Problem set 3

Your name here

Due 10/18/2021 at 5pm

NOTE1: Start with the file `ps3_2022.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F22/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

NOTE2: You will need to have a working LaTeX installation to compile your code.

(The objective with these questions is to illustrate the theorems with examples. We can also do some explicating of a proof.)

Question 1: Expected value

Consider the random variable X characterized by the following PMF:

x	$P(X = x)$
0	.3
1	.3
4	.4

1.1 Compute $\mathbf{E}[X]$. Show your work.

$$
\begin{aligned}
\mathrm{E}[X] &= \sum_x x P(X = x) \\
&= 0 \times .3 + 1 \times .3 + 4 \times .4 \\
&= 1.9
\end{aligned}
$$

1.2 Write a function to compute the expectation of any discrete random variable. The arguments to your function should include the values the random variable can take on (`x`) and the probability it takes on each value (`probs`). Use your function to confirm your answer from question 1.1.

```r
exp_func <- function(x, probs){
  sum(x * probs)
}
exp_func(x = c(0, 1, 4), probs = c(.3, .3, .4))
```

```
## [1] 1.9
```

1.3 Compute the MSE ($\mathbf{E}[(X - c)^2]$) for $c = 1$ and $c = 2$. Show your work.

$$\begin{aligned}
\mathrm{E}\left[(X-1)^2\right] &= \sum_x (x-1)^2 P(X=x) \\
&= (0-1)^2 \times .3 + (1-1)^2 \times .3 + (4-1)^2 \times .4 \\
&= 1 \times .3 + 0 \times .3 + 9 \times .4 \\
&= 3.9 \\
\mathrm{E}\left[(X-2)^2\right] &= \sum_x (x-2)^2 P(X=x) \\
&= (0-2)^2 \times .3 + (1-2)^2 \times .3 + (4-2)^2 \times .4 \\
&= 4 \times .3 + 1 \times .3 + 4 \times .4 \\
&= 3.1
\end{aligned}$$

1.4 Write a function to compute the MSE for any discrete random variable at a value `c`. The arguments to your function should include the values the random variable can take on (`x`), the probability it takes on each value (`probs`), and the value `c` being considered. Use your function to confirm your answers from 1.3.

```
mse_func <- function(x, probs, c){
  sum(probs*(x - c)^2)
}
mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1)
```

```
## [1] 3.9
```

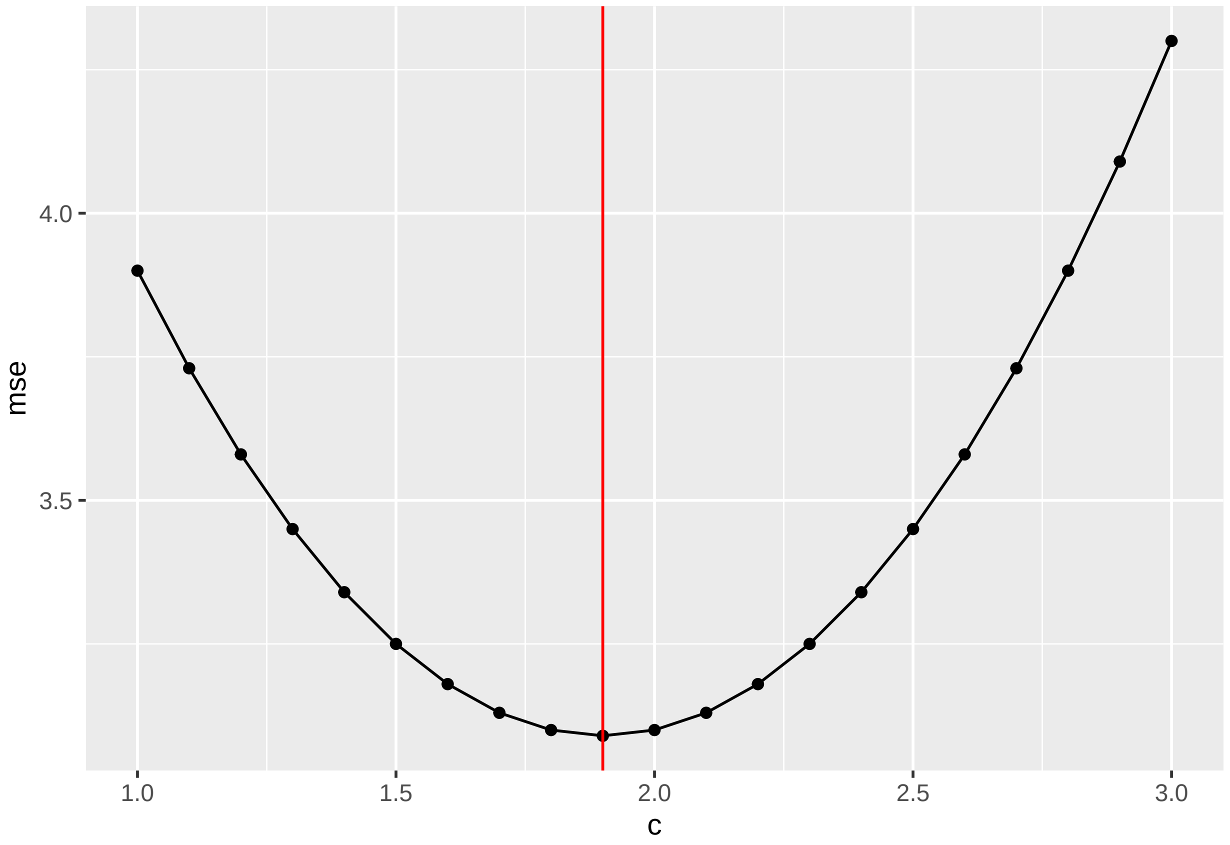
```
mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2)
```

```
## [1] 3.1
```

