考試科目	統計方法	系所別	統計學系	考試時間	2 月 6 日 (五) 第 4 節

Note:

Let $\Phi(z)$ be the cumulative distribution function of a standard normal random variable Z. Please express your answer in $\Phi(\cdot)$ or $\Phi^{-1}(\cdot)$. Please define percentiles of a distribution whenever needed, and express your answer in terms of the notations. For example, define $t_{0.05,\text{df=30}}$ be the number satisfying either $P(T_{\text{df=30}} \le t_{0.05,\text{df=30}}) = 0.05$ or $P(T_{\text{df=30}} \ge t_{0.05,\text{df=30}}) = 0.05$.

一、填充題 (30%)

註：填充題只需寫出答案，不需計算過程。

1. Consider the following hypothesis test: H_0: $\mu = 2026$ v.s. H_a: $\mu \neq 2026$. The population variance is 400. Using a 0.05 level of significance.
   (a) (5%) A sample of 100 provided a sample mean of 2025. What is the p-value of the test?
   (b) (7%) What is the probability of making a Type II error when the actual population mean is 2027.36 ?

   Note: You may use some of the following quantities for computing #1(b).
   $z_{0.001} = 3.09$, $z_{0.005} = 2.575$, $z_{0.01} = 2.33$, $z_{0.025} = 1.96$, $z_{0.05} = 1.645$, $z_{0.1} = 1.28$.

2. (13%) Consider the probability density function
$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}.$$
   A random sample of size 5 is taken from $f(x)$, and the order statistics are denoted by $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, $X_{(4)}$, and $X_{(5)}$. Let the observation of $X_{(5)}$ be y_5. Conduct a test H_0: $\theta = 1$ v.s. H_a: $\theta \neq 1$ by rejecting H_0 if
$$y_5 \le 0.5 \text{ or } y_5 \ge 1.$$
   Please find the power function of the test.

3. (5%) To estimate a population proportion, the sample proportion is usually adopted as the point estimator. What is the probability that the sample proportion will be within $\pm 3\%$ of the population proportion? Suppose that the population proportion is .35, and the sample size is 100.

備註	一、作答於試題上者，不予計分。 二、試題請隨卷繳交。

考試科目	統計方法	系所別	統計學系	考試時間	2 月 6 日 (五) 第 4 節

二、是非題 (15%)

For each of following statements, answer True or False. For each of false statements, please explain why it is false. (請簡單扼要說明)

4. (5%) If the alternative hypothesis is true, then the p-value will always be smaller than the significance level.

5. (5%) The power of a test may be less or greater than the significance level.

6. (5%) The significance level is greater than or equal to the probability of a type I error.

三、計算、推導或簡答題 (55%) 註：請寫出計算過程，否則不予計分。請清楚標示最後的答案。

7. Suppose that a random sample from a normal distribution $N(\mu, \sigma^2)$ results in the following observations: 79, 82, 63, 75, 83, 99, 71, 76, 77, and 85.
   (a) (5%) Construct a 95% confidence interval for σ^2 if μ is unknown.
   (b) (5%) Suppose that μ is known to be 80. Construct a 95% confidence interval for σ^2, using the information of μ. What is your interpretation about this interval estimate?

8. The *Avocado Price* dataset from Kaggle and the Hass Avocado Board website comprises 18,249 observations. It includes several key variables: volume (total sales volume ; Y), price (X_1), date, type, season, year, region (Atlanta, Boston, SanDiego, …), and so on. In the current data analysis, we ignore the time information, limit our dataset to Boston and San Diego (338 observations in total), and conduct the following linear regression analysis.

   Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
   Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
   Model 3: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$

   As the first step, we consider a simple linear regression, Model 1, on the 338 observations, and the results is given in the following tables.

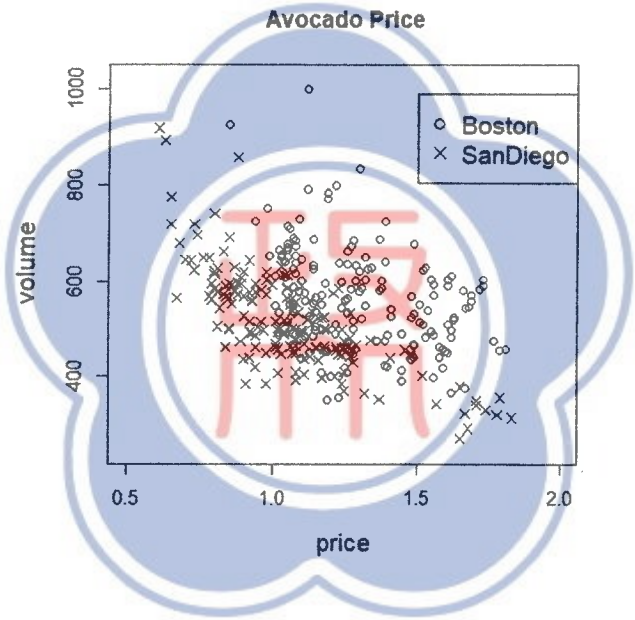
Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Regression	623,173	1	(A2)	(A5)
Error	3,319,382	336	(A3)	
Total	(A1)	337	(A4)	

