Probability and Statistics for Data Analysis

Assignment 1: Probability

Chalkiopoulos Georgios | $p3352124$

7 November 2021

Exercise 1. *The number of incoming calls within one minute to a call center follows the Poisson distribution with parameter $\lambda = 5$. What is the probability of the event:*

1. *No calls arrive within one minute?*

$$P(X = 0) = \frac{e^{-5} \cdot 5^0}{0!} = 6.74e^{-3} \tag{1.1}$$

2. *10 calls arrive within one minute?*

$$P(X = 10) = \frac{e^{-5} \cdot 5^{10}}{10!} = 1.81e^{-2}$$

3. *At least 10 calls arrive within one minute?*

$$P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{i=0}^{10} P(i) = 1 - 0.968 = 3.18e^{-2} \tag{1.2}$$

4. *10 calls arrive within 2 minutes?*

Given that the incoming calls are independent we can model the number of calls within two minutes as a Poisson distribution with $\lambda = 10$. Therefore:

$$P(X = 10) = \frac{e^{-10} \cdot 10^{10}}{10!} = 0.13$$

5. *At least 10 calls arrive within 1 minute, conditional on the event that at least one call arrives (within one minute)?*

We want to calculate the probability: $P(X \geq 10|X = 1)$ for $\lambda = 5$. More specifically:

$$P(X \geq 10|X = 1) = \frac{P(X \geq 10, X = 1)}{P(X = 1)} = \frac{P(X \geq 10)}{P(X = 1)} \tag{1.3}$$

The numerator has already been calculated in (1.2).
For the denominator we calculate:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(0) \overset{(1.1)}{=} 1 - 0.00674 = 0.99 \tag{1.4}$$

Using (1.3), (1.4) we get: $P(X \geq 10|X = 1) = \frac{P(X \geq 10)}{P(X = 1)} = \frac{0.0318}{0.99} = 3.2e^{-2}$

Exercise 2. The duration (in minutes) of a call is a random variable (X) with probability density function

$$f_X(x) = xe^{-x} I_{(0,\infty)}(x) \tag{2.1}$$

The cost (in euro) $c(x)$ of a call with duration x is equal to

$$c(x) = \begin{cases} 2, & 0 < x \leq 3 \\ 2 + 6(x-3), & x > 3. \end{cases} \tag{2.2}$$

Hint: you may find useful to identify the distribution in Equation (2.1), although it is not necessary.

Looking at (2.1) we observe that $f_X(x)$ is an Gamma Distribution with $\alpha = 2$ and $\beta = 1$, thus $X \sim \mathcal{G}(\alpha = 2, \beta = 1)$.

1. *What is the probability that a call lasts at most one minute?*

(Solution a) Analytical approach: By definition,

$$P(X \leq 1) = \int_0^1 f_X(x)dx = \int_0^1 xe^{-x}dx = \left[-xe^{-x}\right]_0^1 + \int_0^1 e^{-x}dx = e - e + 1 = 0.264$$

(Solution b) use R: In order to calculate the probability $P(X \leq 1)$ we can use the following commands in R

```
> a <- 2 ### parameter alpha
> b <- 1 ### parameter beta
> pgamma(1, a, scale=b)
[1] 0.2642411
```

2. *What is the probability that a call lasts at least two minutes?*

(Solution a) Analytical approach: Similar to the previous question:

$$P(X \geq 2) = \int_2^\infty f_X(x)dx = 0.406$$

(Solution b) use R: In this case we use the following commands:

```
> a <- 2 ### parameter alpha
> b <- 1 ### parameter beta
> pgamma(1, a, scale=b, lower.trail = FALSE)
[1] 0.4060058
```

3. *Calculate the mean and variance of call duration.*

Having identified that the Random Variable follows a Gamma Distribution we can calculate the Mean and Variance directly: $E[X] = \alpha\beta$ and $Var[X] = \alpha\beta^2$. Therefore, for $X \sim \mathcal{G}(\alpha = 2, \beta = 1)$ the mean is equal to:

$$E[X] = 1 \cdot 2 = 2 \tag{2.3}$$

and the Variance:

$$Var[X] = 2 \cdot 1^2 = 2$$

4. *Calculate the cost of a call with average duration.*

Having calculated the average duration of a call (2.3), equal to 2, we can replace this value in (2.2) and calculate the cost of an average call:

$$c(x) = \begin{cases} 2, & 0 < x \leq 3 \\ 2 + 6(x-3), & x > 3. \end{cases} \quad \Rightarrow \quad c(x) \overset{x=2}{=} 2$$

