Exam 2016 — Solutions

Probability and Statistics SIC

Solution 1. (a) (3 points) A probability space corresponding to the experiment is $(\Omega, \mathcal{F}, \Pr)$, where the sample space of X is

$$\Omega = \{2, 3, 4, 5, 6, 7, 8\},$$

the event space is the set of all subsets of Ω, that is

$$\begin{aligned}\mathcal{F} &= \{A : A = \cup_{i\in\mathcal{I}}\{X = i\}, \mathcal{I} \subset \{2, 3, 4, 5, 6, 7, 8\}\} \\ &= \{\emptyset, \{2\}, \ldots, \{8\}, \{2, 3\}, \ldots, \{7, 8\}, \{2, 3, 4\}, \ldots, \{6, 7, 8\}, \ldots, \{2, 3, 4, 5, 6, 7, 8\}\},\end{aligned}$$

and the probability distribution $\Pr : \mathcal{F} \to [0, 1]$ is such that, for all $A \in \mathcal{F}$, $\Pr(A) = \sum_{i\in\mathcal{I}} \Pr(X = i)$, where

i	2	3	4	5	6	7	8	Sum
$\Pr(X = i)$	1/16	2/16	3/16	4/16	3/16	2/16	1/16	1

So,

$$\begin{aligned}\Pr(X = 3 \mid X \leq 3) &= \frac{\Pr(\{X = 3\} \cap \{X \leq 3\})}{\Pr(X \leq 3)} = \frac{\Pr(X = 3)}{\Pr(X = 2) + \Pr(X = 3)} \\ &= \frac{2/16}{1/16 + 2/16} = 2/3.\end{aligned}$$

(b) (1 point) The requested probability is

$$\begin{aligned}\Pr(\text{`Bezukhov wins'}) &= 0.2 + (0.8 \times 0.2) \times 0.2 + (0.8 \times 0.2)^2 \times 0.2 + (0.8 \times 0.2)^3 \times 0.2 + \cdots \\ &= 0.2 \sum_{i=0}^{\infty} 0.16^i = \frac{0.2}{1 - 0.16} = 5/21 \approx 0.238.\end{aligned}$$

Alternatively: Let the random variables B and D denote respectively the number of shots that Bezukhov and Dolokhov would need to hit their opponent for the first time. So, $B \sim \text{Geom}(0.2)$ and $D \sim \text{Geom}(0.8)$. Since Bezukhov shots first, the probability that he wins is

$$\begin{aligned}\Pr(B \leq D) &= \sum_{b\leq d} \Pr(B = b, D = d) = \sum_{d=1}^{\infty}\sum_{b=1}^{d} \Pr(B = b) \times \Pr(D = d) \\ &= \sum_{d=1}^{\infty} 0.8 \times 0.2^{d-1} \sum_{b=1}^{d} 0.2 \times 0.8^{b-1} = 0.2 \times 0.8 \sum_{d=1}^{\infty} 0.2^{d-1} \frac{1 - 0.8^d}{1 - 0.8} \\ &= \frac{0.16}{1 - 0.8}\left(\sum_{d=0}^{\infty} 0.2^d - 0.8 \sum_{d=0}^{\infty} 0.16^d\right) \\ &= \frac{0.16}{1 - 0.8}\left(\frac{1}{1 - 0.2} - \frac{0.8}{1 - 0.16}\right) = 5/21 \approx 0.238.\end{aligned}$$

Alternatively: Denote B the event Bezukhov wins the duel, and let N be the random variable indicating the number of rounds (an attempt for each duelist) needed to end the duel. By the law of total probability, the probability that Bezukhov wins is

$$\Pr(B) = \sum_{i=1}^{\infty} \Pr(B \mid N = i) \times \Pr(N = i).$$

The probabilities that Bezukhov and Dolokhov hit their target do not depend on the round. So, letting p_B and p_D denote respectively the probabilities that Bezukhov and Dolokhov hit their target with a single shot, we have

$$\Pr(B \mid N = i) = \frac{p_B}{p_B + (1 - p_B)p_D} = \frac{1/5}{1/5 + 16/25} = 5/21.$$

Since the duel ends as soon as one of the duelists hits his target, the number or rounds N follows a geometric distribution with probability of success $p = 1 - (1 - p_B)(1 - p_D)$. So, $N \sim \text{Geom}(p = 0.84)$, and

$$\sum_{i=1}^{\infty} \Pr(N = i) = 1.$$

Therefore,

$$\Pr(B) = \sum_{i=1}^{\infty} \frac{5}{21} \times \Pr(N = i) = 5/21 \approx 0.238.$$

