TMA03

Hefin Rhys (J4342909)

July 10, 2022

Question 1

a) Calculate the row, column, and overall totals for Table 1, presenting them as part of the table.

Solution:

	Colour				
Type	White	Beige	Navy blue	Black	Total
Coat	4	1	2	6	13
Sweater	3	0	3	5	11
Total	7	1	5	11	24

---

b) i) Calculate the probability that she selected a sweater.

Solution:

The probability of selecting a sweater is

$$\begin{aligned} P(\text{sweater}) &= \frac{\text{number of sweaters}}{\text{number of items}} \\ &= \frac{11}{24} \end{aligned}$$

So the probability of selecting a sweater is $\frac{11}{24}$ or 0.458 (to 3 d.p.).

---

ii) Using the probability rule for complementary events, calculate the probability that the selected item isn't a black coat.

Solution:

The probability rule for complementary events states that if event A is complementary to event B, then $P(A) = 1 - P(B)$. For example, if the probability of selecting a black coat is

$$P(\text{black coat}) = \frac{6}{24}$$

then the probability of selecting an item that isn't a black coat is

$$\begin{aligned} P(\text{not a black coat}) &= 1 - \frac{6}{24} \\ &= 0.75 \end{aligned}$$

(iii) Calculate the probability that the selected item is dark coloured.

Solution:

Taking navy blue and black items to be "dark coloured", the probability of selecting a dark coloured item is

$$\begin{aligned} P(\text{sweater}) &= \frac{\text{number of navy blue items + number of black items}}{\text{number of items}} \\ &= \frac{5+11}{24} \\ &= \frac{2}{3} \end{aligned}$$

So the probability of selecting a dark coloured item is $\frac{16}{24}$ or 0.667 (to 3 d.p.).

---

c) Calculate the probability that they both pick the same type of item, regardless of colour.

Let C and S represent the events of choosing a coat and a sweater, respectively. As the choices are independent of each other, the probability of both friends choosing a sweater is

$$P(SS) = \frac{11}{24} \times \frac{11}{24} = \frac{121}{579}$$

and the probability of both friends choosing a coat is

$$P(CC) = \frac{13}{24} \times \frac{13}{24} = \frac{169}{576}$$

As both of these events are mutually exclusive, the probability of both friends picking the same type of item regardless of colour is

$$P(SS \text{ or } CC) = \frac{121}{579} + \frac{169}{576}$$

$$= \frac{145}{288}$$

$$= 0.503 \text{ (to 3 d.p.)}$$

Question 2

a) Considering the data in Table 2, does this value seem plausible to you? Justify your answer.

Solution:

Considering fluctuation in the economy over a year (the data in Table 2 are for 2018), sampling error, and response bias, a population median income in 2017 of £18,000 seems plausible given the estimated income of England in 2018 is £21,609.

---

b) For each of the English regions, check whether the estimated household income lies above or below £18,000. Ignoring the aggregate value for England, write down the number of regional values lying above £18,000 and the number of regional values lying below £18,000.

Solution:

Place	Estimated income (£)	Sign
North East	16 995	-
North West	18 362	+
Yorkshire and The Humber	17 665	-
East Midlands	18 277	+
West Midlands	18 222	+
East of England	22 205	+
London	29 362	+
South East	24 318	+
South West	20 907	+

As shown in the table, there were 7 regions with values above £18,000 and 2 regions with values below £18,000.

---

c) Based on these numbers, what is the value of the test statistic?

Solution:

The test statistic is the smaller of: the number of values above the value of interest or the number of values below the value of interest. In this case, the test statistic is 2, as there are only two regions with values smaller than £18,000.

---

d) What is the appropriate critical value at the 5% significance level?

Solution:

As the size of the sample is 9, the critical value at the 5% significance level is 1.

---

e) Decide whether or not you would reject the hypothesis at the 5% significance level.

Solution:

As the test statistic (2) is not less than or equal to the critical value (1), then we cannot reject the null hypothesis at the 5% significance level.

---

f) Show by hand that the p-value given by the hypothesis test is 0.180.

