Tutorial questions for exploratory data analysis

Law, Crime

1. Customers of a particular bank rated the service provided by the bank on a scale of one to ten, correct to one decimal point. The bank categorised their customers as either (1) Private Account holders or (2) Business Account holders. The information below summarises customer attitudes towards the quality of service provided by the bank. Use the output to answer the questions below.

a) Briefly describe and compare the distributions of results for the two groups of customers.

You should mention appropriate measures of centre and spread, and any other points you feel may be of interest.

b) The survey results were described as being symmetrically distributed. What does this mean, and what evidence is there below to support this claim?

c) The standard deviation of the ratings for the Private Account holders is 1. 336. What does this value mean?

d) Verify the value of the standard error of the mean (SE Mean) for the Business Account holders.

e) Determine a 90% confidence interval for the true mean of the Business Account holders and interpret the result in the context of the situation.

Descriptive Statistics:

Quality Variable Use N N* Mean SE Mean StDev Minimum Q1 Median Q3 Quality 1 45 0 5. 971 0. 199 1. 336 3. 700 4. 800 6. 000 7. 050 2 30 0 8. 323

0. 172 0. 941 6. 200 7. 675 8. 400 9. 025 Variable Use Maximum Quality 1 8. 400 2 10. 000

Stem-and-Leaf Displays: Quality

Solution:

a) Mean for Business group (8. 3) is higher than that for Private group (5. 97). StDev for Private group (1. 336) is larger than for Business group (0. 941), i. e. there is more variation or less consistency in Private group.

No outliers are evident in either distribution and both seem to be quite symmetrical as evidenced by similarities in means and medians.

b) Symmetry implies similar patterns of variation either side of the median – top half is distributed similarly to the bottom half. This is evidenced by similarities in means and medians and also by the appearance of boxplots – to a lesser extent by stem & leaf plots.

c) The stdev measures the spread of the raw data values around the mean. d) SEMean = 0. 941((30 = 0. 1718

e) [pic] 90% CI = 8. 323 (1. 5(0. 1718) = 8. 323 (0. 258 which gives 8. 065 to 8. 581 Accept t29 = 1. 6991 and/or z = 1. 4(5) for full marks – deduct 1 mark for any other values

This means we can be 90% certain the real mean will lie between 8. 065 to 8. 581.

2. Students enrolled in a university statistics unit were given the choice of three tutorial methods to support the lectures in the unit. These were: on-line tutorials, self-paced tutorials and tutorials conducted by a tutor. At the

end of the semester, a random sample of students was selected from each of the three tutorial methods and their final results (%) compared. Use the output provided below to answer the following questions:

a) Which of the three methods gave the best results overall?

Explain your reasoning.

b) The results for the “ tutor” group were described as being skewed. What does this mean, and what evidence is there below to support this claim?

c) Which group has an inter-quartile range of 7. 25? What does this value mean?

d) Briefly comment on the differences between the three groups.

e) Determine a 95% confidence interval for the true mean of the “ self-paced” group, and interpret the result in the context of the situation.

Descriptive Statistics:

Result in Variable Method N N* Mean SE Mean StDev Minimum Q1 Median Q3 Result On-Line 15 0 68. 40 4. 50 17. 45 32. 0 61. 00 65. 00 83. 00 Self 25 0 55. 76 1. 01 5. 05 46. 00 51. 50 56. 00 59. 00 Tutor 20 0 70. 05 1. 13 5. 06 63. 00 66. 25 68. 50 73. 50 Variable Method Maximum Result On-Line 91. 00 Self 65. 00 Tutor 81. 00

Solution:

a)The “ Tutor” group achieved the best results because they had the highest mean (70. 05) as well as a small side (5. 06).

b)A skewed distribution means a non-symmetrical (unevenly distributed) distribution. The “ Tutor” group is positively skewed because the right half of the boxplot covers a wider range than the left half.

Also, the mean is slightly larger than the median.

c)The “ Tutor” group has an IQR of 7. 25 which means that the middle 50% of the data lies within a range of 7. 25 % (marks). The IQR is represented by the box in the boxplot.

d)The “ On-Line” group and the “ Tutor” group had similar averages, and both were considerably higher than the “ Self-paced” group. The “ On-Line” group had the highest and lowest scores and the largest variation (s = 17. 45) and was, therefore, the most inconsistent.

e)[pic]. (Allow t = 2. 06). This means that there is a 95% chance that the true mean for the “ Self-paced” group will lie between these two values.

3. A report on State School class sizes recently reported that government funding is based on a formula of 1teacherper 21 students. An analysis of the data collected from a random sample of 50 classes from State Schools yielded the following information about class sizes. Use this information to answer the questions below.

Descriptive Statistics

Variable N Mean Median Tr Mean StDev SE Mean class 50 21. 802 23. 450 22. 118 5. 464 0. 773 Variable Min Max Q1 Q3 class 9. 000 28. 500 17. 100 26. 650 Confidence Intervals Variable N Mean StDev SE Mean 95. 0 % CI class 50 21. 802 5. 464 0. 773 (20. 249, 23. 355)

a) Briefly describe the distribution of class sizes including mention of sample size, centre, spread and shape.

b) There have been frequent complaints about large classes. Using the information above, discuss whether announcing that ‘ the average class size is 21’ gives a fair impression of class sizes in Victoria.

c) Determine the value of the Inter-Quartile Range for this data.

d) Explain the difference between the standard deviation and the standard error for this data.

e) Briefly explain the meaning of the 95% confidence interval in this context.

Solution:

a)The sample size is 50; the mean class size is 21. 8 which is less than the median 23. 45. This indicates a negative skew, also seen in the boxplot and histogram. No outliers are present. Spread: Classroom sizes range from 9 to 28. 5. The stdev is 5. 464. Shape: Note the large group of classes at the top end of the scale.

b)Not really a fair impression as the data are negatively skewed; the median may be more appropriate. The range would also be of interest here, perhaps more so than measures of centre for new teachers.

c)IQR = 26. 65 – 17. 1 = 9. 55

d)The StDev measures variation in the raw data, the standard error measures variation (expected) in sample means (based upon samples of n = 50). SE Mean = StDev ((n e)There is a 95% chance that the true mean class size is between 20. 249 and 23. 355.

4. The graphs below illustrate the number of cases of a particular virus reported in Victoria in 1995 and 2001 based on 6 geographical regions.

a)Why have bar graphs been used for this data?

b)Which region had the highest incidence of this virus in 2001?

c)Which region(s) had the lowest incidence of the virus in 1995?

d)Which region experienced the greatest change between 1995 and 2001?

e)Express this change as a percentage change.

Solution

a) Bar graphs have been used because the data is categorical in nature.

b) The Central region had the highest incidence of the virus in 2001.

c) The South region had the lowest incidence of the virus in 1995.

d) The largest increase seems to be in the Central Region, however, be careful of the different vertical scales here. Central increased about 12-fold, North increased about 5-fold, but South increased by about 30-fold. So the largest proportional increase was in the South region. e) As a percentage change, this would be about a 3000% increase.

5. Last month there were 8559 persons registered at the CES as unemployed. Of this number 1379 had been unemployed for less than three months, 1800 had been unemployed for more than three months but less than 6 months, 1513 had been unemployed for more than six but less than

nine months, 1220 had been unemployed for between nine and twelve months, and 2647 had been unemployed for more than twelve months. Present this data using a table and then an appropriate graphical display.

Solution: | Months Unemployed | People | |