Statistical Methods for Data Science

HW 1 Solution

$*$ The solution often refers to distribution tables for computing probabilities. They can be computed using `R`. See the accompanying `R` code for this.

Exercise 2.14

(a) The total number of possible password is

$$P(26,6) = 26 \times 25 \times 24 \times 23 \times 22 \times 21 = \frac{26!}{(26-6)!} = 165,765,600$$

because there are 26 letters in the alphabet, they should be all different in the password, and the order of characters is important. The password is guessed (favorable outcome) if it is among the 1000000 attempted passwords. Then

$$P(guess\ the\ password) = \frac{\#\ of\ favorable\ passwords}{total\ \#\ of\ passwords} = \frac{1,000,000}{165,765,600} = 0.0060$$

(b) Now we can use 52 characters ($b/c\ 2(each\ letter\ has\ upper\ and\ lower\ case) \times 26(26\ letters) = 52$) and the order is still important. Then the total number of passwords is

$$P(52,6) = 52 \times 51 \times 50 \times 49 \times 48 \times 47 = \frac{52!}{(52-6)!} = 14,658,134,400$$

and

$$P(guess\ the\ password) = \frac{1,000,000}{14,658,134,400} = 0.000068$$

(c) Letters can be repeated in passwords, therefore, the total number of passwords is

$$P(52,6) = \underset{\text{each character has 52 choices}}{\underset{\uparrow}{52^6}}$$

and

$$P(guess\ the\ password) = \frac{10^6}{52^6} = 0.000051$$

(d) Adding the digits bring the number of possible characters to 62. ($because$ $52(52\ letters) +$ $10\ (digits\ from\ 0,1,...,9) = 62$). Then the total number of password is

$$P(62,6) = 62^6,$$

and

$$P(guess\ the\ password) = \frac{10^6}{62^6} = 0.000018$$

The more characters we use the lower is the probability for a spy ware to break into the system.

Exercise 3.7 Let X_i be the number of home runs in game i for $i = 1, 2$. Compute

$$\begin{aligned}
E(X) &= \sum_x xp(x) = 0 \times 0.4 + 1 \times 0.4 + 2 \times 0.2 = 0.8; \\
E(X^2) &= \sum_x x^2 p(x) = 0^2 \times 0.4 + 1^2 \times 0.4 + 2^2 \times 0.2 = 1.2; \\
Var(X) &= E(X^2) - E^2(X) = 1.2 - 0.8^2 = 0.56.
\end{aligned}$$

Then use the fact that $Y = X_1 + X_2$, where X_1 and X_2 are independent.

$$\begin{aligned}
E(Y) &= E(X_1 + X_2) = EX_1 + EX_2 = 0.8 + 0.8 = 1.6; \\
Var(Y) &= Var(X_1 + X_2) = VarX_1 + VarX_2 = 0.56 + 0.56 = 1.12.
\end{aligned}$$

Exercise 3.21 Let X be the # of computers entered by the virus. Then X Binomial $(n = 20, p = 0.4)$.
Because each of the 20 computers is either entered or not, X is the # of "successes" in $n = 20$ Bernoulli trials.

$$\begin{aligned}
P(X \geq 10) &= 1 - \underbrace{P(X \leq 9)}_{or\ from\ Table\ A_2,\ we\ get\ P(X \leq 9) = 0.7553} \\
&= 1 - \sum_{x=0}^{9} \underbrace{\binom{20}{x} 0.4^x 0.6^{20-x}}_{P(X=x), x=0,1,\ldots,9} \\
&= 1 - 0.7553 \\
&= 0.2447
\end{aligned}$$

Exercise 3.37 We need $P(X > 4)$, where X = # of breakdowns during 21 weeks. This is the number of rare events, averaging 1 per 3 weeks, or 7 per 21 weeks. Thus, X is Poisson with $\lambda = 7$, from Table A_3, $P(X > 4) = 1 - F(4) = 1 - 0.173 = 0.827$.

