

An Overview: Exploratory Data Analysis - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

1 point possible (graded)

What is “exploratory data analysis”?

- ☐ Quantifying the extent to which data deviates from an expected model
- ☐ Creating graphs made from data which aim to convey a story; often presented to others
- ☐ Cleaning datasets prior to analysis
- ☒ Summarizing data for yourself/your own research ✓

Explanation

As Professor Duflo briefly summarizes, exploratory data analysis aims to summarize a dataset’s main characteristics (often visually), in hopes of revealing interesting information or patterns to the analyst.

Show answer

The Histogram - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

To obtain the density for a histogram, you must divide the number of observations in each “bin” by:

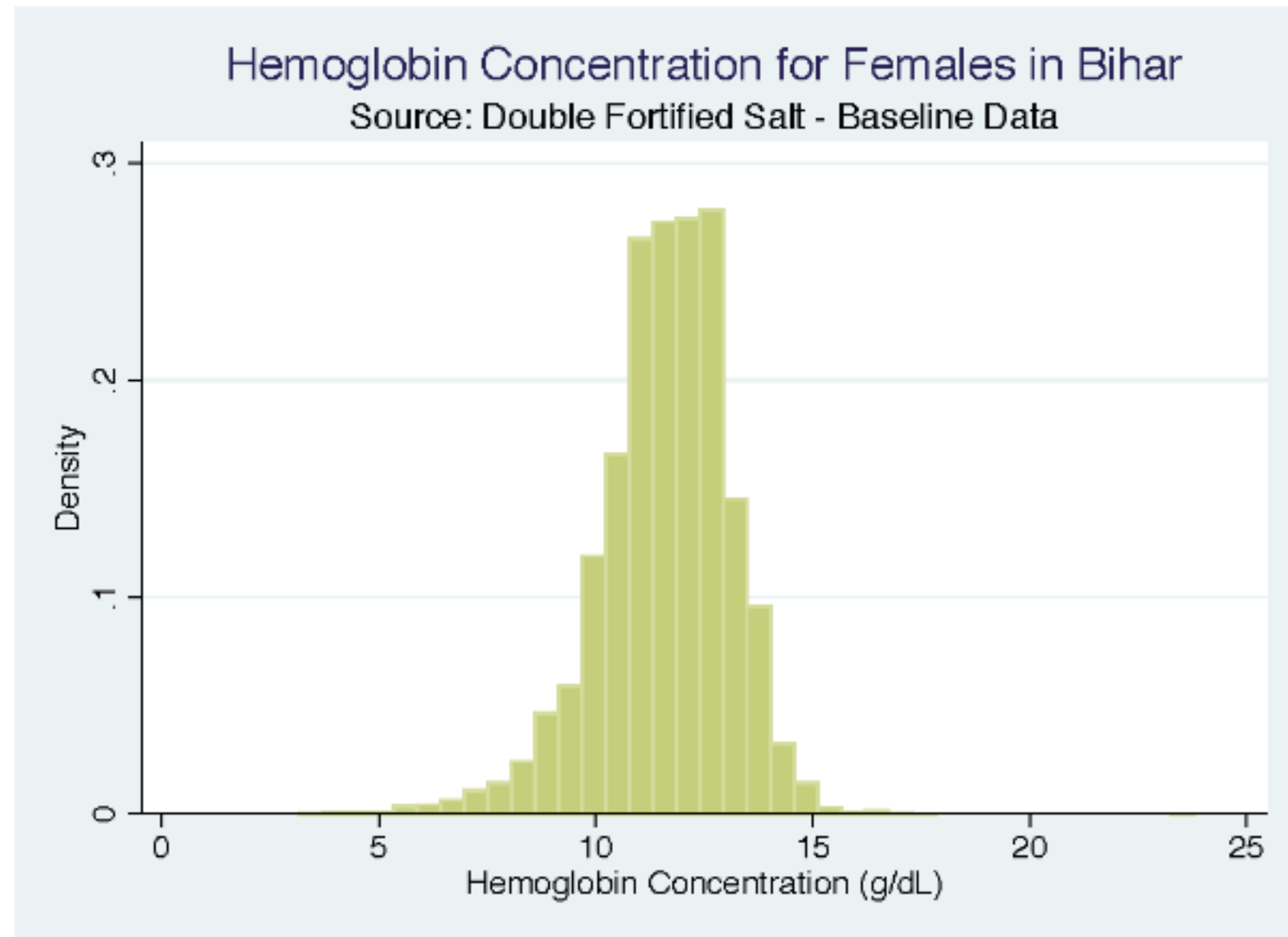
- ☒ The total number of observations ✓
- ☐ 100 in order to obtain a percentage
- ☐ The number of observations in the following bin
- ☐ The number of observations in the previous bin