備註	一、作答於試題上者，不予計分。 二、試題請隨卷繳交。

考試科目	統計方法	系所別	統計學系	考試時間	2 月 6 日 (五) 第 4 節

Predictor	Coef	SE Coef
Constant	732.36	24.90
price	−163.18	20.55

(a) (7%) If the model is fitted directly without preliminary exploratory data analysis, such as scatter plots, it yields the unsatisfactory R^2. The investigation of the scatter plot, given below, of Y versus X_1 for Boston (marked as circle) and SanDiego (marked as cross) suggests us to extend Model 1 to improve the R^2. Please suggest <u>one</u> reasonable extended model, as the extension of Model 1, and clearly define every notation or variable that you use. Also, provide the reason for the extension.



(b) (7%) Consider the above Model 2, and suppose that X_2 is some other variable in the dataset. Is it possible that the R^2 under Model 1 and Model 2 take on identical values? Explain.

(c) (6%) Model 3 includes three additional predictors, X_3, X_4, and X_5, which are associated with Y. Conduct a global test of hypothesis to determine whether any of the regression coefficients under Model 3 are not zero. Use a 0.05 level of significance. The answers should include the followings. Specify your null and alternative hypotheses. What is your conclusion about the test? What is the implication/interpretation about the conclusion of the F test? (The critical value for the global test is 2.39877.)

備註	一、作答於試題上者，不予計分。 二、試題請隨卷繳交。

考試科目	統計方法	系所別	統計學系	考試時間	2 月 6 日 (五) 第 4 節

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Regression	1,093,708	(A8)	(A11)	(A12)
Error	(A6)	(A9)	8,555	
Total	(A7)	(A10)		

Predictor	Coef	SE Coef
Constant	736.57	24.76
price	−211.30	21.68
X_3	61.76	12.88
X_4	76.73	14.27
X_5	135.90	21.25

9. *Time Series Analysis* focuses on the values of a subject over time or certain space. For example, we may be interested in the five-year historical daily stock closing prices (股票收盤價) of a given company. A sequence of random variables $X_1, X_2, \ldots, X_n$ is used to denote the values at time points $1, 2, \ldots, n$ (suppose that we have regular time points).

In order to measure the covariance between different time points, a so-called autocovariance function (ACF) at time point s and t, for $1 \le s, t \le n$, is defined and considered:

$$\gamma_x(s,t) = E[(X_s - \mu_s)(X_t - \mu_t)],$$

where $\mu_s = E(X_s)$. Since we simply obtain one realization of time series in real data, some stationary assumption is necessary for real data analysis. The following is the definition of a weakly stationary process:

(i) $E(X_t)$ is constant and does not depend on time t;

(ii) the ACF $\gamma_x(s,t)$ depends on s and t only through their difference $|s-t|$.

The computation of the sample ACF is a crucial step before fitting some time series models. Sample ACF:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}),$$

where $\bar{x}$ is the sample mean.

Please answer Questions (a) to (d).

備註	一、作答於試題上者，不予計分。 二、試題請隨卷繳交。

考試科目	統計方法	系所別	統計學系	考試時間	2 月 6 日 (五) 第 4 節

(a) (5%) Consider the time series $X_t = \beta_0 + \beta_1 \cdot t + Z_t$, where β_0 and β_1 are known constants and Z_t is a sequence of i.i.d. random variables with mean 0 and variance σ^2. Determine whether X_t is weakly stationary.

(b) (5%) Derive the ACF for the following process
$X_t = Z_t + a \cdot Z_{t-1} + b \cdot Z_{t-2}$, where a and b are parameters, and Z_t is a sequence of i.i.d. random variables with mean 0 and variance σ^2.

(c) (8%) Suppose that we have one observed time series of length 30, denoted by $x_1, x_2, \ldots, x_{30}$, and the value of x_3 and x_{15} are missing. That is, what we observed: $x_1, x_2, x_4, \ldots, x_{14}, x_{16}, \ldots, x_{30}$. In order to compute the sample ACF for this dataset which is with missing data, a student suggested to directly ignore what was missing and join what were observed. In order words, let $y_1 = x_1$, $y_2 = x_2$, $y_3 = x_4, \ldots, y_{13} = x_{14}, y_{14} = x_{16}, \ldots y_{28} = x_{30}$, and the student computed sample ACF on the series $\{y_t\}$ for the sample ACF of X_t. Please comment on the approach for missing data by the student. (Is it appropriate, or are there some drawbacks? Explain.)

(d) (7%) Please suggest an alternative approach to handling the missing data in (c) when computing the sample ACF, and explain its advantages.

備註	一、作答於試題上者，不予計分。 二、試題請隨卷繳交。