1.5 Make a plot showing the MSE for values of $c \in \{1.0, 1.1, 1.2, \ldots, 3.0\}$. Add a vertical red line (see `geom_vline()`) at $E[X]$.

Hint: You should be able to do this in a tedious way using only the above and what you have learned about `ggplot`. To make it more efficient code-wise, you may want to use the `sapply` function or the `map_dbl` function.

```
cs <- seq(1, 3, by = .1)
# tedious way
mses <- c(mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.1),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.2),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.3),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.4),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.5),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.6),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.7),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.8),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 1.9),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.1),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.2),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.3),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.4),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.5),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.6),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.7),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.8),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2.9),
          mse_func(x = c(0, 1, 4), probs = c(.3, .3, .4), c = 2))
# less tedious ways
mses <- sapply(cs, mse_func, x = c(0, 1, 4), probs = c(.3, .3, .4)) # base R
mses <- map_dbl(cs, mse_func, x = c(0, 1, 4), probs = c(.3, .3, .4)) # tidy
tibble(c = cs, mse = mses) |>
  ggplot(aes(x = c, y = mse)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1.9, col = "red")
```



Question 2: Variance

Consider the random variable X characterized by the following PMF:

x	$P(X = x)$
2	.1
3	.2
4	.7

2.1 Show that the variance of X is the same whether we compute it by the formula in Definition 2.1.12 or the Alternative Formula in Theorem 2.1.13. (We want the two variance computations, not the proof.)

First we calculate the mean.

$$\begin{aligned}
\mathrm{E}\left[X\right] &= \sum_x xP(X = x) \\
&= 2 \times .1 + 3 \times .2 = 4 \times .7 \\
&= 3.6
\end{aligned}$$

Using Definition 2.1.12:

$$\begin{aligned}
\mathrm{V}\left[X\right] &= \mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right] \\
&= \mathrm{E}\left[(X - 3.6)^2\right] \\
&= (2 - 3.6)^2 \times .1 + (3 - 3.6)^2 \times .2 + (4 - 3.6)^2 \times .7 \\
&= 0.44
\end{aligned}$$

Using Theorem 2.1.13:

$$\begin{aligned}
\mathrm{V}\left[X\right] &= \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 \\
&= \left(2^2 \times .1 + 3^2 \times .2 + 4^2 \times .7\right) - 3.6^2 \\
&= 0.44
\end{aligned}$$

2.2 Show that the variance is the same if you add a constant to X. (We want the two variance computations, not the proof of the first part of Theorem 2.1.14.)

