**Part 1 (20 pts): Provide written responses in a Word document. Show all work for calculations using either images of pencil and paper work, a separate Excel spreadsheet, or an R script file. Save all work to a folder called “Exam 1” on your computer.**

***1. Describe a population of interest that you might study in the future and how you would seek to draw a sample from this population (1pt).***

Answer: I might study with ‘Buildings in Columbia Downtown areas’ population in the future. If we look into the characteristics of downtown areas, the buildings are generally mixed use (lower floor for commercial and upper floors for residential) in nature. However, due to large number of buildings, it is hard to conduct simulation study with all the buildings which will be enormous time consuming. To make the research feasible, I would prefer to draw random sampling from this population, where each building from the population will have equal chance of being chosen. For instance, Columbia downtown has more than 500 buildings which have similar characteristics in usage and materials. Through computer processed lottery method, I can randomly select 50 buildings out of these buildings.

***2. Describe an experiment that you could conduct with the above sample, being sure to discuss random sampling, the independent variable, and the dependent variable (3 pts).***

Answer: With the above sample, I can conduct, ‘Simulation framework for predicting downtown-wide energy consumption by Photovoltaic panels’. My hypothesis for this research is that, installing photovoltaic panels on building façade can reduce the energy consumption in the buildings of downtown areas.

To conduct the research, I can select Columbia downtown as the case study area. As I explained in the first question, I can randomly select 50 buildings out of 500 buildings through computer processed lottery method where each building will have the same probability of being chosen. This will create a balanced subset (as the specific characteristics of the buildings are similar) that’ll carry the greatest potential for representing the larger group as a whole, free from any bias. Then, I can simulate and get the result of total energy consumption in units for the sample buildings. Finally, I can assign photovoltaic panels on the same building facades through computer simulation to compare the result.

Here, independent variable is ‘Installation of Photovoltaic panels’ which would have a nominal measurement of ‘yes’ or ‘No’. On the other hand, ‘Energy consumption in units’ is dependent variable and its level of measurement would be Ratio. Because, we can have 0-unit or hundreds of units as reduction (therefore, improvement) in energy consumption by installing photovoltaic panel.

***3. Describe a correlational study that you could conduct with the above sample, being sure to note the variables of interest (2 pts).***

Answer: Let us assume that the buildings of Columbia downtown are the units of our analysis. I am wondering if there is any relationship between formal building configurations and crime. I can use the above-mentioned sample buildings for collecting the vast amount of data of the variables to conduct a correlational study which may lead towards a proper recommendation for crime prevention through proper building design.

Here, investigation can be conducted through setting the relationship between the physical variables of the building configurations (building entry orientation, building height, site layout) and the number of crime incidents.

***4. The Achievement variable represents a measure of standardized achievement for students in an elementary school class. What is the mean and standard deviation for this variable (use formulas from the book to calculate your answers and show your work)?***

Answer:

**Mean Value for Achievement variable, = 99.39**

**Standard Deviation for Achievement variable, *SD* = 4.90**

r script for hand calculation:

