MONASH University

James Maurice Apel

Started on	Monday, 3 November 2025, 10:00 AM
State	Finished
Completed on	Monday, 3 November 2025, 11:41 AM
Time taken	1 hour 41 mins
Grade	**16.50** out of 50.00 (**33**%)

Information

You can review your exam instructions by clicking the 'Show Instructions' button above.

Unsure **Response history**

Information

There are some long questions in this exam. It may help to use the split screen functionality for these questions. Click on the  to turn split screen on and off.

Unsure **Response history**

Information

Attempt every question in this exam. The total marks allocated is 50 marks

Unsure **Response history**

Information

Getting help and support

Assessments without online supervision

If you have any difficulties (technical, medical or otherwise), call the Assessment Support hotline.

If your internet connection drops out during your assessment, don't panic – your responses are automatically saved every 30 seconds.

Assessment Support
Australia: +61 3 9905 4300
Malaysia: +60 3 5514 5600
China: +800 666 274 73

Assessments with online supervision

If you have any difficulties (technical, medical or otherwise) use the Raise Hand button or webchat function to alert your online supervisor.

If you've lost your connection to your online supervisor, call the Assessment Support hotline.

Raise hand

Unsure **Response history**

Information

Formula Sheet

Probability	
Conditional Probability	$P(A\|B) = \frac{P(A \cap B)}{P(B)}$
Statistical Independence	$P(A \cap B) = P(A).P(B)$

Descriptive Statistics	
Mean	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Median (odd) position	$\frac{n+1}{2}$
Median (even) position	$\frac{\frac{n}{2}+\left(\frac{n}{2}+1\right)}{2}$
Range	Max - Min
Interquartile Range	Q3 – Q1
Variance	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
Standard deviation	$\sqrt{\text{Variance}}$
Z-score	$Z_{population} = \frac{x-\mu}{\sigma}$ $Z_{sample} = \frac{x-\bar{x}}{s}$

Confidence Intervals	
Confidence Intervals (assume Z for all questions in exam)	$\bar{x} \pm Z^* \times \frac{s}{\sqrt{n}}$ • z* = 1.96 for 95% confidence • z* = 2.58 for 99% confidence

Correlation and Regression	
Covariance	$Cov(X,Y) = \frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})$
Pearson Correlation	$\frac{Cov(X,Y)}{s_X s_Y}$
Linear Regression	$y = \beta_0 + \beta_1 X_1 \ldots \beta_k X_k + \epsilon$
t statistic	$\frac{\beta}{SE(\beta)}$
R^2	$\frac{SS_{Regression}}{SS_{Total}}$
Adjusted R^2	$1 - \left(\frac{MS_{residual}}{MS_{total}}\right)$; where: • $MS_{residual} = \frac{SS_{residual}}{df_{residual}}$ $MS_{residual} = \frac{SS_{total}}{df_{total}}$

Classification and Prediction	
Classification accuracy (overall)	$\frac{TP+TN}{TP+TN+FP+FN}$; TP = True positive, TN = True Negative FP = False Positive, FN = False Negative
False Positive (Type I) Rate	$\frac{FP}{FP+TN}$
False Negative (Type II) Rate	$\frac{FN}{FN+TP}$
Positive Predictive Value (Precision)	$\frac{TP}{TP+FP}$
Sensitivity (Recall)	$\frac{TP}{TP+FN}$

Unsure

Response history

Information

The questions in this exam all relate to the scenario described below:

A retail analytics company is studying shopping behaviour at a large Melbourne mall. Over the course of two weeks, the company collects data from a sample of shoppers (via a survey) who happened to be at the mall and were willing to participate when approached.

For each participating shopper they completed a survey that contained the following

For each participating shopper, they completed a survey that contained the following information:

- Whether they used a shopping list (1 = Yes / 0 = No)
- Whether they spent more than $100 (1 = Yes / 0 = No)
- Their total spending (in dollars)
- Age (in years)
- Household income (in $1,000s per year)
- Distance from the mall (in km)
- Loyalty program membership (1 = Yes / 0 = No)

In total, there were 400 survey responses collected during this time period.

