IS507 - Data, Statistical Models and Information - Assignment 1

1)

3.12 School absences. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.[24]

(a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

(b) What is the probability that a student chosen at random misses no more than one day?

(c) What is the probability that a student chosen at random misses at least one day?

(d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

(e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

(f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

Solution:

a)

Let A be the event when student misses 1 day of school, P(A) = 0.25
Let B be the event when student misses 2 days of school, P(B) = 0.15
Let C be the event when student misses 3 or more days of school, P(C) = 0.28

Since sum of all probabilities is 1,
Probability that student does not miss and day at school = 1 - [P(A) + P(B) + P(C)]
= 1 - [0.25 + 0.15 + 0.28]
= 1 - 0.68 = 0.32

Probability a student does not miss any school day is 0.32

b)

Student does not miss more than 1 day = 1 - Student misses more than 1 day
= 1 - [P(B) + P(C)]
= 1 - [0.15 + 0.28]
= 1 - 0.43 = 0.57

Probability a student misses no more than one school day is 0.57

c)

Student does misses atleast 1 day = Student misses 1 day, 2 days and 3 days
= [P(A) + P(B) + P(C)]
= 0.25 + 0.15 + 0.28
= 0.68

Probability a student misses atleast one school day is 0.68

d)

The first kid misses school is independent of the second school missing school
So, both students miss school = P(missing) * P(missing)
= 0.32 * 0.32
= 0.1024

Probability both kids miss school is 0.1024

e)
The first kid misses school is independent of the second school missing school
So, both students miss school = P(missing atleast 1 day) * P(missing atleast 1 day)
= 0.68 * 0.68
= 0.4624
Probability both kids miss school atleast 1 day is 0.4624

f) The assumption that the kids missing school are independent may not be exactly true because it has been said the parent has 2 kids and so if 1 kid is sick, there is high chance that the other kid might also be sick as sickness may spread.

---

2) Please use this sample space of rolling 2 dice from the lecture slides. Find the following probabilities for rolling two dice (Show your work, do not just provide an answer):

a) The sum of the dice is not 3
b) The sum is at least 8.
c) The sum is no more than 6.

Solution:

A single throw of a die can have 6 possible outcomes from 1 to 6.
When 2 dice are rolled since each has 6 outcomes, the total number of outcomes becomes $6^2 = 36$

Sample space is as follows:

	Die 2						
Die 1		**1**	**2**	**3**	**4**	**5**	**6**
	1	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

a)
Sum of the die is 3 is possible only in 2 cases — (1,2) or (2,1). Let this be event A.
Sum of the die is not 3 will be in all other cases. Let this be event B.

Since the sum of all probabilities is 1, so
P(A) + P(B) = 1
2/36 + P(B) = 1

P(B) = 1 - (2/36)
P(B) = 34/36 = 17/18
Therefore, probability of sum of dice is not 3 is 17/18 or 0.944.

b)
Sum is atleast 8 - means sum can be 8,9,10,11,12
Sum is 8: (2,6), (3,5), (4,4), (5,3), (6,2)
Sum is 9: (3,6), (4,5), (5,4), (6,3)
Sum is 10: (4,6), (5,5), (6,4)
Sum is 11: (5,6), (6,5)
Sum is 12: (6,6)

Total cases = 5+4+3+2+1 = 15
Therefore, probability of sum is at least 8 is 15/36 or 0.417

c)
Sum is no more than 6 - means sum can be 1,2,3,4,5,6
Sum is 1 - No case
Sum is 2: (1,1)
Sum is 3: (1,2), (2,1)
Sum is 4: (1,3), (2,2), (3,1)
Sum is 5: (1,4), (2,3), (3,2), (4,1)
Sum is 6: (1,5), (2,4), (3,3), (4,2), (5,1)

Total cases = 0+1+2+3+4+5 = 15
Therefore, probability of sum is at least 8 is 15/36 or 0.417

---

3)

3.17 Burger preferences. A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.[41]

		Gender		
		Male	Female	Total
Best hamburger place	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
	Fat Burger	10	12	22
	Tommy's Hamburgers	27	27	54
	Umami Burger	5	1	6
	Other	26	20	46
	Not Sure	13	5	18
	Total	248	252	500

(a) Are being female and liking Five Guys Burgers mutually exclusive?
(b) What is the probability that a randomly chosen male likes In-N-Out the best?
(c) What is the probability that a randomly chosen female likes In-N-Out the best?
(d) What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.
(e) What is the probability that a randomly chosen person likes Umami best or that person is female?

