STAT 321: Assignment 3

Saksham Sudershan

19 February 2022

Problem 1

(a)

$$\mathbb{E}(N|R=r) = n \times p \qquad \textit{where } p \textit{ is the probability of success}$$
$$\mathbb{E}(N|R=r) = 10 \times \frac{r}{10}$$
$$\mathbb{E}(N|R=r) = r$$

(b)

$$\mathbb{V}ar[N|R=r] = n \times p \times (1-p)$$
$$\mathbb{V}ar[N|R=r] = 10 \times \frac{r}{10} \times \left(1 - \frac{r}{10}\right)$$
$$\mathbb{V}ar[N|R=r] = r \times \left(1 - \frac{r}{10}\right)$$

(c)

$$\mathbb{E}(R) = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10}$$
$$\mathbb{E}(R) = 5.5 = r$$
$$\mathbb{E}(N) = r$$

$$\mathbb{E}(N) = 5.5$$

(d)

Using the Law of Total Variance,

$$\mathbb{V}ar(N) = \mathbb{E}\left[\mathbb{V}ar(N|R)\right] + \mathbb{V}ar\left(\mathbb{E}[N|R]\right)$$
$$= \mathbb{E}\left[r \times \left(1 - \frac{r}{10}\right)\right] + \mathbb{V}ar(r)$$

We know that $\mathbb{E}[r] = \bar{r} = 5.5$. Now $\mathbb{V}ar(r)$ is given by,

$$\mathbb{V}ar(r) = \frac{\sum_{i=1}^{10}(r_i - \bar{r})^2}{10}$$

$$\mathbb{V}ar(r) = \frac{(1-5.5)^2 + (2-5.5)^2 + (3-5.5)^2 + (4-5.5)^2 + (5-5.5)^2 + (6-5.5)^2 + (7-5.5)^2 + (8-5.5)^2 + (9-5.5)^2 + (10-5.5)^2}{10}$$

$$\mathbb{V}ar(r) = \frac{82.5}{10} = 8.25$$

And $\mathbb{E}\left[r \times \left(1 - \frac{r}{10}\right)\right]$ is given by,

$$\mathbb{E}(r) - \frac{1}{10} \times \mathbb{E}(r^2) = 5.5 - \frac{1}{10} \times 38.5$$

So,

$$\mathbb{V}ar(N) = 5.5 - 3.85 + 8.25$$
$$\Rightarrow \mathbb{V}ar(N) = 9.9$$

Problem 2

(a)

The discrete probability mass function $p(x_1, x_2)$ is given by,

$\frac{X_1=}{X_2=}$	0	1	2	3
0	$\frac{\binom{3}{3}}{\binom{14}{3}}$	$\frac{\binom{5}{1}\times\binom{3}{2}}{\binom{14}{3}}$	$\frac{\binom{5}{2}\times\binom{3}{1}}{\binom{14}{3}}$	$\frac{\binom{5}{3}}{\binom{14}{3}}$
1	$\frac{\binom{6}{1}\times\binom{3}{2}}{\binom{14}{3}}$	$\frac{\binom{5}{1}\times\binom{6}{1}\times\binom{3}{1}}{\binom{14}{3}}$	$\frac{\binom{5}{2}\times\binom{6}{1}}{\binom{14}{3}}$	0
2	$\frac{\binom{6}{2}\times\binom{3}{1}}{\binom{14}{3}}$	$\frac{\binom{6}{2}\times\binom{5}{1}}{\binom{14}{3}}$	0	0
3	$\frac{\binom{6}{3}}{\binom{14}{3}}$	0	0	0

This can be further simplified to,

$\frac{X_1=}{X_2=}$	0	1	2	3
0	$\frac{1}{364}$	$\frac{15}{364}$	$\frac{30}{364}$	$\frac{10}{364}$
1	$\frac{18}{364}$	$\frac{90}{364}$	$\frac{60}{364}$	0
2	$\frac{45}{364}$	$\frac{75}{364}$	0	0
3	$\frac{20}{364}$	0	0	0

(b)