Exercise 4.4

(a) Find K from the condition $\int f(x)\, dx = 1$:

$$\int f(x)\, dx = \int_0^{10} (K - \frac{x}{50})\, dx = Kx - \frac{x^2}{2 \cdot 50}\Big|_0^{10} = 10K - 1 = 1 \qquad \Rightarrow K = 0.2$$

(b)

$$P(X<5)=\int_0^5 (0.2-\frac{x}{50})\,dx=(0.2x-\frac{x^2}{2\cdot 50})\Big|_0^5=1-0.25=0.75$$

(c)

$$E(X)=\int xf(x)\,dx=\int_0^{10} x(0.2-\frac{x}{50})\,dx=\left(\frac{0.2x^2}{2}-\frac{x^3}{3\cdot 50}\right)\Big|_0^{10}=10-\frac{20}{3}=3\frac{1}{3}\ \mathit{or}\ 3.333\ \mathit{years}.$$

Exercise 4.6 Denote Exponential (λ) times for the 3 blocks by X_1, X_2 and X_3, and let $X=\max_i X_i$ be the time it takes to compile the whole program. Find the cdf, then the pdf of X and use the latter to compute the expectation $E(X)$.
For an exponential (λ) time X_i,

$$E(X_i)=\frac{1}{\lambda}=5\ min \qquad \Rightarrow \lambda=0.2\ min^{-1}$$
$$F(X_i)=1-e^{-0.2x}\ (x>0)$$

Now, we compute the cdf of X,

$$F_X(x)=\overbrace{P\{\max_i X_i\le x\}=P\{X_1\le x, X_2\le x, X_3\le x\}}^{\textit{they are the same event}}\ \underset{\substack{\uparrow\\ X_1, X_2,\text{ and }X_3\text{ are independent}}}{=}\prod_{i=1}^3 P(X_i=x)=(1-e^{-0.2x})^3, x>0$$

Differentiate it to find the pdf,

$$f_X(x)=F_X^{'}(x)=0.6(1-e^{-0.2x})^2e^{-0.2x}, x>0$$

Now we can compute $E(X)$ as

$$E(X)=\int xf_X(x)\,dx=0.6\int_0^\infty x(1-e^{-0.2x})^2e^{-0.2x}\,dx$$
$$=0.6\int_0^\infty (xe^{-0.2x}-2xe^{-0.4x}+xe^{-0.6x})\,dx$$

We take the three integrals by parts

$$\int_0^\infty xe^{-0.2x}\,dx=xe^{-0.2x}\Big|_0^\infty+5\int_0^\infty e^{-0.2x}\,dx$$
$$=xe^{-0.2x}\Big|_0^\infty+5(-\frac{1}{0.2}e^{-0.2x}\Big|_0^\infty)$$
$$=0+5\cdot(-5)\cdot(0-1)=25$$

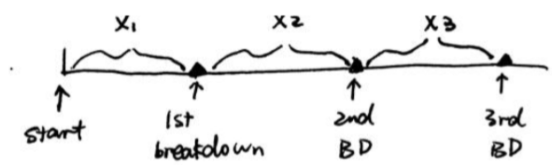
and we get

$$2\int_0^\infty xe^{-0.4x}\,dx = 2 \times \frac{25}{4} = \frac{25}{2}$$
$$\int_0^\infty xe^{-0.6x}\,dx = \frac{25}{9}$$

Then $E(X) = 0.6(25 - \frac{25}{2} + \frac{25}{9}) = 9.167$ min.
Or you can also use the gamma-function ($\Gamma(2) = 1! = 1$) to get these three integrals and have:

$$E(X) = 0.6(\frac{\Gamma(2)}{0.2^2} - \frac{2\Gamma(2)}{0.4^2} + \frac{\Gamma(2)}{0.6^2}) = 15 - 7.5 + 1\frac{2}{3} = \frac{55}{6} \; or \; 9.17 \; min$$

Exercise 4.9



$X_i \sim$ Exponential ($\frac{1}{5}$), let $T = X_1 + X_2 + X_3 \sim$ Gamma ($3, \frac{1}{5}$). Because Exponential (λ) is a special case of Gamma (α, λ) when $\alpha = 1$, and if $X_i \sim$ Gamma $(\alpha_i, \lambda), i = 1, ..., n$, then $\sum_{i=1}^n X_i \sim$ Gamma ($\sum_{i=1}^n \alpha_i, \lambda$).
Here we have $X_i \sim$ Gamma ($1, \frac{1}{5}$), then $T = \sum_{i=1}^3 X_i \sim$ Gamma ($3, \frac{1}{5}$).

(a) By the Gamma-Poisson formula with a Poisson ($\lambda t = \frac{1}{5} \cdot 9 = 1.8$) variable X and Table A_3.

$$P(T \leq 9) = P(X \geq 3) = 1 - F_X(2) = 1 - 0.731 = 0.269$$

$\diamondsuit$ `R` can be used to compute a gamma probability directly, i.e., without converting it into a Poisson probability. See the accompanying `R` code.