Using Definition 2.1.12:

$$\begin{aligned}
\mathrm{V}\left[X + c\right] &= \mathrm{E}\left[(X + c - \mathrm{E}\left[X + c\right])^2\right] \\
&= \mathrm{E}\left[(X + c - (3.6 + c))^2\right] \\
&= \mathrm{E}\left[(X - 3.6)^2\right] \\
&= (2 - 3.6)^2 \times .1 + (3 - 3.6)^2 \times .2 + (4 - 3.6)^2 \times .7 \\
&= 0.44
\end{aligned}$$

Using Theorem 2.1.13:

$$\begin{aligned}
\mathrm{V}\left[X + c\right] &= \mathrm{E}\left[(X + c)^2\right] - \mathrm{E}\left[X + c\right]^2 \\
&= \left((2 + c)^2 \times .1 + (3 + c)^2 \times .2 + (4 + c)^2 \times .7\right) - (3.6 + c)^2 \\
&= \left(\left(2^2 + 4c + c^2\right) \times .1 + (3^2 + 6c + c^2) \times .2 + (4^2 + 8c + c^2) \times .7\right) - (3.6^2 + 7.2c + c^2) \\
&= 0.44 + (7.2c - 7.2c) + (c^2 - c^2) \\
&= 0.44
\end{aligned}$$

2.3 Show that the variance is multiplied by a^2 if X is multiplied by a. (We want the two variance computations, not the proof of the second part of Theorem 2.1.14.)

Using Definition 2.1.12:

$$\begin{aligned}
\mathrm{V}\left[aX\right] &= \mathrm{E}\left[(aX - \mathrm{E}\left[aX\right])^2\right] \\
&= \mathrm{E}\left[(aX - 3.6a)^2\right] \\
&= (2a - 3.6a)^2 \times .1 + (3a - 3.6a)^2 \times .2 + (4a - 3.6a)^2 \times .7
\end{aligned}$$

$$= 0.256a^2 \times .1 + 0.072a^2 + 0.11a^2$$

$$= 0.44a^2$$

Using Theorem 2.1.13:

$$\begin{aligned}
\mathrm{V}\left[aX\right] &= \mathrm{E}\left[aX^2\right] - \mathrm{E}\left[aX\right]^2 \\
&= \left((2a)^2 \times .1 + (3a)^2 \times .2 + (4a)^2 \times .7\right) - (3.6a)^2 \\
&= \left(.4a^2 + 1.8a^2 + 11.2a^2\right) - 12.96a^2 \\
&= 0.44a^2
\end{aligned}$$

Question 3: CDF of a discrete random variable

Consider the discrete random variable characterized by the following PMF:

