Examiners' commentaries 2022

ST3188 Statistical methods for market research: Preliminary examination

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2021). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

General remarks

Learning outcomes

At the end of the course and having completed the essential reading and activities you should be able to:

- define a market research problem and create an appropriate research design
- perform independent data analysis in a market research setting
- determine which statistical method is appropriate in a given situation and be able to discuss the merits and limitations of a particular method
- use statistical software to analyse datasets and be able to interpret output
- draw appropriate conclusions following empirical analysis and use to form the basis of managerial decision-making
- demonstrate greater commercial awareness.

Format of the examination

The examination is two hours long and you must answer the question in Section A and two questions out of three in Section B. The examination is worth 70% of the final grade. The other 30% is determined by the coursework component. (The coursework comprised the production of a market research proposal – see the 'Assessment' section in the VLE for details.)

Examination revision strategy

Many candidates are disappointed to find that their examination performance is poorer than they expected. This may be due to a number of reasons, but one particular failing is '**question spotting**', that is, confining your examination preparation to a few questions and/or topics which have come up in past papers for the course. This can have serious consequences.

We recognise that candidates might not cover all topics in the syllabus in the same depth, but you need to be aware that examiners are free to set questions on **any aspect** of the syllabus. This means that you need to study enough of the syllabus to enable you to answer the required number of examination questions.

The syllabus can be found in the Course information sheet available on the VLE. You should read the syllabus carefully and ensure that you cover sufficient material in preparation for the examination. Examiners will vary the topics and questions from year to year and may well set questions that have not appeared in past papers. Examination papers may legitimately include questions on any topic in the syllabus. So, although past papers can be helpful during your revision, you cannot assume that topics or specific questions that have come up in past examinations will occur again.

If you rely on a question-spotting strategy, it is likely you will find yourself in difficulties when you sit the examination. We strongly advise you not to adopt this strategy.

Examiners' commentaries 2022

ST3188 Statistical methods for market research: Preliminary examination

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2021). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions

Section A

Question 1

(a) **The UK's Department of Health and Social Care (DHSC) wishes to investigate 'vaccine hesistancy' among the general UK population, in particular among parents as the country's Covid-19 vaccination programme extends to ever-younger age groups.**

In order to encourage greater take up of the vaccine, the DHSC is considering a (possibly digital) marketing campaign to encourage any hesitant individuals to get vaccinated. However, the DHSC is unsure who to target as well as the common reasons for vaccine hesitancy.

To better understand potential anxieties, the DHSC has decided to use a survey and has asked you to devise an appropriate sampling scheme. Explain in detail how each of the following sampling methods could be applied to the overall sampling strategy for this study. Make sure you describe the merits and limitations of each as well as how each would be applied in practice.

i. **Judgemental sampling.**

ii. **Snowball sampling.**

iii. **Stratified sampling.**

iv. **Cluster sampling.**

(20 marks)

Reading for this question

Block 9 of the course.

Approaching the question

Standard responses expected here, noting which are probability and non-probability methods. For i. and ii., a brief discussion of selection bias is desirable and the extent to which these sampling techniques can yield representative samples. For iii. and iv., a viable sampling frame should be identified, as well as possible examples of appropriate strata and clusters, respectively. As explicitly mentioned in the question, as well as the mechanics of each method the merits and limitations should be described. Note that for this type of question there is no single 'right' answer.

(b) Suppose we are interested in estimating the proportion of a population using a simple random sample of size n. *In your own words*, answer the following.

i. State a suitable estimator of the population proportion and also state its sampling distribution. Mention any assumptions which you make.

ii. Explain statistically how to determine the minimum sample size necessary to estimate a population proportion to within e units with 99% confidence.

iii. In part ii., discuss how you would choose a numerical value for e. Justify your choice.

iv. Suppose you were told that a 95% confidence interval for a population proportion was computed to be $(0.635, 0.665)$. Explain how this interval should be interpreted in practice.