(b)

$$\begin{aligned} P(T > 16|T > 12) &= \frac{P(T > 16 \cap T > 12)}{P(T > 12)} = \frac{P(T > 16)}{P(T > 12)} \\ &= \frac{P(T_1 < 3)}{P(T_2 < 3)} = \frac{e^{-3.2}(1 + 3.2 + 3.2^2/2)}{0.570} = 0.666 \end{aligned}$$

by the Gamma-Poisson formula, the formula of Poisson pmf and Table A_3, where T_1 has Poisson distribution with parameter $\frac{1}{5} \cdot 16 = 3.2$ and T_2 has Poisson distribution with parameter $\frac{1}{5} \cdot 12 = 2.4$

Exercise 4.23 Apply the Central Limit Theorem. A continuity correction is not needed because the lifetime is a continuous random variable.

$$\begin{aligned} P(\frac{S_{400}}{400} < 5012) =& P(S_{400} < 5012 \times 400) = P(\frac{S_{400} - 400\mu}{\sigma\sqrt{400}} < \frac{5012 \times 400 - 400\mu}{\sigma\sqrt{400}}) \\ &= P(Z < \frac{5012 \times 400 - 400\mu}{\sigma\sqrt{400}}) \approx \Phi(\frac{5012 - 5000}{100/\sqrt{400}}) = \Phi(2.4) = 0.9918(from\ Table\ A_4) \end{aligned}$$

Exercise 4.29 Given $P(PrinterI) = 0.4, P(PrinterII) = 0.6$. For $PrinterI$ with exponential time, $E(X) = 2 = \frac{1}{\lambda}$, hence $\lambda = 0.5$ and $P\{T < 1|PrinterI\} = 1 - e^{-0.5 \cdot 1} = 0.393$.
For $PrinterII$ with uniform time, the density of T is $f(t) = \frac{1}{5}$ for t between 0 and 5, and

$$P(T < 1|PrinterII) = \int_0^1 \frac{1}{5}\,dt = 0.2$$

By the Bayes Rule,

$$\begin{aligned} P(PrinterI|T < 1) &= \frac{P(T < 1|PrinterI)P(PrinterI)}{P(T < 1|PrinterI)P(PrinterI) + P(T < 1|PrinterII)P(PrinterII)} \\ &= \frac{0.393 \times 0.4}{0.393 \times 0.4 + 0.2 \times 0.6} = 0.567 \end{aligned}$$

♢ You can also use *pexp* function in `R` to compute exponential probabilities. Try it.

Exercise 4.31

(a) We have $n = 68, \mu = 15\ sec$, and $\sigma = \sqrt{11}\ sec$.
By the Central Limit Theorem,

$$\begin{aligned} P(S_{68} < 720\ sec) &= P(\frac{S_{68} - n\mu}{\sigma\sqrt{n}} < \frac{720 - n\mu}{\sigma\sqrt{n}}) \\ &= P(Z < \frac{720 - 68 \times 15}{\sqrt{11}\sqrt{68}}) \\ &\approx \Phi(-10.97) = 0.00(practically\ 0) \\ &(see\ the\ last\ line\ of\ Table\ A_4) \end{aligned}$$

(b) We are given that

$$P(S_N < 600\ sec) = 0.95$$

That is,

$$
\begin{aligned}
0.95 &= P\left(\frac{S_N - N\mu}{\sigma\sqrt{N}} < \frac{600 - N\mu}{\sigma\sqrt{N}}\right) = P\left(Z < \frac{600 - N\mu}{\sigma\sqrt{N}}\right) \\
&= \Phi\left(\frac{600 - N\mu}{\sigma\sqrt{N}}\right)
\end{aligned}
$$

On the other hand,

$0.95 = \Phi(1.645) \qquad (from\ Table\ A_4,\ because\ \Phi(1.64) = 0.9495, \quad \Phi(1.65) = 0.9505)$

Therefore,

$$\frac{600 - N\mu}{\sigma\sqrt{N}} = \frac{600 - 15N}{\sqrt{11N}} = 1.645.$$

It remains to solve this equation for N :

$$
\begin{aligned}
&(600 - 15N)^2 = 1.645^2(11N) \\
&360000 - 18000N + 225N^2 = 30N, \\
&225N^2 - 18030N + 360000 = 0, \\
&N = \frac{18030 \pm \sqrt{18030^2 - 4 \times 225 \times 360000}}{2 \times 225} = 40 \pm 2 = 38\ or\ 42
\end{aligned}
$$