*# Hand calculating mean and standard deviation of Achievement variable---*

*# Mean Value---*

*Achievement = Aman\_Part1$Achievment*

*X = sum(Achievement)*

*X*

*N = length(Achievement)*

*N*

*Xbar = X/N*

*Xbar*

*#Variance---*

*Dif = (Achievement-Xbar)^2*

*SS = sum(Dif)*

*S2 = SS/(N-1)*

*S2*

*#Standard deviation---*

*S = sqrt(S2)*

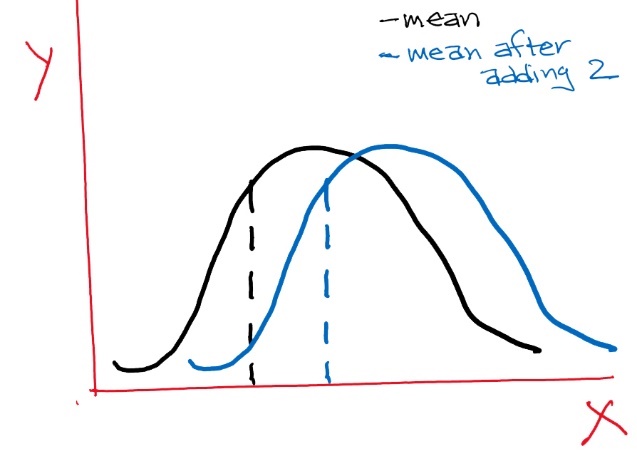
*S*

***5. Without using any discussion of calculations, explain what would happen to the calculated***

***mean if I added 2 additional points to all students’ Achievement score (be sure to provide a***

***clear rationale for your answer)?***

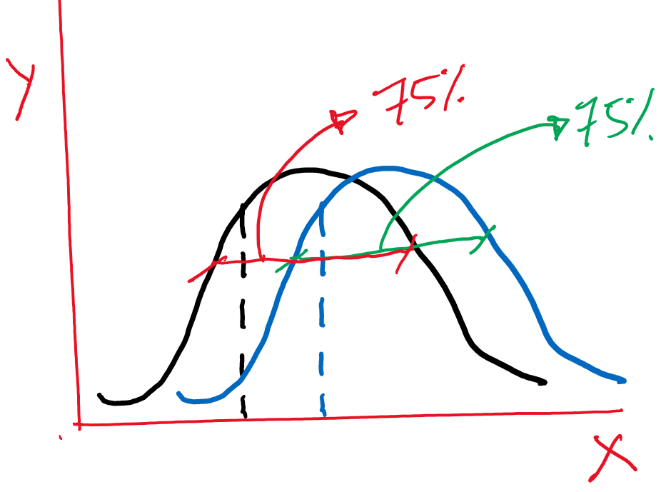
Answer: Adding 2 additional scores to all students’ Achievement score will not affect the overall frequency distribution curve. It will only shift towards the positive X axis by 2. The reason behind of this situation is that, we are not adding data to specific entries, but all of our data entries have increased equally by +2. Figure 1 presents the conceptual representation of previous and present situation. Suppose, the current mean value of Achievement score is **.** if we add 2 additional points to all students’ Achievement score, the new mean will be, .



***Figure 1: Previous and present mean***

***6. Without using any discussion of calculations, explain what would happen to an individual’s percentile rank if I added 2 additional points to all students’ Achievement score (be sure to provide a clear rationale for your answer)?***

Answer: As the additional 2 points have been equally added to everyone, Individual percentile rank will not be affected by the new changes. The frequency (consequently cumulative frequency and percentage) curve stays the same. Let’s say, Jack scores at 75th percentile which equals or exceeds 75 percent of the other individuals in the distribution. If we add 2 extra points to all of them, Jack’s rank will be at the same place (Figure 2). Moreover, we are not changing the number of participants. Therefore, if an individual’s percentile rank would remain in the same ranking after adding the 2 additional points to all students’ Achievement score.

******

***Figure 2: Previous and present percentile rank***

***7. Aptitude is a measure of student aptitude in the elementary school class. If we assume that the variable is normally distributed, what is the z score and aptitude score at the 50th percentile?***

Answer: Here, ***Z*** score at the 50th Percentile (.50) is 0.00 based on the standard normal distribution table.

Mean value of Aptitude variable*,* = 50.77

Standard Deviation of Aptitude variable, = 5.80

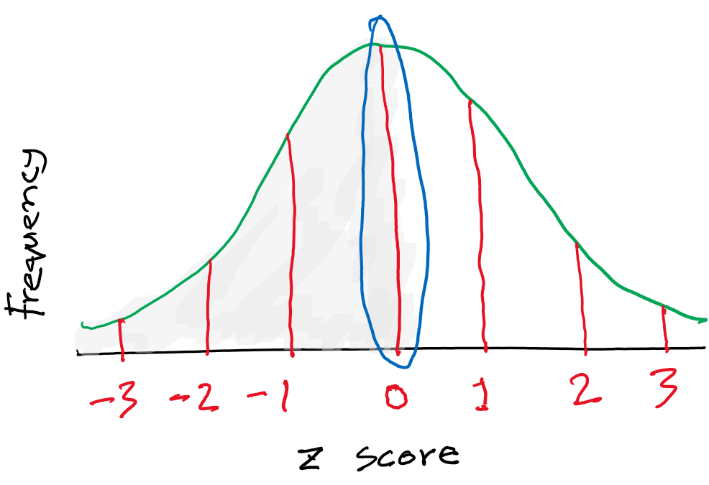
We know that,

Here, ***X*** is aptitude score.

