NEWCASTLE UNIVERSITY

SEMESTER 2 2012/2013

ACE2013

Statistics for Marketing and Management

Time allowed: 2 hours

Candidates should attempt ALL questions. Marks for each question are indicated. However you are advised that marks indicate the relative weight of individual questions, they do not correspond directly to marks on the University scale.

There are FIVE questions on this paper.

Answers to questions should be entered directly on this question paper in the spaces provided. This question paper must be handed in, attached inside an anonymised cover sheet, at the end of the examination.

Statistical tables, and a list of useful formulae, are provided in a separate booklet.

1. An article in the *Journal of Business Venturing* (Vol. 11, 1996) reported on the activities of entrepreneurs during the business creation process. Among the questions investigated were: what activities, and how many activities, do entrepreneurs initiate in attempting to create a new business? A total of 63 entrepreneurs were interviewed and divided into three groups: those who were successful in founding a new business ($n = 30$), those still actively trying to establish a business ($n = 19$), and those who tried to start a new business but eventually gave up ($n = 14$). The total number of activities undertaken (i.e., developed a business plan, sought funding, looked for facilities, etc.) by each group over a specified time period was measured. ANOVA was used to investigate whether there were differences in the total number of activities amongst the three groups of entrepreneurs.

(a) The following is a partial ANOVA table from the study. Complete all the missing entries in the table, except those marked ✘ *[Show your working in the space provided under the table]*

Source	DF	Sum of squares	Mean square	F
Between groups	2	128.70	64.35	0.14
Within Groups	60	27124.52	452.08	✘
Total	62	27257.22	✘	✘

(b) Specify the null and alternative hypotheses that should be tested in the study.

Answer:

H_0 : $\mu_1 = \mu_2 = \mu_3$ (mean number of activities for the three groups is equal)

H_1 : Not all mean number of activities are equal

Question 1 continued on next page

(c) Do the data provide sufficient evidence to indicate that the total number activities undertaken differed among the three groups of entrepreneurs? Explain your answer.

Answer:

No. H_0 is retained because the p–value is greater than 5%.

(d) What is the range for the p–value of the test you conducted in part (c)?

Answer:

$F_{2,60}(0.05) = 3.15$, $F_{2,60}(0.01) = 4.98$, so with a critical value of 0.14, the p–value is greater than 5%.

(e) What assumptions must be met in order for the test in part (c) to be valid?

Answer:

The data for each group are independent and assumed to follow a normal distribution.

(f) Classify this study as observational or experimental. How does this impact the strength of conclusions drawn from the study?

Answer:

Observational study (survey). Because of this, it is difficult to control the effect of other factors (e.g. age, education, industry) on the total number of activities, hence the conclusions may be seen to be weak.

[20 marks]

2. An investment bank is thinking of investing in a start–up alternative energy company. They can become a major investor for £6M, a moderate investor for £3M, or a small investor for £1.5M. The worth of their investment in 12 months will depend on how the price of oil behaves between then and now. A financial analyst produces the following payoff table with the net worth of their investment (predicted worth − initial investment) as the payoff (in £millions):

	Price of oil		
Action	Substantially higher	About the same	Substantially lower
Major investment	£5M	£3M	−£2M
Moderate investment	£2.5M	£1.5M	−£1M
Small investment	£1M	£0.5M	−£0.1M

(a) Construct a decision tree for this payoff table.

Answer:

Question 2 continued on next page

(b) Based on the expected monetary value (EMV) criterion, identify the "best" investment strategy under the assumption that the probability that the price of oil goes substantially higher is 0.4 and that the probability that it goes substantially lower is 0.2. *[Show your working]*

Answer:

$$\begin{aligned} EMV(\text{Major investment}) &= (5)(0.4) + (3)(0.4) + (-2)(0.2) = 2.8 \\ EMV(\text{Moderate investment}) &= (2.5)(0.4) + (1.5)(0.4) + (-1)(0.2) = 1.4 \\ EMV(\text{Small investment}) &= (1)(0.4) + (0.5)(0.4) + (-0.1)(0.2) = 0.58 \end{aligned}$$

According to the EMV criterion major investment is best — highest payoff of £2.8M.