5. *Calculate the average cost of a call.*

The average cost of a call will be calculated as the Expected value of $c(x)$:

$$\begin{aligned} E[c(X)] &= \int_0^\infty c(x) f_X(x) dx \\ &= \int_0^3 2 f_X(x) dx + \int_3^\infty 2 + 6(x-3) f_X(x) dx \\ &= \int_0^3 2xe^{-x} dx + \int_3^\infty 2xe^{-x} dx + \int_3^\infty 6(x-3)xe^{-x} dx \\ &= \int_0^\infty 2xe^{-x} dx + \int_3^\infty 6x^2 e^{-x} dx - \int_3^\infty 18xe^{-x} dx \\ &= \dots \\ &= 3.49 \end{aligned}$$

6. *Calculate the 1st and 3rd quartile of call duration.*

Using R we calculate the 1st (0.25) and 3rd (0.75) quantiles:

```
> qgamma(p = c(0.25, 0.75), a, scale=b)
[1] 0.9612788 2.6926345
```

Exercise 3. The files `call_duration_100.txt` and `call_duration_10000.txt` (located at eclass under Documents/datasets) contain the duration (in minutes) of 100 and 10000 randomly sampled calls from the call center of Exercise 2. For each dataset:

1. *Provide a histogram of the observed data. The vertical axis should correspond to relative frequency. The probability density function of the theoretical distribution given in Equation* (2.1) *should be also displayed at the same graph.*

We used the following commands to generate the Histograms shown in Figures 1 and 2, by loading the corresponding data each time.

```
> hist(x, 40, freq = FALSE, xlab = "Duration",
+ main=paste("Histogram of x for", substr(filename, 3, 50),
+ sep=" "))
> xSeq <- seq(min(x),max(x), length = 1000)
> points(xSeq, dgamma(xSeq, a, scale=b), type = "l", col = "red",
+ lwd = 2)
> legend("topright",inset = 0.05, max(dgamma(xSeq, a, scale=b)),
+lty = 1, col = 2, bty = "n",legend=bquote(italic(f)*"
+ ("*italic(x)*") = "*italic(x)*italic(e)^
+ {- italic(x)}*", x">=0))
```