Solution:

a)
The event of being female and liking Five Guys is not mutually exclusive because there are 6 females which like Five Guys burgers.

b)
Let A be the event where a random male likes In-N-Out
There are 162 males which fall in the event A
The total outcomes are 248 as there are 248 males.
P(A) = 162/248 = 0.6532
Probability of a random male liking In-N-Out is 0.653

c)
Let B be the event where a random female likes In-N-Out
There are 181 females which fall in the event B
The total outcomes are 252 as there are 252 females.
P(B) = 181/252 = 0.7183
Probability of a random female liking In-N-Out is 0.718

d)
We assume that a man liking In-N-Out is independent of a woman liking In-N-Out.
It seems reasonable to assume this fact because a survey of 500 people was carried out and it has not Benn mentioned that people knew each other. So we can assume that they are independent events.

Probability a man liking In-N-Out is 0.653
Probability a woman liking In-N-Out is 0.718
Probability both man and woman dating like In-N-Out = P(A) * P(B)
= 0.653 * 0.718
= 0.469
Probability both man and woman dating like In-N-Out is 0.469

e)
Let C be the event a person likes Umami.
Let D be the event a person is female.
P(C) = (5+1)/500 = 6/500
P(D) = 252/500
There is a female which likes Umami and occurs twice in events C and D.
P(C and D) = 1/500

P(C or D) = P(C) + P(D) - P(C and D)
= 6/500 + 252/500 - 1/500
= 257/500 = 0.514
Probability a person likes Umami or is female is 0.514

4)

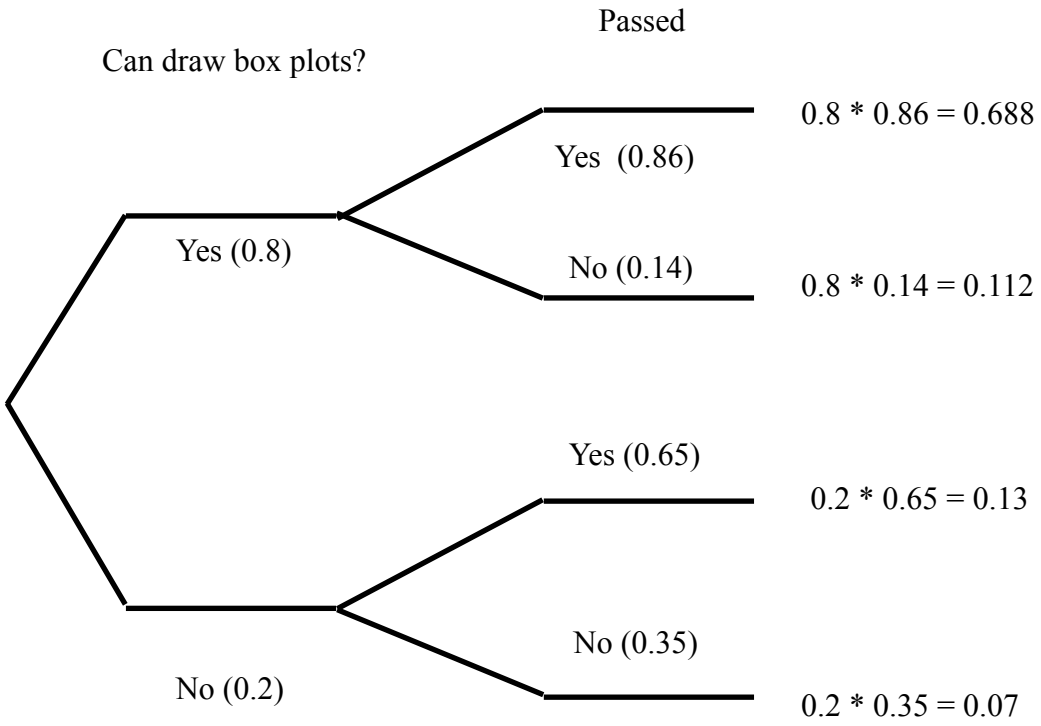
3.19 Drawing box plots. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

