

14.310x: Data Analysis for Social Scientists

Joint, Marginal, and Conditional distributions & Functions of Random Variables

Welcome to your third homework assignment! We encourage you to get an early start, particularly if you still feel you need more experience using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced.

Good luck!

In this problem set we will guide you through different ways of accessing real data sets and how to summarize and describe it properly. First we will go through some of the data that is collected by the World Bank. We will do some cleaning on the data before we start analyzing it. Then, we will try to do a simple web scraping exercise where we will analyze the data as well.

Let's start with this dataset from the World Bank. Please download and save [this file](#) in a folder where you can get it easily.

This dataset is from the [World Bank Gender Statistics website](#). You may visit the website and explore other datasets they offer. **However, for the purposes of this assignment, please use the file in this set of instructions**, since the dataset on the World Bank website may have been updated in the time since this problem set and answer key was posted.

NOTE: It is important to work in the same directory that the files are or to use the whole path when you specify opening a data set. To know which directory you are currently working in, you can use the command `getwd()`. Similarly, in order to set a different directory, you can use the command `setwd()`.

For the purpose of analyzing the data, we are going to use the packages **utils** and **tidyverse**. Once you have uploaded the data to R you are going to see there are multiple indicators of gender, countries and years in the data. In this case we are just interested in analyzing the data for one indicator that is the *Adolescent Fertility Rate*, in the data the indicator code for this variable is called `SP.ADO.TFRT`. This indicator measures the annual number of births to women 15 to 19 years of age per 1,000 women in that age group. It represents the risk of childbearing among adolescent women 15 to 19 years of age. It is also referred to as the age-specific fertility rate for women aged 15-19. Once you have completed this problem set you'll have more information of how this rate has evolved over time and how it varies across different groups of countries.

Take a look at the following lines of code, whose main purpose is to upload the data in a data frame and to choose the proper indicator. Please, try to understand the code and then run it in your computer. Remember to set the directory accordingly to the folder where you saved the files.

```
#Preliminaries
rm(list=ls())
library("utils")
library("tidyverse")

setwd("-----")

#Getting the data
gender_data <- as_tibble(read.csv("Gender_StatsData.csv"))
```

Using the code, answer the following questions:

Question 1

What is the purpose of the line `rm(list = ls())`?

- ☐ To remove all current existing objects in R
- ☐ To change the current directory path
- ☐ To list all the files in the current directory
- ☐ To look in the web for the World Bank dataset

Question 2

The first thing you want to figure out when you look at a new dataset is how it is organized. If your dataset is stored as a tibble, you can simply print the object and it will print in a nice-looking format.

Alternatively, you can also use the built-in R commands such as `str()`, which allows you to see the structure of an object in R. Likewise, the commands `head()` and `tail()` will allow you to see the first six and last six observations of your data frame respectively. Another useful function is the function `dim()`, which will give you the number of rows and columns in your dataset. Take the time to explore the data using these commands and others.

Which of the following statements best describes how your data is organized? Note that we use “indicator” to mean something that is being measured (e.g. fertility rate, enrollment rate, etc.).

- ☐ Each unit of observation (row) is a country/region for a given year.
- ☐ Each unit of observation (row) corresponds to a country/region and an indicator
- ☐ Each unit of observation (row) corresponds to an indicator for a given year.
- ☐ Each unit of observation (row) is a country/region’s indicator for a given year.

Question 3

Now, generate a tibble called “teenage_fr”, which contains only the adolescent fertility rate indicator for each country-year. Please fill in the blank with the correct code.

```
teenager_fr <- _____(gender_data, Indicator.
Code=="SP.ADO.TFRT")
```

- Filter
- Select

What is the equivalent base-R function? (Select one)

- The equivalent base-R function is `match()`.
- The equivalent base-R function is `which()`.
- The equivalent base-R function is `subset()`.

Question 4

Since we are not interested in any other variables and the `gender_data` dataset is quite large, you might want to get rid of it instead of asking R to keep it stored in memory.

Which of the following is the correct code for getting rid of the object `gender_data`?

- `data$gender_data <- NULL`
- `rm(gender_data)`
- `delete(gender_data)`
- `gender_data <- 0`

Now that you have loaded the data we want to analyze and have familiarized yourself with the structure, it is time to get our hands dirty!

A second exploratory thing to do once we have organized a data set is to get basic summary statistics of the data. Now let's do this! To print summary statistics directly in your console, you can use any of the basic summary functions in R(`mean()`, `sd()`, `min()`, `max()`, `sum()` ...). The basic summary functions take vectors as an input, and output a single value.

For example, if you were interested in obtaining the sample mean of the Adolescent Fertility Rate in 1975, one way of doing this is as follows:

```
mean(teenager_fr$X1975, na.rm=TRUE)
```

Question 5

Why is it necessary to add the option `na.rm=TRUE` to the above command? (Select all that apply)

- ☐ The default option of `na.rm` is set to `FALSE`. Therefore, if we don't specify this, R will try to calculate the mean using all the observations in the data.
- ☐ This part is necessary since otherwise R would duplicate some of the observations in the data set when it calculates the sample mean. In particular, the observations with missing values would have higher weights than the observations without missing values.
- ☐ It is not necessary to add this option to the command to obtain the mean of this variable.
- ☐ Otherwise we will obtain a missing value (`na` in R) since not all the countries in the data have information on the adolescent fertility rate in 1975.

To calculate summary statistics for a group of variables, there are a few different commands. The command `mean()` is just one example of the different options available. Now, we ask you to go through the R documentation and explore some of the other commands by yourself.

If you want to store the output as values in your dataset, or if you want to do something more complicated (ex. Generate these by group, or use one of the `dplyr` summary functions (ex. `n_distinct()`), you can use any of the basic summary functions as well as others, in combination with `mutate()` and `summarise()` to generate variables in your dataset containing summary values.

Now that you've learned how to look at and generate summary statistics, answer the following questions.

Question 6

What is the sample mean and standard deviation of the adolescent fertility rate in 1960?

Please round your answers to the second hundredth decimal place, i.e. if your answer is 2.356, round it to 2.36.

Sample mean:

Standard deviation:

Question 7

What is the sample mean and standard deviation of the adolescent fertility rate in 2000?

Please round your answers to the second hundredth decimal place, i.e. if your answer is 2.356, round it to 2.36.

Sample mean:

Standard deviation:

Question 8

True or False? From the values that you have calculated above we can conclude that the Adolescent Fertility Rate has had a permanent decreasing (i.e. only decreases and never increases during this period) trend from 1960-2000, and that the dispersion of this variable has decreased over time.

- ☐ True
- ☐ False

Now, we are interested in plotting the evolution of the Adolescent Fertility Rate from 1960 to 2015. In addition, we are interested in having different information in the same plot. First, we want to plot the sample mean of all the data set, but also we want to add more information such as the rate for low, middle and high income countries (an indicator for country code is stored in the variable `"Country.Code"`).

Inspect this variable to get a sense of what it contains. Note that it includes indicators for both countries, regions, and income group. Since we are only interested in the trends by income group, we want to filter the data to contain only the fertility rate for high, middle, and low income countries as well as the world average.

Question 9

Use the `dplyr filter()` command and the logical `%in%` to keep only the relevant `Country.Code` observations in `teenager_fr`. Make sure you name the new dataset “`byincomelevel`”. Choose the line of code below:

- `byincomelevel <- filter(teenager_fr, Country.Code%in%c(LIC, MIC, HIC, WLD))`
- `teenager_fr <- filter(byincomelevel, Country.Code%in%(LIC, MIC, HIC, WLD))`
- `byincomelevel <- filter(Country.Code%in%c("LIC", "MIC", "HIC", teenager_fr))`
- `byincomelevel <- filter(teenager_fr, Country.Code%in%c("LIC", "MIC", "HIC", "WLD"))`

Notice, there are still two problems with the resulting data:

1. It contains additional variables that we don't need or are meaningless at this level of aggregation.
2. It is not organized in a very intuitive way. A more natural way to organize this data, and prepare it for plotting, is to have each observation represent either a year or a country group-year, and each of the columns represent either the fertility rate for a given group, or if the data is at the country-group year level, then just the fertility rate.

Question 10

Suppose you decide you prefer to have one observation per income group and year. The `dplyr` command `gather()` can help you achieve this. Look up the command in the help files. Select the set of arguments that belong in the blanks below.

Note: Depending on your operating system, you may encounter an encoding error when trying to reference the variable “`Country.Name`”, if that is the case, add the following line before running the code:

```
byincomelevel <- colnames(byincomelevel)[1]="Country.Names")
```

OR

```
colnames(byincomelevel)[1]<- "Country.Name"
```

```
plotdata_bygroupyear <- gather(byincomelevel, ____ (1) ____,  
____ (2) ____,  
____ (3) ____ ) %>%  
select(Year, Country.Name, Country.Code, FertilityRate)
```

- Select from `Country.Code`, `Year`, `FertilityRate`, `X1960:X2015`
- Select from `Country.Code`, `Year`, `FertilityRate`, `X1960:X2015`
- Select from `Country.Code`, `Year`, `FertilityRate`, `X1960:X2015`

Question 11

Suppose you take a look at the data and change your mind. You decided you prefer to look at the data at the year level and have the fertility rates for each income-group as separate variables. The `dplyr` command `spread()` can help you achieve this. Look up the command in the help files. Select the set of arguments that belong in the blanks below:

```
plotdata_byyear <- select(plotdata_bygroupyear, Country.Code,
Year, FertilityRate) %>%
spread(____(1)____, ____ (2) ____)
```

1. Select from Year, Country.Code, FertilityRate
2. Select from Year, Country.Code, FertilityRate

Question 12

True or False? The `select` statement in the code for question 11 is redundant, since we already selected the variables we wanted in generating `plotdata_byyear`.

- True
- False

Question 13

Good news. We are finally ready to plot the data! Let's begin by plotting the fertility rate over time, separately for each income level. To do this, we can use the basic `ggplot` syntax Prof. Duflo explained in lecture.

Let's start by trying to generate this plot using the `plotdata_bygroupyear` tibble we generated earlier. Here is the code to generate this plot. Select the set of arguments that belong in the blanks to generate the desired plot.

```
ggplot(plotdata_bygroupyear, aes(x=____(1)____, y=____(2)____,
group=____(3)____) Q
+ geom_line
```

1. Select from Year, Country.Code, FertilityRate
2. Select from Year, Country.Code, FertilityRate
3. Select from Year, Country.Code, FertilityRate

Question 14

It would be nicer if the different plot lines had different colors. You can add the argument `color=Country.Code` to the code you generated in question 13. Where do you need to specify this argument? Select one of the Roman numeral blanks in the code below to replace with `, color=Country.Code` or `color=Country.Code` in order for each of the lines to have a different color.

Note that the unnumbered blanks are from Question 13.

```
ggplot(plotdata_bygroupyear, aes(x=____, y=____, group=____ I)
II) + III
o geom_line(IV) VI
```

- II
- III
- IV
- V

Question 15

It is good practice to include titles in your plot. To do this, look up the `ggplot labs()`. Select one of the Roman numeral blanks in the code below to replace with (possibly preceded by `.` or `+`) `labs(title='Fertility Rate by Country-Income-Level over Time')`.

Note that the unnumbered blanks are from Question 13.

```
ggplot(plotdata_bygroupyear, aes(x=____, y=____, group=_____ I)
II) +
  geom_line(III) IV
```

- I
- II
- III
- IV

Question 16

One more thing we could improve in this plot is the x-axis labels. First, we can remove the leading “X”. Second, by storing them as numeric, `ggplot` can use its “optimal” scaling to make a prettier plot instead of having a label for each year. To do this, we can transform the `Year` variable using `dplyr`’s `mutate` function and a combination of the functions `as.numeric()` and the `stringr` package. Try to figure out a few ways to do this.

Which of the following statements are equivalent and can be used in the blank below to complete it? Select all that apply

```
plotdata_bygroupyear <- mutate(plotdata_bygroupyear,
Year=as.numeric(____))
☐ str_sub(year, -4)
☐ str_sub(Year, 2, 5)
☐ str_replace(Year, "X", "")
```

Question 17

Which of the following statements can you conclude from the plot?

- ☐ The rate of the world average is always below the rate of high and middle income countries.
- ☐ While the rate for high income countries presents a decreasing trend in the period, the rate for low income countries fluctuates until the mid-nineties when it becomes to decrease significantly.

- The gap between high and middle income countries is lower in 2014 than in 1960, while the gap between low and middle income countries is actually larger in 2014 than it is in 1960.
- Since the mid-nineties, the rate for low income countries has decreased more than it has for high and middle income countries.

Now, we are not going to consider the trends of the different categories over the years. Instead, we are going to compare how the distribution of the Adolescent Fertility Rate is different between 1960 and 2000.

We have provided the R script here to help you answer the next set of questions.