Explanation

To obtain the proportion of cases that fall into each bin, you must divide the the number of cases in that bin by the total number of observations.

1 point possible (graded)

The following is histogram of female hemoglobin concentration (an indicator for anemia) from an experiment conducted in Bihar, India. The unit for hemoglobin level is grams per deciliter (g/dL).



According to this histogram, the majority of respondents have hemoglobin levels between:

According to this histogram, the majority of respondents have hemoglobin levels between:

☐ 15 and 18 g/dL

☐ 8 and 11 g/dL

☒ 11 and 15 g/dL ✓

☐ 5 and 8 g/dL

Explanation

From the histogram, we can see that the density (which is the number of observations within a bin divided by the total number of observations) peaks somewhere between 10 g/dL and 15 g/dL. Therefore, of the choices provided, the greatest proportion of respondents have hemoglobin levels between 11 and 15 g/dL.

[Show answer](#)

Submit

You have used 0 of 2 attempts

Question 1

1 point possible (graded)

Line 13 contains the following R script:

```
bihar_adult_females <- filter(bihar_data, adult==1, female==1)
```

Suppose we only want to look at adult males. Which of the following scripts would select only adult males?

☐ `bihar_adult_males <- filter(bihar_data, adult==0, female==0)`

☐ `bihar_adult_males <- filter(bihar_data, adult==1, male==1)`

☒ `bihar_adult_males <- filter(bihar_data, adult==1, female==0)` ✓

☐ `bihar_adult_males <- filter(bihar_data, adult=1, male=1)`

Explanation

In the dataset which Professor Duflo uses in this lecture, the “female” and the “adult” variables are binary variables. This means that “female” is represented as either “yes” or “no” (or “true” or “false”). There is no variable “male” so if we wanted to select only adult males, we would use either `female==0`.

Plotting Histograms in R: Part Two - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

What aesthetic problems were discussed in the original graph output? [Select all that apply]

☐ The colors were too light


☒ The bins are too large ✓

☒ There are too many outliers ✓

☐ The y-axis needs to be re-scaled

Explanation

As the students in the lecture have pointed out, large bins and too many outliers seem to be pertinent issues. For example, there are data points that exist which indicate that some adult females in Bihar are 0 centimeters tall which cannot be correct. Additionally, the larger bin size may hide interesting patterns in the data.

 Answers are displayed within the problem

Question 2

1 point possible (graded)

True or False: It is considered bad practice to create subsets of datasets (i.e., truncating data).

☐ True

☒ False ✓

Explanation

No, this is not considered bad practice. In fact, Professor Duflo creates a subset of the Bihar dataset to remove the outliers by using the `filter()` method in R. She explains that doing this does not waste computer memory. Alternatively, one could remove outliers by defining a range in the histogram.

[Show answer](#)

Submit

Plotting Histograms in R: Part Three - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

1 point possible (graded)

True or false: A bin width of 10 means that there are 10 total bins in the histogram.

☐ True

☒ False ✓

Explanation

A bin width of 10 means that the data is graphed in groups of 10 by whatever the x-axis variable is. In this example, this means that each bin represents a range in centimeters, which go up by 10 centimeters in each column. Histograms show trends in data; they do not display individual centimeters.