Unsure **Response history**

6.5 Marks

Suppose Table 1 below represents a contingency table for the first two items of the survey:

Table 1. Survey results for "Used a shopping list" and "Spent more than $100"

Shopping List	Spent > $100	Spent < $100	Total
Used a list	80	120	200
Did not use a list	140	60	200
Total	220	180	400

1a) 0.5 Marks

What proportion of all shoppers spent more than $100?

0.55

Resubmit to Turnitin

Report question issue

1b) 0.5 Marks

If a shopper used a shopping list, what is the probability they spent more than $100?

Pr(spent more than 100 | used a shopping list)=

80/200=0.4

Resubmit to Turnitin

Report question issue

1c)

0.5 Marks

If a shopper spent more than $100, what is the probability they used a shopping list?

Pr(used a shopping list | more than 100)

80/220=0.3636

Resubmit to Turnitin

Report question issue

1 of 6

1d)

2 Marks

Are 'Using a shopping list' and 'Spending more than $100' independent? Justify with calculations.

Pr (intersection of used a shopping list and spending more than 100) = 0.2

Pr (A) x Pr(B) = 0.5 x 0.55 = 0.275

Using a shopping list and spending more than 100 is not independent due to 0.2 not equalling 0.275.

Turnitin ID: 2801989752

0%

Report question issue

1e)

3 Marks

The marketing team claims that 'shopping lists reduce overspending (where overspending is defined as spending > $100).' Does the table support this claim? Explain using probabilities.

The marketing team's claim is generally correct in this instance as with the information provided using a shopping list does reduce over spending (spending more than 100 dollars). Exactly 50% of the study uses a shopping lists and the other 50% don't use a shopping list thus making the comparisons much easier. 70% of the group that don't use shopping lists overspend compared to just 40% of people that use shopping lists overspending. Thus, only 30% of non shopping list users don't overspend compared to 60% of people with shopping lists who underspend. Therefore we can conclude that using a shopping list will increase your probability of saving money when shopping but it doesn't guarantee that you will reduce your overspending.

Turnitin ID: 2801990034

0%

Report question issue

Report question issue Notes

Question 1 Notes

Comment:

Unsure **Response history**

9 Marks

Table 2 below summarises total spending (in dollars) by shoppers in the sample, separated by whether they used a shopping list.

Table 2. Descriptive Statistics of total spending ($)		
Statistic	**Used a List**	**No List**
Mean	98.5	126.3
Median	95.0	120.0
Standard Deviation	28.4	40.5
Minimum	45.0	60.0
Maximum	185.0	240.0
IQR	35.0	50.0

2a)

2 Marks

Interpret the mean and median spending separately for shoppers with and without a shopping list.

Of shoppers who shopped with a shopping list, they averaged spending $98.50 per trip with a median of all the figures of $95.00

Compared to shoppers without a shopping list who average $126.30 per shopping trip with a median of those figures of $120.00 per trip.

Turnitin ID: 2801990159

0%

Report question issue

2b)

2 Marks

Compare the mean and median for each group. What do the differences imply about high-spending shoppers in each group?

The mean of shoppers with a shopping list is $27.80 less than the mean of shoppers who shopped without a shopping list. The median of shoppers with a shopping list is $25.00 less than the median of shoppers who shopped without a shopping list.

Thus, the differences in the mean and median between groups indicate that the average person with a shopping list will spend 25-27.80 dollars less than a person without a shopping list per trip. It also means that the average person with a shopping list won't overspend whereas the average person with no shopping list will overspend per trip.

Turnitin ID: 2801990112

0%

Report question issue

2 of 6

2c)

1 Marks

Compare the standard deviations across the two groups. What does this tell you about the variability in spending?

The standard deviation of shoppers with a shopping list is 28.4 compared to the standard deviation of shoppers with no shopping list which is 40.5. The standard deviation for shoppers with no shopping list is 12.1 larger than the with shopping list. This means that the shoppers with no shopping list has more variability in their spending because of the larger standard deviation which adds up due to there being a wide range of random things they might buy that isn't needed.

Turnitin ID: 2801990465

0%

Report question issue

2d)

2 Marks

Suppose a shopper without a shopping list spent $200. Compute their z-score using the sample mean and sample standard deviation provided in Table 2, relative to the "No List" group and interpret.