Solution:

a)



b)

Let A be the event that a student is able to construct the graph

Let B be the event that the student has passed

So, we need to find P(A | B) where B has occurred.

P(A | B) = P(A and B) / P(B)

= P(Student is able to construct graph and has passed) / P(Student has passed)

P(A and B) = 0.8 * 0.86 = 0.688

P(B) = 0.8 * 0.86 + 0.2 * 0.65 (Students drew and passed + Students did not draw but passed)

= 0.688 + 0.13

= 0.818

So P(A | B) = 0.688 / 0.818

= 0.841

The probability that a student is able to construct a graph given that they passed is 0.841

5)

4.1 Area under the curve, Part I. What percent of a standard normal distribution $N(\mu = 0, \sigma = 1)$ is found in each region? Be sure to draw a graph.

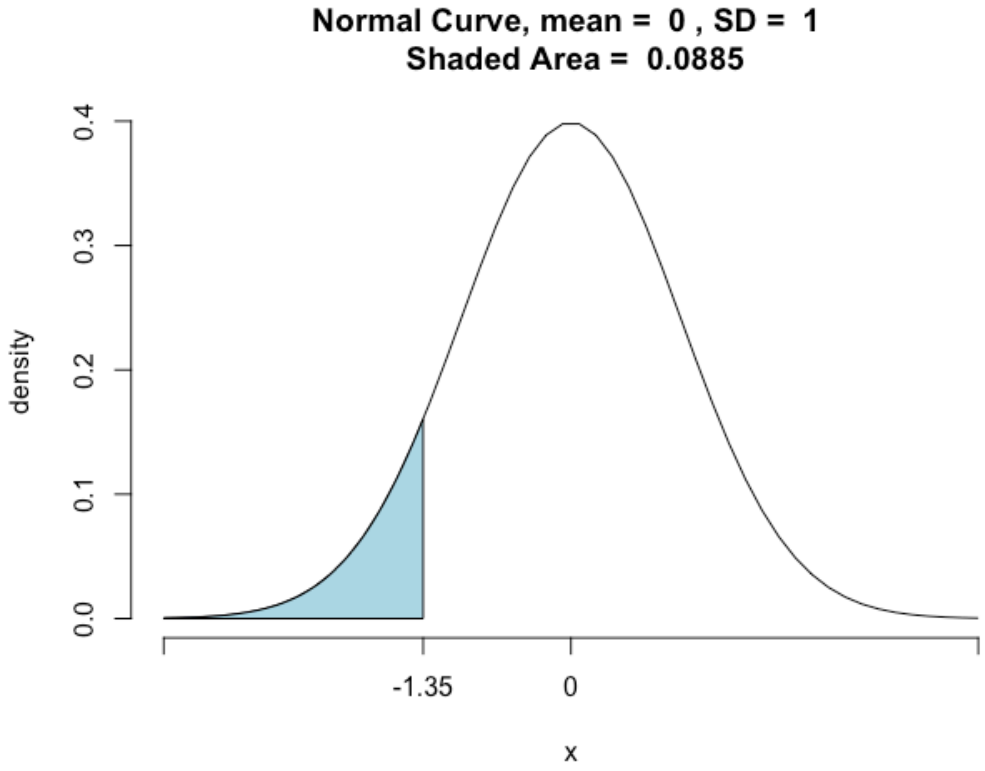
(a) $Z < -1.35$ (b) $Z > 1.48$ (c) $-0.4 < Z < 1.5$ (d) $|Z| > 2$

