國立成功大學 108 學年度碩士班招生考試試題

系　　所：統計學系

考試科目：統計學

考試日期：0224，節次：3

※ 考生請注意：本試題不可使用計算機。請於答案卷(卡)作答，於本試題紙上作答者，不予計分。

一、單選題(請選出最適合問題的答案, 每題 3 分，共 48 分)

1. Adding additional independent variables in a regression equation will
   (A) increase the value of the coefficient of multiple determination.
   (B) improve the predictive power of the equation.
   (C) reduce the predictive power of the equation.
   (D) Additional independent variables may not be added to a regression equation.
   (E) none of the above

2. Dummy variables are usually used to incorporate _____ into a regression analysis.
   (A) dependent variables
   (B) nominal variables
   (C) interval/ratio variables
   (D) multicollinearity
   (E) beta coefficients

3. Null and alternative hypotheses are statements about:
   (A) population parameters.
   (B) sample parameters.
   (C) sample statistics.
   (D) it depends - sometimes population parameters and sometimes sample statistics.
   (E) estimation

4. Fisher's Exact Probability Test is used when:
   (A) The calculations for chi-square test are too difficult
   (B) You have more cells with expected frequencies of less than 5
   (C) You have a 3 by 2 contingency table
   (D) You have non-categorical data
   (E) For a given data set, the underlying distribution is normal

5. Based on a regression model, which of the following statistics is **NOT** design for detecting the outliers?
   (A) Cook's distance (B) Standardized residual (C) DFFITS
   (D) VIF(variance inflation factor) (E) Hat matrix

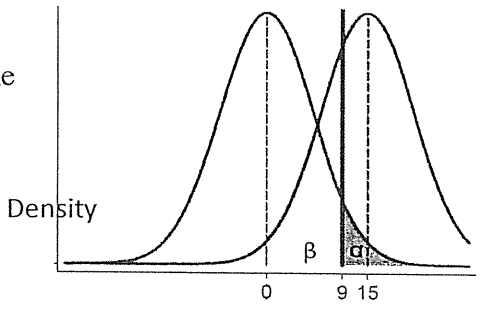
6. Early studies showed that the time between arrivals at the car wash follows the exponential distribution with mean 4 hours. What is the probability that no arrival during a two-hour period?
   (A) e^{-8} (B) e^{-2} (C) e^{-1} (D) $e^{-0.5}$ (E) $e^{-0.25}$

7. Suppose we have a sample of the heights of students and want to use the sample mean to get a confidence interval for the mean height in the population. Which of the following would increase the width of this confidence interval?
   (A) Increasing the sample size used to calculate the sample mean.
   (B) Switching from a 95% confidence interval to a 99% confidence interval.
   (C) Switching from a 95% confidence interval to a 90% confidence interval.
   (D) Increasing the sample mean.
   (E) None of the above.

8. Suppose X ~Uniform[0, 4]. Suppose a random sample of 100 observations is taken from each population. How is the sample mean of X distributed?
(A) Uniform[0, 4] (B) Uniform[0, 0.4] (C) N(0, 0.4) (D) N(2, 1.333) (E) N(2, 0.013)

9. Simpson's Paradox occurs when
(A) No baseline risk is given, so it is not known whether or not a high relative risk has practical importance.
(B) A confounding variable rather than the explanatory variable is responsible for a change in the response variable.
(C) The direction of the relationship between two variables changes when the categories of a confounding variable are taken into account.
(D) The results of a test are statistically significant but are really due to chance.
(E) This is not a randomized experiment.

10. The following diagram graphically describes the probability of type I error (α) and the probability of type II error (β) of a hypothesis test regarding the population mean μ when the sample size is 100. Given the significance level α, the critical value is 9. Choose a set of hypotheses from the following alternatives that corresponds to this diagram.
(A) $H_0: \mu = 15, H_a: \mu < 15$
(B) $H_0: \mu = 15, H_a: \mu > 15$
(C) $H_0: \mu = 0, H_a: \mu \neq 0$
(D) $H_0: \mu = 0, H_a: \mu > 0$
(E) $H_0: \mu = 0, H_a: \mu < 0$