Solution:

The probability of observing x values above the median in a sample of size n is

$$ {}^{n}C_{x} \times \left(\frac{1}{2}\right)^{n} $$

and the two-tailed probability of observing a test statistic as extreme or more extreme if the null hypothesis were true is

$$ p = 2 \times \sum_{i=0}^{x} {}^{n}C_{i} \times \left(\frac{1}{2}\right)^{n} $$

which in this case expands out to

$$p = 2 \times \left[\left({}^{9}C_0 \times \left(\frac{1}{2}\right)^9 \right) + \left({}^{9}C_1 \times \left(\frac{1}{2}\right)^9 \right) + \left({}^{9}C_2 \times \left(\frac{1}{2}\right)^9 \right) \right]$$

$$= 2 \times \left[\frac{1}{512} + \frac{9}{512} + \frac{36}{512} \right]$$

$$= 2 \times \frac{46}{512}$$

$$= 0.1796...$$

Therefore, the p value given by the hypothesis test is $p = 0.180$ (to 3 d.p.).

---

g) By looking at this p-value, and using Table 10 of Unit 6 (Subsection 5.1), what conclusion can be drawn from the hypothesis test?

Solution:

The conclusion that can be drawn form this hypothesis test is that there is little evidence against the null hypothesis that the population median income for England in 2018 was different from £18,000.

---

h) How does this conclusion sit with the result of part (e)?

Solution:

The conclusion stated in (g) is concordant with the decision not to reject the null hypothesis at the 5% significant level.

---

i) Based on these results, what should the economist conclude? Make specific references to the context and consider whether the sign test was appropriate.

Solution:

Based on the results, the economist should conclude that it is quite plausible that the population median income for English regions is £18,000. As the population median income for England in 2017 was £18,000, this result suggests it is also plausible that the gross disposable household income did not change in England between 2017 and 2018. However this analysis only allows us to draw conclusions based on the household income in England as a whole. It is quite possible that some regions were worse off, and some better off, but that this balanced out

when the data are reduced to signs above and below the reference value of £18,000.

It is also possible that all or many of the regions experienced a small but consistent change in household income, but not enough to change the sign when compared to the reference value. In this scenario, the sign test is insensitive to detecting such small changes, leading to a type 2 error on the behalf of the economist.

Question 3

a) For the normal distribution shown in Figure 1, find approximate values for its mean and standard deviation. Explain how you obtained your answers.

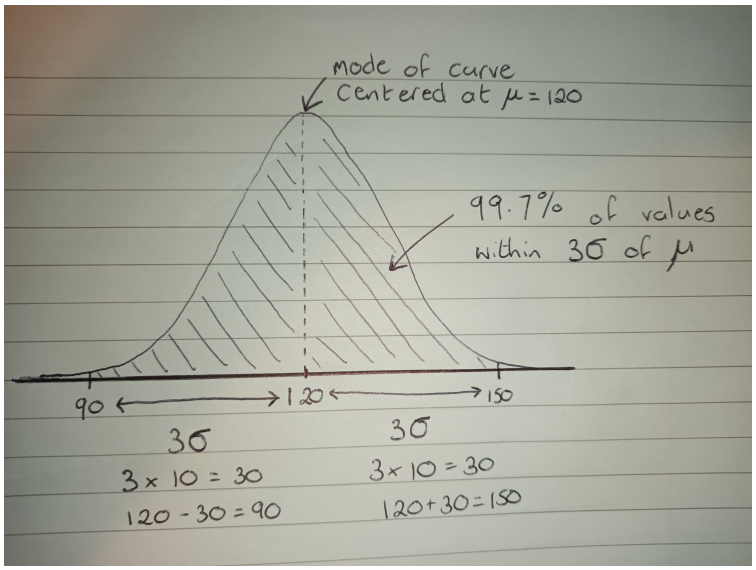
Solution:

The mean of the normal distribution is approximately -40, as this is the value the bell curve is centered on. The standard deviation is approximately 5, as 99.7% of the values distributed according to a normal distribution are within 3 standard deviations of the mean and almost all of the distribution lies on the interval $[-55, -25]$ or $[-40 - 15, -40 + 15] = [\mu - 3\sigma, \mu + 3\sigma]$.