Notice that $600 - 15N$ is positive, so, $N < 40$.
Thus the new version of the package requires 38 new files.

```
#########################################################
# R code  #
#########################################################

########
# 3.21 #
########
#Probability of X >= 10, equivalent to calculating the probability of X > 9

pbinom(q = 10 - 1, size = 20, prob = 0.4, lower.tail = FALSE)

# > pbinom(q=10-1, size=20, prob=0.4, lower.tail = FALSE)
# [1] 0.2446628
# >
```

```
# Alternatively, we can compute P(X >= 10) = 1 - P(X < 10) = 1 - P(X <= 9)

# > 1 - pbinom(q = 9, size = 20, prob = 0.4) # lower.tail = TRUE by default
# [1] 0.2446628
# >

########
# 3.37 #
########

#Probability of X > 4

ppois(q = 4, lambda = 7 * 1, lower.tail = FALSE)

# > ppois(q = 4, lambda = 7 * 1, lower.tail = FALSE)
# [1] 0.8270084
# >

# As before, we can compute P(X > 4) = 1 - P(X <= 4)

# > 1 - ppois(q = 4, lambda = 7 * 1)
# [1] 0.8270084
# >

#######
# 4.9 #
#######

# Approach 1: Computing Poisson probabilities

#(a)#
#Probability of X >= 3,equivalent to calculating the probability of X > 2

ppois(q = 3 - 1, lambda = 9/5, lower.tail = FALSE)

# > ppois(q = 3 - 1, lambda = 9/5, lower.tail = FALSE)
# [1] 0.2693789
# >

# As before, P(X >= 3) = 1 - P(X < 3) = 1 - P(X <= 2)
```

```
# > 1 - ppois(q = 3 - 1, lambda = 9/5)
# [1] 0.2693789
# >

#(b)#

#Probability of T1 < 3,equivalent to calculating the probability of T1 <= 2

ppois(q = 3 - 1, lambda = 16/5)

# > ppois(q = 3 - 1, lambda = 16/5)
# [1] 0.3799037
# >

#Probability of T2 < 3,equivalent to calculating the probability of T2 <= 2

ppois(q = 3 - 1, lambda = 12/5)

# > ppois(q = 3 - 1, lambda = 12/5)
# [1] 0.5697087
# >

# Ratio of the above two probabilities

ppois(q = 3 - 1, lambda = 16/5)/ppois(q = 3 - 1, lambda = 12/5)

# > ppois(q = 3 - 1, lambda = 16/5)/ppois(q = 3 - 1, lambda = 12/5)
# [1] 0.6668385
# >


# Approach 2: Directly computing gamma probabilities, i.e., without
# coverting them into Poisson probabilities

#(a)#

#P(T <= 9), where T follows Gamma (3, 1/5) distribution

pgamma(q = 9, shape = 3, rate = 1/5)

# > pgamma(q = 9, shape = 3, rate = 1/5)
```

```
# [1] 0.2693789
# >

#(b)#

#P(T > 16)/P(T > 12), where T follows Gamma (3, 1/5) distribution

(1 - pgamma(q = 16, shape = 3, rate = 1/5))/(1 - pgamma(q = 12, shape = 3, rate = 1/5)

# > (1 - pgamma(q = 16, shape = 3, rate = 1/5))/(1 - pgamma(q = 12, shape = 3, rate =
# [1] 0.6668385
# >



########
# 4.23 #
########

pnorm(q = 5012, mean = 5000, sd = 100/sqrt(400))

# > pnorm(q = 5012, mean = 5000, sd = 100/sqrt(400))
# [1] 0.9918025
# >


########
# 4.31 #
########
# (a) #

pnorm(q = 720, mean = 68 * 15, sd = sqrt(11) * sqrt(68))

# > pnorm(q = 720, mean = 68 * 15, sd = sqrt(11) * sqrt(68))
# [1] 2.69067e-28
# >

# (b) #

#95-th percentile of standard normal distribution
```

```
qnorm(p = 0.95, mean = 0, sd = 1)

# > qnorm(p = 0.95, mean = 0, sd = 1)
# [1] 1.644854
# >

########## END #################
```