The marginal probability mass function $p(x_1)$ is given by,

$X_1 =$	0	1	2	3
	$\frac{84}{364}$	$\frac{180}{364}$	$\frac{90}{364}$	$\frac{10}{364}$

And the marginal probability mass function $p(x_2)$ is given by,

$X_2 =$	0	1	2	3
	$\frac{56}{364}$	$\frac{168}{364}$	$\frac{120}{364}$	$\frac{20}{364}$

(c)

To calculate the correlation between X_1 and X_2, we can first calculate the covariance,

$$Cov(X_1, X_2) = \sum_{(x,y)\in S} (x_1 - \mu_{x_1})(x_2 - \mu_{x_2}) f(x_1, x_2)$$

From the marginal probability mass function, we know that $\mu_{x_1} = 1.07$ and $\mu_{x_2} = 1.28$.

$$\Rightarrow Cov(X_1, X_2) = (0-1.0714)(0-1.2857)\left(\frac{1}{364}\right)+(1-1.0714)(0-1.2857)\left(\frac{15}{364}\right)+(2-1.0714)(0-1.2857)\left(\frac{30}{364}\right)$$

$$+(3-1.0714)(0-1.2857)\left(\frac{10}{364}\right)+(0-1.0714)(1-1.2857)\left(\frac{18}{364}\right)+(1-1.0714)(1-1.2857)\left(\frac{90}{364}\right)$$

$$+(2-1.0714)(1-1.2857\left(\frac{60}{364}\right)+(0-1.0714)(2-1.2857)\left(\frac{45}{364}\right)+(1-1.0714)(2-1.2857)\left(\frac{75}{364}\right)$$

$$+(0-1.0714)(3-1.2857)\left(\frac{20}{364}\right)$$

$$\Rightarrow Cov(X_1, X_2) \approx -0.388$$

We also have to find the standard deviations of X_1 and X_2,

$$\sigma_{X_1} = \sum_{i=1}^{4}(x_{1_i} - \mu_{x_1}) = \sqrt{\frac{1458}{2548}}$$

$$and\ \sigma_{X_2} = \sum_{i=1}^{4}(x_{2_i} - \mu_{x_2}) = \sqrt{\frac{396}{637}}$$

$$Therefore\ Corr(X_1, X_2) \approx \frac{-0.388}{0.596} = -0.651$$

Problem 3

(a)

We can find c by equating the total probability equal to 1. So,

$$\int_0^{30} c(6-t)^2 dt = 1$$

$$\Rightarrow c\int_0^{30}(6-t)^2dt = 1$$

$$\Rightarrow c\int_0^{30} 36+t^2-12tdt = 1$$

$$\Rightarrow c\left[36t+\frac{t^3}{3}-6t^2\right]_0^{30} = 1$$

$$\Rightarrow c\times 4680 = 1$$

$$\Rightarrow c = \frac{1}{4680}$$

(b)

The cumulative distribution function is given by,

$$F(t) = P(T \leq t)$$

$$\int_0^t f(t)dt = \frac{1}{4680}\int_0^t (6-t)^2dt$$

$$= \frac{1}{4680}\left[36t+\frac{t^3}{3}-6t^2\right]_0^t$$

$$= \frac{1}{4680}\left[36t+\frac{t^3}{3}-6t^2\right]$$

$$\Rightarrow F(t) = 0, when\ t < 0$$

$$and\ F(t) = \frac{1}{4680}\left[36t+\frac{t^3}{3}-6t^2\right], when\ 0 \leq t \leq 30$$

$$and\ F(t) = 1, when\ t > 30$$

(c)

The probability that a student will wait more than 10 minutes is given by,

$$P(T > 10) = 1 - P(T \leq 10)$$

$$= 1 - F(10)$$

$$= 1 - \frac{1}{4680}\left[36t+\frac{t^3}{3}-6t^2\right]$$

$$\Rightarrow P(T > 10) \approx 0.9801$$

(d)

