A9 Statistics 2019 – solutions

1. (a) (i) [B, substantial/harder as no help given]

$Z_i = (X_i - \mu)/\sigma \overset{\text{iid}}{\sim} N(0,1)$ so have joint pdf

$$f(\mathbf{z}) = (2\pi)^{-n/2} e^{-\sum_i z_i^2/2} \quad \mathbf{z} \in \mathbb{R}^n. \tag{1}$$

Let $\mathbf{y} = A\mathbf{z}$ where A is an orthogonal $n \times n$ matrix with first row $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$.
Since $A^T A = I$, we have $\mathbf{z} = A^T \mathbf{y}$ and $|\det(A)| = 1$.
So $\partial z_i / \partial y_j = a_{ji}$ and the Jacobian J satisfies $|J| = |\det(A^T)| = 1$.
Also

$$\sum_{i=1}^n y_i^2 = \mathbf{y}^T \mathbf{y} = \mathbf{z}^T A^T A \mathbf{z} = \mathbf{z}^T \mathbf{z} = \sum_{i=1}^n z_i^2. \tag{2}$$

Hence the pdf of $\mathbf{Y}$ is

$$\begin{aligned} g(\mathbf{y}) &= f(\mathbf{z}(\mathbf{y})) \cdot |J| \\ &= (2\pi)^{-n/2} e^{-\sum_i y_i^2/2} \cdot 1 \quad \text{using (1), (2) and } |J| = 1. \end{aligned}$$

Hence $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(0,1)$. **[4]**
Now

$$Y_1 = (\text{first row of } A) \times \mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i = \sqrt{n}\, \overline{Z}$$

and then

$$\sum_{i=1}^n (Z_i - \overline{Z})^2 = \sum_{i=1}^n Z_i^2 - n\overline{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2. \quad \textbf{[2]}$$

So

- $Y_1, \ldots, Y_n$ are independent
- $\overline{Z}$ is a function of Y_1 only
- $\sum_{i=1}^n (Z_i - \overline{Z})^2$ is a function of $Y_2, \ldots, Y_n$ only

and hence $\overline{Z}$ and $\sum_{i=1}^n (Z_i - \overline{Z})^2$ are independent.
Therefore $\overline{X}$ and S^2 are independent since $\overline{X} = \sigma \overline{Z} + \mu$ and $S^2 = \frac{\sigma^2}{n-1} \sum_{i=1}^n (Z_i - \overline{Z})^2$. **[3]**

(ii) [B]

- $\overline{X} = \sigma \overline{Z} + \mu = \frac{\sigma}{\sqrt{n}} Y_1 + \mu \sim N(\mu, \frac{\sigma^2}{n})$ since $Y_1 \sim N(0,1)$
- $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n (Z_i - \overline{Z})^2 = \sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$ since $Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} N(0,1)$, which gives $S^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$
- The normal and the χ^2 just above are independent by (i). So, independently,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}.$$

So from the definition of a t-distribution,

$$\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \right) \Big/ (S/\sigma) = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \quad \textbf{[3]}$$

(iii) [N, but have seen similar]

We have $\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$ and independently $\frac{1}{2\sigma^2}\sum Y_j^2 \sim \chi^2_m$.

So $\frac{1}{\sigma^2}\left[(n-1)S_X^2 + \sum Y_j^2/2\right] \sim \chi^2_{m+n-1}$ and

$$P\left(c_1 < \frac{1}{\sigma^2}\left[(n-1)S_X^2 + \sum Y_j^2/2\right] < c_2\right) = 1-\alpha$$

where $\alpha/2 = P(\chi^2_{m+n-1} < c_1) = P(\chi^2_{m+n-1} > c_2)$. Hence

$$P\left(\frac{1}{c_2}\left[(n-1)S_X^2 + \sum Y_j^2/2\right] < \sigma^2 < \frac{1}{c_1}\left[(n-1)S_X^2 + \sum Y_j^2/2\right]\right) = 1-\alpha$$

gives the required confidence interval. **[4]**

(b) (i) [B, here to setup the next part]

If $X \sim N(0,1)$ then $\Phi(X) \sim U(0,1)$, so $X \sim \Phi^{-1}(U)$. Hence

$$\begin{aligned} E[X_{(k)}] &= E[\Phi^{-1}(U_{(k)})] \\ &\approx \Phi^{-1}(E[U_{(k)}]) \\ &= \Phi^{-1}(\tfrac{k}{n+1}). \end{aligned}$$

So if $x_1, \ldots, x_n$ are from $N(0,1)$ then we should have $x_{(k)} \approx \Phi^{-1}(\frac{k}{n+1})$.