First, we want to generate `histdata_twoyears`.

```
histdata_twoyears <- select(teenager_fr, Country.Name,
  Country.Code, Indicator.Name,
    Indicator.Code, X1960, X2000)

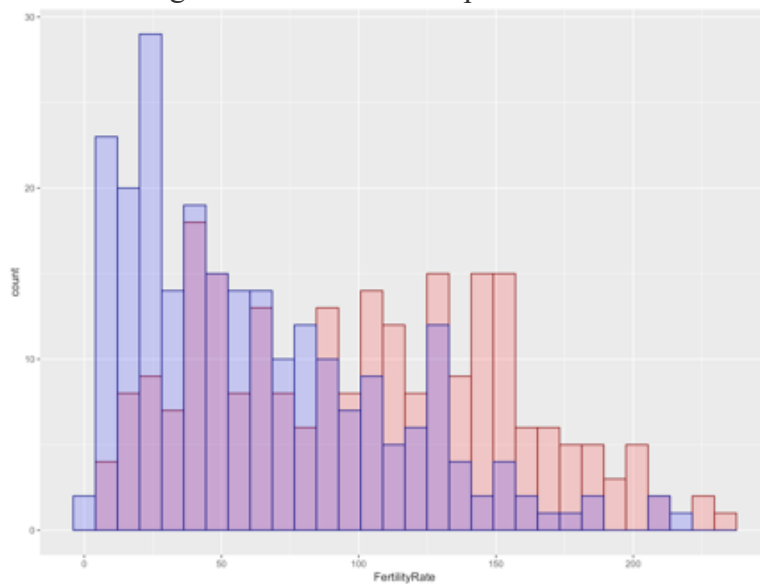
histdata_twoyears <- gather(teenager_fr, Year, FertilityRate,
  X1960, X2000) %>%
  select(Year, Country.Name, Country.Code, FertilityRate)

histdata_twoyears <- filter(histdata_twoyears,
  !is.na(FertilityRate))
```

We want to plot a histogram of the two variables. The following code in R plots the histogram of these two variables in the same graph. Please take a look at the code and try to understand what it is doing.

```
ggplot(histdata_twoyears, aes(x=FertilityRate)) +
  geom_histogram(data=subset(histdata_twoyears,
    Year=="X1960"),
    color="darkred", fill="red", alpha=0.2) +
  geom_histogram(data=subset(histdata_twoyears,
    Year=="X2000"),
    color="darkblue", fill="blue", alpha=0.2)
ggsave("hist.png")
```


Here is the figure that this code has produced:



Question 18

What does the argument `alpha` dictate?

- ☐ The width of the bins.
- ☐ The width of the outline of the bins.
- ☐ The extent to which the plot colors are different.
- ☐ The level of transparency in the color of bins.

Question 19

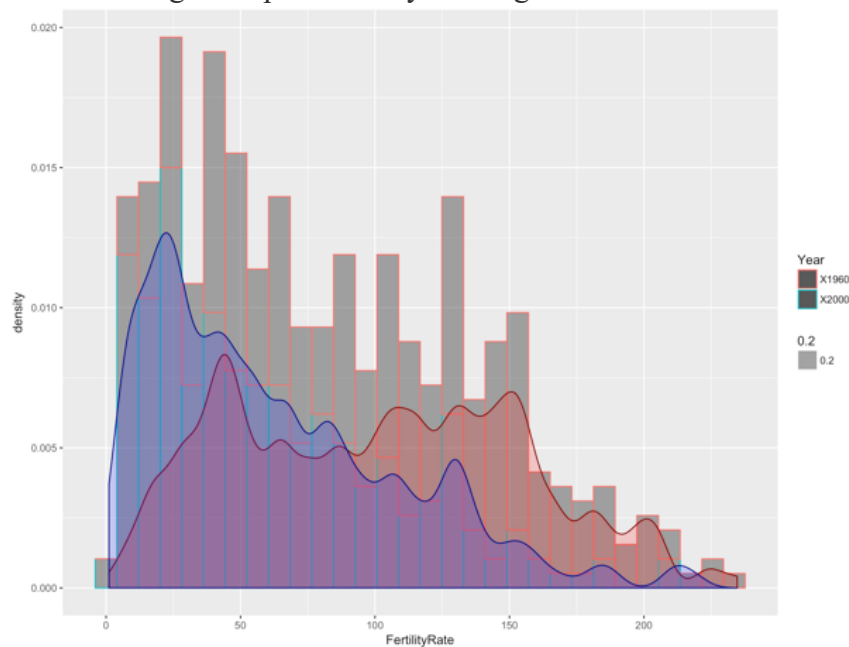
As you can see, we have certain number of bins in the figure. Go to the R documentation and look for the option in the command `geom_histogram()` that will allow you to change the number of bins in the figure. Select all that apply.

- ☐ `binwidth`
- ☐ `position`
- ☐ `breaks`
- ☐ `bins`
- ☐ `center`

Now, we are going to add some kernels to the histogram. The kernels were done using the command `density`, and all the default options in R. Again, take a look at the code, run it on your computer and try to understand what it is doing.

```
ggplot(histdata_twoyears, aes(x=FertilityRate, group=Year, color=Year,
alpha=0.2)) +
  geom_histogram(aes(y=..density..)) +
  geom_density(data=subset(histdata_twoyears, Year=="X1960"),
color="darkred", fill="red", alpha=0.2, bw=5)+
  geom_density(data=subset(histdata_twoyears, Year=="X2000"),
color="darkblue", fill="blue", alpha=0.2, bw=5)
```

The below figure is produced by running the code above.



Question 20

As it was stated before, the plot was done using the default options in R. For the kernel, the default option is to use `gaussian`. There are other options that the user can state when running the density command in R. Of the following list, which of the following weighting function is not bell-shaped? In other words, which one doesn't underweight observations at the boundaries of each bin.

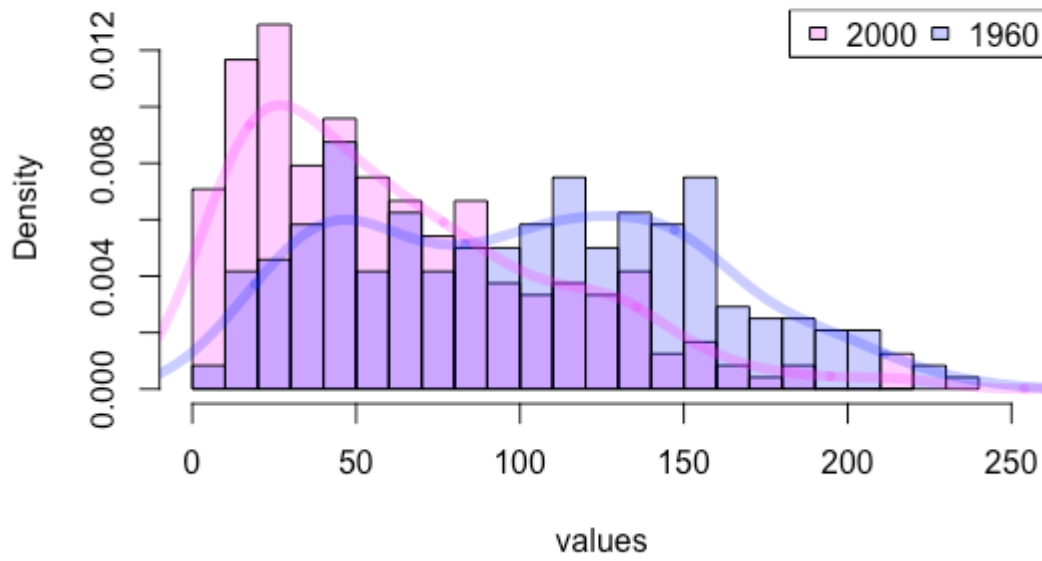
- ☐ `gaussian`
- ☐ `epanechnikov`
- ☐ `rectangular`
- ☐ `triangular`
- ☐ `biweight`
- ☐ `cosine`
- ☐ `optcosine`

Question 21

The following plots were made by changing the bandwidth of the kernel function in R. Which one of them was made with the largest bandwidth?

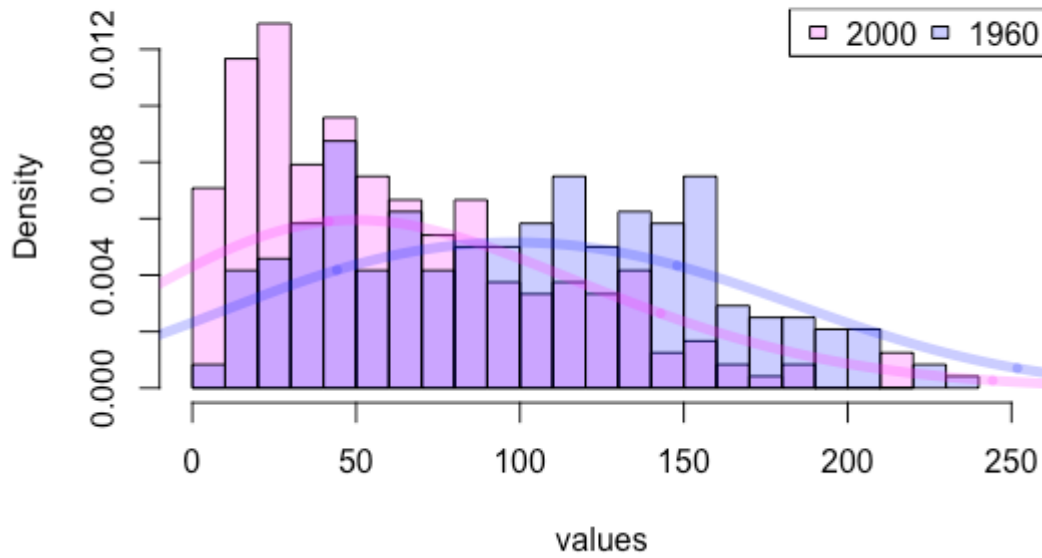
- ☐ It is not possible to tell just by looking at the figure.

Change in the distribution

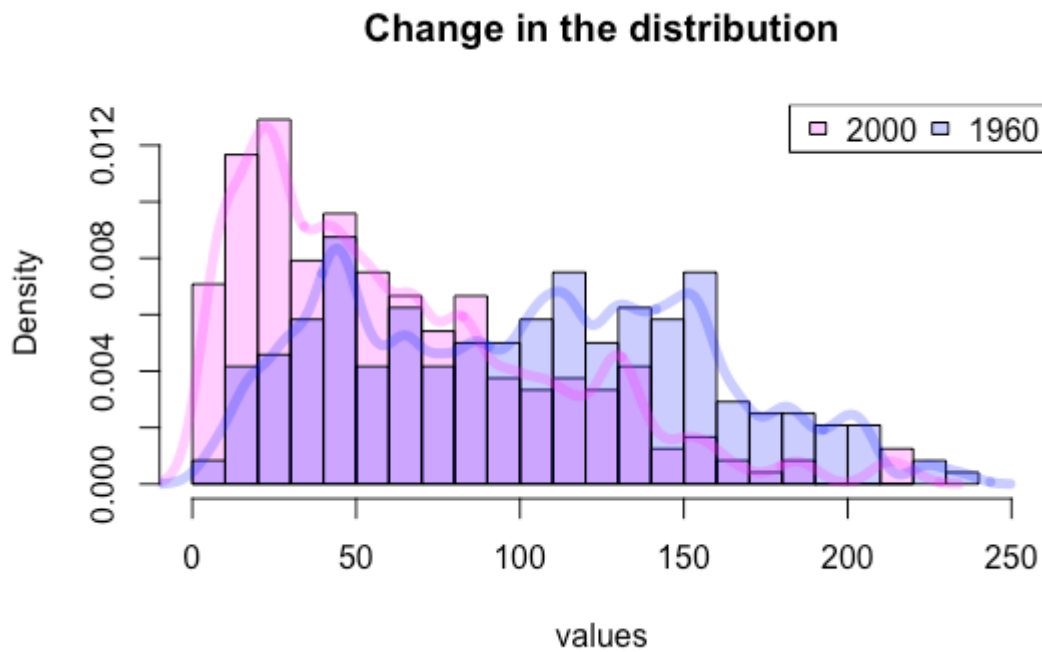


○

Change in the distribution

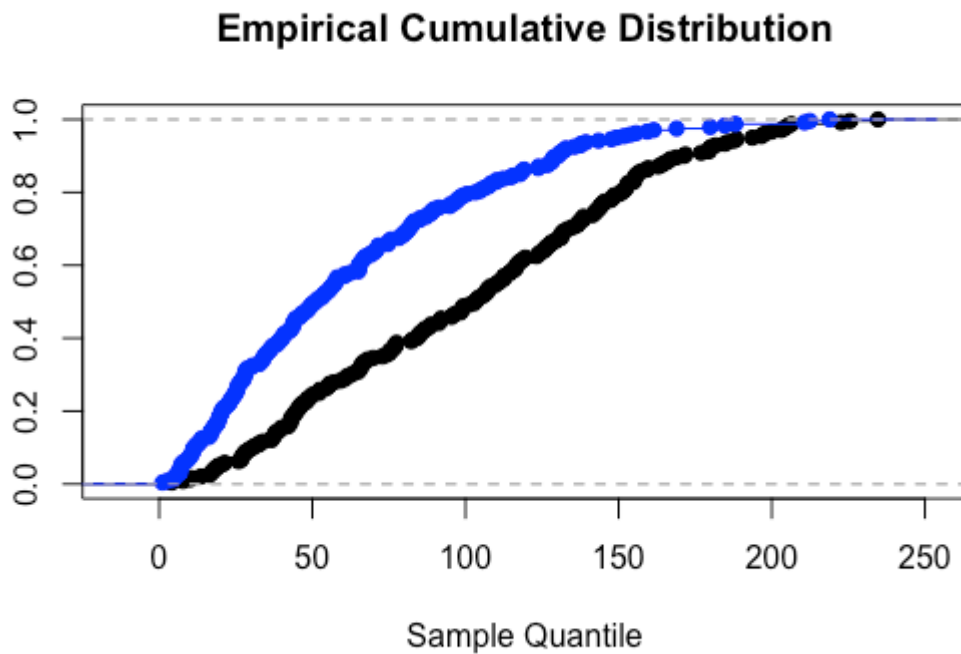


○



○

One of the things that Professor Duflo also discussed in the lecture was the construction of the Empirical Cumulative Distribution (ECD). The following figures shows the ECD for the Adolescent Fertility Rate in the World in 1960 and in 2000. However, as you can see, the person who made the graph forgot to properly label it.



