Solutions to the extraordinary exam for Probability
Master in Statistics for Data Science
25 June, 2019

P1. (1.5 points) One hundred people line up to board an airplane with one hundred seats. Each has a boarding pass with assigned seat. However, the first person to board has lost his boarding pass and takes a random seat. After that, each person takes the assigned seat if it is unoccupied, and one of unoccupied seats at random otherwise.

a) (0.25 points) What is the probability that the first person to board gets to sit in his assigned seat?

b) (0.5 points) What is the probability that the second person to board gets to sit in his assigned seat?

c) (0.25 points) What is the probability that the third person to board gets to sit in his assigned seat?

d) (0.5 points) What is the probability that the last person to board gets to sit in his assigned seat?

Solution. Denote by $S_{i,j}$ the event that the i-th passenger takes the seat assigned to the j-th one.

a) Since the first person to board takes a random seat and there are one hundred seats, $P(S_{1,1}) = 1/100$.

b) The second passenger to board will take the seat assigned to him as long as it is unoccupied when he boards it, which will occur as long as the first passenger does not take it, $P(S_{2,2}) = 1 - P(S_{1,2}) = 1 - 1/100 = 99/100$.

c) Observe that if the first passenger takes the seat of the third one, then the second passenger cannot take it, so the events $S_{1,3}$ and $S_{2,3}$ are mutually exclusive. Further, the only way that the second passenger takes the sea of the third one is that his own seat is occupied when he enters the plane, which only occurs if the first passenger has previously taken it.

$$\begin{aligned} P(S_{3,3}) &= 1 - P(S_{1,3} \cup S_{2,3}) \\ &= 1 - P(S_{1,3}) - P(S_{2,3}) \\ &= 1 - 1/100 - P(S_{1,2} \cap S_{2,3}) \\ &= 1 - 1/100 - P(S_{2,3}|S_{1,2})P(S_{1,2}) \\ &= 1 - 1/100 - 1/99 \times 1/100 = 98/99. \end{aligned}$$

d) Notice that the 100-th passenger can only either take the seat of the 1st passenger or his own seat. In case any other sea was empty at any stage, the corresponding passenger would have taken it, and it cannot be empty after such passenger has entered the airplane. Further those two seats will be empty with the same probability (the only way at which they might have been occupied is completely at random), so $P(S_{100,100}) = P(S_{100,1}) = 1/2$ and $P(S_{100,i}) = 0$ for $i = 2, 3, \ldots, 99$.

P2. (2 points) The cdf of the shifted (or two-parameter) exponential distribution with origin parameter $\theta \in \mathbb{R}$ (sometimes restricted to $\theta \geq 0$) and scale parameter $\lambda > 0$ is given by

$$F(x) = 1 - e^{-(x-\theta)/\lambda}, \quad x > \theta,$$

a) (0.25 points) Show that if X is an exponential random variable with parameter $\lambda > 0$ and $\theta \in \mathbb{R}$, then $X + \theta$ is a shifted exponential random variable with parameters θ and λ.

b) (0.25 points) Determine the density mass function of a shifted exponential random variable.

c) (0.25+0.25 points) Determine the mean and variance of a shifted exponential random variable.

d) (0.5 points) The shifted exponential distribution is commonly used in reliability studies to model lifetime data. In this setting, the origin parameter θ can be interpreted as the length of the guarantee period. If the lifetime of a device (in years) is modelled as a shifted exponential random variable with parameters $\theta = 1$ and $\lambda = 0.5$, what lifetime is exceeded by 75% of the devices?

e) (0.5 points) Write a piece of code to simulate 1000 observations of a shifted exponential random variable with origin parameter $\theta = 1$ and scale parameter $\lambda = 0.5$. Use your simulations to approximate the answer to part d).

Solution. a) Consider X an exponential random variable with parameter λ. We know that its cdf is $F_X(x) = 1 - e^{-x/\lambda}$ if $x > 0$ and $F_X(x) = 0$ otherwise. The cdf of $Y = X + \theta$ evaluated on x is

$$F_Y(x) = P(Y \leq x) = P(X + \theta \leq x) = P(X \leq x - \theta) = F_X(x - \theta) = 1 - e^{-(x-\theta)/\lambda}$$

if $x - \theta > 0$, so $x > \theta$, while $F_Y(x) = 0$ otherwise.

b) Differentiating the cdf, we obtain that the density mass function is $f_Y(x) = F_Y'(x) = \lambda^{-1} e^{-(x-\theta)/\lambda}$ if $x > \theta$, while $f_Y(x) = 0$ otherwise.

c) $\mathbb{E}[Y] = \mathbb{E}[X + \theta] = \mathbb{E}[X] + \theta = \lambda^{-1} + \theta$ and $\mathrm{Var}[Y] = \mathrm{Var}[X + \theta] = \mathrm{Var}[X] = \lambda^{-2}$.

d) Take $\theta = 1$ and $\lambda = 0.5$. We must determine t such that $P(Y > t) = 0.75$, so $P(X + \theta > t) = 0.75$, and $P(X < t - \theta) = 0.25$, so $t - \theta = t - 1$ is the 0.25-quantile of an exponential random variable with parameter $\lambda = 0.5$, and $t =$ `1+qexp(0.25,rate=0.5)` $= 1.575364$.

e)
```
set.seed(123)
sim.se<-rexp(1000,rate=0.5)+1
quantile(sim.se,0.25)
```
and the approximate answer to part d) is `1.613385`.

