Problem Set 2

Daniel Shapiro

Question 1 Background:[1]

In a parliamentary system, newly elected members of parliament choose a prime minister and a cabinet ("the government"). Typically the government stays in power for a fixed term; however, a no-confidence vote from the parliament or a decision by the prime minister can lead to early elections. Suppose we want to understand the duration of governments in a parliamentary democracy with five-year terms – that is, elections must be held at least every five years, but might be held earlier (in the case of a no-confidence vote or a prime ministerial decision to dissolve the government).

1a) If we can measure the government duration in infinitely small units of time (ie, continuously), what is the sample space for this experiment?

The sample space, if continuous, must be greater than 0 (more than no time in office) and less than or equal to five (it can technically last five years). So it would look like: $\Omega = (0, 5]$.

1b) The random variable $X(\omega)$ is the amount of time in years or fractions of years) between the last election and the calling of the next election. Suppose we know that $X(\omega)$ has the probability density function:

$$f(x) = \begin{cases} kx^3 & 0 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant. What is k (hint: remember the rules of pdfs)?

In this case, k = $\frac{4}{625}$. When the highest x-value gets plugged into the CDF, the result should be 1. Our integral will thus look like: $\int kx^3 \, dx$. We can integrate that to equal $k\frac{x^4}{4}$. We know that since this has to equal 1 at the highest value (5), we can do $k\frac{5^4}{4} = 1$; $625k = 4$; $k = \frac{4}{625}$.

1c) What is the CDF of X?

We know that the CDF can be defined as $F(x) = \int_{-\infty}^{\infty} f(x) \, dx$. Thus, here we can say that for the first interval we have:

$\int_0^x \frac{4}{625}x^3 \, dx$. This equals $\frac{1}{625}x^4\Big|_0^x$, or just $\frac{1}{625}x^4$.

For the remainder, we can say the following:

$F(x) = \int_{-\infty}^{y} 0 \, dx$

Also, we already know that F(5) = 1, so F(x) = 1 for x > 5.

[1]For code, see my Github at https://github.com/dfshapir/ps2.

1d) What is the probability that the government remains in power for exactly 3 years? Why?

Well, technically, it should be 0. This is continuous, not discrete, so really the probability of the event happening at one specific point is 0. If we show it mathematically, we would take the integral from value "3" to value "3", which would end up just being 0.

1e) What is the probability that the government remains in power between 2 and 4 years?

Here, we actually need to map from x = 2 to x = 4. So this would look like:

$\int_2^4 \frac{4}{625}x^3\,dx$. Then, we get $\frac{1}{625}(4)^4 - \frac{1}{625}(2)^4$. This equates to $\frac{256}{625} - \frac{16}{625} = \frac{240}{625}$ or about 38.4%.

1f) What is the probability that the government remains in power for < 1 or > 4 years?

Here, I will do two integrals, one from 0 to 1 and the other from 4 to 5. So:

$\int_0^1 \frac{4}{625}x^3\,dx$ and $\int_4^5 \frac{4}{625}x^3\,dx$.