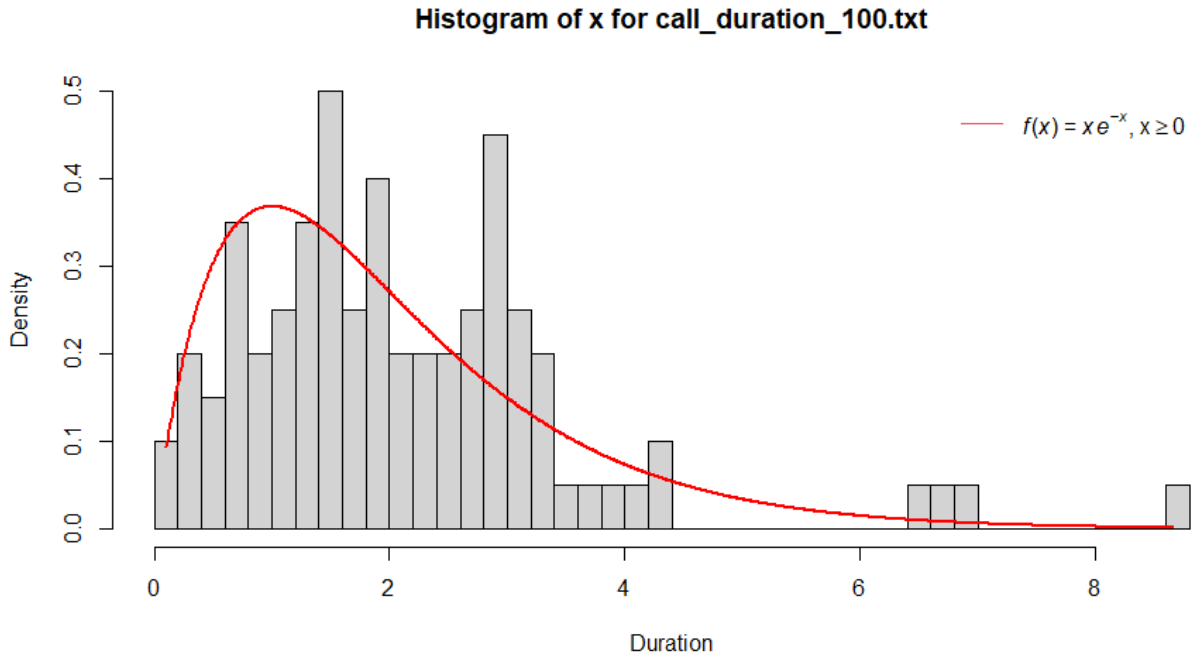

Figure 1: Histogram for `call_duration_100.txt`

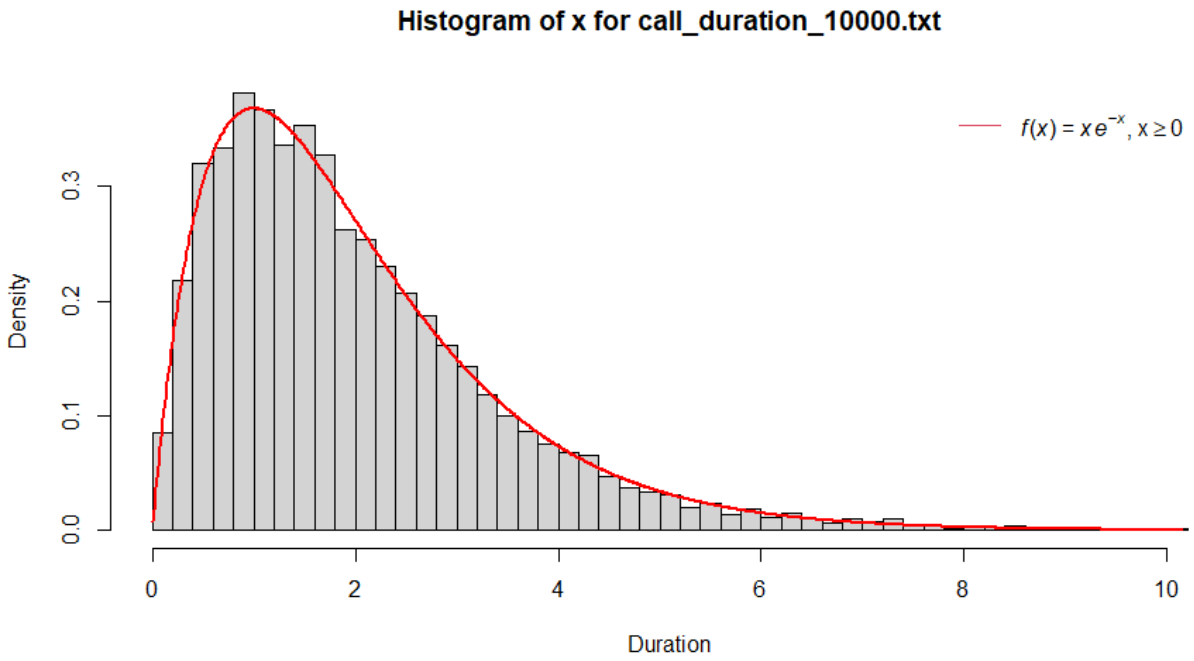

Figure 2: Histogram for `call_duration_10000.txt`

2. *Get an estimate for all quantities described in Exercise 6, that is,*

(a) *Estimate the probability that a call lasts at most one minute.*

(b) *Estimate the probability that a call lasts at least two minutes.*

(c) *Estimate the mean and variance of call duration.*

(d) *Estimate the cost of a call with average duration.*

(e) *Estimate the average cost of a call.*

(f) *Estimate the 1st and 3rd quartile of call duration.*

We first need to define a function that would give the cost of a call, based on Equation (2.2). That is:

```
myfun <- function(x) {
   if(x<=3){ return(2)
      } else {
          return(2+6*(x-3))
      }
}
```