[Show answer](#)

Submit

You have used 0 of 1 attempt

Question 2

1 point possible (graded)

True or False: Making bins narrower is always a good thing because the graph shows more specific data patterns.

☐ True

☒ False ✓

Explanation

Defining bin width is a choice between how much information to convey versus how much knowledge can be tolerated. Bins which are too narrow show too much individual data (some of which happens by chance!) and does not allow an underlying pattern to be identified easily because there may be holes or outliers. Bins which are too wide may also camouflage interesting patterns.

[Show answer](#)

Submit

You have used 0 of 1 attempt

i Answers are displayed within the problem

Question 1

0.0/1.0 point (graded)

Which of the following could be suitable kernel functions? (Select all which apply)

☒ A hump-shaped (inverted U) function ✓

☐ A V-Shaped function

☒ A pyramid-shaped (inverted V) function ✓

☒ A bell-shaped function ✓

Explanation

The goal of kernel density estimation is to estimate random variables' probability density functions. We turn to kernel density estimates to obtain a smoother, less variable representation of the underlying data than a histogram. Intuitively, any function that weights observations on the boundary of the intervals more than observations at the center of the interval surrounding a given point, will lead to higher variance. This would defeat the purpose of a kernel, as it would result in a less smooth estimator.

Question 2

1 point possible (graded)

For a given functional form of the Kernel weighting function, what determines its height? Note that height is defined as the maximal height of kernel weighting function across the domain of the PDF.

- ☐ The height should be chosen optimally
- ☐ The number of observations
- ☒ The bandwidth ✓
- ☐ The probability density function of the underlying random variable

Explanation

Since the kernel function integrates to 1, and the bandwidth represents the (fixed) width of the interval over which it is evaluated, the bandwidth determines the limits of the integral, and thus determines the height of the kernel function.

[Show answer](#)

Submit

You have used 0 of 2 attempts

Kernel Density Estimation in R - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

1 point possible (graded)

True or False: The default `geom_histogram()` output is a count histogram (frequency on the y-axis).

☒ True ✓

☐ False


Explanation

To specify a density histogram, the `aes()` argument must contain: `..density..`

[Show answer](#)

Submit

You have used 0 of 1 attempt

 Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

When creating a kernel density graph in R, what are the two most important method arguments discussed in lecture? [Select all that apply]

☒ Type of kernel ✓

☐ Range

☒ Bandwidth ✓

☐ Color

Explanation

As discussed in lecture, the two most important arguments for a kernel method are (1) type of kernel (i.e., Gaussian) and (2) bandwidth (i.e., height). Of course, additional arguments can be inputted as well.

[Show answer](#)

Bandwidth in Kernel Functions - Quiz

[Bookmark this page](#)