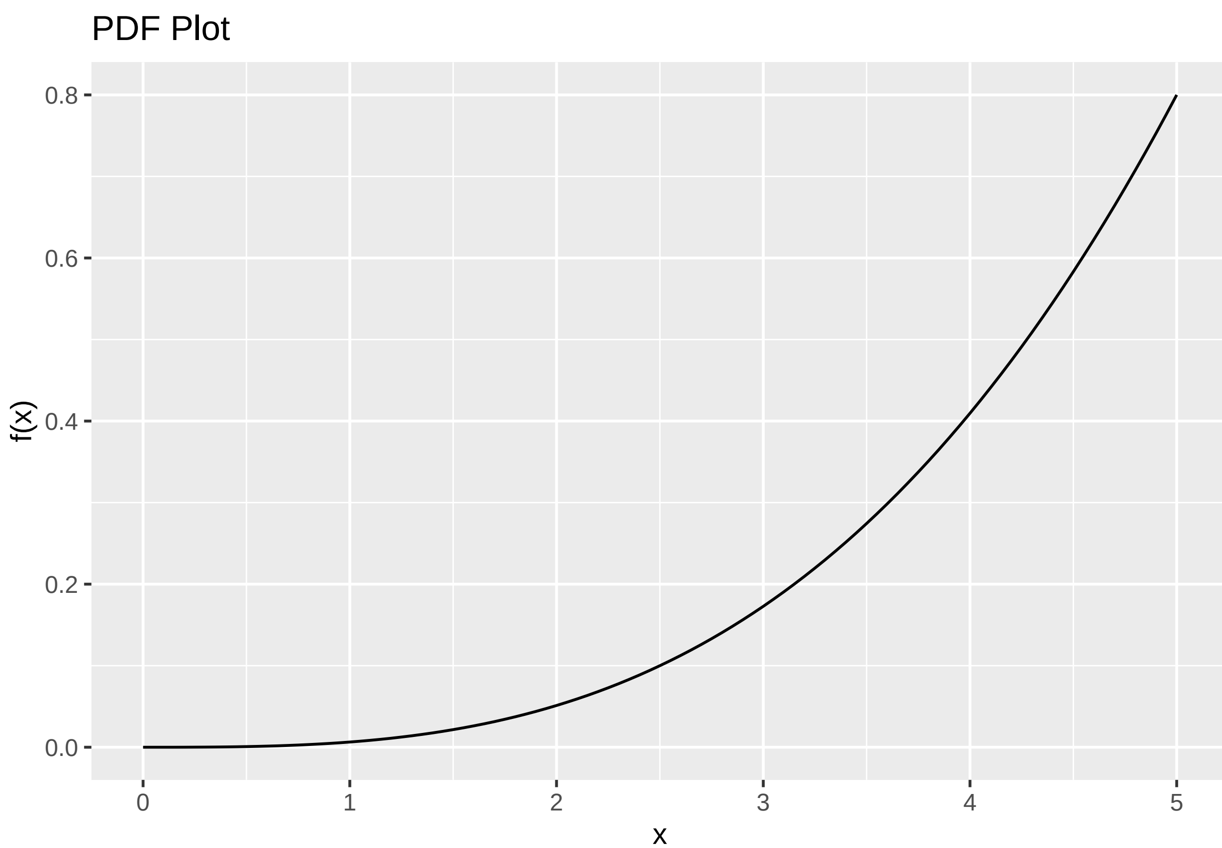
For the first integral, from this, we get $\frac{1}{625}(1)^4 - \frac{1}{625}(0)^4$. This equates to $\frac{1}{625} - 0 = \frac{1}{625}$ or about .16%.

For the second integral, from this, we get $\frac{1}{625}(5)^4 - \frac{1}{625}(4)^4$. This equates to $1 - \frac{256}{625} = \frac{369}{625}$ or about 59.04%.

