DSC 423: Data Analysis and Regression

Assignment 1:

Name: Adarsh Shankar

Student ID: 2117611

Honor Statement: "I have completed this work independently. The solutions given are entirely my own work."

---

1. Short Essay (10 pts.) Consider the following two scenarios. A) take a simple random sample of 100 graduate students at DePaul university and b) take a simple random sample of 100 graduate students studying Data Science. For each sample you record the amount spent on textbooks used for classes. Which sample do you expect to have the smaller standard deviation? Explain your answer.

Ans:

a) DePaul University has 100 graduate students.

Since DePaul University offers a large number of graduate courses, we can state that a student's textbook selection will vary depending on their major, as well as the quantity of textbooks they are required to read, which further widens the range of choices.

b) 100 graduate students studying

Now we are going to consider the amount spend by the students of Data Science, as all Data Science students have same subjects is their courses, the amount spent on textbooks will be similar or have less variation.

When comparing the two scenarios, it is clear that case (a) has a greater variety of textbooks, increasing both the money paid and the standard deviation.

In contrast, all of the students in instance (b) are enrolled in the same major and are using identical, equally priced textbooks.

Therefore, based on our observations, we can conclude that, when compared to 100 graduate students studying at DePaul University, the standard deviation of the amount spent on 100 graduate students studying Data Science will be lower.

2. Empirical rule (20 pts.) The 222 students enrolled in online-learning courses offered by a college ranged from 18 to 64 years of age. The mean age was 28 with standard deviation equal to 4. Use the 68-95-99.7 rule to answer the following questions:

a. (10 pts.) Compute the percentage of students that are between 24 and 32 years old. Show your work.

b. (10 pts.) Compute the percentage of students that are older than 36 years. Show your work.

Ans:

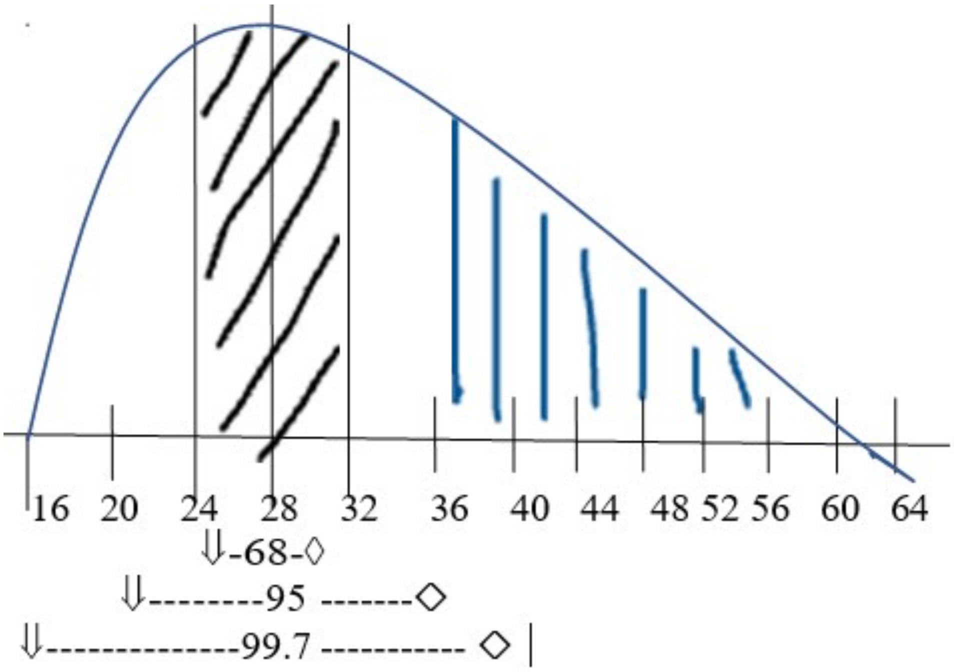
Given data:

Number of students: 222

Range of age: 18-64

Mean: 28

Standard Deviation: 4



a) The percentage of students aged 24 to 32

By taking into account the black-shaded area and applying the empirical rule, we may determine that 68% of the students are located there.

b). The proportion of students who are over 36 years old Consider about the shaded area (blue)

Students who are older than 36 years lay after the second standard deviation

= 99.7 - 95 = 4.7%

= 4.7/2 = 2.35

The final quartile also equals 100-99.7 = 0.30 percent.

=0.30/2 \s=0.15%

The proportion of students over the age of 36 is therefore 2.35 + 0.15 = 2.50%

3. Z-scores (10 pts.) Monthly sale figures for a particular e-retailer tend to be normally distributed with mean equal to 150 thousand dollars and a standard deviation of 35 thousand dollars. Use the normal distribution to determine the top 1% monthly sale figure (a.k.a. 99th percentile)? Show your work.

Ans:

Given:

Mean = $150,000

Standard Deviation = $35,000

We Know Z-score = $z = \frac{X-\mu}{\sigma}$

X =?

P= 0.99

Qnorm (0.99) Equals 2.326348 in Rstudio.

Additionally, we can infer from the z table that 2.3 and 0.03 (2.33) = 0.99.

Using (1) = 2.33 = $\frac{X-150{,}000}{35{,}000}$

= 81,550 = X-150,000
= X = 150,000 + 81,550
= X = 231,550

Therefore, the top 1% monthly sale figure = **$231,550**

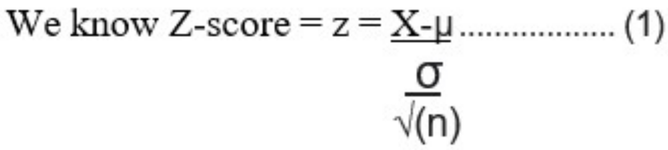
4. Hypothesis Testing (10 pts.) A network provider investigated the number of blocked intrusions to its network, and found that there were, on average, 45 blocked intrusions per day. After a change in firewall settings, the mean number of intrusions during the next 35 days was 42 with a standard deviation equal to 15.5. Perform a hypothesis test to determine if the change in firewall settings reduced the number of intrusions. Show your work.

Ans:

Given:

Null hypothesis (Ho) = 45
Alternate hypothesis (Ha) < 45

A = 0.05



We know Z-score = $z = \dfrac{X-\mu}{\dfrac{\sigma}{\sqrt{(n)}}}$ .... (1)

$$Z = \frac{42-45}{\frac{15.5}{\sqrt{35}}}$$

The equation's solution is z = -1.145.

Pnorm (-1.145) = 0.126 = p using rstudio p = 0.126 and

A = 0.05 are now known to us.

We can state that we fail to reject the null hypothesis since p > A. (Ho).

$$Z = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}}$$