11. Which statement(s) is (are) **CORRECT**?
(I) The difference between the sample mean and the population mean is called the margin of error.
(II) The difference between the upper limit of a confidence interval and the point estimate used in constructing the confidence interval is called the margin of error.
(III) The margin of error equals to half the width of a confidence interval.
(IV) The sampling error can either be positive or negative.
(A) I , III (B) I, II, III (C) II, III, IV (D) II, IV (E) I, II, III, IV

12. In a survey sampling, what is the smallest sample size n required if the margin of error (suppose α is set to be 0.05), in estimating the population proportion p, is set to be less than 0.03? Assume that $0.2 \leq p \leq 0.4$.
(A) 1068 (B) 1025 (C) 1000 (D) 996 (E) None for all the above four.

13. From a population with a variance of 900, a sample of 225 items is selected. At 95% confidence, the margin of error for estimating the population mean is
(A)15 (B) 2 (C) 3.28 (D) 4 (E) 3.92

14. A normal population has mean μ and standard deviation 12. The hypotheses to be tested are $H_0: \mu = 40$ versus $H_1: \mu \neq 0$. Which would result in the highest probability of a Type II error?
(A) $\mu = 38; n = 10$ (B) $\mu = 38; n = 100$ (C) $\mu = 39; n = 10$
(D) $\mu = 42; n = 10$ (E) $\mu = 43; n = 100$

15. Suppose we have a random variable X where $P(X = k) = \binom{15}{k} (.29)^k (.71)^{15-k}$, for $k = 0,1,\ldots,15$. What is the mean of X?
(A) 0.29 (B) 0.71 (C) 4.35 (D) 10.65 (E) None of the above

國立成功大學 108 學年度碩士班招生考試試題

系　　所：統計學系

考試科目：統計學

考試日期：0224，節次：3

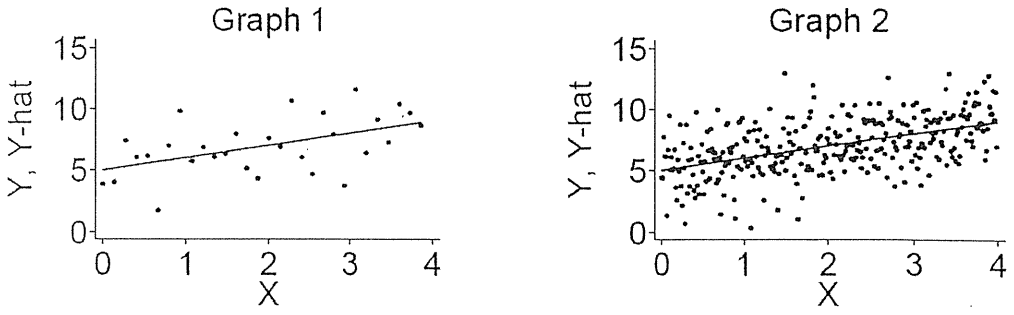
16. Suppose the R^2 for a bivariate regression is equal to 1. This tells us that:
    (A) The correlation between the dependent and independent variables is equal to 1.
    (B) The slope coefficient is equal to 1 or -1.
    (C) The error sum of squares is equal to the total sum of squares.
    (D) The dependent and independent variables are perfectly correlated.
    (E) A smaller F statistic for the regression.

二、填充題(每題 4 分，共 28 分)

1. A retailer of electronic equipment received six VCRs from the manufacturer. Three of the VCRs were damaged in the shipment. The retailer sold two VCRs to two customers. What is the probability that both customers received damaged VCRs? _____

2. Scores on a statistics exam were normally distributed with a mean of 80 and a standard deviation of 6. If the top 2.5% of test scores receive merit awards, what is the lowest score eligible for an award? _____

3. Consider these two graphs.



Compared to Graph 1, Graph 2 clearly shows a regression analysis where the coefficient of determination is _____ (**smaller**, **similar**, or **larger**) and the F test statistic is _____ (**smaller**, **similar**, or **larger**).