P3. (3.25 points) The number of items bought by each client visiting a small shop follows a Poisson distribution with mean 3.5 items.

a) (0.25 points) What is the probability of having a client that buys more than 3 items?

b) (0.25 points) What is the probability that, out of all clients visiting the shop in a given day, the first one that buys more than 3 items is the fourth one?

c) (0.25 points) What is the probability that at least 2 of the last 10 clients buy more than 3 items each?

d) (0.5 points) Approximate the probability that the total number of items bought by the last 10 clients is greater than 30.

e) (0.5 points) How many visits are needed so that the probability of selling more than 100 items is at least 0.95?

f) (0.5 points) What is the probability that at least 20 of the last 100 clients visiting the shop bought more than 3 items each?

g) (0.5 points) What is the probability that in a group of 10 clients taken at random, there are exactly 2 clients that bought less than two items and 4 clients that bought two or three items.

h) (0.5 points) A group of 100 clients visiting the shop has been selected in order to study all available data about them. A total number of 42 of those 100 clients bought more than 3 items. If a set of 10 visits is taken at random from the group of 100, what is the probability that more than 5 of them are of clients that bought more than 3 items.

Solution. a) Denote by X the number of items bought by 1 client, $X \sim \mathcal{P}(\lambda = 3.5)$, $P(X > 3) = 1 - P(X \leq 3) =$ `1-ppois(3,lambda=3.5)` $= 0.4633673$.

b) Denote by Y the number of clients until one buys more than

c) Denote by N the number of clients that buy more than 3 items in a group of 10 clients. Random variable N clearly follows the binomial distribution $N \sim \mathrm{B}(n = 10, p = 0.4633673)$, and $P(N \geq 2) = 1 - P(N \leq 1) =$ `1-pbinom(1,size=10,prob=0.4633673)` $= 0.9809186$.

d) Denote by M the number of items bought by 10 clients, then $M \sim \mathcal{P}(\lambda = 35)$ and $P(M > 30) = 1 - P(M \leq 30) =$ `1-ppois(30,lambda=35)` $= 0.7730576$.

e) Denote by H_n the number of items sold in n visits, then $H \sim \mathcal{P}(\lambda = 3.5n)$ whose distribution can be approximated by means of a normal one as $H \approx \mathrm{N}(\mu = 3.5n, \sigma = \sqrt{3.5n})$. We must determine n such that $P(H_n > 100) = P(H_n > 100.5) = 0.95$, so $P((H - 3.5n)\sqrt{3.5n} > (100.5 - 3.5n)/\sqrt{3.5n}) = 0.95$. In conclusion $(100.5 - 3.5n)/sqrt3.5n =$ `qnorm(0.05)` $= -1.645$. We obtain an equation of degree 2 on $\sqrt{n}$, and conclude $n = 33.8$, so the number of needed visits is $n = 34$.

f) Denote by L the number of clients, out of 100, buying more than 3 items each, then $L \sim \mathrm{B}(n = 100, p = 0.46337)$, and $P(X \geq 20) = 1 - P(X \leq 19) =$ `1-pbinom(19,size=100,prob=0.46337)` $= 1$.

g) Denote by J_1 the number of clients, out of 10, that bought 0 or 1 item, by J_2 the number of clients, out of 10, that bought 2 or 3 items, and J_3 the number of clients, out of 10, that bought more than 3 items, then the random vector (J_1, J_2, J_3) follows a multinomial distribution with parameters $(J_1, J_2, J_3) \sim \mathrm{M}(n = 10, \mathbf{p} = (0.13589, 0.40074, 0.46337))$, where 0.13589, 0.40074, and 0.46337 are respectively the probabilities of 0 or 1 item, 2 or 3 items, and more than 3 items and were obtained from a $\mathcal{P}(\lambda = 3.5)$ distribution model. Finally $P(J_1 = 2, J_2 = 4) =$ `dmultinom(c(2,4,4),size=10,prob=c(0.13589,0.40074,0.46337))` $= 0.069159$.

h) Denote by K the number of clients in the set of 10 selected visits that bought more than 3 items. Recall that 42 clients out of the group of 100 clients bought more than 3 items, so K follows a hypergeometric distribution with parameters $K \sim \mathrm{H}(M = 42, N = 58, k = 10)$, and $P(K > 5) =$ `1-phyper(5,m=42,n=58,k=10)` $= 0.18951$.