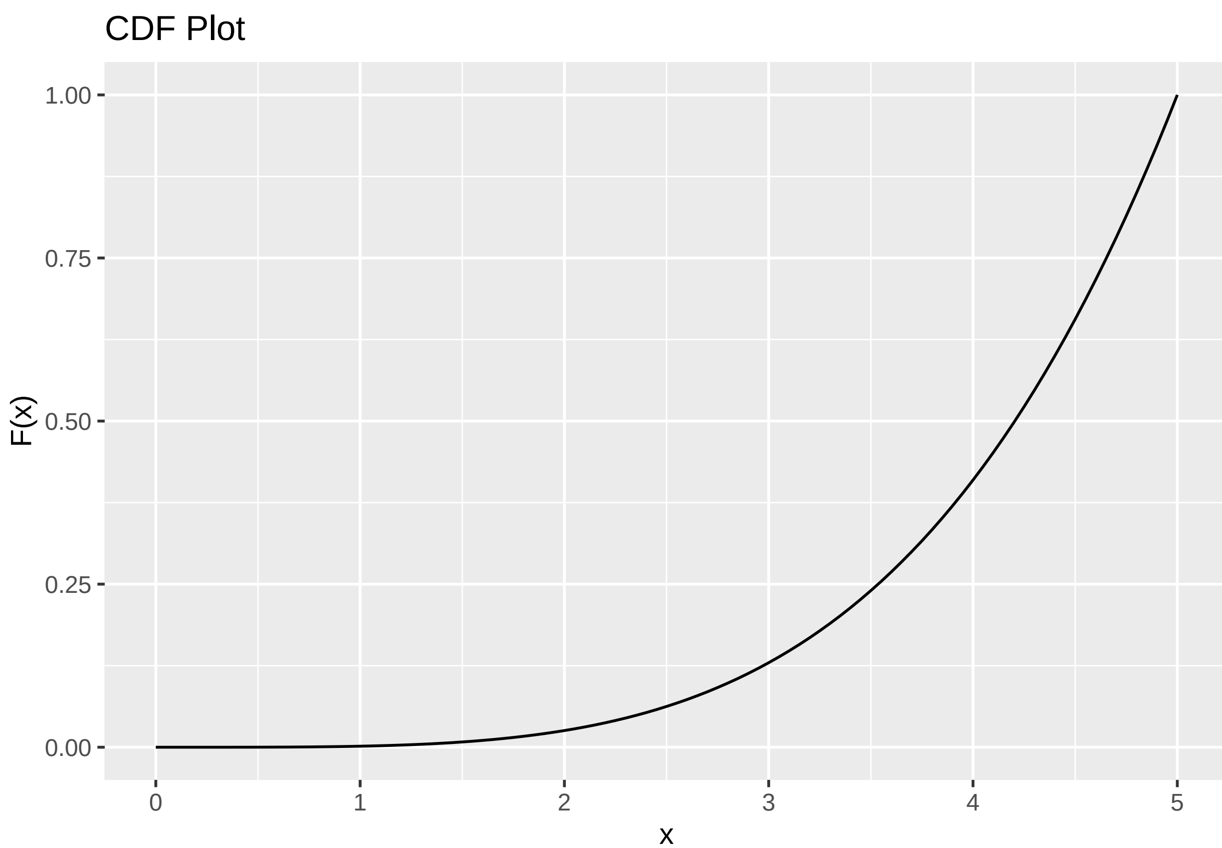
So, in total, we add .16% to 59.04, and we get 59.2%.

1g) Plot the pdf and cdf in ggplot2, with axis labels. Describe, in words, what your findings mean for the survival of parliamentary cabinets.

```
# Set limits

x <- 0:5

df <- data.frame(x)

# Create function and plot for pdf

pdfplot <- ggplot(df, aes(x)) +
  stat_function(fun = function (x) (4/625) * x ^ 3) +
  labs(title = "PDF Plot",
       x = "x",
       y = "f(x)")

pdfplot
```




```r
# Create function and plot for cdf

cdfplot <- ggplot(df, aes(x))+
stat_function(fun = function (x) (1/625) * x ^ 4) +
  labs(title = "CDF Plot",
       x = "x",
       y = "F(x)")


cdfplot
```



Question 2 Background:

In a donation experiment, two participants ($P1$ and $P2$) can give up to $5 to the other player, in $1 dollar increments. Let's say we expect that participants donate randomly between 0 and 5, and let's define the random variables X, Y, Z as follows:

X=The amount that $P1$ gives

Y=The number of participants that give $5.

$$Z = \begin{cases} 0 & \text{if the two give the same amount} \\ 1 & \text{if P2 gives more than P1} \\ 2 & \text{if P1 gives more than P2} \end{cases}$$

For example, if $P1$ gives $5 and $P2$ gives $3, then $X = 5$, $Y = 1$, and $Z = 2$.

