Problem Set 3 Solutions

Lars Lien Ankile

02.26.2020

Problem 1.

a) This just the sum of the probability for $P(X \geq 2) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7$.

b) $P(X \geq 2 | X \geq 1) = \dfrac{P(X \geq 2)}{P(X \geq 1)} = \dfrac{0.7}{0.9} = 0.778$.

c) $E[X] = 0 \cdot 0.1 + 1 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.4 = 2.3$.

```
#use r as a calculator
0*0.1 + 1*0.2 + 3*0.3 + 3*0.4
```

```
## [1] 2.3
```

d)

$$
\begin{aligned}
Var(X) &= \sum_{i=1}^{4} P(X = x_i)(x_i - \mu)^2 \\
&= 0.1 \cdot (0 - 2.3)^2 \\
&+ 0.2 \cdot (1 - 2.3)^2 \\
&+ 0.3 \cdot (2 - 2.3)^2 \\
&+ 0.4 \cdot (3 - 2.3)^2 = 1.09.
\end{aligned}
$$

```
#use r as a calculator
0.1 * (0-2.3)^2 + 0.2 * (1-2.3)^2 + 0.3 * (2-2.3)^2 + 0.4 * (3-2.3)^2
```

```
## [1] 1.09
```

Problem 2.

a) The probabilities of the different outcomes (zero wins, 1 win, 2 wins) is summarized in the below table.

$X = x$	$P(X = x)$
0	$P(\text{lose 1st race})P(\text{lose 2nd race}) = 0.8 \cdot 0.7 = 0.56$
1	$P(\text{win 1st})P(\text{lose 2nd}) + P(\text{lose 1st})P(\text{win 2nd}) = 0.2 \cdot 0.7 + 0.8 \cdot 0.3 = 0.38$
2	$P(\text{win 1st race})P(\text{win 2nd race}) = 0.2 \cdot 0.3 = 0.06$

Expected profit will be $E[p] = 0.56 \cdot 10000 + 0.38 \cdot 50000 + 0.06 \cdot 100000 - 20000 = 10600$.

Standard deviation is the square root of the variance.

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{0.56(10000 - 10600)^2 + 0.38(50000 - 10600)^2 + 0.06(100000 - 10600)^2} \\ &= 32705.35.\end{aligned}$$

```
# use r as a calculator
# Find probabilities
0.8 * 0.7
```

```
## [1] 0.56
```

```
0.2 * 0.7 + 0.8 * 0.3
```

```
## [1] 0.38
```

```
0.2 * 0.3
```

```
## [1] 0.06
```

```
# Calculate expected profit
0.56 * 10000 + 0.38 * 50000 + 0.06 * 100000 - 20000
```

```
## [1] 10600
```

```
# Calculate the standard deviation
(0.56*(10000-10600)^2 + 0.38*(50000-10600)^2 + 0.06*(100000-10600)^2)^0.5
```

```
## [1] 32705.35
```

b) It depends, but in general I would thus assume that the races are independent. If the horse does badly in the first race, it could be pure chance, but it's very likely that it is because the horse isn't that good compared to the competition, which would mean that the horse will prabably do badly in the next race too.

Problem 3.

a) To find this probability I've looked at the binomial probabilities. In this case, we have n trials, where the probability of success is 0.94 in each trial. We want at least $400 \cdot 0.96 = 384$ people to be vaccinated. Since `pbinom` calculates either $P(X \leq x) or P(X > x)$, but I want $P(X \geq x)$ I need to flip the question, and rather ask, what is the probability that 16 people, or less, are not vaccinated. According to my below R calculations that probability is 0.051, or 5.1%.

```
# Probability that any given undergraduate is vaccinated
p <- 0.94
n <- 400
# The required proportion of people who must be vaccinated
herd <- 0.96

house_is_herd_immune = pbinom((1 - herd) * n, size = n, prob = 1 - p)
house_is_herd_immune
```

```
## [1] 0.05100541
```

b) For this problem, I assume that each house is isolated from the others, so that each house has an equal probability of being herd immune, independent from the other houses immunity status. Given that assumption, the probability that all houses have herd immunity is just the product of all the individual her immunities, i.e. $P(\text{herd immunity in a house})^{\text{number of hosues}} = 0.05100541^12 = 3.1 \cdot 10^{-16}$.

```
number_of_houses <- 12

house_is_herd_immune^number_of_houses
```

```
## [1] 3.100239e-16
```

c) Just because the national rate of vaccination is 94% it doesn't necessarily mean that that directly applies to Harvard undergraduates. Hopefully, the specific number for undergrads here is higher, but it could also be lower. Even if it's a little lower, lets say 92%, the chance that a house achieves herd immunity drops to 0.00097, or 0.097%. Also, without knowing too much about herd immunity, I'm guessing the herd needs a certain size and level of isolation from other herds to work that I'd think that a Harvard house won't necessarily achieve.