(c) Calculate the expected value of *perfect information* and explain what it means to the investment bank.

Answer:

$$\begin{aligned} EVPI &= \text{Expected profit under certainty} \\ &\quad -\text{EMV of the best decision (i.e. Major investment).} \end{aligned}$$

EMV(Major investment) is obtained in part (b) above as £2.8M. Expected profit under certainty is weighted (weighted by probabilities) sum of the values of the best decision under each state of nature

$$= (5)(0.4) + (3)(0.4) + (0.1)(0.2) = £3.18M.$$

Thus

$$EVPI = 3.18 - 2.8 = £0.38M.$$

The EVPI represents the additional expected value that can be obtained if perfect information were available about the behaviour of oil prices and hence revenue stream from the three investment plans. In this case, the investment bank should be willing to pay a maximum of £380,000 for this information e.g. through market research aimed at providing information that could help them improve probability assessments of the states of nature (changes in oil prices).

Question 2 continued on next page

[Some space is left here for your solution to the previous question]

[**20 marks**]

3. *AB Vista* produce feed for farm animals. A new type of pig feed containing the enzyme *Econase XT* has been developed with the aim of increasing the rate of weight gain of *Landrace X* pigs. In an upcoming trial, the company are interested in the response variable y, where

$$y = \begin{cases} 1 & \text{if a pig has reached slaughter weight within 4 months} \\ 0 & \text{if a pig has } \textit{not} \text{ reached slaughter weight within 4 months} \end{cases}$$

They believe this might be related to the following predictor variables:

$$\begin{aligned} x_1 &= \text{Starting weight of a pig (kilograms)} \\ x_2 &= \text{Number of pigs in a pen} \\ x_3 &= \begin{cases} 1 & \text{if a pig is given feed containing } \textit{Econase XT} \\ 0 & \text{if a pig is } \textit{not} \text{ given feed containing } \textit{Econase XT} \end{cases} \end{aligned}$$

(a) Briefly explain why it would not be appropriate to fit a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

Answer:

Multiple linear regression would not be appropriate because the response variable is *binary*. Fitted values from the model above would not necessarily be 0 or 1.

(b) `Minitab` was used to fit a multiple logistic regression model, giving the (edited) output shown below. Look at it, and then answer the following questions.

Binary Logistic Regression: y versus x1, x2, x3

```
Logistic Regression Table

 Predictor          Coef      SE Coef      Z      P

 Constant       -10.9209      6.03375  -1.81  0.070

 x1           ?    0.5493     0.206509   2.66  0.008

 x2            -0.738846     0.439324  -1.68  0.093

 x3              3.71372  ?    1.9444   1.91  0.056
```

(i) Complete the regression output above by filling in the blanks indicated by the question marks.

(ii) Give brief practical interpretations of the *direction* of the relationship between $\Pr(y = 1)$ and variables x_1 and x_2, as suggested by the estimated coefficients for these variables.

Answer:

The coefficient of x_1 is positive, suggesting that the heavier the starting weight of a pig, the more likely it is that it will reach it's slaughter weight in the allotted time. Conversely, the coefficient of x_2 is negative, suggesting that the more pigs there are in a pen, the *less* likely it is that a pig will reach it's slaughter weight within 4 months.

(iii) Both x_2 and x_3 have coefficients with p–values greater than 5%. Why should *both* variables not be removed *together* at this stage?

Answer:

The p–values shown indicate the significance of each predictor variable *in the presence of all the other predictors*. Removing a single variable and then re–fitting will result in a different fitted model, with different coefficients and thus different p–values. For example, removing x_2 and then re–fitting could result in the p–value for x_3 falling below 5%.

(c) The output from `Minitab` shown below represents the "final model". Look at it, and then answer the questions that follow.