(c) (2 points) Since $Y \mid X = x \sim \text{Pois}(x + 1)$, we have $\text{E}(Y \mid X = x) = x + 1$. Thus, by the law of iterated expectations,

$$\text{E}(Y) = \text{E}_X\{\text{E}(Y \mid X)\} = \text{E}_X(X + 1) = \text{E}(X) + 1 = 3/2.$$

Alternatively: The requested expectation is

$$\begin{aligned}\text{E}(Y) &= \sum_{x=0}^{\infty} \text{E}(Y \mid X = x) \times \Pr(X = x) \\ &= \text{E}(Y \mid X = 0)\Pr(X = 0) + \text{E}(Y \mid X = 1)\Pr(X = 1) \\ &= 1 \times 1/2 + 2 \times 1/2 = 3/2.\end{aligned}$$

Bayes' theorem gives

$$\begin{aligned}\Pr(X = 1 \mid Y = y) &= \frac{\Pr(Y = y \mid X = 1)\Pr(X = 1)}{\Pr(Y = y \mid X = 1)\Pr(X = 1) + \Pr(Y = y \mid X = 0)\Pr(X = 0)} \\ &= \frac{2^y e^{-2}/y! \times 1/2}{2^2 e^{-2}/y! \times 1/2 + 1^y e^{-1}/y! \times 1/2} = \frac{2^y}{2^y + e}, \quad y = 0, 1, 2, \ldots.\end{aligned}$$

(d) (2 points) If $X_1 = X_2$, then $Z = 0$, which happens with probability $p^2 + (1 - p)^2 = 1 - 2p(1-p)$. If $X_1 \neq X_2$, then $Z = 1$, which happens with probability $p(1-p) + (1-p)p = 2p(1 - p)$. Thus,

$$\Pr(Z = z) = \begin{cases} 1 - 2p(1 - p), & z = 0, \\ 2p(1 - p), & z = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, the probabilities sum to one. So, Z is a Bernoulli random variable with parameter $2p(1 - p)$.

(e) (1 point) We have

$$M_Z(t) = \mathrm{E}\left\{e^{tZ}\right\} = \mathrm{E}\left\{e^{t(a+bX)}\right\} = e^{ta}\mathrm{E}\left(e^{btX}\right) = e^{ta}M_X(bt).$$

(f) (1 point) This is not possible. Indeed, if the sample variance is 0, then all observations are equal (to the sample mean), in which case the lower and upper quartiles are equal yielding an inter-quartile range of zero.

(g) (2 points) Since $X \sim \exp(\lambda)$, we have $F_X(x) = 1 - e^{-\lambda x}$, $x > 0$. Letting $Y = 1/X$, we have

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(1/X \le y) = \Pr(1/y \le X) = 1 - \Pr(X < 1/y) \\ &= 1 - F_X(1/y) = e^{-\lambda/y}, \quad y > 0, \lambda > 0. \end{aligned}$$

The p quantile of $1/X$ is the value y_p that solves $F_Y(y_p) = p$. Thus $y_p = -\lambda/\log p$.

Alternatively: Since $X \sim \exp(\lambda)$, we have $f_X(x) = \lambda e^{-\lambda x}$, $x > 0$. We have $Y = g(X)$ where the transformation $g(x) = 1/x$, $x > 0$ and its inverse $g^{-1}(y) = 1/y$ are both monotonic. Then

$$f_Y(y) = f_X\{g^{-1}(y)\} \times \left|\frac{dg^{-1}(y)}{dy}\right| = \lambda e^{-\lambda/y} \times \frac{1}{y^2}, \quad y > 0, \lambda > 0.$$

Thus,

$$F_Y(y) = \int_0^y \lambda e^{-\lambda/t} \times \frac{1}{t^2}\, dt = \left[e^{-\lambda/t}\right]_0^y = e^{-\lambda/y}, \quad y > 0, \lambda > 0.$$

The p quantile of $1/X$ is the value y_p that solves $F_Y(y_p) = p$. Thus $y_p = -\lambda/\log p$.