2.1 Answers for **call_duration_100.txt**

```
> n <- length(x)
# (a) P(call lasts at most one minute).
> sum((x <= 1)*1)/n
[1] 0.2
# (b) P(call lasts at least two minutes).
> sum((x >= 2)*1)/n
[1] 0.45
# (c) Mean and Variance of call duration.
> mean(x)
[1] 2.141249
> var(x)
[1] 2.098749
# (d) Cost of a call with average duration.
> lapply(mean(x), FUN = myfun)
[[1]]
[1] 2
# (e) Average cost of a call.
> sum(unlist(lapply(x, FUN = myfun)))/n
[1] 3.451476
# (f) 1st and 3rd quartile of call duration.
> quantile(x, prob = c(1/4,3/4))
     25%      75%
1.215675 2.834025
```

2.2 Answers for **call_duration_10000.txt**

```
> n <- length(x)
# (a) P(call lasts at most one minute).
> sum((x <= 1)*1)/n
[1] 0.2676
# (b) P(call lasts at least two minutes).
> sum((x >= 2)*1)/n
[1] 0.4034
# (c) Mean and Variance of call duration.
> mean(x)
[1] 1.992259
> var(x)
[1] 2.0095
# (d) Cost of a call with average duration.
> lapply(mean(x), FUN = myfun)
[[1]]
[1] 2
# (e) Average cost of a call.
```

```
> sum(unlist(lapply(x, FUN = myfun)))/n
[1] 3.490902
# (f) 1st and 3rd quartile of call duration.
> quantile(x, prob = c(1/4,3/4))
     25%      75%
0.948882 2.676362
```

3. *Compare the theoretic and sample evaluated statistics and comment on the discrepancies between them.*

The results we see comparing the theoretical values of Exercise 2 to that of Exercise 3 are expected. Since the samples, in the files provided, are randomly sampled calls from the call center of Exercise 2, we expect to approach the theoretical ones the more samples we have. Looking at the histograms we may also observe that there aren't too many deviations from the actual values, in the dataset with 10000 rows, while the smaller dataset doesn't follow the theoretical line well. Finally, this exercise points the importance of looking at multiple metrics when doing a statistical analysis. For examples, while the average values of all calculations was close, looking at the Quantiles, we see that the dataset with few samples has a big deviation, compared to the other two.

Exercise 4. *The lifetime of a mechanical assembly is fairly described by the N(10, 2) distribution.*

1. *Calculate the probability that the lifetime is larger than 11.5 but smaller than 13 time units.*

(Solution a) Using Z tables: Since $\mu = 10$ and $\sigma = 2$ we can calculate:

$$\begin{aligned} P(11.5 \leq X \leq 13) &\overset{\mu \geq 0}{=} P(11.5 - \mu \leq X - \mu \leq 13 - \mu) \\ &\overset{\sigma \geq 0}{=} P\left(\frac{11.5 - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{13 - \mu}{\sigma}\right) \\ &= P(0.75 \leq Z \leq 1.5) \end{aligned}$$

Using the standard normal table we conclude that:

$$P(0.75 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z \leq 0.75) = 0.9331928 - 0.7733726 = 0.16$$

(Solution a) Using R: The same result can be calculated using the following commands in R:

```
> m <- 10 ### mean parameter
> s <- 2 ### standard deviation parameter
> pnorm(13, m, s) - pnorm(11.5, m, s)
[1] 0.1598202
```

2. *The manufacturer aims to replace all purchased items failing before a guaranteed minimum lifetime of t timeunits. Calculate the maximum value of t under the restriction that at most 2% of the purchased products can be replaced.*

We are looking to replace at most 2% of all products, for an unkown time of t. Therefore:

$$P(X \leq t) = 0.02 \Rightarrow P\left(\frac{X - \mu}{\sigma} \leq \frac{t - \mu}{\sigma}\right) = 0.02 \Rightarrow P\left(Z \leq \frac{t - \mu}{\sigma}\right) = 0.02 \tag{4.1}$$