(20 marks)

Reading for this question

Block 10 of the course.

Approaching the question

i. When sampling from a Bernoulli population:

$$P = \bar{X} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

approximately, by the central limit theorem as $n \to \infty$.

ii. Should state the condition on finding n as:

$$z_{\alpha/2} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq e$$

and hence:

$$n \geq \frac{(z_{\alpha/2})^2 \pi(1-\pi)}{e^2}.$$

For a 99% confidence level, we would set $z_{\alpha/2} = 2.576$. The value of π should either be an assumed value, an estimate based on a pilot study, or set equal to 0.50 as a coservative estimate which provides the maximum standard error.

iii. Any reasonable argument accepted. Ultimately, it would depend on the nature of the market research problem and how accurate parameter estimation needed to be.

iv. We are 95% confident that the true population proportion, π, lies between 0.635 and 0.665. As the point estimate of π is the midpoint of the confidence interval, we can deduce that $p = (0.635 + 0.665)/2 = 0.65$. When constructing 95% confidence intervals, 95% of the time we expect the interval endpoints to span (or contain) the parameter being estimated.

Section B

Answer two questions. Each question carries equal weight.

Question 2

(a) An e-commerce retailer wants to investigate whether there is any relationship between order values and when orders are made based on a sample of $n = 400$ recent transactions. The day of the week (Monday through to Sunday) was recorded along with the time of day (classified as 'Morning', 'Afternoon' and 'Evening'). A two-way analysis of variance was conducted.

Analyse the selected SPSS output in Figure 1 (on the next page) and discuss what conclusions can be drawn from the data. In your analysis, be sure to address at least the following:

- **Describe the strength of the joint effect of the factors.**
- **Test the significance of the variables individually and the interaction between them and interpret the results.**
- **What conclusions, if any, could be drawn about online spending patterns.**

(20 marks)

Reading for this question

Block 13 of the course.

Approaching the question

Two-way analysis of variance (ANOVA) has a model assumption of the error variance of the dependent variable being equal across groups and this is tested using Levene's test of equality of error variances. We test:

$$\mathrm{H}_0 : \text{equal error variances} \quad \text{vs.} \quad \mathrm{H}_1 : \text{unequal error variances.}$$

The F test statistic value (based on the mean, say) is 1.303 drawn from the $F_{20,\,379}$ distribution under H_0, giving a p-value of 0.172. Since this is *greater* than 0.05, say, we do *not* reject H_0 at the 5% significance level, hence the two-way ANOVA model variance assumption seems to be satisfied.

Turning to the 'Tests of Between-Subjects Effects' table, the multiple η^2 value is:

$$\frac{SS_{Day} + SS_{Time} + SS_{Day*Time}}{SS_{Order\ value}} = \frac{SS_{Model}}{SS_{Order\ value}} = \frac{9{,}573{,}556.01}{13{,}247{,}115.80} = 0.7227$$

hence 72.27% of the variation in order value can be jointly explained by the day and time factors. This indicates a fairly strong joint effect of the factors.

We now test the significance of the model overall. H_0 is that the model has no overall effect, while H_1 is that the model has some effect. The test statistic value is:

$$F = \frac{MS_{Day,Time,Day*Time}}{MS_{Error}} = \frac{455{,}883.620}{9{,}692.770} = 47.033$$

with a corresponding p-value of < 0.001 indicating that the model is highly significant. Turning to the factors separately, 'Day' is highly significant (F-value of 140.459; p-value of < 0.001), while 'Time' and the interaction of 'Day' and 'Time' are both insignificant (p-values of 0.717 and 0.986, respectively). Hence 'Day' is the significant determinant of order value.

Looking at the means plot, it is clear that Thursdays seem to generate the highest order values (regardless of time of day), while Saturdays seem to generate the lowest order values, on average. So it appears that there are patterns in online spending, hence further market research could be conducted to better understand the reason(s) behind *why* such variation exists across days of the week.