(h) (3 points) The joint density of the data may be written

$$f(y \mid \theta) = f(y_1, \ldots, y_n \mid \theta) = \prod_{j=1}^n \theta^{y_j}(1-\theta)^{1-y_j} = \theta^s(1-\theta)^{n-s},$$

say, where $s = \sum_{j=1}^n y_j$. Then the posterior density of θ is, by Bayes' theorem,

$$\begin{aligned} \pi(\theta \mid y) &= \frac{f(y \mid \theta)\pi(\theta)}{\int f(y \mid \theta)\pi(\theta)\, d\theta}, \\ &= \frac{\theta^s(1-\theta)^{n-s} \times \theta^{a-1}(1-\theta)^{b-1}/B(a,b)}{\int_0^1 \theta^s(1-\theta)^{n-s} \times \theta^{a-1}(1-\theta)^{b-1}/B(a,b)\, d\theta} \\ &\propto \theta^{a+s-1}(1-\theta)^{b+n-s-1} \\ &= \theta^{a+s-1}(1-\theta)^{b+n-s-1}/B(a+s, b+n-s), \quad 0 < \theta < 1, \end{aligned}$$

where the normalising constant follows from the fact that the posterior is a density for any $a, b > 0$. So, the posterior density of θ is a Beta distribution with parameters $a + s$ and $b + n - s$.

(i) (2 points) The marginal density of X is

$$f_X(x) = c\int_{y=0}^1 (1 + 4xy)\, dy = c\left[y + 2xy^2\right]_0^1 = c(1 + 2x), \quad 0 < x < 1,$$

and is otherwise zero. The constant c is determined by the equation

$$1 = \int_{x=0}^1 f_X(x)\, dx = c\left[x + x^2\right]_0^1 = c \times 2,$$

so $c = 1/2$, and therefore $f_X(x) = (1 + 2x)/2 \times I(0 < x < 1)$.

Solution 2. (a) (2 points) The random variable $T = X_1 + X_2$ follows a normal distribution since both X_1 and X_2 are normally distributed and the sum of normally distributed random variables is normally distributed. We have

$$\mathrm{E}(T) = \mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 8 + 16 = 24,$$

and by the independence of X_1 and X_2,

$$\mathrm{var}(T) = \mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) = 9 + 16 = 25.$$

So, $T \sim \mathcal{N}(24, 5^2)$.

(b) (1.5 points) From (a), the random variable $Z = (T - 24)/5$ follows a standard normal distribution. Then, the probability that the total download time exceeds 30 minutes is

$$\Pr(T > 30) = \Pr\left(Z > \frac{30 - 24}{5}\right) = 1 - \Pr(Z \leq 1.2) = 1 - \Phi(1.2) = 1 - 0.88493 \approx 0.115,$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution.

(c) (2 points) The probability that the total download time T exceeds 30 minutes given that $X_1 = 10$ is

$$\begin{aligned}\Pr(T > 30 \mid X_1 = 10) &= \Pr(X_1 + X_2 > 30 \mid X_1 = 10) \\ &= \Pr(X_2 > 20 \mid X_1 = 10) = \Pr(X_2 > 20)\end{aligned}$$

by the independence of X_1 and X_2. The random variable $Z_2 = (X_2 - 16)/4$ follows a standard normal distribution. Thus,

$$\Pr(X_2 > 20) = 1 - \Pr\left(Z_2 \leq \frac{20 - 16}{4}\right) = 1 - \Phi(1) = 1 - 0.84134 \approx 0.152.$$

Alternatively: Similarly to the development in (d), we have $T \mid X_1 = 10 \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ where $\tilde{\mu} = 24 + 9 \times (10 - 8)/9 = 26$ and $\tilde{\sigma}^2 = 25 - 9^2/9 = 4^2$. Thus, $Z_T = (T - \tilde{\mu})/\tilde{\sigma} \mid X_1 = 10 \sim \mathcal{N}(0, 1)$, and therefore,

$$\Pr(T > 30 \mid X_1 = 10) = \Pr\left(Z_T > \frac{30 - 26}{4}\right) = 1 - \Phi(1) = 1 - 0.84134 \approx 0.152.$$

(d) (3.5 points) The random vector $Y = (X_1, T)^{\mathrm{T}} = (X_1, X_1 + X_2)^{\mathrm{T}} = B(X_1, X_2)^{\mathrm{T}}$ is a linear combination of normal variables, so it has a bivariate normal distribution, with mean and covariance matrix

$$\mu = \begin{pmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(T) \end{pmatrix} = \begin{pmatrix} 8 \\ 24 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, T) \\ \mathrm{cov}(X_1, T) & \mathrm{var}(T) \end{pmatrix} = \begin{pmatrix} 9 & 9 \\ 9 & 25 \end{pmatrix},$$

since $\mathrm{cov}(X_1, T) = \mathrm{var}(X_1) = 9$ by the independence of X_1 and X_2. Thus

$$\begin{pmatrix} X_1 \\ T \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} 8 \\ 24 \end{pmatrix}, \begin{pmatrix} 9 & 9 \\ 9 & 25 \end{pmatrix} \right\}.$$

