CIE A-Level Further Mathematics (9231/42)

Probability & Statistics (Complete Solutions)

Prepared by Capybara

Question 1

The manager of a village store is interested in the shopping habits of his customers. From previous records, the median number of items bought per customer was found to be 8. The manager believes the median number of items bought per customer has increased. He selects a random sample of 15 customers and records the number of items each customer buys at his store. The results are shown in the stem-and-leaf diagram.

key 2	3 represents 23 items
0	6 6 7 9 9
1	0 5 7 7
2	3 6
3	4
4	0
5	6
6	
7	4

(a) Suggest a reason why using a Wilcoxon signed-rank test to examine the manager's belief might not be appropriate in this situation. [1]

(b) Use a sign test at the 5% significance level to investigate the manager's belief. [6]

Solution & Analysis

(a) Assumption Check:
The Wilcoxon signed-rank test requires the underlying population distribution to be symmetrical. Observing the stem-and-leaf diagram, the data is distinctly right-skewed (positively skewed) with a long tail extending towards the 70s. Therefore, the assumption of symmetry is severely violated, making the Wilcoxon test inappropriate.
(b) Sign Test Application:
Step 1: Hypotheses formulation
Let M denote the population median number of items purchased.
$H_0 : M = 8$
$H_1 : M > 8$ (One-tailed test)
Step 2: Define test statistic
Let X represent the number of customers who purchased more than 8 items. Under the null hypothesis, X follows a binomial distribution: $X \sim B(n, 0.5)$.
From the data, values greater than 8 are: 9, 9, 10, 15, 17, 17, 23, 26, 34, 40, 56, 74.
Number of values > 8: $x = 12$.
Number of values < 8: $y = 3$.
Total observations $n = 15$. So, $X \sim B(15, 0.5)$.
Step 3: Calculate p-value
We want the probability of getting a result as extreme or more extreme than observed:

$$\begin{aligned} P(X \geq 12) &= P(X = 12) + P(X = 13) + P(X = 14) + P(X = 15) \\ &= \binom{15}{12}(0.5)^{15} + \binom{15}{13}(0.5)^{15} + \binom{15}{14}(0.5)^{15} + \binom{15}{15}(0.5)^{15} \\ &= 0.5^{15}(455 + 105 + 15 + 1) \\ &= 576 \times (0.5)^{15} \approx 0.017578 \end{aligned}$$

Step 4: Conclusion
Comparing the p-value to the significance level $\alpha = 0.05$:
$0.0176 < 0.05$.
Hence, we reject the null hypothesis H_0. There is sufficient statistical evidence at the 5% level to support the manager's belief that the median number of items bought per customer has increased.

Question 2

Three different bus companies, A, B and C, record the reliability of their service. Each bus journey is recorded as being on time, late or cancelled. A reliable bus service is considered to be one that is on time. The records from a random sample of buses for each company are taken. These records are summarised in the table.

	Company A	Company B	Company C	Total
On time	7	22	20	49
Late	14	11	12	37
Cancelled	4	6	4	14
Total	25	39	36	100

(a) Test at the 5% significance level whether the reliability of the service is independent of the bus company. [8]

(b) Further evidence indicates that one company may be particularly unreliable compared with the other two companies. Use your calculations in 2(a) to suggest which company this is likely to be. [2]

Solution & Analysis

(a) Chi-Square Test for Independence:
Step 1: State hypotheses
H_0: The reliability of the service is independent of the bus company.
H_1: The reliability of the service is not independent of the bus company.
Step 2: Evaluate expected frequencies & adjustment
The expected frequency E is given by $\frac{\text{Row Total}\times\text{Col Total}}{\text{Grand Total}}$.
For (Company A, Cancelled), $E = \frac{14\times 25}{100} = 3.5$.
Since an expected frequency is less than 5, the standard χ^2 approximation is invalid. We must combine the 'Late' and 'Cancelled' rows into a single 'Not on time' row.
Merged Observed Frequencies (O):

	Company A	Company B	Company C	Total
On time	7	22	20	49
Not on time	18	17	16	51

Expected Frequencies (E):

	Company A	Company B	Company C
On time	12.25	19.11	17.64
Not on time	12.75	19.89	18.36