Solution:

A normal distribution N has μ = 0 and σ = 1.

a) Z < -1.35

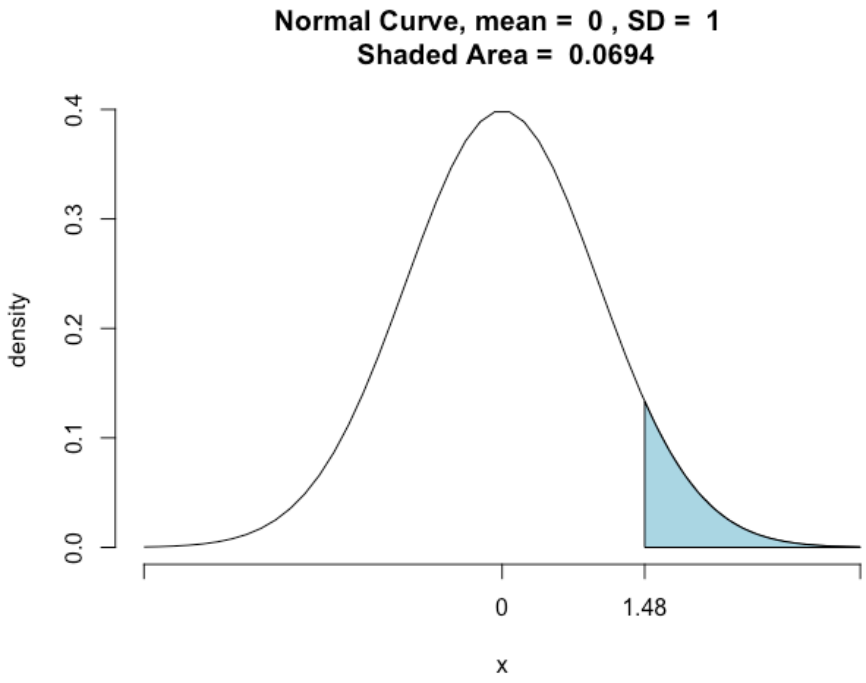
This represents the area which lies below the -1.35 Z value on the left hand side of the curve N.



The percentage of N found in the region Z < -1.35 is 8.85%

b) Z > 1.48

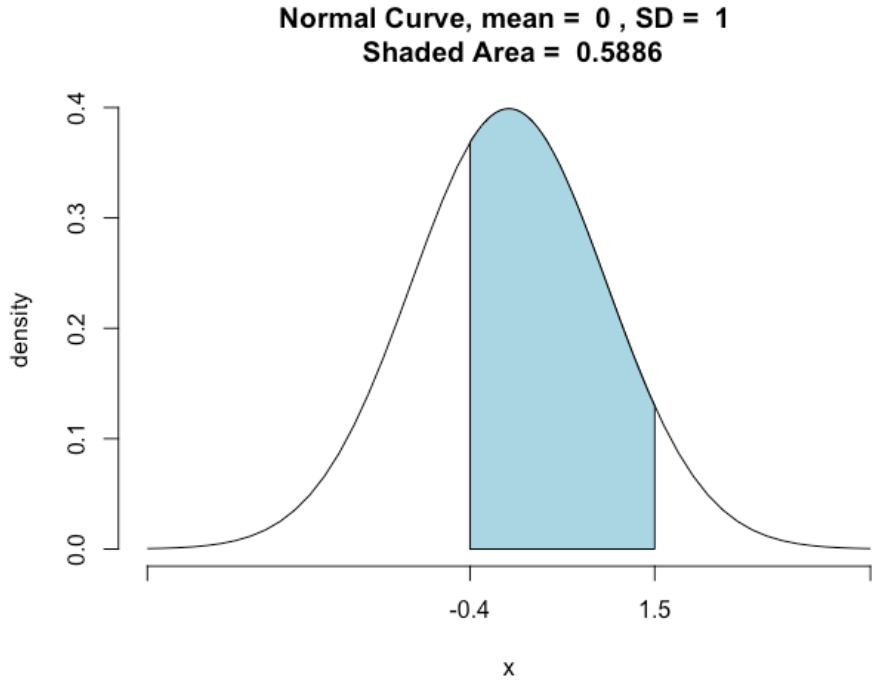
This represents the area which lies above the 1.48 Z value on the right hand side of the curve N.



