Assignment 1
DA 241 Statistical Foundations for Data Science
Instructor: Rhythm Grover
Teaching Assistant: Kartikay Agarwal

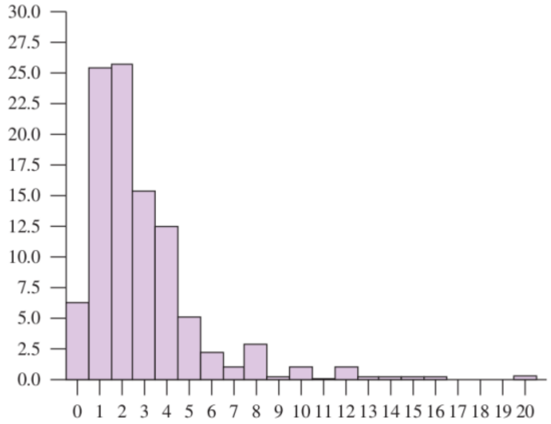
1. The following are the quartiles of a large data set:
   First Quartile = 33
   Second Quartile = 54
   Third Quartile = 73.

   a) Give an interval in which approximately 50 % of the data lie.

   b) Give a value which is greater than approximately 50 % of the data.

   c) Give a value such that approximately 25 % of the data values are greater than it.

2. During a census survey, Annapalai asked people "On an average day, about how many hours do you watch T.V.?". Below is the histogram recording recording the number of hours versus percentage of 899 responses.



   a) What was the most common outcome?

   b) What percentage of people reported watching TV no more than 2 hours per day?

3. Data were taken from an experinlent on three groups of mice. The measurements are amounts of nitrogen-bound bovine serum albumen produced by normal mice treated with a placebo (i.e. an inert substance), alloxan-diabetic mice treated with a placebo, and alloxan-diabetic mice treated with insulin. The data are given in the following table:

Normal	Alloxan-diabetic	Insulin treatment
156	391	82
282	46	100
197	469	98
297	86	150
116	174	243
127	133	68
119	13	228
29	499	131
253	168	73
122	62	18
349	127	20
110	276	100
143	176	72
64	146	133
26	108	465
86	276	40
122	50	46
455	73	34
655		44
14		

a) Summarize the three groups in terms of their five-figure summaries.

b) Calculate the mean and standard deviation for each group.

c) Calculate the sample skewness for each group.

d) Obtain a comparative boxplot for the three groups. Are any differences apparent between the three treatments?

4. In the table below, you are given the top 20 movies worldwide along with the gross amount that was earned from ticket sales for each of these movies. Summarise the data in the form of a frequency table (class intervals, frequency, relative frequency). Group the data into classes according to their values such that the class midpoints are convenient numbers.

Title	Worldwide gross
Dangal	₹2,024 crore
Baahubali 2 The Conclusion	₹1,810 crore
RRR	₹1,200 crore
K.G.F: Chapter 2	₹1,200 crore
Pathaan	₹1,050.3 crore
Bajrangi Bhaijaan	₹918.18 crore
Secret Superstar	₹858.43 crore
PK	₹769.89 crore
2.0	₹655.81 crore
Sultan	₹623.33 crore
Baahubali: The Beginning	₹600 crore
Sanju	₹586.85 crore
Padmaavat	₹585 crore
Tiger Zinda Hai	₹565.10 crore
Dhoom 3	₹556.74 crore
Ponniyin Selvan: I	₹495.50 crore
War	₹475.50 crore
3 Idiots	₹460 crore
Andhadhun	₹456.89 crore
Vikram	₹432.50 crore

a) Plot a histogram to show the distribution of earned gross amounts using the frequencies obtained in the previous part.

b) Again plot a histogram for data in part (c) but this time with relative frequencies on the y-axis.

c) Draw a frequency polygon by connecting the mid-points of the tops of the rectangles in the histogram obtained in part (d).

d) Find the mean of the histogram:

- Calculate x_i = The midpoint of the i^{th} bin
- Find f_i = The frequency of the i^{th} bin
- Mean = $(\sum_{i=1}^{k} x_i f_i)/n$, where k = number of bins, $n = \sum_{i=1}^{k} f_i$ = sample size

e) Median of a histogram is the value with half the area to the left and half to the right. Identify the median group. Find the median of the histogram by doing the following computations:

- L = The lower limit of the median group
- n = sample size
- f = The frequency of the median group
- F = The cumulative frequency up to the median group
- w = width of the bin of the median group
- $$\text{Median } = L + \frac{(n/2 - F)}{f} \times w$$