---

b) i) Sketch this distribution by hand (i.e. not using any computer software). Explain and show any calculations used to determine the key features of the distribution.

Solution:



ii) Write down the formula for z that converts each value of the variable x so that z follows a standard normal distribution. Your answer should contain both the general formula and the specific formula for this example.

Solution:

The formula for scaling and centering any normally distributed variable x such that it follows a standard normal distribution is

$$z = \frac{x - \mu}{\sigma}$$

where z is the scaled and centered value, x is the original value, and μ and σ are the mean and standard deviation of the original distribution, respectively. Therefore, the specific formula for this example is

$$z = \frac{x - 120}{10}$$

---

iii) Calculate the value of z corresponding to $x = 90$.

Solution:

Substituting $x = 90$ into the formula stated above gives

$$z = \frac{90 - 120}{10}$$

$$= -3$$

So the value of z corresponding to $x = 90$ is -3.

---

iv) Interpret the value of z by completing the sentence.

Solution:

The value of $x = 90$ is 3 standard deviations below its mean.

---

v) Suppose you are obtaining a sample of size 64 from this distribution. What will be the standard deviation of the sampling distribution of the mean?

Solution:

The standard deviation of the sampling distribution of the mean, also called the standard error of the mean, is given by

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size. Substituting $\sigma = 10$ and $n = 64$ gives

$$\text{SE} = \frac{10}{\sqrt{64}}$$

$$= 1.25$$

So the standard deviation of the sampling distribution of the mean is 1.25.

Question 4

a) i) Using appropriate notation, which you should define, specify the null and alternative hypotheses associated with the test.

Solution:

The null (H_0) and alternative (H_1) hypotheses are:

H_0: the population density among Council Areas of Scotland was equal to 430 people per square km in 2018.

H_1: the population density among Council Areas of Scotland was not equal to 430 people per square km in 2018.

---

ii) According to the Minitab results, what are the mean and standard deviation of this data? Do not forget to mention units corresponding to the values.

Solution:

The mean and standard deviation reported by Minitab are 653 people per Km and 962 people per Km, respectively. A screenshot of the Minitab output is shown below.

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Density	20	0	653	215	962	9	44	270	683	3586

iii) Using these numbers, use Minitab to obtain the test statistic z corresponding to the aforementioned hypothesis. Include a copy of the relevant Minitab output in your answer and state the z-value.

Solution:

The test statistic z corresponding to the hypothesis is 1.04. A screenshot of the Minitab output is shown below.

Test

Null hypothesis	H_0: $\mu = 430$
Alternative hypothesis	H_1: $\mu \neq 430$

Z-Value	P-Value
1.04	0.301

---

iv) Compare z with the appropriate critical values (in Subsection 5.2 of Unit 7) to say what can be concluded from the z-test in terms of evidence about the hypotheses. Make specific references to the context.

Solution:

The absolute value of the z statistic is <1.96, so the null hypothesis cannot be rejected at the 5% significance level. Based on this result, there is little evidence to suggest the population density among Council Areas of Scotland was not equal to 430 in 2018.

---

v) Should there be any reservations about the use of this test here?

Solution:

The sample size is just 20, which is lower than is recommended for the z-test, especially as the distribution of values is considerably skew, looking at the mean, median, and quartiles. It's likely that the estimated standard error of the mean used in the calculation of the z statistic is not a good estimate of the population standard error of the mean.

---

b) i) Use these numbers to calculate by hand the estimated standard error (ESE) necessary for the two-sample z-test, following Subsection 6 of Unit 7.

Solution:

The estimated standard error of the difference between two means is

$$\text{ESE} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

where s_a and s_b are the sample standard deviations of groups a and b, respectively, and where n_a and n_b are the sample sizes for groups a and b, respectively.

Substituting $s_a = 962$, $n_a = 20$, $s_b = 572$, and $n_b = 20$ gives

$$\text{ESE} = \sqrt{\frac{962^2}{20} + \frac{572^2}{20}}$$

$$= \sqrt{46272.2 + 16359.5}$$

$$= 250.2627$$

So the ESE for z test is 250.26 (to 2 d.p.).