The percentage of N found in the region Z > 1.48 is 6.94%

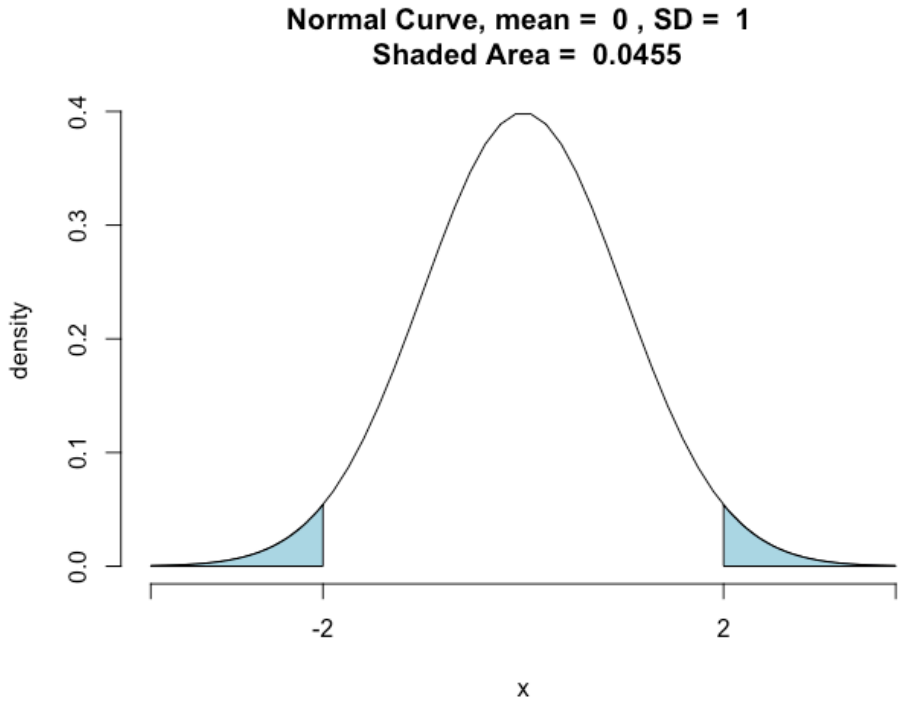
c) -0.4 < Z < 1.5

This represents the area which lies in between the Z values of -0.4 and 1.5 on the curve N.



The percentage of N found in the region -0.4 < Z < 1.5 58.86%

d) |Z| > 2
This modulus sign can lead to 2 different interpretations
=> Z < -2 or Z > 2



The percentage of N found in the region |Z| > 2 is 4.55%

In all the solutions above, we have used the pnormGC function in R to calculate the probabilities for the are under the curve. pnormGC function provides an easy way to compute the cumulative density function of normal distribution with graphs. It takes parameters like bound, mean, standard deviation, the type of region and graph=TRUE.
In our case, we have used the Z values in bounds and they act like x for the function.

Refer the R syntax file for the solution and code snippets.

6)

4.3 GRE scores, Part I. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

(a) Write down the short-hand for these two normal distributions.

(b) What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.

(c) What do these Z-scores tell you?

(d) Relative to others, which section did she do better on?

(e) Find her percentile scores for the two exams.

(f) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

(g) Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.