A Q-Q plot is a plot of $x_{(k)}$ against $\Phi^{-1}(\frac{k}{n+1})$ for $k = 1, \ldots, n$. **[3]**

If the observations really are from $N(0,1)$ then the plotted points should lie approx on the line $y = x$. **[1]**

(ii) [N] If the observations really are from F then, replacing Φ by F above, we should have

$$y_{(k)} \approx F^{-1}\left(\frac{k}{n+1}\right) \quad \text{or} \quad F(y_{(k)}) \approx \frac{k}{n+1}.$$

That is,

$$\begin{aligned} \frac{1}{1+(y_{(k)}/\alpha)^{-\beta}} &\approx \frac{k}{n+1} \\ \left(\frac{y_{(k)}}{\alpha}\right)^{-\beta} &\approx \frac{n+1-k}{k} \\ \log y_{(k)} &\approx \log\alpha + \frac{1}{\beta}\log\left(\frac{k}{n+1-k}\right). \end{aligned}$$

So if we plot the values of $y_{(k)}$ against $\log\left(\frac{k}{n+1-k}\right)$ we should see an approximate straight line (intercept $\log\alpha$, gradient $1/\beta$) if the assumption is reasonable. **[5]**

2. (a) [B]

(i) The size of the test is $P(\text{reject } H_0 \,|\, H_0 \text{ true})$, or $P(\mathbf{X} \in C \,|\, \theta_0)$ where C is the critical/rejection region. **[1]**

The power of the test is $P(\text{reject } H_0 \,|\, H_1 \text{ true})$, or $P(\mathbf{X} \in C \,|\, \theta_1)$. **[1]**

(ii) *NP lemma*: Consider testing the given null H_0 against the given alternative H_1. Let $L(\theta; \mathbf{x})$ be the likelihood function. Define critical region C by

$$C = \left\{ \mathbf{x} : \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})} \leqslant k \right\}$$

and suppose constants k and α are such that $P(\mathbf{X} \in C \,|\, H_0) = \alpha$. Then among all tests of size $\leqslant \alpha$, the test with critical region C has maximum power. **[2]**

(b) [S]

(i) The likelihood is

$$L(\lambda) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod x_i!}.$$

First: let λ_1 be a fixed value of λ satisfying $\lambda_1 > 1$. By the NP lemma, the most powerful test of H_0 against the alternative $\lambda = \lambda_1$ is of the form:

$$\begin{aligned} \text{reject } H_0 &\iff \frac{L(1; \mathbf{x})}{L(\lambda_1; \mathbf{x})} \leqslant k \\ &\iff \frac{e^{-n}}{e^{-n\lambda_1} \lambda_1^{\sum x_i}} \leqslant k \\ &\iff \lambda_1^{\sum x_i} \geqslant k_1 \\ &\iff \sum x_i \geqslant c \quad \text{since } \lambda_1 > 1. \end{aligned}$$ **[3]**

For this test to have size 0.01 we need

$$\begin{aligned} 0.01 &= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\ &= P(\textstyle\sum X_i \geqslant c \,|\, \lambda = 1) \\ &= P(\text{Poisson}(20) \geqslant c) \\ &= g_{20}(c). \end{aligned}$$

So from the hint we want $c = 32$ and the most powerful test has critical region $\{\mathbf{x} : \sum x_i \geqslant 32\}$. **[3]**

Second: now observe that the critical region just found is the same for all $\lambda_1 > 1$. Hence it gives a uniformly most powerful test of H_0 against the alternative $H_1 : \lambda > 1$. **[1]**