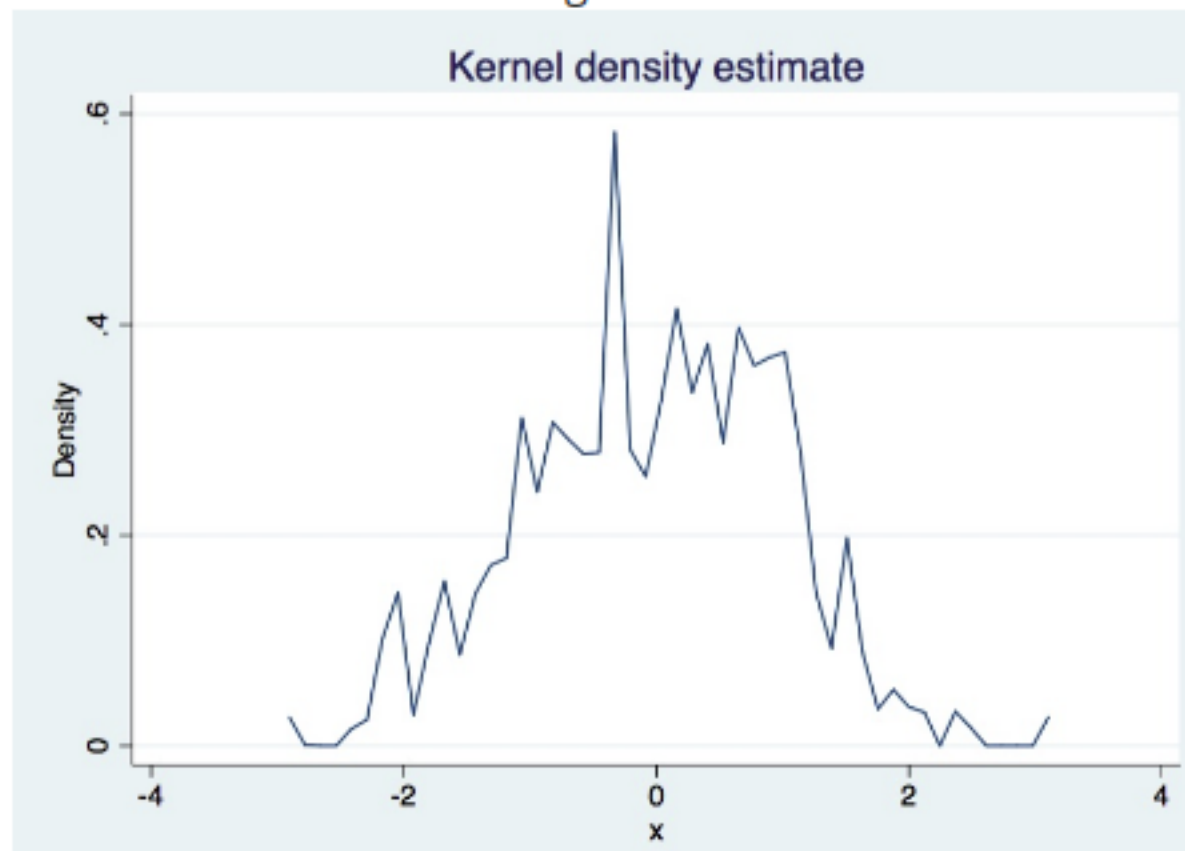
Finger Exercises due Sep 28, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

The set of figures shown below present the kernel probability density estimates from data sampled from a standard normal distribution. What parameter is changing from Figure A to Figure B to Figure C?

