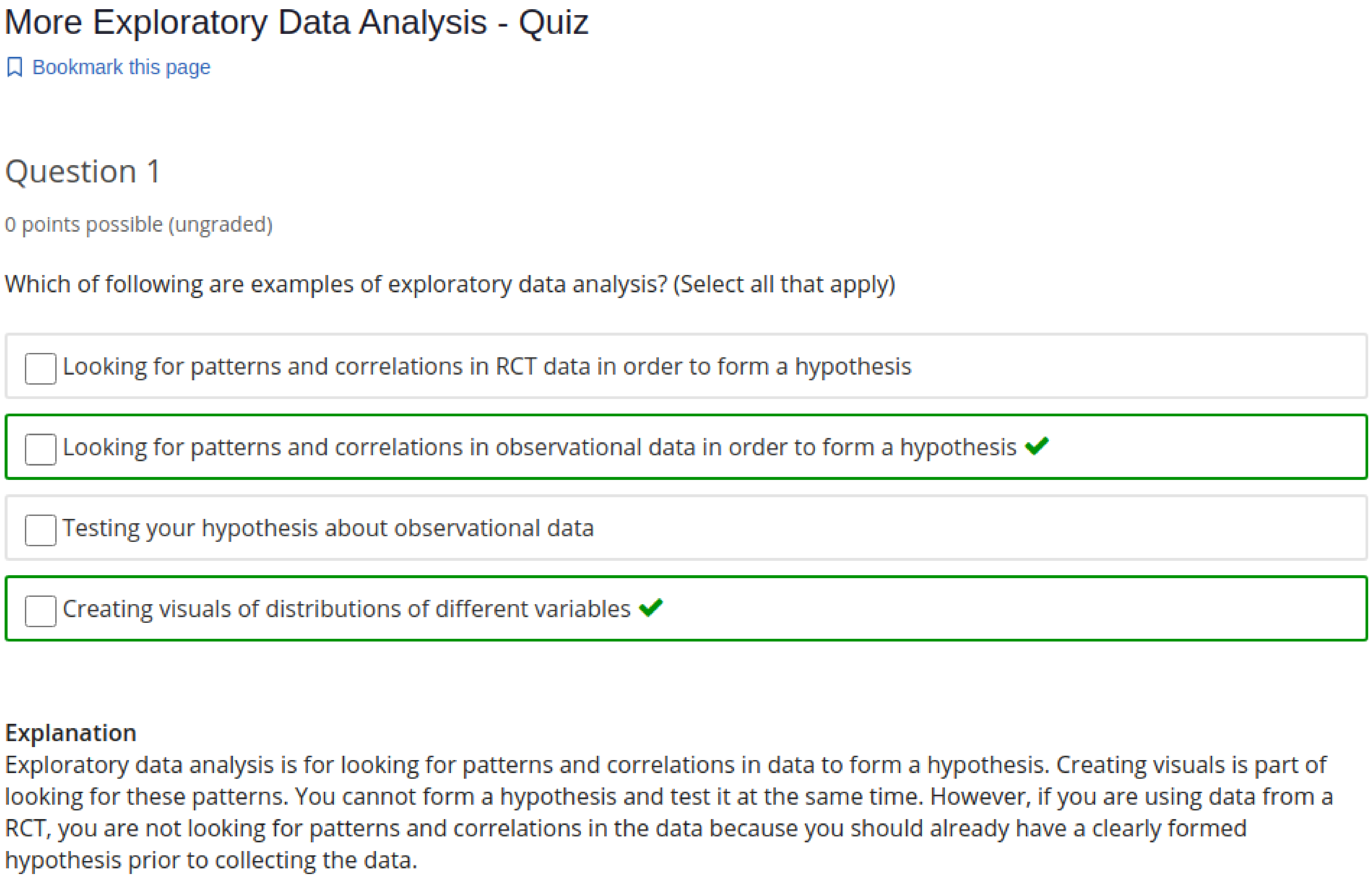


More Exploratory Data Analysis - Quiz

Bookmark this page

Question 1

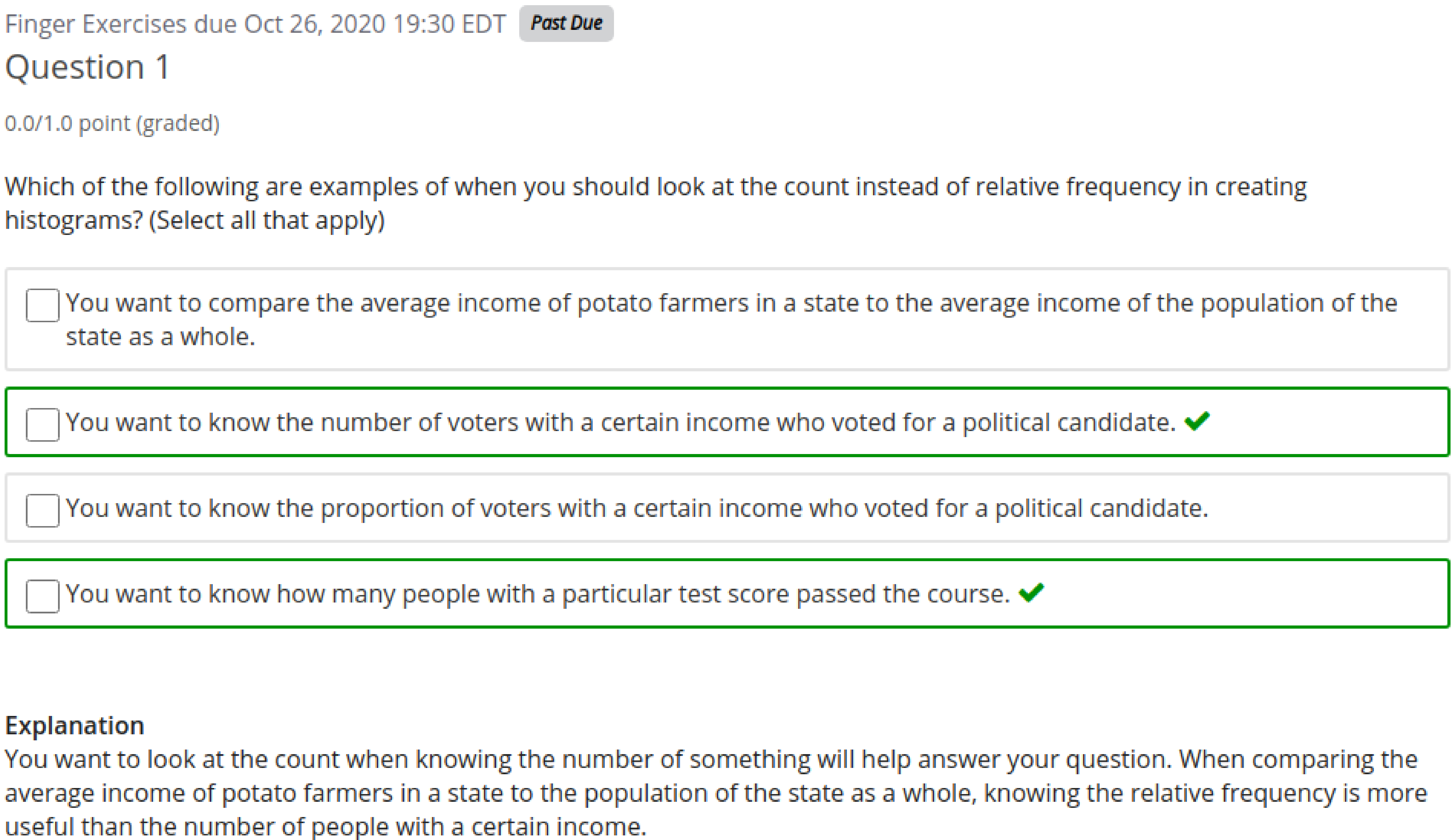
0 points possible (ungraded)

Which of following are examples of exploratory data analysis? (Select all that apply)

- □ Looking for patterns and correlations in RCT data in order to form a hypothesis
- □ Looking for patterns and correlations in observational data in order to form a hypothesis ✓
- □ Testing your hypothesis about observational data
- □ Creating visuals of distributions of different variables ✓

Explanation

Exploratory data analysis is for looking for patterns and correlations in data to form a hypothesis. Creating visuals is part of looking for these patterns. You cannot form a hypothesis and test it at the same time. However, if you are using data from a RCT, you are not looking for patterns and correlations in the data because you should already have a clearly formed hypothesis prior to collecting the data.



