國立成功大學

111學年度碩士班招生考試試題

編　號：236

系　所：統計學系

科　目：統計學

日　期：0220

節　次：第 3 節

備　註：不可使用計算機

編號： 236

國立成功大學 111 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：0220，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。

A. True or False (Please answer "T" for true and "F" for false statement, 2% × 20 = 40%)

1. A statistic is a function of the data.
2. Bar chart is an appropriate tool to describe the household income.
3. The lower and upper bound of a confidence interval are parameters.
4. The intersection of two independent sets is the empty set.
5. The assumption of normality is not required to estimate the slope with least-square method in linear regression.
6. The data would be left-skewed if the median is less than the mean.
7. The central limit theorem guarantee that the sample would be approximately normally distributed when the sample size is large enough given the population variance is finite.
8. The primary purpose of the cluster sampling design is to save the sampling cost.
9. If X is a discrete random variable, then the value of its density function cannot be greater than 1.
10. The median of a distribution is unique.
11. The one-sample t-test to examine the population mean μ can be used even if the population is not normally distributed as long as the sample size n is large enough, so can the one-sample test for population variance based on $\frac{(n-1)S^2}{\sigma^2}$, where S^2 is the sample variance,
12. Let U^{α} be the upper α quantile, that is $P(U > U^{\alpha}) = \alpha$. Suppose that $X \sim F_{m,k}$, then $X^{1-\alpha} = (Y^{\alpha})^{-1}$, where $Y \sim F_{k,m}$.
13. The geometric distribution in the Bernoulli process is an analogy of the Exponential distribution in the Poisson process.
14. The nonparametric version of the parametric one-sample t-test is the Mann-Whitney test.
15. The daily temperature is an example of interval-scale variable.
16. The independence between two events A and B cannot be guaranteed by neither the independence between A^c and B, nor the independence between A and B^c.
17. If random variables X, Y are independently and identically distributed as the standard normal distribution, then the median of $\frac{X}{Y}$ is zero.
18. The household income data is often heavily right-skewed, hence a better location measurement of the center of the data would be the median.
19. At least about 89% of the data would be located within 4 standard deviations of the sample mean is guaranteed by the Chebyshev's Theorem.

20. If the correlation coefficient between random variables X, Y is 1, then X can be expressed as $X = aY + b$, $a > 0$ for sure, where a and b are arbitrary constants.

B. Blank Filling (Please provide your answer orderly as (A), (B), ...,(L). 3x12 = 36%)

1. A factory would like to evaluate the efficiency of three different production methods, denoted as A, B, and C. Each of these three methods are used to produce the products for 10 production periods. The amounts of products produced are recorded, the data is

	Amount	Method
1	500	B
2	300	C
3	200	A
4	100	A
5	400	C
6	400	A
.	.	.
.	.	.
.	.	.
30	200	B

The sum of squares $\sum(y_i - \bar{y})^2$ is 441667 of the 30 items of data. One would like to evaluate if there are significant difference among these three methods, and simple linear regression with dummy coding is used. The coding system is:

Method	X1	X2	X3
A	1	0	0
B	0	1	0
C	0	0	0

The followings are the results of the regression analysis and the associated ANOVA table, with a few missings:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	350.00	34.69	10.088	1.18e-10 ***
X1	-150.00	49.07	-3.057	0.00499 **
X2	-50.00	49.07	-1.019	0.31722

Residual standard error: 109.7 on 27 degrees of freedom

Multiple R-squared: 0.2642, **Adjusted R-squared: (A)**

國立成功大學 111 學年度碩士班招生考試試題

Analysis of Variance Table
Response: Amount

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	104167	104167	8.6538	0.006625 **
X2	1	12500	12500	1.0385	0.317224
Residuals	27	**(B)**	12037		

Please answer the following questions: (For the non-integer answers, please round to the second digit after the decimal point)

i. Please fill in the missing items __(A)__ and __(B)__ of the results above.

ii. The estimated average amount of products produced by method C is __(C)__, and the estimated overall average amount produced by these three methods is __(D)__.

iii. The value of the test statistic to examine if these is any different among these methods is __(E)__, and it should be compared to the __(F)__ distribution for a proper conclusion.

iv. If one would like to further examine if any pair of two methods are significantly different with the procedure of Fisher's least significant distance, the Least Significant Distance is __(G)__ (The exact number is not required, a proper expression would be good enough). If the significant level is set to be $\alpha = 0.1$ for each pairwise comparison, the overall type-I error rate would be __(H)__.

2. One would like to study if the beer preference is independent of gender, and data of 100 beer drinkers are collected, the data yields:

		Beer Preference (Z)	
		Light; $Z = 1$	Dark; $Z = 2$
Gender (Y)	Female; $Y = 1$	12	13
	Male; $Y = 2$	24	51

A test to examine $H_o: p_1 = P(Z = 1|Y = 1) = P(Z = 1|Y = 2) = p_2$ against Ha: $p_1 \neq p_2$ is performed, and the value of the test statistic $(\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot (1/n_1 + 1/n_2)} = 1.44$, which gives a p-value of 0.15.

Suppose we reevaluate this data with the goodness-of-fit test to examine if beer preference is independent of gender, please answer the following questions:

i. The expected counts of the female beer drinkers who prefer light beer under H_o is __(I)__.

ii. The value of the Pearson residual of the female beer drinkers who prefer dark beer under H_0 is __(J)__.

iii. The value of the test statistic of this goodness-of-fit test is __(K)__.

iv. The p-value of the test statistic of this goodness-of-fit test is __(L)__.

C. Short Essay Question (24%)

1. (8%) For the two hypothesis testing methods in Question 2, Part B, please explain the reason why you get the same or different conclusions.

2. (8%) Pease explain why the width of the **prediction band** is wider than which of the **confidence band** in linear regression model **without** using any formula.

3. (8%) Based on the data collected from her department, Jane obtained the 95% confidence interval of the daily expenditure of the students in her department as 250 NTD to 500 NTD based on the sampling distribution of the sample mean. Please interpret the meaning of this confidence interval.