Looking at the normal tables, we will find the Z value that is closer to the 98^{th} percentile, and revert the equation. That is $Z = 2.06$. That way, by solving Equation (4.1) we get:

$$Z \leq \frac{t - \mu}{\sigma} \Rightarrow 2.06 \leq \frac{t - 10}{2} \Rightarrow t \leq 5.88 \tag{4.1}$$

3. *Calculate the probability that in 100 independent purchases there will be at least two items that will fail before 7.1 units of time.*

We will assume that we have 100 independent and identically distibuted purchases $X_i, i = 1, \ldots, n$. Let's calculate the probability that a product fails before 7.1 units of time. Similar to Question 1 we get that:

$$P(X \leq 7.1) = 0.074 \Rightarrow P(X \geq 7.1) = 0.926$$

We can then say that $X \sim Bernoulli(0.926)$ with $\mu = 0.926$ and $\sigma = 0.262$. For $n = 100$ independent trials and by virtue of the Central Limit Theorem we have:

$$\frac{\sum_{i=1}^{n} X_i/n - \mu}{\sigma/\sqrt{n}} \overset{\text{n}\to\infty}{\to} \mathcal{N}(0,1)$$

Thus,

$$\begin{aligned}
P\left(\sum_{i=1}^{100} X_i \geq 2\right) &= 1 - P\left(\sum_{i=1}^{100} X_i \leq 2\right) \\
&= 1 - P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq \frac{2 - n\mu}{\sigma\sqrt{n}}\right) \\
&= 1 - P\left(\frac{\sum_{i=1}^{n} X_i - 92.6}{2.62} \leq \frac{2 - 92.6}{2.62}\right) \\
&\overset{\text{n}\to\infty}{\to} 1 - P(Z \leq 2.06), \text{ where } Z \sim \mathcal{N}(0,1) \\
&\approx 0.0197
\end{aligned}$$

So the probability of the event that at in 100 independent purchases there will be at least two items that will fail before 7.1 units of time is approximately equal to 1.97%.

Exercise 5. *A target moves within region A, shown in the graycoloured area (Figure 3), where $c > 0$ denotes a given constant. Let X and Y denote the two random variables corresponding to the coordinates of the target at a given instance. The joint probability density function of X and Y is given by*

$$f_{X,Y}(x,y) = \frac{2}{c^2} I_A(x,y) \tag{1}$$

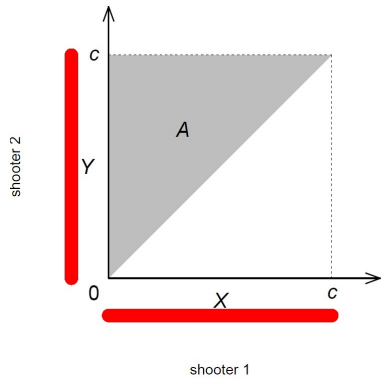

Figure 3: Target movement(graycolored area).

1. *Show that f is a probability density function for all $c > 0$.*

First of all we need to identify the possible values for x, y. Given that $c > 0$ and based on Figure 3 we may write that: $A = \{(x,y) \in R^2 : x \geq 0, y \geq 0, y \geq x\}$. Obviously $f_{X,Y}(x,y) > 0$ for all $(x,y) \in A$. Next we have to show that $\iint_A f_{X,Y}(x,y)dxdy = 1$. Indeed:

$$\begin{aligned}
\iint_A f_{X,Y}(x,y)dxdy &= \int_0^c \left(\int_0^y \frac{2}{c^2} dx \right) dy \\
&= \frac{2}{c^2} \int_0^c [x]_0^y \, dy \\
&= \frac{2}{c^2} \int_0^c (y) dy \\
&= \frac{2}{c^2} \left[\frac{y^2}{2} \right]_0^c \\
&= \frac{2}{c^2} \left(\frac{c^2}{2} \right) \\
&= 1
\end{aligned}$$