Finger Exercises due Oct 26, 2020 19:30 EDT **_Past Due_**

Question 1

0.0/1.0 point (graded)

Which of the following are examples of when you should look at the count instead of relative frequency in creating histograms? (Select all that apply)

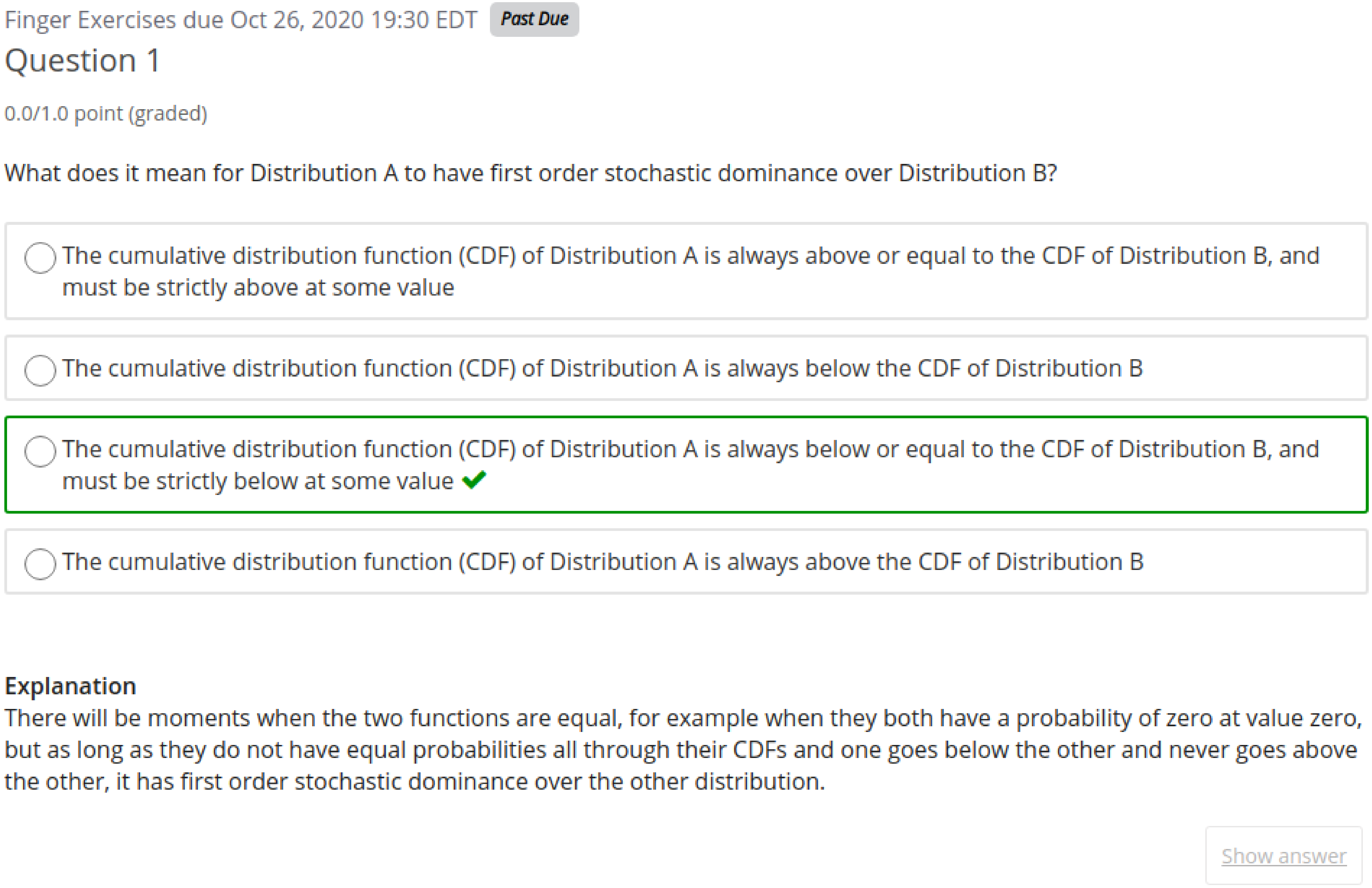
□ You want to compare the average income of potato farmers in a state to the average income of the population of the state as a whole.

□ You want to know the number of voters with a certain income who voted for a political candidate. ✓

□ You want to know the proportion of voters with a certain income who voted for a political candidate.

□ You want to know how many people with a particular test score passed the course. ✓

Explanation
You want to look at the count when knowing the number of something will help answer your question. When comparing the average income of potato farmers in a state to the population of the state as a whole, knowing the relative frequency is more useful than the number of people with a certain income.



Finger Exercises due Oct 26, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

What does it mean for Distribution A to have first order stochastic dominance over Distribution B?

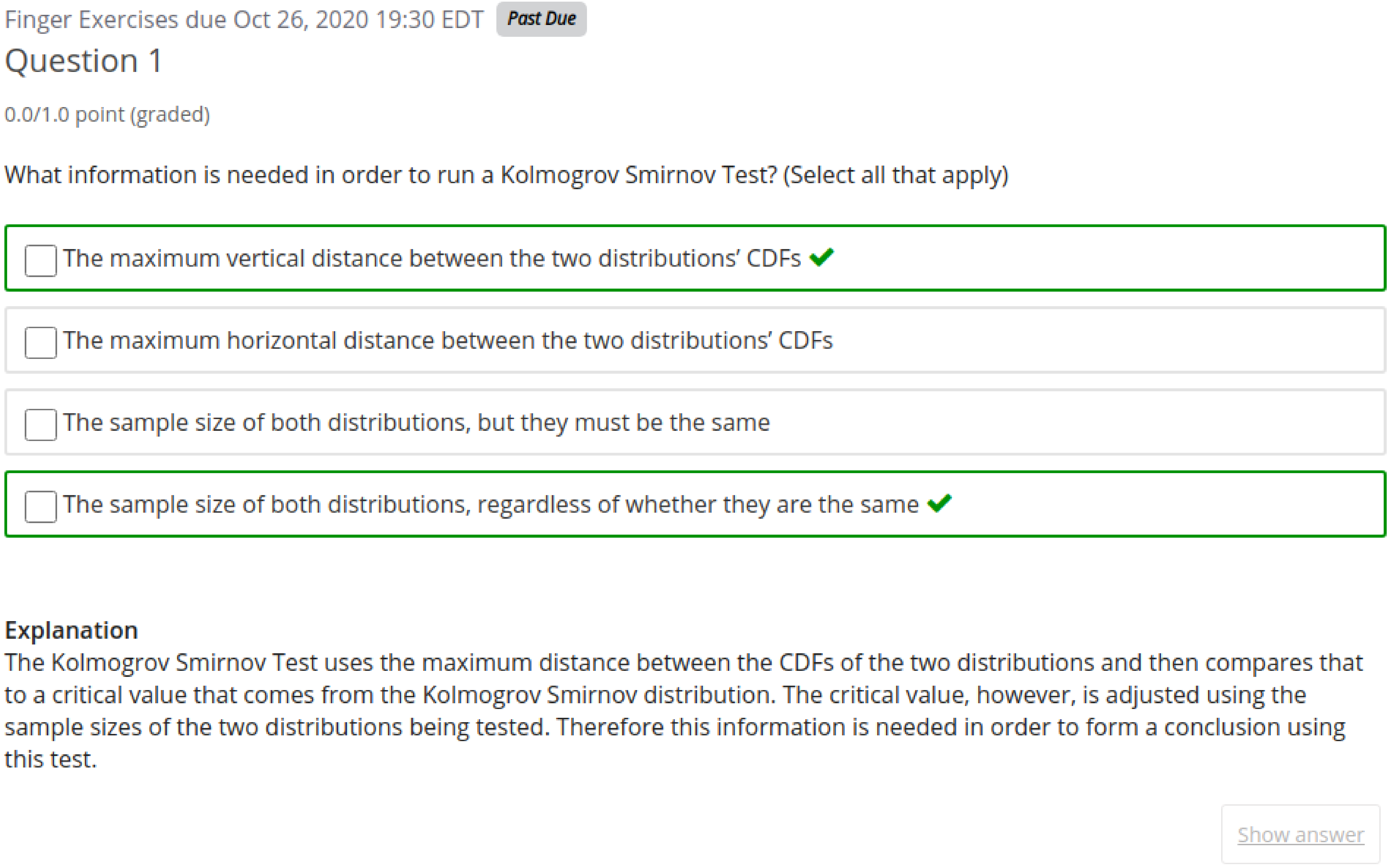
○ The cumulative distribution function (CDF) of Distribution A is always above or equal to the CDF of Distribution B, and must be strictly above at some value

○ The cumulative distribution function (CDF) of Distribution A is always below the CDF of Distribution B

○ The cumulative distribution function (CDF) of Distribution A is always below or equal to the CDF of Distribution B, and must be strictly below at some value ✓

○ The cumulative distribution function (CDF) of Distribution A is always above the CDF of Distribution B

Explanation
There will be moments when the two functions are equal, for example when they both have a probability of zero at value zero, but as long as they do not have equal probabilities all through their CDFs and one goes below the other and never goes above the other, it has first order stochastic dominance over the other distribution.



