TMA02

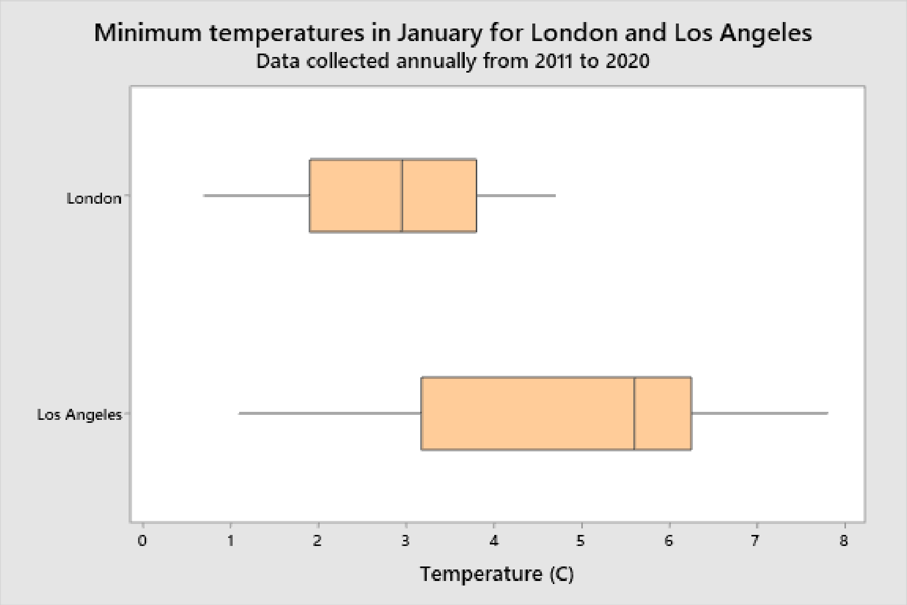
Hefin Rhys (J4342909)

March 29, 2022

Question 1

a) Produce horizontal boxplots of the data for the minimum temperatures in January in London and Los Angeles on the same diagram using Minitab.

Solution:



3/3

b) On the basis of the boxplots obtained in part (a), are the minimum temperatures in January in Los Angeles left-skew, right-skew, or neither?

3/3

Solution:

The distribution of minimum January temperatures in Los Angeles are left-skew. The boxplot shows this because the median is much closer to the 3rd quartile than the 1st quartile, and the lower whisker is longer than the upper whisker. This pattern shows that lower values are more spread out than higher values, and that the distribution is left-skew. ✓

---

c) Using the two boxplots you obtained in (a), compare the minimum temperatures in January in London and Los Angeles.

Solution:

3/3

Total for Q1. 9/9

The median minimum January temperature in London is ≈ 2.7°C lower than the median minimum January temperature in Los Angeles. ✓ The distribution for London is approximately symmetrical, while the distribution for Los Angeles is both left-skew and more variable, having a range of ≈ 6.7°C compared to ≈ 4°C in London, ✓ and an interquartile range of ≈ 3°C compared to ≈ 2°C in London. Despite London typically having lower minimum temperatures in January than Los Angeles, the lowest temperature recorded in this period is similar for both cities. ✓

Question 2

a) Calculate by hand to two decimal places, the mean and standard deviation of the number of goals scored in matches played in February and March 2020.

Solution:

These data represent grouped data, and so I start by constructing a table of values that will facilitate the calculation of the mean and standard deviation of the batch. In the table below, x represents the number of goals and f represents the number of matches where x goals were scored.

x	x^2	f	xf	$x^2 f$
0	0	5	0	0
1	1	10	10	10
2	4	9	18	36
3	9	9	27	81
4	16	8	32	128
5	25	5	25	125
6	36	1	6	36
		$\sum f = 47$	$\sum xf = 118$	$\sum x^2 f = 416$

✓

As shown in the table, the batch size is $\sum f = 47$, and the sum of the ungrouped data values is $\sum xf = 118$. The mean is therefore $118/47 = 2.5106...$ or 2.51 goals per match (to 2 d.p.). ✓

To calculate the standard deviation of the grouped data, I first calculate the sum of squares of the deviations, which is given by the formula

$$\sum(x-\bar{x})^2 f = \sum x^2 f - \frac{(\sum xf)^2}{n}$$

where $\sum x^2 f$ is the sum of the squares of the data values, and n is the batch size. Substituting $\sum x^2 f = 416$, $\sum xf = 118$, and $n = 47$ gives

$$\sum(x-\bar{x})^2 f = 416 - \frac{118^2}{47}$$

$$= 119.7446...$$ ✓

So the sum of squares of the deviations is 119.74 (to 2 d.p.). The variance s^2 is given by

$$s^2 = \frac{\sum(x-\bar{x})^2 f}{n-1}$$

$$= \frac{119.746...}{46}$$

$$= 2.6031...$$ ✓

4/4

So the variance is 2.60 goals^2 (to 2.d.p). The standard deviation s is simply the square root of the variance:

$$\begin{aligned} s &= \sqrt{s^2} \\ &= \sqrt{2.6031...} \\ &= 1.6134... \end{aligned}$$

Therefore, the standard deviation is 1.61 goals (to 2 d.p.). ✓

Total for Q1. 4/4

Question 3

a) For each of May 2020 and June 2020, calculate, as a percentage, the real earnings for that month compared with one year earlier.