Calculation:

***Z***= 50.77

Therefore, at 50th Percentile, ***Z*** score is Zero and Aptitude score is 50.77.



r script for hand calculation:

*# Hand calculating mean and standard deviation of Aptitude variable---*

*# Mean Value---*

*Aptitude = Aman\_Part1$Aptitude*

*X1 = sum(Aptitude)*

*X1*

*N1 = length(Aptitude)*

*N1*

*Xbar1 = X1/N1*

*Xbar1*

*#Variance---*

*Dif1 = (Aptitude-Xbar1)^2*

*SS1 = sum(Dif1)*

*S21 = SS1/(N1-1)*

*S21*

*#Standard deviation---*

*S1 = sqrt(S21)*

*S1*

***8. What is the z score and percentile rank for an aptitude of 55?***

Answer: For aptitude,

Mean value of Aptitude variable, = 50.77

Standard Deviation of Aptitude variable, = 5.80

(Hand Calculated through r script for previous answer)

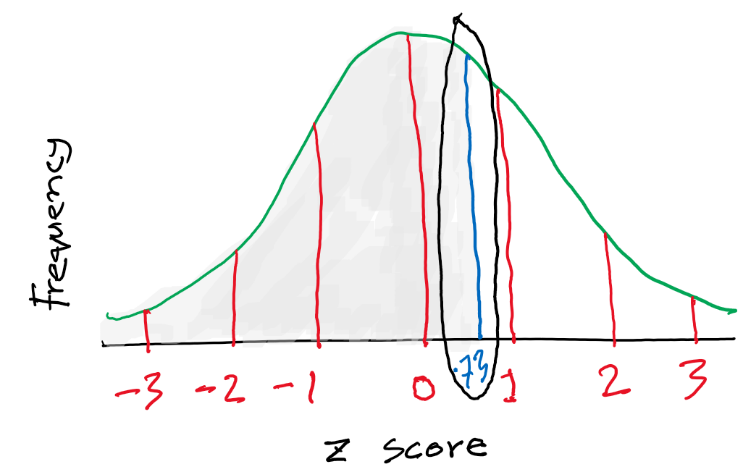
Here, aptitude raw value is 55

We know that,

Calculation:

***Z*** = 0.73

According to the standard normal distribution table, percentile rank on calculated ***Z*** value (+0.73) is 76.73%



***9. If the mean of a continuous distribution of calculated statistics is 39 and the standard deviation is 3.8, provide the following (4 pts total):***

1. ***The probability of a calculated statistic being above 50***

Answer: Here,

Mean value of Aptitude variable, = 39

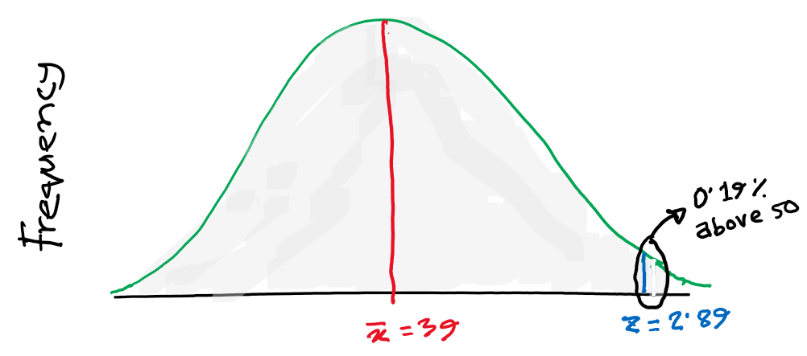
Standard Deviation of Aptitude variable, = 3.80

We know that,

Calculation

***Z*** = 2.89

According to the distribution table, we get 49.81% area between the mean and ***Z*** score, which means that, the probability of a calculated statistic being above 50 is 0.19%.