(200-126.3)/40.5 = 1.8198

This means that the shopper without a list spending 200 dollars is 1.8198 standard deviations away from the mean

Resubmit to Turnitin

Report question issue

2e)

2 Marks

If the company wanted to report the "typical" spending amount for each group, which measure(s) from Table 2 would you recommend and why?

I would recommend the mean as this determines the average of each customers purchase rather than purely just the middle number in the median. Therefore it takes into consideration any high spending customer and low spending customers to give an accurate figure as a typical spending figure by using an average across every single purchase.

Turnitin ID: 2801990557

0%

Report question issue

Report question issue Notes

Question 2 Notes

Comment:

Unsure **Response history**

8.5 Marks

The SCENARIO section described how the data for this scenario was collected. This information is provided again below:

A retail analytics company is studying shopping behaviour at a large Melbourne mall. Over the course of two weeks, the company collects data from a sample of shoppers who happened to be at the mall and were willing to participate when approached.

For each participating shopper, they completed a survey that contained the following information:

- Whether they used a shopping list (1 = Yes / 0 = No)
- Whether they spent more than $100 (1 = Yes / 0 = No)
- Their total spending (in dollars)
- Age (in years)
- Household income (in $1,000s per year)
- Distance from the mall (in km)
- Loyalty program membership (1 = Yes / 0 = No)

In total, there were 400 survey responses collected during this time period.

3a)

1.5 Marks

Identify the population, sampling frame, and sample in this study.

The population is 400 people.

Sampling frame was over the course of a 2 week period.

Sample is the person and their characteristics with their spending.

Turnitin ID: 2801990739

0%

Report question issue

3b)

2 Marks

Explain what type of sampling approach has been used in this study.

A random sample approach has occurred as the marketing team chose 400 random people who entered the mall to participate in the survey if they chose to. They ensured it wasn't over 1 day but instead over 2 weeks to best get people that come in on each day and no double ups of people occurred.

Turnitin ID: 2801990712

0%

Report question issue

3 of 6

3c)

2 Marks

Based on the sampling approach, describe two sources of bias that might arise.

A source of bias that definitely could occur is the fact that is the same mall. The same mall could result in potential biased and skewed results due to the fact that you are only collating data from the one mall of melbourne. One way to fix this is to do the same survey at a different mall and compare results.

Another form of bias is people not giving honest answers due to embarrassment of being judged for over spending or not using a shopping list. This can affect the results as people giving false answers will skew the averages to make them look more favourable for the customers.

Turnitin ID: 2801990852

0%

Report question issue

3d)

1 Marks

Suppose the mean spending is $112.40 with a sample standard deviation of $25.20 and n = 400. Compute a 95% confidence interval for mean spending.

The confidence interval will be 112.40 - 25.20 and 112.40 + 25.20

95% CI = (87.2 , 137.6)

Resubmit to Turnitin

Report question issue

3e)

2 Marks

The mall claims the true mean spending is $120. Based on your confidence interval, comment on whether this claim is supported.

The mall's claim that the true mean spending is $120 could very well be true due to the fact that the 95% confidence interval is between $87.20 and $137.60. This means we are 95% confidence that the true mean lies between these figures which $120 does. Thus the claim can be supported with the standard deviation being a high number meaning theres reduced accuracy in the figures gathered.

Turnitin ID: 2801991074

23%

Report question issue

Report question issue Notes

Question 3 Notes

Comment:

Unsure

Response history

7 Marks

As part of their study, the retail analytics team wanted to understand how different factors relate to one another in the dataset. To do this, they calculated Pearson's correlation coefficients between key variables (see the table below):

Table 3. Correlation Matrix of selected study variables

	1	2	3	4	5
1. Income	--				
2. Spending	0.78	--			
3. Age	0.35	0.10	--		
4. Distance	-0.25	-0.40	0.05	--	
5. Loyalty Member	0.05	0.30	-0.02	-0.08	--

Note:

- $|r| < 0.20$ = very weak,
- $0.21 < |r| < 0.39$ = weak,
- $0.40 < |r| < 0.59$ = moderate.

- 0.60 < |*r*| < 0.79| = strong,
- |*r*| > 0.80 = very strong

4a)

1 Marks

Interpret the correlation between Income and Spending.