Binary Logistic Regression: y versus x1, x3

```
Logistic Regression Table

Predictor       Coef    SE Coef      Z      P
Constant    -17.6706    5.84175  -3.02  0.002
x1          0.604886   0.204517   2.96  0.003
x3           2.74803    1.19247   2.30  0.021
```

(i) Briefly explain why this output is that for the "final model".

Answer:

We have reached the "final model" because all coefficients are significantly different from zero; thus, all remaining predictors are important in the model.

(ii) Complete the fitted logistic regression equation below.

Answer:

$$\Pr(y = 1) = \frac{e^{-17.671+0.605x_1+2.748x_3}}{1 + e^{-17.671+0.605x_1+2.748x_3}}$$

(iii) Briefly explain why the logistic regression equation ensures that fitted values lie within the required range.

Answer:

The exponential function always returns positive values; dividing a positive value by "one plus itself" will then always give an answer in the range $[0, 1]$, the required range for probabilities.

(iv) Two pigs both have a starting weight of exactly 27.5 kilograms. For each, find the probability that they will reach their slaughter weight within four months if one *is* given pig feed containing the special enzyme, and the other is *not*. Comment.

Answer:

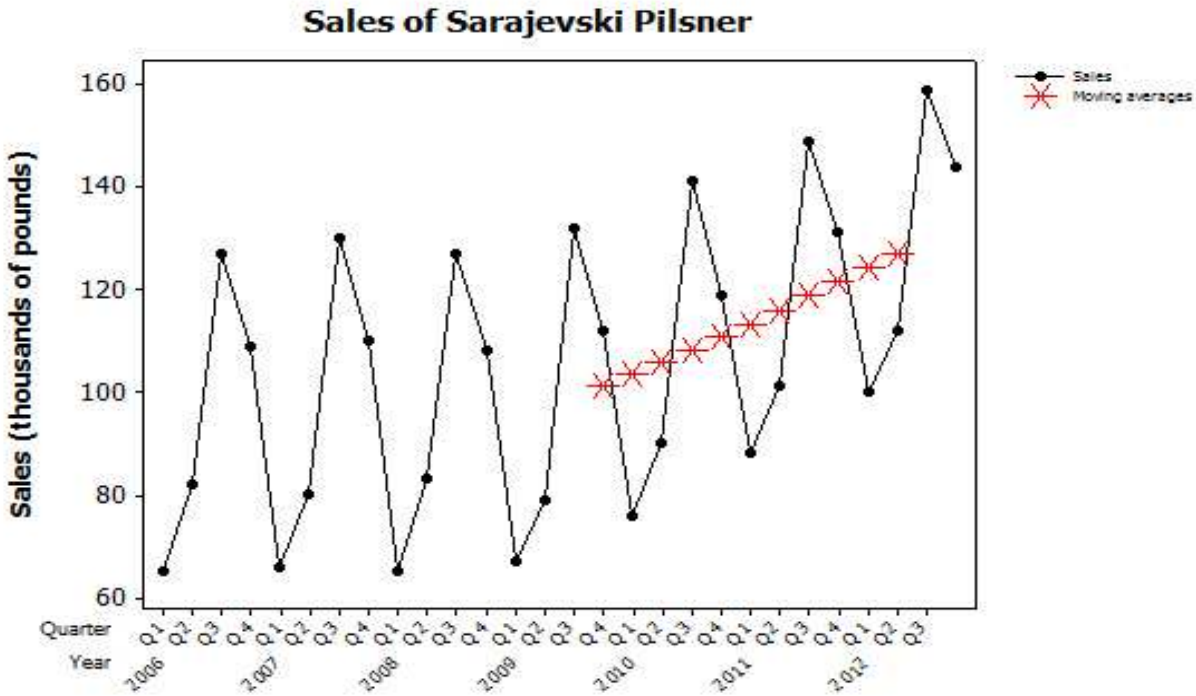
$$\begin{aligned}\Pr(y = 1|x_1 = 27.5, x_3 = 1) &= \frac{e^{-17.671+0.605\times 27.5+2.748\times 1}}{1 + e^{-17.671+0.605\times 27.5+2.748\times 1}} \\ &= 0.847 \quad \text{(for a pig given the enzyme)}\end{aligned}$$

$$\begin{aligned}\Pr(y = 1|x_1 = 27.5, x_3 = 0) &= \frac{e^{-17.671+0.605\times 27.5+2.748\times 0}}{1 + e^{-17.671+0.605\times 27.5+2.748\times 0}} \\ &= 0.262 \quad \text{(for a pig } \textit{not} \text{ given the enzyme)}\end{aligned}$$

Comment: Much more likely to reach slaughter weight within four months if given the special feed containing the enzyme – in fact, more than three times as likely!

