F20SA/F21SA: Statistical Modelling and Analysis

Exam Solutions 2021-22

1. a) (similar to tutorial exercises)
We calculate

$$P(\text{faulty}) = \frac{1}{10} \cdot \frac{2}{100} + \frac{3}{10} \cdot \frac{5}{1000} + \frac{6}{10} \cdot \frac{1}{100} = \frac{95}{10000}.$$

[2 marks]

b) (similar to tutorial exercises)
We have

$$P(\text{not faulty}) = 1 - P(\text{faulty}) = \frac{9905}{10000}.$$

Moreover,

$$P(A|\text{not faulty}) = \frac{P(A \cap \text{not faulty})}{P(\text{not faulty})} = \frac{P(\text{not faulty}|A) \cdot P(A)}{P(\text{not faulty})}$$

hence

$$P(A|\text{not faulty}) = \frac{\frac{98}{100} \cdot \frac{1}{10}}{\frac{9905}{10000}} = \frac{980}{9905} \approx 0.0989.$$

[3 marks]

c) (non-standard)
To solve the last part, we will calculate the expected total cost of the order, separately for all three companies. If there are no faulty pipes in the order, the cost is equal to the total price of the pipes. If there is even one faulty pipe, the total cost has to be increased by \$ 100000. Hence we have

$$E(\text{cost of order from A}) = 8000 \cdot \left(\frac{98}{100}\right)^{100} + 108000 \cdot \left(1 - \left(\frac{98}{100}\right)^{100}\right) = 94738.04$$

Similarly,

$$E(\text{cost of order from B}) = 15000 \cdot \left(\frac{995}{1000}\right)^{100} + 115000 \cdot \left(1 - \left(\frac{995}{1000}\right)^{100}\right) = 54422.96$$

and

$$E(\text{cost of order from C}) = 12000 \cdot \left(\frac{99}{100}\right)^{100} + 112000 \cdot \left(1 - \left(\frac{99}{100}\right)^{100}\right) = 75396.77$$

The total expected cost is minimized by choosing to order from B. [4 marks]

2. a) (similar to tutorial exercises)
We have

$$E(X) = 3\theta + 4\theta + 10\theta + 3 - 6\theta + \frac{7}{2} - 21\theta = \frac{13}{2} - 10\theta\,.$$

Now we can compute the sample mean

$$3 \cdot \frac{4}{40} + 4 \cdot \frac{3}{40} + 5 \cdot \frac{9}{40} + 6 \cdot \frac{15}{40} + 9 \cdot \frac{7}{40} = \frac{222}{40}\,.$$

The method of moments estimator $\hat{\theta}$ is obtained by solving

$$\frac{13}{2} - 10\hat{\theta} = \frac{222}{40}$$

and is equal

$$\hat{\theta} = \frac{19}{200} = 0.095\,.$$

[2 marks]

b) (similar to tutorial exercises, but a bit more difficult than the previous question)
The likelihood function is given by

$$L(\theta) = \theta^{n_1}\theta^{n_2}(2\theta)^{n_3}\left(\frac{1}{2} - \theta\right)^{n_4}\left(\frac{1}{2} - 3\theta\right)^{n_5}$$

and the log-likelihood by

$$l(\theta) = (n_1 + n_2)\log(\theta) + n_3\log(2\theta) + n_4\log(\frac{1}{2} - \theta) + n_5\log(\frac{1}{2} - 3\theta)\,.$$

Computing the derivative

$$l'(\theta) = \frac{n_1 + n_2}{\theta} + \frac{n_3}{\theta} - \frac{n_4}{\frac{1}{2} - \theta} - \frac{3n_5}{\frac{1}{2} - 3\theta}$$

we can now obtain the maximum likelihood estimator $\bar{\theta}$ by solving

$$l'(\bar{\theta}) = 0\,,$$

that is

$$(n_1 + n_2 + n_3)(\frac{1}{2} - \bar{\theta})(\frac{1}{2} - 3\bar{\theta}) - n_4\bar{\theta}(\frac{1}{2} - 3\bar{\theta}) - 3n_5\bar{\theta}(\frac{1}{2} - \bar{\theta}) = 0\,,$$

which is equivalent to

$$120\bar{\theta}^2 - 53\bar{\theta} + 4 = 0\,.$$

This equation has two roots

$$\bar{\theta} = \frac{53 - \sqrt{889}}{240} \qquad \text{and} \qquad \bar{\theta} = \frac{53 + \sqrt{889}}{240}\,.$$

However, note that for the second root, the expression $\frac{1}{2} - 3\bar{\theta}$ takes a negative value and hence the probability distribution given in the problem would not make sense. Hence we choose $\bar{\theta} = \frac{53-\sqrt{889}}{240} \approx 0.0966$

[4 marks]