******

1. ***The probability of a calculated statistic being below 25***

Answer: Here,

Mean value of Aptitude variable, = 39

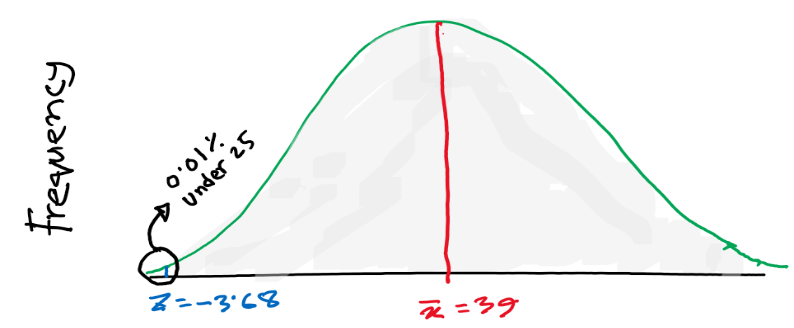
Standard Deviation of Aptitude variable, = 3.80

We know that,

Calculation: ***Z***

***Z*** = -3.68

According to the distribution table, we get 49.99% area between the mean and ***Z*** score, which means that, the probability of a calculated statistic being below 25 is 0.01%.

******

1. ***The probability of a calculated statistic being between 31.4 and 46.6***

Answer: Here,

Mean value of Aptitude variable, = 39

Standard Deviation of Aptitude variable, = 3.80

We know that,

For ***X*** value of 31.40, Calculation:

***Z*** = -2.00

According to the distribution table, we get 47.72% area between the mean and ***Z*** score

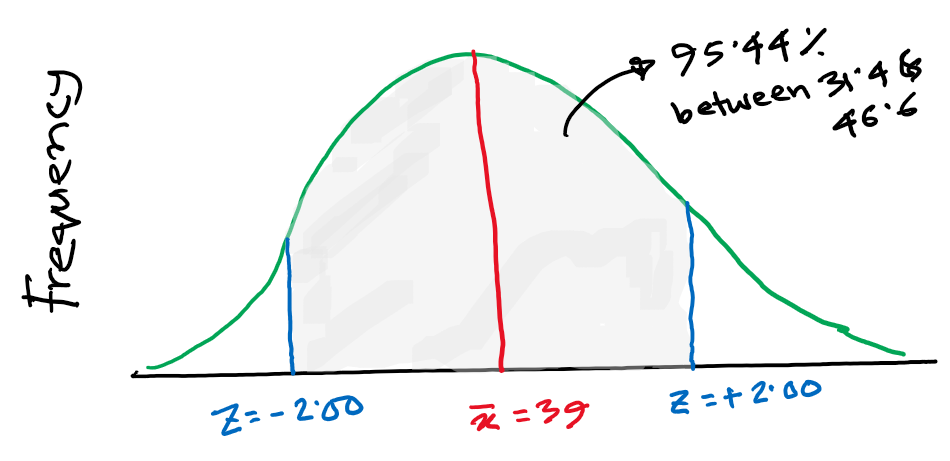
For ***X*** value of 46.60, Calculation:

***Z*** = +2.00

According to the distribution table, we get 47.72% area between the mean and ***Z*** score.

