Examen 2014 — Corrigé

Probabilité et Statistiques SIC

Corrigé 1. (16 points)

(a) *(2 points)* There are C_{50}^5 ways to choose 5 numbers out of 50, and C_{11}^2 for 2 stars out of 11, so the total number of combinations is $N = C_{50}^5 \times C_{11}^2 \approx 1.2 \times 10^8$. The probability of winning the jackpot is thus $1/N$, i.e., approximately 8.6×10^{-9}. The winnings are -3 francs in $N-1$ cases and $1.5 \times 10^8 - 3$ francs in one case (winning the jackpot), giving expectation

$$-3 \times (N-1)/N + (1.5 \times 10^8 - 3)/N \approx -1.71 \text{ francs.}$$

(b) *(2 points)* Write B for the event "an American woman aged 40 has breast cancer" and M for "the mammography gives a positive outcome". Following Bayes' rule, the probability we seek is

$$\mathrm{P}(B \mid M) = \frac{\mathrm{P}(M \mid B)\mathrm{P}(B)}{\mathrm{P}(M \mid B)\mathrm{P}(B) + \mathrm{P}(M \mid B^c)\mathrm{P}(B^c)},$$

with $\mathrm{P}(B) = 1 - \mathrm{P}(B^c) = 0.01$, $\mathrm{P}(M \mid B) = 0.8$, and $\mathrm{P}(M \mid B^c) = 0.1$. With these figures we get $\mathrm{P}(B \mid M) \approx 0.075$, i.e., only 7.5% of women with positive mammographies actually have breast cancer. Writing M_i, $i = 1, 2$, for "the ith mammography gives a positive outcome", we get

$$\begin{aligned}\mathrm{P}(B \mid M_1 \cap M_2) &= \frac{\mathrm{P}(M_1 \cap M_2 \mid B)\mathrm{P}(B)}{\mathrm{P}(M_1 \cap M_2 \mid B)\mathrm{P}(B) + \mathrm{P}(M_1 \cap M_2 \mid B^c)\mathrm{P}(B^c)} \\ &= \frac{\mathrm{P}(M_1 \mid B)\mathrm{P}(M_2 \mid B)\mathrm{P}(B)}{\mathrm{P}(M_1 \mid B)\mathrm{P}(M_2 \mid B)\mathrm{P}(B) + \mathrm{P}(M_1 \mid B^c)\mathrm{P}(M_2 \mid B^c)\mathrm{P}(B^c)},\end{aligned}$$

using the conditional independence of mammography outcomes. The probability for this woman to have breast cancer is about 39%.

(c) *(2 points)* The median is the value of x satisfying $F(x) = 0.5$, i.e., $1 - (\beta/x)^\alpha = 0.5$, so $x = 2^{1/\alpha}\beta$.

(d) *(2 points)* Writing $f_X(x) = 1_{(0,1)}(x)$ for the uniform density, $g(x) = 1/x^2$ for the transformation, and $f_Y(y)$ for the density of Y, we have

$$f_Y(y) = \left|\frac{dg^{-1}}{dy}(y)\right| \times f_X\left\{g^{-1}(y)\right\}.$$

We derive $g^{-1}(y) = \sqrt{1/y} \in (0,1)$ for $y > 1$, and thus $dg^{-1}(y)/dy = -0.5x^{-3/2}$. We end up with $f_Y(y) = 0.5y^{-3/2} \times 1_{(1,\infty)}(x)$.
Alternatively we might write

$$\mathrm{P}(Y \le y) = \mathrm{P}(1/X^2 \le y) = \mathrm{P}(X \ge y^{-1/2}) = 1 - y^{-1/2}, \quad y > 1,$$

since $X \sim U(0,1)$, and differentiation yields $f_Y(y) = y^{-3/2}/2$, for $y > 1$.

(e) *(2 points)* We first calculate the moment-generating function of X_1 :

$$M_{X_1}(t) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \int_0^\infty \lambda e^{-(\lambda - t)x}dx = \left[-\frac{\lambda}{\lambda - t}e^{-(\lambda-t)x}\right]_{x=0}^{x=\infty} = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

The moment-generating function for Z can be expressed in terms of the moment-generating functions of X_1 and X_2 as follows, using at the second line the independence of X_1 and X_2 :

$$\begin{aligned}M_Z(t) &= \mathrm{E}\left(e^{tZ}\right) = \mathrm{E}\left\{e^{t(X_1 - X_2)}\right\} = \mathrm{E}\left(e^{tX_1}e^{-tX_2}\right) \\ &\overset{\text{ind}}{=} \mathrm{E}\left(e^{tX_1}\right) \times \mathrm{E}\left(e^{-tX_2}\right) = M_{X_1}(t) \times M_{X_2}(-t) \\ &= \frac{\lambda}{\lambda - t} \times \frac{\lambda}{\lambda + t} = \frac{\lambda^2}{\lambda^2 - t^2}, \quad |t| < \lambda.\end{aligned}$$

(f) *(2 points)* The variance can be simplified as :

$$\begin{aligned}\text{var}(4+2X-5Y) &= 2^2\text{var}(X)+5^2\text{var}(Y)-2\times 2\times 5\text{cov}(X,Y)\\ &= 4\text{var}(X)+25\text{var}(Y)-20\text{cor(X, Y)}\sqrt{\text{var(X)var(Y)}}.\end{aligned}$$

With the figures we are given, we derive $\text{var}(4+2X-5Y) = 4\times 4+25\times 9-20\times(-0.5)\times 6 = 301$.