2. *What is the probability* $P(X \leq c/2, Y \leq c/2)$*?*

$$\begin{aligned}
\iint_A f_{X,Y}(x,y)dxdy &= \int_0^{c/2} \left(\int_0^y \frac{2}{c^2} dx \right) dy \\
&= \frac{2}{c^2} \int_0^{c/2} [x]_0^y \, dy \\
&= \frac{2}{c^2} \int_0^{c/2} (y) \, dy \\
&= \frac{2}{c^2} \left[\frac{y^2}{2} \right]_0^{c/2} \\
&= \frac{2}{c^2} \left(\frac{c^2}{4} \right) \\
&= \frac{1}{2}
\end{aligned}$$

3. *Find the marginal distributions of* X *and* Y*. Are* X *and* Y *independent random variables? Are* X *and* Y *identical random variables?*

For $y \in (0, c)$ we have that $0 < x < y$ and $y < c$. Thus:

$$\begin{aligned}
f_X(x) &= \int_0^c f_{X,Y}(x,y)dy \\
&= \int_x^c \frac{2}{c^2} dy \\
&= \frac{2}{c^2} [y]_x^c \\
&= \frac{2}{c^2} (c - x)
\end{aligned}$$

Similarly:

$$\begin{aligned}
f_Y(y) &= \int_0^c f_{X,Y}(x,y)dx \\
&= \int_0^y \frac{2}{c^2} dx \\
&= \frac{2}{c^2} [y]_0^y \\
&= \frac{2}{c^2} y
\end{aligned}$$

X and Y are not independent since $f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$. Moreover, X and Y are not identical random variables since $f_X(\cdot)$ is different from $f_Y(\cdot)$.

4. *Calculate the marginal probabilities* $P(X \leq c/2)$ *and* $P(Y \leq c/2)$.

$$\begin{aligned} P(X \leq c/2) &= \int_0^{c/2} f_X(x)dx \\ &= \int_0^{c/2} \frac{2}{c^2}(c-x)\,dx \\ &= \frac{2}{c^2}\left[cx - \frac{x^2}{2}\right]_0^{c/2} \\ &= \frac{2}{c^2}\left(\frac{c^2}{2} - \frac{c^2}{8}\right) \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} P(Y \leq c/2) &= \int_0^{c/2} f_Y(y)dy \\ &= \int_0^{c/2} \left(\frac{2}{c^2}y\right)dy \\ &= \frac{2}{c^2}\left[\frac{y^2}{2}\right]_0^{c/2} \\ &= \frac{1}{2} \end{aligned}$$

5. *Calculate* $Cov(X,Y)$ *and* $\rho(X,Y)$.

The covariance function is given by:

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

We calculate that:

$$\begin{aligned} E(X,Y) &= \iint_A xyf_{X,Y}(x,y)dxdy \\ &= \int_0^{c/2}\left(\int_0^y \frac{2xy}{c^2}dx\right)dy \\ &= \frac{1}{c}\int_0^{c/2} y\left(\int_0^y 2xdx\right)dy \\ &= \frac{1}{c}\int_0^{c/2} y^3dy \\ &= \frac{1}{c^2}c^4 \\ &= c^2 \end{aligned} \tag{5.1}$$

$$
\begin{aligned}
E(X) &= \int_0^c \frac{2x}{c^2}\,(c-x)\,dx \\
&= \frac{1}{c^2}\left[cx^2 - \frac{2x^3}{3}\right]_0^c dx \\
&= \dots \\
&= \frac{c}{3}
\end{aligned} \tag{5.2}
$$

$$
\begin{aligned}
E(Y) &= \int_0^c \frac{2y^2}{c^2} dy \\
&= \dots \\
&= \frac{2c}{3}
\end{aligned} \tag{5.3}
$$

Using (5.1), (5.2), (5.3) we obtain that:

$$
Cov(X,Y) = E(XY) - E(X)E(Y) = c^2 - \frac{c}{3}\cdot\frac{2c}{3} = -\frac{7c^2}{9}
$$

The Correlation is calculated based on:

$$
\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]}\sqrt{Var[Y]}} \tag{5.4}
$$

