F20SA/F21SA:
Statistical Modelling and Analysis

Exam Solutions 2022-23

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

$$\begin{aligned} P(\text{faulty component}) &= P(\text{faulty component}|A)P(A) + P(\text{faulty component}|B)P(B) \\ &= 0.06 \cdot 0.6 + 0.09 \cdot 0.4 \\ &= 0.072 \end{aligned}$$

Hence

$$P(B|\text{faulty component}) = (0.09 * 0.4)/0.072 = 0.5\,.$$

[3 marks]

b) (similar to tutorial exercises)
Let X be the number of faulty components in a group of 80 components from the order.

Then $X \sim \mathrm{Bin}(80, 0.072)$ and it can be approximated, due to CLT, by $Y \sim N(5.76, 5.34528)$. Hence

$$P(X \le 3) \approx P(Y < 3.5) = P(Z < \frac{3.5 - 5.76}{\sqrt{5.34528}}) = P(Z < -0.9775) = 0.1642\,.$$

[4 marks]

c) (similar to tutorial exercises)
Let Y be the number of faulty components produced by company A that we gathered in our group of size n. Then $Y \sim \mathrm{Bin}(n, 0.06)$. We have

$$P(Y \ge 1) = 1 - P(Y = 0) = 1 - (0.94)^n\,,$$

and we want

$$1 - (0.94)^n > 0.95\,,$$

which gives

$$(0.94)^n < 0.05$$

and hence

$$n > \log(0.05)/\log(0.94) \approx 48.42$$

so there have to be at least 49 components in the group. [3 marks]

2. (this is similar to examples seen previously, the numbers and the setting is different)

a) Here the likelihood function is

$$L(\lambda; x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}.$$

So the log likelihood is given by:

$$l(\lambda; x) = \ln(L(\lambda, x)) = n\ln(\lambda) - \sum_{i=1}^{n} \lambda x_i$$

We differentiate with respect to lambda to obtain the score function:

$$U(\lambda; x) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

To find the MLE for λ we use this to find the minimum which is given by:

$$\bar{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

[4 marks]

b) We have $\sum_{i=1}^{5} x_i = 20.84$, so the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{20.84}{5} = 4.168.$$

Now we know that $X \sim Exp(\lambda)$, so

$$E(X) = 1/\lambda.$$

To find the MME we set $\bar{x} = E(X)$, re-arranging to obtain

$$\hat{\lambda} = \frac{1}{\bar{x}} = 0.2399.$$

[3 marks]

c) Let Y be the number of buses arriving in a day. We are interested in

$$P(Y = 4 | Y > 1) = \frac{P(Y = 4)}{P(Y > 1)},$$

by the definition of conditional probability. From the formula sheet we have the probability mass function is

$$P(Y = y) = \frac{e^{-4.6}(4.6)^y}{y!},$$

so

$$P(Y > 1) = 1 - P(Y = 0) - P(Y = 1) = 1 - e^{-4.6} - e^{-4.6}(4.6) = 0.9437$$

and

$$P(Y = 4) = (4.6)^4 e^{-4.6}/24 = 0.1875$$

So $P(Y = 4|Y > 1) = 0.1987$ [4 marks]

3. a) Sample mean: $\overline{x} = \frac{\sum x_i}{n} = \frac{612}{16} = 38.25$

Sample variance: $s^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) = \frac{1}{15}\left(25014 - \frac{612^2}{16}\right) = 107$ [2 marks]

b) A 95% CI for σ^2 is given by $\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right)$, where a and b are the 97.5% and 2.5% points of χ^2_{15}

From Tables: $a = 6.262$ and $b = 27.49$

We obtain the interval: $(58.385, 256.308)$. [4 marks]

c) σ^2 is unknown, so the test statistic is $t_s = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{15}$ distribution

The observed value of the test statistic under H_0 is:

$t_s = \frac{38.25-36}{\sqrt{\frac{107}{16}}} = 0.87$

The p-value is: $P(t_{15} > 0.87) = 1 - P(t_{15} < 0.87) > 1 - P(t_{15} < 0.9) = 0.1912 > 0.05$, so we do not reject H_0 at 5% level. [4 marks]

4. a) $S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 5538.02 - \frac{222.6^2}{9} = 32.38$

$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1582.33 - \frac{117.3^2}{9} = 53.52$

$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 2942.08 - \frac{(222.6)(117.3)}{9} = 40.86$ [3 marks]

b) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{40.86}{32.38} = 1.26$

$\hat{\alpha} = \overline{y} - \hat{\beta} \times \overline{x} = \frac{117.3}{9} - 1.26 \times \frac{222.6}{9} = -18.13$