(ii) The power function is

$$\begin{aligned} w(\lambda) &= P(\text{reject } H_0 \,|\, \lambda \text{ is the true parameter value}) \\ &= P\left(\textstyle\sum X_i \geqslant 32 \,|\, \lambda\right) \\ &= P(\text{Poisson}(20\lambda) \geqslant 32) \\ &= g_{20\lambda}(32). \end{aligned}$$ **[2]**

(iii) Let $t = \sum_{i=1}^{20} x_i$. The p-value is

$$\begin{aligned} p &= P\left(\sum X_i \geqslant t \,|\, H_0\right) \\ &= P\left(\sum X_i \geqslant t \,|\, \lambda = 1\right) \\ &= P(\text{Poisson}(20) \geqslant t) \\ &= g_{20}(t) = g_{20}(\textstyle\sum x_i). \end{aligned}$$

[2]

(c) [N] The likelihood and log-likelihood are

$$\begin{aligned} L(\beta,\gamma) &= \frac{\gamma^n}{\beta^n}\left(\prod x_i\right)^{\gamma-1}\exp\left(-\sum x_i^\gamma/\beta\right) \\ \ell(\beta,\gamma) &= n\log\gamma - n\log\beta + \gamma\sum\log x_i - \frac{1}{\beta}\sum x_i^\gamma + \text{constant}. \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= -\frac{n}{\beta} + \frac{1}{\beta^2}\sum x_i^\gamma \\ &= 0 \qquad \text{when } \beta = \beta_\gamma = \frac{1}{n}\sum x_i^\gamma \end{aligned}$$

and

$$\frac{\partial \ell}{\partial \gamma} = \frac{n}{\gamma} + \sum \log x_i - \frac{1}{\beta}\sum x_i^\gamma \log x_i$$

and so $\widehat{\gamma}$ solves $\frac{n}{\gamma} + \sum\log x_i - \frac{1}{\beta_\gamma}\sum x_i^\gamma \log x_i = 0$ (numerical solution required). **[3]**

Under H_0: $\gamma = 1$ and MLE of β is $\widehat{\beta} = \overline{x}$ with maximised likelihood

$$L_0 = L(\widehat{\beta}_0, 1) = \frac{1}{\overline{x}^n}e^{-n}.$$

Under H_1: general γ, MLEs of $\widehat{\beta}, \widehat{\gamma}$ with maximised likelihood

$$L_1 = L(\widehat{\beta}, \widehat{\gamma}) = \frac{\widehat{\gamma}^n}{\widehat{\beta}^n}e^{-n}\prod x_i^{\widehat{\gamma}-1}.$$

And the likelihood ratio test is

$$\begin{aligned} \text{reject } H_0 &\iff \frac{L_0}{L_1} \leqslant k \\ &\iff \frac{1/\overline{x}^n}{\widehat{\gamma}^n\left(n\Big/\sum x_i^{\widehat{\gamma}}\right)^n \prod x_i^{\widehat{\gamma}-1}} \leqslant k \\ &\iff \overline{x}^n g(\widehat{\gamma}) \geqslant c = 1/k \end{aligned}$$

where $g(\widehat{\gamma}) = \left(n\widehat{\gamma}\Big/\sum x_i^{\widehat{\gamma}}\right)^n \prod x_i^{\widehat{\gamma}-1}$. **[5]**

We have $-2\log(L_0/L_1) = 2\log\left(\overline{x}^n g(\widehat{\gamma})\right) \approx \chi_1^2$ under H_0.

