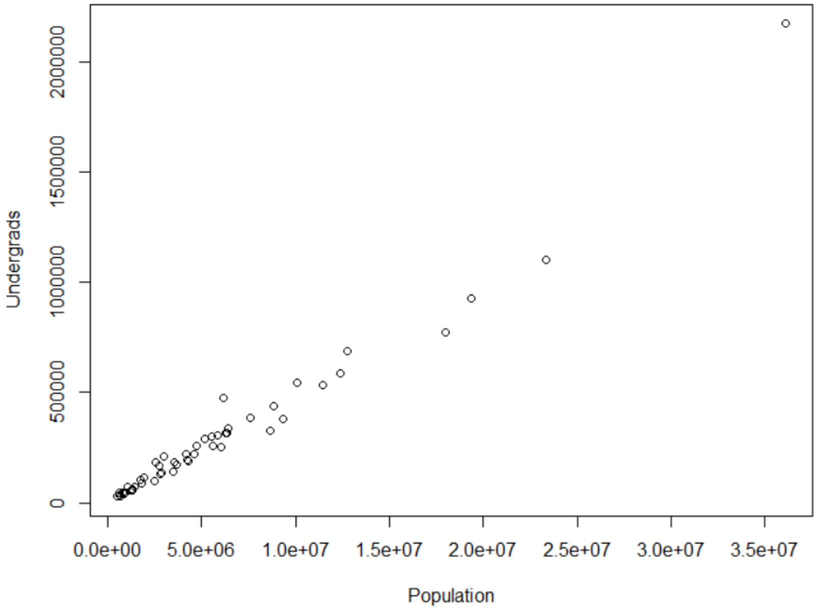
```
> cor(LogUsers,LogBirth,method = c("pearson"))

Correlation = -0.7213689
```

(b) Is the correlation a good numerical summary for this relationship? Explain your answer.

Since the plot of data is a curved non linear form, correlation is not a good numerical summary for this relationship because it does not describe curved relationships between variables, no matter how strong the relationship is.

Ques 6. Text Book Exercise 2.75

a) Make a scatterplot with population on the x axis and number of undergraduates on the y axis



(b) Describe the form, direction and strength of the relationship. Are there any outliers?

The relationship is linear, positive and strong. There are several outliers.

(c) Textbook exercise 2.75. Find least square regression line.

Let undergrads be Y

Let population be X.

Mean for undergrads :302,136

Standard deviation for undergrads : 358,460

Mean for population : 5,955,551

Standard deviation for population : 6,620,733

Correlation between undergrads and population is 0.98367

Least squares regression line :

Y=b0+b1X

b1= (r*Standard deviation for undergrads)/Standard deviation for population

b1=.05326

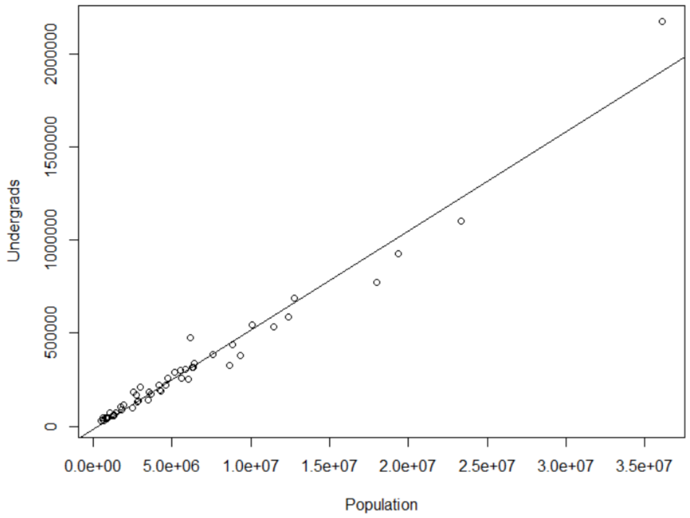
b0=Mean for undergrads – (b0*mean for population)

b0 = -15057

y= -15057+0.05326x

(d) add regression line to the plot

```
> mydata275=read.csv("mydata2_75.csv")
> mydata275
```



Ques 7. Textbook Exercise 2.76

Let undergrads be Y

Let population be X.