Step 3: Calculate the test statistic
$\chi^2 = \sum \frac{(O-E)^2}{E}$
Company A contributions: $\frac{(7-12.25)^2}{12.25} + \frac{(18-12.75)^2}{12.75} = 2.2500 + 2.1618 = 4.4118$
Company B contributions: $\frac{(22-19.11)^2}{19.11} + \frac{(17-19.89)^2}{19.89} = 0.4369 + 0.4199 = 0.8568$
Company C contributions: $\frac{(20-17.64)^2}{17.64} + \frac{(16-18.36)^2}{18.36} = 0.3157 + 0.3033 = 0.6190$
Total $\chi^2 = 4.4118 + 0.8568 + 0.6190 \approx 5.888$.
Step 4: Critical value and conclusion
Degrees of freedom: $df = (\text{rows} - 1)(\text{cols} - 1) = (2-1)(3-1) = 2$.
Critical value at 5% significance level, $\chi^2_{\text{crit}}(2) = 5.991$.
Since $5.888 < 5.991$, we do not reject H_0.
There is insufficient evidence at the 5% level to conclude that the reliability of the service depends on the bus company.

(b) Identifying the Unreliable Company:
Based on the χ^2 breakdown, Company A is the most likely candidate. It is responsible for the majority of the χ^2 test statistic (4.41 out of 5.89). Observing the raw data, Company A recorded 18 'not on time' journeys, which is substantially higher than its theoretically expected value of 12.75.

Question 3

Students studying geography at a college choose either human geography or physical geography. A teacher is investigating the times taken by students to complete assignments of equal difficulty. He selects a random sample of 7 human geography students (times x) and a random sample of 9 physical geography students (times y). The results are summarised as follows:

$$\Sigma x = 445, \quad \Sigma x^2 = 29144$$

$$\Sigma y = 310, \quad \Sigma y^2 = 11622$$

You should assume that the underlying distributions are normal and have equal population variances. Find a 95% confidence interval for the difference in population mean times taken to complete assignments in human geography and physical geography. [6]

Solution & Analysis

Confidence Interval Calculation:

Step 1: Compute sample means

$\bar{x} = \frac{445}{7} \approx 63.571$

$\bar{y} = \frac{310}{9} \approx 34.444$

Step 2: Compute unbiased variance estimates

$s_x^2 = \frac{1}{n_x-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n_x}\right) = \frac{1}{6}\left(29144 - \frac{445^2}{7}\right) \approx 142.452$

$s_y^2 = \frac{1}{n_y-1}\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n_y}\right) = \frac{1}{8}\left(11622 - \frac{310^2}{9}\right) \approx 118.028$

Step 3: Compute the pooled variance

Assuming equal population variances, we calculate s_p^2:

$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2} = \frac{6(142.452)+8(118.028)}{14} = \frac{854.712+944.224}{14} \approx 128.495$

Step 4: Determine the t-critical value

Degrees of freedom $df = 7 + 9 - 2 = 14$.

For a 95% confidence interval, the two-tailed critical t-value is $t_{14,0.025} = 2.145$.

Step 5: Construct the Confidence Interval

$$CI = (\bar{x} - \bar{y}) \pm t\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

$$CI = (63.571 - 34.444) \pm 2.145\sqrt{128.495\left(\frac{1}{7} + \frac{1}{9}\right)}$$

$$CI = 29.127 \pm 2.145\sqrt{32.634}$$

$$CI = 29.127 \pm 2.145(5.713) = 29.127 \pm 12.254$$

$$\text{Lower Bound} = 16.873, \quad \text{Upper Bound} = 41.381$$

The 95% confidence interval is $[16.9, 41.4]$ (correct to 3 s.f.).

Question 4

The continuous random variable X has cumulative distribution function F given by

$$F(x) = \begin{cases} 0 & x < 0, \\ -\frac{1}{108}x^3 + \frac{1}{12}x^2 & 0 \leq x \leq 3, \\ -\frac{1}{32}x^2 + \frac{7}{16}x + a & 3 < x \leq b, \\ 1 & x > b. \end{cases}$$

(a) Verify that the median of X is 3. [1]
(b) Show that $a = -\frac{17}{32}$ and find the value of b. [3]
(c) Find the probability density function of X. [2]
(d) Find $E\left(\frac{1}{X}\right)$. [3]

Solution & Analysis

(a) Verifying the Median:
The median m is the value such that $F(m) = 0.5$. Let's test $x = 3$ using the first non-zero segment:

$$F(3) = -\frac{1}{108}(3)^3 + \frac{1}{12}(3)^2 = -\frac{27}{108} + \frac{9}{12} = -\frac{1}{4} + \frac{3}{4} = \frac{1}{2} = 0.5$$