Finger Exercises due Oct 26, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

What information is needed in order to run a Kolmogrov Smirnov Test? (Select all that apply)

□ The maximum vertical distance between the two distributions' CDFs ✓

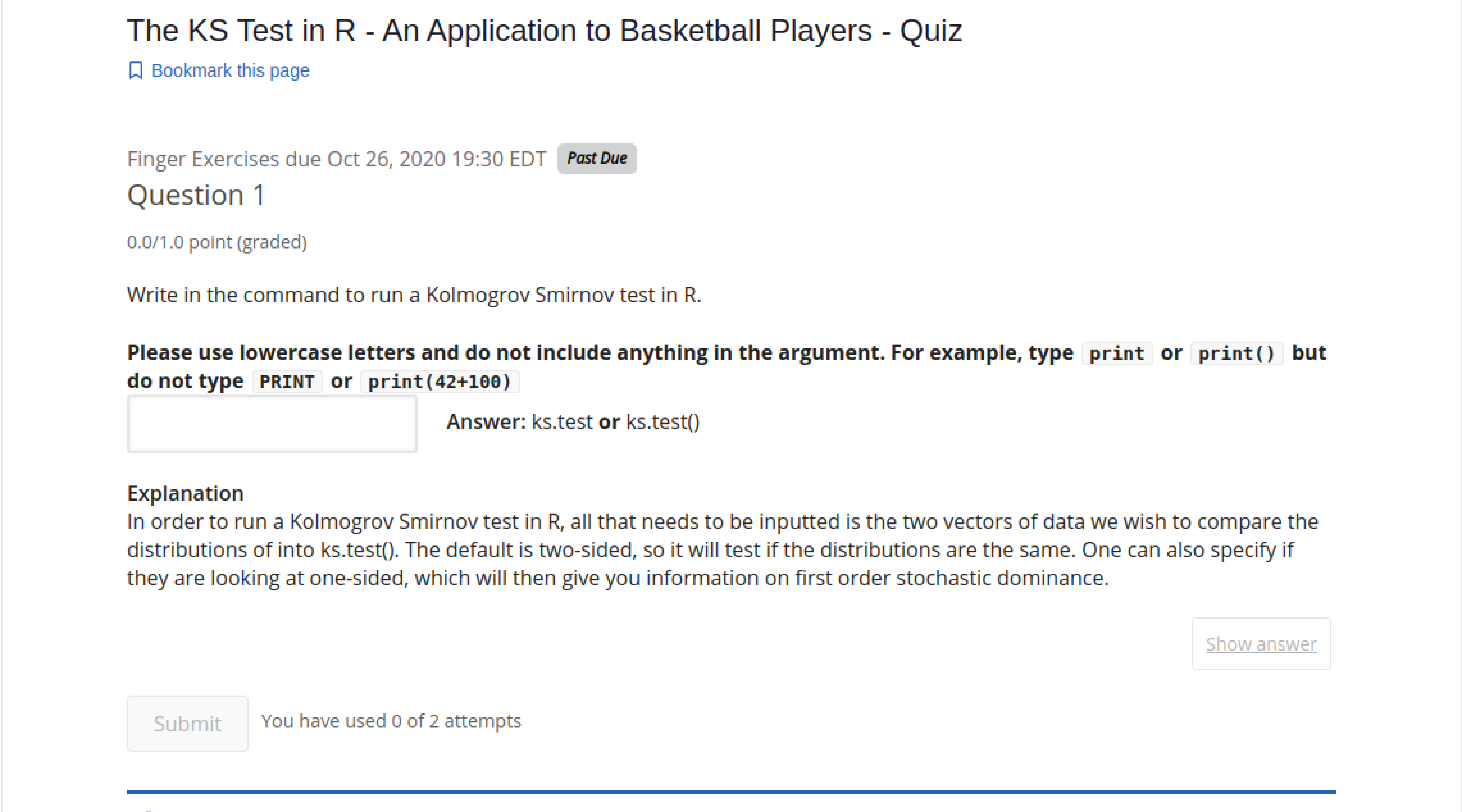
□ The maximum horizontal distance between the two distributions' CDFs

□ The sample size of both distributions, but they must be the same

□ The sample size of both distributions, regardless of whether they are the same ✓

Explanation

The Kolmogrov Smirnov Test uses the maximum distance between the CDFs of the two distributions and then compares that to a critical value that comes from the Kolmogrov Smirnov distribution. The critical value, however, is adjusted using the sample sizes of the two distributions being tested. Therefore this information is needed in order to form a conclusion using this test.



The KS Test in R - An Application to Basketball Players - Quiz

Finger Exercises due Oct 26, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

Write in the command to run a Kolmogrov Smirnov test in R.

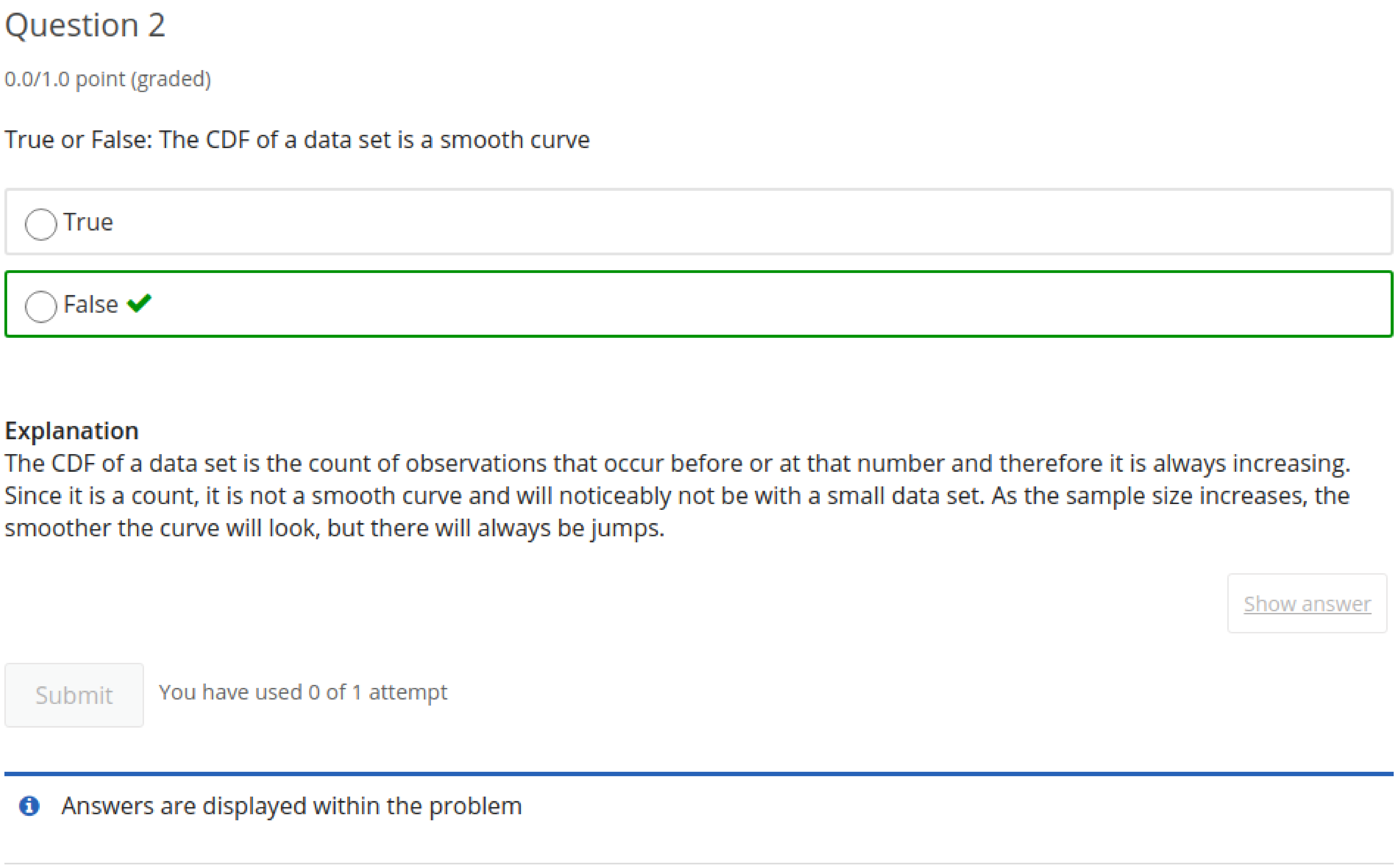
Please use lowercase letters and do not include anything in the argument. For example, type `print` or `print()` but do not type `PRINT` or `print(42+100)`

Answer: ks.test **or** ks.test()

Explanation

In order to run a Kolmogrov Smirnov test in R, all that needs to be inputted is the two vectors of data we wish to compare the distributions of into ks.test(). The default is two-sided, so it will test if the distributions are the same. One can also specify if they are looking at one-sided, which will then give you information on first order stochastic dominance.

