MAST90105 Methods of Mathematical Statistics
Assignment 4, Semester 1 2019 **Solutions**

Note: You may use R and/or Mathematica for any questions but must include your commands and reasoning.

Problems:

1. Carbon monoxide concentrations (in $\mu g/m^3$) are measured on 6 different days in city A and on 4 different days in city B. The following measurements have been obtained in cities A and B, respectively: 3.9, 4.7, 7.1, 6.9, 4.3, 6.3 and 7.3, 6.9, 7.6, 9.1. We assume that these two samples are independent.

   (a) Assume that both samples are from normally distributed populations. We also assume that the same type of measurement device was used in both cities so that the measurement error (variance) is the same for both samples. Construct the 95% confidence interval for $\delta = \mu_A - \mu_B$ where μ_A and μ_B are mean concentrations of carbon monoxide in cities A and B, respectively.

   - *Let* $X_1, \ldots, X_5$ *and* $Y_1, \ldots, Y_4$ *are the two samples. We find* $\bar{X} = 5.533, s_X^2 = 1.959$ *and* $\bar{Y} = 7.725, s_Y^2 = 0.9225$.
   - *An estimate of* δ *is* $\hat{\delta} = \bar{X} - \bar{Y} = -2.192$.
   - *The pooled variance estimate is* $s_p^2 = (5s_X^2 + 3s_Y^2)/8 = 1.57$.
   - *The 95% confidence interval is*

   $$\left(\hat{\delta} - t_{0.975}(8)s_p\sqrt{\frac{1}{6}+\frac{1}{4}}, \hat{\delta} + t_{0.975}(8)s_p\sqrt{\frac{1}{6}+\frac{1}{4}}\right) = (-4.057, -0.326).$$

   [2]

   (b) Do we reject $H_0 : \delta = 0$ in favor of $H_1 : \delta \neq 0$ at the 5% significance level? Why or why not? Find the p-value of this test.

   - *The 95% confidence interval for* δ *does not contain zero and therefore* H_0 *is rejected at the 5% significance level.* [1]
   -

   $$\begin{aligned} P &= \Pr\left\{|T| > \frac{|\hat{\delta}|}{s_p/\sqrt{1/6+1/4}} = 2.71 \,\middle|\, H_0 : T \sim t(6+4-2)\right\} \\ &= 2\Pr(T > 2.71 | T \sim t(8)) = 0.027. \end{aligned}$$

   [1]

   (c) Now assume that both samples are from normally distributed populations, as before, but two different measurement devices were used and their measurements errors are known. We assume that the variances are 2 and 1, respectively, for the two samples. Construct the 95% confidence interval for δ.

- *An estimate of variance is* $\hat{\sigma}^2 = \frac{\sigma_X^2}{6} + \frac{\sigma_Y^2}{4} = \frac{2}{6} + \frac{1}{4} = \frac{7}{12}$.
- *The 95% confidence interval is*

$$\left(\hat{\delta} - q_{0.975}\hat{\sigma}, \hat{\delta} + q_{0.975}\hat{\sigma}\right) = (-3.689, -0.695).$$

[2]

(d) Compare confidence intervals you obtained in (a) and (c). Which one is narrower? Briefly explain why.

- *The interval in (c) is narrower. The variances of the two samples are known so we don't have to estimate them. This results in less variability of* $\hat{\delta}$ *and hence in a narrower confidence interval.* [1]

(e) Show that $(3.9, 7.1)$ is an approximate 97% confidence interval for the median concentration of carbon monoxide in city A.

- *Let* $Y_1 < \cdots < Y_6$ *be the order statistics for measured concentrations of carbon monoxide in city A. Let* $W \sim Binom(n = 6, p = 0.5)$ *and* m *be the median concentration. Then*

$$\begin{aligned} \Pr(3.9 < m < 7.1) &= \Pr(0 < W < 6) = 1 - \Pr(W = 0) - \Pr(W = 6) \\ &= 1 - 0.5^6 - 0.5^6 \approx 0.97. \end{aligned}$$

[2]

`Total marks = 9`

2. Let $X_1, X_2, \ldots, X_{10}$ be a sample of size 10 from an exponential distribution with the density function

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We reject $H_0 : \lambda = 1$ in favor of $H_1 : \lambda = 2$ if the observed value of $Y = \sum_{i=1}^{10} X_i$ is smaller than 6.