So, Vol$= -18.13 + 1.26$Temp. [3 marks]

c) $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{40.86}{\sqrt{(32.38)(53.52)}} \simeq 0.98$

r is very close to 1, which shows a very strong positive correlation between the temperature at noon and the volume of water sold. [3 marks]

d) A 95%-CI for $\hat{\beta}$ is given by $\left(\hat{\beta} \pm t \times ese\left(\hat{\beta}\right)\right)$, where t is the 2.5% point of $t_{n-2} = t_7$

$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{7}\left(53.52 - \frac{40.86^2}{32.38}\right) = 0.28$

$ese(\hat{\beta}) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{0.28^2}{32.38}} = 0.0492$

$t = 2.365$ from the percentage points table of student t-distribution

A 95%-CI for $\hat{\beta}$ is $(1.26 \pm 2.365 \times 0.0492) = (1.144, 1.376)$. [6 marks]

e) The estimates for the volume of water expected to be sold in the 3 days are:

$y_1 = -18.13 + 1.26 \times 24 = 12.11$ kl

$y_2 = -18.13 + 1.26 \times 25 = 13.37$ kl

$y_3 = -18.13 + 1.26 \times 26.5 = 15.26$ kl

The total volume is $12.11 + 13.37 + 15.26 = 40.74$ kl. [2 marks]

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

$$p(\theta|x_1, \ldots, x_n) \propto \theta^n (1-\theta)^{\sum_{k=1}^n x_k} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \\ \propto \theta^{n+\alpha-1}(1-\theta)^{\sum_{k=1}^n x_k + \beta - 1},$$

and hence

$$\theta|x_1, \ldots, x_n \sim \text{Beta}(\alpha', \beta')$$

with parameters $\alpha' = \alpha + n$ and $\beta' = \beta + \sum_{k=1}^n x_k$.

Conjugate priors are prior distributions that lead to posterior distributions that are in the same parametric family, in this case the Beta distribution is conjugate.

[4 marks]

b) (Similar to class examples) Using the fact that he mean and variance of a beta r.v. $Z \sim \text{Beta}(\alpha, \beta)$ are $\text{E}(Z) = \alpha/(\alpha+\beta)$ and $\text{Var}(Z) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$ we obtain that

$$\text{E}(\theta|\underline{x}) = \frac{\alpha + n}{\alpha + \beta + n + \sum_{k=1}^n x_k}$$

and

$$\text{Var}(\theta|\underline{x}) = \frac{(\alpha+n)(\beta + \sum_{k=1}^n x_k)}{(\alpha + n + \beta + \sum_{k=1}^n x_k + 1)(\alpha + n + \beta + \sum_{k=1}^n x_k)^2}.$$

[2 marks]

c) (Unseen, but related to class examples and tutorial exercises)

Let $\hat{\theta} = (1 + \bar{X})^{-1}$ denote the maximum likelihood estimator (MLE) for θ. We know that $\hat{\theta}$ is asymptotically unbiased. So to show that the posterior mean $\text{E}(\theta|\underline{X})$ is an asymptotically unbiased estimator for θ it suffices to show that it converges to $\hat{\theta}$ as n increases (i.e., that the difference between $\text{E}(\theta|\underline{X})$ and $\hat{\theta}$ vanishes as n increases).

$$
\begin{aligned}
\mathrm{E}(\theta|\underline{X}) &= \frac{\alpha+n}{\alpha+\beta+n+\sum_{k=1}^{n} X_k}\,, \\
&= \frac{\alpha/n+1}{(\alpha+\beta)/n+1+\sum_{k=1}^{n} X_k/n}\,, \\
&= \frac{1}{(\alpha+\beta)/n+1+\sum_{k=1}^{n} X_k/n} + \frac{\alpha/n}{(\alpha+\beta)/n+1+\sum_{k=1}^{n} X_k/n}\,, \\
&= \frac{1}{1+\sum_{k=1}^{n} X_k/n} - \frac{(\alpha+\beta)/n}{[(\alpha+\beta)/n+1+\sum_{k=1}^{n} X_k/n](1+\sum_{k=1}^{n} X_k/n)} \\
&\quad + \frac{\alpha/n}{(\alpha+\beta)/n+1+\sum_{k=1}^{n} X_k/n} \\
&= \frac{1}{1+\bar{X}} + \mathcal{O}(n^{-1})
\end{aligned}
$$

where $\mathcal{O}(n^{-1})$ gathers all terms that vanishes as n increases at a rate $1/n$.

Hence, as $n \to \infty$, the posterior mean coincides with the MLE and is asymptotically unbiased as a result. [5 marks]