Now that you have uploaded the data to R that we are interested in analyzing, it is time to get our hands dirty! First, let's try to explore the data a bit more. The command str() will allow you to see the structure of an object in R. Type str(teenager_fr) to get a sense of the variables that we are currently using. Likewise, the command head() and tail(), will allow you to see the first six and last six observations of your data frame, respectively. Try to add this into your R code and explore the data set a bit more by yourself.

A second exploratory thing to do once we have organized a data set is to get basic summary statistics of the data. So, let's do this! One way to do this is to use the command `mean()`, which allows you to get the average of the variables in your data. For example, if you were interested in obtaining the sample mean of the Adolescent Fertility Rate in 1975, you can do this by running the following code.

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

Question 4

(1/1 point)

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

Finger Exercises due Oct 17, 2016 at 05:00 IST

Module 3: Homework
Homework due Oct 10, 2016 at 05:00 IST

▸ Exit Survey

- [x] a. 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. ✔
- [ ] b. 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.
- [ ] c. It is not necessary to add this option to the command to obtain the mean of this variable.
- [x] d. Otherwise we will obtain a missing value since not all the countries in the data have information on the adolescent fertility rate in 1975. ✔
- [ ] e. 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.

✔

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 for the calculation.

You have used 2 of 2 submissions

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.

This next set of questions you should be able to answer after you have taken sometime to explore the R documentation.

Question 5

(1 point possible)
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 you answer is 2.356 round it 2.36.

Sample mean:

101.17 ✔ **Answer:** 101.17

\(\)

Standard deviation:

53.7 ✔ **Answer:** 53.70

\(\)

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.

You have used 1 of 2 submissions

Question 6

(1 point possible)

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:

62.72 ✔

\(\)

Standard deviation:

46.61 ✔

\(\)

You have used 1 of 2 submissions

Question 7

(1/1 point)

From the values that you have calculated above, can you conclude that the Adolescent Fertility Rate has had a permanent decreasing trend over this period of time and that the dispersion of this variable has decreased?

- Yes
- No ✔

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 decreasing.

You have used 1 of 1 submissions