Section I. Multiple Choice Questions (單選題, 每題 4 分, 共 60 分)

1. An auditor inspects 25 transactions processed by the business office of a company. The auditor selects these transactions at random from the thousands that were processed in the most recent three months. In the past, the auditor has found 10% of transactions at this type of company to have been processed incorrectly.

    (a) Assuming Bernoulli trials, there is a 1/1000 chance for at least one error among the first three transactions that the auditor checks.

    (b) Assuming Bernoulli trials, the auditor would expect to find more than 3 errors among these transactions.

    (c) It would be unlikely for the auditor to discover more than 10 errors among these transactions because such an event lies more than 4 standard deviations above the mean.

    (d) All of the above.

    (e) None of the above.

2. Which of the following statement is correct?

    (a) In a simple random sample, every member of the population is equally likely to be selected.

    (b) The size of the sample for a survey should be a fixed percentage of the population size in order to produce representative results.

    (c) Voluntary response samples and convenience samples are likely to be unrepresentative.

    (d) All of the above.

    (e) Only A and C.

3. Which of the following statement is correct?

    (a) If variable X is associated with variable Y, then Y is caused by X.

    (b) A large chi-squared tells us that there is strong association between two categorical variables.

    (c) If the percentage of defective items produced by a manufacturing process is about the same on Monday, Tuesday, Wednesday, Thursday, and Friday, then the day of the week is associated with defective items.

    (d) All of the above.

    (e) Only B and C.

4. Which of the following statement is NOT correct?

    (a) A time series is a sequence of data that records an attribute at different times.

    (b) A Likert scale produces an ordinal variable even though numbers label the choices.

    (c) Nominal variables name categories without implying an ordering, whereas the categories of an ordinal variable can be ordered.

    (d) It makes sense to add or subtract ratio variables.

    (e) Interval variables allow multiplication and division.

5. Which of the following statement is correct?

    (a) The value of the t-statistic in a two-sample test does not depend on the units of the comparison. (We could, for example, measure the data in meters or centimeters.)

    (b) If the standard two-sample t-test reject $H_0 : \mu_1 - \mu_2 \leq \$100$, then we know FOR SURE that μ_1 is more than \$100 larger than μ_2.

    (c) If we double the size of both samples in a two-sample t-test, we reduce the chance of making type I error.

    (d) Pooling the two samples to estimate a common variance σ^2 avoids complications due to confounding.

    (e) None of the above.

● 可以使用不具儲存程式功能之電子計算機

6. A retailer maintains a Web site that it uses to attract shoppers. The average purchase amount is $100. The retailer is evaluating a new Web site that would, it hopes, encourage shoppers to spend more. Let μ represent the average amount spent per customer at its redesigned Web site.

   (a) The appropriate null hypothesis for testing the profitability of the new design is $H_0 : \mu > 100$.

   (b) If the level of significance $\alpha = 0.05$, then there is at most a 5% chance of incorrectly rejecting H_0.

   (c) If the p-value of the test of H_0 is less than α, then the test has produced a Type II error.

   (d) If the test used by the retailer rejects H_0 with $\alpha = 0.05$, then it would also reject H_0 with $\alpha = 0.01$

   (e) None of the above.

7. Which of the following statement is correct?

   (a) The sum of the residuals within a category used in an ANOVA is zero.

   (b) The F-test in an ANOVA tests the null hypothesis that all of the groups have equal variance.

   (c) The one-way analysis of variance requires balanced data, with an equal number of observations in each group.

   (d) All of the above.

   (e) None of the above.

8. As a baseline when planning future advertising, retail executive treat the dollar values of sales on consecutive weekends as iid (independently and identically distributed) random variables. The amounts sold on two consecutive weekends (call these X_1 and X_2) are independent and identically distributed random variables with mean μ and standard deviation σ. Then,

   (a) The standard deviation of total sales over consecutive weekends(連續兩個週末的總收入) is twice σ.

   (b) If a promotion were to introduce negative dependence between the amounts sold on two weekends, then we would need the correlation in order to find $E(X_2 - X_1)$.

   (c) If a promotion were to introduce negative dependence between the amounts sold on two weekends, then we would need the correlation or covariance in order to find $Var(X_2 - X_1)$.

   (d) All of the above.

   (e) None of the above.

9. The current age (in years) of 400 clerical employees at an insurance company is normally distributed with mean 38 and standard deviation 6. Then,

   (a) More employees at the company are older than 44 than between 38 and 44.

   (b) If the company were to convert these ages from years to days, then the ages in days would also be normally distributed.

   (c) If none of these employees leaves the firm and no new hires are made, then the distribution a year from now will be normal with mean 39 and standard deviation 7.

   (d) All of the above.

   (e) Only B and C.

10. Which of the following statement is NOT correct?

   (a) In a scatter plot, if all the data lie along a single line with nonzero slope, then the R^2 of the regression is 1. (Assume the values of the explanatory variable are not identical.)