(g) *(2 points)* The density functions are $f_X(x_i) = \lambda e^{-\lambda x_i}$ and $f_Y(y_i) = \lambda^{-1}e^{-y_i/\lambda}$ respectively, yielding likelihood function

$$L(\lambda; x_1,\ldots,x_n,y_1,\ldots,y_n) = \prod_{i=1}^{n} e^{-\lambda x_i}e^{-y_i/\lambda}$$

andcorresponding log-likelihood

$$\ell(\lambda; x_1,\ldots,x_n,y_1,\ldots,y_n) = -\lambda n\bar{x} - n\bar{y}/\lambda,$$

which is maximised when

$$\frac{d\ell(\hat{\lambda})}{d\lambda} = 0 \iff -n\bar{x}+\frac{n\bar{y}}{\lambda^2} = 0 \iff \lambda = \pm\sqrt{\frac{n\bar{y}}{n\bar{x}}};$$

we take the positive root because $\lambda > 0$. We finally verify that the log-likelihood function is concave,

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = -2\frac{n\bar{y}}{\lambda^3} < 0, \quad \lambda > 0,$$

so the value $\hat{\lambda}$ gives a maximum and is unique.

(h) *(2 points)* The significance level is the probability of observing the value of a test statistic T as large as the value t_{obs} calculated from the data, or larger, calculated under the assumption that the null hypothesis is true. Thus this tells us that the event $T \geq t_{\text{obs}}$ has a probability of 0.04, if the null hypothesis is true. The value 0.04 casts some doubt on the truth of the null hypothesis, but not a lot.

Corrigé 2. *(8 points)*

(a) *(4 points : 1 for c, 1 for the marginal density, 1 for the conditional density, 1 for lack of independence)* We first obtain c :

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{S,T}(s,t)dtds = 1 \iff c\int_0^{\infty}\left[-(1+s)^{-3}e^{-t/(1+s)}\right]_{t=0}^{t=\infty} ds = 1$$
$$\iff c\int_0^{\infty}(1+s)^{-3}ds = c\left[-\frac{1}{2}(1+s)^{-2}\right]_0^{\infty} = 1,$$

i.e., $c = 2$. The marginal density of S is therefore

$$f_S(s) = \int_{-\infty}^{\infty} f_{S,T}(s,t)dt = 2(1+s)^{-3}, \quad s > 0,$$

using the previous computations. Download time depends on file size, as

$$f_{T|S}(t \mid s) = \frac{f_{S,T}(s,t)}{f_S(s)} = \frac{1}{1+s} \times e^{-t/(1+s)}, \quad t > 0, s > 0,$$

depends on s and hence does not equal the marginal density $f_T(t)$.
(It helps below if we notice that this conditional density is exponential with mean $1+s$ minutes.)

(b) *(1 points)* The mean file size is

$$\begin{aligned}\mathrm{E}(S) &= \int_{-\infty}^{\infty} s f_S(s)ds = \int_0^{\infty} 2s(1+s)^{-3}ds \\ &= \left[2s \times \frac{-1}{2}(1+s)^{-2}\right]_0^{\infty} - \int_0^{\infty} 2\frac{-1}{2}(1+s)^{-2}ds \\ &= \left[-1 \times (1+s)^{-1}\right]_0^{\infty} = 1 \text{ GB}.\end{aligned}$$

(c) *(1 points)* Using (a), we get

$$f_{T|S}(t \mid s = 1) = \frac{1}{2} \times e^{-t/2}, \quad t > 0,$$

which is exponential with mean 2 minutes.

(d) *(1 points)* If we recognise the exp(1/3) density, we easily see that the conditional expectation is 3 minutes. Otherwise, we have to compute (integration by parts)

$$\mathrm{E}(T \mid S = 2) = \int_{-\infty}^{\infty} t f_{T|S}(t \mid s = 2)dt = \int_0^{\infty} \frac{t}{3} \times e^{-t/3}dt = \left[\frac{t}{3} \times (-3)e^{-t/3}\right]_0^{\infty} - \int_0^{\infty} \frac{1}{3} \times (-3)e^{-t/3}dt = 3.$$

(e) *(1 points)* Since (a) implies that $\mathrm{E}(T \mid S = s) = 1 + s$, we get from (b) that

$$\mathrm{E}(T) = \mathrm{E}_S\{\mathrm{E}(T \mid S)\} = \mathrm{E}_S(1+S) = 1 + \mathrm{E}(S) = 2 \text{ minutes}.$$

Otherwise one can do the integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t f_{S,T}(s,t)dtds = 2\int_0^{\infty}\int_0^{\infty} t(1+s)^{-4}e^{-t/(1+s)}dtds = \cdots = 2,$$

after some struggle.