Problem 4.

a) To find the probability of the stock yielding negative returns, I'll use `pnorm` to find the integral under the probability density function of the stock (defined by the mean and standard deviation), from negative infinity to 0, because that's all the values for which the stock's yields are negative. From the below calculation we see the probability is 0.328, or 32.8%.

```
pnorm(0, mean = 0.147, sd = 0.33)
```

```
## [1] 0.3279957
```

b) This is very similar to the calculation in a), only that we now look at the area under the normal distribution that is to the right of a point, not to the left, thus the `lower.tail = FALSE`-argument. From running the below command we find that the probability of this stock yielding more than 50% returns in any given year is 0.142, or 14.2%.

```
return_over_50 <- pnorm(0.5, mean = 0.147, sd = 0.33, lower.tail = FALSE)
return_over_50
```

```
## [1] 0.1423779
```

c) In this case we need to look at the area under the curve between two points, and therefore we subtract the probability of the stock yielding less than 25% from the probability that the stock will yield less than 75%, and we're left with the probability that it will yield somewhere within that range. From running the below code we get that this probability is 0.344, or 34.4%.

```
pnorm(0.75, mean = 0.147, sd = 0.33) - pnorm(0.25, mean = 0.147, sd = 0.33)
```

```
## [1] 0.3436448
```

d) To find the return that marks off the 10th quantile in this distribution I'll use the `qnorm`-function as in the code chunk below. When used with the arguments for the quantile I want, the mean, and SD, I get out that -0.276, or -27.6%, returns and below are the 10% worst returns for this stock.

```
qnorm(0.1, mean = 0.147, sd = 0.33)
```

```
## [1] -0.275912
```

e) For this calculation I assume that the returns for a specific year is independent from the previous years and won't impact the following years.

Here, I've used the binomial function `dbinom`, with 4 as the amount of successes we want, 10 as the amount of trials, and the probability equals the probability we found in b) for the probability of success. From the below calculation we can see that we get a probability of 0.034, or 3.4%, for the stock yielding more than 50% in exactly 4 our of the next 10 years.

The below result requires the yield from year to year to be independent, which is very unlikely. Therefore, the result must be viewed with skepticism.

```
dbinom(x = 4, size = 10, prob = return_over_50)
```

```
## [1] 0.03433722
```

Problem 5.

a) To solve this, I essentially split the calculation into 2 mutually exclusive calculations. First, what is the probability of a return over 5%, given that they beat the expectation, multiplied with the probability of beating the expectations. Second, what is the same probability, given that they do not beat the expectation, multiplied with the probability of not beating the expectations. Lastly, I sum these probabilities. For each of the parts I used `pnorm` with `lower.tail = FALSE` to sum from 5% yield and up (to the right of 5%). The resulting probability comes out to be 0.657, or 65.7%, chance of yielding more than 5%.

```
# Define and assign the variables used in the calculation
# Probability of beating the earnings expectations
p_beat <- 0.75

# The mean and SD of the probability distribution over returns
# if the expectation is beat
beat_mean <- 0.1
beat_sd <- 0.05

# The mean and SD of the probability distribution over returns
# if the expectation is not beat
fail_mean <- -0.05
fail_sd <- 0.08

# The yield we would like to know if the stock will be greater than
yield1 <- 0.05

# Probability that the stock will have a return greater than 5%
(p_beat * pnorm(yield1, mean = beat_mean, sd = beat_sd, lower.tail = FALSE)
 + (1 - p_beat) * pnorm(yield1,
                        mean = fail_mean,
                        sd = fail_sd,
                        lower.tail = FALSE))
```

```
## [1] 0.657421
```

b) Just like in a), I've split the calculation into two cases. This time, however, I've looked at the portion of the distributions to the left of the -5%-mark by using the argument `lower.tail = TRUE` (implicitly). This time the probability comes out to be 0.126, or 12.6%. This is a much smaller probability, which makes sense, because the positive outcome is much mor likely than the negative outcome, and the probability of getting a positive yield is very high when they're beating expectations, and low for having a negative yield (mean is positive and -5% is three SDs from the mean).

```
# The yield we would like to know if the stock will be less than
yield2 <- -0.05

(p_beat * pnorm(yield2, mean = beat_mean, sd = beat_sd)
 + (1 - p_beat) * pnorm(yield2, mean = fail_mean, sd = fail_sd))
```

```
## [1] 0.1260124
```

Problem 6.