(h) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

Solution:

a)

μ is mean and σ is standard deviation

Short-hand for Verbal Reasoning Section - **N(μ = 151, σ = 7)**

Short-hand for Quantitative Reasoning Section - **N(μ = 153, σ = 7.67)**

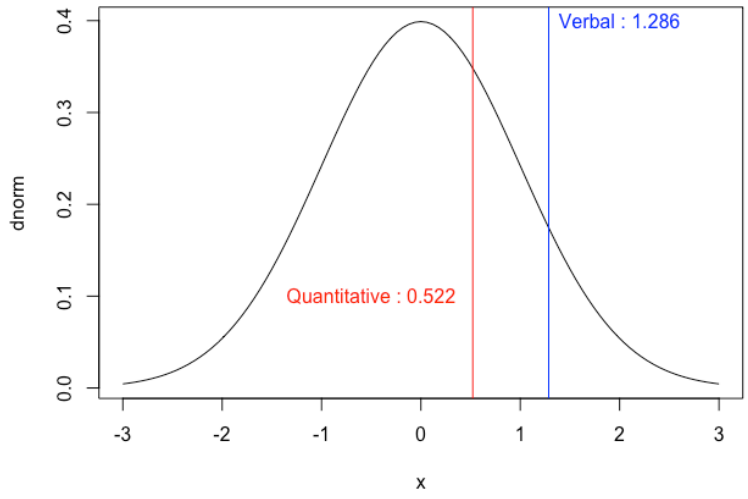
b)

Z-score formula = (X - μ) / σ ….. X is the score obtained by Sophia

Z-score for Verbal section = (160 - 151) / 7
= 9/7 = 1.286

Z-score for Quantitative section = (157 - 153) / 7.67
= 4/7.67 = 0.522

Z-score for Verbal Reasoning is 1.286 and for Quantitative Section is 0.522



c)
The Z-scores obtained by Sophia tell us that she has scored 1.286 standard deviations more than the mean in Verbal Reasoning and 0.522 standard deviations more than the mean in Quantitative Section.

d)
Sophia has scored more Z-score in the Verbal Reasoning than the others. So she has done well in Verbal section.
Also the question seems vague, because when compared with other people she has Z-scores which are above the mean in both Verbal and Quantitative sections.

e)
We have used the pnorm function and passed the Z-score calculated in part b) to get percentile.

Refer the R syntax file.

The Verbal percentile is 90 and Quantitative percentile is 69.9

f)
Since Sophia's percentile on Verbal is 90, the percentage of people who did better than her in Verbal is 100 - 90 = 10%
Since Sophia's percentile on Quantitative is 69.9, the percentage of people who did better than her in Verbal is 100 - 69.9 = 31.1%
Actually there is no data on the number of people, so I have assumed that percentage and percentile are similar.

10% of people did better in Verbal and 31.1% of people did better in Quantitative than Sophia.

g)
We need to normalise the raw scores before making comparisons because there might be certain cases where a student might have a higher Z-score in a section but in reality the raw score might be lesser in that section as compared to the other section.
Like if we do reverse analysis,

If say Z-score of Verbal is 0.6 which is slightly greater than 0.522, then the corresponding raw score might be $X = \text{Z-score} * \sigma + \mu$

$= 0.6 * 7 + 151$

$= 4.2 + 151 = 155.2$

So for Sophia, if we consider this case she has still done better in Verbal as compared to Quantitative even when her score is less.

h)
Since the type of distribution is not known, we need more information. Z-score answers will change but we can calculate them for non-normal distributions, so we can work with parts (b) to (d). But the parts (e) and (f) require to find us percentiles for which we need the distribution type. Otherwise, the questions (e) and (f) cannot be answered if we do not have details about the distribution and its statistics.

References:

Using pnormgc(). (n.d.). Retrieved September 14, 2022, from https://homerhanumat.github.io/tigerstats/pnormGC.html

Adding straight lines to a plot in R programming - abline() function. GeeksforGeeks. (2020, July 14). Retrieved September 14, 2022, from https://www.geeksforgeeks.org/adding-straight-lines-to-a-plot-in-r-programming-abline-function/

Plot: Generic X-y plotting. RDocumentation. (n.d.). Retrieved September 14, 2022, from https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/plot

Text: Add text to a plot. RDocumentation. (n.d.). Retrieved September 14, 2022, from https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/text