Can you infer from the histograms that were plotted before, which one corresponds to the Adolescent Fertility Rate in 2000 and which one to the same indicator in 1960? (Select all that apply)

- ☐ Blue corresponds to 2000
- ☐ Black corresponds to 2000
- ☐ Blue corresponds to 1960
- ☐ Black corresponds to 1960
- ☐ It is not possible to tell from the plot

Question 23

Using the figure, can you determine whether the distribution used to construct the black series satisfies the First Order Stochastic Dominance property over the distribution used to construct the blue series? Assume that the blue plot is always above (has a value greater than) the black plot.

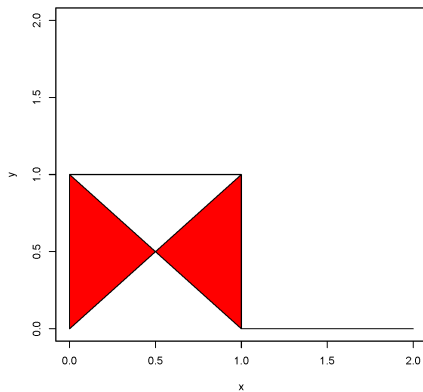
- ☐ Yes
- ☐ No

14.310x: Data Analysis for Social Scientists
Describing Data, Joint and Conditional Distributions – Part II

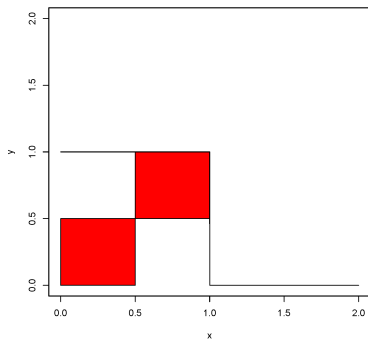
Suppose two sisters, Caroline and Anna, sleep in adjoining rooms. Each has a speaker with which she plays music, and each speaker has a volume dial going from 0 to 1. The joint distribution of the volumes of the two speakers is $f_{XY}(x, y) = c(x + y^2)$ over the unit square, 0 otherwise. Caroline's volume is denoted by X , Anna's by Y .

Question 1

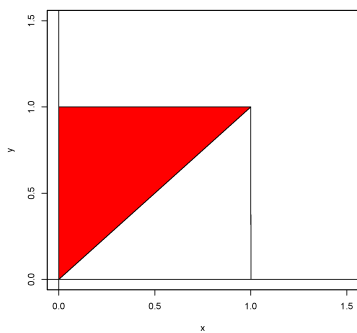
Which of the following figures represent the domain (in red) in which the density function is defined as $f_{XY}(x, y) = c(x + y^2)$?



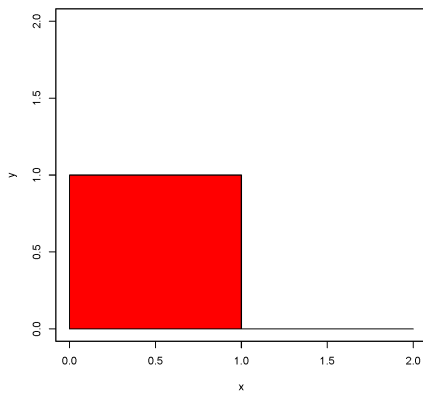
☐



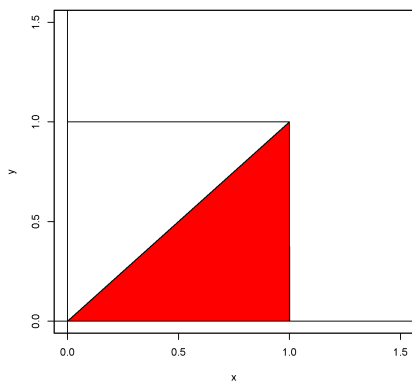
☐



☐



☐



☐

Question 2

What does the constant c represent? (Select all that apply)

- ☐ The constant c is a parameter whose value assures that the joint PDF integrates to 1.
- ☐ The constant c represents a parameter that changes both the joint PDF and the joint CDF of the random variables X and Y .
- ☐ The constant c is an irrelevant parameter in the shape of the joint CDF of the random variables X and Y .
- ☐ The constant c is a parameter that helps to infer whether the random variables X and Y are independent.

Question 3

What is the value of the constant c in this case?

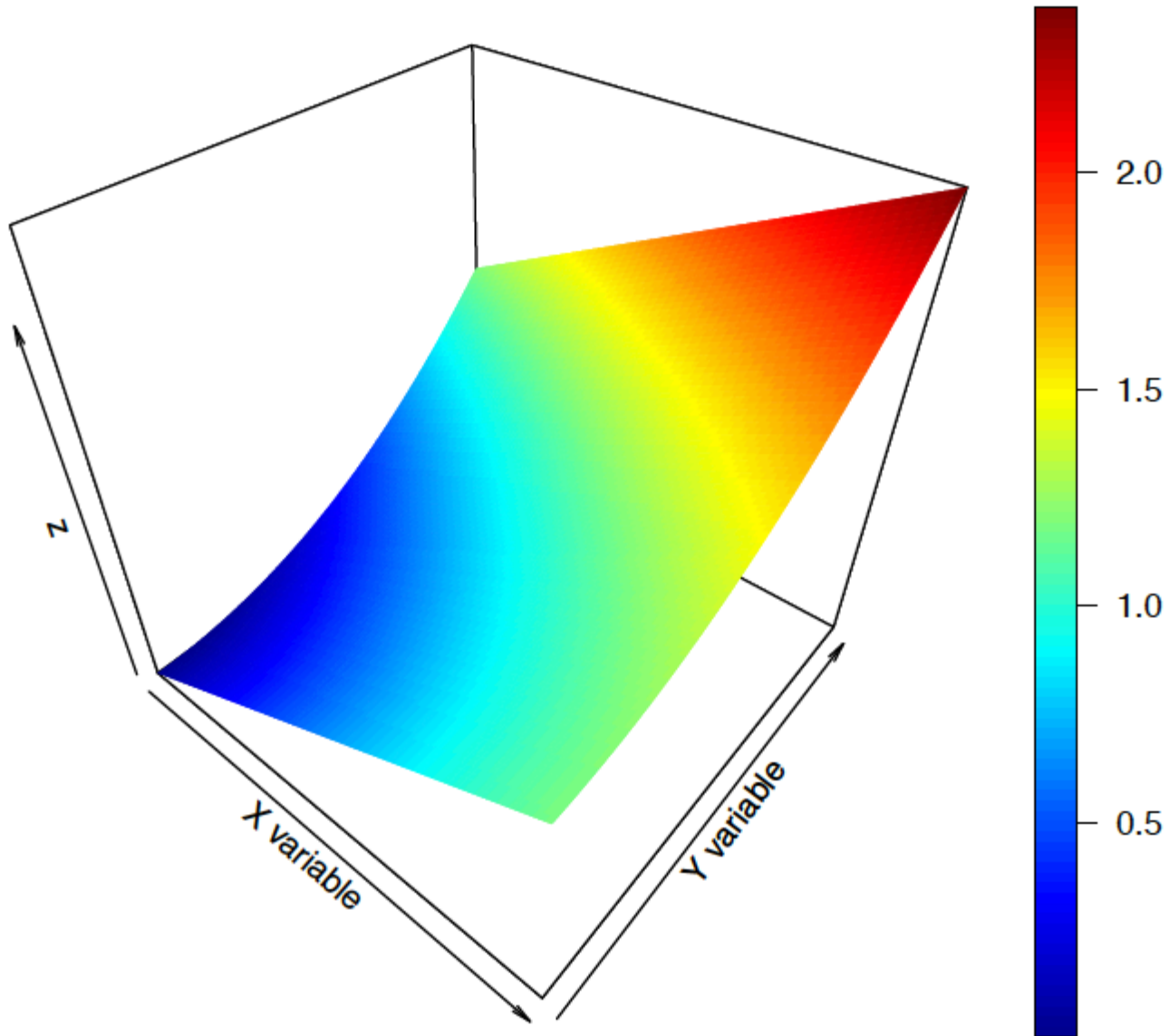
Note: Please review our guidelines on precision regarding rounding answers [here](#).

Now we are going to work in R to plot the bivariate PDF. Download the code [here](#) and take a look at the following code in order to create a grid and a 3-dimensional plot of the PDF. Please note that you might need to install the package `plot3D`.

Question 4

The following plot was created by running the code. A student is claiming that this plot is wrong since there are certain regions in which the PDF shows values larger than 1. Is this student correct that there is a mistake and therefore the plot does not correspond to the information given in the problem?

Plotting joint pdf



- ☐ Yes
- ☐ No

Question 5

Are the volumes of the two speakers independent random variables?

- ☐ Yes

- No

Question 6

Recall that Caroline's speaker volume is denoted by X and Anna's speaker volume is denoted by Y . What is the formula for the marginal distribution of Anna's speaker volume?

- $\frac{5}{6}\left(\frac{1}{2} + y^2\right)$
- $\frac{6}{5}\left(\frac{1}{2} + y^2\right)$
- $\frac{6}{5}\left(\frac{1}{2} + \sqrt{y}\right)$
- $\frac{5}{6}\left(\frac{1}{2} + \sqrt{y}\right)$

Question 7

Recall that Caroline's volume is denoted by X and Anna's volume is denoted by Y . What is the conditional distribution of Caroline's volume as a function of Anna's?

- $\frac{x+y^2}{\frac{1}{2}+y^2}$
- $\frac{\frac{5}{6}(x+y^2)}{\frac{6}{5}\left(\frac{1}{2}+y^2\right)}$
- $\frac{x+\sqrt{y}}{\frac{1}{2}+y^2}$
- $\frac{\frac{6}{5}(x+y^2)}{\frac{1}{2}+y^2}$

Question 8

From this conditional distribution can you infer whether Caroline likes Anna's music or not?

Hint: Think about the probability that Caroline's volume is high when the volume of Anna's music increases.

- Caroline does like Anna's music
- Caroline does not like Anna's music

Question 9

What is the probability that Caroline's volume is less than $\frac{1}{2}$ if Anna's volume is $\frac{1}{2}$?

Note: Please review our guidelines on precision regarding rounding answers [here](#).

Question 10

Recall that Caroline's speaker volume is denoted by X and Anna's speaker volume is denoted by Y . What is the marginal distribution of Caroline's speaker volume?

- ☐ $\frac{5}{6}\left(x + \frac{2}{3}\right)$
- ☐ $\frac{5}{6}\left(x + \frac{1}{3}\right)$
- ☐ $\frac{6}{5}\left(x + \frac{2}{3}\right)$
- ☐ $\frac{6}{5}\left(x + \frac{1}{3}\right)$

Question 11

Is there a First Order Stochastic Dominance (FOSD) relationship between the random variables X and Y ? (We suggest you compute the CDF's of both variables and plot them in R.)

- ☐ The distribution of X FOSD the distribution of Y
- ☐ The distribution of Y FOSD the distribution of X
- ☐ There is no clear relationship

Question 12

From this information, does Anna or Caroline prefer higher volumes?

- ☐ Anna
- ☐ Caroline
- ☐ We can't say

In this problem set we will guide you through different ways of accessing real data sets and how to summarize and describe it properly. First we will go through some of the data that is collected by the World Bank. We will do some cleaning on the data before we start analyzing it. Then, we will try to do a simple web scraping exercise where we will analyze the data as well.

Let's start with this dataset from the World Bank. Please download and save [this file](#) in a folder where you can get it easily.

This dataset is a truncated version of one you can find on the [World Bank Gender Statistics website](#). You may visit the website and explore other datasets they offer. **However, for the purposes of this assignment, please use the file in this set of instructions**, since the dataset on the World Bank website may have been updated in the time since this problem set and answer key was posted.

NOTE: It is important to work in the same directory that the files are or to use the whole path when you specify opening a data set. To know which directory you are currently working in, you can use the command `getwd()`. Similarly, in order to set a different directory, you can use the command `setwd()`.

For the purpose of analyzing the data, we are going to use the packages **utils** and **tidyverse**. Once you have uploaded the data to R you are going to see there are multiple indicators of gender, countries and years in the data. In this case we are just interested in analyzing the data for one indicator that is the *Adolescent Fertility Rate*, in the data the indicator code for this variable is called `SP.ADO.TFRT`. This indicator measures the annual number of births to women 15 to 19 years of age per 1,000 women in that age group. It represents the risk of childbearing among adolescent women 15 to 19 years of age. It is also referred to as the age-specific fertility rate for women aged 15-19. Once you have completed this problem set you'll have more information of how this rate has evolved over time and how it varies across different groups of countries.