[20 marks]

4. The time series plot below shows quarterly sales (in thousands of pounds) of *Sarajevksi Pilsner*, a premium lager, at a chain of pubs in the U.K. (for the years 2006–2012 inclusive). Since 2009, the company have used television advertising to market *Sarajevski Pilsner*.



(a) Comment on any interesting features revealed by this plot.

Answer:

Seasonal variation, with clear positive trend from about 2009 – exactly when the company started to use TV advertising!

(b) Some data from the year 2009 onwards are shown in the table below.

	Jan–Mar	Apr–Jun	Jul–Sep	Oct–Dec
2009	67	79	132	112
2010	76	90	141	119
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Let T represent time, and let $t = 1, 2, \ldots$ correspond to Jan–Mar (2009), Apr–Jun (2009), ... (respectively). We can use centred moving averages (y_t^*) to isolate any trend, where, for quarterly data,

$$y_t^* = \frac{y_{t-2} + 2(y_{t-1} + y_t + y_{t+1}) + y_{t+2}}{8}, \quad t = 3, 4, \ldots \tag{1}$$

(i) Briefly explain why observations at time–points $\{t-1\}$, $\{t\}$ and $\{t+1\}$ are given twice as much weight as those at the other time–points included in equation (1).

Answer:

We use five observations so the moving averages are centred around an integer time–point. However, this will always result in two observations coming from the same 'season'; thus, the remaining three observations are given twice as much weight to 'balance things out', and to avoid one season having twice as much weight as the others.

(ii) Use this formula to find y_3^*, and superimpose this on the time series plot.

Answer:

$$y_3^* = \frac{67 + 2(79 + 132 + 112) + 76}{8} = 98.625.$$

(c) `Minitab` was used to fit a simple linear regression model of the form

$$Y^* \quad = \quad \beta_0 + \beta_1 T + \epsilon,$$

where Y^* represents the centred moving averages and $\epsilon \sim N(0, \sigma^2)$. The resulting output is shown below. Look at it, and then answer the following questions.

Regression analysis: Y* versus T

```
The regression equation is
Y* = 90.4 + 2.58T

Predictor      Coef   SE Coef       T      P
Constant    90.4258    0.3110  290.80  0.000
T           2.58348   0.03389   76.22  0.000

S = 0.405314   R-Sq = 99.8%   R-Sq(adj) = 99.8%
```

(i) Comment on the success of the company's television advertising campaign, with specific reference to the output shown above.

Answer:

The company's TV advertising campaign has obviously been successful. From 2009 onwards we have an underlying positive trend in sales; further, this positive trend is significant since the slope in the above fitted regression model is significant, having a p–value of 0.000 (3 d.p.).

(ii) Use the fitted model to estimate the trend for the fourth quarter of this year, 2013.

Answer:

We are treating Jan–Mar (2009) as time–point 1. Thus, Oct–Dec (2013) will be time–point 20, giving

$$\hat{y}^*_{20} = 90.428 + 2.58 \times 20 = 142.028,$$

i.e. £142,028.

(d) The following (edited) output has been obtained from `Minitab` for sales of *Sarajevski Pilsner*:

```
Seasonal Indices

 Period        Index

      1   ?   -25.75

      2      -15.625

      3   ?   +32.5

      4        8.875
```

One of the remaining seasonal effects is –25.75. Insert this in the correct space above, and then find the remaining seasonal effect. *[Use the space above to show your working, if required]*

(e) The residual series was obtained by subtracting the estimated trend in part (c), and the seasonal effects in part (d), from the original data (for observations from the year 2009 onwards). The company intends to use an autoregressive model to capture any remaining autocorrelation.