Solution:

The real earnings for a month A compared to month B one year earlier is given by

$$\text{real earnings}_A = \frac{\text{AWE}_A}{\text{AWE}_B} \times \frac{\text{CPI}_{\mathbf{B}}}{\text{CPI}_A}$$

where AWE is the Average Weekly Earnings index and CPI is the Consumer Price Index for the appropriate month. Substituting the appropriate values for AWE and CPI from Table 2 gives

$$\text{real earnings}_{\text{May 2020}} = \frac{136.1}{136.6} \times \frac{105.8}{107.5}$$
$$= 0.9805...$$

$$\text{real earnings}_{\text{June 2020}} = \frac{137.2}{136.3} \times \frac{106.3}{107.5}$$
$$= 0.9953...$$

Both calculations inverted.

The real earnings for May 2020 is:

$$\frac{107.5}{105.8} \times \frac{136.6}{136.1} = 1.019801 \simeq 102.0\%$$

The real earnings for June 2020 is:

$$\frac{107.5}{106.3} \times \frac{136.3}{137.2} = 1.004655 \simeq 100.5\%$$

0/2

So the the real earnings in May and June 2020 were 98.1% and 99.5% compared to one year earlier (as percentages to 1 d.p.), respectively.

---

b) Were workers in the construction sector of the economy generally better or worse off in May and June of 2020 than they were a year earlier?

2/2

Solution:

As the real earnings were less than 100% compared to the year before, Workers were generally worse off in May and June 2020 compared to one year earlier. This is because there was a larger increase in CPI over this period than the increase in AWE. This indicates the cost of living has increased faster than wages in this period, and therefore real earnings have decreased.

✓ Correct conclusion for your results in a).

Total for Q3. 2/4

Question 4

a) Use the pairs of digits in row 4 (from left to right) of the random number table to identify which of the first 10 people listed in the table you should use as the starting point for your systematic sample.

Solution:

The first pair of random digits between 01 and 10 on line 4 of the random number table is 03, so the first person in the sample will be Archer, Simon with Label 03.

✓ 1/1

b) Select the rest of the sample and write down the labels of the people selected in the order that they were selected.

Solution:

Using 03 as the random start and selecting every 10th individual gives the following sample labels in the order they were selected: 03, 13, 23, 33, 43.

1 tenth of 86 is 8.6 so you should obtain 8 staff numbers. So also, 53, 63, 73, 83.

✗ 0/1

c) List the names of all the sampled individuals in the order in which they were chosen, along with their gender and occupation.

Solution:

The sampled individuals are listed in the following table in the order in which they were chosen (from top to bottom).

Name	Label	Declared gender	Occupation
Archer, Simon	03	M	M
Cameron, Lynne	13	F	P
Daley, Stuart	23	M	P
Gowan, Dai	33	M	P
Hewitt, Ray	43	M	P

✓ 3/3

d) Comment on the representativeness of the sample with respect to occupation, assuming that the full staff list is the population.

Solution:

The sample is not a good representation of the distribution of occupations within the population. Only Professional and Manual occupations are present in the sample, despite Administrative and Secretarial occupations constituting $\approx 10\%$ of the population each. However, the Professional occupation is the most frequent occupation in both the sample and the population, making up 80% in the former and $\approx 72\%$ in the latter.

2/3

Given the very small sample size it is fairly representative.

Total for Q4. 6/8

(a) *Unit 4 Subsection 3.2*

The number of respondents is 356, so the median is the $(356+1)/2 = 178.5$, halfway between the 178th and 179th values.

Question 5

a) What is the media response for this population? Show your working.

Solution:

The total number of responses to the survey is 356, therefore the median rating is the rating of the $\frac{356}{2} = 178$th response. Starting from ratings of 1, the cumulative sum of the number of responses with a rating of 1 or 2 is 112, and is 279 for a rating of 1, 2, or 3. Therefore, the 178th response has a rating of 3, so the median response is "Neutral". ✓

1/2

b) Giving your reasons, say which figure (A, B, or C) relates to which sample (1, 2, or 3).

Solution:

Sample 1 corresponds to figure B and sample 2 corresponds to figure A. ✓ This is because the median response in the population was 3, and both of these figures show a sampling distribution of the median consistent with the 3 being the population value. As the samples that make up sample 2 are larger than those that make up sample 1, it is more likely to be represented in figure A where the proportion of sample medians equal to the population median is higher (0.840) than in figure B (0.747). ✓

Sample 3 corresponds to figure C because despite the samples being larger in size than sample 1, the most frequent sample median is different to the population median. ✓

6/6

Total for Q5. 7/8

Question 6

a) Interpret the scatterplot in Figure D giving reasons for your answers.