Take a look at the following lines of code, whose main purpose is to upload the data in a data frame and to choose the proper indicator. Please, try to understand the code and then run it in your computer. Remember to set the directory accordingly to

indicator. Please, try to understand the code and then run it in your computer. Remember to set the directory accordingly to the folder where you saved the files.

```
#Preliminaries
rm(list=ls())
library("utils")
library("tidyverse")

setwd("-----")

#Getting the data
gender_data <- as_tibble(read.csv("Gender_StatsData.csv"))
```

Using the code, answer the following questions:

Question 1

0.0/1.0 point (graded)

What is the purpose of the line `rm(list = ls())` ?

☐ To look in the web for the World Bank dataset.

☐ To change the current directory path

☐ To list all the files in the current directory

☐ To remove all the current existing objects in R ✓


Explanation

You should be able to look for the help file of the command `rm()`. In particular you should run `help("rm")` in the command window of R-studio. The main argument for `rm()` correspond to the objects that you want to remove from your current session in R. In this case we can also specify a list, by using inside the command `rm()`, the command `ls()` -which lists all the files-, we are able to remove all the current existing objects in our R session.

[Show answer](#)

Submit

You have used 0 of 2 attempts

 Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

The first thing you want to figure out when you look at a new dataset is how it is organized. If your dataset is stored as a tibble, you can simply print the object and it will print in a nice-looking format.

Alternatively, you can also use the built-in R commands such as `str()`, which allows you to see the structure of an object in R. Likewise, the commands `head()` and `tail()` will allow you to see the first six and last six observations of your data

frame respectively. Another useful function is the function `dim()`, which will give you the number of rows and columns in your dataset. Take the time to explore the data using these commands and others.

Which of the following statements best describes how your data is organized? Note that we use "indicator" to mean something that is being measured (e.g. fertility rate, enrollment rate, etc.).

- ☐ Each unit of observation (row) is a country/region's indicator for a given year.
- ☒ Each unit of observation (row) corresponds to a country/region and an indicator. ✓
- ☐ Each unit of observation (row) corresponds to an indicator for a given year.
- ☐ Each unit of observation (row) is a country/region for a given year.

Explanation

If you look at the data using the suggested commands, you will note that the unit of observation (each row) corresponds to a country-indicator and the years are represented in columns. For instance, if you run `head(gender_data)`, you will see that the rate of female access to anti-retroviral drugs is measured by year for the "Arab World."

[Show answer](#)

Submit

You have used 0 of 2 attempts

Question 3

2 points possible (graded)

Now, generate a tibble called `"teenager_fr"`, which contains only the adolescent fertility rate indicator for each country-year. Please fill in the blank with the correct code.

`teenager_fr <-` Select an option ▼ Answer: `filter` `(gender_data, Indicator.Code == "SP.ADO.TFRT")`

What is the equivalent base-R function ? (select one)


☒ The equivalent base-R function is `subset()`. ✓

☐ The equivalent base-R function is `which()`.

☐ The equivalent base-R function is `match()`.

Explanation

Given the structure of the data, you want to get rid of the rows which contain indicators other than the one we are interested in. The dplyr command to select observations based on row values is `filter()`, if you have not done so already, we suggest you look it up in the help files as it is very useful. The equivalent command in base-R is `subset()`, and the basic syntax is identical. The `select()` command from `dplyr`, allows you to select certain columns. If you are curious, look up the `match()` and `which()` commands in the helpfiles.

 Answers are displayed within the problem

Question 4

0.0/1.0 point (graded)

Since we are not interested in any other variables and the `gender_data` dataset is quite large, you might want to get rid of it instead of asking R to keep it stored in memory.

Which of the following is the correct code for getting rid of the object `gender_data` ?

☐ `data$gender_data <- NULL`

☒ `rm(gender_data)` ✓

☐ `delete(gender_data)`

☐ `gender_data <- 0`

Explanation

As noted in question 1, the R command that allows you to remove objects is called `rm()`. Here we just specify that we want to remove the data frame that we called `gender_data`.

Now that you have loaded the data we want to analyze and have familiarized yourself with the structure, it is time to get our hands dirty!

A second exploratory thing to do once we have organized a data set is to get basic summary statistics of the data. Now let's do this! To print summary statistics directly in your console, you can use any of the basic summary functions in R (`mean()`, `sd()`, `min()`, `max()`, `sum()`...). The basic summary functions take vectors as an input, and output a single value.

For example, if you were interested in obtaining the sample mean of the Adolescent Fertility Rate in 1975, one way of doing this is as follows:

```
mean(teenager_fr$X1975, na.rm = TRUE)
```

Question 5

0.0/1.0 point (graded)

Why it is necessary to add the option `na.rm = TRUE` to the above command? (Select all that apply)

☒ The default option of `na.rm` is set to FALSE. Therefore, if we don't specify this, R will try to calculate the mean using all the observations in the data. ✓

☐ This part is necessary since otherwise R would duplicate some of the observations in the data set when it calculates the sample mean. In particular, the observations with missing values would have higher weights than the observations

☐ It is not necessary to add this option to the command to obtain the mean of this variable.

☒ Otherwise we will obtain a missing value (`na` in R) since not all the countries in the data have information on the adolescent fertility rate in 1975. ✓

☐ This option is necessary since there are missing values in the data set. Thus, when R tries to calculate the mean it assumes that the result is not a number (described as `NaN` in R).

Explanation

The default option of the mean command regarding missing values is set to `FALSE`. Thus, when R tries to calculate the sample mean, it is considering the missing values as well. As any operation with missing values, R assumes that the result is also a missing value. For this reason, it is necessary to specify `na.rm = TRUE` so that missing values are not taken into account into the calculation.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

To calculate summary statistics for a group of variables, there are a few different commands. The command `mean()` is just one

To calculate summary statistics for a group of variables, there are a few different commands. The command `mean()` is just one example of the different options available. Now, we ask you to go through the R documentation and explore some of the other commands by yourself.

If you want to store the output as values in your dataset, or if you want to do something more complicated (ex. Generate these by group, or use one of the `dplyr` summary functions (ex. `n_distinct()`), you can use any of the basic summary functions as well as others, in combination with `mutate()` and `summarise()` to generate variables in your dataset containing summary values.

Now that you've learned how to look at and generate summary statistics, answer the following questions.

Question 6

0.0/1.0 point (graded)

What is the sample mean and standard deviation of the adolescent fertility rate in 1960?

Please round your answers to the second hundredth decimal place, i.e. if your answer is 2.356, round it to 2.36.

Sample mean:

Answer: 101.33

Standard deviation:

Answer: 54.21

Explanation

One way of getting this into R is by running the following code:

```
mean(teenager_fr$X1960, na.rm = TRUE)
```

```
sd(teenager_fr$X1960, na.rm = TRUE)
```

Then you should be able to obtain these numbers.

Note: you can also round directly in R:

```
round(mean(teenager_fr$X1960, na.rm = TRUE), 2)
```

```
round(sd(teenager_fr$X1960, na.rm = TRUE), 2)
```

[Show answer](#)

Submit

You have used 0 of 2 attempts

Question 7

0.0/1.0 point (graded)

What is the sample mean and standard deviation of the adolescent fertility rate in 2000?

Please round your answers to the second hundredth decimal place, i.e. if you answer is 2.356 round it 2.36.

Sample mean:

Answer: 63.15

Standard deviation:

Answer: 46.92

Explanation

One way of getting this into R is by running the following code:

```
mean(teenager_fr$X2000, na.rm = TRUE)
```

```
sd(teenager_fr$X2000, na.rm = TRUE)
```


Then you should be able to obtain these numbers.

Note: you can also round directly in R:


```
round(mean(teenager_fr$X2000, na.rm = TRUE), 2)
```

```
round(sd(teenager_fr$X2000, na.rm = TRUE), 2)
```

[Show answer](#)

Submit

You have used 0 of 2 attempts

 Answers are displayed within the problem

Question 8

0.0/1.0 point (graded)

True or False? Based on the results from Questions 6 and 7, we can conclude that the Adolescent Fertility Rate has had a permanent decreasing (i.e. only decreases and never increases during this period) trend from 1960-2000, and that the dispersion of this variable has decreased over time.

☐ True

☒ False ✓

Explanation

From the data above you know that both the mean and the standard deviation of the rate have decreased. This implies that in 2000 the average rate is lower and there seems to be less dispersion than in 1960. However, with this information it is not possible to conclude that the trend has always been permanently decreasing over this period. For example, it may have increased before decreasing.

Now, we are interested in plotting the evolution of the Adolescent Fertility Rate from 1960 to 2015. In addition, we are interested in having different information in the same plot. First, we want to plot the sample mean of all the data set, but also we want to add more information such as the rate for low, middle and high income countries (an indicator for country code is stored in the variable "Country.Code").

Inspect this variable to get a sense of what it contains. Note that it includes indicators for both countries, regions, and income group. Since we are only interested in the trends by income group, we want to filter the data to contain only the fertility rate for high, middle, and low income countries as well as the world average.

Question 9

0.0/1.0 point (graded)

Use the dplyr `filter()` command and the logical `%in%` to keep only the relevant `Country.Code` observations in `teenager_fr`. Make sure you name the new dataset `"byincomelevel"`. Choose the line of code below:

☐ `byincomelevel <- filter(teenager_fr, Country.Code %in% c(LIC, MIC, HIC, WLD))`

☐ `teenage_fr <- filter(byincomelevel, Country.Code %in% c(LIC, MIC, HIC, WLD))`

☐ `byincomelevel <- filter(Country.Code %in% c("LIC", "MIC", "HIC", teenager_fr))`

☒ `byincomelevel <- filter(teenager_fr, Country.Code %in% c("LIC", "MIC", "HIC", "WLD"))` ✓

Notice, there are still two problems with the resulting data:

1. It contains additional variables that we don't need or are meaningless at this level of aggregation.
2. It is not organized in a very intuitive way. A more natural way to organize this data, and prepare it for plotting, is to have each observation represent either a year or a country group-year, and each of the columns represent either the fertility rate for a given group, or if the data is at the country-group year level , then just the fertility rate.

Question 10

0.0/1.0 point (graded)

Suppose you decide you prefer to have one observation per income group and year. The tidyr command `gather()` can help you achieve this. Look up the command in the help files. Select the set of arguments that belong in the blanks below.

NOTE: Depending on your operating system, you may encounter an encoding error when trying to reference the variable "Country.Name", if that is the case, add the following line before running the code:

```
byincomelevel<-colnames(byincomelevel)[1]="Country.Name")
```

OR

```
colnames(byincomelevel)[1]<-"Country.Name"
```

```
plotdata_bygroupyear <- gather(byincomelevel, 
```

Answer: Year

Answer: FertilityRate

Answer: X1960:X2015

Explanation

The `tidyr` function `gather()` basically works by moving column names specified in the 4th argument (3rd blank), in this case the year columns `X1960:X2015` into a new variable, whose name is specified as the first argument, in this case `"Year"` (as you can see from the specified arguments of `select()`). It stores variable values into a new variable, whose name is specified as the second argument, in this case `"FertilityRate"`. So in short, `gather` creates a separate observation for each year-country.code, from the year columns `X1960:X2015`.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 11

0.0/1.0 point (graded)

Suppose you take a look at the data and change your mind. You decided you prefer to look at the data at the year level and have the fertility rates for each income-group as separate variables. The `tidyr` command `spread()` can help you achieve this. Look up the command in the help files. Select the set of arguments that belong in the blanks below:

```
plotdata_byyear<-select(plotdata_bygroupyear, Country.Code, Year, FertilityRate) %>%
```

`spread(` `Answer: Country.Code` `,` `Answer: FertilityRate` `)`

Explanation

Question 12

0.0/1.0 point (graded)

True or False? The select statement in the code for question 11 is redundant since we already selected the variables we wanted in generating `plotdata_byyear`.

☐ True

☒ False ✓


Explanation

Try and view the data and see what happens. Recall that we also included the variable `Country.Name` in `plotdata_bygroupyear`. Since this variable is not constant by year (it varies by country-group), and we are spreading the data by country-group; if you omit the select statement, you will end up with stacked observations, and the data still at the country-group level.

[Show answer](#)

Submit

You have used 0 of 1 attempt

 Answers are displayed within the problem

i Answers are displayed within the problem

Question 13

0.0/1.0 point (graded)

Good news. We are finally ready to plot the data! Let's begin by plotting the fertility rate over time, separately for each income level. To do this, we can use the basic `ggplot` syntax Prof. Duflo explained in lecture.

Let's start by trying to generate this plot using the `plotdata_bygroupyear` tibble we generated earlier. Here is the code to generate this plot. Select the set of arguments that belong in the blanks to generate the desired plot.

```
ggplot(plotdata_bygroupyear, aes(x=  Answer: Year , y=   
Answer: FertilityRate , group=  Answer: Country.Code ))  
+ geom_line()
```

Explanation

We want to plot the Fertility rate (our y-variable) over time (our x-variable), separately by income group (our group variable, in this case Country.Code)). Therefore we want `x="Year", y="FertilityRate", group="Country.Code"`.