Figure 1

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
Order value	Based on Mean	1.303	20	379	.172
	Based on Median	.841	20	379	.663
	Based on Median and with adjusted df	.841	20	335.619	.663
	Based on trimmed mean	1.232	20	379	.224

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

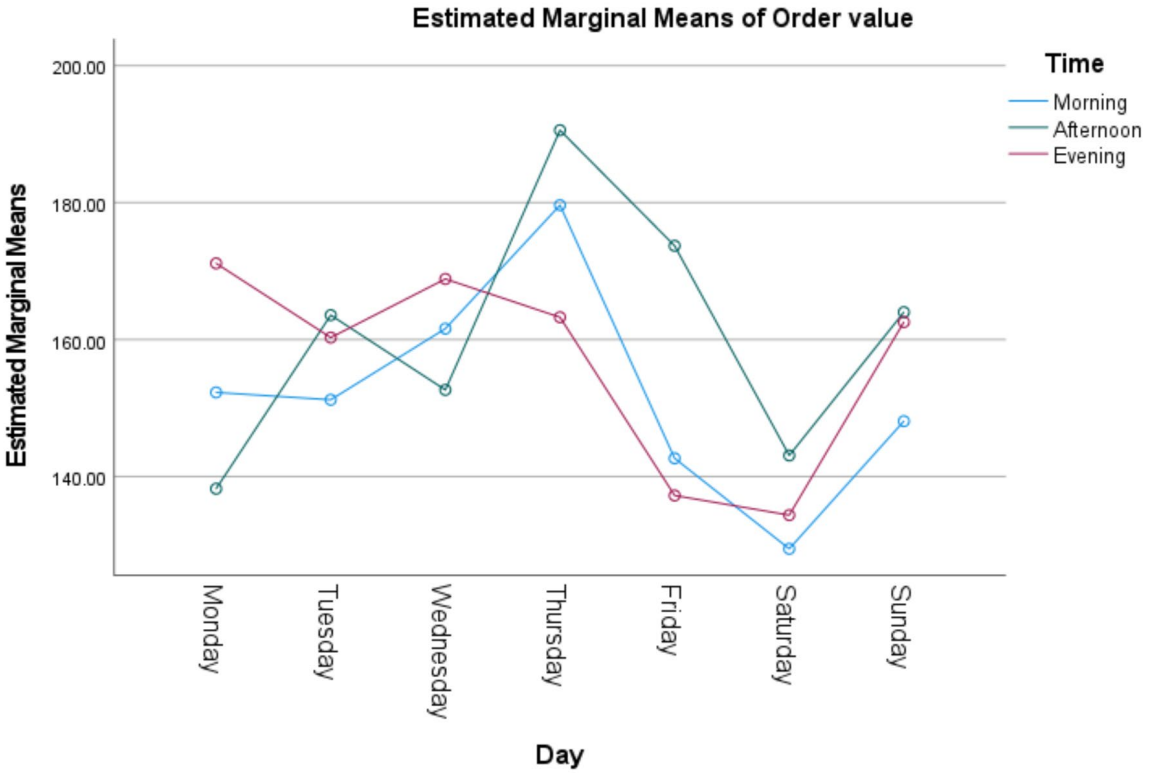
a. Dependent variable: Order value

b. Design: Day + Time + Day * Time

Tests of Between-Subjects Effects

Dependent Variable: Order value

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Model	9573556.01[a]	21	455883.620	47.033	<.001
Day	9530060.133	7	1361437.162	140.459	<.001
Time	6462.988	2	3231.494	.333	.717
Day * Time	37032.893	12	3086.074	.318	.986
Error	3673559.784	379	9692.770		
Total	13247115.80	400			



(b) ***In your own words*, answer the following.**

In a questionnaire design, suppose you wanted to collect information on the respondent's age. Explain any advantages and disadvantages associated with the choice of the following three questions:

i. **'What is your age?'**

ii. **'What is your year of birth?'**

iii. **'To which age group do you belong? Under 18, 18–29, 30–49, 50–69, or 70 and over.'**

(10 marks)

Reading for this question

Block 7 of the course.