You have used 0 of 2 attempts



Question 2

0.0/1.0 point (graded)

True or False: The CDF of a data set is a smooth curve

○ True

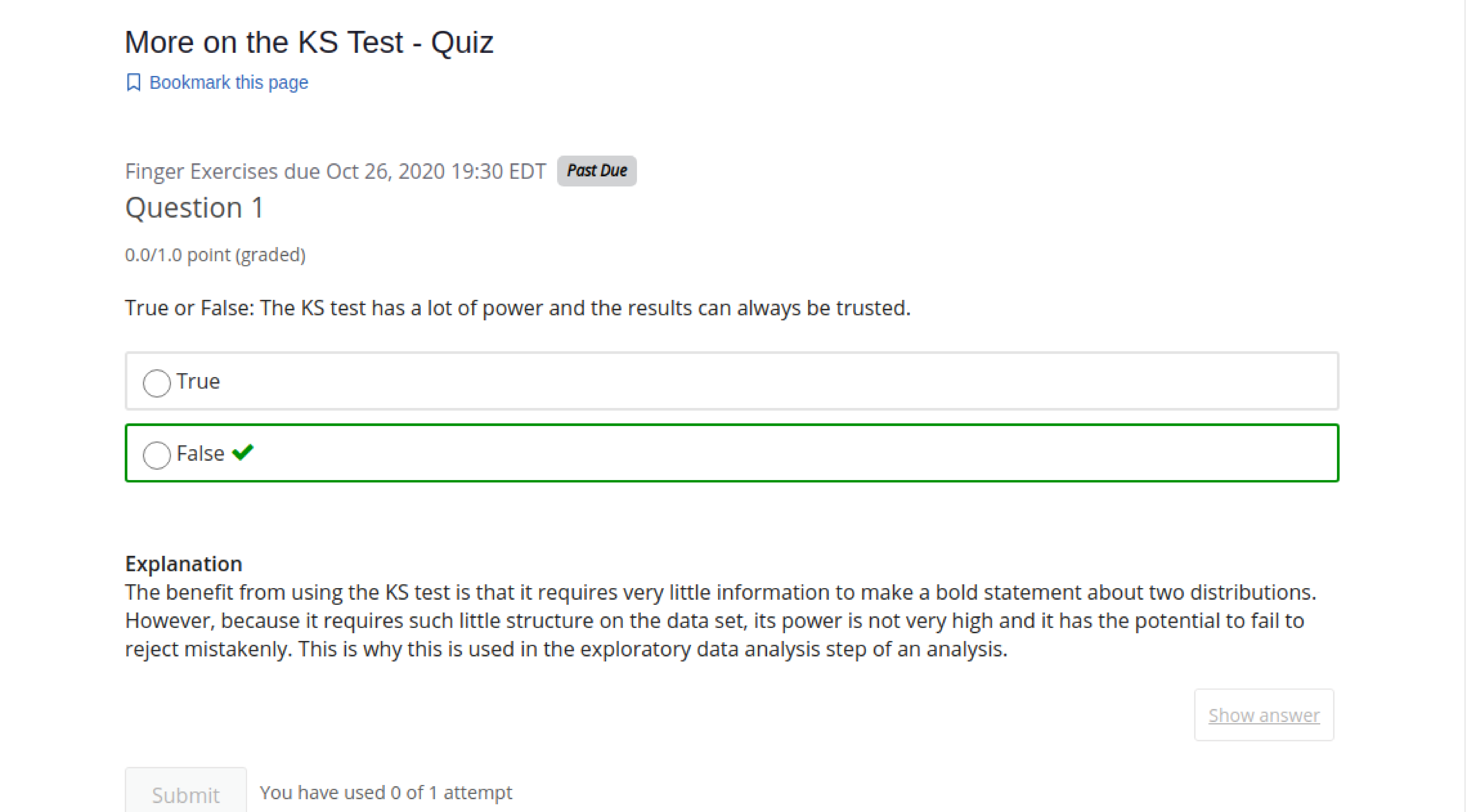
○ False ✓

Explanation

The CDF of a data set is the count of observations that occur before or at that number and therefore it is always increasing. Since it is a count, it is not a smooth curve and will noticeably not be with a small data set. As the sample size increases, the smoother the curve will look, but there will always be jumps.

You have used 0 of 1 attempt

Answers are displayed within the problem



More on the KS Test - Quiz

Finger Exercises due Oct 26, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

True or False: The KS test has a lot of power and the results can always be trusted.

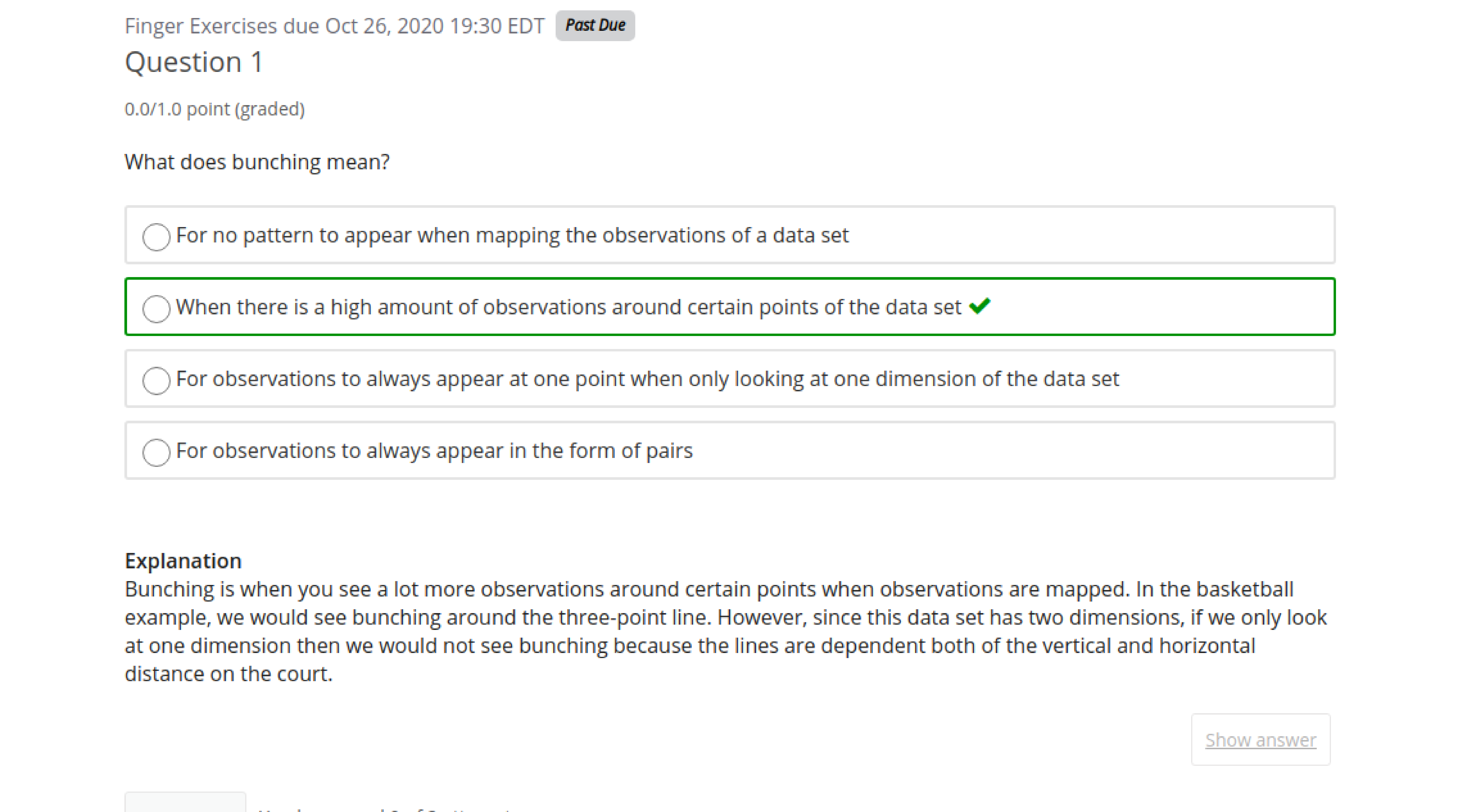
○ True

○ False ✓

Explanation

The benefit from using the KS test is that it requires very little information to make a bold statement about two distributions. However, because it requires such little structure on the data set, its power is not very high and it has the potential to fail to reject mistakenly. This is why this is used in the exploratory data analysis step of an analysis.