[Show answer](#)

Submit

You have used 0 of 2 attempts

It would be nicer if the different plot lines had different colors. You can add the argument `color=Country.Code` to the code you generated in question 13. Where do you need to specify this argument? Select one of the Roman numeral blanks in the code below, to replace with `,color=Country.Code` or `color=Country.Code` in order for each of the lines to have a different color.

Note that the unnumbered blanks are from Question 13.

```
ggplot(plotdata_bygroupyear, aes(x=_____, y=_____, group=_____ I) II) + III  
  geom_line(IV)V
```

☒ I ✓

☐ II

☐ III

☐ IV

☐ V

Explanation

The `color()` argument is an `aes()` option, therefore it belongs before the end of the parentheses where "I" is.

Question 15

0.0/1.0 point (graded)

It is good practice to include titles in your plot. To do this, look up the ggplot `labs()` . Select one of the Roman numeral blanks in the code below to replace with (possibly preceded by `,` or `+`)

```
labs(title='Fertility Rate by Country-Income-Level over Time')
```

Note that the unnumbered blanks are from Question 13.

```
ggplot(plotdata_bygroupyear, aes(x=_____, y=_____, group=_____ I) II) +  
geom_line(III) IV
```

☐ I

☐ II

☐ III

☒ IV ✓

Explanation

Since title is a plot option, it should be added as

```
+ labs(title='Fertility Rate by Country-Income-Level over Time')
```

 in IV.

Question 16

0.0/1.0 point (graded)

One more thing we could improve in this plot is the x-axis labels. First, we can remove the leading “X”. Second, by storing them as numeric, `ggplot` can use its “optimal” scaling to make a prettier plot instead of having a label for each year. To do this, we can transform the `Year` variable using `dplyr`’s `mutate` function and a combination of the functions `as.numeric()` and the `stringr` package. Try to figure out a few ways to do this.

Which of the following statements are equivalent and can be used in the blank below, to complete it? Select all that apply.

```
plotdata_bygroupyear <- mutate(plotdata_bygroupyear, Year=as.numeric(_____))
```

☐ `str_sub(Year, -4)` ✓

☐ `str_sub(Year, 2, 5)` ✓

☐ `str_replace(Year, "X", "")` ✓

Explanation

Since the numeric part of the year variable contains the last 4 parts, the function `str_sub()` can be used to return the last four characters of a string (`str_sub(Year, -4)`) or the second through fifth characters (`str_sub(Year, 2, 5)`). Alternatively, we can use `str_replace()` to replace “X” with nothing (“”) to achieve the same thing.

Which of the following statements can you conclude from the plot? Select all that apply.

- ☐ The rate of the world average is always below the rate of high and middle income countries.
- ☒ While the rate for high income countries presents a decreasing trend in the period, the rate for low income countries fluctuates until the mid-nineties when it becomes to decrease significantly. ✓
- ☒ The gap between high and middle income countries is lower in 2014 than in 1960, while the gap between low and middle income countries is actually larger in 2014 than it is in 1960. ✓
- ☒ Since the mid-nineties, the rate for low income countries has decreased more than it has for high and middle income countries. ✓

Explanation

A is false since the MIC rate is actually lower than the the WLD rate.

B is true and you can infer that from looking at the HIC and LIC lines. The HIC line decreases continuously over time. The LIC line is almost flat until the mid-nineties, and then decreases much faster.

C is true. You can take a look at the difference between the HIC and the MIC lines in 2014 compared to 1960. The difference is smaller in 2014 than it is in 1960. When we look at the difference between LIC and MIC in 2014 and 1960, we see the difference is larger in 2014 than it is in 1960.

D is true. When we compare the average slope for LIC countries, we see it is steeper (more negative) than the average slope for MIC and HIC countries during that same period.

Now, we are not going to consider the trends of the different categories over the years. Instead, we are going to compare how the distribution of the Adolescent Fertility Rate is different between 1960 and 2000.

We have provided the R script [here](#) to help you answer the next set of questions.

First, we want to generate `histdata_twoyears`.

```
histdata_twoyears <- select(teenager_fr, Country.Name, Country.Code, Indicator.Name,  
  Indicator.Code, X1960, X2000)
```

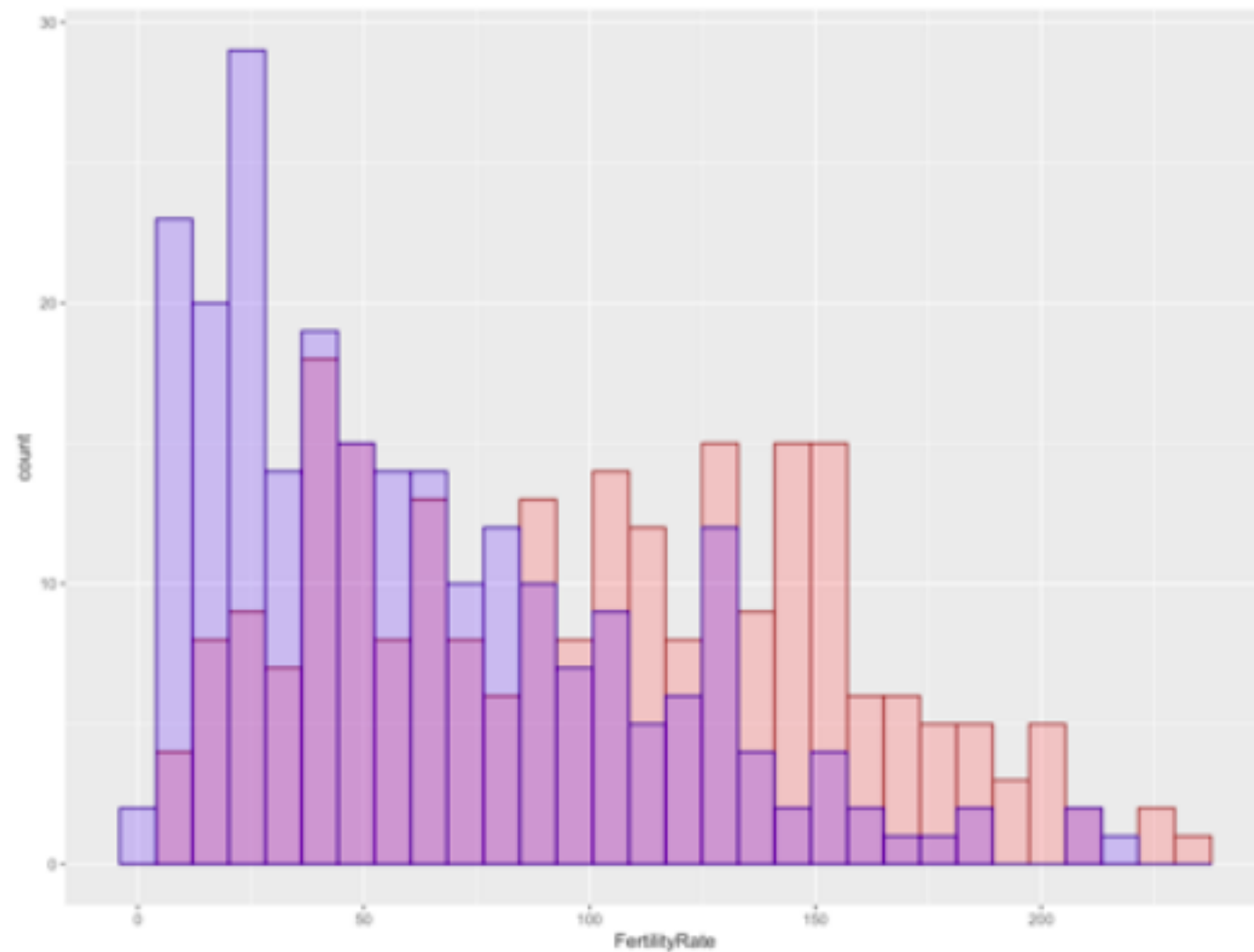
```
histdata_twoyears <- gather(teenager_fr, Year, FertilityRate, X1960, X2000) %>%  
  select(Year, Country.Name, Country.Code, FertilityRate)
```

```
histdata_twoyears <- filter(histdata_twoyears, !is.na(FertilityRate))
```

We want to plot a histogram of the two variables. The following code in R plots the histogram of these two variables in the same graph. Please take a look at the code and try to understand what it is doing.

```
ggplot(histdata_twoyears, aes(x=FertilityRate)) +  
  geom_histogram(data=subset(histdata_twoyears, Year=="X1960"),  
    color="darkred", fill="red", alpha=0.2) +  
  geom_histogram(data=subset(histdata_twoyears, Year=="X2000"),  
    color="darkblue", fill="blue", alpha=0.2)  
ggsave("hist.png")
```

Here is the figure that this code has produced:



Question 18

0.0/1.0 point (graded)

What does the argument `alpha` dictate?

- ☐ The width of the bins.
- ☐ The width of the outline of the bins.
- ☐ The extent to which the plot colors are different.
- ☒ The level of transparency in the color of bins. ✓

Explanation

if you go to the R documentation for the `ggplot` function, you will find that `alpha` dictates the transparency of the line and color fills.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 19

Question 19

0.0/1.0 point (graded)

As you can see, we have certain number of bins in the figure. Go to the R documentation and look for the option in the command `geom_histogram()` that will allow you to change the number of bins in the figure. Select all that apply.

☒ `binwidth` ✓

☐ `position`

☒ `breaks` ✓

☒ `bins` ✓

☐ `center`

Explanation

Refer to the R documentation for `geom_histogram()`. You can control the number of bins by specifying the `bins`, `binwidth` – the width of each bin, or the number of > `breaks` on the x-axis.

[Show answer](#)

Explanation

Refer to the R documentation for `geom_histogram()`. You can control the number of bins by specifying the `bins`, `binwidth` – the width of each bin, or the number of `> breaks` on the x-axis.

[Show answer](#)

Submit

You have used 0 of 2 attempts

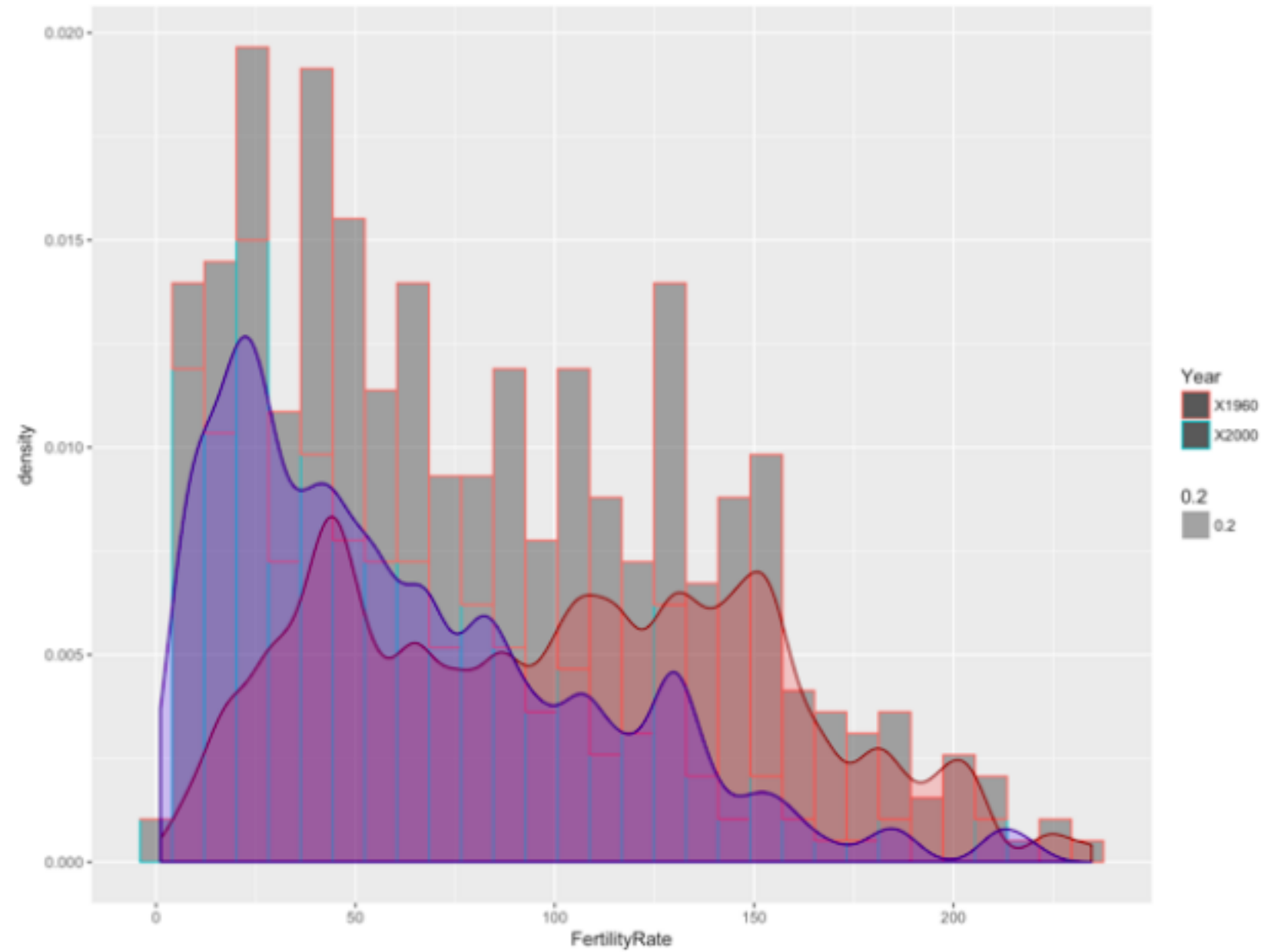
i Answers are displayed within the problem

Now, we are going to add some kernels to the histogram. The kernels were done using the command `density`, and all the default options in R. Again, take a look at the code, run it on your computer and try to understand what it is doing.

```
ggplot(histdata_twoyears, aes(x=FertilityRate, group=Year, color=Year, alpha=0.2)) +  
  geom_histogram(aes(y=..density..)) +  
    geom_density(data=subset(histdata_twoyears, Year=="X1960"), color="darkred", fill="red",  
alpha=0.2, bw=5)+  
    geom_density(data=subset(histdata_twoyears, Year=="X2000"), color="darkblue",  
fill="blue", alpha=0.2, bw=5)
```

The below figure is produced by running the code above.

