Academic year 23-24
Subject: Probability and Statistics (01CT1401)
Tutorial 1 as a part of Term work assessment

T1LA.1

Problem Statement:

A wholesale distributor operating in different regions of India has information on annual spending of several items in their stores across different regions and channels. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Gujarat, West Bengal, Other) and across different sales channel (Hotel, Retail).

1.1 Use methods of descriptive statistics to summarize data. Which Region and which Channel seems to spend more? Which Region and which Channel seems to spend less?

1.2 There are 6 different varieties of items are considered. Do all varieties show similar behaviour across Region and Channel?

1.3 based on a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

1.4 Are there any outliers in the data?

1.5 based on this report, what are the recommendations?

Data sheet is attached in separate CSV/Excel file (T1LA_data_1.csv/excel)

T1LA.2

(A) Write a code in C/C++/Java/Python to take numbers as an input and gives output five points summary of box plot. It should also declare outlier if any.

(B) The nine measurements that follow are furnace temperature recorded on successive batches in a semiconductor manufacturing process (Units are in °C): 511, 513, 509, 515, 510.5, 509.4, 514, 512, 513.3

(i) Calculate the sample mean, sample variance and sample standard deviation.

(ii) Find the median. How much could the largest temperature measurement increase without changing the median value?

(iii) Construct a box plot of the data.

(C) The following data are the joint temperature of the O-rings (°F) for each test firing or actual launch of the space shuttle rocket motor (from the presidential commission on the space shuttle challenger accident, Vol 1, pp. 129-131):

84, 49, 64, 63, 40, 70, 83, 78, 67, 52, 45, 67, 68, 53, 70, 67, 69, 75, 79, 61, 58, 70, 68, 81, 58, 76, 67, 79, 72, 75, 73, 76, 70, 58, 57, 31

(i) Compute the mean, mode and standard deviation. Construct the dot plot of the graph

(ii) Computer the median, upper and lower quartile and prepare box plot.

(iii) Is there any potential outlier? if yes, find the outlier data

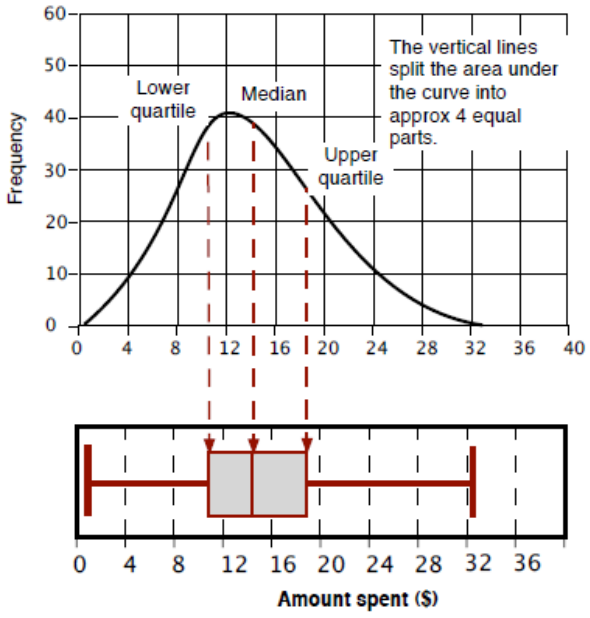
(iv) For this data suggest the what are the two extremes out of which data will be considered as extreme outliers?

(v) Set aside the lowest observation (31°F) and recompute the quantities in part (ii). Comment on your findings. How "different are the other temperature from this lowest value?

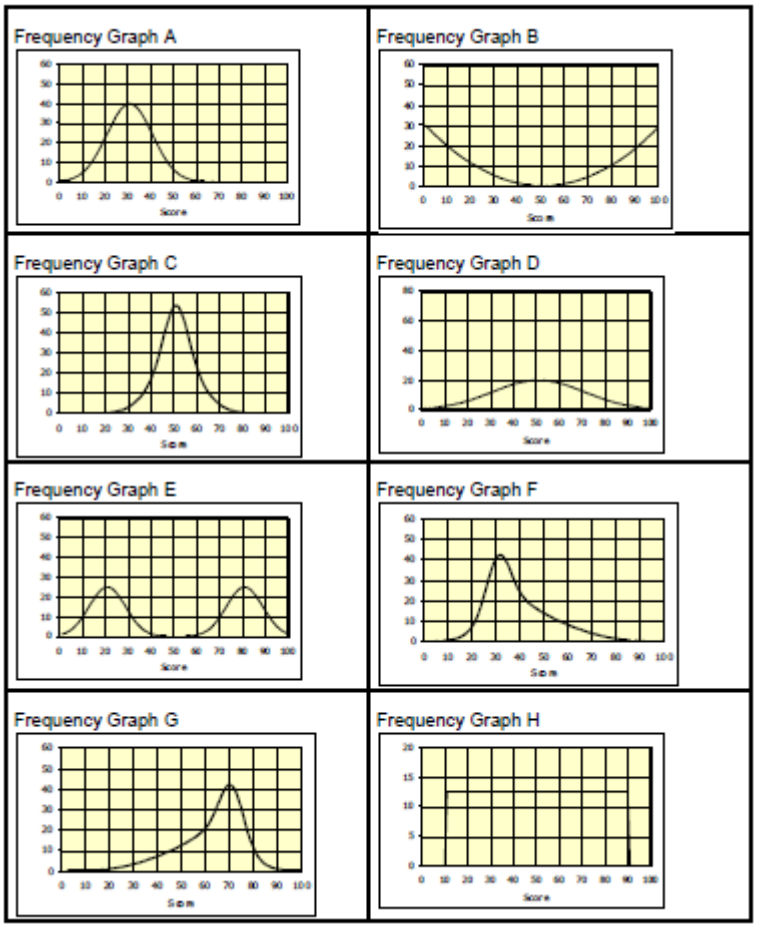
(D) Write a script in python to find SD and Variance of any given data, test it by applying three different data set. Submit script and output of the code