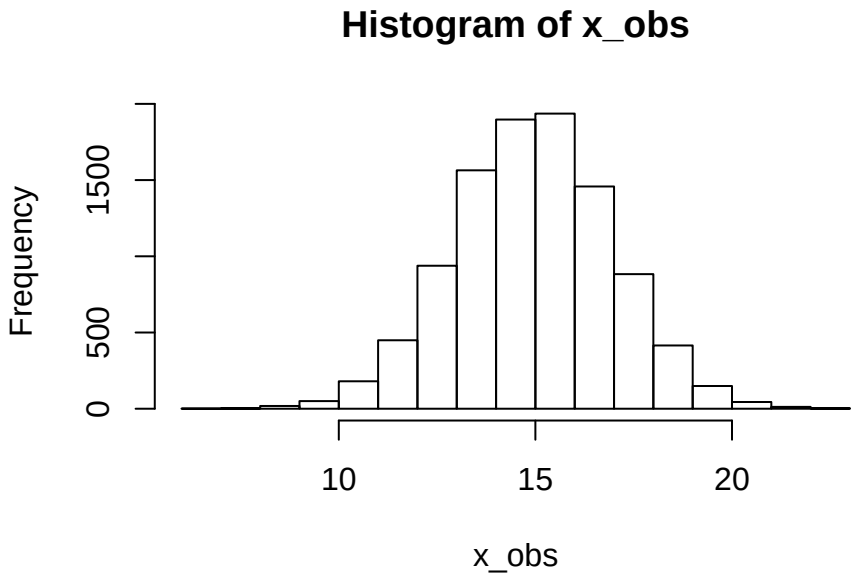
a) Below I've generated the specified 10,000 observations for the two distributions. From looking at the two histograms created, both look very nicely normally distributed. Also, we see that the first appears to be centered around 15, while the last is centered around 10. Lastly, the last histogram is much more spread out than the first, which makes sense seeing that Y has a larger SD than X.

We can also see from the numerical summaries, we see that both are pretty close to the specified values for mean and SD, but not quite. After a little testing with higher numbers for sample size, I see that the values converge to the true values.

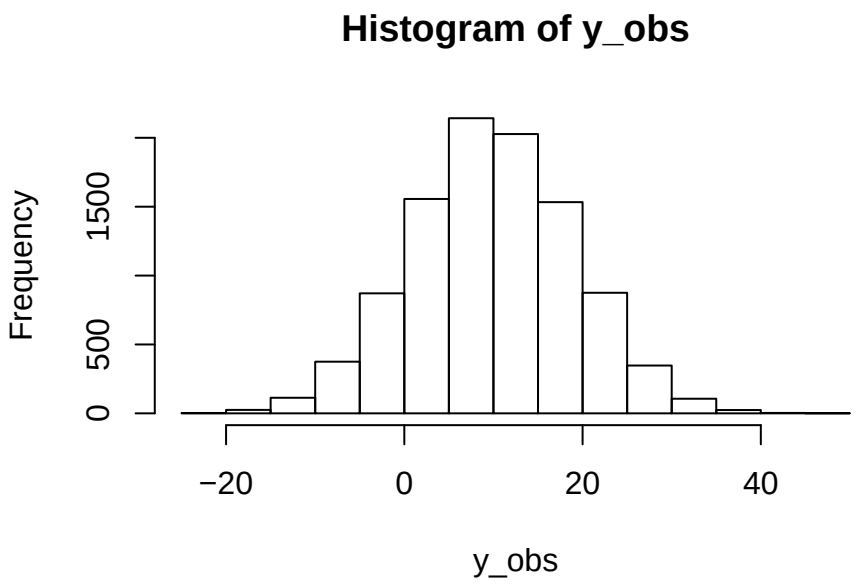
```
#set seed
set.seed(2019)

#generate observations
x_obs <- rnorm(10000, mean = 15, sd = 2)
y_obs <- rnorm(10000, mean = 10, sd = 9)

#graphical summaries
hist(x_obs)
```




```
hist(y_obs)
```




```
#numerical summaries
sprintf("Distribution X: Mean: %.3f, SD: %.3f.", mean(x_obs), sd(x_obs))
```

```
## [1] "Distribution X: Mean: 14.935, SD: 2.001."
```

```
sprintf("Distribution Y: Mean: %.3f, SD: %.3f.", mean(y_obs), sd(y_obs))
```

```
## [1] "Distribution Y: Mean: 9.882, SD: 9.057."
```