Knowing the numerator we can calculate the variance of X and Y. Similar to (5.2):

$$
\begin{aligned}
E(X^2) &= \int_0^c \frac{2x^2}{c^2}\,(c-x)\,dx \\
&= \dots \\
&= \frac{c^2}{6}
\end{aligned} \tag{5.5}
$$

$$
\begin{aligned}
E(Y) &= \int_0^c \frac{2y^3}{c^2} dy \\
&= \dots \\
&= \frac{c^2}{2}
\end{aligned} \tag{5.6}
$$

Using (5.4), (5.5), (5.6):

$$\begin{aligned}\rho(X,Y) &= \frac{Cov(X,Y)}{\sqrt{Var[X]}\sqrt{Var[Y]}} \\ &= -\frac{7c^2}{9\sqrt{\frac{c^2}{6}}\sqrt{\frac{c^2}{2}}} \\ &= -\frac{7c^2}{9\sqrt{\frac{c^2}{6}}\sqrt{\frac{c^2}{2}}} \\ &= \dots \\ &= -2.6943 \end{aligned} \tag{5.4}$$

6. *Find the conditional distributions of X given Y and Y given X.*

The conditional distributions are:

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \dots \\ &= \frac{1}{y} \end{aligned}$$

and

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \dots \\ &= \frac{1}{(c-x)} \end{aligned}$$

Exercise 6. Let $X_1, ..., X_n$ denote an iid sample of n observations from the distribution

1. $\mathcal{B}(10, 0.1)$
2. $\mathcal{P}(1.9)$
3. $\mathcal{NB}(5, 0.65)$
4. $\mathcal{G}(5, 0.5)$
5. $\mathcal{X}_1^2$

For each case calculate the probability $P(\bar{X}_n \leq 2)$, where $\bar{X}_n = \sum_{i=1}^n X_i/n$ using

1. exact calculations

2. Central Limit Theorem approximation

for $n = 2, 5, 10, 20, 40$.
Hint: For the exact calculation use the following properties

1. If $Xi \sim \mathcal{B}(\nu_i, p)$, independent for $i = 1, \ldots n$, then $\sum_{i=1}^n Xi \sim \mathcal{B}(\sum_{i=1}^n \nu_i, p)$.
2. If $Xi \sim \mathcal{P}(\lambda_i)$, independent for $i = 1, \ldots n$, then $\sum_{i=1}^n Xi \sim \mathcal{P}(\sum_{i=1}^n \lambda_i)$.
3. If $Xi \sim \mathcal{NB}(\nu_i, p)$, independent for $i = 1, \ldots n$, then $\sum_{i=1}^n \mathcal{NB}(P \sum_{i=1}^n \nu_i, p)$.
4. If $Xi \sim \mathcal{G}(\alpha_i, \beta)$, independent for $i = 1, \ldots n$, then $\sum_{i=1}^n \mathcal{G}(\sum_{i=1}^n \alpha_i, \beta)$.
5. The X^2 distribution is a special case of the Gamma distribution $G(\nu/2, 2)$ (shape scale parameterization) with $\nu \in Z_+$.

Solution

Exact calculations: For each distribution we will use the properties mentioned. First of all we will transform the needed probability:

$$P(\bar{X}_n \leq 2) = P\left(\sum_{i=1}^n X_i/n \leq 2\right) = P\left(\sum_{i=1}^n X_i \leq 2n\right) \tag{6.1}$$

We will now write a script in R that will use the properties mentioned and for the various values of n will calculate the needed properties (example below is for Bernoulli but the function changes each time).

```
> for(n in c(2,5,10,20,40)){
+ q <- 2*n #same for all distributions
+ size <- 10*n #parameters change depending on the distribution
+ prob <- 0.1
+ cat("Probability for n =",n,":",pbinom(q, size, prob),'')
+ }
```

For each distribution we calculate:

1. $\mathcal{B}(10, 0.1)$

```
Probability for n = 2 :  0.9568255
Probability for n = 5 :  0.9906454
Probability for n = 10 :  0.9991924
Probability for n = 20 :  0.9999928
Probability for n = 40 :  1
```

2. $\mathcal{P}(1.9)$

```
Probability for n = 2 :  0.6678436
Probability for n = 5 :  0.6453284
Probability for n = 10 :  0.6471744
Probability for n = 20 :  0.6656928
Probability for n = 40 :  0.7020047
```