4. A study of 1,000 families gave the following results:
   $\bar{X} = 68$: average height of husband in inches, $s_X = 2.7$ inches;
   $\bar{Y} = 63$: average height of wife in inches, $s_Y = 2.5$ inches;
   $s_{XY} = 1.6875$ inches
   If her husband is 72 inches tall, on average the height of a wife in inches is_____

5. A researcher wants to know whether men and women in a particular community differ in their political party preferences. She collects data from a random sample of 200 registered voters, and observes the following:

political party	Party A	Party B
Male	55	65
Female	50	30

   To answer the question "Do men and women significantly differ in their political preferences", the appropriate test statistic is _____.

6. Let X have range [0,3] and density $f_x(x) = kx^2$, Find the 12.5th percentile of X.

7. A sample of 41 observations yielded a sample standard deviation of 5. If we want to test $H_0: \sigma^2 = 20$, the test statistic is _____.

國立成功大學 108 學年度碩士班招生考試試題

系　　所：統計學系

考試科目：統計學

考試日期：0224，節次：3

三、計算問答題 (請於答案卷(卡)作答，2 大題，共 24 分)

1. (12%) An experiment is conducted to study the treatments of diabetes. Twenty-five subjects are randomized to the five treatment groups described below. The blood sugar levels are checked to make sure that the patients are under control.

Treatment Group	Sample mean	Sample std. deviation
1	43	4.4
2	51	9.7
3	39	12.1
4	46	11.2
5	23	?

(A) Specify a model for this experiment that allows for treatment effects. State all the usual assumptions required for the analysis of variance.

(B) The error mean square is given by MSE= 90. Deduce the sample variance of treatment group 5.

(C) We know that the sum of squares due to the treatment is SSt=2181. Using the summary statistics above, complete an appropriate ANOVA table. Formulate and test the hypothesis that the mean longevity is constant across the five treatment groups. Use level $\alpha = 0.05$ and draw a brief conclusion.

(D) Construct and estimate a contrast that compares sample mean for group 4 to that for group 5. Report a standard error. Do the two sample mean differ significantly?

2. (12%) The economic structure of Major League Baseball (MLB) allows some teams to make substantially more money than others, which in turn allows some teams to spend much more on player salaries. These teams might therefore be expected to have better players and win more games on the field as a result. Over the course of the last 4 years (2015-2018), each of the 30 MLB teams were measured each year and the following data were collected for these 120 observations:

wins: number of games the team won for a specific year

payroll: opening day payroll, in millions of dollars, for the team for a specific year

year: the year in which the measurement was taken

A regression model was run to predict the number of wins a team had from the ln(payroll), and the results are shown below, along with some summary statistics:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.076     11.654   3.782 0.000246 ***
log(payroll) 7.963      2.505   3.179 0.001887 **
---
Residual standard error: 10.5649 on 118 degrees of freedom
Multiple R-squared: 0.06250, Adjusted R-squared: 0.06709
F-statistic: 10.11 on 1 and 118 DF, p-value: 0.001887
```

國立成功大學 108 學年度碩士班招生考試試題

系　　所：統計學系

考試科目：統計學

考試日期：0224，節次：3

The mean and standard deviation of log(payroll) are 4.6371 and 0.3867.

(A) What is the estimated correlation between wins and log(payroll)?

(B) Calculate $\bar{Y}$, the mean number of wins, and s_Y, the standard deviation for the number of wins among the 120 observations.

(C) The Red Sox are projected to spend about $180 million dollars on payroll in 2019. Provide a formula for the 95% interval for the number of games they will win next year (You don't need to calculate it).

(D) List out all of the assumptions for this regression model, and comment on how to check if they are reasonable.

參考值:

$Z_{0.2/2} = 0.842$
$Z_{0.1/2} = 1.645$
$Z_{0.05/2} = 1.96$
$Z_{0.01/2} = 2.576$

F(0.05, 4, 25) = 2.76
F(0.05, 5, 25) = 2.60
F(0.05, 4, 24) = 2.78
F(0.05, 5, 24) = 2.62
F(0.05, 4, 20) = 2.87
F(0.05, 5, 20) = 2.71

F(0.025, 4, 25) = 3.35
F(0.025, 5, 25) = 3.13
F(0.025, 4, 24) = 3.38
F(0.025, 5, 24) = 3.15
F(0.025, 4, 20) = 3.51
F(0.025, 5, 20) = 3.29