Now $X_1 \mid T = 30 \sim \mathcal{N}(\mu, \sigma^2)$ where the hint in the question gives $\mu = 8 + 9 \times (30 - 24)/25 = 10.16$ and $\sigma^2 = 9 - 9^2/25 = 2.4^2$. Thus, $Z_1 = (X_1 - \mu)/\sigma \mid T = 30 \sim \mathcal{N}(0, 1)$, so

$$\begin{aligned}\Pr(X_1 < 7 \mid T = 30) &= \Pr\left(Z_1 < \frac{7 - 10.16}{2.4}\right) \approx \Phi(-1.32) = 1 - \Phi(1.32) \\ &= 1 - 0.90658 \approx 0.093.\end{aligned}$$

Solution 3. (a) (2.5 points) Let the random variable $X \sim \text{Pois}(\lambda t)$ denote the number of failures over a period of time t, so

$$\Pr(X = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0.$$

The likelihood for λ based on independent observations is therefore

$$L_1(\lambda; t_1, \ldots, t_n, x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda t_i}\frac{(\lambda t_i)^{x_i}}{x_i!}, \quad \lambda > 0,$$

and the log-likelihood is

$$\ell_1(\lambda) = \sum_{i=1}^{n} \left\{-\lambda t_i + x_i \log(\lambda t_i) - \log(x_i!)\right\}, \quad \lambda > 0.$$

Differentiation with respect to λ yields

$$\frac{d}{d\lambda}\ell_1(\lambda) = -\sum_{i=1}^{n} t_i + \frac{1}{\lambda}\sum_{j=1}^{n} x_j, \quad \frac{d^2}{d\lambda^2}\ell_1(\lambda) = -\frac{1}{\lambda^2}\sum_{j=1}^{n} x_j, \quad \lambda > 0.$$

Noting that $\ell_1(\lambda)$ is concave since its second derivative is everywhere negative, unless $\sum x_j = 0$, the maximum likelihood estimate of λ is obtained by solving $d\ell_1(\lambda)/d\lambda = 0$, which yields

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{\sum_{j=1}^{n} t_j}.$$

(b) (2 points) A two-sided equi-tailed confidence interval for λ with approximate level $(1-\alpha)$ is

$$\mathcal{I}_{1-\alpha}^{\hat{\lambda}} = [\hat{\lambda} - J(\hat{\lambda})^{-1/2} z_{1-\alpha/2}; \hat{\lambda} + J(\hat{\lambda})^{-1/2} z_{1-\alpha/2}],$$

where $J(\cdot)$ denotes the observed information, and z_p denotes the p-quantile of the standard normal distribution. We have $z_{0.95} = 1.64$, and the data yield $\sum_{i=1}^{n} x_i = 24$ and $\sum_{i=1}^{n} t_i = 36$. Thus

$$\hat{\lambda} = \frac{24}{36} = 2/3 \text{ days}^{-1}, \quad J(\hat{\lambda}) = -\left.\frac{d^2}{d\lambda^2}\ell_1(\lambda)\right|_{\lambda=\hat{\lambda}} = \left(\frac{24}{36}\right)^{-2} \times 24 = 54,$$

and therefore $\mathcal{I}_{0.9}^{\hat{\lambda}} = [0.44, 0.89]$ days^{-1}. On average we expect 2 failures every 3 days.

(c) (1.5 points) Using the distribution above, the probability of no failure in the interval $(0, t)$ is $\exp(-\lambda t)$, so the probability that the first failure Y takes place before t is $\Pr(Y \leq t) = 1 - \exp(-\lambda t)$, for $t > 0$; this is the exponential distribution, with density $\lambda \exp(-\lambda t)$. Thus, the likelihood and log-likelihood for λ based on the entire dataset are

$$L_2(\lambda; t_1, \ldots, t_n, x_1, \ldots, x_n, y_1, \ldots, y_m) = L_1(\lambda) \times \prod_{j=1}^{m} \lambda e^{-\lambda y_j},$$

and

$$\ell_2(\lambda) = \ell_1(\lambda) + \sum_{j=1}^{m} (\log \lambda - \lambda y_j), \quad \lambda > 0.$$