Figure A



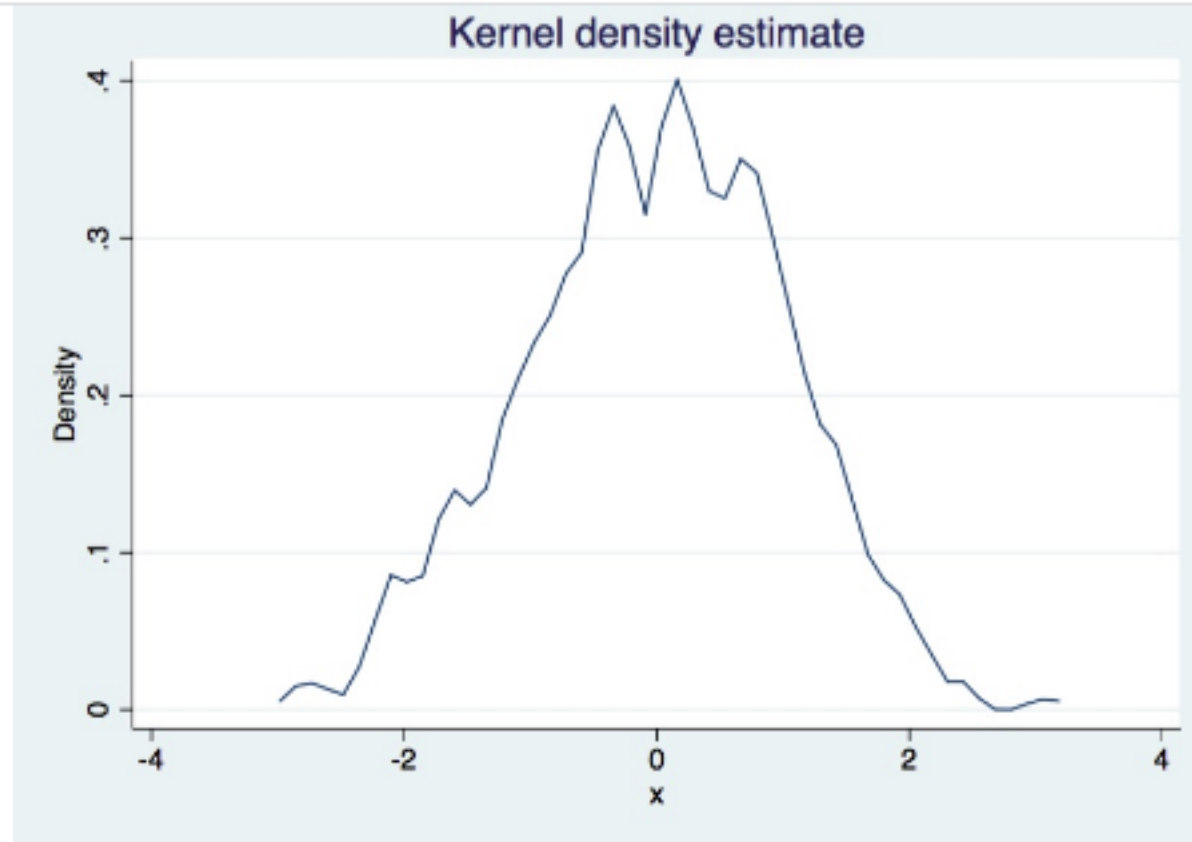
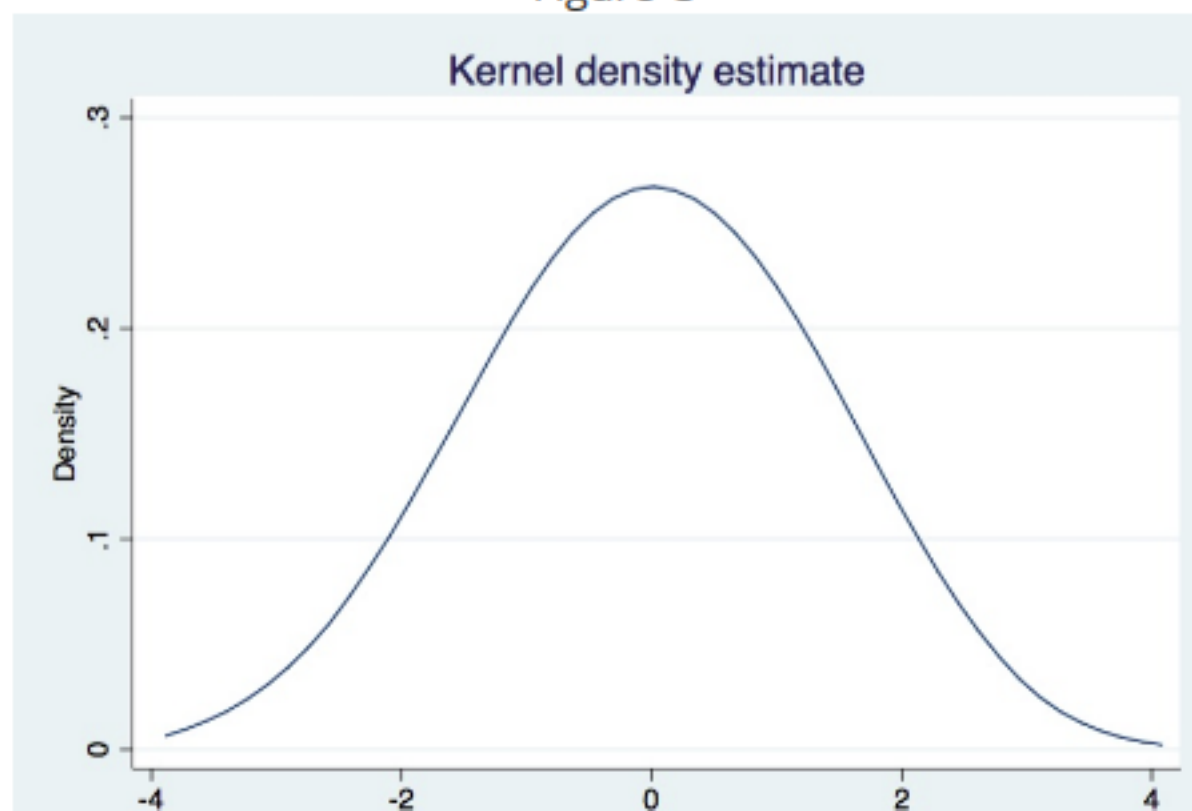