You have used 0 of 1 attempt



Finger Exercises due Oct 26, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

What does bunching mean?

○ For no pattern to appear when mapping the observations of a data set

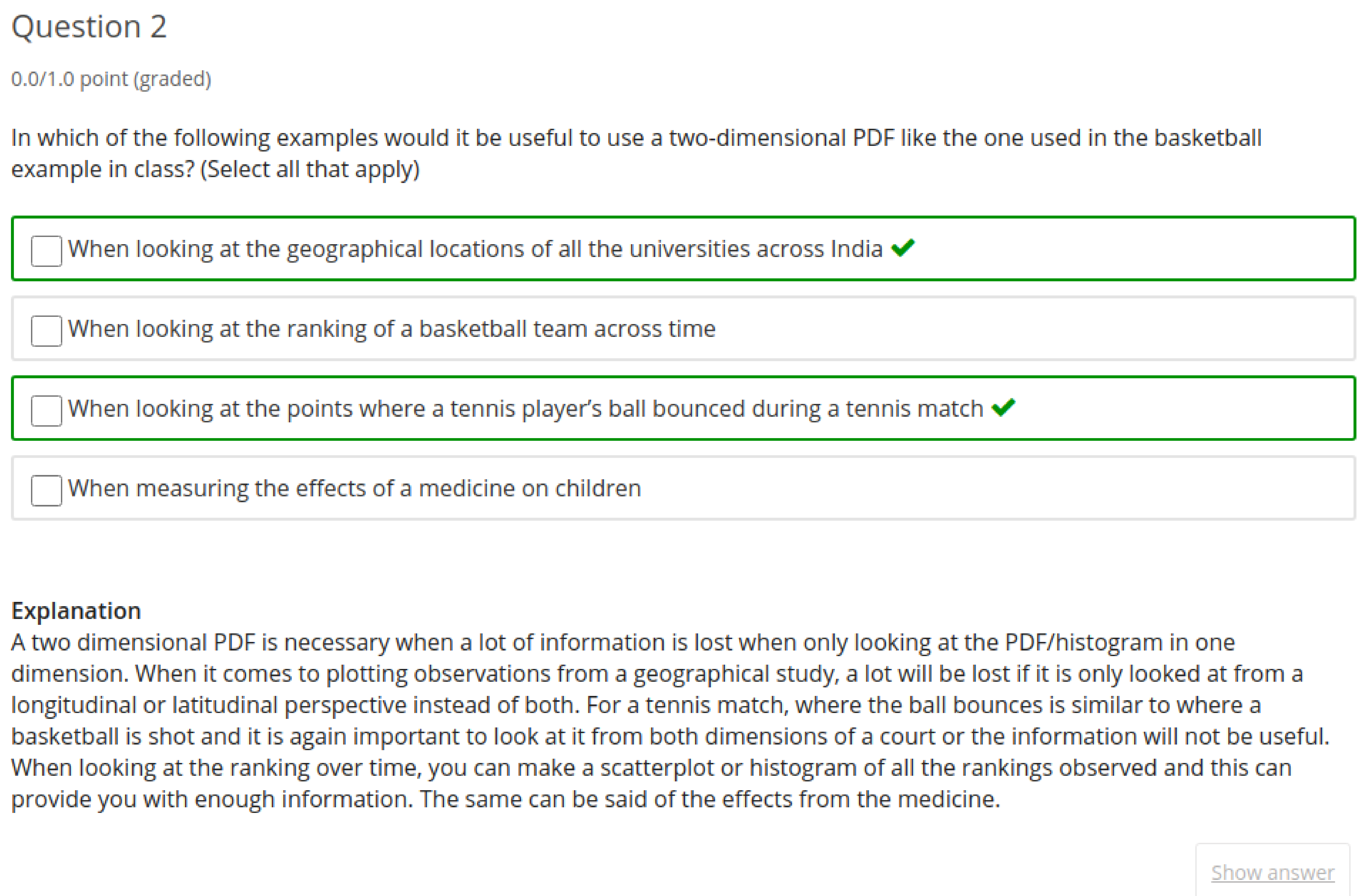
○ When there is a high amount of observations around certain points of the data set ✓

○ For observations to always appear at one point when only looking at one dimension of the data set

○ For observations to always appear in the form of pairs

Explanation

Bunching is when you see a lot more observations around certain points when observations are mapped. In the basketball example, we would see bunching around the three-point line. However, since this data set has two dimensions, if we only look at one dimension then we would not see bunching because the lines are dependent both of the vertical and horizontal distance on the court.



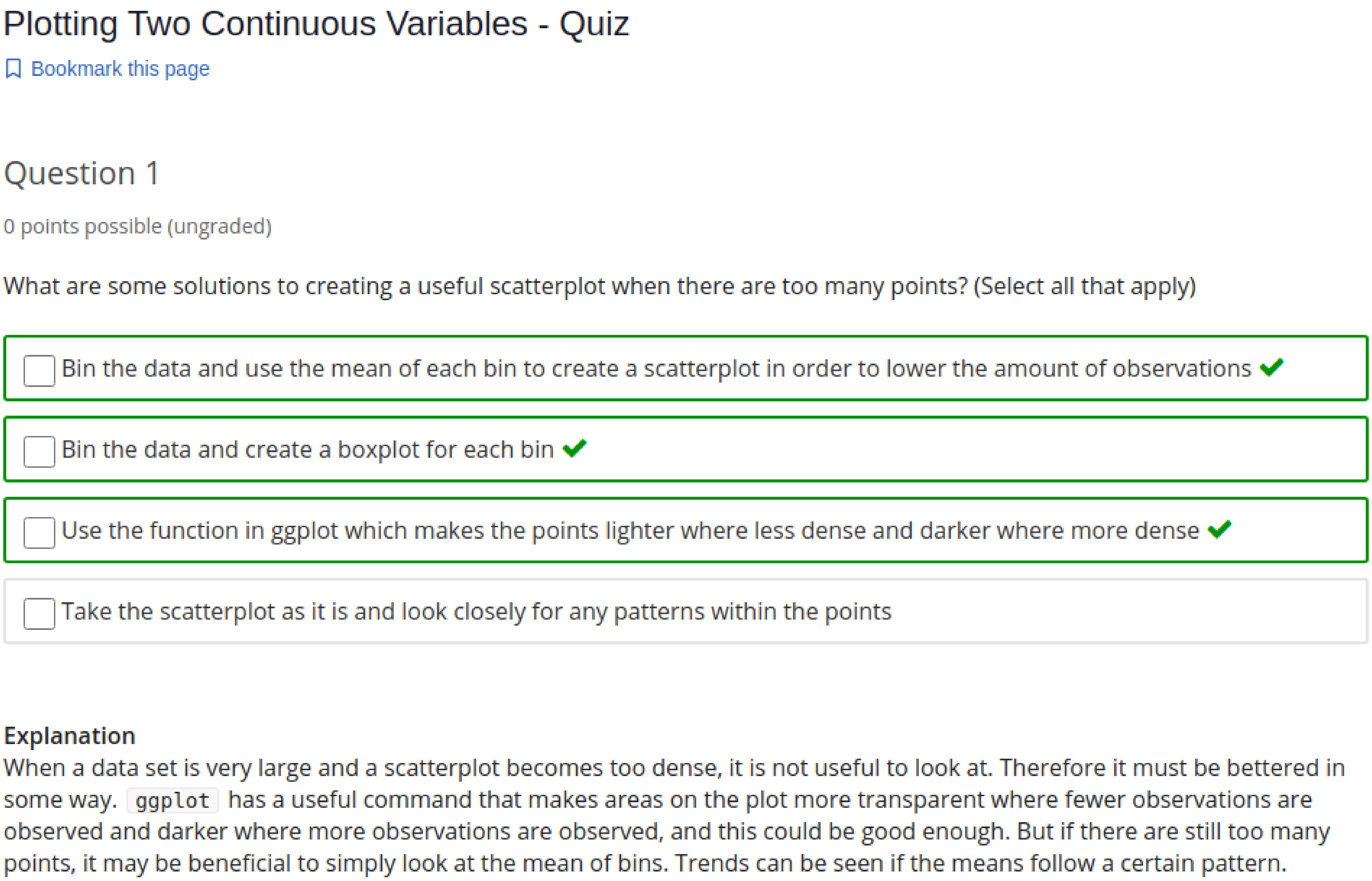
Question 2

0.0/1.0 point (graded)

