# UNIVERSITY COLLEGE LONDON

## EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **MATH7501**

ASSESSMENT PATTERN : **MATH7501A**

MODULE NAME : **Probability and Statistics**

DATE : **18-May-11**

TIME : **10:00**

TIME ALLOWED : **2 Hours 0 Minutes**




**TURN OVER**

*All questions may be attempted but only marks obtained on the best* four *solutions will count.*
*The use of an electronic calculator* is *permitted in this examination.*

New Cambridge Statistical Tables are provided.

1. (a) State the definition of independence of two events $E$ and $F$.
   (b) What does it mean to say that events $E$ and $F$ are disjoint?
   (c) Suppose that $A_1$, $A_2$, $A_3$ are three events such that $P(A_1 \cap A_2 \cap A_3) > 0$. Show that
   $$P(A_1 \cap A_2 \cap A_3) = P(A_3|A_1 \cap A_2)P(A_2|A_1)P(A_1).$$
   (d) Assuming that $P(A) > 0$ and $P(B) > 0$, show that if $P(A|B) > P(A)$ then $P(B|A) > P(B)$.
   (e) A fair coin is tossed three times. Let $X$ be the number of heads in the first two tosses, and let $Y$ be the number of heads in the last two tosses.
      (i) Show that the events $\{X = 0\}$ and $\{Y = 1\}$ are independent.
      (ii) Determine whether the events $\{X = 0\}$ and $\{Y = 2\}$ are independent or dependent.
      (iii) What does it mean to say that two discrete random variables are independent? Are $X$ and $Y$ independent?

2. (a) (i) Define a partition of a probability space $\Omega$.
      (ii) Show that if $\{B_1, B_2, \ldots, B_n\}$ is a partition such that $P(B_i) > 0$ for all $i = 1, \ldots, n$, then for any event $A$, $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$.
      (iii) Assuming that $P(A) > 0$, show further that $P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$ for $j = 1, \ldots, n$.
   (b) A hamster lives in a cage and is looked after by three brothers, $B_1$, $B_2$ and $B_3$. The probabilities of each of the brothers taking the hamster out to play are 0.2, 0.5 and 0.3. Every time the hamster is taken out, there is a danger it gets trodden on by the brother who took it out. The probabilities of the hamster being trodden on by $B_1$, $B_2$ or $B_3$ are 0.3, 0.4 and 0.3. One day the hamster is found as flat as a pancake on the carpet, but the brothers are nowhere to be found. Which brother is most likely to be the culprit?

3. Assume that $X$ follows a Geometric distribution with parameter $p \in (0,1)$. That is,

$$P(X = i) = (1-p)^{i-1}p \quad \text{for} \quad i = 1, 2, \ldots$$

(a) Show that its probability generating function, $\Pi_X(s) = E(s^X)$, is given by

$$\Pi_X(s) = \frac{ps}{1 - s(1-p)} \quad \text{for} \quad s \in [0,1].$$

(b) Using $\Pi_X(s)$, derive the mean and variance of $X$.

(c) Fix $k$ as a positive integer, find $P(Y \geqslant n | X \geqslant k)$ where $Y = X - k + 1$ and $n = 1, 2, 3, \ldots$, and state the distribution of its argument.

(d) Find $E(Y^2 | X \geqslant k)$.

4. Let $X_1, \ldots, X_n$ be *iid* random quantities, each having probability density function

$$f(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), \; 0 < x < \infty.$$

Suppose that $T = T(X_1, \ldots, X_n)$ is an estimator of the unknown parameter $\theta$, and denote the bias of the estimator $T$ by $b(T, \theta)$ and the mean square error of $T$ by $MSE(T)$.

(a) State the definition of both $b(T, \theta)$ and $MSE(T)$.

(b) Show that $MSE(T) = Var(T) + b^2(T, \theta)$.

(c) Given the $n$ random quantities defined above, each following $f(x)$, find an estimator $T$ of $\theta$.

[Hint: find a simplified expression for $L(\theta) = \prod_{i=1}^{n} f(x_i)$, and then take the logarithm of $L(\theta)$. Let $l(\theta)$ be the logarithm of $L(\theta)$, an estimator $T$ of $\theta$ can be found by differentiating $l(\theta)$ with respect to $\theta$, setting the result to zero and solving for $\theta$.]

(d) Assume that $E(X_i) = \theta$ and $E(X_i^2) = 2\theta^2$. Using the results above find the bias and mean square error of the estimator found in (c).