(E) Using Python programming script and data (T1LA_dataset_2_diabetes.csv/excel) plot the box and whisker plot for all features and class. Write your inference from the box plots. Submit scripts and outputs.

(F) We can correlate box plot and frequency plot as follows



(i) With the above reference draw the box plot for the given frequency plot.



(G)

The data of 30 customers on credit card usage in INR1000, sex (1:male, 2: female) and whether they have done shopping or banking (1: yes , 2: no) with credit card are given in file (T1LA_dataset_3_CC_Expenses_Exercise.csv/excel)

a. Import the file to Python
b. Compute descriptive summary of variable Credit Card Usage
c. Check whether the average usage varies with sex?
d. Check whether the average credit card usage vary with those who do shopping with credit card and those who don't do shopping?
e. Check whether the average credit card usage vary with those who do banking with credit card and those who don't do banking?
f. Compute the aggregate average of usage with sex & shopping?
g. Compute the aggregate average of usage with all three factors?

(H) Refer the research paper "Data analysis using Box and Whisker Plot for Lung Cancer" presented in *International Conference on Innovations in Power and Advanced Computing Technologies [i-PACT2017]*
Write down five important/useful learning (which was not covered in any of the problem in this assignment)

Suggested extra reading : Multidimensional box plot, different box plot width types

T1LA.3

(A)

Population. The Census of India provides a variety of statistical information on different aspects of the Indian population. According to the Population Enumeration Data for the 2011 census, provided on the official website for the Indian census censusindia.gov.in, the top 9 metro cities on the basis of their population are as follows.

City	**Population (in million)**
Mumbai	18.4
Delhi	16.3
Kolkatta	14.1
Chennai	8.7
Bangalore	8.5
Hyderabad	7.8
Ahemdabad	6.4
Pune	5.1
Surat	4.6

a. Compute the population mean enrolment, μ, of the cities. (Round your answer to two decimal places.)
b. Compute σ. (Round your answer to two decimal places.)
c. Letting x denote a city, specify the standardized variable, z, corresponding to x.
d. Without performing any calculations, give the mean and standard deviation of z. Explain your answers.

(B) Write a code in C/C++/Java/Python which can take input numbers and calculate Z score (Zi)against each data(Xi).

(C) Refer the link

https://www.linkedin.com/advice/1/what-benefits-drawbacks-using-z-scores-standardize

Topic "What are the benefits and drawbacks of using z-scores to standardize your data for predictive modeling?"

Write down learning from it.

T1LA.4

(A) Write a code in C/C++/Java/Python which can take input numbers and declared the following requirement

(i) Mean, median, mode

(ii) Distribution is unimodal, bimodal or multimodal

(iii) Declare the shape of distribution normal(central)/left skewed/right skewed by comparing mean and median

(B) Refer the activity applet (http://www.shodor.org/interactivate/activities/SkewDistribution/)
Explore the relationship between central tendency, skewness, relation of mean and median for various shapes. Paste at least 8 different scenario with two -three line conclusion on each. Guideline for the activity is mention in instructor and help menu. Also answer the following questions.

This applet lets you see how the results of an experiment with a lot of trials might look if the mean, median, and mode aren't the same. Experiment with moving the median and standard deviation lines to change the shape of the graph. Try to answer these questions:

(i) Move the median line a small amount away from center, about half a box. Where is the mode? Did it move farther than the median? The same amount? Not as far?

(ii) Create a 100-trial histogram with the median off-center. Is the mode of the histogram where you expect it to be? The median? Repeat the trial a few times. Also try with a 1000-trial histogram, and experiment with the bin size. Are the median and mode close to where the curve has them?

(iii) Do you think it would be easier or harder to recognize this kind of distribution than the normal distribution in an experiment?

(C) Using the link (https://www.geogebra.org/m/BxqJ4Vag) give three case study by entering suitable data(min 30 data).

(i) Mean = Median (ii) Mean > Median (iii) Mean >> Median
(iv) Mean < Median (v) Mean << Median (vi) Unimodal Distribution (vii) Bimodal distribution (viii) multimodal distribution

Paste the result screenshot in pdf format.