---

ii) Regardless of any reservations you may have about using the z-test, calculate by hand the test statistic z for the two-sample z-test.

Solution:

The z statistic for a two-sample z test is given by

$$z = \frac{\bar{x}_a - \bar{x}_b}{\text{ESE}}$$

where $\bar{x}_a$ and $\bar{x}_b$ are the sample means of groups a and b, respectively, and ESE is the estimated standard error of the difference between the means of groups a and b.

Substituting $\bar{x}_a = 653$, $\bar{x}_b = 447$, and ESE = 250.2626... gives

$$z = \frac{653 - 447}{250.2626...}$$

$$= 0.8231....$$

So the z statistic for the two-sample z test is 0.82 (to 2 d.p.).

Question 5

a) Write the information given in the two bullet points above in symbolic form, using the notation of the two events mentioned.

Solution:

The probability of seeing at least one goldfinch is $P(G) = 0.15$. On walks where they see at least one goldfinch, the probability of seeing at least one robin is $P(R|G) = 0.4$.

---

b) From these two probabilities, calculate the probability that on a randomly chosen walk they see at least one goldfinch and at least one robin.

Solution:

The joint probability of G and R is given by

$$\begin{aligned} P(G \text{ and } R) &= P(G) \times P(R|G) \\ &= 0.15 \times 0.4 \\ &= 0.06 \end{aligned}$$

So the probability of seeing at least one goldfinch and at least one robin is 0.06.

---

c) Additional information is now given that they see at least one robin on 45% of their walks. What is the probability that on a randomly chosen walk they do not see any robin?

Solution:

If the probability of seeing at least one robin is 0.45, then the complementary probability of not seeing any robin is 1 - 0.45 = 0.55.

---

d) On a walk where they see at least one robin, what is the probability that they see at least one goldfinch?

Solution:

The probability of seeing at least one goldfinch, given they have seen at least one robin is given by

$$P(G|R) = \frac{P(G \text{ and } R)}{P(R)}$$

$$= \frac{0.06}{0.55}$$

$$= 0.1090...$$

So the probability of seeing at least one goldfinch where they see at least one robin is 0.109 (to 3 d.p.).

---

e) Calculate the probability that on a randomly selected walk they see either at least one goldfinch, or at least one robin, or both.

Solution:

The probability that any of these three events will occur is given by

$$\begin{aligned} P(G \text{ or } R \text{ or } (G \text{ and } R)) &= P(G) + P(R) + P(G \text{ and } R) \\ &= 0.15 + 0.45 + 0.06 \\ &= 0.66 \end{aligned}$$

So the probability that they see either at least one goldfinch, at least robin, or both is 0.66.

---

f) Are the events corresponding to seeing at least one goldfinch and at least one robin independent? Give a reason for your answer.

Solution:

The events G and R are statistically independent if $P(G|R) = P(G)$ and $P(R|G) = P(R)$. As $P(R|G) = 0.4$ and $P(R) = 0.45$, these events are not independent.

Question 6

a) Check whether the table is a contingency table by checking whether it meets all three criteria. Explain your reasons in the context of this investigation.

Solution:

The table is a contingency table for the following reasons. Both the row and column variables are categorical (hair shape and country are not continuous variables). The categories for both variables are mutually exclusive (a person has only a single hair shape and comes from one country). The entry in each cell of the table is a count of the number of subjects with that combination of hair shape and country of origin.

---

b) Specify the null and alternative hypotheses associated with the χ^2 test.

Solution:

The null (H_0) and alternative (H_1) hypotheses are:

H_0: hair shape and country of origin are independent.

H_1: hair shape and country of origin are not independent.

---

c) Perform the χ^2 test using Minitab. Include a copy of the Minitab output in your answer showing the table of observed and expected frequencies.