Differentiation with respect to λ yields

$$\frac{d}{d\lambda}\ell_2(\lambda) = -\sum_{i=1}^{n} t_i + \frac{1}{\lambda}\sum_{j=1}^{n} x_j + \frac{m}{\lambda} - \sum_{j=1}^{m} y_j, \quad \lambda > 0.$$

Noting that $\ell_2(\lambda)$ is concave since

$$\frac{d^2}{d\lambda^2}\ell_2(\lambda) = -\frac{1}{\lambda^2}\left(m + \sum_{j=1}^{n} x_j\right) < 0, \quad \lambda > 0,$$

the maximum likelihood estimate of λ is obtained by solving $d\ell_2(\lambda)/d\lambda = 0$, which yields

$$\hat{\lambda} = \frac{m + \sum_{i=1}^{n} x_i}{\sum_{j=1}^{n} t_j + \sum_{k=1}^{m} y_k}.$$

Solution 4. (a) (4 points)

(i) A boxplot is a graphical summary of a sample of observations. It has the following elements:
- A (central) box which shows the sample lower and upper quartiles ($\hat{q}(0.25)$ and $\hat{q}(0.75)$).
- A line inside the box which shows the sample median ($\hat{q}(0.5)$).
- The whiskers which show the most extreme observations lying inside the numbers $\hat{q}(0.25) - C$ and $\hat{q}(0.75) + C$, where $C = 1.5 \times \{\hat{q}(0.75) - \hat{q}(0.25)\}$. Observations that are more extreme than the whiskers are shown individually.

(ii) The left panel shows that at least 75% of observations in each sample are below 15ppb (the upper part of boxes are below the horizontal line). Also, the spreads of samples 2 and 3 might be slightly larger than that for sample 1.

In addition, (log) lead levels' tend to decrease when one increases the duration where tap water is let run, with the exception of the 3 largest observations in sample 2. This trend is at the distribution level; one cannot draw conclusions at the house level. Indeed, (log) lead levels could increase for houses with low lead levels and decrease (drastically) for houses with high lead levels.

(b) (3 points)

(i) The calculation corresponds to $\Pr(X \geq 45)$ for a random variable $X \sim \mathrm{B}(n = 271, p = 0.1)$. In the context of Flint's tap water, this is the probability of observing 45 samples or more among the 271 for which the level of lead exceeds 15ppb, assuming that lead levels are independent between houses and have a probability of 0.1 to exceed 15ppb.

(ii) The proportion of elements of sample 1 that exceed 15ppb is $45/271 = 16.61\%$. However, the fact that the observed proportion exceeds the limit of 10% could be due to chance, even if the true proportion is less than 10%.

To determine if action should be taken, one can perform a hypothesis test on the proportion of samples that exceed 15ppb. Consider the null hypothesis $H_0 : p = 0.1$ and the alternative hypothesis $H_1 : p > 0.1$. The given calculation provides the p-value of the test. Since the p-value is equal to 0.0003, either the null hypothesis holds but a very unlikely event has occurred, or the null hypothesis is untrue and action should be taken. Since the signficance probability is so small, there is strong evidence against the null hypothesis, and we conclude that action should be taken.

(c) (1 point) This is a normal QQplot, and a perfect normal sample would fall on a straight line. The data seem to follow a slightly convex shape, at least for the lower portion, so the data seem to be slightly skewed to the right relative to a normal sample. However there is no evidence of outliers or heavy tails.

(d) (2 points) Let μ denote the mean log level of lead. From (c) we assume that the average of the log lead levels should have close to a normal distribution, using the central limit theorem. The variance is unknown, so we would usually base a confidence interval on the t statistic, but in this case the sample is so large that we can take the variance as known. A two-sided confidence interval for μ is

$$J_{1-\alpha} = \left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, and z_p is the p-quantile of a standard normal distribution. The confidence interval with $\alpha = 0.05$, $\bar{x} = 1.4$, $s^2 = 1.68$, and $z_{0.975} = 1.96$ is

$$J_{0.95} = [1.2457; 1.5543].$$

(e) (2 points) The interval in (d) is based on the assumption that the central limit theorem is applicable to the average of the log lead levels, and that they are independent.

The normality of the average seems reasonable from (c).

The independence of the samples seems plausible, if the houses were chosen at random. The level of non-response is fairly low (29/300), so it probably does not affect any conclusions drawn.