Statistics (II) Midterm Examination

Version A Reference Solution

(I) Multiple Choice (7 pts for Q1 & Q2 ; 8 pts each for Q3 & Q4)

1. (C) Its exact shape depends strictly on the degrees of freedom, becoming less skewed and more symmetric as the degrees of freedom increase.

2. (B) It ensures that the calculated test statistic is always greater than or equal to 1, restricting the critical region entirely to the upper tail of the F distribution.

3. (C) χ^2_B will be 100 times larger than χ^2_A, making the test on Dataset B much more likely to reject the null hypothesis, regardless of the practical significance of the association.

4. (B) Homogeneity involves drawing independent random samples from predetermined sub-populations and comparing the distribution of a single categorical response across them.

(II) Fill in the Blanks (10 pts each)

5. multiple comparison

6. expected frequencies

(III) Short Answer (10 pts each)

7. **Difference between independent and matched sample designs**

   An independent sample design draws two separate, unrelated random samples from two populations. The comparison between the two population means is based on the difference between the two sample means.

   In a matched sample design, one simple random sample is selected from a population, and each observational unit provides a pair of data values, one for each method. The analysis is then based on the sample of differences.

   Primary statistical advantage:

   In a matched sample design, the primary statistical advantage is that it often leads to a smaller sampling error. The primary reason is that variation between subjects is eliminated because the same subjects are used for both methods.

8. **Reason why the confidence interval for a population variance is not symmetric**

   A confidence interval for a population mean is symmetric because it is based on the z distribution or the t distribution, both of which are symmetric distributions.

   However, a confidence interval for a population variance is based on the chi-square distribution:

   $$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

   and the chi-square distribution is not symmetric; it is right-skewed.

   Therefore, the resulting confidence interval

   $$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

   is not centered symmetrically around the sample variance s^2.

(IV) Computation (30 pts)

9. **Given:** $\bar{x} = 22.8,\ s = 6.27,\ n = 30,\ \alpha = 0.05$

   Use six classes, each with probability $0.1667 \approx \dfrac{1}{6}$

   $P(Z \le -0.97) = 0.1667,\ P(Z \le -0.43) = 0.3334,\ Z \sim N(0,1)$

   Hypotheses

   $$H_0 : \text{The data come from a normal distribution}$$

   $$H_a : \text{The data do not come from a normal distribution}$$

   Class Boundaries

   Using

   $$x_i = \bar{x} + z_{i/6}s, \quad i = 1, 2, \ldots, 5,$$

   where $z_{i/6}$ denotes the $(i/6)$th quantile of the standard normal distribution, we obtain:

   $$x_1 = 22.8 + (-0.97)(6.27) = 16.7181$$
   $$x_2 = 22.8 + (-0.43)(6.27) = 20.1039$$
   $$x_3 = \bar{x} = 22.8$$
   $$x_4 = 22.8 + (0.43)(6.27) = 25.4961$$
   $$x_5 = 22.8 + (0.97)(6.27) = 28.8819$$

   Hence, the six classes are:

   $$(-\infty, x_1],\ (x_1, x_2],\ (x_2, x_3],\ (x_3, x_4],\ (x_4, x_5],\ (x_5, \infty)$$

   Observed Frequencies

   First, sort the data in ascending order:

   $$11,\ 13,\ 15,\ 17,\ 18,\ 18,\ 18,\ 19,\ 20,\ 20,\ 21,\ 21,\ 21,\ 22,\ 22,$$
   $$23,\ 23,\ 23,\ 23,\ 23,\ 24,\ 24,\ 26,\ 27,\ 28,\ 29,\ 29,\ 30,\ 33,\ 43$$

   Counting the number of observations in each class, we obtain:

   $$f_1 = 3, \quad f_2 = 7, \quad f_3 = 5, \quad f_4 = 7, \quad f_5 = 3, \quad f_6 = 5$$

   Expected Frequencies

   $$e_1 = e_2 = e_3 = e_4 = e_5 = e_6 = 30 \times \frac{1}{6} = 5$$

   Chi-Square Test Statistic

   $$\chi_0^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

   $$= \frac{(3-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(5-5)^2}{5}$$

   $$= \frac{4}{5} + \frac{4}{5} + 0 + \frac{4}{5} + \frac{4}{5} + 0 = 3.2000$$

   Degrees of Freedom

   $$df = k - 1 - 2 = 6 - 1 - 2 = 3$$

Critical Value Approach

Given $\chi^2_{0.05,3} = 7.815$,

reject H_0 if

$$\chi^2_0 \geq 7.815$$

Since

$$3.2000 < 7.815,$$

we do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the data do not come from a normal distribution.

(V) Bonus Question (+20 pts)

10. **Difference in sampling design between homogeneity and independence tests**

    The fundamental difference in the sampling design is as follows:

    In a chi-square test for homogeneity, independent samples are drawn from different populations, and the objective is to compare whether the population proportions of the categorical variable are the same across these populations.

    In a chi-square test of independence, a single sample is drawn from one population, and each observational unit is classified according to two categorical variables to determine whether the variables are independent.