First, let's make a table showing all possible outcomes. We get a 6x6 table of (x, y, z) values:

```
P1_0 <- c(("(0, 0, 0)"), ("(0, 0, 1)"), ("(0, 0, 1)"), ("(0, 0, 1)"), ("(0, 0, 1)"), ("(0, 1, 1)"))
P1_1 <- c(("(1, 0, 2)"), ("(1, 0, 0)"), ("(1, 0, 1)"), ("(1, 0, 1)"), ("(1, 0, 1)"), ("(1, 1, 1)"))
P1_2 <- c(("(2, 0, 2)"), ("(2, 0, 2)"), ("(2, 0, 0)"), ("(2, 0, 1)"), ("(2, 0, 1)"), ("(2, 1, 1)"))
P1_3 <- c(("(3, 0, 2)"), ("(3, 0, 2)"), ("(3, 0, 2)"), ("(3, 0, 0)"), ("(3, 0, 1)"), ("(3, 1, 1)"))
P1_4 <- c(("(4, 0, 2)"), ("(4, 0, 2)"),("(4, 0, 2)"),("(4, 0, 2)"), ("(4, 0, 0)"), ("(4, 1, 1)"))
P1_5 <- c(("(5, 1, 2)"), ("(5, 1, 2)"), ("(5, 1, 2)"), ("(5, 1, 2)"), ("(5, 1, 2)"), ("(5, 2, 0)"))
```

```r
outcomes <- data.frame(P1_0, P1_1, P1_2, P1_3, P1_4, P1_5)
rownames(outcomes) <- c("P2_0", "P2_1", "P2_2", "P2_3", "P2_4", "P2_5")

outcomes
```

```
##           P1_0      P1_1      P1_2      P1_3      P1_4      P1_5
## P2_0 (0, 0, 0) (1, 0, 2) (2, 0, 2) (3, 0, 2) (4, 0, 2) (5, 1, 2)
## P2_1 (0, 0, 1) (1, 0, 0) (2, 0, 2) (3, 0, 2) (4, 0, 2) (5, 1, 2)
## P2_2 (0, 0, 1) (1, 0, 1) (2, 0, 0) (3, 0, 2) (4, 0, 2) (5, 1, 2)
## P2_3 (0, 0, 1) (1, 0, 1) (2, 0, 1) (3, 0, 0) (4, 0, 2) (5, 1, 2)
## P2_4 (0, 0, 1) (1, 0, 1) (2, 0, 1) (3, 0, 1) (4, 0, 0) (5, 1, 2)
## P2_5 (0, 1, 1) (1, 1, 1) (2, 1, 1) (3, 1, 1) (4, 1, 1) (5, 2, 0)
```

2a) Write down a table showing the joint probability mass function for X and Y.

Joint probability is calculated with the following equation: $P(AandB) = P(A|B) * P(B)$. We will define A as X, and B as Y, so it will look like $P(X|Y) * P(Y)$.

So, there are 3 possibilities for Y: 0, 1, and 2. There are 6 possibilities for X: 0, 1, 2, 3, 4, and 5. Luckily, I made this handy table above with all of the combinations, so we can really just look visually to give us a good sense as to all of these values. For P(Y = 0), the probability is 25/36; in 25 of our 36 instances, Y = 0. P(Y = 1) = 10/36 – only one person chooses "5" in ten instances. Finally, P(Y = 2) = 1/36 – when both people give 5 dollars.

P(A|B) for each value is a bit more complicated. Mainly, the big differences here come when X = 5. Below, in the code, I explain how I came up with my vectors.

```r
# First, for P(A = 0|B = 0): There are 25 instances when B = 0. Of those,
# five occur when X = 0. Thus, P(A|B) in this instance = 5/25, or 1/5.
# 1/5 * 25/36 (P(B = 0)) = 5/36. A similar operation is done for when P(A = 0|B = 1).
# There are 10 instances where B = 1. Of those, one occurs when X = 1. Thus, P(A|B)
# in this instance = 1/10. 1/10 * 10/36 (P(B = 1)) = 1/36. Finally,
# there are no instances of Y = 2 outside of where X = 5. So our vector is:

X0 <- c((5/36), (1/36), (0))

# The next four vectors (X = 1, X = 2, X = 3, X = 4) are all the same.

X1 <- c((5/36), (1/36), (0))
X2 <- c((5/36), (1/36), (0))
X3 <- c((5/36), (1/36), (0))
X4 <- c((5/36), (1/36), (0))

# When X = 5, there are 0 of 25 instances of Y = 0, 5 of 10 instances
# of Y = 1, and 1/1 instance of Y = 2. So, combining them by their respective
# P(Y) we get:

X5 <- c((0), (5/36), (1/36))

twoa <- data.frame(X0, X1, X2, X3, X4, X5)

rownames(twoa) <- c("Y0", "Y1", "Y2")

twoa
```

```
##            X0         X1         X2         X3         X4         X5
## Y0 0.13888889 0.13888889 0.13888889 0.13888889 0.13888889 0.00000000
## Y1 0.02777778 0.02777778 0.02777778 0.02777778 0.02777778 0.13888889
## Y2 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.02777778
```

2b, 2c) What is the marginal probability of Y? What is the marginal probability of X?