Approaching the question

A good answer here would reflect on the different levels of measurement of answers across the questions, as well as the potential for non-response. For example, 'what is your age?' would return a ratio level variable, but as age may be a sensitive matter for some people non-response could be an issue. 'What is your year of birth?' is a subtle way of obtaining age information without explicit reference to 'age'. Note the response to this question would not necessarily return the respondent's exact age, depending on whether their birthday had already occurred that year at the time of responding. The age group question is perhaps most likely to obtain responses, since respondents are not revealing their precise age, however such data would be ordinal and so offer more limited ways to analyse the data.

Question 3

(a) **A large company wanted to explain the variation in annual salaries of its managerial staff. It collected data on the following variables:**

- **annual salary, in £**
- **years of experience**
- **tenure with the company**
- **managerial level (in the company), coded 1 to 4 (level 1 = most junior; level 4 = most senior).**

You are told the correlation coefficient between years of experience and tenure with the company is 0.978.

Selected SPSS output is provided in Figure 2 (on the next page). Analyse the regression results. In your analysis, be sure to address at least the following:

- **Write out the full regression model, including any assumptions, and the estimated model.**
- **Comment on whether it is appropriate to include 'years of experience' *and* 'tenure with the company' as independent variables in the regression model.**
- **Comment on whether it is appropriate to have included 'managerial level' in the way it has been modelled.**

(20 marks)

Figure 2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.947[a]	.897	.893	5835.178

a. Predictors: (Constant), Managerial level, Tenure with the company, Years of experience

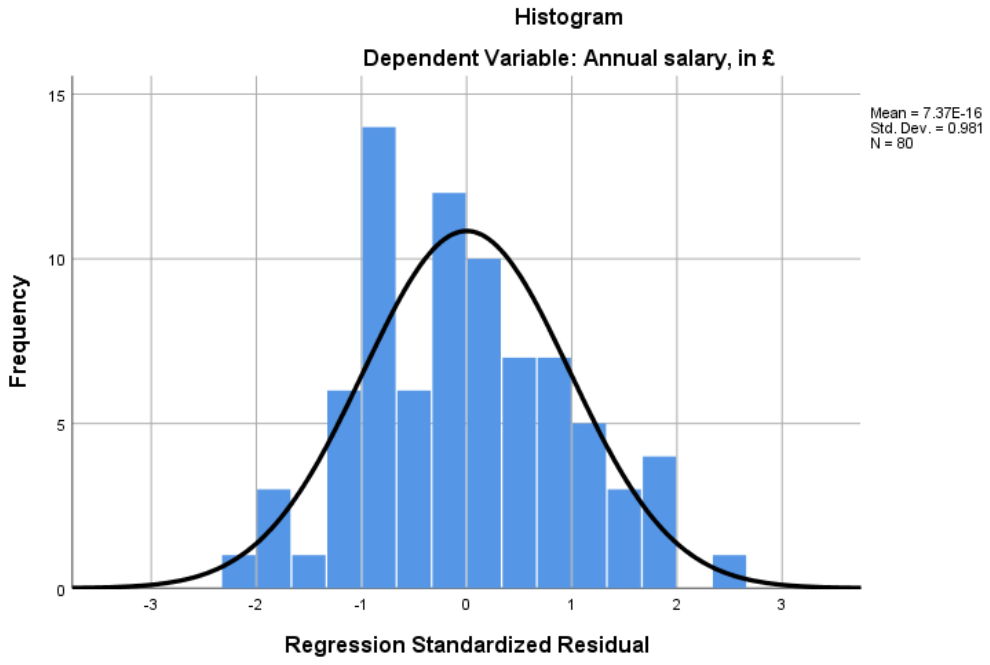
b. Dependent Variable: Annual salary, in £

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.259E+10	3	7530169273	221.155	.000[b]
	Residual	2587747230	76	34049305.66		
	Total	2.518E+10	79			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12878.763	1792.258		7.186	.000
	Years of experience	935.949	428.025	.385	2.187	.032
	Tenure with the company	144.536	453.359	.056	.319	.751
	Managerial level	15652.681	680.463	.848	23.003	.000



Reading for this question

Block 14 of the course.