Solution:

i) There is a positive relationship between house prices and GDHI because generally, as one increases, so does the other. ✓

ii) The relationship appears linear as it could be summarised reasonably well by a straight line. ✓

iii) The relationship appears strong as all the points lie close to a line (such as a least squares line). ✓

iv) The datum representing London could be considered an outlier as its house prices and GDHI are both unusually large compared to the rest of the batch (although it would still lie close to a least squares line fit through the data). The datum representing Scotland could also be considered an outlier, not because it house prices or GDHI are unusual in isolation, but because Scotland has an ✓

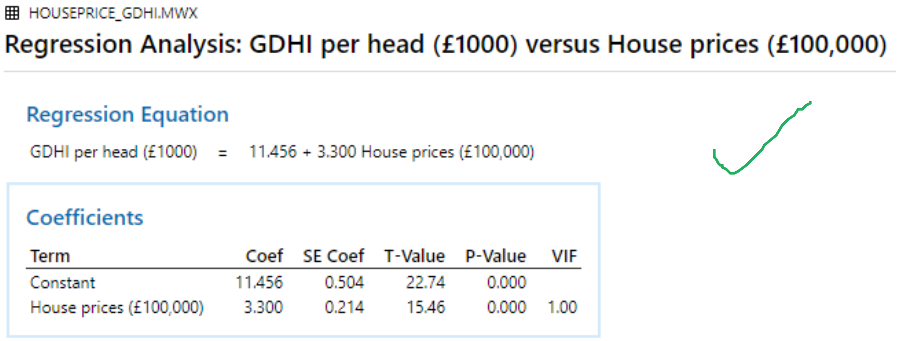
8/8

As London is roughly on the regression line, just much further along from the other data we actually call it an influencer rather than an outlier. Something we'll look at later in the unit. Well observered!

unusually high GDHI for its house prices, placing it further from a least squares line than the other data.

b) Use Minitab to fit a least squares regression line to the data. Include the appropriate Minitab output to show the regression equation in your answer. What is the equation of the least squares regression line in terms of x and y?

Solution:



HOUSEPRICE_GDHI.MWX

Regression Analysis: GDHI per head (£1000) versus House prices (£100,000)

Regression Equation

GDHI per head (£1000) = 11.456 + 3.300 House prices (£100,000)

✓

Coefficients

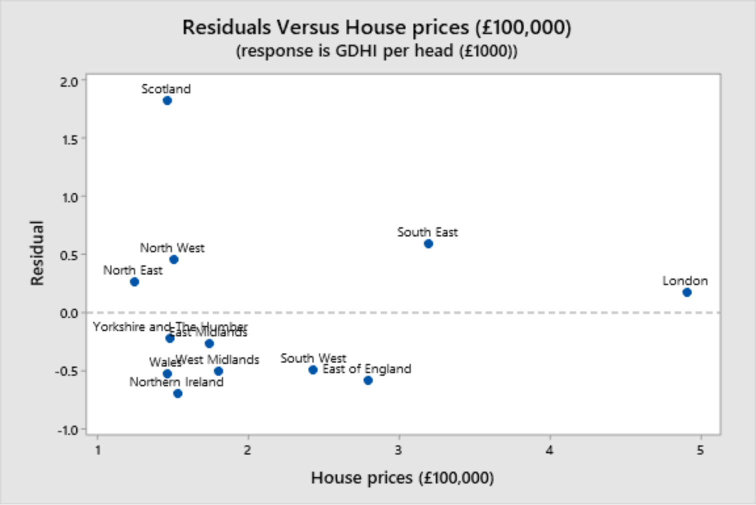
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	11.456	0.504	22.74	0.000	
House prices (£100,000)	3.300	0.214	15.46	0.000	1.00

2/2

The equation of the least squares regression line is $y = 11.456 + 3.300x$ where x and y are House prices (£100,000) and GDHI per head (£1000), respectively. ✓ ✓

c) Produce a residual plot using Minitab. Comment on what the residual plot tells you.

Solution:



✓

4/4

The residual plot shows no relationship between the residuals of the least squares fit, and the explanatory variable, suggesting the straight line model gives an adequate explanation of the patterns in the data. The residual plot also supports the suggestion that Scotland is an outlier in this batch as its residual has a greater magnitude than any of the other data. ✓

---

d) If the average house price in a region is £200,000, estimate the GDHI per head of population for that region.

Solution:

The least squares model is GDHI per head (£1000) = 11.456+3.300×House prices (£100,000). Substituting a house price of 2 (£100,000) gives

2/2

$$\begin{aligned}\text{GDHI per head (£1000)} &= 11.456 + 3.300 \times 2 \\ &= 18.056\end{aligned}$$

✓

Therefore, based on this model, a region with an average house price of £200,000 is predicted to have a GDHI per head of £18,056. ✓

---

e) If the average house price in a region fell to £90,000, would it be appropriate to use the equation of the least squares regression line from part (b) to estimate the GDHI per head of the population for that region?

1/1

Solution:

It would not be appropriate to use the equation of the least squares regression to predict the GDHI per head for this region, because it would require extrapolating beyond the range spanned by the explanatory variable when the model was trained. ✓

Total for Q6. 17/17

Great work