Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 4 (Data Collection, Sampling and Pre-processing - part 2)

Q1. The first three class marks for a frequency distribution of "weights of college men" recorded to the nearest pound are 105, 115, and 125.
(a) What is the class interval?
(b) What is the lower limit for the third class?
(c) What is the upper limit for the third class?
(d) What are the class limits for the fourth class?

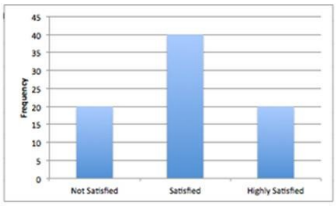
Ans:
(a) The class interval is the difference between the class marks or midpoints of adjacent classes. (115 - 105 = 10, or 125 - 115 = 10).
(b) The class interval is 10. The class marks are the same as the class midpoints. Therefore, the upper and lower limits for a class are the class mark plus or minus half of the class interval. For the third class, the lower limit is 125 - ½(10) = 120.
(c) The class interval is 10. The class marks are the same as the class midpoints. Therefore, the upper and lower limits for a class are the class mark plus or minus half of the class interval. For the third class, the upper limit is 125 + ½(10) = 130.
(d) 130 - 140. The class interval is 10. The class marks are the same as the class midpoints. Therefore, the upper and lower limits for a class are the class mark plus or minus half of the class interval. For the third class, the upper limit is 125 + ½(10) = 130. The upper limit for the third class becomes the lower limit for the fourth class. The class interval, 10, is then added to get the upper limit for the fourth class.

Q2. Refer to the following breakdown of responses to a survey of room cleanliness in a hotel.

Response	Frequency
Not Satisfied	20
Satisfied	40
Highly Satisfied	20

(a) What is the class interval for the following frequency table?
(b) Draw a bar graph that illustrates the preceding frequency table.
(c) Draw a pie chart that illustrates the relative frequencies.

Ans:
(a) There is no class interval. The variable is qualitative. There is no class interval for data measured on an ordinal scale.
(b)



A graph with appropriate labels on the horizontal (satisfaction) and vertical (frequency) axes. The bar for "satisfied" should be twice as high as the "not satisfied and highly satisfied" categories, and these categories should be equal in height.

BACS3013 Data Science

(c)

The pie chart should be divided into three slices. The "satisfied" slice should be one half of the pie, and the "not satisfied" and "highly satisfied" slices should each be one quarter of the pie. The slices should be labeled.

Q3. A data set has 100 observations. In the data, a quantitative variable's highest value is 117 and its lowest value is 47. What is the minimum class interval that you would recommend?

The intermediate answer is 7 classes. The difference between the high and low is 70. So, the class interval is 10.

Note: The class interval would be (maximum - minimum)/number of classes. Using the "2 to the k rule," there would be 7 classes (100 is less than 27 = 128. So (117 - 47)/7 = 10.

Q4. A data set has 200 observations. In the data, a qualitative variable's highest value is "extremely satisfied" and its lowest value is "extremely dissatisfied." What is the minimum class interval that you would recommend?

There is no class interval because the variable is qualitative, not quantitative.

Note: Qualitative data does not have class intervals or class limits.

Q5. The following set of data represents the distribution of annual salaries earned by a random sample of 100 project managers in a country:

Annual salary (£000)	f
0 but under 20	1
20 but under 40	32
40 but under 60	45
60 but under 80	16
80 but under 100	5
100 but under 120	1

Using any appropriate methods, find the mean, median and mode for this distribution and comment on your results.

x	f	fx	fx^2	CF
10	**1**	**10**	**100**	**1**
30	**32**	**960**	**28800**	**33**
50	**45**	**2250**	**112500**	**78**
70	**16**	**1120**	**78400**	**94**
90	**5**	**450**	**40500**	**99**
110	**1**	**110**	**12100**	**100**

total:	**100**	**4900**	**272400**

mean = 49
median = 48
mode = 46

mean > median > mode, the distribution is positively skewed.

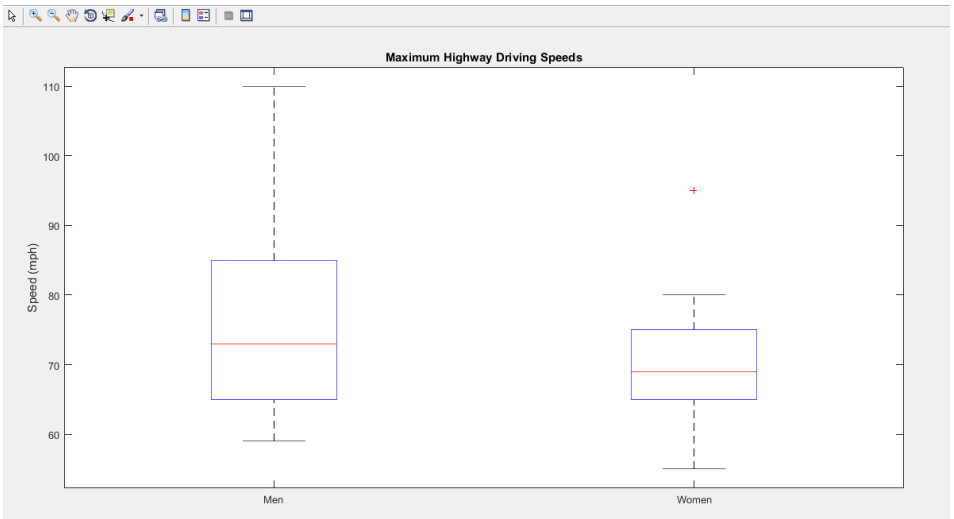
Q6. The maximum highway driving speeds of 10 men and 10 women were recorded.

Men	70	63	90	59	110	75	75	85	65	71
Women	65	70	55	65	68	62	70	95	80	75

Find the five-number summary, and draw the comparative boxplots of the two data sets. Compare the two data sets. (note: Five-number summary consists of the minimum, first quartile, median, third quartile, and maximum. A boxplot is a visual display of the five-number summary.)

Ans:

Men (59,65,73,85,110) Women (55,65,69,75,95). Similar medians but (spread)
IQR is much less in women, note one outlier at 95 for women as per test of 1.5IQR from each quartile)]



Q7. For the following data, are any of the observations an outlier?

1	5	6	6	6	7	7	7	8	8	8
8	8	9	9	9	9	9	9	9	9	

Yes, 1 is an outlier; 1 < 2.75 (6.5 - 1.5*2.5 = 2.75)

BACS3013 Data Science