The probability that a student will wait more than 20 minutes given that they have waited for 10 minutes is given by,

$$P(T \geq 20 | T > 10) = \frac{P(T \geq 20 \cap T > 10)}{P(T > 10)}$$

Similar to the previous question, this can be written as,

$$\frac{1-F(20)}{1-F(10)} = \frac{277/351}{344/351} = 0.80523$$

(e)

The probability that a student will be late for class is given by,

$$P(T > 20) = 1 - F(20)$$

$$\Rightarrow P(T > 20) = \frac{277}{351} = 0.789$$

Using the binomial distribution, setting the probability of being late as the probability of success, $p = 0.789$. Then,

$$P(3\ students\ are\ late) = \binom{n}{r}(p)^r(1-p)^{n-r} = \binom{10}{3}(0.789)^3(1-0.789)^7$$

$$So,\ P(3\ students\ are\ late) \approx 0.0011$$

Problem 4

(a)

We know that $Y \sim exp(\lambda)$. So, the probability distribution function of Y is given by,

$$f_y(y) = \lambda e^{-\lambda y}$$

And the cumulative distribution function of Y is given by,

$$F_y(y) = 1 - e^{-\lambda y} = P(Y \leq y)$$

And $X = \lfloor Y \rfloor + 1$. So, the probability distribution function of X is given by,

$$f_x(x) = P(X = x)$$

We know that X takes the value k, whenever $k - 1 \leq Y \leq k$, so

$$f_x(x) = P(x - 1 \leq Y \leq x)$$

Thus, we are trying to find the probability that Y lies between x and $x - 1$. This can also be written as,

$$P(Y \leq x) - P(Y \leq k - 1)$$

$$= F_y(k) - F_y(k-1)$$

Using the cdf of Y, this can be written as

$$\left(1 - e^{-\lambda x}\right) - \left(1 - e^{-\lambda(x-1)}\right)$$

$$= 1 - e^{-\lambda x} - 1 + -e^{-\lambda(x-1)}$$

$$= e^{-\lambda(x-1)} - e^{-\lambda x}$$

$$= e^{-\lambda x} \cdot e^{\lambda} - e^{-\lambda x}$$

$$= e^{-\lambda x}[e^{\lambda} - 1]$$

Now, multiplying and dividing by e^{λ}, this becomes

$$= e^{-\lambda x} \cdot e^{\lambda}\left[\frac{e^{\lambda}}{e^{\lambda}} - \frac{1}{e^{\lambda}}\right]$$

$$\Rightarrow f_x(x) = e^{-\lambda(x-1)}[1 - e^{-\lambda}]$$

This resembles the probability distribution function of a geometric distribution, where the probability of success is given by $1 - e^{-\lambda}$. So we can say that,

$$X \sim Geom(p) \qquad , where\ p = 1 - e^{-\lambda}$$

Problem 5

(a)

```
temp_val <- data.frame()
sn_save <- data.frame()
for(j in 1:10){
  for(i in 1:10000){
    temp_val[i,1] = sum(rnorm((10*j), mean = 0, sd = 1))
    temp_val[i,2] = abs(temp_val[i,1])
    }
  sn_save[j,1] <- (mean(temp_val[,2]))
}
row.names(sn_save) <- c("n=10","n=20","n=30","n=40",
                        "n=50","n=60","n=70","n=80","n=90","n=100")
colnames(sn_save) <- c("Expected Value of |Sn|")
sn_save
```

```
##       Expected Value of |Sn|
## n=10                2.563111
## n=20                3.564391
## n=30                4.363401
## n=40                5.030196
## n=50                5.570841
## n=60                6.256573
## n=70                6.699018
## n=80                7.101563
## n=90                7.523201
## n=100               7.941398
```

(b)

The distribution of S_n is based upon the value of n, where n takes the values $10, 20, 30 \cdots$