Mean for undergrads : 220,134

Standard deviation for undergrads : 165,270

Mean for population : 4,367,448

Standard deviation for population : 3,310,957

Correlation between undergrads and population is 0.97081

Least squares regression line :

Y=b0+b1X

b1=(r*Standard deviation for undergrads)/Standard deviation for population

b1=.0485

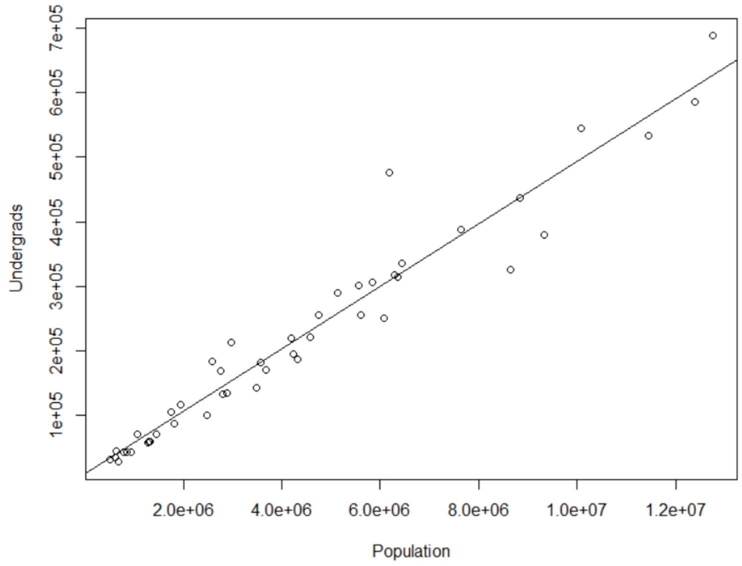
b0=Mean for undergrads – (b0*mean for population)

b0 = 8312.772

y= 8312.77+0.0485x

(d) add regression line to the pot

```
> mydata276=read.csv("mydata2_76.csv")
> mydata276
> attach(mydata276)
> plot(Population,Undergrads)
> abline(8312.77,0.0485)
```



Ques 8. Textbook Exercise 2.77

Answer .

(a) From 2.75, y= -15057+0.05326x

X=4,000,000. So y=-15057+0.05326*4000000

X=197983

(b) From 2.76, y= 8312.77+0.0485x

X=4,000,000. So y=8312.77+0.0485*4000000

X=202312.77

(c) After comparing the data with and without outliers, the predicted number of undergraduate students are near about similar. The presence of outliers in the prediction equation does not affect the predicted number of undergraduate students.

Ques 9. Textbook Exercise 2.128

Answer.

(a) Death rate for hospital A patients with 'poor' condition before surgery : 57/1500=0.038
Death rate for hospital B patients with 'poor' condition before surgery : 8/200=0.04

(b) Death rate for hospital A patients with 'good' condition before surgery : 6/600=0.01
Death rate for hospital B patients with 'good' condition before surgery : 8/600=0.013

(c) Recommendation to someone facing surgery and choosing between these two hospitals would be to choose hospital A as Hospital A has lower death rate for both 'good' and 'poor' patients.

(d) Hospital A does better in both groups 'good' and 'poor' yet does worse overall because the overall death rate of hospital A and B is misleading when we discover the values for individual death rate for the same data . This is a classic case of Simpson's paradox.
The number of people with 'poor' condition who went for hospital A is a lot more than the number of people who were in good condition and chose hospital A. Although fewer people went for hospital B in poor condition, the number of deaths were the same when compared for good condition. If the lurking variable here 'Condition of patient before surgery(good/poor) is not considered, the overall summary incorrectly depicts Hospital B had a lower death rate. However, when we split the data around the condition and find individual death rate of each condition, it helps to determine the correct death rate for each hospital.