In which of the following examples would it be useful to use a two-dimensional PDF like the one used in the basketball example in class? (Select all that apply)

- □ When looking at the geographical locations of all the universities across India ✓
- □ When looking at the ranking of a basketball team across time
- □ When looking at the points where a tennis player’s ball bounced during a tennis match ✓
- □ When measuring the effects of a medicine on children

Explanation
A two dimensional PDF is necessary when a lot of information is lost when only looking at the PDF/histogram in one dimension. When it comes to plotting observations from a geographical study, a lot will be lost if it is only looked at from a longitudinal or latitudinal perspective instead of both. For a tennis match, where the ball bounces is similar to where a basketball is shot and it is again important to look at it from both dimensions of a court or the information will not be useful. When looking at the ranking over time, you can make a scatterplot or histogram of all the rankings observed and this can provide you with enough information. The same can be said of the effects from the medicine.



Plotting Two Continuous Variables - Quiz

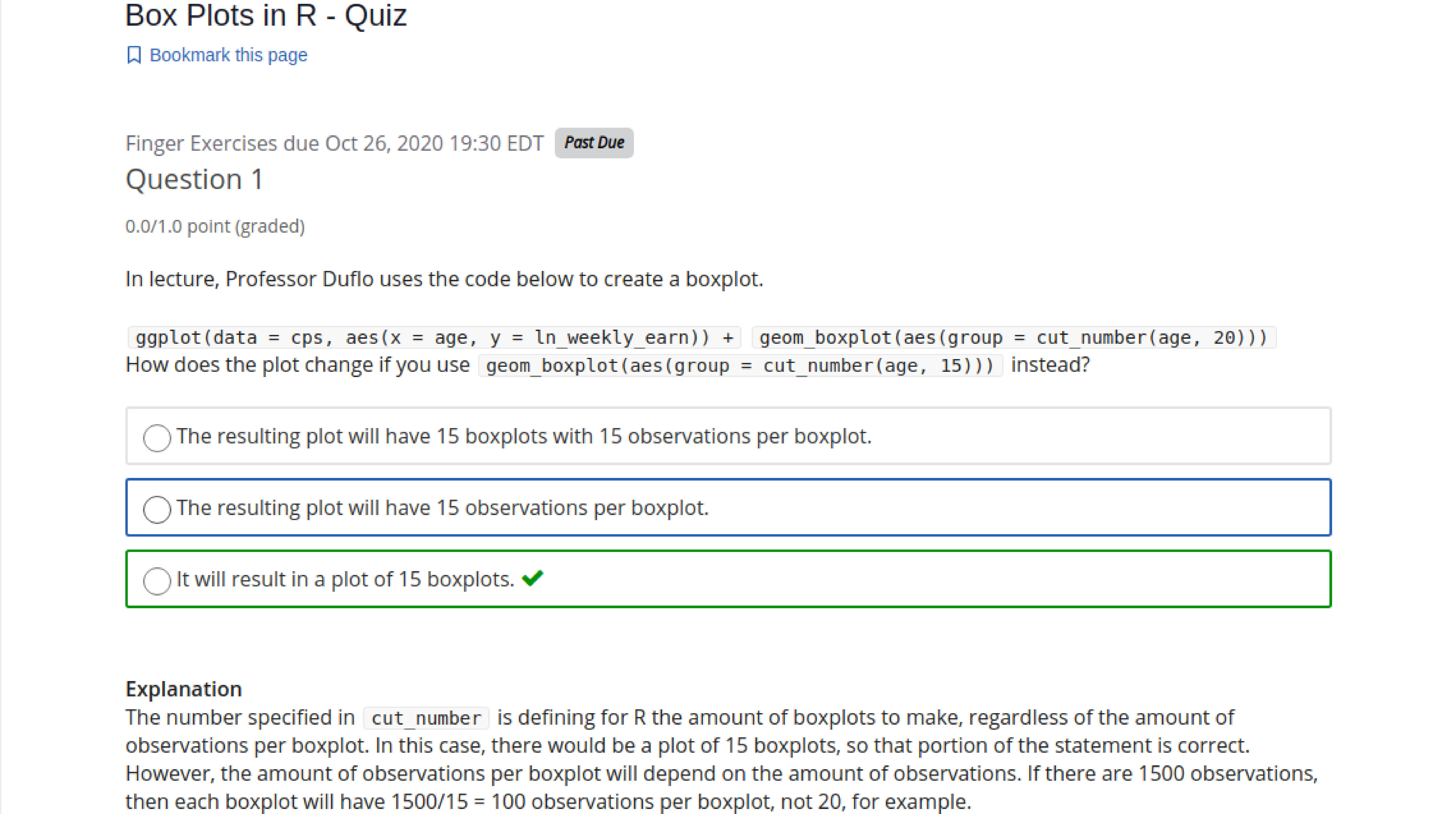
Question 1

0 points possible (ungraded)

What are some solutions to creating a useful scatterplot when there are too many points? (Select all that apply)

- □ Bin the data and use the mean of each bin to create a scatterplot in order to lower the amount of observations ✓
- □ Bin the data and create a boxplot for each bin ✓
- □ Use the function in ggplot which makes the points lighter where less dense and darker where more dense ✓
- □ Take the scatterplot as it is and look closely for any patterns within the points

Explanation
When a data set is very large and a scatterplot becomes too dense, it is not useful to look at. Therefore it must be bettered in some way. `ggplot` has a useful command that makes areas on the plot more transparent where fewer observations are observed and darker where more observations are observed, and this could be good enough. But if there are still too many points, it may be beneficial to simply look at the mean of bins. Trends can be seen if the means follow a certain pattern.



Box Plots in R - Quiz

Bookmark this page

Finger Exercises due Oct 26, 2020 19:30 EDT **Past Due**

Question 1

0.0/1.0 point (graded)

In lecture, Professor Duflo uses the code below to create a boxplot.