The correlation between income and spending indicates that 78% of the change in spending is caused by the change in income. This is a strong positive linear relationship indicating that it is very likely that these two variables impact each other.

Turnitin ID: 2801991052

0%

Report question issue

4b)

2 Marks

Discuss two factors that could make the correlation between Income and Spending be misleading if interpreted as causal.

If it is interpreted casually this could mean that people could interpret this as being a very very good number as 78% is very good. But in reality there is still 22% that doesn't cause change. Yes it is strong but it could be interpreted as very strong.

As well as this, the correlation could be misleading as because of the correlation people could think that there is a complete association between the two where there is an association but there are still nearly a quarter of the data that can't be explained by a change in the variable.

4 of 6

Turnitin ID: 2801991214

0%

Report question issue

4c)

2 Marks

Consider the correlation between Distance from the mall and Spending. Suggest two

plausible reasons that could explain this finding.

The correlation is -0.40 which makes sense as this indicates a moderate negative linear association between distance from the mall and spending. This indicates that 40% of the change in spending can be explained by the change in distance from the mall. The negative indicates that as distance from the mall is further away, the spending will decrease. This finding makes sense as someone who lives further away from the mall will most likely visit the mall less frequently and therefore not spend as much as someone who lives closer and might visit more frequently. In conjunction, there still might be people who do big purchases when they come in their infrequent trips causing very varied data collated. That would explain why it is only a moderate association.

Turnitin ID: 2801991262

10%

Report question issue

4d)

2 Marks

Which correlation(s) in the table should we be most cautious about interpreting, and why?

All of the figures that are very weak. As these figures are usually hard to see on graphs and are difficult as since there is very little correlation between the two variables it can be tricky to interpret. Such as the correlations -0.02, 0.05, 0.05, -0.08 etc. All show close to no correlation between the two variables and it is almost not even worth explaining that there is some sort of explanation between the two variables.

Turnitin ID: 2801991351

0%

Report question issue

Report question issue Notes

Question 4 Notes

Comment:

Unsure

Response history

10 Marks

To understand drivers of shopper spending, the retail analytics team fitted two linear models with Spending (in dollars) as the outcome (see Tables 4 – 7 below):

Table 4. Analysis of variance output for Model 1

Source	Sum of Squares	df	MS	F
Model	207,638	3	69,213	95.60
Error	286,739	369	725	
Total	494,377	399		

Table 5. Regression coefficients output for Model 1

Variable	Coefficient	Std. Error	*t*	*p*
Constant	50.20	12.10	4.15	.0001
Household Income	1.25	0.30	4.1	.0001
Age	0.80	0.25	3.20	.0020
Distance	-2.40	0.90	-2.67	.0080

Table 6. Analysis of variance output for Model 2

Source	Sum of Squares	df	MS	F
Model	222,470	3	74,157	108.1
Error	271,907	396	686	
Total	494,377	399		

Table 7. Regression coefficients output for Model 2

Variable	Coefficient	Std. Error	*t*	*p*
Constant	47.80	11.90	4.02	.0001
Household Income	1.15	0.31	3.71	.0001
Distance	-2.25	0.88	-2.56	.0110
Loyalty Member	13.50	3.70	3.65	.0001

Note: for the questions below (5a to 5d), report coefficients to 2 decimal places, and *p*-values to 4 decimal places.

5 of 6

5a) 1 Marks

Interpret the coefficient for Distance in Model 1.

For every one additional unit increase in distance, the spending in dollars decreases by 2.40 dollars.

Resubmit to Turnitin

Report question issue

5b) 2 Marks

Interpret the coefficient for Loyalty Member in Model 2.

For every one additional unit increase in loyalty member the spending in dollars is expected to increase by 13.50

Resubmit to Turnitin

Report question issue

5c) 4 Marks

Using the information provided, evaluate how well each model explains variation in spending. Show your working and interpret the results in context.