Solution:

Rows: Country Columns: Worksheet columns

	Straight	Wavy	Curly	All
Brazil	654	585	308	1547
	705.6	601.4	239.9	
Chile	894	532	94	1520
	693.3	590.9	235.7	
Colombia	603	639	383	1625
	741.2	631.8	252.0	
Mexico	740	678	195	1613
	735.7	627.1	250.2	
Peru	394	366	137	897
	409.1	348.7	139.1	
All	3285	2800	1117	7202

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	284.417	8	0.000
Likelihood Ratio	296.837	8	0.000

d) Calculate by hand the expected value for the number of people with curly hair in Brazil.

Solution:

The expected frequency of people from Brazil with curly hair is the product of the total number of people from Brazil and the total number of people with curly hair, divided by the total number of people in the study. This gives:

$$E(\text{Brazil and Curly}) = \frac{1547 \times 1117}{7202}$$

$$= 239.9332$$

So the expected frequency of people from Brazil with curly hair is 239.9 (to 1 d.p.).

---

e) Explain why the χ^2 test that you performed in part (c) is valid in this case.

Solution:

The χ^2 test is valid in this case because all of the expected counts are considerably greater than 5.

---

f) Minitab gives the degrees of freedom associated with this χ^2 test as 8. Show how this value arises.

Solution:

The number of degrees of freedom for a χ^2 test is $(n-1)(m-1)$, where n and m and the number of rows and columns of the contingency table, respectively. In this case we have

$$\begin{aligned} \text{df} &= (5-1)(3-1) \\ &= 4 \times 2 \\ &= 8 \end{aligned}$$

Hence, 8 degrees of freedom.

---

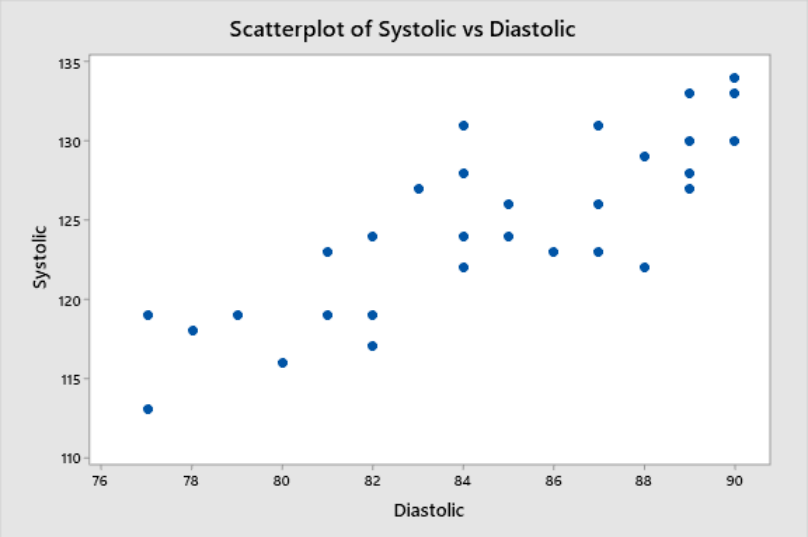
g) Interpret the results of the χ^2 test, making specific references to the context.

Solution:

The p value is < 0.01 and so the null hypothesis that country and hair shape are independent can be rejected as being highly unlikely. While the result doesn't indicate causality, a person's hair shape is unlikely to influence where they are from, but the country a person is from is likely to influence their hair shape.

Question 7

a) Use Minitab to make a scatterplot of Systolic as the response variable, y, against Diastolic as the explanatory variable, x.



---

b) Is the correlation coefficient between the two variables likely to be positive or negative? What is your best guess for the value of the correlation coefficient? Explain your answers

Solution:

The correlation coefficient is likely to be positive because, subjects with a high Diastolic value tend to have a high Systolic value (generally as one increases so does the other). My best guess for the value of the correlation coefficient is 0.8 as the relationship is positive and very strong (all the points lie close to a line through the data), but is not a 1:1 relationship, so the coefficient must be close to but less than 1.

---

c) Suppose you had been asked instead to produce a scatterplot with Systolic as the explanatory variable, x, and Diastolic as the response variable, y. Would this change the sign of the correlation coefficient? Explain your answer.