Figure C



☒ the bandwidth is increasing ✓

☐ the kernel function is different

☐ the bandwidth is decreasing

Explanation

As discussed in this segment, the parameter h , the bandwidth of the estimating function, controls the smoothness and corresponds to the bin width of the histogram. If h is too small, the estimate is too rough; if it is too large, then the resulting estimate of the function is too smooth, since it obscures the shape, and spreads the probability mass out too much, and hence, is biased.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

True or False: Error due to bias is caused from being too sensitive to fluctuations in a dataset; it can cause additional noise and may fail to model accurate future observations (i.e., overfitting).

☐ True

☒ False ✓

Explanation

This is false—this is true for error due to variance. Bias is the difference between the model's expected predictions and the correct values. Error due to bias is caused by false, simplified assumptions in the model that do not understand the dataset's complexity, which leads to underfitting. This dilemma is referred to as the bias-variance tradeoff.

[Show answer](#)

Submit

You have used 0 of 1 attempt

Question 1

0.0/1.0 point (graded)


The histograms of adult female height in the US and Bihar show that:

- ☐ Women from the US are shorter on average than those from Bihar, and the distribution of heights has more dispersion than Bihar
- ☒ Women from the US are taller on average than those from Bihar, and the distribution of heights has more dispersion than Bihar ✓
- ☐ Women from the US are shorter on average than those from Bihar, and the distribution of heights has less dispersion than Bihar
- ☐ Women from the US are taller on average than those from Bihar, and the distribution of heights has less dispersion than Bihar

Explanation

The plot shown in class suggests that women in the US are on average taller than women in Bihar as the US graph peaks farther to the right. Furthermore, it is more dispersed as there seems to be a wider range.

[Show answer](#)

 Answers are displayed within the problem

Question 2

1 point possible (graded)

True or False: When comparing distributions, density histograms should be used if the distributions have a different number of observations.

☒ True ✓

☐ False

Explanation

This is true as discussed in class and in the question/answer above. The density histogram will better scale both distributions as only the proportions matter.

[Show answer](#)

Submit

You have used 0 of 1 attempt

Plotting Cumulative Distribution Functions - Quiz

[Bookmark this page](#)

Finger Exercises due Sep 28, 2020 19:30 EDT *Past Due*

Question 1

1 point possible (graded)

True or False: The cumulative distribution function (CDF) is the integral of the probability density function (PDF).

☒ True ✓

☐ False

Explanation

Knowing this, the opposite is also true: the PDF is the derivative of the CDF.

[Show answer](#)

Submit

You have used 0 of 1 attempt

Submit

You have used 0 of 1 attempt

i Answers are displayed within the problem

Question 2

0.0/1.0 point (graded)

A cumulative histogram is to the

Answer: CDF what the histogram is to the

Answer: PDF .

Explanation

A histogram contains information about the frequency of observations within each interval. Dividing the frequency by the total number of observations gives us the density. On the other hand, a cumulative histogram conveys information on the “cumulative” frequency/density and hence provides an idea of what the CDF would look like. You can think of it as a running count as you move across bins, whereas a normal histogram resets the count for each bin.

[Show answer](#)

Submit

You have used 0 of 2 attempts

i Answers are displayed within the problem