b) i. The expectation of the sum of these simulated observations is $E[X+Y] = 24.81754$. Likewise, the variance of the sum of these sumilated observations is $Var(X+Y) = 86.10896$.

ii. Below I've plotted a histogram of the sum of the two random samples. As we see, the mean appears to be roughly at somewhere between 20 and 30, which is in accordance with the calculations above. It is also relatively normally distributed around the mean, with a SD that looks like it could be arund 10. Again, this makes sense since $SD(X+Y) = \sqrt{Var(X+Y)} = \sqrt{86.10896} = 9.27949$.

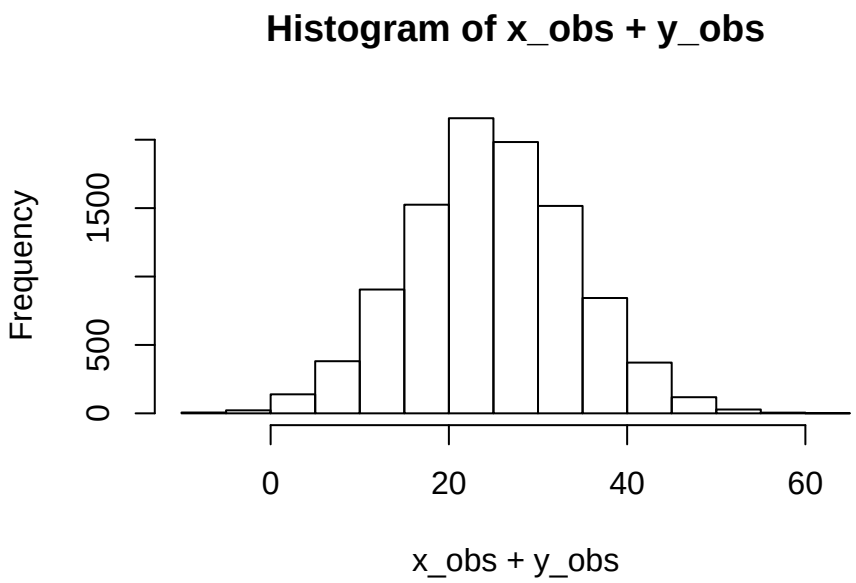
```
#part i
mean(x_obs + y_obs)
```

```
## [1] 24.81754
```

```
var(x_obs + y_obs)
```

```
## [1] 86.10896
```

```
#part ii
hist(x_obs + y_obs)
```



c) Doing essentially the same as in b), I get $E[X-Y] = 5.053127$ and $Var(X-Y) = 85.96098$. See below for the R-code.

```
mean(x_obs - y_obs)
```

```
## [1] 5.053127
```

```
var(x_obs - y_obs)
```

```
## [1] 85.96098
```

d) From theory, we know that the expectation of the of the sum of two random variables is the sum of their expectations, i.e. $E[X+Y] = E[X] + E[Y] = 15 + 10 = 25$, and likewise with the difference (just regarding the second variable as a negative and compute from the sum), $E[X-Y] = E[X] + E[-Y] = E[X] - E[Y] = 15 - 10 = 5$. Variance is in general a bit more complicated, but since the variables are independent the covariance term disappears, and thusfollows this formula $Var(X+Y) = Var(X) + Var(Y) = SD(X)^2 + SD(Y)^2 = 4 + 81 = 85$, and $Var(X-Y) = SD(X)^2 + SD(Y)^2 = 4 + 81 = 85$.

```
#use r as a calculator
2^2 + 9^2
```

```
## [1] 85
```