Since $F(3) = 0.5$, the median is indeed 3.
(b) Finding Parameters a and b:
Due to the continuity of the CDF at $x = 3$, the second segment must also equal 0.5 when $x = 3$:

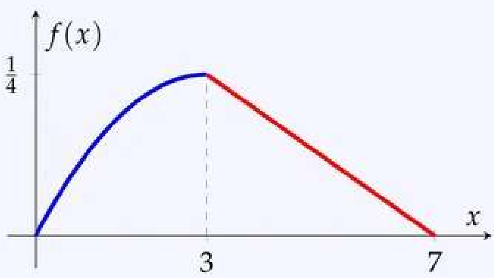
$$-\frac{1}{32}(3)^2 + \frac{7}{16}(3) + a = 0.5 \implies -\frac{9}{32} + \frac{42}{32} + a = \frac{16}{32} \implies \frac{33}{32} + a = \frac{16}{32} \implies a = -\frac{17}{32} \text{ (Shown)}$$

For b, we use the property that $F(b) = 1$:

$$-\frac{1}{32}b^2 + \frac{7}{16}b - \frac{17}{32} = 1 \implies -b^2 + 14b - 17 = 32 \implies b^2 - 14b + 49 = 0 \implies (b-7)^2 = 0 \implies b = 7$$

(c) Probability Density Function $f(x)$:
We find $f(x)$ by differentiating $F(x)$ with respect to x: For $0 \le x \le 3$: $f(x) = \frac{d}{dx}\left(-\frac{1}{108}x^3 + \frac{1}{12}x^2\right) = -\frac{1}{36}x^2 + \frac{1}{6}x$
For $3 < x \le 7$: $f(x) = \frac{d}{dx}\left(-\frac{1}{32}x^2 + \frac{7}{16}x - \frac{17}{32}\right) = -\frac{1}{16}x + \frac{7}{16}$

$$f(x) = \begin{cases} -\frac{1}{36}x^2 + \frac{1}{6}x & 0 \le x \le 3 \\ -\frac{1}{16}x + \frac{7}{16} & 3 < x \le 7 \\ 0 & \text{otherwise} \end{cases}$$



(d) Finding the Expected Value $E(1/X)$:
Using LOTUS, $E(g(X)) = \int g(x)f(x)dx$:

$$\begin{aligned}
E\left(\frac{1}{X}\right) &= \int_0^3 \frac{1}{x}\left(-\frac{1}{36}x^2 + \frac{1}{6}x\right)dx + \int_3^7 \frac{1}{x}\left(-\frac{1}{16}x + \frac{7}{16}\right)dx \\
&= \int_0^3 \left(-\frac{1}{36}x + \frac{1}{6}\right)dx + \int_3^7 \left(-\frac{1}{16} + \frac{7}{16}x^{-1}\right)dx \\
&= \left[-\frac{1}{72}x^2 + \frac{1}{6}x\right]_0^3 + \left[-\frac{1}{16}x + \frac{7}{16}\ln(x)\right]_3^7 \\
&= \left(-\frac{9}{72} + \frac{3}{6} - 0\right) + \left[\left(-\frac{7}{16} + \frac{7}{16}\ln 7\right) - \left(-\frac{3}{16} + \frac{7}{16}\ln 3\right)\right] \\
&= \left(-\frac{1}{8} + \frac{1}{2}\right) - \frac{4}{16} + \frac{7}{16}\ln\left(\frac{7}{3}\right) \\
&= \frac{3}{8} - \frac{1}{4} + \frac{7}{16}\ln\left(\frac{7}{3}\right) = \frac{1}{87} + \frac{7}{16}\ln\left(\frac{7}{3}\right) \approx 0.496
\end{aligned}$$

Question 5

Margaret claims that her new revision course will increase a student's score by more than 5 marks on average. For a random sample of 8 students, the test scores before and after the revision course are given below.

Student	A	B	C	D	E	F	G	H
Test score before	43	29	67	54	34	29	52	60
Test score after	49	53	79	62	39	55	65	62

(a) Use a t-test at the 10% significance level to investigate Margaret's claim. [7]
(b) State an assumption that is necessary for the test in 5(a) to be valid. [1]

