BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI,
K K BIRLA GOA CAMPUS
Open Book, Open Laptop, **No Internet**

Subject Name: MATH F432 - Applied Statistical Methods, Date: 04 November 2022
Examiner Name: Sravan Danda Marks: 30
Duration: 1.5 hours

Attempt all questions. Marks corresponding to each question is highlighted in bold within square braces at the end of the question. In case of ambiguities in any of the questions, clearly state your assumptions and attempt the question(s).

1. Suppose $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ are realizations of i.i.d. samples from an unknown probability distribution. Let F denote the cumulative distribution function (CDF) which is unknown. Assume $n \geq 10^8$. For testing $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ (assume that for both the tests, type 1 error is set to $\alpha = 0.05$) where F_0 is given by

$$F_0(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{x}{100} & \text{if } 0 \leq x \leq 100 \\ 1 & \text{if } x > 100 \end{cases} \tag{1}$$

i.e. F_0 is the CDF of $Uniform(0, 100)$. We construct two asymptotic chi-squared tests obtained by bucketing the intervals as follows: for the first test, the buckets are given by $B_1 = (-\infty, 10), B_2 = [10, 20), B_3 = [20, 30), \cdots, B_9 = [80, 90), B_{10} = [90, \infty)$ i.e.

$$T_1 = \sum_{i=1}^{10} \frac{(C_i - n/10)^2}{n/10} \tag{2}$$

(with $C_i = \sum_{j=1}^{n} I_{\{X_j \in B_i\}}$ for each $1 \leq i \leq 10$ i.e. C_i denotes the count of observed values in bucket B_i). The second set of bucketed intervals are $D_1 = (-\infty, 1), D_2 = [1, 2), D_3 = [2, 3), \cdots, D_{99} = [98, 99), D_{100} = [99, \infty)$ i.e. the test statistic is given by

$$T_2 = \sum_{i=1}^{100} \frac{(E_i - n/100)^2}{n/100} \tag{3}$$

(with $E_i = \sum_{j=1}^{n} I_{\{X_j \in D_i\}}$ for each $1 \leq i \leq 100$ i.e. E_i denotes the count of observed values in bucket D_i). You may assume $\chi^2_{9,0.95} \approx 17$ and $\chi^2_{99,0.95} \approx 123$ for all your arguments.

   (a) Denote the rejection regions of test statistics T_1 and T_2 as $R_1 = [t_{critical,1}, \infty)$ and $R_2 = [t_{critical,2}, \infty)$ respectively. Does $t_{critical,2} > t_{critical,1}$ hold true? Justify your answer based on the asymptotic probability distributions of T_1 and T_2 under the null hypothesis.

   (b) Argue that if each C_i for $1 \leq i \leq 10$ is a multiple of 10, it is possible that for a given set of data, the values of the test statistics T_1 and T_2 are identical.

   (c) For a given set of data, using (a) and (b) or otherwise argue (with proper justification) that it is possible that T_1 rejects the null hypothesis while T_2 does not reject the null hypothesis.

[(1+1)+3+4]

2. The following data represent the successive quality levels (the higher the better) of 11 items: $81, 123, 134, 142, 311, 362, 375, 381, 405, 413, 422$. Assume that the data from a continuous distribution hence the probability of all values being identical is zero. To check if these data are a random

sample from some unknown population, a statistician decides to construct a sequence of zeros and ones (exactly as done in class) as follows: Let $X_1, \cdots, X_{11}$ be the samples, construct

$$I_j = \begin{cases} 1 & \text{if } X_j \leq \textit{sample median} \\ 0 & \text{if } X_j > \textit{sample median} \end{cases} \tag{4}$$

and apply the runs test.

(a) Let N_0 and N_1 denote the number of runs of 0 and 1 respectively. Find the number of runs $N_0 + N_1$ based on the data.

(b) Let r denote the number obtained by data in (a). Find $P_{H_0}\{N_0 + N_1 = r\}$ i.e. under the null hypothesis, the probability that the number of runs equal r. You need not simplify the numbers in the expression.

(c) Is $min\{P_{H_0}\{N_0 + N_1 \leq r\}, P_{H_0}\{N_0 + N_1 \geq r\}\} < 0.025$? Justify.

(d) Using (c) or otherwise determine whether or not the test rejects the null hypothesis (when type 1 error is restricted to $\alpha \leq 0.05$).

[1+2+2+2]

3. Suppose we wish to simulate a random variable X from the discrete distribution whose probability mass function (p.m.f.) is given by

$$P\{X = i\} = \begin{cases} \dfrac{1}{6n} & \text{if } 1 \leq i \leq n \\ \dfrac{1}{3n} & \text{if } n+1 \leq i \leq 2n \\ \dfrac{1}{2n} & \text{if } 2n+1 \leq i \leq 3n \end{cases} \tag{5}$$

Design a constant-time algorithm (w.r.t. n) to simulate one realization of a random variable with the p.m.f. given by Eq 5. You will be awarded full marks only if your algorithm has a deterministic constant-time algorithm. If your algorithm has expected constant-time but does not have a worst-case constant-time, only partial marks will be awarded.

[7]

4. Suppose a random variable with the cumulative distribution function (CDF) F can be generated in constant time i.e. $\mathcal{O}(1)$.

(a) One can easily verify that the function given by Eq 6 is right continuous, non-decreasing, has the limits 0 and 1 at $-\infty$ and ∞ respectively. Hence, G is a CDF. Construct a random variable whose CDF is given by Eq 6.

$$G(x) = (F(x))^{10} + 10(1 - F(x))(F(x))^9 \tag{6}$$

(b) Using (a) or otherwise, devise a constant-time algorithm to simulate a random variable whose CDF is given by Eq 6. Justify your answer.

[3+4]