F20SA/F21SA: Statistical Modelling and Analysis

Exam Solutions 2020-21

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

$$A := \text{ phone produced by } A$$
$$B := \text{ phone produced by } B.$$

Hence

$$P(\text{faulty}) = P(\text{faulty}|A)P(A) + P(\text{faulty}|B)P(B) = \frac{9}{100} \cdot \frac{2}{3} + \frac{1}{10} \cdot \frac{1}{3} = \frac{7}{75}.$$

Furthermore,

$$P(B|\text{faulty}) = \frac{P(\text{faulty}|B)P(B)}{P(\text{faulty})} = \frac{1}{30} \Big/ \frac{7}{75} = \frac{15}{42}.$$

[3 marks]

b) (almost identical to an example from lecture notes)
Let $X \sim Exp(1/100)$ be the waiting time with mean 100. We know that the cdf $F_X(x) = 1 - e^{-x/100}$. Let Y be the remaining waiting time after 120 days. We have

$$P(Y \leq y) = 1 - P(Y > y) = 1 - P(X > y + 120|X > 120)$$

and

$$P(X > y+120|X > 120) = \frac{P(X > y + 120)}{P(X > 120)} = e^{-(y+120)/100}e^{120/100} = e^{-y/100}.$$

Hence we see that $Y \sim Exp(1/100)$ and thus $E(Y) = 100$. [2 marks]

c) (similar to tutorial exercises)
Denote $Y := \sum_{i=1}^{1000} X_i$, where $X_i \sim Bernoulli(7/75)$ are i.i.d. Hence Y is the number of faulty phones in the batch. We have

$$E(X_i) = 7/75\,, \qquad E(X_i^2) = 7/75\,, \qquad Var(X_i) = \frac{7}{75} - \left(\frac{7}{75}\right)^2 \approx 0.0846.$$

From CLT we know that Y has approximately $N(1000 \cdot \frac{7}{75}, 1000Var(X_i))$ distribution. Hence

$$Z := \left(Y - \frac{7000}{75}\right) \Big/ 10\sqrt{10Var(X_i)}$$

has approximately $N(0,1)$ distribution. As a consequence, after applying a continuity correction we obtain

$$P(Y \leq 80) = P(Z \leq -1.3953) = 1 - 0.9192 = 0.0808.$$

[4 marks]

2. a) (more difficult, combines several exercises from lectures)
Denote by X the number of the bacteria in 1 litre of water. We know that $X \sim Poi(\lambda)$ and hence $P(X = 0) = e^{-\lambda}$ and $P(X > 0) = 1 - e^{-\lambda}$. Similarly, the probability that there are no bacteria in 2 litres of water is equal to $e^{-2\lambda}$. Hence the likelihood function for our sample x is given by

$$L(\lambda; x) = e^{-38\lambda}(1 - e^{-\lambda})^{12}e^{-10\lambda}$$

Denote for convenience $q := e^{-\lambda}$ and compute the log-likelihood function

$$l(\lambda) = 48 \log q + 12 \log(1 - q)\,.$$

Hence

$$l'(\lambda) = \frac{48}{q} - \frac{12}{1-q}\,.$$

Solving $l'(\lambda) = 0$ we obtain

$$q = 4/5$$

which gives $\lambda = \log(5/4)$. Now we know that the number of the bacteria in 100 litres of water has the $Poi(100 \log(5/4))$ distribution, and, because the expectation of a Poisson random variable is equal to the parameter of the Poisson distribution, we conclude the proof.

[5 marks]

3. a) (similar to examples discussed in lectures)
We will use the test statistic

$$(\bar{x} - 20)/(s/\sqrt{n})\,,$$

which, for $n = 50$, has approximately the $N(0,1)$ distribution. For $Z \sim N(0,1)$ we have

$$P(Z > 2.3263) = 0.01\,,$$

hence our critical region is $(2.3263, \infty)$. On the other hand, the value of our test statistic is

$$\frac{0.68}{2.59}\sqrt{50} \approx 1.8565 < 2.3263$$

hence there is not enough evidence to reject H_0. [3 marks]

b) (similar to tutorial exercises)

The p-value is $P(Z > 1.8565) \approx 0.032$. Hence there is enough evidence to reject H_0 at significance level 5%.