e) In the below R-code, I've created a new vector containing whether any given X-observation was smaller than the corresponding Y-observation. To fine the estimate of $P(X < Y)$, I count the occurrences of `TRUE` in the vector and devide by the total number of elements. The result is an estimated probability of 0.294, or 29.4%.

```
x_less_than_y <- x_obs < y_obs

# Calculate the simulated probability
length(x_less_than_y[x_less_than_y == TRUE]) / length(x_less_than_y)
```

```
## [1] 0.2935
```

f) These are equivalent: $P(X < Y)$ and $P(X - Y < 0)$, which is a question R can answer with `pnorm` with mean equal to 5 and SD equal to $\sigma = \sqrt{85} = 9.219544$. The calculation yields a probabilit of X being less than Y of 0.294, or 29.4%. From this we see that the estimate is extremely close, and is the same to 3 decimal places. This is quite cool and shows that the pseudo-random number generator R uses is pretty good.

```
s_deviation <- 85^0.5
new_mean <- 15 - 10

# Calculate the probability
pnorm(0, mean = new_mean, sd = s_deviation)
```

```
## [1] 0.2937969
```

Problem 7.

```
# Define the probabilities
probs <- matrix(c(0.05, 0.23, 0.20, 0.30, 0.20, 0.02), ncol = 3, byrow = TRUE)
rownames(probs) <- c(100, 250)
colnames(probs) <- c(0, 100, 300)

# Define the values of X and Y, respectively
x <- c(100, 250)
y <- c(0, 100, 300)

# Print the table for reference
addmargins(probs)
```

```
##         0  100  300  Sum
## 100 0.05 0.23 0.20 0.48
## 250 0.30 0.20 0.02 0.52
## Sum 0.35 0.43 0.22 1.00
```

a) From the below code we see that the mean of X is 178, while the standard deviation of X is 74.9.

```
# Calculate the marginal probability of X by using the
# apply function over each row in the table
px <- apply(probs, 1, sum)

# Calculate and print the expectation/mean of X
ex <- sum(px * x)
ex
```

```
## [1] 178
```

```
# Calculate the expection of X squared for use in calculation of SD(X)
ex2 <- sum(px * x^2)

# Standard deviation of X
x_sd <- sqrt(ex2 - ex^2)
x_sd
```

```
## [1] 74.93998
```

b) From the below code we see that the mean of Y is 109, while the standard deviation of Y is 110.5.

```
# Calculate the marginal probability of Y by using the apply
# function over each column in the table
py <- apply(probs, 2, sum)

# Calculate and print the expectation/mean of Y
ey <- sum(py * y)
ey
```

```
## [1] 109
```

```
# Calculate the expection of Y squared for use in calculation of SD(Y)
ey2 <- sum(py * y^2)

# Standard deviation of Y
y_sd <- sqrt(ey2 - ey^2)
y_sd
```

```
## [1] 110.5396
```

c) Below is the calculation of the covariance of X and Y. The result came out to be -4602. This is a negative number, which means that they tend to move in the opposite direction of each other.

```
#use r as a calculator
# Calculate the expected value of the product of X and Y
# Initialize the variable to hold the result
e_xy <- 0

# For-loop that runs over every entry in the join distribution table
# and multiplies the probability P(X=x_i, Y=y_j) with x_i and y_j
# and adds the result to the expectation
for (i in 1:2) {
  for (j in 1:3) {
    e_xy <- e_xy + probs[i, j] * x[i] * y[j]
  }
}

# Here we calculate the actual covariance between X and Y
# according to the formula for covariance
cov_xy <- e_xy - ex*ey
cov_xy
```

```
## [1] -4602
```

d) The to get correlation from covariance one can just devide ou the standard devation of each variable, as I've done below. This gives a correlation between X and Y of -0.27, i.e. they tend to move weakly different directions.

```
# Correlation is covariance devided by the product of the two SD's
corr_xy <- cov_xy / (x_sd * y_sd)
corr_xy
```

```
## [1] -0.5555399
```

e) Below is the calculation of $E[X + Y]$. It is just the sum of the individual expectations and came out to be 287.

```r
mean_x_plus_y <- ex + ey
mean_x_plus_y
```

```
## [1] 287
```

f) The covariance of a sum of variables is a little more tricky and is the square of one plus the square of the other, plus two times the covariance. This is implemented below and comes out to be 47336.57 for this X and Y.

```r
var_x_plus_y <- x_sd^2 + y_sd^2 + 2 * cov_xy
var_x_plus_y
```

```
## [1] 8631
```

g) They are not independent because their covariance is non-zero. We can also see that the marginal probaility is different from the joint probabilities, which means that e.g. the probability of X being 100 will change depending on whether we condition on a specific value of Y or not.

h) Below I've calculated $E[Y|X = 250]$, which came out to be 50.

```r
# This is the probability distribution over Y, conditioned on X being 250
y_given_x_250 <- probs[2, ] / px[2]

# This calculated the conditional expected value of Y conditioned on X beign 250
sum(y * y_given_x_250)
```

```
## [1] 50
```

Problem 8.

