Indian Statistical Institute
Bangalore Centre
B.Math (Hons.) II Year 2018-2019
Back paper Examination

Statistics I Date 03 January 2019

Answer as many questions as possible. The maximum you can score is 100.
The notation used have their usual meaning unless stated otherwise.
State clearly the assumptions you make and the results you use.
No numerical computation is required. It is enough to present you answer as an algebraic expression.

1. A coin was tossed three times and the results noted. This process was continued 100 times, i.e. there were 100 sets of three tosses. In 69 cases first toss showed head, in 49 cases heads obtained in second trial and in 53 cases third trial showed head. In 33 cases heads were obtained in both first and second tosses and in 21 cases both second and third trial resulted in head.

   Show that there were at most 15 occasions when tails occurred all three times. [5]

2. (a) Define mode of data set.

   (b) 100 pebbles were collected from a sea beach. The masses of them were measured and summarized in a table. Suppose there are k classes, the ith class has boundaries a_{i-1} and a_i and frequency f_i, $1 \leq i \leq k$. If the rth class has maximum frequency and $f_{r-1} > f_{r+1}$, show how you can find the mode of the collected pebbles. [2 + 6 = 8]

3. (a) When is an unimodal probability density function is said to be symmetric, positively skewed or negatively skewed ? Illustrate with graph.

   (b) Show that if a unimodal probability density function is positively skewed, then the mean is greater than the mode. [4 + 6 = 10]

4. In a study the Systolic blood pressure of n male workers in a factory in the age group 30-40 were observed. The data had mean $\bar{X}$ and variance V. Using appropriate statistical tables,

   (a) estimate the probability that a randomly selected worker would have Systolic blood pressure between 100 and 130 and

   (b) obtain a 95% confidence interval for the true variance of the Systolic blood pressure of all males in the age group 30-40. [3 + 5 = 8]

5. A chemist wants to study the relationship between the drying time of a paint and the concentration of a chemical additive. k concentration levels of the chemical additive were selected. These were added to the paint, which were used to paint $n(> k)$ metallic plates of the same size. Then the drying time of the metallic plates were noted.

   (a) To begin with she tried a linear model. Describe the model, stating clearly all the assumptions. Derive least square estimates of the coefficients. What is meant by residual sum of squares (R_0^2)? Show that

$$E[R_0^2] = (n-2)\sigma^2. \tag{1}$$

(b) Later, however, she felt that a quadratic model held. Is equation (1) still true ? justify.

(c) Suppose Y_{ij} denote the drying time of the jth metallic plate which was painted with the ith level of concentration, $j = 1, \cdots n_i$, $i = 1, \cdots k$, $\sum_{i=1}^{k} n_i = n$.

Define $SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2$ and $SS_B = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2$, where $\bar{Y}_i$ and $\bar{Y}$ have their usual meaning. Assuming that every Y_{ij} follow follow normal distribution with constant variance σ^2 show the following.

(i) $E(SS_W) = (n-k)\sigma^2$.

(ii) SS_W and SS_B are independent. $[(2 + 6 + 2 + 8) + (1 + 3) + (4 + 6) = 32]$

6. Suppose $U_1, U_2, \cdots U_n$ are i.i.d $N(0, \sigma_u^2)$ variables and $V_1, V_2, \cdots V_n$ are i.i.d. $N(0, \sigma_v^2)$ variables. Consider the following statistics.

$$\begin{aligned} T_1 = \sqrt{n}\bar{U}/\sigma_u, \quad T_2 &= S_{uv}/(\sigma_u \sqrt{S_{vv}}), \\ T_3 = \quad (S_{uu} &- S_{uv}^2/S_{vv})/\sigma_u^2, \\ T_4 = \sqrt{n}\bar{V}/\sigma_v \quad \text{and} \quad T_5 &= S_{vv}/\sigma_v^2, \end{aligned}$$

$$\text{where } S_{uv} = \sum_{i=1}^{n} (U_i - \bar{U})(V_i - \bar{V}) \quad \text{and} \quad S_{qq} = \sum_{i=1}^{n} (Q_i - \bar{Q})^2, Q = U, V.$$

(a) Show that T_1 and T_4 follow standard normal distribution.

(b) If U_i is independent of V_j for each $i, j = 1, \cdots n$, then show that

(i) T_2 follows standard normal distribution, T_3 follows $\chi^2(n-2)$ and T_5 follows $\chi^2(n-2)$.

(c) Suppose not all U_i's are independent of V_j's. Derive the distribution of T_3 ? $[(2 + 2) + (4 + 7 + 5) + 8 = 28]$

7. (a) Suppose $X_1, \cdots X_n$ are i.i.d random variables having probability density $f(x, \theta)$. Consider the problem of testing the hypothesis $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. What are meant by type 1 and type 2 errors. Describe a test procedure such that the probability type 1 error is not more than α.

(b) Formulate the following as statistical testing of hypothesis problems and suggest a test procedure for each of them.

(i) During a flu epidemic, 20% of the population suffers from flu attacks. A physician claims that regular users of vitamin C are less susceptible. She takes a random sample of 500 regular users of vitamin C, finds how many of them had flu to support her claim.

(ii) A publisher claims that in the books he published, the average number of misprints per page is not more than 0.2. A book with 346 pages was chosen and the number of misprints on a random sample of 30 pages of this book were found.

(iii) The following statement is printed on the packs of a certain brand of cigarettes. "The average nicotine content is not more than 0.60 milligrams per cigarette". A government agency decided to verify this and chemically analyzed a random sample of 100 cigarettes of this brand. $[(3 + 4) + 3 \text{ x } 3 = 16]$