(i) Name the diagnostic plot that can be produced by `Minitab` to suggest the *order* of the autoregressive model.

Answer:

Either a correlogram/autocorrelation function, or a plot of the *partial* autocorrelation function.

(ii) After further investigations in `Minitab`, it was found that an autoregressive model of order 1, given by

$$y_t = 0.059 + 0.423(y_{t-1} - 0.059) + \epsilon_t,$$

was the most suitable model for the residual series. Use the fitted model to complete the output from `Minitab` overleaf (fill in the blank at `?`). *[Use the space underneath to show your working]*

Question 4 continued on next page

```
Forecasts from period 16


                                95% Limits
 Period       Forecast      Lower     Upper

     17       -1.33891   -8.64386   8.67585

     18       -0.53232   -9.11466   8.97027

     19       -0.19113   -9.17493   8.97753

     20   ?    -0.04680   -9.18593   8.97260
```

$$y_{20} = 0.059 + 0.423(-0.19113 - 0.059) = -0.04680.$$

(iii) Add the trend estimate you obtained in part c(ii), and the relevant seasonal effect from part (d), to the appropriate forecasted residual above to obtain a full forecast for sales of *Sarajevski Pilsner* in the fourth quarter (October–December) of 2013.

Answer:

$$\text{Full forecast} = -0.04680 + 142.028 + 8.875 = 150.8562,$$

i.e. £150,856.

(iv) Obtain a 95% confidence interval for your forecast in part (iii).

Answer:

$$\begin{aligned} \text{Lower bound} &: -9.18593 + 142.028 + 8.875 = 141.717, \\ \text{Upper bound} &: 8.97260 + 142.028 + 9.875 = 159.876, \end{aligned}$$

giving (£141,717, £159,876).

[20 marks]

5. (a) In the context of *Statistical Process Control*, briefly explain *level shift* and *instability*.

Answer:

Level shift: This is a change in the mean of the process distribution.

Instability: This is an increase in the standard deviation of the process distribution.

(b) You are an operations manager for *Service Air*, a manufacturing firm producing parts for aircraft. A current contract with *Boeing* requires the production of special screws for the pitot tubes for their new *Dreamliner* aircraft. These screws should be 22mm in length.

To determine whether your production process for these screws is under control, random samples of four screws are taken every 30 minutes over the period of 5 hours, and the length of each screw is measured.

(i) Complete the table of results shown below for the last sample.

Answer:

Sample (j)	1st	2nd	3rd	4th	$\bar{x}_j$	s_j
1	22.03	21.08	22.16	21.18	21.613	0.561
2	21.83	22.59	22.49	21.80	22.178	0.421
3	21.26	22.60	21.62	22.82	22.075	0.753
4	21.86	21.55	22.32	22.89	22.155	0.583
5	22.32	20.91	21.43	21.75	21.603	0.590
6	22.36	22.20	22.58	21.93	22.268	0.274
7	21.97	22.17	21.01	22.15	21.825	0.551
8	22.30	22.39	23.60	22.52	22.703	0.605
9	22.56	21.66	23.80	22.70	22.680	0.877
10	21.34	22.98	22.65	24.03	22.750	1.109

Question 5 continued on next page

(ii) It can be shown that $\bar{\bar{x}} = 22.185$ mm. Find the pooled standard deviation S, and hence find the 2–sigma control limits for an $\bar{x}$–chart.

