SINGAPORE INSTITUTE OF MANAGEMENT

UNIVERSITY OF LONDON
PRELIMINARY EXAM 2022

MODULE CODE : ST3188

MODULE TITLE : Statistical methods for market research

DATE OF EXAM :

TOTAL NUMBER OF PAGES (INCLUDING THIS PAGE) : 9

INSTRUCTIONS TO CANDIDATES :-

Candidates should answer the **ONE** question in **Section A** and **TWO** questions from **Section B**. Section A carries 40 marks. Questions in Section B carry 30 marks each.

Candidates are strongly advised to divide their time accordingly.

SECTION A: Compulsory

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)

   (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)

SECTION B: Answer two questions. Each question carries equal weight.

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)

   (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)

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.

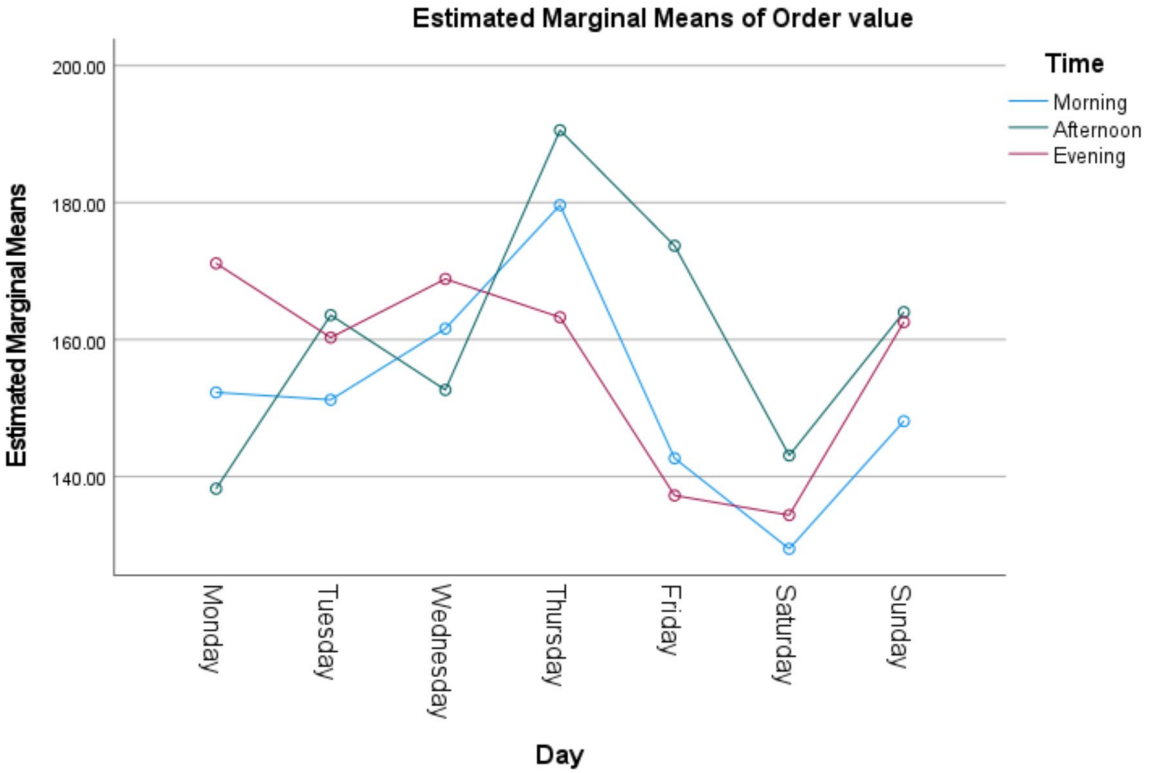
a. Dependent variable: Order value

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			



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)

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

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

(10 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

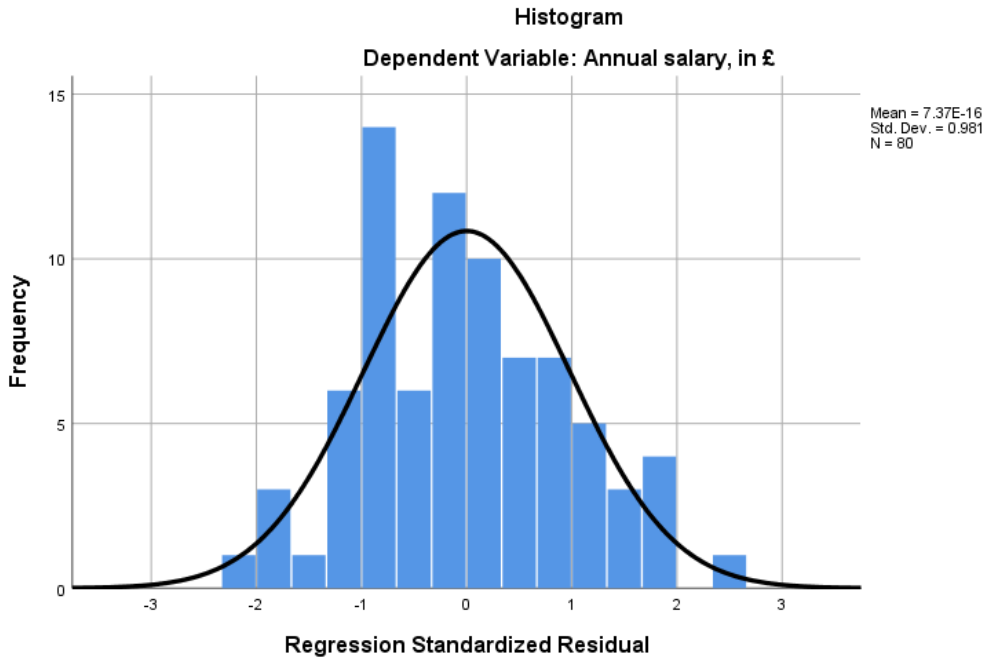
b. Dependent Variable: Annual salary, in £

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		
		B	Std. Error	Beta	t	Sig.
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



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)

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

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

(10 marks)

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

[END OF PAPER]