Problem Set 3

Daniel Shapiro

9/13/2022

Question 1 Background:

Suppose that X and Y are identical and independently distributed (i.i.d.) random variables distributed $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$, respectively. Note that σ_X represents the standard deviation of X, and σ_X^2 the variance. Find the following, expressed in terms of μ and σ:

1a) $E(7X - 6Y + 12)$

Additivity and homogeneity are two properties of expectations. Thus, we can write this expression as $7E[X] - 6E[Y] + 12$, or $7\mu_X - 6\mu_Y + 12$.

1b) Var$(X + 5Y)$

We know that $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$. Inserting a 1 and 5 as a and b, we get $\sigma_X^2 + 25\sigma_Y^2 + 10Cov(X, Y)$. $Cov(X, Y) = E[XY] - E[X]E[Y] = E[(X - \mu_X)(Y - \mu_Y)]$. So then we get: $Var(X + 5Y) = \sigma_X^2 + 25\sigma_Y^2 + 10E[(X - \mu_X)(Y - \mu_Y)]$.

Because we know that the two variables are so-called "i.i.d." variables, we know that the co-variance is 0. Thus, we can remove the last term entirely and end up with $Var(X + 5Y) = \sigma_X^2 + 25\sigma_Y^2$.

1c) $E(5X^2 - 12XY + 16Y^2)$

First, we can split these up according to properties shown in 1a. So we get $5E[X^2] - 12E[XY] + 16E[Y^2]$.

Now, we can already deal with the first and last bits of this expression. We know that $\sigma_X^2 = E[X^2] - \mu_X^2$ and $\sigma_Y^2 = E[Y^2] - \mu_Y^2$. Thus, $E[X^2] = \sigma_X^2 + \mu_X^2$ and $E[Y^2] = \sigma_Y^2 + \mu_Y^2$. So we can set this as $5\sigma_X^2 + 5\mu_X^2 - 12E[XY] + 16\sigma_Y^2 + 16\mu_Y^2$.

Finally, the question tells us that X and Y are independent random variables. So we know therefore that $E[XY] = E[X]E[Y]$. We can write this as $\mu_X\mu_Y$. bbbbbbbbbbb b Our final expression: $5\sigma_X^2 + 5\mu_X^2 - 12\mu_X\mu_Y + 16\sigma_Y^2 + 16\mu_Y^2$.

Question 2 Background:

Papers like this one have found that married couples tend to sort along ideological lines ("assortative mating"). Say we know the true distribution of ideology among heterosexual married couples along a scale of 1 (very conservative) to 4 (very liberal). The joint distribution for X (the woman's ideology) and Y (the man's ideology) is:

	X			
Y	1	2	3	4
1	.12	.05	.01	.01
2	.08	.17	.05	.02
3	.02	.01	.23	.05
4	.02	.02	.04	.10

2a) What is the expected value of X and the expected value of Y?

To find the expected value of X, we can run the formula: $E[X] = \Sigma(x)P(x)$. So we multiply the value x by each corresponding probability and add them up. So $E[X] = ((1 * .12) + (1 * .08) + ... + (4 * .10))$. The final answer is **2.45**.

To find the expected value of Y, we can run the same formula, but for Y instead of X. So we multiply the value y by each corresponding probability and add them up. So $E[Y] = ((1 * .12) + (1 * .05) + ... + (4 * .10))$. The final answer is **2.48**.

2b) What are the variances of X and of Y?

The variance of X or $(Var(X))$ can be defined as $\Sigma(x - \mu_x)^2 p(x)$. The variance of Y or $(Var(Y))$ can be defined as $\Sigma(y - \mu_y)^2 p(y)$. So we already know that $\mu_x = 2.45$ and $\mu_y = 2.48$; now, we need to find the probability of X at each point and the probability of Y at each point. To do this, we can just add up the proper rows and columns.

For X, we get: (.24 (when x = 1), .25 (when x = 2), .33 (when x = 3), and .18 (when x = 4)).

For Y, we get: (.19 (when y = 1), .32 (when y = 2), .31 (when y = 3), and .18 (when y = 4)).

Now, equation time:

$Var(X) = (1 - 2.45)^2 * .24 + (2 - 2.245)^2 * .25 + (3 - 2.45)^2 * .33 + (4 - 2.45^2 * .18)$. This ends up equaling **1.0875**.

$Var(Y) = (1 - 2.48)^2 * .19 + (2 - 2.248)^2 * .32 + (3 - 2.48)^2 * .31 + (4 - 2.48^2 * .18)$. This ends up equaling **0.9896**.