   (b) Regression predictions become less reliable as we extrapolate farther from the observed data.

   (c) Prediction intervals get wider as you extrapolate outside the range of the data.

   (d) The simple linear regression model requires that a histogram of the response look like a normal distribution.

   (e) None of the above.

100097

●可以使用不具儲存程式功能之電子計算機

11. Which of the following statement is correct?

  (a) We usually use a bar chart to show shares and a pie chart to show frequencies of a categorical variable.

  (b) Pareto charts or Pareto diagrams are popular in quality control to identify problems in a business process.

  (c) The histogram balances on the mean.

  (d) All of the above.

  (e) Only B and C.

12. Coefficient of correlation

  (a) is a more easily interpreted measure of linear association than is the covariance.

  (b) does not have units.

  (c) always lies between -1 and 1

  (d) All of the above.

  (e) Only A and C.

13. Which of the following statement is NOT correct?

  (a) All other things the same, a 90% confidence interval is shorter than a 95% confidence interval.

  (b) By increasing the sampling size from n = 100 to n = 400, we can reduce the margin of error by 50%.

  (c) To guarantee a margin of error of 5% for the population proportion p, a survey needs to have at lest 500 respondents.

  (d) All of the above.

  (e) None of the above.

14. Which of the following statement is correct?

  (a) The best plot for checking the constant variance assumption (or condition) is the scatter plot of y on x.

  (b) A leveraged outlier (i.e., an influential point) has an unusually large or small value of the explanatory variable.

  (c) The presence of an outlier in the data used to fit a regression causes the estimated model to have a larger R^2 than it should.

  (d) In regression modeling, getting the highest R^2 is the most important issue.

  (e) None of the above.

15. Which of the following statement is correct?

  (a) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

  (b) $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

  (c) If event A is independent of event B, then $\mathbb{P}(A|B) = \mathbb{P}(B|A)$

  (d) If an event A is independent of the event B, then B is independent of A.

  (e) None of the above.

Section II. Problems (計算題與申論題, 可使用中文答題, 但請清楚寫出過程, 否則不予計分)

1. When fitting the regression of y on x for two groups, we can estimate the slope and intercept within each group either by fitting two simple regression or by fitting one multiple regression (in which a dummy variable is used). If simple regressions are so much easier to interpret, why combine them into one multiple regression? (8 %)

●可以使用不具儲存程式功能之電子計算機

2. Consider the partial printout for an interaction regression analysis of the relationship between a dependent variable y and two independent variables x_1 and x_2.

	SS	df	MS	F	p-value
Regression	3393.0	3	1131	9425	2.10E-11
Residual	0.7	6	0.12		
Total	3393.7				

	Coefficients	Std Error	t-value	p-value
Intercept	16.72	8.28	2.02	0.0900
x1	-3.04	2.68	-1.13	0.3001
x2	-1.05	1.55	-0.67	0.5239
x1*x2	4.08	0.45	9.16	9.47E-05

(1) Write the prediction equation for the interaction model. (3 %)

(2) Test the overall utility of the interaction model using the global F-test at $\alpha = 0.05$. (3 %)

(3) Test the hypothesis (at $\alpha = 0.05$) that x_1 and x_2 interact positively. (3 %)

(4) Estimate the change in y for each additional 1-unit increase in x_1 when $x_2 = 10$. (3 %)

3. The printout below shows part of the least squares regression analysis for the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ fit to a set of data. The model attempts to predict a score on the final exam in a statistics course based on the scores on the first two tests in the class.

	SS	df	MS	F	p-value
Regression	1293.1	2	646.55	21.27	2.36E-05
Residual	516.7	17	30.39		
Total	1809.8	19			

	Coefficients	Std Error	t-value	p-value
Intercept	-4.409	16.72	-0.26	0.7951
x1	0.397	0.343	1.16	0.2626
x2	0.638	0.385	1.66	0.1158

Is there evidence of multicollinearity in the printout? Explain. (8 %)

4. An important factor in selecting software for word-processing is the time required to learn how to use the system. To evaluate three word processing softwares, a firm designed a test involving four word-processing operators. Because operator variability was believed to be a significant factor, each of the four operators was trained on each of the three word processing softwares. The data (training time in hours) obtained follow.

Operator	Software I	II	III
1	19	19	16
2	17	13	15
3	15	15	15
4	13	13	10

(Hint: 本題之數字經特別設計以方便計算, 建議可透過平方和及其拆解之基本定義求算, 不必使用複雜的求解公式。)

(1) Construct the ANOVA table (6 %)

(2) Is there a significant difference in the mean training time (in hours) for the three softwares? (Please clearly state the null, the alternative hypothesis, and the test statistic, and use $\alpha = 0.05$. Note also that $F_{0.05}(2, 6) = 5.14$ and $F_{0.05}(3, 6) = 4.76$) (6 %)

100099