The marginal probability $P_Y(y)$ can be described as $\sum_x Pr(X = x, Y = y)$

The marginal probability $P_X(x)$ can be described as $\sum_y Pr(Y = y, X = x)$.

So we simply need to add across or down for each value of X or Y. First, I add a column to express the marginal probability of Y:

```
twoa <- twoa %>%
  rowwise %>%
  mutate("Xmarginal" = sum(across(X0:X5)))
```

Next, I create a new vector showing the totals for X and use rbind() to bind it together with the broader dataset.

```
# Used an extra rep of 1/6 as a placeholder -- will remove at end.

df <- rep(1/6, 7)
dg <- c("X0", "X1", "X2", "X3", "X4", "X5", "Xmarginal")

data <- data.frame(df, dg) %>%
  pivot_wider(names_from = dg, values_from = df)

twoupdate <- rbind(twoa, data)

twoupdate <- data.frame(twoupdate)

rownames(twoupdate) <- c("Y0", "Y1", "Y2", "Ymarginal")

# As mentioned, I set the bottom right corner to 1. -- rows and columns
# add up to this number.

twoupdate[4,7] = 1
```

2d) Write down a table showing the joint probability mass function for X and Z.

Here we will perform a similar procedure as in 2a. Joint probability is calculated with the following equation: $P(AandB) = P(A|B) * P(B)$. We will define A as X, and B as Z, so it will look like $P(X|Z) * P(Z)$.

There are 3 possibilities for Z: 0, 1, and 2. There are 6 possibilities for X: 0, 1, 2, 3, 4, and 5. Going off of the table again, we can just look visually to give us a good sense as to all of these values. For P(Z = 0), the probability is 6/36 or 1/6; in 6 of our 36 instances, Z = 0 (P2 = P1). P(Z = 1) = 15/36 – P2 > P1 in 15 instances. Finally, P(Z = 2) = 15/36 – when P1 > P2.

Calculating P(A|B) (otherwise known as P(X|Z)) will be a bit more difficult; more difficult than 2a. Below, in the code, I explain how I came up with my first vector; the remaining ones follow the same logic but are not explained in detail.

```r
# First, for P(A = 0|B = 0): There are 6 instances when B = 0. Of those,
# one occurs when X = 0. Thus, P(A|B) in this instance = 1/6. 1/6 * 1/6 (P(B = 0))
# = 1/36. A similar operation is done for when P(A = 0|B = 1).
# There are 15 instances where B = 1. Of those, five occur when X = 1. Thus, P(A|B)
# in this instance = 5/15, or 1/3. 1/3 * 15/36 (P(B = 1)) = 5/36. Finally,
# there are no instances of Y = 2 where X = 0. So our vector is:

X_0 <- c(1/36, 5/36, 0)

# Apart from Z = 0, Z values differ a fair amount. The next vector will look
# like:

X_1 <- c(1/36, 4/36, 1/36)

# And so on:

X_2 <- c(1/36, 3/36, 2/36)

X_3 <- c(1/36, 2/36, 3/36)

X_4 <- c(1/36, 1/36, 4/36)

X_5 <- c(1/36, 0, 5/36)

twod <- data.frame(X_0, X_1, X_2, X_3, X_4, X_5)

rownames(twod) <- c("Z0", "Z1", "Z2")

twod
```

```
##           X_0        X_1        X_2        X_3        X_4        X_5
## Z0 0.02777778 0.02777778 0.02777778 0.02777778 0.02777778 0.02777778
## Z1 0.13888889 0.11111111 0.08333333 0.05555556 0.02777778 0.00000000
## Z2 0.00000000 0.02777778 0.05555556 0.08333333 0.11111111 0.13888889
```

2e) What is the marginal probability mass function for Z?