Therefore, the probability of a calculated statistic being between 31.40 and 46.60 is:

(47.72+47.72) %= 95.44%



**Part 2 (20 pts): Download your safety.csv (i.e., name\_safety.csv) data file from the Exam 1 assignment in Canvas and save it to your “Exam 1” folder on your computer. Use this file to complete tasks below.**

***3. In your Exam 1 Word document from Part 1 of this exam:***

1. ***provide a table for the frequencies and percentages of the appropriate variables and another table with the means, medians, and standard deviations for the appropriate variables.***

Answer:

***Table 1: Frequencies and Percentages for nominal/ordinal variables***

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | | **Frequency** | **Percent** |
| Overall (n) = 234 | | | |
| **Size** | <300 | 42 | 17.95% |
| 300-499 | 82 | 35.04% |
| 500-999 | 95 | 40.60% |
| 1000+ | 15 | 6.41% |
| **Crime** | Low | 51 | 21.80% |
| Moderate | 53 | 22.60% |
| High | 130 | 55.60% |
| **Position** | Other | 38 | 16.24% |
| principal | 159 | 67.95% |
| Security | 3 | 1.28% |
| Vice | 34 | 14.53% |
| **Corporal** | No | 178 | 76.10% |
| Yes | 56 | 23.90% |
| **OSS** | No | 75 | 32.10% |
| Yes | 159 | 67.90% |

***Table 2: Means, Medians, and Standard Deviations for interval/ratio variables***

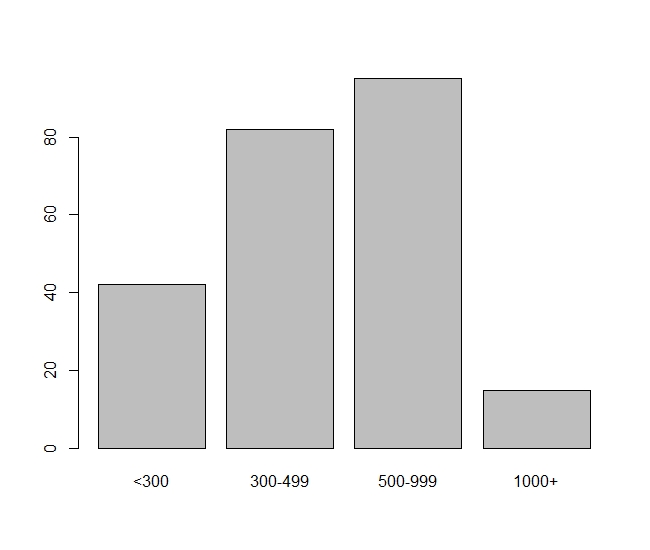
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | ***M*** | ***SD*** | ***MDN*** |
| **Violence** | 6.80 | 8.80 | 2.00 |
| **Disrupt** | 73.27 | 63.47 | 69.50 |

1. ***provide an APA write up detailing the appropriate descriptive statistics for at least one nominal/ordinal variable and at least one interval/ratio variable in the data file.***

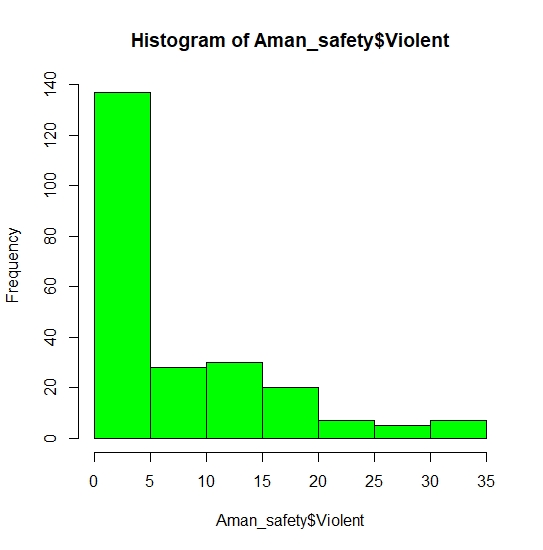
Answer: A total of 234 schools were studied. Frequencies and percentages for categorical variables are presented in Table 1. For the Size category, 42 (17.95%) were less than 300, 82 (35.04%) between 300 and 499, 95 (40.60%) were 500- 999 and 15 (6.41%) larger than 1000 were reported. We also collected data about Violence (*M*= 6.80, *SD*= 8.80), presented in Table 2.

1. ***provide a graph for each variable discussed in (b).***

Answer:

******

***Figure 1: Bar chart for ‘Size’ variable***

******

***Figure 2: Histogram for ‘Violence’ variable***