[2 marks]

c) (similar to tutorial exercises)
Denote the new sample after removing $x_1 = 9.02$ and $x_{50} = 29.95$ by $\mathbf{y}$. We need to compute the mean $\bar{y}$ and the standard deviation s_y of $\mathbf{y}$. We have

$$\bar{x} = \frac{x_2 + \ldots + x_{49} + x_1 + x_{50}}{50} = 20.68$$

hence

$$x_2 + \ldots + x_{49} = 50 \cdot 20.68 - 9.02 - 29.95 = 995.03$$

and

$$\bar{y} = \frac{x_2 + \ldots + x_{49}}{48} \approx 20.73.$$

Next we calculate

$$2.59^2 = s^2 = \frac{1}{49}\left(\sum_{i=1}^{50} x_i^2 - \frac{(\sum_{i=1}^{50} x_i)^2}{50}\right)$$

and hence

$$\sum_{i=1}^{50} x_i^2 = 49 \cdot 2.59^2 + \frac{(50 \cdot 20.68)^2}{50} \approx 21711.817$$

Thus we have

$$\sum_{i=2}^{49} x_i^2 = 21711.817 - 9.02^2 - 29.95^2 = 20733.454$$

which allows us to compute

$$s_y^2 = \frac{1}{47}\left(20733.454 - \frac{995.03^2}{48}\right) \approx 2.27$$

and

$$s_y \approx 1.51.$$

Hence the value of our test statistic for the new sample is

$$(\bar{y} - 20)/(s_y/\sqrt{48}) \approx 3.3494 > 2.3263$$

and now there is enough evidence to reject H_0. [4 marks]

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

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2 / n = 2062.5$$
$$S_{yy} = \sum y_i^2 - \left(\sum y_i\right)^2 / n = 1288.5$$
$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right) / n = -1622.5$$

[2 marks]

b) (similar to tutorial exercises)

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -0.7867$$
$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} = 209.8067$$

The fitted linear regression equation is $y = 209.8067 - 0.7867x$.

[3 marks]

c) (similar to tutorial exercises)
The estimated value of y for $x_0 = 200$ is $\hat{y} = 52.4667$. From the discussion in Chapter 9 of the lecture notes, we know that the 99% one-sided confidence interval for $\hat{y}$ is given as

$$\left[\hat{y} - ts \times \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \infty\right),$$

where $t = 2.896$ is the 1% point of t_8, and

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 1.5167\,.$$

Hence the resulting 99% CI is given by $[52.4667 - 1.1448, \infty) = [51.3219, \infty)$. We see that the measured melting time for the temperature $200°C$ was 51 min, which does not belong to the 99% CI that we calculated. This suggests that the linear regression model does not fit the data well. [4 marks]

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

$$p(\theta|X_1, \ldots, X_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{k=1}^{n}\left(\theta(1-\theta)^{X_k}\right)$$
$$= \theta^{\alpha+n-1}(1-\theta)^{\beta-1+\sum_{k=1}^{n} X_k}\,.$$

Hence

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

[3 marks]

b) (similar to tutorial exercises)
The posterior mean and variance for this model are

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

[2 marks]

c) (more difficult, requires combining several facts from lectures)
Note that the prior $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ corresponds to taking $\alpha = 0$, $\beta = 0$ in (a). Hence the posterior mean is

$$E(\theta|\underline{X}) = \frac{n}{n + \sum_{k=1}^n X_k} = \frac{1}{1 + \bar{X}} ,$$

where

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k .$$

By the law of large numbers, we have

$$E(\theta|\underline{X}) \to \frac{1}{1 + E(X_1)} = \frac{\theta}{1 + \theta} ,$$

as $n \to \infty$, since $E(X_1) = 1/\theta$. Hence we have shown that the posterior mean $E(\theta|\underline{X})$ is not a consistent estimator of θ. [3 marks]