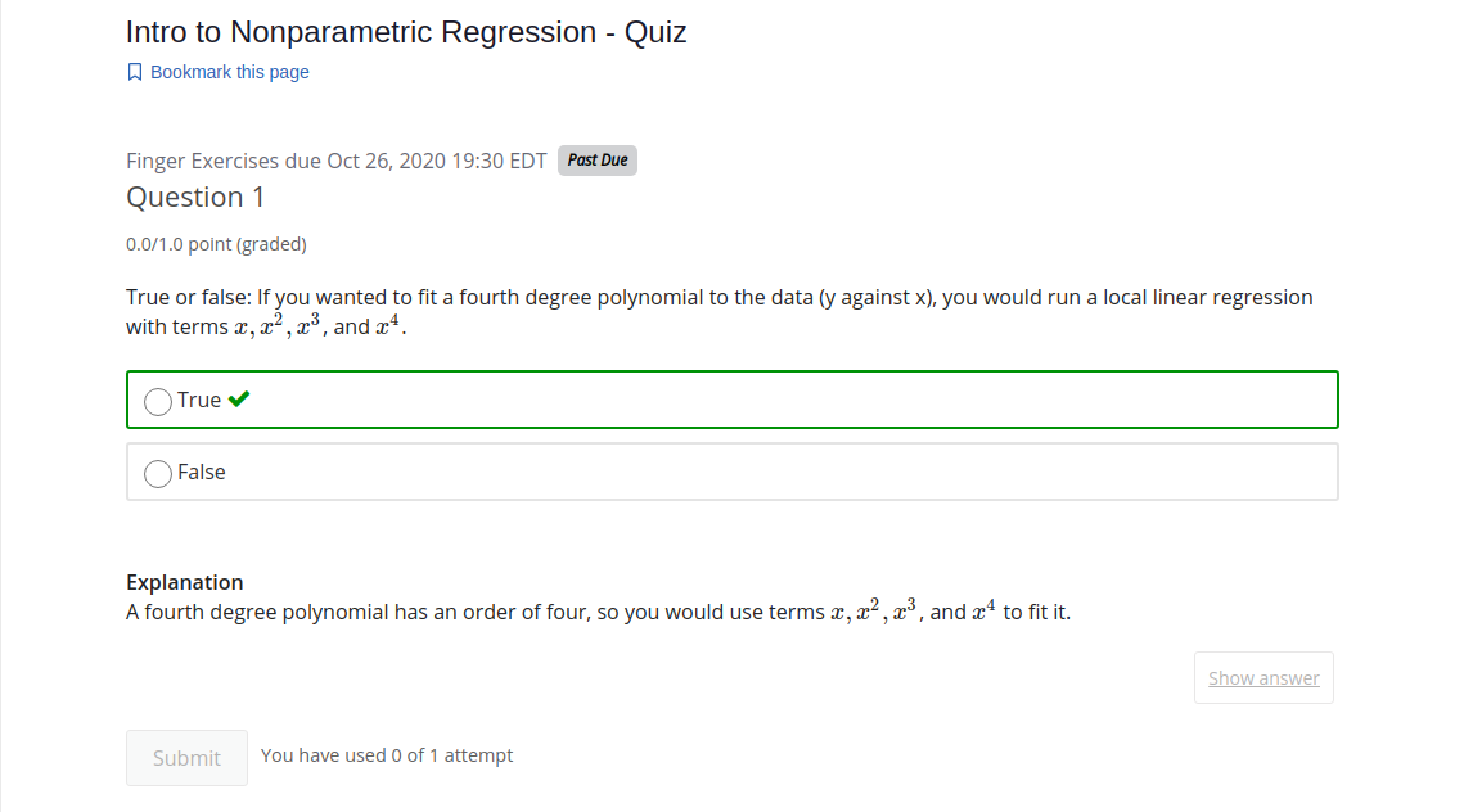
`ggplot(data = cps, aes(x = age, y = ln_weekly_earn)) +` `geom_boxplot(aes(group = cut_number(age, 20)))`
How does the plot change if you use `geom_boxplot(aes(group = cut_number(age, 15)))` instead?

○ The resulting plot will have 15 boxplots with 15 observations per boxplot.

○ The resulting plot will have 15 observations per boxplot.

○ It will result in a plot of 15 boxplots. ✓

Explanation
The number specified in `cut_number` is defining for R the amount of boxplots to make, regardless of the amount of observations per boxplot. In this case, there would be a plot of 15 boxplots, so that portion of the statement is correct. However, the amount of observations per boxplot will depend on the amount of observations. If there are 1500 observations, then each boxplot will have 1500/15 = 100 observations per boxplot, not 20, for example.



Intro to Nonparametric Regression - Quiz

Finger Exercises due Oct 26, 2020 19:30 EDT *Past Due*

Question 1

0.0/1.0 point (graded)

True or false: If you wanted to fit a fourth degree polynomial to the data (y against x), you would run a local linear regression with terms x, x^2, x^3, and x^4.

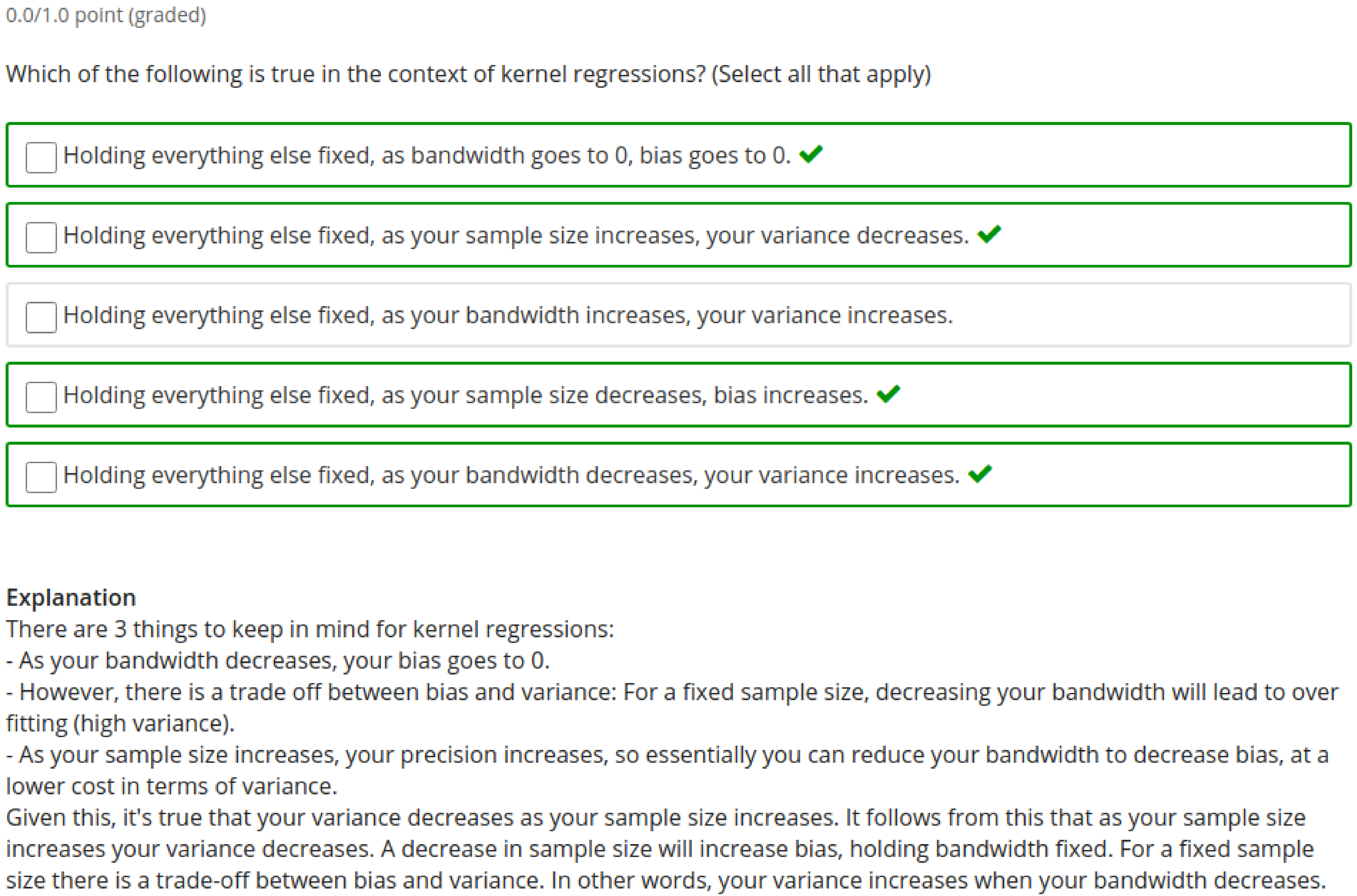
○ True ✓

○ False

Explanation

A fourth degree polynomial has an order of four, so you would use terms x, x^2, x^3, and x^4 to fit it.

You have used 0 of 1 attempt



0.0/1.0 point (graded)

Which of the following is true in the context of kernel regressions? (Select all that apply)