Solution & Analysis

(a) Paired Sample t-test:
Step 1: Compute the differences
Let d = Score After − Score Before.
The set of differences d is: $6, 24, 12, 8, 5, 26, 13, 2$.
Step 2: Hypotheses formulation
Let μ_d be the true population mean difference.
$H_0 : \mu_d = 5$
$H_1 : \mu_d > 5$
Step 3: Descriptive Statistics
$n = 8$
$\Sigma d = 96 \implies \bar{d} = 12$
$\Sigma d^2 = 36 + 576 + 144 + 64 + 25 + 676 + 169 + 4 = 1694$
Sample variance: $s_d^2 = \frac{1}{n-1}\left(\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right) = \frac{1}{7}\left(1694 - \frac{9216}{8}\right) = \frac{542}{7} \approx 77.428$
Standard deviation: $s_d = \sqrt{77.428} \approx 8.80$
Step 4: Compute test statistic
$t = \frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} = \frac{12-5}{8.80/\sqrt{8}} = \frac{7}{3.111} \approx 2.25$
Step 5: Conclusion
For a 10% significance level and a one-tailed test with degrees of freedom $df = 8 - 1 = 7$, the critical t-value is $t_{7,0.10} = 1.415$.
Since $2.25 > 1.415$, we fall into the rejection region.
We reject H_0. There is sufficient evidence at the 10% level to support Margaret's claim that the revision course increases a student's score by more than 5 marks on average.
(b) Required Assumption:
To legitimately carry out this paired t-test, we must assume that the population of differences between the paired scores follows a normal distribution.

Question 6

The discrete random variable X has probability generating function $G_X(t)$ defined by

$$G_X(t) = \frac{k}{t}(at+1)^2$$

It is given that $E(X) = \frac{1}{3}$.

(a) Show that $a = 2$ and find the value of k. [5]
(b) Find $Var(X)$. [2]
(c) Three independent observations of X are taken. The random variable Y is the sum of these observations. Find $P(Y = 0)$. [3]

Solution & Analysis

(a) Parameter Derivation:
For any valid PGF, the probabilities must sum to 1, thus $G_X(1) = 1$.
$G_X(1) = k(a+1)^2 = 1 \quad \ldots(\text{Eq } 1)$
Furthermore, the expected value is derived from the first derivative evaluated at $t = 1$:
$G'_X(1) = E(X) = 1/3$.
Rewrite $G_X(t) = k(a^2 t + 2a + t^{-1})$.
$G'_X(t) = k(a^2 - t^{-2})$.
$G'_X(1) = k(a^2 - 1) = \frac{1}{3} \quad \ldots(\text{Eq } 2)$
Dividing (Eq 2) by (Eq 1):

$$\frac{k(a^2-1)}{k(a+1)^2} = \frac{1/3}{1} \implies \frac{(a-1)(a+1)}{(a+1)^2} = \frac{1}{3} \implies \frac{a-1}{a+1} = \frac{1}{3}$$

$$3(a-1) = a+1 \implies 3a - 3 = a + 1 \implies 2a = 4 \implies a = 2 \text{ (Shown)}$$

Substitute $a = 2$ back into (Eq 1): $k(2+1)^2 = 1 \implies 9k = 1 \implies k = \frac{1}{9}$.

(b) Finding Variance:
The variance using PGFs is given by $Var(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2$.
We already know $G'_X(1) = 1/3$.
Now evaluate $G''_X(t)$ using $k = 1/9$:
$G'_X(t) = \frac{1}{9}(4 - t^{-2})$
$G''_X(t) = \frac{1}{9}(2t^{-3}) = \frac{2}{9t^3}$.
At $t = 1$, $G''_X(1) = \frac{2}{9}$.
Substitute into the variance formula:

$$Var(X) = \frac{2}{9} + \frac{1}{3} - \left(\frac{1}{3}\right)^2 = \frac{2}{9} + \frac{3}{9} - \frac{1}{9} = \frac{4}{9}$$

(c) Evaluating the Sum of Independent Variables:
Let $Y = X_1 + X_2 + X_3$. For sums of independent variables, the PGF is the product of their individual PGFs.

$$G_Y(t) = [G_X(t)]^3 = \left[\frac{1}{9t}(2t+1)^2\right]^3 = \frac{1}{729t^3}(2t+1)^6$$

To find $P(Y = 0)$, we need the coefficient of t^0 in the expansion of $G_Y(t)$. Because the denominator contains t^3, this is equivalent to finding the coefficient of t^3 in the binomial expansion of the numerator $(2t+1)^6$.
Using the Binomial Theorem for $(2t+1)^6$, the term containing t^3 is:

$$\binom{6}{3}(2t)^3(1)^3 = 20 \times 8t^3 = 160t^3$$

Thus, replacing this in $G_Y(t)$, the constant term is $\frac{160t^3}{729t^3} = \frac{160}{729}$.

$$P(Y = 0) = \frac{160}{729}$$