2c) What is the covariance of X and Y?

We know that the covariance of X and Y can be defined by the function:

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

First, we know $E[X]E[Y]$ – it's 2.48*2.45. This is 6.076.

To find E[XY], we need to multiply each value of Y and X by their probability, and then add everything together. Like so:

$((1 * 1 * .12) + (1 * 2 * .08) + (1 * 3 * .03) + ... + (4 * 4 * .10))$.

After a somewhat time-consuming process, we end up with 6.70.

Then, we subtract: 6.70-6.076 = **0.624**.

2d) What is the correlation between X and Y?

The function for correlation is:

$Cor[X, Y] = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}}$.

Luckily, we know all of these values already. Plugging in our numbers, we get:

$Cor[X,Y] = \frac{0.624}{\sqrt{1.0875 * 0.9896}} = \mathbf{0.6015}$.

2e) Find $E[Y|X=x]$ for $x = 1,2,3,4$. Find the probability mass function of the random variable $E[Y|X]$.

For purposes of this question, we will set E[X|Y] = Z. So for this question, we need to take each value of X=x and multiply it by the given probability value divided by the marginal probability of Y=y. For $E[Y|X=1]$, we thus get the following:

$((1 * \frac{.12}{.24}) + (2 * \frac{.08}{.24}) + (3 * \frac{.02}{.24}) + (4 * \frac{.02}{.24})) = 1.75$

For $E[Y|X=2]$, we get:

$((1 * \frac{.05}{.25}) + (2 * \frac{.17}{.25}) + (3 * \frac{.01}{.25}) + (4 * \frac{.02}{.25})) = 2$

For $E[Y|X=3]$, we get:

$((1 * \frac{.01}{.33}) + (2 * \frac{.05}{.33}) + (3 * \frac{.23}{.33}) + (4 * \frac{.04}{.33})) = 2.91$

For $E[Y|X=4]$, we get:

$((1 * \frac{.01}{.18}) + (2 * \frac{.02}{.18}) + (3 * \frac{.05}{.18}) + (4 * \frac{.10}{.18})) = 3.33$

Writing out the probability mass function is relatively easy if we've already figured out Z at x = 1,2,3,4. Below:

$$p(z) = \begin{cases} .24 & \text{if Z} = 1.75 \\ .25 & \text{if Z} = 2 \\ .33 & \text{if Z} = 2.91 \\ .18 & \text{if Z} = 3.33 \end{cases}$$

2f) Find $Var(Y|X)$.

We know:

$$Var(Y|X) = \begin{cases} Var(Y|X=1) & \text{if X} = 1 \\ Var(Y|X=2) & \text{if X} = 2 \\ Var(Y|X=3) & \text{if X} = 3 \\ Var(Y|X=4) & \text{if X} = 4 \end{cases}$$

=

$$Var(Y|X) = \begin{cases} Var(Y|X=1) & \text{w/prob of .24} \\ Var(Y|X=2) & \text{w/prob of .25} \\ Var(Y|X=3) & \text{w/prob of .33} \\ Var(Y|X=4) & \text{w/prob of .18} \end{cases}$$

$Var(Y|X=1) = E[Y^2|X=1] - E[Y|X=1]^2$

$= ((1^2 * \frac{.12}{.24}) + (2^2 * \frac{.08}{.24}) + (3^2 * \frac{.02}{.24}) + (4^2 * \frac{.02}{.24})) - 1.75^2 = \frac{47}{12} - 3.0625 = 0.854166.$

$Var(Y|X=2) = E[Y^2|X=2] - E[Y|X=2]^2$

$= ((1^2 * \frac{.05}{.25}) + (2^2 * \frac{.17}{.25}) + (3^2 * \frac{.01}{.25}) + (4^2 * \frac{.02}{.25})) - 2^2 = \frac{114}{25} - 4 = 0.56.$

$Var(Y|X=3) = E[Y^2|X=3] - E[Y|X=3]^2$

$= ((1^2 * \frac{.01}{.33}) + (2^2 * \frac{.05}{.33}) + (3^2 * \frac{.23}{.33}) + (4^2 * \frac{.04}{.33})) - 2.91^2 = \frac{292}{33} - 8.4628 = 0.3857.$

$Var(Y|X=4) = E[Y^2|X=4] - E[Y|X=4]^2$

$= ((1^2 * \frac{.01}{.18}) + (2^2 * \frac{.02}{.18}) + (3^2 * \frac{.05}{.18}) + (4^2 * \frac{.10}{.18})) - 3.33^2 = \frac{214}{18} - 11.1 = 0.78.$

So, this can be expressed as:

$$Var(Y|X) = \begin{cases} 0.854166 & \text{w/prob of .24} \\ 0.56 & \text{w/prob of .25} \\ 0.3857 & \text{w/prob of .33} \\ 0.78 & \text{w/prob of .18} \end{cases}$$

2g) Show that $E[Y] = E[E[Y|X]]$.