x	$P(X = x)$
2	.1
3	.2
4	.7

3.1 What is the cumulative distribution function F evaluated at 2.5 (i.e. $F(2.5)$)?

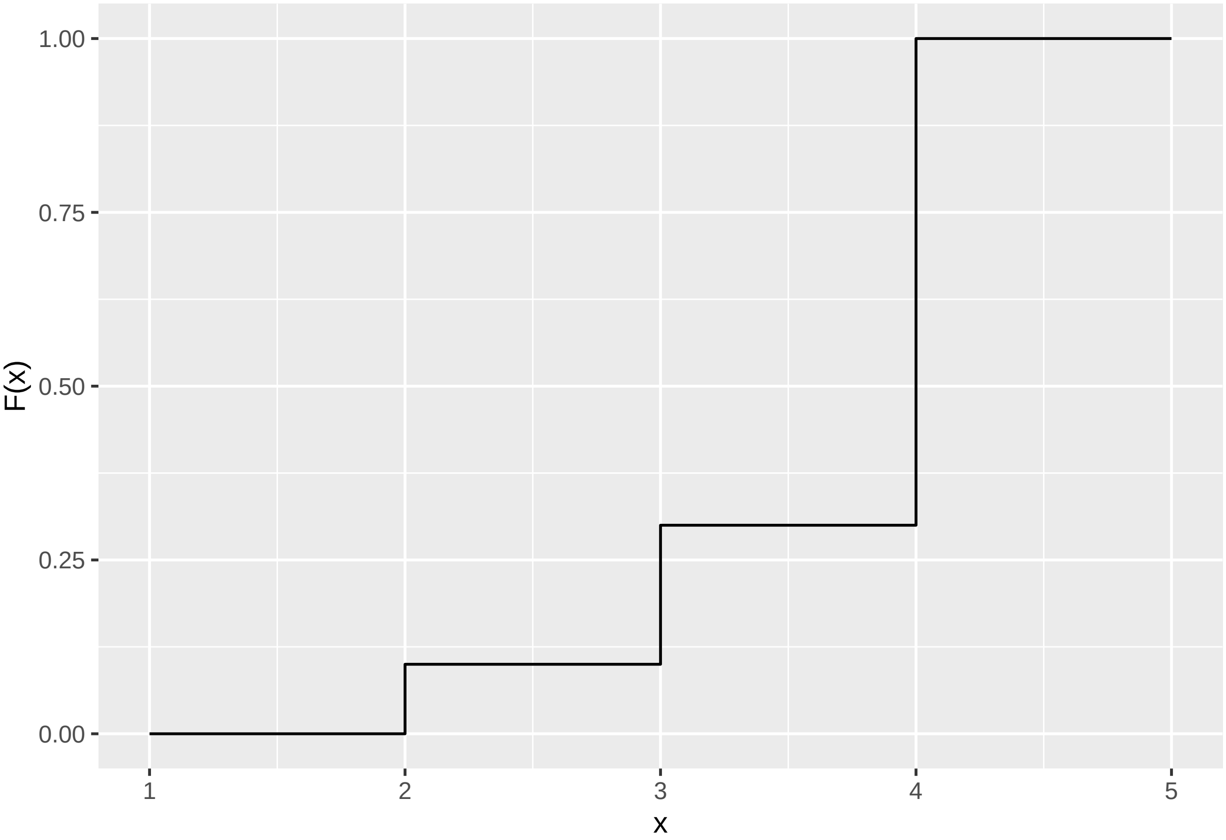
$$P(X \leq 2.5) = 0.1$$

3.2 Plot the cumulative distribution function $F(x)$ for $x \in [1, 5]$. It should look like Figure 1.2.1 in Aronow & Miller. For simplicity you may leave out the open and closed circles and connect the horizontal segments with vertical lines.

Hint: The code below may help you get started.

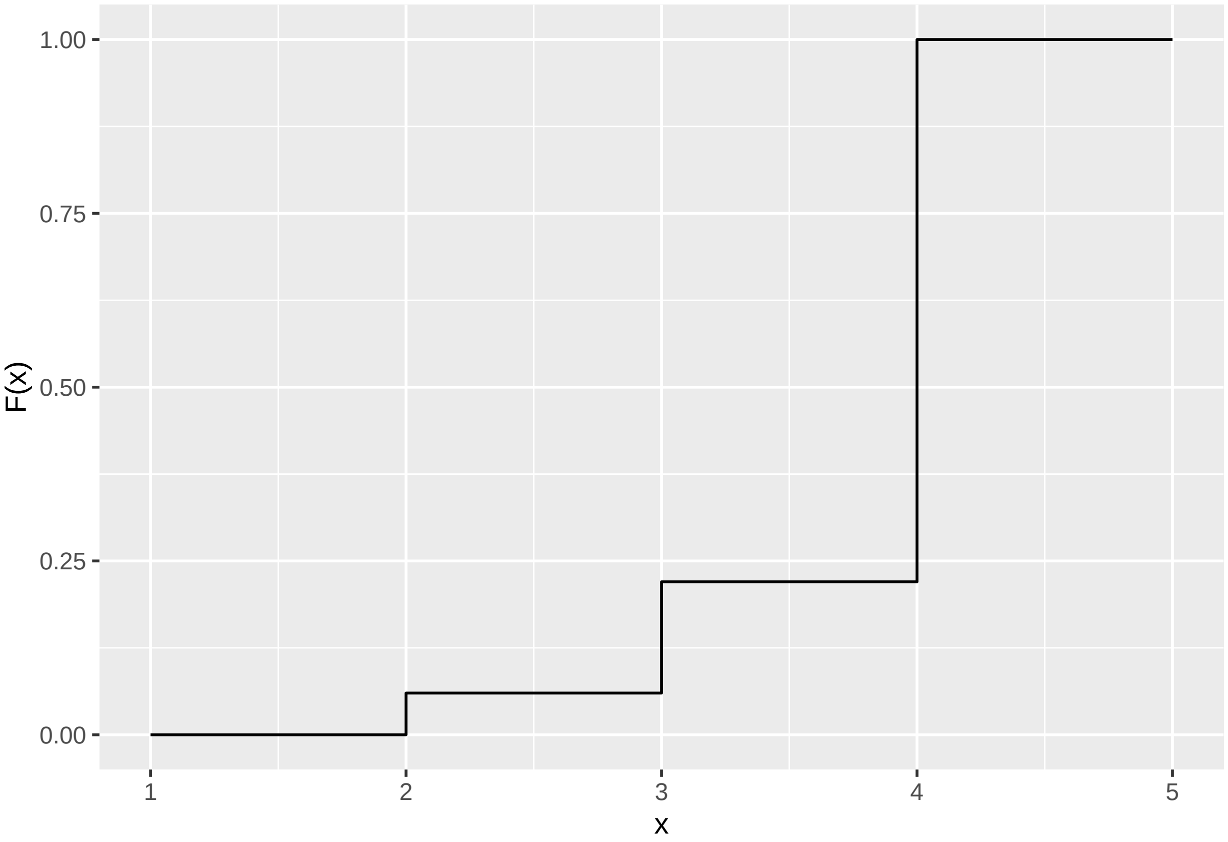
```
# the tibble defines the corners of the step function
tibble(x = c(0, 1, 1, 2, 2, 3),
       y = c(0, 0, 1, 1, 2, 2)) |>
  ggplot(aes(x = x, y = y)) +
  geom_path()
```

```
cdf <- tibble(x = c(1, 2, 2, 3, 3, 4, 4, 5),
       y = c(0, 0, .1, .1, .3, .3, 1, 1))
cdf |>
  ggplot(aes(x = x, y = y)) +
  geom_path() +
  labs(x = "x", y = "F(x)")
```