Going off of the information that we got from 2d, we can add across and get the marginal probability at each value of Z. So:

$$p_Z(z) = \begin{cases} 1/6 & \text{if Z} = 0 \\ 15/36 & \text{if Z} = 1 \\ 15/36 & \text{if Z} = 2 \end{cases}$$

2f) Find the conditional probability mass function for X given $Z = 1$.

We know that the conditional probability mass function in this instance can be defined as:

$P_{X|Z}(x|z) = \frac{P_{Z,X}(z,x)}{P_Z(z)}$

or

$P_{X|Z}(x|z = 1) = \frac{P_{Z,X}(z=1,x)}{P_Z(z=1)}$

This should give us six separate probabilities – when x = 0, x = 1, x = 2, x = 3, x = 4, and x = 5. We can run the equation here for each option.

$P(x = 0|z = 1) = \frac{(\frac{5}{36})}{(\frac{15}{36})} = \frac{1}{3}$.

$P(x = 1|z = 1) = \frac{(\frac{4}{36})}{(\frac{15}{36})} = \frac{4}{15}$.

$P(x = 2|z = 1) = \frac{(\frac{3}{36})}{(\frac{15}{36})} = \frac{1}{5}$.

$P(x = 3|z = 1) = \frac{(\frac{2}{36})}{(\frac{15}{36})} = \frac{2}{15}$.

$P(x = 4|z = 1) = \frac{(\frac{1}{36})}{(\frac{15}{36})} = \frac{1}{15}$.

$P(x = 5|z = 1) = \frac{0}{(\frac{15}{36})} = 0$.

So, the function will look like:

$$p_{X|Z}(x|z = 1) = \begin{cases} 1/3 & \text{if x = 0} \\ 4/15 & \text{if x = 1} \\ 1/5 & \text{if x = 2} \\ 2/15 & \text{if x = 3} \\ 1/15 & \text{if x = 4} \\ 0 & \text{if x = 5} \end{cases}$$

Question 3 Background:

Let's replicate the donation experiment empirically, through sampling. First, create a dataset of trials p1 and p2. Then define the variables X, Y , and Z, based on the definitions above.

```
p1 <- sample(0:5, size = 1000, replace = TRUE)
p2 <- sample(0:5, size = 1000, replace = TRUE)

# I didn't set the probability, because all outcomes are equally likely.

# Next, I create empty vectors for X, Y, and Z. These will all be defined further down.

X <- c(rep(1, 1000))
Y <- c(rep(1, 1000))
Z <- c(rep(1, 1000))

newdata <- data.frame(p1, p2, X, Y, Z)

# First, X is set equal to the p1 value.

newdata <- newdata %>%
  mutate(X = p1)

# Second, I run a loop first defining where both p1 and p2 are 5 as 2, then where either p1 or p2 are 5

for(i in 1:1000) {
  if(newdata$p1[i] == 5 & newdata$p2[i] == 5) {newdata$Y[i] <- 2}
  else if(newdata$p1[i] == 5 | newdata$p2[i] == 5) {
```

```r
    newdata$Y[i] <- 1}
  else{newdata$Y[i] <- 0}
}

# Now, it's time to define Z. This is a much simpler for loop.

for(i in 1:1000) {
  if(newdata$p1[i] == newdata$p2[i]) {newdata$Z[i] <- 0}
  else if(newdata$p1[i] < newdata$p2[i]) {newdata$Z[i] <- 1}
  else(newdata$Z[i] <- 2)
}
```

3a) Replicate the joint probability mass function for X and Y using the table() command in R. How closely do the results align?

```r
prop.table(table(newdata$Y, newdata$X))
```

```
##
##         0     1     2     3     4     5
##   0 0.135 0.123 0.155 0.146 0.129 0.000
##   1 0.028 0.029 0.026 0.027 0.033 0.136
##   2 0.000 0.000 0.000 0.000 0.000 0.033
```

```r
# First, printed it for the answer, then saved it so I could use it for later

table <- prop.table(table(newdata$Y, newdata$X))
```

The results are not identical, but they are quite close to the ones we observed in 2a.