We already know that E[Y] = 2.48. Now, to get E[E[Y|X]], we multiply each value of of E[Y|X] (which we called Z) by its corresponding probability and add them together. We thus end up with:

$1.75 * .24 + 2 * .25 + 2.91 * .33 + 3.33 * .18.$

Added together, these equal 2.4797. While 2.4797 is *technically* .0003 off from 2.48, this is almost certainly due to my rounding throughout the previous few problems. Thus, we can say that 2.4797 is essentially the same as 2.48 here, and thereby confidently say: $2.48 = 2.48; thus, E[Y] = E[E[Y|X]]$.

2h) Find $Var(Y)$ using the law of total variance. Explain in words what the law of total variance does.

The law of total variance can be expressed as: $V[Y] = E[V[Y|X]] + V[E[Y|X]]$.

We already have V[Y|X], so here we can find E[V[Y|X]] by multiplying each value of V[Y|X] by its corresponding probability and adding them all together. So we get

$.854166 * .24 + .56 * .25 + .3857 * .33 + .78 * .18$. This equals about .613.

Next, we need to find the variance of E[Y|X], or as we define it, Z. $V[Z] = E[Z^2] - E[Z]^2$. $E[Z] = 2.48$, so $E[Z]^2 = 6.1504$. For $E[Z^2]$, we just have to do something similar as in 2g, but we square each value z in each part of the expression. So:

$1.75^2 * .24 + 2^2 * .25 + 2.91^2 * .33 + 3.33^2 * .18 = 6.529473.$

So $V[Z] = 6.529473 - 6.1504 =$ about 0.379. .379 + .613 = **about .99.** This makes sense, because I got .9896 for V[Y] in 2b.

The law of total variance essentially combines "within variance" and "between variance" to give us the complete picture. In this instance, it shows us average variance in associative mating patterns within the two gender groups and between the two gender groups.

2i) Explain, in words, what your findings mean for the relationship between the ideology of husbands and wives.

Well, for the most part, we see that the idea of "associative mating" holds true, although it's definitely not 100% black and white. There are certainly very few couples whose ideologies consist of one who is very liberal and one who is very conservative. And it's not like the data was taken from a biased sample (i.e. the participants were overwhelmingly liberal or overwhelmingly conservative); the expected values of both men

and women were very close to the center, which would be defined as 2.50. For the most part, "birds of a feather flock together."

It's definitely not 100% the case, though. Our correlation value was around .6, which definitely shows a correlation between the two variables, but it does not show a directly linear relationship. There is certainly room for variation in couples, and many couples do have differences in ideology. Obviously, a correlation value of "1" is unrealistic, but still, it certainly could have been higher.

One block of responses that I found interesting was the center bit in the table – the center 4 ((2, 3), (3, 2), (2, 2), (3, 3). It really surprised me how moderate liberals and moderate conservatives definitely tended to stay very much more with their own ((2, 2), (3, 3)) than with moderates on the other side. The percentages for (3, 2) and (2, 3) were way smaller than I would have expected. So I would be curious to see more research regarding this phenomenon.

In general, however, we can say that at least in this sample, "associative mating" definitely does have a degree of legitimacy as a concept. It's not a perfect fit, but in general, people seem to tend to group together with similarly-minded people.

Question 3 Background:

Now let's explore assortative mating in R! First, based on the probabilities in Problem 2, create a dataset of 1,000 married couples' ideologies using sample().

```
# First, I create the "women" dataset. Most of what we looked at in question 2
# was Y|X, or men|women. So it's easier to create "women" first and then base the men off
# of that because the calculations are already done.

women <- sample(1:4, size = 1000, replace = TRUE, prob = c(.24, .25, .33, .18))

# Create an empty vector for men that I can modify with a for loop.

men <- c(rep(1, 1000))

couples <- data.frame(women, men)

# For loop time. For the probabilities, I used the probability at each value (X, Y)
# and divided it by the marginal probability for (X = x) for each value.

for(i in 1:1000){
  if(couples$women[i] == 1){
    couples$men[i] <- sample(1:4, replace = TRUE, prob = c((.12/.24), (.08/.24),
                                                           (.02/.24), (.02/.24)))
  }
  else if(couples$women[i] == 2){
    couples$men[i] <- sample(1:4, replace = TRUE, prob = c((.05/.25), (.17/.25),
                                                           (.01/.25), (.02/.25)))
  }
    else if(couples$women[i] == 3){
      couples$men[i] <- sample(1:4, replace = TRUE, prob = c((.01/.33), (.05/.33),
                                                             (.23/.33), (.04/.33)))
    }
    else(couples$men[i] <- sample(1:4, replace = TRUE, prob = c((.01/.18), (.02/.18),
                                                                (.05/.18), (.10/.18))))
}
```

3a) Using the data you created, define a function that takes a value of X and Y and returns the probability that two individuals with their ideologies are married. Show the results for $\{X = 1, Y = 4\}$, $\{X = 2, Y = 2\}$, $\{X = 4, Y = 3\}$.

First, I create the function (below) then run it for the coordinates indicated. The results are show below.

```
ideology_prob <- function(x, y){
 newmen <- couples %>%
   filter(women == x) %>%
   filter(men == y)
 return <- nrow(newmen)/nrow(couples)
 return
}

ideology_prob(1, 4)
```

```
## [1] 0.023
```

```
ideology_prob(2, 2)
```

```
## [1] 0.171
```

```
ideology_prob(4, 3)
```

```
## [1] 0.039
```

3b) Now define a function that takes as an input a value of *either* X or Y and returns the conditional probability distribution of the other variable (e.g., you input a value of X and get the conditional probability distribution of Y at that value of X). Use it to find the conditional probability of Y at $X = 1$ and the conditional probability of X at $Y = 4$.

This problem presented me with some difficulties. At first, I tried to define the function as a function of x and y. However, that approach required me to input *both* variables, which defeats the purpose of the problem and makes the function un-runnable for the purposes of this question. So instead, I have the user define the input variable first ("x" or "y") and then input the value (1, 2, 3, or 4). The return will be the conditional probability distribution of the other variable given the value of the input variable, expressed as a series of four numbers. I took all probability figures from the table that we were given before Problem 2. Incidentally, note the fun message that I have R display if the user of the function defines the arguments as anything other than those values specified in the function.

```
conditional_prob <- function(input_variable, value){
  if(input_variable == "x" & value == 1){return <- c((.12/.24), (.08/.24), (.02/.24), (.02/.24))}
  else if(input_variable == "x" & value == 2){return <- c((.05/.25), (.17/.25), (.01/.25), (.02/.25))}
  else if(input_variable == "x" & value == 3){return <- c((.01/.33), (.05/.33), (.23/.33), (.04/.33))}
  else if(input_variable == "x" & value == 4){return <- c((.01/.18), (.02/.18), (.05/.18), (.1/.18))}
  else if(input_variable == "y" & value == 1){return <- c((.12/.19), (.05/.19), (.01/.19), (.01/.19))}
  else if(input_variable == "y" & value == 2){return <- c((.08/.32), (.17/.32), (.05/.32), (.02/.32))}
  else if(input_variable == "y" & value == 3){return <- c((.02/.31), (.01/.31), (.23/.31), (.05/.31))}
  else if(input_variable == "y" & value == 4){return <- c((.02/.18), (.02/.18), (.04/.18), (.10/.18))}
  else(return <- "Sadly, some relationships are not meant to be.")
```

```
return
}

conditional_prob(input_variable = "x", value = 1)
```

```
## [1] 0.50000000 0.33333333 0.08333333 0.08333333
```

```
conditional_prob(input_variable = "y", value = 4)
```

```
## [1] 0.1111111 0.1111111 0.2222222 0.5555556
```

3c) Use ggplot to create a bar chart that visualizes the relationship between X and Y (hint: the bar chart set-up in ggplot is a bit different than geom_line().

Obviously, there are a ton of different ways to "visualize the relationship" between X and Y. I think an interesting way to show it would be to show X on the X axis (1, 2, 3, 4) and then the expected value of Y|X, as we found in Question 2e. Below, I visualize this relationship.