Approaching the question

The full regression model is:

$$\text{Salary}_i = \beta_0 + \beta_1 \times \text{Years}_i + \beta_2 \times \text{Tenure}_i + \beta_3 \times \text{Level}_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. We assume $\varepsilon_i \sim N(0, \sigma^2)$. The estimated model is:

$$\widehat{\text{Salary}}_i = 12{,}878.8 + 935.9 \times \text{Years}_i + 144.5 \times \text{Tenure}_i + 15{,}652.7 \times \text{Level}_i.$$

Since 'years of experience' and 'tenure with the company' have a reported correlation of 0.978, there is a high risk of multicollinearity, i.e. it becomes very difficult to distinguish the effects of one variable from another on the dependent variable, since highly correlated variables move closely together. This tends to result in imprecise parameter estimates (i.e. large estimated standard errors) which in turn tends to lead to small t values (since $t = \widehat{\beta}_i/\text{E.S.E.}(\widehat{\beta}_i)$) and hence large p-values, so more likely to be insignificant.

'Managerial level' is an ordinal variable with four levels, but here has been included in the regression model as a measurable variable. The correct treatment of categorical independent variables is by using dummy variables. So, ideally, 'managerial level' should be linked to three dummy variables (i.e. one less than the number of levels), with the reference category being any of the four, but when the variable is ordinal it seems reasonable to have the lowest level as the base category, in this case level 1.

A good response would also report the R^2 value (here 0.897) and use this to indicate the strength of the explanatory power of the regression model. A joint test of significance could be conducted, reporting the F-value and corresponding p-value (here 221.155 and 0.000, respectively). Thereafter, t tests of the individual variables should be conducted. Note that 'tenure with the company' is insignificant, unlike 'years of experience', possibly due to multicollinearity as discussed above. The standardised coefficients can also be discussed to infer the relative importance of the independent variables (but mindful of the multicollinearity issue). Finally, a brief discussion of the standardised residual plot would be beneficial to indicate the extent to which the error term assumptions seem to be satisfied.

(b) ***In your own words*****, answer the following.**

What benefits do focus groups provide to a market researcher? Are there any drawbacks?

(10 marks)

Reading for this question

Block 4 of the course.

Approaching the question

A discussion of some of the benefits, which include synergy, snowballing, stimulation, security, spontaneity, serendipity, specialisation, scientific scrutiny, structure and speed; as well as some of the drawbacks, which include misjudgement, moderation, messiness, misrepresentation and meeting.

Question 4

(a) A retailer is concerned about a high level of *churn* recently (i.e. losing its customers to its competitors). It has decided to use discriminant analysis to construct a model capable of predicting which existing customers may churn. It decides to use the following predictor variables (all with respect to the customer):

- **Age in years.**
- **Gender.**
- **Household income, in thousands.**
- **Household size.**
- **Years at current address.**

Analyse the selected SPSS output in Figure 3 (spread over the next two pages) and discuss what conclusions can be drawn from the data. Keep in mind the retailer's desire to predict which customers are likely to churn. In your analysis, be sure to address at least the following:

- **State the theoretical and estimated discriminant analysis models.**
- **Comment on the relative importance of the predictor variables.**
- **Comment on the suitability of including the gender variable.**
- **Determine the predictive accuracy of the model.**

(20 marks)

Reading for this question

Block 15 of the course.