The below figure is produced by running the code above.



Question 20

0.0/1.0 point (graded)

Question 20

0.0/1.0 point (graded)

As it was stated before, the plot was done using the default options in R. For the kernel, the default option is to use gaussian. There are other options that the user can state when running the density command in R. Of the following list, which of the following weighting function is not bell-shaped? In other words, which one doesn't underweight observations at the boundaries of each bin.

☐

gaussian

☐

epanechnikov

☒

rectangular ✓

☐

triangular

☐

biweight

☐

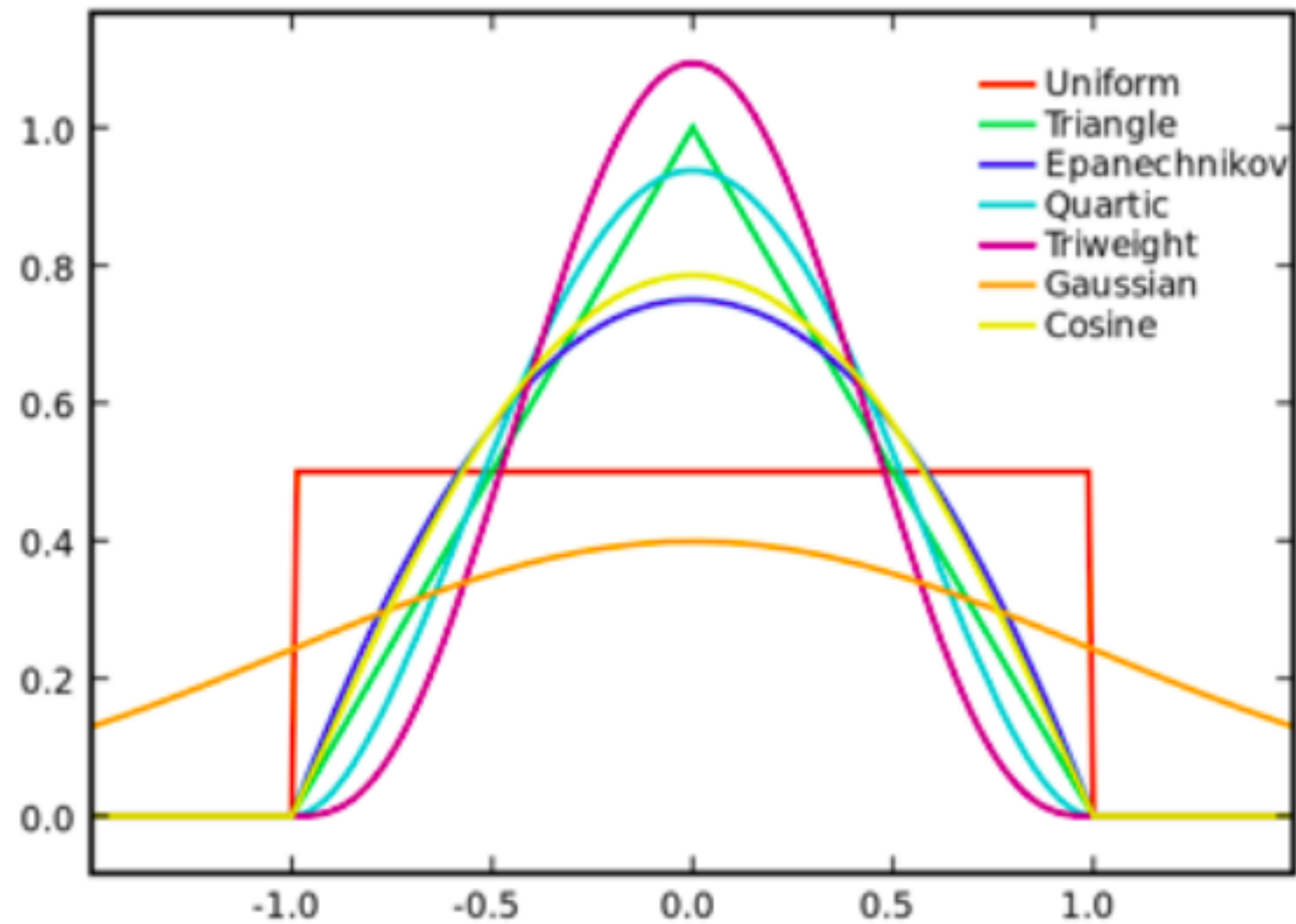
cosine

☐

optcosine

Explanation

The following plot shows the different shapes of the kernel functions. As you can see the only one without a bell-shaped function is the rectangular one. This kernel is also called the uniform kernel.



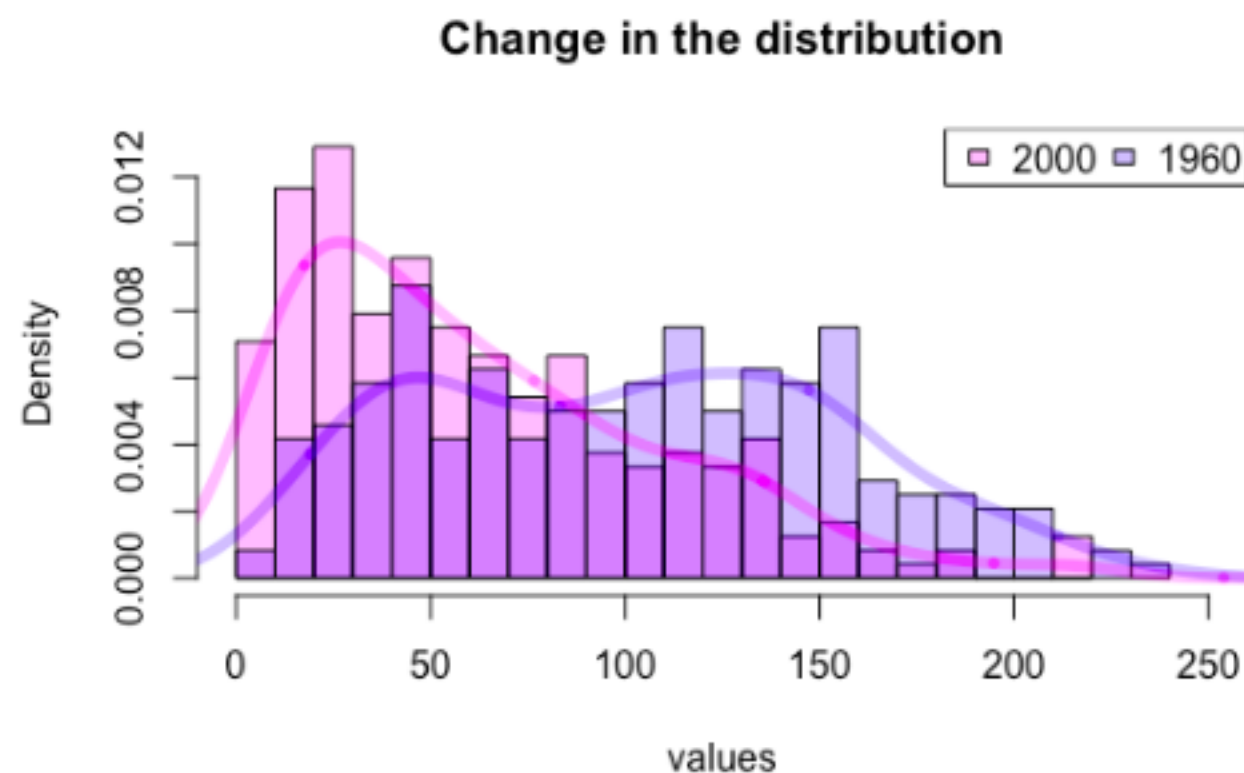
Question 21

0.0/1.0 point (graded)

The following plots were made by changing the bandwidth of the kernel function in R. Which one of them was made with the largest bandwidth?

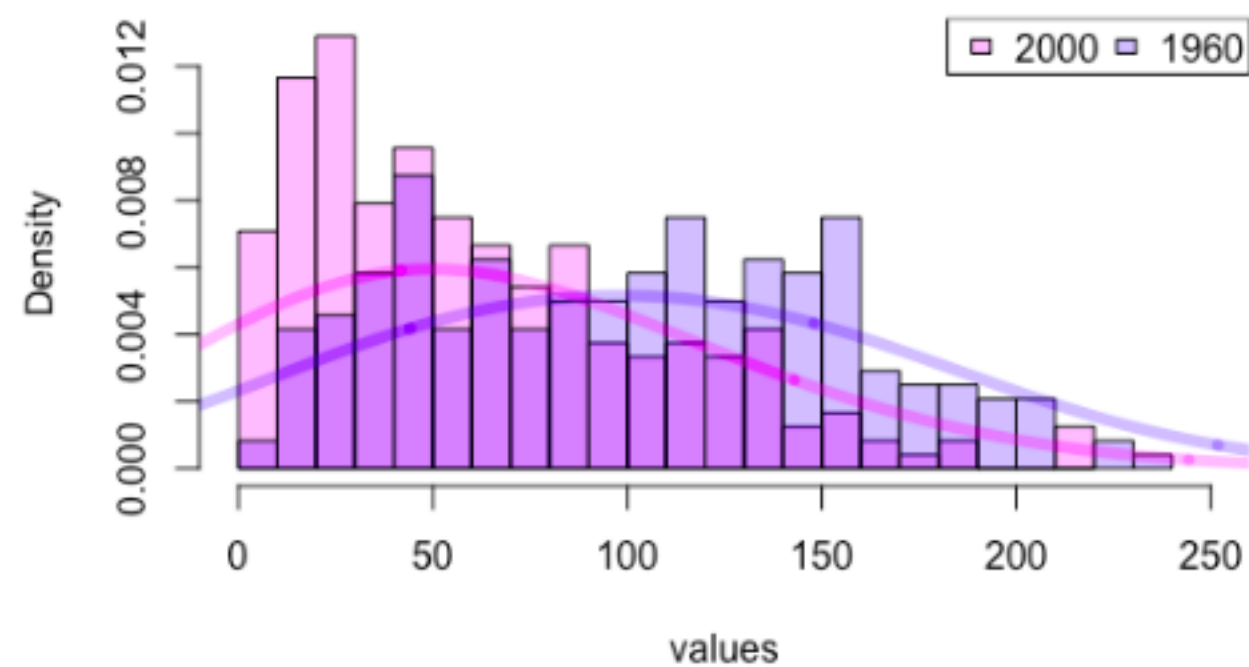
☐ It is not possible to tell just by looking at the figure.

☐

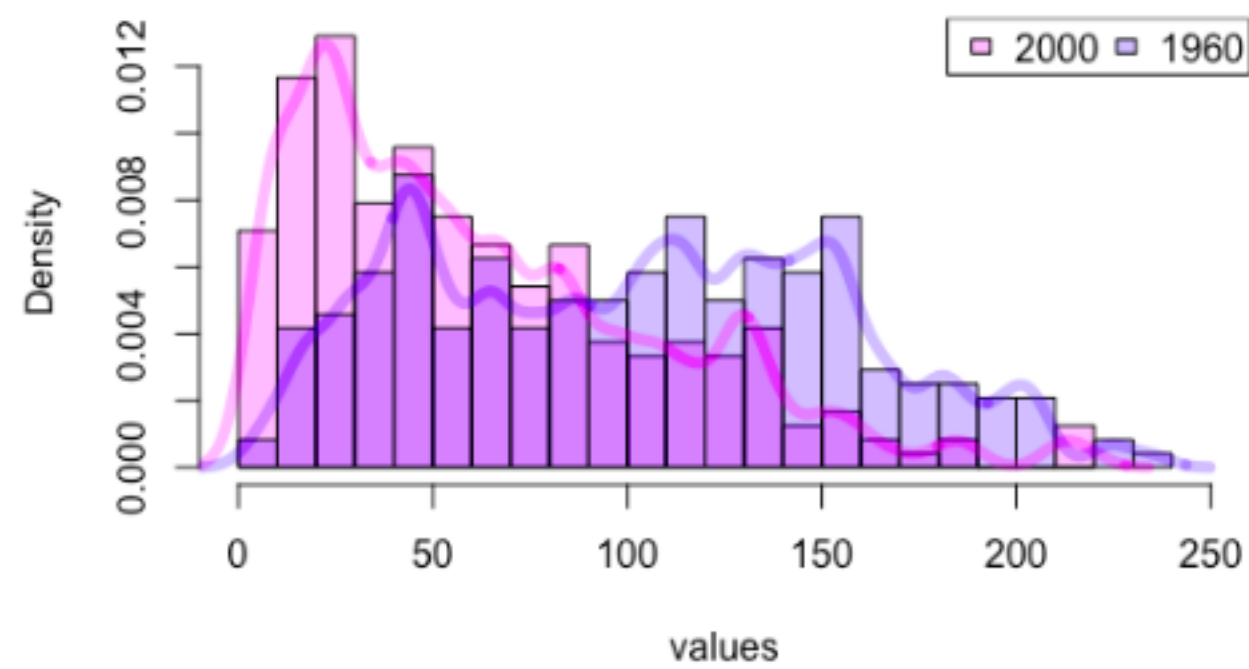


☐

Change in the distribution



Change in the distribution



Explanation

As Professor Duflo discussed in the class, the optimal bandwidth balances a trade-off between bias and variance. The larger the bandwidth is the largest the bias of the density is which results in a smoother looking function. As you can see, the kernel in answer (c) is the furthest from the histogram, which suggests is the one done with the largest bandwidth.