3. $\mathcal{NB}(5, 0.65)$

```
Probability for n = 2 :  0.422723
Probability for n = 5 :  0.2715955
Probability for n = 10 :  0.1580295
Probability for n = 20 :  0.06435676
Probability for n = 40 :  0.01320122
```

4. $\mathcal{G}(5, 0.5)$

```
Probability for n = 2 :  0.2833757
Probability for n = 5 :  0.1567726
Probability for n = 10 :  0.07033507
Probability for n = 20 :  0.01710831
Probability for n = 40 :  0.001269693
```

5. $\mathcal{X}_1^2$

```
Probability for n = 2 :  0.9544997
Probability for n = 5 :  0.9984346
Probability for n = 10 :  0.9999923
Probability for n = 20 :  1
Probability for n = 40 :  1
```

Using the central limit theorem: For each disrtibution we will use the central limit theorem. Using (6.1) we get:

$$P\left(\sum_{i=1}^{n} X_i \leq 2n\right) = P\left(\sum_{i=1}^{n} \frac{X_i - n\mu}{\sqrt{n}\sigma} \leq \frac{2n - n\mu}{\sqrt{n}\sigma}\right)$$

$$\overset{\text{n}\to\infty}{\to} P\left(Z \leq \frac{2n - n\mu}{\sqrt{n}\sigma}\right), \text{ where } Z \sim \mathcal{N}(0,1) \tag{5.3}$$

We, therefore, need to calculate the Mean and standard deviation for each distribution.

Table 1: Mean and Standard deviation of given distributions

Distribution	Theoretical Mean	Theoretical Standard Deviation	Mean	Standard Deviation
1. $\mathcal{B}(10, 0.1)$	np	$\sqrt{np(1-p)}$	1	0.974004
2. $\mathcal{P}(1.9)$	λ	$\sqrt{\lambda}$	1.9	1.3784
3. $\mathcal{NB}(5, 0.65)$	n(1-p)/p	$\sqrt{n(1-p)/p^2}$	2.69231	2.03519
4. $\mathcal{G}(5, 0.5)$	$\alpha\beta$	$\sqrt{\alpha\beta^2}$	2.5	1.11803
5. $\mathcal{X}_1^2$	ν	$\sqrt{2\nu}$	1	$\sqrt{2}$

Having all the info we need we can write a code in R in order to calculate the needed probablilities.

```
> for(n in c(2,5,10,20,40)){
+ mu <- 1  mean
+ s <- 0.974004  change according to table
+ q <- (2*n - n*mu)/(sqrt(n)*s)
+ cat("Probability for n =",n,":",pnorm(q, 0, 1),'')
+ }
```

1. $\mathcal{B}(10, 0.1)$

```
Probability for n = 2 :  0.9267435
Probability for n = 5 :  0.9891549
Probability for n = 10 :  0.9994162
Probability for n = 20 :  0.9999978
Probability for n = 40 :  1
```

2. $\mathcal{P}(1.9)$

```
Probability for n = 2 :  0.5408591
Probability for n = 5 :  0.5644345
Probability for n = 10 :  0.5907274
```

```
Probability for n = 20 :  0.627199
Probability for n = 40 :  0.676823
```

3. $\mathcal{NB}(5, 0.65)$

```
Probability for n = 2 :  0.3152324
Probability for n = 5 :  0.2234353
Probability for n = 10 :  0.1410282
Probability for n = 20 :  0.06409414
Probability for n = 40 :  0.01572145
```

4. $\mathcal{G}(5, 0.5)$

```
Probability for n = 2 :  0.2635439
Probability for n = 5 :  0.1586544
Probability for n = 10 :  0.07864886
Probability for n = 20 :  0.02274975
Probability for n = 40 :  0.002338794
```

5. $\mathcal{X}_1^2$

```
Probability for n = 2 :  0.8413447
Probability for n = 5 :  0.9430769
Probability for n = 10 :  0.9873263
Probability for n = 20 :  0.9992173
Probability for n = 40 :  0.9999961
```

Comparing the results from the exact calculation we observe the larger the sample the closer the Central Limit Theorem approximated the exact values.