Both models explain the variations in spending really well using many different tools to determine the variation and describing the features of the statistics collated. Both models providing the sum of squares to determine the amount of room from line of best fit to the actual plots. The constant coefficient being provided is important as this shows the intercept when everything being = 0 for model 1 this is 50.20 so spending is larger and model 2 shows 47.80. The coefficients are the most important with the variation as this provides an intel of how much

each variable changes for each additional unit increase in the variables. With model 1 having coefficients of 1.25, 0.80 and -2.40 and model 2 having 1.15, -2.25 and 13.50. Model 2 having much larger coefficients determines that there is more variation in model 2 with their coefficients. The standard errors in model 2 being much larger also explains a much larger variation in model 2 on comparison to model 1.

Turnitin ID: 2801988986

0%

Report question issue

5d)

3 Marks

If you were advising the Head of the retail analytics team, which model would you recommend and why? In your response, consider both the statistical evidence and the practical usefulness of the predictors included.

If i was advising the head of the retail analytics team I would definitely recommend using model 1 as in this model all the p-values generated are less than 0.05 which means that the data and the coefficients are significant. With the p value for the constant being 0.001 household income is 0.0001 as well, age is 0.0020 and distance is 0.0080. Therefore all of these being under 0.05 indicates that this can be a significant indicator to use for predictions and models in the future. In Model 2, the distance p value is 0.0110 which is greater than 0.05 so therefore that makes that insignificant so it is best not to use it. As well as this, model 2 has some greater standard errors for their coefficients whereas in model 1, the standard errors are a lot smaller. Ultimately I would recommend Model 1 to be used to the head of the retail analytics team.

Turnitin ID: 2801989215

0%

Report question issue

Report question issue Notes

Question 5 Notes

Comment:

Unsure **Response history**

9 Marks

The retail analytics team has constructed another **linear regression model**, where the outcome variable is whether a shopper spends **more than $100**. The predictors included in this model are **Income, Distance from the mall, and Loyalty Membership**. The model was trained on 300 shoppers and tested on a separate 100 shoppers. The resulting **confusion matrices** are shown in Tables 8 and 9 below.

Table 8. Confusion matrix for the training set (n = 300)

		Actual		
		Spent > $100	**Spent < $100**	Total
Predicted	**Spent > $100**	140	20	160
	Spent < $100	30	110	140
	Total	170	130	300

Table 9. Confusion matrix for the testing set (n = 100)

		Actual		
		Spent > $100	**Spent < $100**	Total
Predicted	**Spent > $100**	25	25	50
	Spent < $100	25	25	50
	Total	50	50	100

6a)

2 Marks

Why have the analysts split the data into training and testing sets? What are the implications if they do not do this?

The training set was used as a trial run for the analysts to practice the model that they had generated to predict the results of the 300 people that they interviewed and then the testing set was to actually properly use the model for research data on 100 new people. This model is created in hope to reduce the interviews needed as the analysts can almost predict everyone's answer. The implication if they don't do this is they will implement a new strategy and model to use to predict people's spending without ensuring they know that it works and is accurate.

Turnitin ID: 2801989592

0%

Report question issue

6b)

1 Marks

Determine the overall accuracy for the training set.

For the training set, there was 0.8333 accuracy overall in the data set with the analysis's correctly predicting 250 out of 300 samples.

Turnitin ID: 2801989828

0%

Report question issue

6c)

1 Marks

Determine the Type I and Type II error rates for the testing set.

The error rates for Type I is 0.5 as 25 out of 50 were predicted wrong.

The error rates for Type II is also 0.5 as 25 out of 50 were also predicted wrong.

Turnitin ID: 2801989969

0%

Report question issue

6d)

3 Marks

Compare the training and testing performances. What does this suggest about how well the model generalises?

The training model uses triple the sample size (300 compared to 100) and also has a much higher accuracy

rate.

The accuracy rate for the training performance was 0.8333 compared to the testing set which was 0.5 even with a much higher sample size. Therefore the model wasn't great when it was being tested but it was really good when it was in the training stages.

Turnitin ID: 2801990073

0%

Report question issue

6e)

2 Marks

Briefly explain one limitation of the analysis chosen for this model.

A limitation is that the testing and training performances used a different number of people for each sample. With the training sample using 300 people and the testing sample using 100 people. Therefore there wasn't as many people for the testing sample to test on and the same amount of people should be used in the test.

Turnitin ID: 2801990238

0%

Report question issue

Report question issue Notes

Question 6 Notes

Comment:

Unsure **Response history**