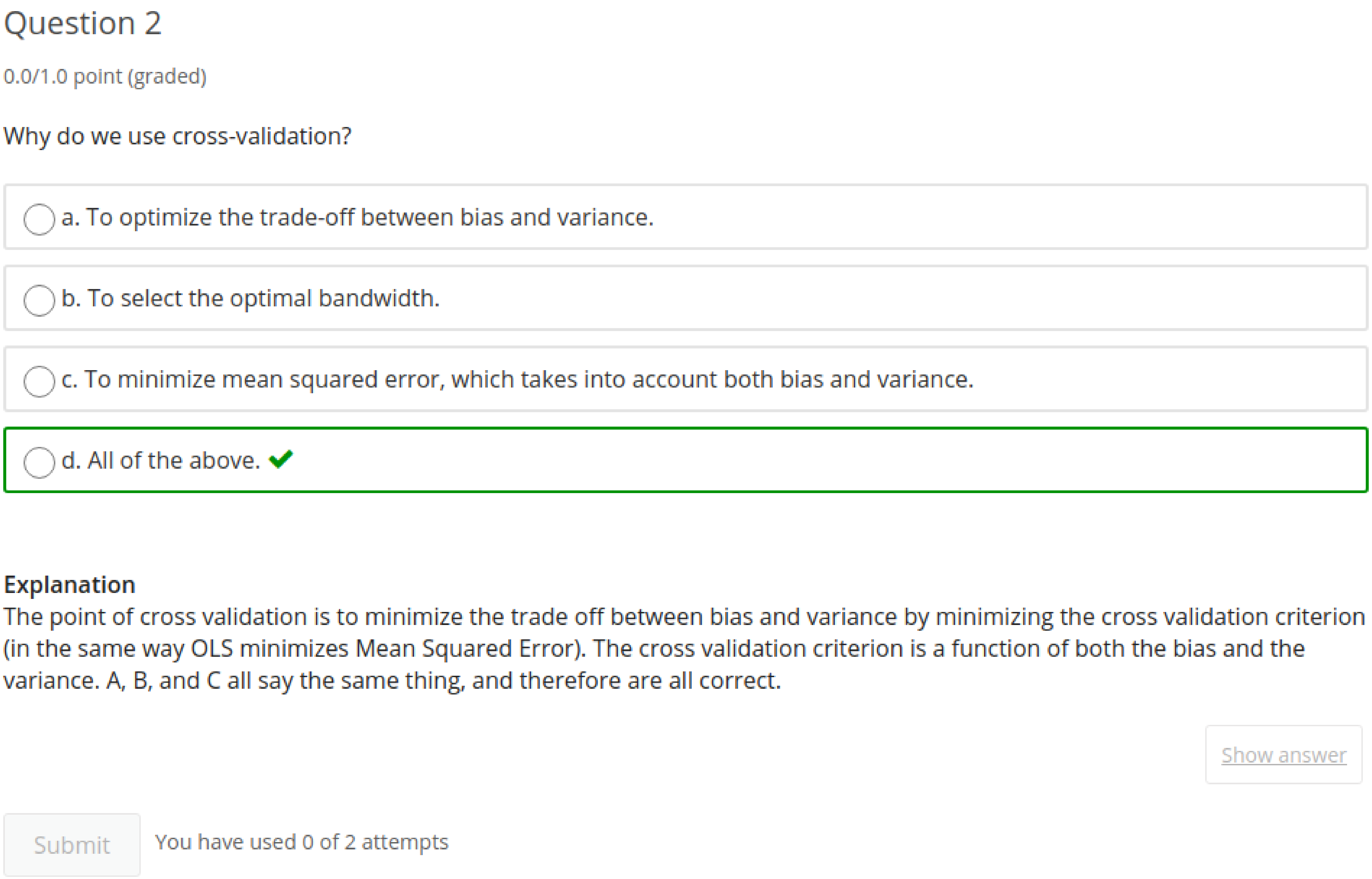
- □ Holding everything else fixed, as bandwidth goes to 0, bias goes to 0. ✓
- □ Holding everything else fixed, as your sample size increases, your variance decreases. ✓
- □ Holding everything else fixed, as your bandwidth increases, your variance increases.
- □ Holding everything else fixed, as your sample size decreases, bias increases. ✓
- □ Holding everything else fixed, as your bandwidth decreases, your variance increases. ✓

Explanation

There are 3 things to keep in mind for kernel regressions:

- As your bandwidth decreases, your bias goes to 0.
- However, there is a trade off between bias and variance: For a fixed sample size, decreasing your bandwidth will lead to over fitting (high variance).
- As your sample size increases, your precision increases, so essentially you can reduce your bandwidth to decrease bias, at a lower cost in terms of variance.

Given this, it's true that your variance decreases as your sample size increases. It follows from this that as your sample size increases your variance decreases. A decrease in sample size will increase bias, holding bandwidth fixed. For a fixed sample size there is a trade-off between bias and variance. In other words, your variance increases when your bandwidth decreases.



Question 2

0.0/1.0 point (graded)

Why do we use cross-validation?

○ a. To optimize the trade-off between bias and variance.

○ b. To select the optimal bandwidth.

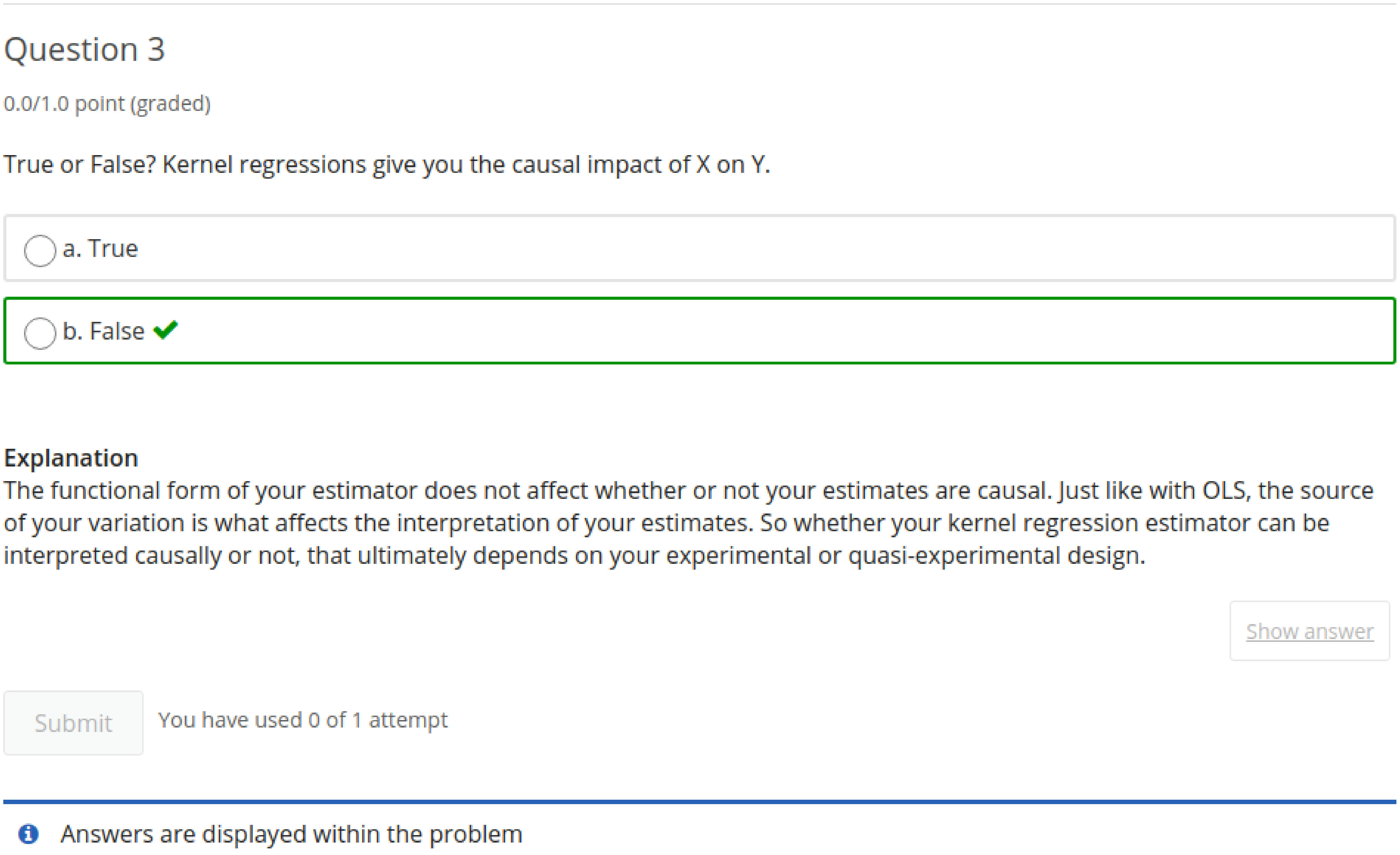
○ c. To minimize mean squared error, which takes into account both bias and variance.

○ d. All of the above. ✓

Explanation

The point of cross validation is to minimize the trade off between bias and variance by minimizing the cross validation criterion (in the same way OLS minimizes Mean Squared Error). The cross validation criterion is a function of both the bias and the variance. A, B, and C all say the same thing, and therefore are all correct.

You have used 0 of 2 attempts



Question 3

0.0/1.0 point (graded)

True or False? Kernel regressions give you the causal impact of X on Y.

○ a. True

○ b. False ✓

Explanation

The functional form of your estimator does not affect whether or not your estimates are causal. Just like with OLS, the source of your variation is what affects the interpretation of your estimates. So whether your kernel regression estimator can be interpreted causally or not, that ultimately depends on your experimental or quasi-experimental design.

You have used 0 of 1 attempt

Answers are displayed within the problem