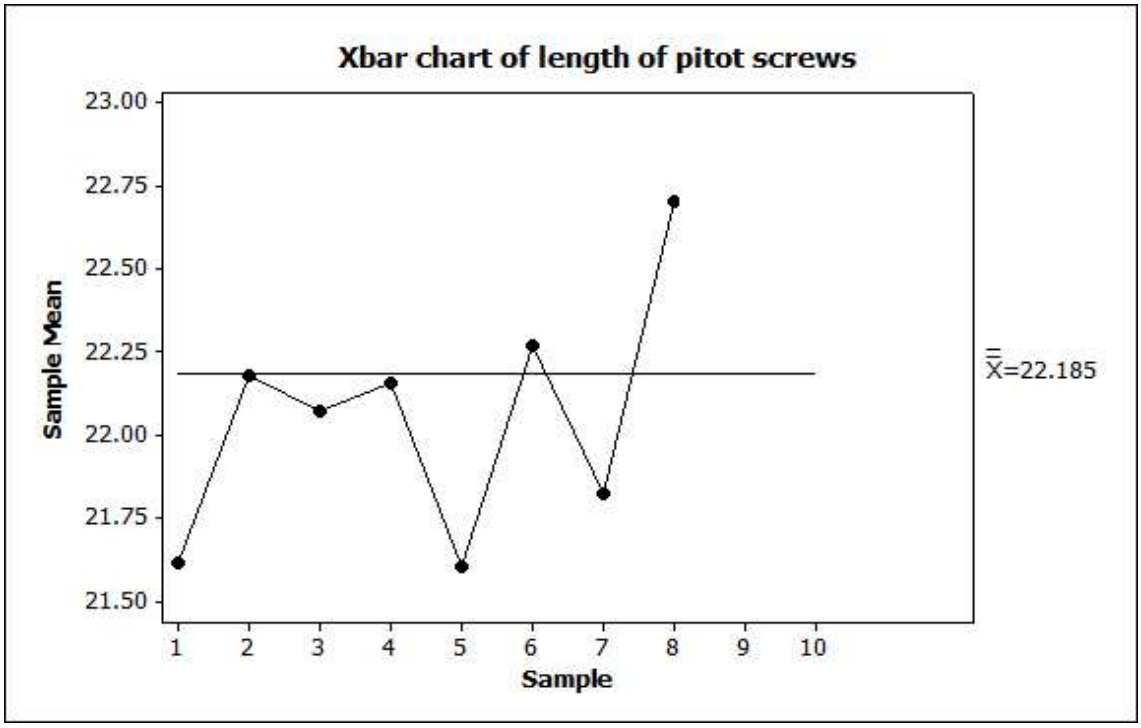
Answer:

$$S = \sqrt{\frac{\sum_{j=1}^{10} s_j^2}{10}} = \sqrt{\frac{0.561^2 + 0.421^2 + \ldots + 0.877^2}{10}} = \sqrt{\frac{4.49203}{10}} = 0.6702.$$

Thus, we have

$$\begin{aligned} LCL &= 22.185 - 2 \times \frac{0.6702}{\sqrt{4}} = 21.5184 \qquad \text{and} \\ UCL &= 22.185 + 2 \times \frac{0.6702}{\sqrt{4}} = 22.8552. \end{aligned}$$

(iii) Complete the graph produced by `Minitab` below.



(c) What is meant by the phrase *average run length*? Find the average run length in this example, and show that we might expect a *false alarm*, on average, about once every 11 hours.

Answer:

The *average run length* is the number of samples we can expect to be taken before the control chart indicates that the process may be out of control. In this example, we have

$$\begin{aligned} \Pr(\text{sample mean falls outside the control limits}) &= 2 \times \Pr(Z < -2) \\ &= 2 \times 0.0228 \\ &= 0.0456; \end{aligned}$$

thus, the average run length is

$$ARL = 1/0.0456 \approx 22 \text{ samples.}$$

Since we are told that samples are taken every 30 minutes, this would give us a false alarm , on average once every 11 hours, as required.

(d) The next two samples give

$$\bar{x}_{11} = 22.90\text{mm} \quad \text{and} \quad \bar{x}_{12} = 23.00\text{mm}.$$

(i) Update the $\bar{x}$–chart in part (b)(iii).

(ii) Without performing any further calculations, comment on whether or not you think the process has gone out of control.

Answer:

Looking at the updated chart, it would appear that the process has indeed gone out of control. However, we would need to recalculate the control limits based on these new samples, and these will shift.

[20 marks]

THE END