3b) Using a for loop, find the marginal probability for Y. based on your simulation data.

```r
tabledata <- data.frame(table) %>%
  pivot_wider(names_from = Var2, values_from = Freq) %>%
  mutate("Y_Marg" = NA)

# Above, I put in an empty column "Y_Marg" as NA. This way, I can manage to
# do this with a for loop below.I personally would have just preferred the
# mutate() command, but this works.

for(i in 1:nrow(tabledata)) {
  tabledata$Y_Marg[i] <- sum(tabledata[i,2:7])
}

tabledata %>%select(Var1, Y_Marg)
```

```
## # A tibble: 3 x 2
##   Var1  Y_Marg
##   <fct>  <dbl>
```

```
## 1 0       0.688
## 2 1       0.279
## 3 2       0.033
```

3c) Using a for loop, show the conditional probability mass function for X given that $Z = 1$.

We know that the conditional probability mass function in this instance can be defined as:

$P_{X|Z}(x|z) = \frac{P_{Z,X}(z,x)}{P_Z(z)}$

or

$P_{X|Z}(x|z=1) = \frac{P_{Z,X}(z=1,x)}{P_Z(z=1)}$

So how do we write this in R?

```r
# First, let's create a constant -- the probability that Z = 1. This will be
# the denominator.

threecdata <- newdata %>%
  filter(Z == 1)

zprob <- nrow(threecdata)/nrow(newdata)

# Now, for the hard part. First, we need to use the whole data, instead of just threecdata.
# Then, create an empty vector to make the for loop easier.

newdata <- newdata %>%
  mutate("prob" = 1)

# Here, I want to select only Z values that equal 1 and classify the new column
# as the corresponding X value.

for(i in 1:1000) {
  if(newdata$Z[i] == 1) {
    newdata$prob[i] <- newdata$X[i]
  }
  else(newdata$prob[i] <- NA)
}

# Now, I can set up probabilities. This is going to be a bit ungainly, but it'll do
# the job.

givenz0 <- newdata %>%
  filter(prob == 0) %>%
  nrow()

givenz1 <- newdata %>%
  filter(prob == 1) %>%
  nrow()

givenz2 <- newdata %>%
  filter(prob == 2) %>%
  nrow()
```

```r
givenz3 <- newdata %>%
  filter(prob == 3) %>%
  nrow()

givenz4 <- newdata %>%
  filter(prob == 4) %>%
  nrow()

givenz5 <- newdata %>%
  filter(prob == 5) %>%
  nrow()

# Below is how I set up P_{Z, X}(z = 1,x).

p0 <- givenz0/nrow(newdata)
p1 <- givenz1/nrow(newdata)
p2 <- givenz2/nrow(newdata)
p3 <- givenz3/nrow(newdata)
p4 <- givenz4/nrow(newdata)
p5 <- givenz5/nrow(newdata)

# Then, I run the equation for each value of X below.

conditional0 <- p0/zprob
conditional1 <- p1/zprob
conditional2 <- p2/zprob
conditional3 <- p3/zprob
conditional4 <- p4/zprob
conditional5 <- p5/zprob

# Here, I check if they add up correctly -- they should equal 1. They do, so my math
# checks out.

check <- conditional0 + conditional1 + conditional2 + conditional3 + conditional4 + conditional5

# Now, we can print the values to get a sense of what they look like. I will use 4 decimal places in th

conditional0
```

```
## [1] 0.3333333
```

```r
conditional1
```

```
## [1] 0.2364066
```

```r
conditional2
```

```
## [1] 0.2174941
```

```r
conditional3
```

```
## [1] 0.1347518
```

```
conditional4
```

```
## [1] 0.07801418
```

```
conditional5
```

```
## [1] 0
```

So, below, we can write the function. I use the above printed values out to four decimal places:

$$p_{X|Z}(x|z=1) = \begin{cases} .3114 & \text{if x} = 0 \\ .2749 & \text{if x} = 1 \\ .2165 & \text{if x} = 2 \\ .1241 & \text{if x} = 3 \\ .0730 & \text{if x} = 4 \\ 0 & \text{if x} = 5 \end{cases}$$

The numbers are quite close to what we got for question 2f, but not the same. This makes sense, due to some natural discrepancies due to sampling.