Corrigé 3. *(6 points)*

(a) *(4 points : 1 each for correct computation of means and variances, 1 for normality (mention of CLT ?), and 1 for the probability calculation)*

Write $S_1, \ldots, S_{50}$ for the amounts of soup put into small bowls and $L_1, \ldots, L_{40}$ for the amounts of soup put into large bowls. We can use the central limit theorem by assuming that 50 and 40 are sufficiently large to get a good normal approximation for the sums. If so, we have that, approximately,

$$\sum_{i=1}^{50} S_i \sim \mathcal{N}(50 \times 300, 50 \times 30^2) = \mathcal{N}(15, 0.045), \quad \sum_{i=1}^{40} L_i \sim \mathcal{N}(40 \times 600, 40 \times 60^2) = \mathcal{N}(24, 0.144),$$

where the last numbers are in litres and litres^2.

By properties of independent normal variables, the daily total amount of soup consumed has approximate distribution $T \sim \mathcal{N}(15+24, 0.045+0.144) = \mathcal{N}(39, 0.199)$, in litres and litres^2. The probability of not having enough soup is

$$\begin{aligned} \mathrm{P}(T > 40) &= \mathrm{P}\{(T-39)/\sqrt{0.199} \geq (40-39)/\sqrt{0.199}) \\ &\approx \mathrm{P}(Z \geq 2.24) \\ &= 1 - \mathrm{P}(Z \leq 2.24) \\ &= 1 - \Phi(2.24) \\ &\approx 1 - 0.987 \\ &= 0.013, \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the cumulative distribution functon of Z.

(b) *(2 points : 1 for binomial, 1 for calculation)* Sandwiches correspond to independent Bernoulli variables $X_1, \ldots, X_{100}$ with probabilities $1 - e^{-t}$ of being sold after time t, so the number left at time t is $Y = \sum_i X_i \sim B(100, 1 - e^{-t})$. The probability of having sold all sandwiches after 4 hours is therefore

$$\mathrm{P}(Y = 100) = \left(1 - e^{-4}\right)^{100} \approx 0.16.$$

Corrigé 4. *(10 points)*

(a) *(4 points : 1 for the (log) likelihood, 1 for the MLE, 1 for the observed information, 1 for checking the maximum)* The likelihood corresponding to this model is

$$L(\theta; x_1, \ldots, x_{20}) = \prod_{i=1}^{20} e^{-\theta} \frac{\theta_i^x}{x_i!}, \quad \theta > 0,$$

with x_i the number of spam e-mails received on day i, $i = 1, \ldots, 20$. The log-likelihood is

$$\ell(\theta; x_1, \ldots, x_{20}) = -20\theta + \sum_{i=1}^{20} x_i \log(\theta), \quad \theta > 0,$$

apart from an additive constant. The maximum likelihood estimate for θ is derived as follows :

$$\frac{d\ell(\hat{\theta})}{d\theta)} = 0 \iff -20 + \sum_{i=1}^{20} x_i/\hat{\theta} = 0 \iff \hat{\theta} = \bar{x},$$

with $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 12.9$. We verify that $\ell(\theta)$ is concave :

$$\frac{d^2\ell(\theta)}{d\theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{20} x_i < 0, \quad \theta > 0,$$

so $\hat{\theta}$ gives the only maximum.

The observed information is easily derived as $J(\hat{\theta}) = -\dfrac{d^2\ell(\hat{\theta})}{d\theta^2} \approx 1.55$.

(b) *(2 points : 1 for the calculation, 1 for the interpretation)* The asymptotic distribution of $\hat{\theta}$ is $\mathcal{N}\left\{\theta, J(\hat{\theta})^{-1}\right\}$, yielding the 95% confidence interval

$$[\hat{\theta} - J(\hat{\theta})^{-1/2} z_{0.975}, \hat{\theta} + J(\hat{\theta})^{-1/2} z_{0.975}] \approx [11.33, 14.47] \text{ spams per day.}$$

If we repeated this experiment many times, and computed such an interval each time, we would expect (approximately) 95% of our intervals to contain the true value of θ.

(c) *(1 points)* We would reject $\theta = 15$ at the 95% level based on the confidence interval computed in the previous point (and would still reject it even at the 99% level). It seems clear that the mean daily number of spams has dropped.

(d) *(3 points)* (i) A normal QQ-plot compares sample quantiles for a set of data (the order statistics of the data) with the theoretical quantiles of the standard normal distribution. It helps in comparing the data with the $N(0, 1)$ distribution, with perfect agreement when all points lie on the diagonal line. Outliers and curvature may appear if the model is not a good fit. The intercept and the slope can be used to estimate μ and σ respectively, in the case of $N(\mu, \sigma^2)$ data.

(ii) In this case, the two distributions do not seem to agree, since the slope of the graph exceeds unity, so the variance of the data seems to be greater than it should be. However, the small number of data points means that the graph does not give strong evidence against the model.