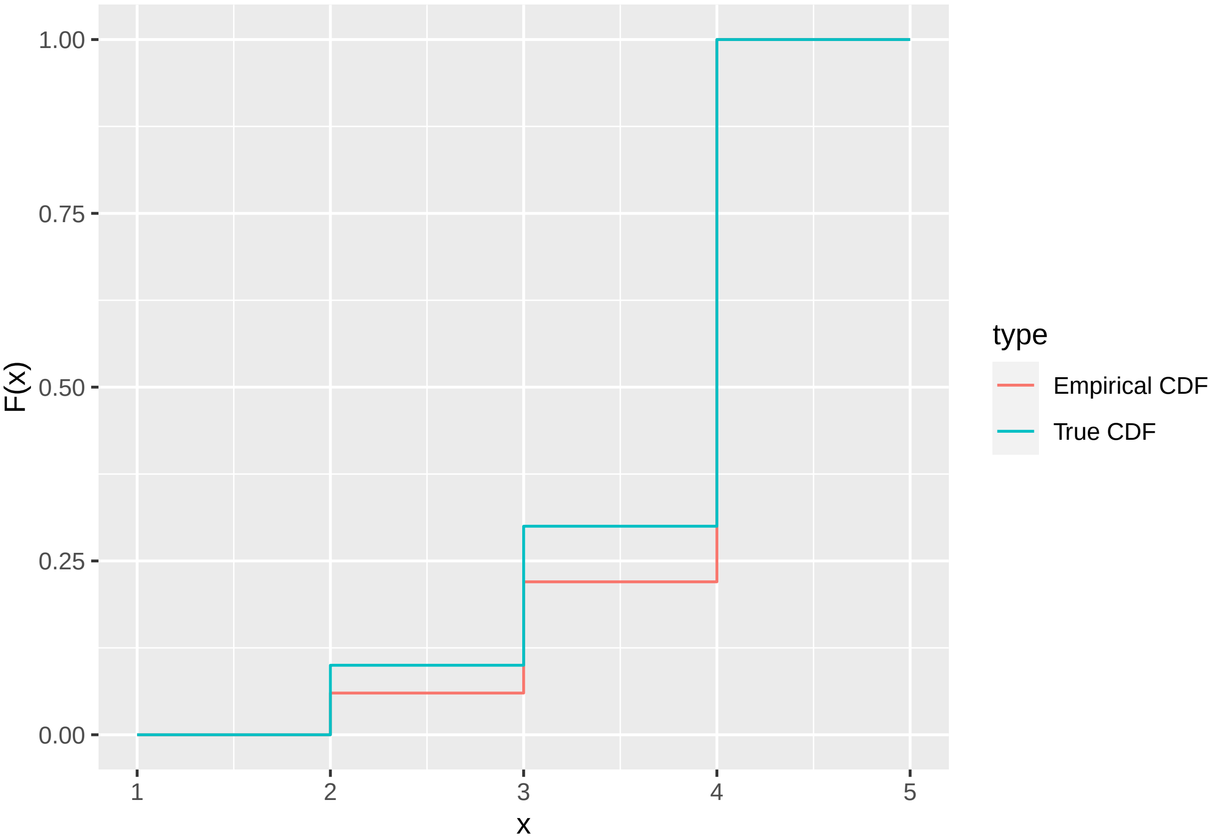
3.3 Draw a sample of size 100 from the PMF above. Plot the empirical CDF from this sample using the same approach you used above.

```
samp <- sample(c(2,3,4), size = 100, replace = T, prob = c(.1, .2, .7))
ecdf <- tibble(x = c(1, 2, 2, 3, 3, 4, 4, 5),
       y = c(0, 0, rep(mean(samp <= 2), 2), rep(mean(samp <= 3), 2), 1, 1))
ecdf |>
  ggplot(aes(x = x, y = y)) +
  geom_path() +
  labs(x = "x", y = "F(x)")
```