(D) The following data are direct solar intensity measurement (watts per meter square). On different days at a location in southern spain :
562, 869, 708, 775, 775, 704, 809, 856, 655, 805, 878, 909, 918, 558, 768, 870, 918, 940, 946, 660, 820, 898, 935, 952, 957, 693, 835, 905, 939, 955, 960, 499, 653, 730, 753.
Calculate sample mean, mode and median. Prepare dot diagram manually and indicate mean, median, mode in it.

(E) Write a code to generate 12 different numbers randomly from the range entered by user. (you can also ask to enter choice of integer, fractional value with one digit in fraction, fractional value with two digit in fraction).
(i) Calculate mean,median and mode Integer outcome of this code manually on paper
(ii) Calculate mean median and mode of choice 2/3 of the code.

T1LA.5

(A) Refer the following link and enlist all possible types of graphs used in statistical data analysis. Write a brief conclusion about each graph in your own words to represent the advantages/characteristics of particular graph. Take one case study data for each graph separately and represent it using any online/offline tool. Write inference from the graph below the chart. (You can refer other learning resources as well for the topic)

(i) https://www.intellspot.com/types-graphs-charts/
(ii) https://piktochart.com/blog/types-of-graphs/
(iii) https://www.datapine.com/blog/different-types-of-graphs-charts-examples/
(iv) https://www.studysmarter.co.uk/explanations/math/statistics/statistical-graphs/

(B) The following are figure on an oil well's daily production in barrels : 214, 203, 226, 198, 243, 225, 207, 203, 208, 200, 217, 202, 208, 212, 205 and 220. Construct a stem-and-leaf display with the stem labels 19, 20,…., and 24.

(C) The following are determinations of a river's annual maximum flow in cubic meters per second: 405, 355, 419, 267, 370, 391, 612, 383, 434, 462, 288, 317, 540, 295, and 508. Construct a stem-and-leaf display with two-digit leaves.

(D) List the data that correspond to the following stems of stem-and-leaf displays:

(i) 1 | 1 2 3 4 5 7 8 . Leaf unit = 1.0
(ii) 23 | 0 0 1 4 6. Leaf unit = 1.0
(iii) 2 |03 18 35 57. First leaf digit unit =10.0
(iv) 3.2 | 1 3 4 4 7. Leaf unit = 0.01

(E) The following are the IQs of 20 applicants to an undergraduate engineering program: 109, 111, 106, 106, 125, 108, 115, 109, 107, 109, 108, 110, 112, 104, 110, 112, 128, 106, 111, 108. Construct a five-stem display with one digit leaves.

(F) Use the link (https://www.calculatorsoup.com/calculators/statistics/stemleaf.php) Enter five different use case (which represent various types of stem-leaf) data and take screenshot of plot and write inference from each of it.

T1LA.6

(A) **Exam Scores.** Consider the following sample of exam scores, arranged in increasing order.

28 57 58 64 69 74
79 80 83 85 85 87
87 89 89 90 92 93
94 94 95 96 96 97
97 97 97 98 100 100

The sample mean and sample standard deviation of these exam scores are 85 and 16.1, respectively. solve the following problems.

a. Compare the percentage of the observations that actually lie within two standard deviations to either side of the mean with that given by Chebyshev's rule with $k = 2$.

b. Repeat part (a) with $k = 3$.

(B) **Alcohol Consumption.** In the *Global Status Report on Alcohol and Health 2014*, the World Health Organization reports that in 2010, the worldwide consumption was 6.2 liters of pure alcohol per person (15 years or older), which amounts to 13.5 grams of pure alcohol a day. The countrywise total, recorded and unrecorded alcohol per capita consumption (APC), in 2010 (in liters of pure alcohol) is given in the report. The following data is the APC for the African region (45 Countries).

0.1 0.2 0.6 0.7 1 1.1 2.1 2.3 3.4
3.6 3.6 3.9 4 4.3 4.4 4.8 5.6 5.7
6 6.8 7.5 8.4 8.7 10.8 11 10.9 10.1
9.8 9.8 9.3 8.4 7.7 7.1 6.9 6.6 6.5
4.7 4.2 4 3.8 2.5 2.3 1.8 1.1 0.3

The sample mean and sample standard deviation of the consumption are 5.08 liters and 3.26 liters, respectively. A histogram of the consumption is bell-shaped. Modeling your solutions, solve the following problems.

a. Is it reasonable to apply the empirical rule to estimate the percentages of observations that lie within one, two, and three standard deviations to either side of the mean?
b. Use the empirical rule to estimate the percentages of observations that lie within one, two, and three standard deviations to either side of the mean.
c. Use the data to obtain the exact percentages of observations that lie within one, two, and three standard deviations to either side of the mean.
d. Compare your answers in parts (b) and (c).

(c) Refer the following link and conclude your learning from it in maximum five lines. (https://www.linkedin.com/pulse/application-chebyshevs-theorem-swanand-marathe)

T1LA.7

(A) Refer the Lesson 3 of module 6 of the course Statistics and Probability on khan academy.

https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/v/probability-decisions

Submit the quiz/test result proof in pdf form.