(a) Find the probability of type 1 error for this test.

- *We find* $\alpha = \Pr(Y < 6 | H_0 \textit{ is true}) = \Pr(Y < 6 | Y \sim Gamma(10, 1)) = 0.084$. [2]

(b) Find the probability of type 2 error for this test.

- *We find* $\beta = \Pr(Y \geq 6 | H_1 \textit{ is true}) = \Pr(Y \geq 6 | Y \sim Gamma(10, 2)) = 0.242$. [2]

(c) Let $y = 5$ be the observed value of Y. Find the p-value for this test.

- *We find* $P = \Pr(Y < 5 | H_0 \textit{ is true}) = \Pr(Y < 5 | Y \sim Gamma(10, 1)) = 0.032$. [1]

(d) Let $y = 5$ be the observed value of Y. Construct the exact 95% confidence interval for λ. *Hint:* $\lambda Y \sim Gamma(10, 1)$.

- *Let $\Gamma_{0.025}, \Gamma_{0.975}$ be the 2.5% and 97.5% quantiles of the gamma distribution with shape and rate parameters 10 and 1, respectively. We find*

$$\Pr(\Gamma_{0.025} < \lambda Y < \Gamma_{0.975}) = \Pr(\Gamma_{0.025}/Y < \lambda < \Gamma_{0.975}/Y) = 0.95.$$

- *The 95% confidence interval is therefore*

$$(\Gamma_{0.025}/y, \Gamma_{0.975}/y) = (0.96, 3.42).$$

[1]

`Total marks = 6`

3. The number of faults over a period of time was collected for a sample of 100 data-transmission lines. We want to test if the data come from a Poisson distribution.

number of faults	0	1	2	3	4	5	> 5
number of lines	38	30	16	9	5	2	0

(a) Assuming the number of faults for a data-transmission line, Y_i, $i = 1, \ldots, 100$, follows a Poisson distribution with parameter λ, find the maximum likelihood estimate of λ for these data.

- *The likelihood function is* $L(\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}$
- $\ell(\lambda) = \ln L(\lambda) = -n\lambda + (\sum_{i=1}^n y_i) \ln \lambda - \sum_{i=1}^n \ln y_i!$,
- $\partial \ell(\lambda)/\partial\lambda = -n + \frac{1}{\lambda}(\sum_{i=1}^n y_i) = 0 \quad \Rightarrow \quad \hat{\lambda} = \bar{y}$,
- $\partial^2 \ell(\lambda)/\partial\lambda^2 = -\frac{1}{\lambda^2}(\sum_{i=1}^n y_i) < 0$ *and hence $\hat{\lambda}$ is the maximum likelihood estimate of λ.*
- *For the observed data,* $\hat{\lambda} = \frac{1}{100}(0 \cdot 38 + 1 \cdot 30 + 2 \cdot 16 + 3 \cdot 9 + 4 \cdot 5 + 5 \cdot 2) = 1.19$ [2]

(b) Test the hypothesis that the number of faults for a data-transmission line follows a Poisson distribution using the chi-squared test. Do we reject the null hypothesis at the 5% significance level? *Hint:* You should combine the observed data into several groups such that expected frequencies are greater or equal to 5 for each group.

- *Let* $p_k = e^{-\hat{\lambda}} \frac{\hat{\lambda}^k}{k!}$ *and* $E_k = 100 p_k$. *Notice that* $E_k < 5$ *for* $k \geq 4$ *so we combine the observed data for* $y_i \geq 3$ *into a single, fourth, category, with 16 lines in total.*
- *We find:* $E_0 = 30.4, E_1 = 36.2, E_2 = 21.5, E_3 = 11.8$
- *We compute*

$$\chi^2 = \frac{(38 - 30.4)^2}{30.4} + \frac{(30 - 36.2)^2}{36.2} + \frac{(16 - 21.5)^2}{21.5} + \frac{(16 - 11.8)^2}{11.8} = 5.864$$

- $\chi^2 < \chi^2_{0.95}(2) = 5.99$ *so we do not reject H_0 at the 5% significance level* [3]

`Total marks = 5`