P4. (2.25 points) A statistics class takes two exams X (Exam 1) and Y (Exam 2) where the scores follow a bivariate normal distribution with parameters:

$$\mu_X = 70 \text{ and } \mu_Y = 60 \text{ are the marginal means,}$$

$$\sigma_X = 10 \text{ and } \sigma_Y = 15 \text{ are the marginal standard deviations,}$$

$$\rho = 0.6 \text{ is the correlation coefficient.}$$

Suppose we select a student at random. What is the probability that...

a) (0.25 points) ...the student scores over 75 on Exam 2?

b) (0.5 points) ... the student scores over 75 on Exam 2, given that the student scored $x = 80$ on Exam 1?

c) (0.5 points) ... the sum of his/her Exam 1 and Exam 2 scores is over 150?

d) (0.5 points) ... the student did better on Exam 1 than Exam 2?

e) (0.5 points) Assume the final grade is computed as $0.4X + 0.6Y$. What final grade is exceeded by 65% of the students?

Solution. a) $P(Y > 75) =$ `1-pnorm(75,mean=60,sd=15)` $= 0.1586553$.

b) The distribution of $Y|_{X=80}$ is $Y|_{X=80} \sim \mathrm{N}(\mu = 60 + (15/10)0.6(80-70) = 69, \sigma = 15\sqrt{1-0.6^2} = 12)$, so $P(Y > 75|X = 80) =$ `1-pnorm(75,mean=69,sd=12)` $= 0.30854$.

c) The distribution of $X+Y$ is $X+Y \sim \mathrm{N}(\mu = 70+60 = 130, \sigma = \sqrt{10^2 + 15^2 + 2 \times 0.6 \times 10 \times 15} = 22.47)$, so $P(X + Y > 150) =$ `1-pnorm(150,mean=130,sd=22.47)` $= 0.18671$.

d) The distribution of $X-Y$ is $X-Y \sim \mathrm{N}(\mu = 70-60 = 10, \sigma = \sqrt{10^2 + 15^2 - 2 \times 0.6 \times 10 \times 15} = 12.04)$, so $P(X - Y > 0) =$ `1-pnorm(0,mean=10,sd=12.04)` $= 0.79689$.

e) The distribution of $0.4X + 0.6Y$ is $0.4X + 0.6Y \sim \mathrm{N}(\mu = 0.4 \times 70 + 0.6 \times 60 = 64, \sigma = \sqrt{0.4^2 \times 10^2 + 0.6^2 \times 15^2 + 2 \times 0.6 \times 4 \times 9} = 11.84)$, and que must compute t such that $P(04X + 0.6Y > t) = 0.65$, so $t =$ `qnorm(0.35,mean=64,sd=11.84)` $= 59.43781$.

P5. (1 point) During the Middle Ages, coins struck by the Royal Mint in England were evaluated for their metal content on a sample basis, in a ceremony called the Trial of the Pyx. One hundred gold coins were chosen at random from all of the coins made at the Mint every given year, put in the Pyx (a ceremonial box), and weighed. Each of these gold coins was supposed to weigh 128 grains, so the 100 coins in the Pyx should have weighed about 12800 grains. A margin of error of 32 grains was allowed at the trial, but if the actual weight of the coins in the Pyx was less than $12800 - 32 = 12768$ grains, the Master of the Mint was exposed to serious penalties. Assume that the Master of the Mint is honest and manufactures gold coins with a mean weight of 128 grains and standard deviation of 1 grain.

a) (0.5 points) What are the mean and variance of the total weight of the 100 coins?, and the (approximate) distribution of the total weight of the 100 coins?

b) (0.25 points) What is the chance that the Master of Mint survives the Trial of the Pyx?

c) (0.25 points) What margin of error should be fixed so that the probability that a honest Master of Mint survives the Trial of the Pyx is 0.99?

Solution. a) Denote by X_i the weight (in grains) of the i-th coin, then $\mathbb{E}[X_i] = 128$ and $\mathrm{Var}[X_i] = 1$, while the total weight of the 100 coins, denoted by T, is $T = \sum_{i=1}^{100} X_i$. We have

- $\mathbb{E}[T] = \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathbb{E}[X_i] = 100 \times 128 = 12800$
- $\mathrm{Var}[T] = \mathrm{Var}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathrm{Var}[X_i] = 100 \times 1 = 100$, where the fact that the variance of the sum of a set of random variables is the sum of variances follows from their independence.

Further, by the Central Limit Theorem, T is approximately normally distributed, $T \approx \mathrm{N}(\mu = 12800, \sigma = \sqrt{100} = 10)$.

b) $P(T > 12768) =$ `1-pnorm(12768,mean=12800,sd=10)` $= 0.9993129$.

c) In order to determine t such that $P(T > t) = 0.99$, we compute $t =$ `qnorm(0.01,mean=12800,sd=10)` $= 12776.74$. So the margin of error is $12800 - 12776.74 = 23.26$ grains.