3. a) (similar to tutorial exercises)
The confidence interval is given as

$$\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a} \right) ,$$

where S is the sample standard deviation and a, b are appropriate percentage points of the χ^2_{n-1} distribution. For $n = 15$, from the tables we get

$$a = 5.629 \qquad \text{and} \qquad b = 26.12 \, .$$

Moreover, we have

$$S^2 = \frac{1}{14} \left(12086.41 - \frac{(424.22)^2}{15} \right) \approx 6.3502$$

and hence the desired confidence interval is $(3.4036, 15.7937)$. As a consequence, the belief that the population variance equals 5 seems justified. [3 marks]

b) (similar to tutorial exercises, but a bit more difficult than the previous question)
We now have $n = 13$ and the modified data

$$\sum_{i=1}^{13} x_i = 424.22 - 33.97 - 34.04 = 356.21 \qquad \text{and}$$

$$\sum_{i=1}^{13} x_i^2 = 12086.41 - 33.97^2 - 34.04^2 = 9773.7275 \, .$$

Hence for this new data

$$S^2 = \frac{1}{12} \left(9773.7275 - \frac{(356.21)^2}{13} \right) \approx 1.1083$$

and the percentage points of the χ^2_{12} distribution are $a = 4.404$ and $b = 23.34$ which leads to the confidence interval $(0.5698, 3.0199)$. For this new interval, the belief that the population variance equals 5 is no longer justified. [4 marks]

4. a) (similar to tutorial exercises)
We calculate summary statistics for this data:

$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{15} = 2010.8 \qquad \text{and} \qquad S^2 = 5964.6$$

We design the following hypothesis test: $H_0 : \mu = 2000$, $H_1 : \mu \neq 2000$. The test statistic is

$$t = \frac{\bar{x} - 2000}{\sqrt{S^2/15}} = \frac{10.8}{\sqrt{5964.6/15}} \approx 0.5416$$

and hence the p-value is

$$2 \times P(t_{14} > 0.5416) \approx 0.6$$

where t_{14} has the t-distribution with 14 degrees of freedom, and $P(t_{14} > 0.5416) \approx 0.3$ since $P(t_{14} < 0.5416) \approx 0.7$ based on the tables, given that $P(t_{14} < 0.5) = 0.6876$ and $P(t_{14} < 0.6) = 0.7210$ in the tables. Hence there is insufficient evidence to reject H_0. [4 marks]

5. a) (similar to tutorial exercises)
We have

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2 /n = 82.5$$
$$S_{yy} = \sum y_i^2 - \left(\sum y_i\right)^2 /n = 944.1$$
$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) /n = -261.5$$

Hence

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.1697$$
$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} = 95.6607$$

The fitted linear regression equation is $y = 95.6607 - 3.1697x$.

[3 marks]

b) (similar to tutorial exercises) Pearson's coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \approx -0.937$$

This indicates a strong negative correlation between the two variables. [2 marks]

c) (similar to tutorial exercises)
A 99% confidence interval for $\hat{\alpha}$ is given by $(\hat{\alpha} - t_{0.5} \times ese(\hat{\alpha}), \hat{\alpha} + t_{0.5} \times ese(\hat{\alpha}))$, where

$$t_{0.5} = 3.355$$

is the 0.5% point of t_8, and

$$ese(\hat{\alpha}) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

We have

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 14.403$$

and hence

$$ese(\hat{\alpha}) = 6.2053\,,$$

which gives the confidence interval

$$(95.6607 - 20.8188, 95.6607 + 20.8188) = (74.8419, 116.4795)\,.$$

[3 marks]

6. a) (similar to tutorial exercises)
From Bayes' theorem,

$$\begin{aligned} p(\theta|X_1,\ldots,X_n) &\propto \theta^{\alpha-1}\exp(-\beta\theta)\theta^{\sum_{k=1}^n X_k}\exp(-n\theta) \\ &\propto \theta^{\alpha-1+\sum_{k=1}^n X_k}\exp(-(\beta+n)\theta)\,. \end{aligned}$$

Hence

$$\theta|X_1,\ldots,X_n \sim \text{Gamma}\left(\alpha + \sum_{k=1}^n X_k, \beta + n\right).$$

[3 marks]

b) (a bit more difficult) The posterior mean and variance for this model are

$$\begin{aligned} E(\theta|\underline{X}) &= \frac{\alpha + \sum_{k=1}^n X_k}{\beta + n} \\ Var(\theta|\underline{X}) &= \frac{\alpha + \sum_{k=1}^n X_k}{(\beta+n)^2}\,. \end{aligned}$$

Since $E(X_k) = \theta$ for each k, we have

$$E(E(\theta|\underline{X})) = \frac{\alpha + n\theta}{\beta + n} = \frac{\frac{\alpha}{n} + \theta}{\frac{\beta}{n} + 1}$$

Hence the posterior mean is not an unbiased estimator of θ, but it is asymptotically unbiased. On the other hand,

$$E(Var(\theta|\underline{X})) = \frac{\frac{\alpha}{n} + \theta}{(\frac{\beta+n}{\sqrt{n}})^2}$$

hence $E(Var(\theta|\underline{X})) \to 0$ as $n \to \infty$, so the posterior variance is not even asymptotically unbiased as an estimator of θ and it should not be used to estimate θ.
[3 marks]