So for a test of size α we reject iff $2\log\left(\overline{x}^n g(\widehat{\gamma})\right) > c_\alpha$, where $P(\chi_1^2 \geqslant c_\alpha) = \alpha$. **[2]**

3. (a) [B]

$$\pi(\theta \,|\, \mathbf{x}) = \frac{\pi(\theta) \prod_{i=1}^{n} f(x_i \,|\, \theta)}{\int \pi(\theta) \prod_{i=1}^{n} f(x_i \,|\, \theta)\, d\theta} \qquad \textbf{[2]}$$

(b) [B/S]

(i)

$$\begin{aligned} \pi(\theta \,|\, \mathbf{x}) &\propto f(\mathbf{x} \,|\, \theta)\pi(\theta) \\ &= \left(\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \right) \beta e^{-\beta\theta} \\ &\propto \theta^r e^{-(n+\beta)\theta} \end{aligned}$$

where $r = \sum x_i$. Hence $\theta \,|\, \mathbf{x} \sim \text{Gamma}(r+1, n+\beta)$. **[4]**

(ii)

$$\begin{aligned} P(X_{n+1} = k \,|\, \mathbf{x}) &= \int P(X_{n+1} = k \,|\, \theta)\pi(\theta \,|\, \mathbf{x})\, d\theta \\ &= \int \frac{e^{-\theta}\theta^k}{k!} \frac{(n+\beta)^{r+1}}{\Gamma(r+1)} \theta^r e^{-(n+\beta)\theta}\, d\theta \\ &= \frac{(n+\beta)^{r+1}}{k!\,\Gamma(r+1)} \int \theta^{r+k} e^{-(n+\beta+1)\theta}\, d\theta \\ &= \frac{(n+\beta)^{r+1}}{k!\,\Gamma(r+1)} \frac{\Gamma(r+k+1)}{(n+\beta+1)^{r+k+1}} \int \text{Gamma pdf} \\ &= \frac{(n+\beta)^{r+1}}{k!\,\Gamma(r+1)} \frac{\Gamma(r+k+1)}{(n+\beta+1)^{r+k+1}}. \end{aligned} \qquad \textbf{[5]}$$

(c) (i) [B]

$$\begin{aligned} \pi(\theta \,|\, \mathbf{y}) &\propto f(\mathbf{y} \,|\, \theta)\pi(\theta) \\ &\propto \exp\left[-\frac{1}{2} \sum (y_i - \theta)^2 \right] \exp\left[-\frac{1}{2} \frac{(\theta - \mu)^2}{\sigma^2} \right] \end{aligned}$$

Now

$$\begin{aligned} \frac{(\theta-\mu)^2}{\sigma^2} + \sum (y_i - \theta)^2 &= \theta^2 \left(\frac{1}{\sigma^2} + n \right) - 2\theta \left(\frac{\mu}{\sigma^2} + n\bar{y} \right) + \text{constant} \\ &= \frac{1}{\sigma_1^2}(\theta - \mu_1)^2 + \text{constant} \end{aligned} \qquad \textbf{[3]}$$

where

$$\mu_1 = \frac{\frac{1}{\sigma^2}\mu + n\bar{y}}{\frac{1}{\sigma^2} + n} \quad \text{and} \quad \frac{1}{\sigma_1^2} = \frac{1}{\sigma^2} + n. \qquad (3)$$

Hence

$$\pi(\theta \,|\, \mathbf{x}) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\theta - \mu_1)^2 \right)$$

and so $\pi(\theta \,|\, \mathbf{y})$ is a $N(\mu_1, \sigma_1^2)$ pdf. That is $\theta \,|\, \mathbf{y} \sim N(\mu_1, \sigma_1^2)$. **[3]**

(ii) [S] The log-likelihood is

$$\ell(\theta) = \text{constant} - \frac{1}{2}\sum(y_i - \theta)^2$$

and so

$$\ell''(\theta) = \text{constant}.$$

So the Fisher information $I(\theta)$ is a constant.
Hence the Jeffreys prior is $\pi(\theta) \propto I(\theta)^{1/2} \propto 1$, a constant improper prior. **[3]**

(iii) [N] The Jeffreys prior is improper and can be approached by considering a sequence of priors for θ with $\pi(\theta) \sim N(\mu, \sigma^2)$ and $\sigma \to \infty$. In this limit, from (3),

$$\pi(\theta \,|\, \mathbf{y}) \sim N\left(\overline{y}, \frac{1}{n}\right).$$

For this posterior the highest posterior density interval is symmetric around $\overline{y}$, so the interval is

$$\left(\overline{y} \pm \frac{1.96}{\sqrt{n}}\right).$$ **[5]**