However, since we know that S_n is comprised of i.i.d. $X_i s$ which themselves are normally distributed, we can say that S_n is also normally distributed, where $\mu_{S_n} = 0$, and so,

$$S_n \sim \mathbb{N}(0, n)$$

Now, the 95% confidence interval for S_n is given by,

```
for(i in 1:10){
    print(paste("For n =", i*10,",the 95% is (", round(-1.96*sqrt(i*10),
                digits =4),",",round(1.96*sqrt(i*10),digits = 4),")"))
}
```

```
## [1] "For n = 10 ,the 95% is ( -6.1981 , 6.1981 )"
## [1] "For n = 20 ,the 95% is ( -8.7654 , 8.7654 )"
## [1] "For n = 30 ,the 95% is ( -10.7354 , 10.7354 )"
## [1] "For n = 40 ,the 95% is ( -12.3961 , 12.3961 )"
## [1] "For n = 50 ,the 95% is ( -13.8593 , 13.8593 )"
## [1] "For n = 60 ,the 95% is ( -15.1821 , 15.1821 )"
```

```
## [1] "For n = 70 ,the 95% is ( -16.3985 , 16.3985 )"
## [1] "For n = 80 ,the 95% is ( -17.5308 , 17.5308 )"
## [1] "For n = 90 ,the 95% is ( -18.5942 , 18.5942 )"
## [1] "For n = 100 ,the 95% is ( -19.6 , 19.6 )"
```

For the absolute value of S_n, the lower bound for the confidence interval will be 0, while the upper bound remains the same, and this will contain 95% of the values. This is because due to the 'folding' or the use of the absolute values, the 95% CI folds in, and the 95% upper bound remains the same while the lower bound becomes 0. So, the confidence interval for $|S_n|$ is given by,

```
for(i in 1:10){
    print(paste("For n =", i*10,",the 95% is ( 0,",round(1.96*sqrt(i*10),digits = 4),")"))
}
```

```
## [1] "For n = 10 ,the 95% is ( 0, 6.1981 )"
## [1] "For n = 20 ,the 95% is ( 0, 8.7654 )"
## [1] "For n = 30 ,the 95% is ( 0, 10.7354 )"
## [1] "For n = 40 ,the 95% is ( 0, 12.3961 )"
## [1] "For n = 50 ,the 95% is ( 0, 13.8593 )"
## [1] "For n = 60 ,the 95% is ( 0, 15.1821 )"
## [1] "For n = 70 ,the 95% is ( 0, 16.3985 )"
## [1] "For n = 80 ,the 95% is ( 0, 17.5308 )"
## [1] "For n = 90 ,the 95% is ( 0, 18.5942 )"
## [1] "For n = 100 ,the 95% is ( 0, 19.6 )"
```

(c)

We can verify our findings by editing our previous simulation to also record the CI bound values for each n.

```
temp_val <- data.frame()
sn_save <- data.frame()
for(j in 1:10){
  for(i in 1:10000){
    temp_val[i,1] = sum(rnorm((10*j), mean = 0, sd = 1))
    temp_val[i,2] = abs(temp_val[i,1])
  }
  temp_val[,2] <- temp_val[order(temp_val[,2]),2]
  sn_save[j,1] <- mean(temp_val[,2])
  sn_save[j,2] <- temp_val[1,2]
  sn_save[j,3] <- temp_val[9501,2]
}
row.names(sn_save) <- c("n=10","n=20","n=30","n=40",
                        "n=50","n=60","n=70","n=80","n=90","n=100")
colnames(sn_save) <- c("Expected Value of |Sn|","Lower Bound","Upper Bound")
sn_save
```

```
##       Expected Value of |Sn|  Lower Bound Upper Bound
## n=10                2.526063 1.694815e-04    6.198745
## n=20                3.596346 1.147185e-03    8.741096
## n=30                4.354548 2.368436e-03   10.740161
## n=40                5.031778 9.768985e-05   12.486107
## n=50                5.662284 9.894232e-04   14.023396
```

```
## n=60                 6.146714 5.049906e-04   15.100767
## n=70                 6.729930 4.747185e-04   16.520609
## n=80                 6.993729 3.283182e-04   17.147619
## n=90                 7.751934 1.707308e-03   18.783362
## n=100                7.937878 4.512142e-04   19.262166
```

We find that our bounds found through the simulation are very similar to the theoretical bounds.