3.4 Combine the two CDFs in the same plot, with labels to identify the true CDF and the empirical CDF from your sample.

```
bind_rows(
  cdf |> mutate(type = "True CDF"),
  ecdf |> mutate(type = "Empirical CDF")
) |>
  ggplot(aes(x = x, y = y, col = type)) +
  geom_path() +
  labs(x = "x", y = "F(x)")
```



Question 4: Covariance

Consider the following joint PMF of two random variables, X and Y:

x	y	$P(X = x, Y = y)$
0	0	1/4
1	0	1/4
1	1	1/2

4.1 What is the marginal PMF of X? What is the marginal PMF of Y?

$$f(x) = \begin{cases} 1/4 & x = 0 \\ 3/4 & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f(y) = \begin{cases} 1/2 & y = 0 \\ 1/2 & y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance of X and Y in three ways:

4.2 By hand, using the formula $E[XY] - E[X]E[Y]$. (Show work.)

We need to compute three objects: $\mathrm{E}[X]$, $\mathrm{E}[Y]$, and $\mathrm{E}[XY]$.

$$
\begin{aligned}
\mathrm{E}\left[X\right] &= \sum_x x f(x) \\
&= 1/4 \times 0 + 3/4 \times 1 \\
&= 3/4
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[Y\right] &= \sum_y y f(y) \\
&= 1/2 \times 0 + 1/2 \times 1 \\
&= 1/2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[Y\right] &= \sum_x \sum_y xy f(x, y) \\
&= 0 \times 1/4 + 0 \times 1/4 + 1 \times 1/2 \\
&= 1/2
\end{aligned}
$$

So

$$
\begin{aligned}
\mathrm{Cov}(X, Y) &= 1/2 - 3/4 \times 1/2 \\
&= 1/2 - 3/8 = 1/8
\end{aligned}
$$

4.3 By hand, using the formula $E[(X - E[X])(Y - E[Y])]$. (Show work.)

We have $\mathrm{E}\left[X\right]$ and $\mathrm{E}\left[Y\right]$ from the previous question. So

$$
\begin{aligned}
\mathrm{Cov}[X, Y] &= \mathrm{E}\left[(X - \mathrm{E}\left[X\right])(Y - \mathrm{E}\left[Y\right])\right] \\
&= \sum_x \sum_y (X - 3/4)(Y - 1/2) f(x, y) \\
&= 1/4 \times (-3/4)(-1/2) + 1/4 \times (1/4)(-1/2) + 1/2(1/4)(1/4) \\
&= 3/32 - 1/32 + 1/32 \\
&= 1/8
\end{aligned}
$$

4.4 Using `R`, by creating a population with the above frequencies and using the `cov()` function.

```
dat <- tibble(x = c(0, 1, 1, 1), y = c(0, 0, 1 , 1))
cov3 <- cov(dat$x, dat$y)
# tidy alternative:
cov3 <- dat |> summarize(cov(x, y))
cov3
```

```
## # A tibble: 1 x 1
##   `cov(x, y)`
##         <dbl>
## 1       0.167
```

4.5 (Bonus) You should find that your answer to 4.4 is not exactly the same as the other two. Why is this? Show how to make your answer to 4.4 the same as the others.

The reason is that `R`'s `cov()` function applies a degrees-of-freedom correction, dividing by $n - 1$ instead of by n. We can make the responses the same by multiplying by $\frac{n-1}{n}$:

```
cov3*(nrow(dat)-1)/nrow(dat)
```

```
##   cov(x, y)
## 1     0.125
```