5. Let us look at the `murders` dataset in RStudio. You can load the dataset into your R console with the following code:

```
install.packages("dslabs")
library(dslabs)
data(murders)
```

   a) Use the function names to extract the variable names

   ```
   names(murders)
   ```

   b) Compute the per 100,000 murder rate for each state and store it in an object called murder_rate.

   c) Create a histogram of the state populations.

   d) Compute the average murder rate for each state. How many states are below the average?

   e) Plot a bar graph representing the murder rates per 100,000 population for each state. Identify the states with the highest and lowest murder rates from the graph and discuss any regional patterns observed.

   f) What is the median population size among all states in the dataset? Discuss the significance of using the median instead of the mean in this context.

   g) Generate boxplots of the state populations by region.

   h) Compare and contrast the bar graph and box plot visualizations of murder rates in different states:

      I What unique insights can each plot provide, and how do they complement each other in understanding the distribution of murder rates?

      II Discuss the advantages and limitations of each plot in conveying information about the dataset.

   i) Compute the murder rate range (difference between the highest and lowest rates) among all states in the dataset. How does this range reflect the diversity of murder rates across the United States?

6. One surveyor takes a sample of 100 women age $18 - 35$ in a certain town. Another takes a sample of 1,000 such women.

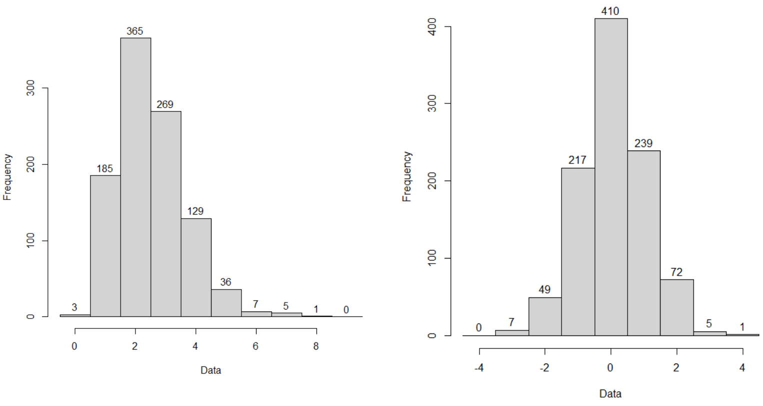
   a) Which surveyor will get a bigger average for the heights of the women in her sample? or should the averages be about the same?

   b) Which surveyor will get a bigger SD for the heights of the women in her sample? or should the SDs be about the same?

   c) Which surveyor is likely to get the tallest of the sample women? or are the chances about the same for both surveyors?

   d) Which investigator is likely to get the shortest of the sample women? or are the chances about the same for both surveyors?

7. The figure below shows histograms of three different data sets.

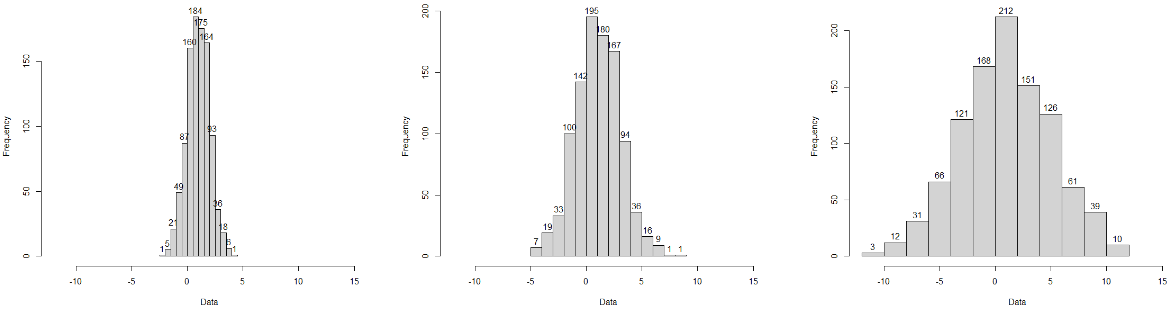


a) Find the approximate means of the three histograms. Compare and order them.

b) Find the approximate medians of the three histograms. Compare and order them.

c) Compare the mean and median for each histogram.

8. The figure below shows histograms of three different data sets.



Order the standard deviations of the three data sets.