[Show answer](#)

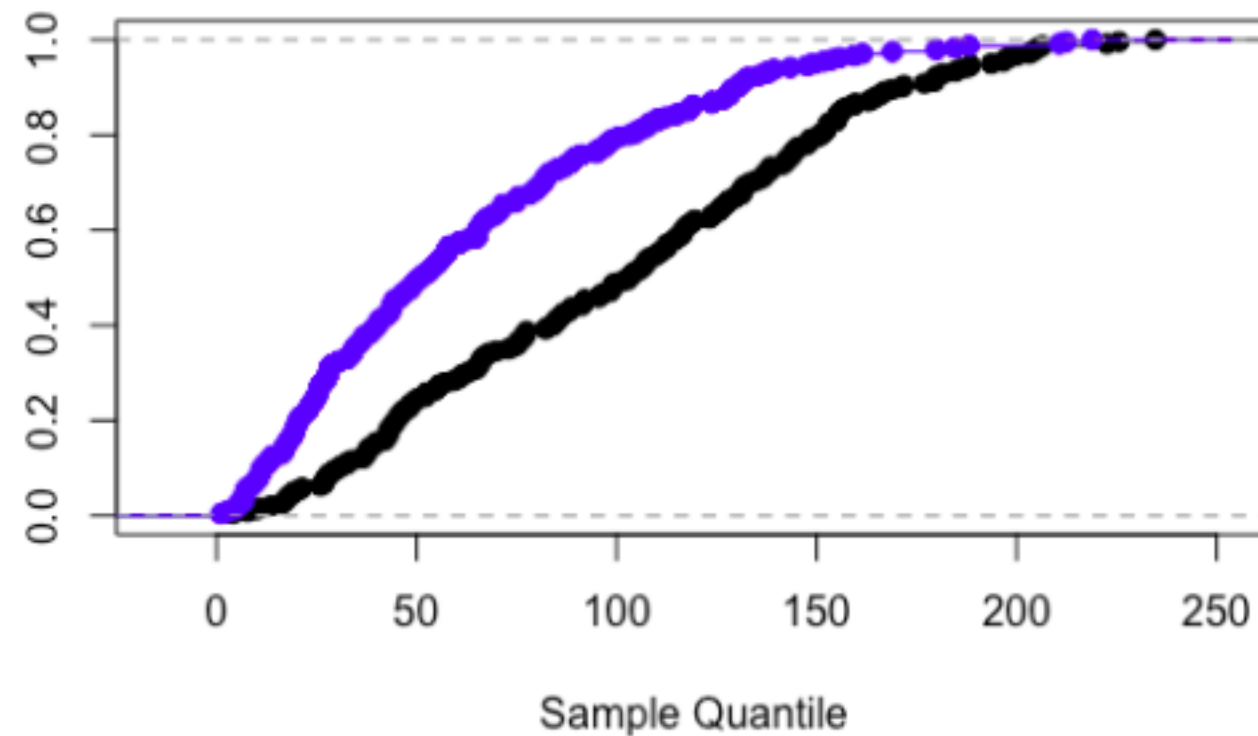
Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

One of the things that Professor Duflo also discussed in the lecture was the construction of the Empirical Cumulative Distribution (ECD). The following figures shows the ECD for the Adolescent Fertility Rate in the World in 1960 and in 2000. However, as you can see, the person who made the graph forgot to properly label it.

Empirical Cumulative Distribution



Question 22

0.0/1.0 point (graded)

Can you infer from the histograms that were plotted before, which one corresponds to the Adolescent Fertility Rate in 2000 and which one to the same indicator in 1960? (Select all that apply)

☐ Blue corresponds to 2000 ✓

☐ Black corresponds to 2000

☐ Blue corresponds to 1960

☒ Black corresponds to 1960 ✓

☐ It is not possible to tell from the plot

Explanation

The empirical PDF of the Adolescent Fertility Rate in 2000 is "to the left" of the one in 1960. Because the mode in 2000 is less than the mode in 1960, it is accumulating mass faster. Thus, the empirical CDF of the Adolescent Fertility Rate in 2000 is always higher than the empirical CDF of 1960. Thus, the blue plot corresponds to the distribution in 2000 and the black plot to the distribution in 1960.

[Show answer](#)


Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 23


0.0/1.0 point (graded)

 Answers are displayed within the problem

Question 23

0.0/1.0 point (graded)

Using the figure, can you determine whether the distribution used to construct the black series satisfies the First Order Stochastic Dominance property over the distribution used to construct the blue series? Assume that the blue plot is always above (has a value greater than) the black plot.

☒ Yes 

☐ No

Explanation

From the figure you can see that the black plot always has a value lower than the blue line. This is precisely the definition of first order stochastic dominance since for all value of k , it is satisfied that $Pr(x \leq k|black) \leq Pr(x \leq k|blue)$.

[Show answer](#)

Submit

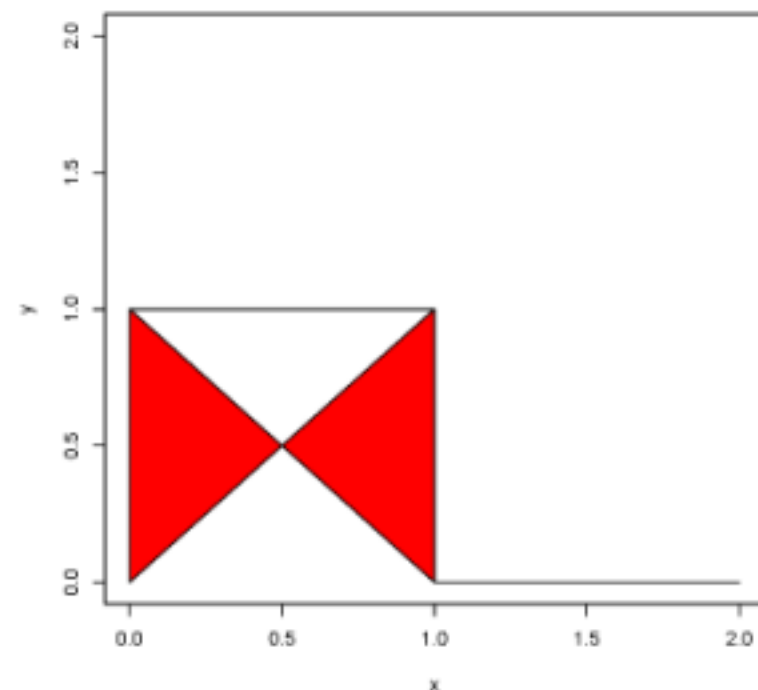
You have used 0 of 1 attempt

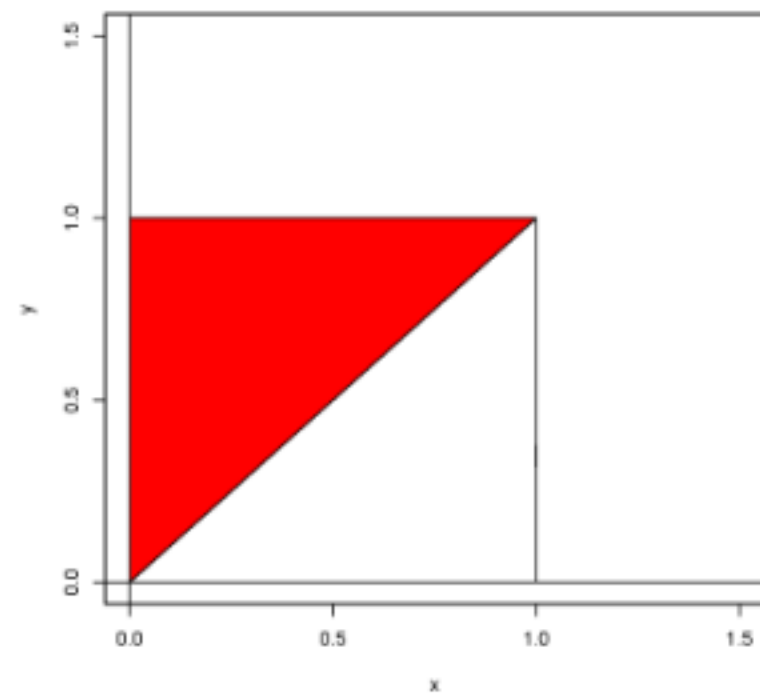
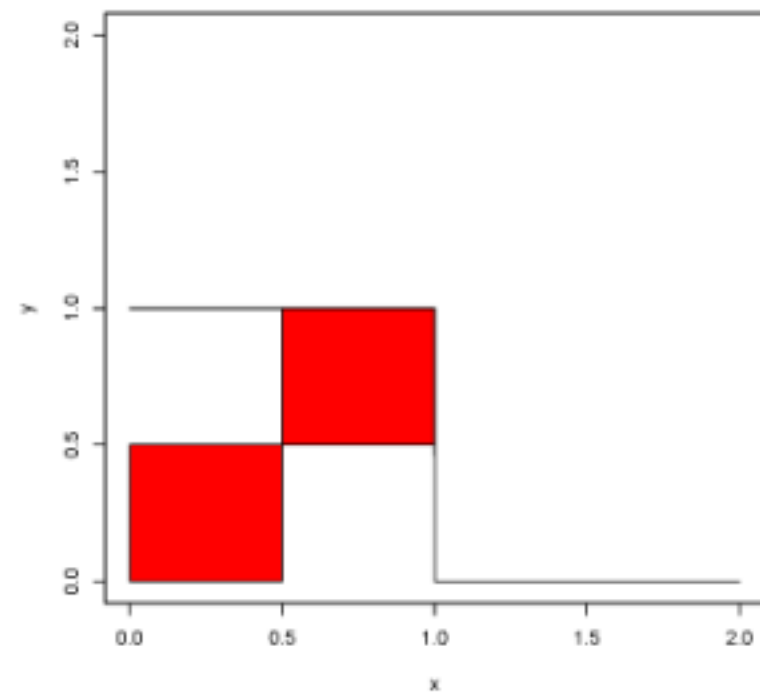
Suppose two sisters, Caroline and Anna, sleep in adjoining rooms. Each has a speaker with which she plays music, and each speaker has a volume dial going from 0 to 1. The joint distribution of the volumes of the two speakers is $f_{XY}(x, y) = c(x + y^2)$ over the unit square, 0 otherwise. Caroline's volume is denoted by X , Anna's by Y .