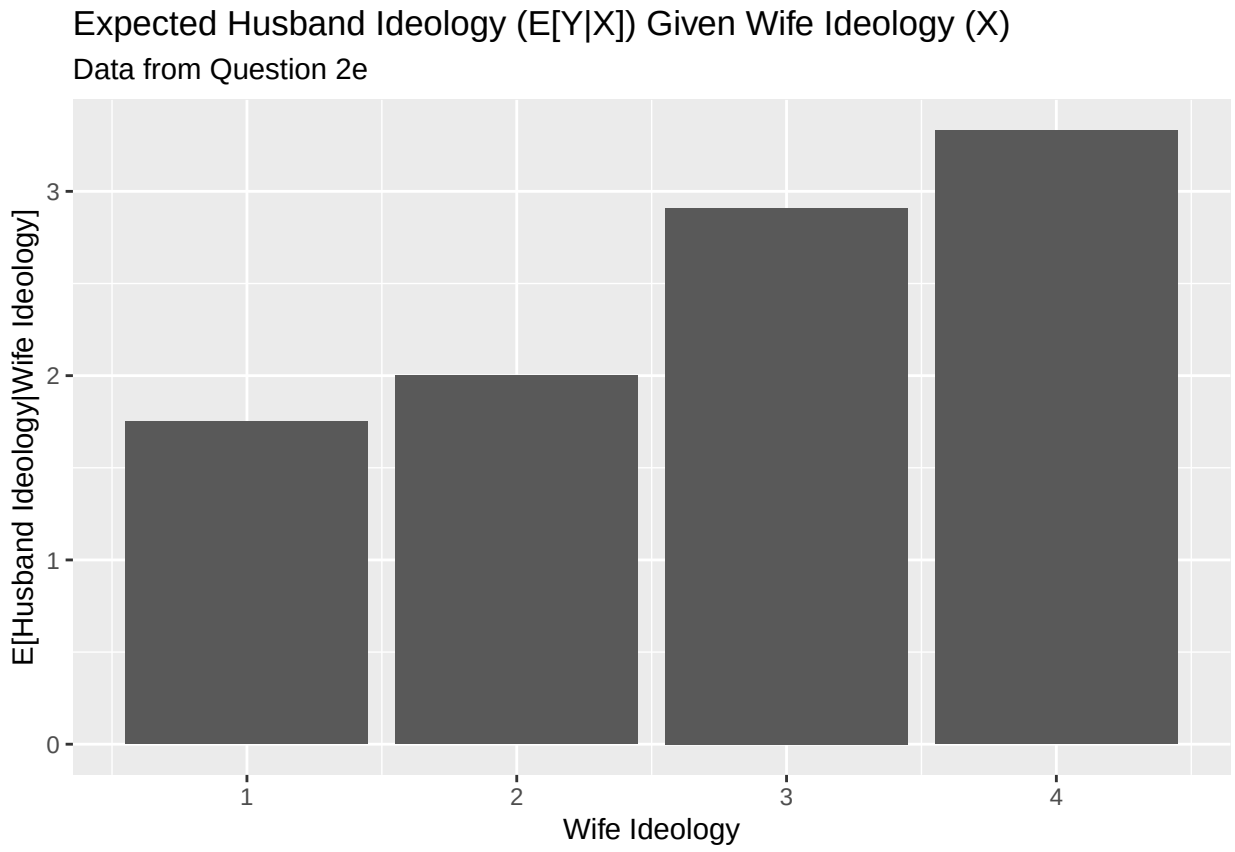
```
# First, I have to set up a dataset.

plotx <- c(1, 2, 3, 4)
ploty <- c(1.75, 2, 2.91, 3.33)

plot <- data.frame(plotx, ploty)


plot1 <- ggplot(plot, aes(x = plotx, y = ploty)) +
  geom_col() +
  labs(title = "Expected Husband Ideology (E[Y|X]) Given Wife Ideology (X)",
       subtitle = "Data from Question 2e",
       y = "E[Husband Ideology|Wife Ideology]",
       x = "Wife Ideology")

plot1
```



3d) Use your data and the functions `mean()` and `var()` to find $E(X), E(Y), Var(X)$, and $Var(Y)$. How does this compare to the findings you calculated by hand?

```
mean(couples$women)
```

```
## [1] 2.409
```

```
mean(couples$men)
```

```
## [1] 2.444
```

```
var(couples$women)
```

```
## [1] 1.124844
```

```
var(couples$men)
```

```
## [1] 1.01588
```

From the results we get here, we see that the mean values for men and women are slightly lower than what we calculated by hand in our initial table. But they're quite similar; just a bit lower.

On the variance side, the variance is a bit higher than in the original table that we calculated by hand. Again, though, they're quite similar to the initial table data.