Approaching the question

Here we are discriminating between two groups, and so we require only one discriminant function. The theoretical model is:

$$D = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Gender} + \beta_3 \times \text{Income} + \beta_4 \times \text{Size} + \beta_5 \times \text{Years}$$

with the estimated model using the output table with the unstandardised coefficients, hence:

$$D = -2.896 + 0.033 \times \text{Age} + 0.319 \times \text{Gender} + 0.000 \times \text{Income} + 0.239 \times \text{Size} + 0.071 \times \text{Years}$$

which we note is a significant model, based on Wilks' lambda.

The relative importance of the predictor variables can be assessed using the standardised coefficients and/or the structure correlations. Using either, it seems that 'years at current address' is the best predictor, followed by 'age in years'. The ANOVA F tests also indicate that these are the only two significant predictor variables. Examining the correlation matrix, the strongest correlation coefficient is between these two variables ($r = 0.660$), which is not too strong, so we would not expect multicollinearity to be much of an issue.

In a discriminant analysis model, we would prefer quantitative variables as the predictor variables. Gender, of course, is categorical so its inclusion in the model is questionable. It could be argued that this may not be too problematic since sufficient variation would come from the other four predictors.

Looking at the group centroids, on average those who churned in the dataset had an average discriminant score of -0.488 versus 0.177 for those who did not churn. Hence the classification rule can be obtained from computing the mid-point of these centroids: $(-0.488 + 0.177)/2 = -0.1555$, such that:

$$\text{If } D_i \begin{cases} < -0.1555 & \text{classify as 'churn'} \\ > -0.1555 & \text{classify as 'not churn'} \\ = -0.1555 & \text{equally likely to churn or not churn.} \end{cases}$$

The predictive accuracy of the model can be calculated with the hit ratio using the cross-validated cases. The proportion of correctly classified cases is:

$$\frac{246 + 110}{600} = 0.5933$$

hence the model has 59.33% predictive accuracy. A good answer would also comment on the relative amounts of false positives and false negatives.

Figure 3

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age in years	.951	30.982	1	598	.000
Gender	.998	1.253	1	598	.263
Household income, in thousands	.995	3.104	1	598	.079
Household size	1.000	.259	1	598	.611
Years at current address	.935	41.860	1	598	.000

Pooled Within-Groups Matrices

		Age in years	Gender	Household income, in thousands	Household size	Years at current address
Correlation	Age in years	1.000	.001	.303	-.293	.660
	Gender	.001	1.000	.046	.013	-.012
	Household income, in thousands	.303	.046	1.000	-.101	.219
	Household size	-.293	.013	-.101	1.000	-.224
	Years at current address	.660	-.012	.219	-.224	1.000

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.087[a]	100.0	100.0	.283

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.920	49.563	5	.000

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age in years	.407
Gender	.160
Household income, in thousands	-.007
Household size	.346
Years at current address	.710

Figure 3 (continued)

Structure Matrix

	Function 1
Years at current address	.898
Age in years	.773
Household income, in thousands	.245
Gender	.155
Household size	.071

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Canonical Discriminant Function Coefficients

	Function 1
Age in years	.033
Gender	.319
Household income, in thousands	.000
Household size	.239
Years at current address	.071
(Constant)	-2.896

Unstandardized coefficients

Functions at Group Centroids

Churn within last month	Function 1
No	.177
Yes	-.488

Unstandardized canonical discriminant functions evaluated at group means

Classification Results[a,c]

		Churn within last month	Predicted Group Membership: No	Yes	Total
Original	Count	No	249	191	440
		Yes	47	113	160
	%	No	56.6	43.4	100.0
		Yes	29.4	70.6	100.0
Cross-validated[b]	Count	No	246	194	440
		Yes	50	110	160
	%	No	55.9	44.1	100.0
		Yes	31.3	68.8	100.0

(b) ***In your own words*, answer the following.**

Discuss the possible impacts of introducing a third variable in cross-tabulation.

(10 marks)

Reading for this question

Block 12 of the course.

Approaching the question

Here it is expected that there should be a discussion of:

i. refining an initial relationship
ii. detecting a spurious initial relationship
iii. revealing a suppressed association.