5. Let $X_1, \ldots, X_n$ be *iid* $N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_n$ be *iid* $N(\mu_Y, \sigma^2)$ and suppose that the $X_i$ are independent of the $Y_j$. Let us define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2,$$

and assume that these four random quantities are all mutually independent.

(a) State the definition of the $t$-distribution with $n$ degrees of freedom.

(b) Explain why $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{s_P\sqrt{\frac{1}{n}+\frac{1}{m}}}$ follows a distribution $t_{n+m-2}$, where $s_P^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$.

(c) A study was carried out to investigate whether or not there was a difference in the examination score between two groups of students. At the beginning of the course, each student was randomly allocated into one of two groups. Students in the first group were taught using computer based methods whilst students in the second group via traditional lectures. The data are as shown below

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 89 | 81 | 92 | 94 | 74 | 56 | 77 | 85 | 78 | 69 | 89 | 88 | 45 | 83 | 95 |
| $y_i$ | 91 | 57 | 89 | 83 | 80 | 83 | 91 | 84 | 84 | 94 | 96 | 88 | 97 | 91 | 94 |

where $x_i$ denotes the examination score for the $i^{th}$ student taught using computer based methods and $y_i$ the examination score for the $i^{th}$ student taught via traditional lectures. The data can be summarised as follows:

$$\sum_{i=1}^{15} x_i = 1195, \quad \sum_{i=1}^{15} x_i^2 = 97997, \quad \sum_{i=1}^{15} y_i = 1302, \quad \sum_{i=1}^{15} y_i^2 = 114344.$$

(i) Stating any assumptions you make, test the hypothesis that the underlying mean examination score is the same in both groups against the hypothesis that they differ, at the 5% significance level.

(ii) Under similar assumptions, derive a 95% confidence interval for $\delta$, the difference in the underlying mean examination score for the two groups. Comment briefly upon the relationship between the confidence interval you derive and the hypothesis test you performed in part (c)(i).

6. Assume that you wish to analyse, using a linear regression, the relationship between a response variable $Y$ and an explanatory variable $x$, employing observations $\{y_i, x_i;\ i = 1, \ldots, n\}$.

   (a) Show that the *residuals* $r_i$ defined by $r_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ with estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, have the following properties:

$$\sum_{i=1}^{n} r_i = 0. \qquad \sum_{i=1}^{n} x_i r_i = 0.$$

   (b) Why the estimate of $\sigma^2$ has $n-2$ degrees of freedom in a regression problem?

   (c) Given independent $n$ observations, find the least squares estimate of $\beta$ in the linear model $y_i = \mu_i + \epsilon_i$, where $\mu_i = \beta$ and the $y_i$ are from distributions with the same mean and constant variance.

   (d) Which of the following common linear model assumptions are required for $\widehat{\beta}$ to be unbiased: (i) The $Y_i$ are independent, (ii) the $Y_i$ all have the same variance, (iii) the $Y_i$ are normally distributed?

   (e) Consider the following data

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $l_i$ | 0.2 | 1 | 3 | 4 | 5 | 7 |
| $h_i$ | 0.02 | 0.1 | 0.4 | 0.5 | 0.6 | 0.9 |

   where $l_i$ and $h_i$ denote the length and number of hours for the $i^{th}$ train journey. (i) Produce a plot of the data above. (ii) On the basis of your plot, propose a model that represents the pattern in the data and that is suitable to predict $h_i$. (iii) Recalling that the average speed of an object in an interval of time is the distance traveled by the object divided by the duration of the interval, find a least squares estimate of the mean journey speed.

   (f) A laboratory experiment was carried out to determine a calibration relationship that can be used to predict fat content, $y_i$, from light absorption, $x_i$. The summary statistics for the experiment were as follows:

$$\begin{aligned} n &= 20 \\ \bar{x} &= 0.11, \quad \bar{y} = 1.25 \\ C_{xx} &= 0.083, \quad C_{xy} = 0.45, \quad C_{yy} = 7.12. \end{aligned}$$

   (i) Use the method of least squares to fit a regression line to these data. (ii) On the assumption that the observations are independent and are normally distributed about the regression line with common variance $\sigma^2$, determine 95% confidence intervals for the slope and intercept of the regression line, and test the null hypothesis that the regression line passes through the origin $x = y = 0$. (iii) If a light absorption of 0.9 is obtained, what is the predicted concentration of fat content? (iv) How useful is this procedure for predicting fat content?

MATH7501 END OF PAPER