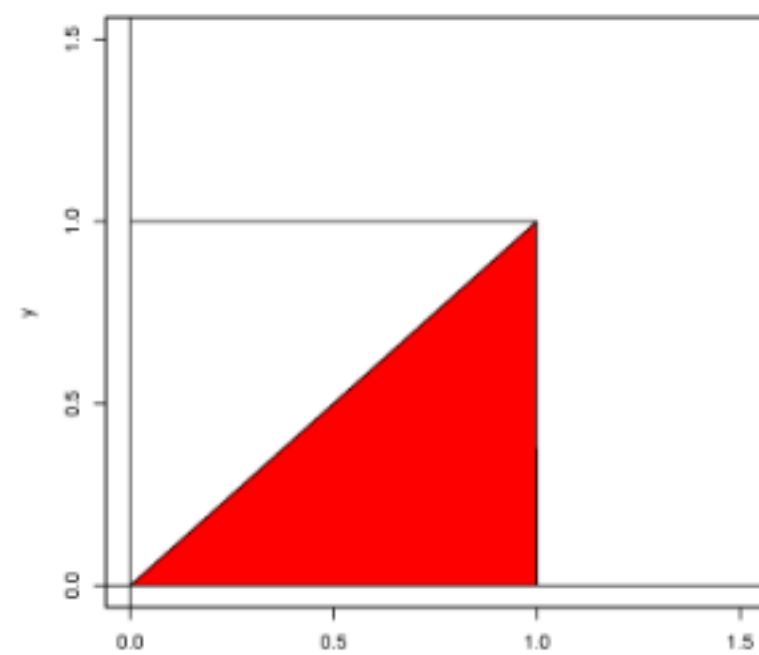
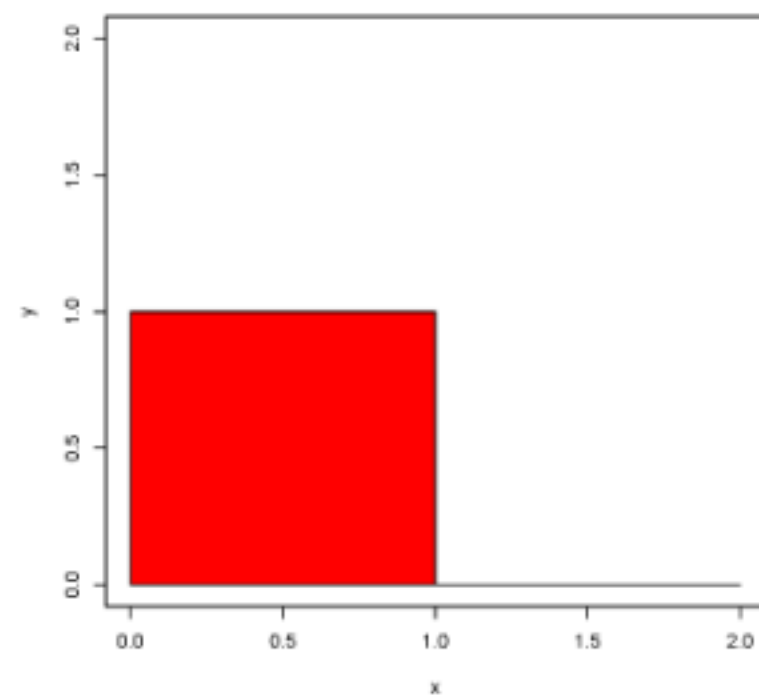
Question 1

0.0/1.0 point (graded)

Which of the following figures represent the domain (in red) in which the density function is defined as $f_{XY}(x, y) = c(x + y^2)$?








Explanation

As the problem states the domain in which the bivariate density function is defined as $f_{XY}(x, y) = c(x + y^2)$ corresponds to the unit square. The two-dimensional plot of the square is in option (d)

[Show answer](#)

Submit

You have used 0 of 2 attempts

 Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

What does the constant c represent? (Select all that apply)

☐ The constant c is a parameter whose value assures that the joint PDF integrates to 1. ✓

☐ The constant c represents a parameter that changes both the joint PDF and the joint CDF of the random variables X and Y . ✓

☐ The constant c is an irrelevant parameter in the shape of the joint CDF of the random variables X and Y .

☐ The constant c is a parameter that helps to infer whether the random variables X and Y are independent.

Explanation

As discussed by Professor Ellison in lecture, the value of the parameter c must assure that the joint PDF of the random variables X and Y integrates to 1. Thus, it affects the shape of both the joint PDF and the joint CDF of these random variables. This implies that it is not irrelevant. It doesn't tell us anything on whether the random variables X and Y are independent, and therefore option (d) is incorrect.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 3

0.0/1.0 point (graded)

What is the value of the constant c in this case?

Note: Please review our guidelines on precision regarding rounding answers [here](#).

Answer: 6/5

Explanation

$$\int_0^1 \int_0^1 c(x + y^2) dy dx = \int_0^1 c\left(xy + \frac{y^3}{3}\right) \Big|_0^1 dx = \int_0^1 c\left(x + \frac{1}{3}\right) dx = c\left(\frac{x^2}{2} + \frac{1}{3}x\right) \Big|_0^1 = c\left(\frac{1}{2} + \frac{1}{3}\right) = \frac{5}{6}c = 1 \implies c =$$

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Now we are going to work in R to plot the bivariate PDF. Download the code [here](#) and take a look at the following code in order to create a grid and a 3-dimensional plot of the PDF. Please note that you might need to install the package `plot3D`.

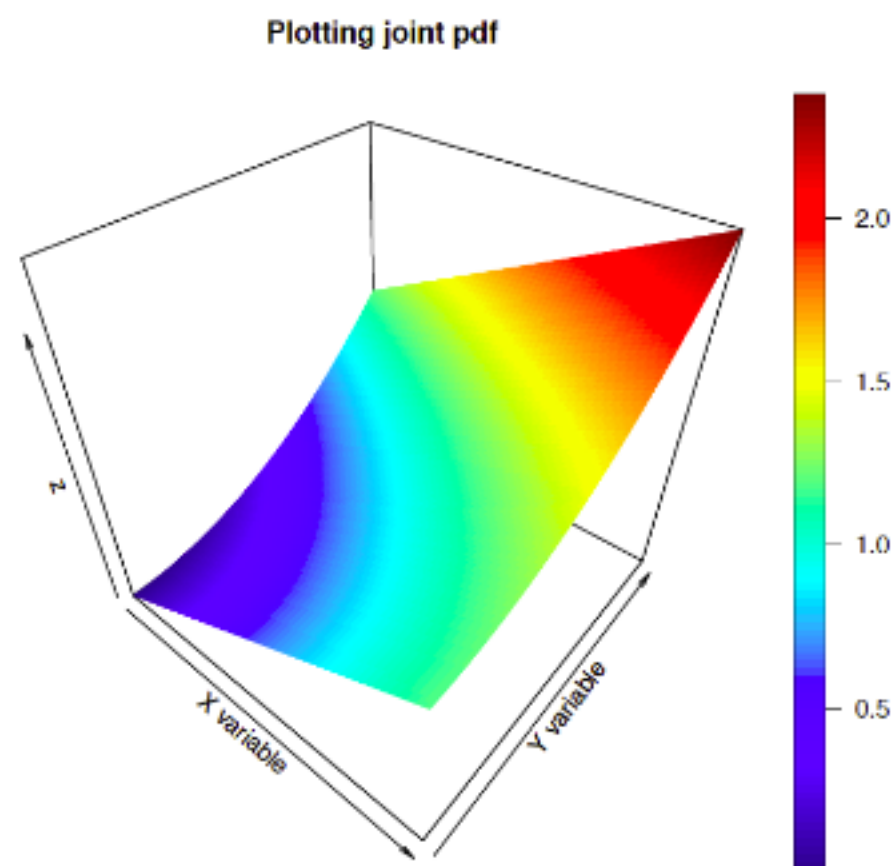
Question 4

0.0/1.0 point (graded)

Question 4

0.0/1.0 point (graded)

The following plot was created by running the code. A student is claiming that this plot is wrong since there are certain regions in which the PDF shows values larger than 1. Is this student correct that there is a mistake and therefore the plot does not correspond to the information given in the problem?



☐ Yes

☒ No ✓

Explanation

The volumes are independent if $f_{XY}(x, y)$ can be factored into an X component and a Y component, i.e.

$f_{XY}(x, y) = f_X(x) f_Y(y)$. It does not appear to be the case that $f_{XY}(x, y) = \frac{6}{5} (x + y^2)$ can be factored into an X and Y component. More formally, if the volumes are independent, the conditional distribution of X given Y cannot depend on Y and vice-versa.

[Show answer](#)

Submit

You have used 0 of 1 attempt

i Answers are displayed within the problem

Question 6

0.0/1.0 point (graded)

Question 6

0.0/1.0 point (graded)

Recall that Caroline's speaker volume is denoted by X and Anna's speaker volume is denoted by Y . What is the formula for the marginal distribution of Anna's speaker volume?

☐ $\frac{5}{6} \left(\frac{1}{2} + y^2 \right)$

☒ $\frac{6}{5} \left(\frac{1}{2} + y^2 \right)$ ✓

☐ $\frac{6}{5} \left(\frac{1}{2} + \sqrt{y} \right)$

☐ $\frac{5}{6} \left(\frac{1}{2} + \sqrt{y} \right)$

Explanation

We know that we have to integrate over the complete domain of X for each potential value y that the random variable Y can take. Then, we have that: $f_Y(y) = \int_0^1 f_{XY}(x, y) dx = \int_0^1 \frac{6}{5} (x + y^2) dx = \frac{6}{5} \left(\frac{x^2}{2} + xy^2 \right) \Big|_0^1 = \frac{6}{5} \left(\frac{1}{2} + y^2 \right)$

Recall that Caroline's volume is denoted by X and Anna's volume is denoted by Y . What is the conditional distribution of Caroline's volume as a function of Anna's?

☐ $\frac{\binom{x+y^2}{\frac{1}{2}+y^2}}{\binom{\frac{1}{2}+y^2}{\frac{1}{2}+y^2}}$ ✓

☐ $\frac{\frac{5}{6} \binom{x+y^2}{\frac{1}{2}+y^2}}{\frac{6}{5} \binom{\frac{1}{2}+y^2}{\frac{1}{2}+y^2}}$

☐ $\frac{\binom{x+\sqrt{y}}{\frac{1}{2}+y^2}}{\binom{\frac{1}{2}+y^2}{\frac{1}{2}+y^2}}$

$$\frac{\frac{6}{5} \binom{x+y^2}{}}{\binom{\frac{1}{2}+y^2}{}}$$

Explanation

We have to use the previous information, then we have that:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\frac{6}{5} \binom{x+y^2}{}}{\binom{\frac{1}{2}+y^2}{}} = \frac{\binom{x+y^2}{}}{\binom{\frac{1}{2}+y^2}{}}$$

[Show answer](#)

Submit

You have used 0 of 2 attempts

Question 8

0.0/1.0 point (graded)

From this conditional distribution can you infer whether Caroline likes Anna's music or not?

Hint: Think about the probability that Caroline's volume is high when the volume of Anna's music increases.

☒ Caroline does like Anna's music ✓

☐ Caroline does not like Anna's music

Explanation

It seems they do like each others' music. Suppose Anna goes from having her music off ($y = 0$) to full blast ($y = 1$). Then you can check using the conditional distribution that the probability that Caroline's music volume is high (say between 0.75 and 1) goes down.

[Show answer](#)

Submit

You have used 0 of 1 attempt

What is the probability that Caroline's volume is less than $\frac{1}{2}$ if Anna's volume is $\frac{1}{2}$?

Note: Please review our guidelines on precision regarding rounding answers [here](#).

Answer: 1/3



Explanation

We are interested in the probability that Caroline's volume is less than $\frac{1}{2}$ if Anna's volume is $\frac{1}{2}$. This is given by the conditional probability $P(X < \frac{1}{2} | Y = \frac{1}{2})$.

This conditional probability is computed in Question 7 using the following formula:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

The numerator is the joint probability: $f_{XY}(x, y) = \frac{6}{5}(x + y^2)$. The denominator is the marginal distribution for Y : $f_Y(y) = \frac{6}{5}(\frac{1}{2} + y^2)$. Note that the marginal distribution for Y is computed by integrating the joint PDF over the entire domain of X . You can see the computation in the answer explanation for Question 6.

Using this, we know the conditional probability is: $f_{X|Y}(x|y) = \frac{(x+y^2)}{(\frac{1}{2}+y^2)}$.

Now, to compute the probability that Caroline's volume X is less than $\frac{1}{2}$, conditional on the probability that Anna's volume Y is equal to $\frac{1}{2}$, we plug in $y = \frac{1}{2}$ into the formula above and integrate from $x = 0$ to $x = \frac{1}{2}$.

$$P\left(X < \frac{1}{2} \middle| Y = \frac{1}{2}\right) = \int_0^{\frac{1}{2}} f_{X|Y=\frac{1}{2}}\left(x \middle| Y = \frac{1}{2}\right) dx = \int_0^{\frac{1}{2}} \left(\frac{4}{3}x + \frac{1}{3}\right) dx = \left(\frac{4}{3} \frac{x^2}{2} + \frac{1}{3}x\right) \bigg|_0^{\frac{1}{2}} = \frac{1}{3}$$

Question 10

0.0/1.0 point (graded)

Recall that Caroline's speaker volume is denoted by X and Anna's speaker volume is denoted by Y . What is the marginal distribution of Caroline's speaker volume?

☐ $\frac{5}{6} \left(x + \frac{2}{3} \right)$

☐ $\frac{5}{6} \left(x + \frac{1}{3} \right)$

☐ $\frac{6}{5} \left(x + \frac{2}{3} \right)$

☒ $\frac{6}{5} \left(x + \frac{1}{3} \right)$ ✓

Explanation

We need to calculate $f_X(x)$ which we can do by integrating over all the domain of Y . Then we have that:

$$f_X(x) = \int_0^1 f_{XY}(x, y) dy = \int_0^1 \frac{6}{5} \left(x + y^2 \right) dy = \frac{6}{5} \left(xy + \frac{y^3}{3} \right) \Big|_0^1 = \frac{6}{5} \left(x + \frac{1}{3} \right)$$

Question 11

0.0/1.0 point (graded)

Is there a First Order Stochastic Dominance (FOSD) relationship between the random variables X and Y ? (We suggest you compute the CDF's of both variables and plot them in R.)

☐ The distribution of X FOSD the distribution of Y

☐ The distribution of Y FOSD the distribution of X

☒ There is no clear relationship ✓

Explanation

If we computed the CDF's, then we know that:

$$P(X \leq x) = \frac{6}{5} \left(\frac{x^2}{2} + \frac{x}{3} \right) P(Y \leq y) = \frac{6}{5} \left(\frac{y^3}{3} + \frac{y}{2} \right)$$

We can run the following [code](#) in R to plot these functions. Please note that your graph would be saved in the working directory. The plot is shown by the following figure:

CDF plot