In the following R-code chunk I just set up my joint probability distribution table for easier calculations later.

```
# Define the join probability distribution table
p <- matrix(c(0.08, 0.02, 0.36, 0.24, 0.10, 0.10, 0.02, 0.08),
            ncol = 2,
            byrow = TRUE)
p
```

```
##      [,1] [,2]
## [1,] 0.08 0.02
## [2,] 0.36 0.24
## [3,] 0.10 0.10
## [4,] 0.02 0.08
```

```
# Define N and F
n <- c(0:1)
f <- c(1:4)

# Marginal probability for N
pn <- apply(p, 2, sum)
```

a) To find $P(N = 1, F = 2)$ I can go to the second row and second column of the distribution table and find the value 0.24, or 24%. This is done in R below. This means that there's a 24% chance that any given customer has bought 2-10 things in the last year and will buy something again.

```
# P(N=1, F=2)
p[2, 2]
```

```
## [1] 0.24
```

b) The probability $P_N(1) = 0.44$. This means that within the group of all previous customers, i.e. the people having made at least one purchase within the last year, there's a 44% chance they will make a new purchase. Below is the calculation I did in R to find the value. Essentially I just got the second element of the marginal probabilities for N.

```
# Marginal probability for P(N=1)
pn[2]
```

```
## [1] 0.44
```

c) Below is a code chunk calculating the marginal probabilites of F. I've used the `apply`-function which applies the specified function, `sum` along the axis given, here 1, i.e. it'll sum over each row. We see that the marginal probabilities of $F = 1, 2, 3, 4$, is 0.1, 0.6, 0.2, and 0.1, respectively.

```
# Marginal probability for F
pf <- apply(p, 1, sum)
pf
```

```
## [1] 0.1 0.6 0.2 0.1
```

d) Below is a code cunk calculating the probability distribution over F, given that $N = 1$. What I did is that I took the second column of the distribution table and devided each element of the column by the marginal probability for $P(N=1). This shows us that out of the people who will buy something again, 4.5% of them bought 1 item within the last year, while 54.5% bought 1-10 items, 22.7% bought 11-20 items, and 18.2% bought more than 20 items within the last year.

```
# Conditional distribution of F given that N=1
p[, 2] / pn[2]
```

```
## [1] 0.04545455 0.54545455 0.22727273 0.18181818
```

e) The calculation below is essentially the same as the above, except that we've flipped what variable we're looking at. Here I've taken the 4th row of the table, and divided each element by the marginal probability of $P(F = 4)$. This gives me the percentage of people who have bought more than 20 items within the last year who will by something again. 80% will and 20% will not.

```
# Conditional distribution of N given that F=4
n_given_f_4 <- p[4, ] / pf[4]
n_given_f_4
```

```
## [1] 0.2 0.8
```

f) The expectation $E[N|F = 4] = 0 \cdot 0.20 + 1 \cdot 0.8 = 0.8$. This calculation I've done below too, just using the table defined at the top. This falls in line with the explaination above, and means that we expect 80% of the people who've bought more than 20 items within the last year to buy something again.

```
# E(N|F=4)
sum(n * n_given_f_4)
```

```
## [1] 0.8
```

g) i. I don't think N and F would be independent, because I think that people who have bought a lot in the past will be more likely to buy more in the future. Therefore, increased buying frequency will most likely mean higher probability of sale in the future, and the variables are not independent.

ii. They're not independent because we can e.g. see that when we condition on $PF = 4$, the probability of $N = 1$ changes. This means that the variables are dependent on each other.

iii. We see that the customers in frequency category 4 have a conditioned probability of making a purchase of 80%, while that same number is only 20%, 40%, and 50% for categories F=1,2, and 3, respectively. Therefore, the marketing department could focus their energy on people in frequency category 1, because we see that once people go from F=1 to F=2 (essentially one sale) their conditional probability of making a purchase jumps from 20% to 40%, which is quite a lot in my opinion.