Solution:

Swapping the axes would not change the magnitude or sign of the correlation coefficient. This is because the correlation coefficient doesn't have a dependent or independent variable, and doesn't depend on how the variables are plotted.

---

d) Use Minitab to calculate and report the correlation coefficient between the two variables. You do not need to include a copy of the Minitab output in your answer.

Solution:

Minitab calculated the correlation coefficient between the two variables to be 0.816.

---

e) By looking at the scatterplot from part (a), do you think there are any outliers? Explain your answer.

Solution:

I would not consider any of the observations to be outliers because none sit away from the main pattern of the data, and none are remote from the rest of the observations.

---

f) Would an increase in Diastolic necessarily cause an increase in Systolic?

Solution:

An increase in Diastolic would not necessarily cause an increase in Systolic because pairs of observations can be found for which one has a higher value for Diastolic but a lower value of Systolic.

---

g) If all patients in the sample had received this (hypothetical) treatment, would the correlation between the resulting systolic and diastolic blood pressure values decrease by 10%, by 7.5%, or by 5% of the original correlation value, or would it remain the same as the original correlation value? Explain your answer.

Solution:

The correlation coefficient would stay the same because it does not depend on the scale of the individual variables but on the relationship between them. As these transformations simply linearly scale each variable independently of each other, there is no change in the correlation coefficient.

Question 8

a) i) Using the 1-Sample Z... option in Minitab, calculate and report a 95% confidence interval for the population mean of the height of adult men of Peru. In addition, include a copy of the Minitab output in your answer.

N	Mean	SE Mean	95% CI for μ
599	171.170	0.256	(170.668, 171.672)

μ: mean of Sample
Known standard deviation = 6.27

Minitab calculated the 95% confidence interval for the mean height in cm for adult men of Peru to be (170.668, 171.672).

---

ii) Interpret this confidence interval in terms of all possible random samples of heights of adult men of Peru.

Solution:

About 95% of the possible random samples we could select will give rise to an interval containing the population mean of adult men's heights from Peru, while only about 5% of the possible random samples we could select will give rise to an interval that does not contain the population mean of adult men's heights from Peru.

---

b) i) Using appropriate notation, which you should define, specify the null and alternative hypotheses that would be associated with this test.

Solution:

The null (H_0) and alternative (H_1) hypotheses are:

H_0: The mean heights of adult men and women from Peru are equal.

H_1: The mean heights of adult men and women from Peru are not equal.

---

ii) Calculate by hand the value of the estimated standard error (ESE) of the difference between the sample means.

Solution:

Let s_F and s_M be the sample standard deviation of the Female and Male groups, respectively, and let n_F and n_M be the sample sizes of the Female and Male groups, respectively. Then the estimated standard error of the difference between the two means is

$$\text{ESE} = \sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}}$$

$$= \sqrt{\frac{6.67^2}{866} + \frac{6.27^2}{599}}$$

$$= 0.3126...$$

So the estimated standard error of the difference between the sample means is 0.313 (to 3 d.p.).

---

iii) Calculate by hand a 99% confidence interval for the difference between the population mean between men and women.

Solution:

The 99% confidence interval for $\mu_{\text{Male}} - \mu_{\text{Female}}$ is

$$\begin{aligned}
&= (\bar{x}_{\text{Male}} - \bar{x}_{\text{Female}} - 2.58 \times \text{ESE}, \bar{x}_{\text{Male}} - \bar{x}_{\text{Female}} + 2.58 \times \text{ESE}) \\
&= (171.17 - 158.34 - 0.8065..., 171.17 - 158.34 + 0.8065...) \\
&= (12.0234..., 13.6365...)
\end{aligned}$$

So the 99% confidence interval for the difference between the population means of men and women's height in Peru is (12.02, 13.64) to 2 d.p.

---

iv) On the basis of this confidence interval, what can you conclude in terms of evidence about the hypotheses? Interpret the results with specific reference to the context of the population mean heights of adult men and women in Peru.

Solution:

As the 99% confidence interval does not include the value of 0, we can reject the null hypothesis at the 1% significance level. This tells us there is strong evidence that the heights of men and women from Peru are not equal, and that mean are, on average, taller.