CSPB 3022 - Craven - Introduction to Data Science Algorithms

Started on	Friday, 22 March 2024, 6:02 PM
State	Finished
Completed on	Friday, 22 March 2024, 7:28 PM
Time taken	1 hour 25 mins
Marks	94.50/111.00
Grade	**85.14** out of 100.00

Question **1**

Correct

Mark 16.00 out of 16.00

A **discrete random variable** X is a function that maps the elements of the sample space Ω to a **finite number** of values $a_1, a_2, \ldots, a_n$ **or** a ***countably*** infinite number of values.

A **probability distribution** of a **discrete** random variable is a function, f, that maps a random variable's values $a_1, a_2, a_3, \ldots$ to the probabilities of those values:

$$f(a_k) = P(X = a_k) \quad \text{for } k = 1, 2, 3, \ldots$$

A discrete random variable X has a **Bernoulli distribution,** denoted $X \sim Ber(p)$, where $0 \le p \le 1$, if its probability distribution is given by

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

A discrete random variable X has a **Binomial Distribution,** denoted $X \sim Bin(n, p)$, with $n = 1, 2, \ldots$ and $0 \le p \le 1$, if its probability distribution is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } 0 \le k \le n$$

p = probability of success for each of the n trials
k = number of successful trials out of n total

A discrete random variable X has a **Poisson distribution** denoted $X \sim Pois(\mu)$ with parameter $\mu > 0$, if its distribution is given by

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

(i.e. probability of k events per unit of measurement)

μ is the expected number of events per unit of measurement (time period, length, space, volume,etc)

A continuous random variable X has a **normal (or Gaussian) distribution** with mean μ and standard deviation σ if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We say $X \sim N(\mu, \sigma^2)$

Definition: *Exponential Distribution*

A continuous rv X is said to have an *exponential distribution* with rate parameter λ if the pdf of X is:

$$f(x) = \lambda e^{-\lambda x}; \quad x \ge 0$$

NOTATION: We write $X \sim exp(\lambda)$ to indicate that X is an exponential rv with rate λ.

Function	Description
X.pmf(k)	$P(X = k)$
X.cdf(k)	$P(X \le k)$
X.mean()	$E[X]$
X.var()	$\text{Var}(X)$
X.std()	$\text{SD}(X)$

What are these?

a. The sky

b. Some definitions, formulas, and topics we have covered in class. ✔

c. A plant

Your answer is correct.

Question **2**

Correct

Mark 5.00 out of 5.00

The **expectation** of a random variable is a number, not a random variable.

True or False

Select one:

True ✔

False

Question **3**

Incorrect

Mark 0.00 out of 5.00

Consider the following simulation. What distribution does the return value of the function belong to? Note that the np.random.exponential function's input $1/q$ corresponds to an exponentially-distributed random variable with rate $\lambda = q$

```
def simrv(q):
    j = 0
    t = np.random.exponential(1/q)
    while t <= 1:
        t += np.random.exponential(1/q)
        j += 1
    return j
```

- a. Uniform
- b. Binomial
- c. Poisson
- d. Exponential ✖
- e. Bernoulli
- f. Normal

Your answer is incorrect.

Question **4**

Correct

Mark 4.00 out of 4.00

Which of the following are true? Select all that apply.

Select one or more:

- [x] a. Cov(X,Y) = $E[XY] - E[X]E[Y]$ ✔
- [x] b. Cov(X,Y) = $E[(X - E[X])(Y - E[Y])]$ ✔
- [ ] c. Cov(X,Y) = $E[(X - E[X])^2(Y - E[Y])^2]$
- [ ] d. Cov(X,Y) = $E[XY] - (E[X]E[Y])^2$

Your answer is correct.

Question **5**

Correct

Mark 3.00 out of 3.00

For a certain random variable X, it is know that $E[X] = 1, Var[X] = 2$.

What is $E[X^2]$? 3

Your last answer was interpreted as follows: 3

Correct answer, well done.
Since $Var[X] = E[X^2] - (E[X])^2$, then $E[X^2] = Var[X] + (E[X])^2$.

Question **6**

Correct

Mark 3.00 out of 3.00

True or False: If Cov$(X, Y) = 0$, this tells us that the random variables X and Y are independent.

Select one:

- ( ) True
- (•) False ✔

Question **7**

Partially correct

Mark 6.00 out of 10.00

Given the following probability distribution for P(X=a) and P(Y=b):

b \ a	0	1	2	P(Y=b)
-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{4}$
1	0	$\frac{1}{4}$	0	$\frac{1}{4}$
P(X=a)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

(a) What is E[XY]? -1/2

Your last answer was interpreted as follows: $\frac{-1}{2}$

Correct answer, well done.

(a) What is E[X]? -1/2

Your last answer was interpreted as follows: $\frac{-1}{2}$

Incorrect answer.

(a) What is E[Y]? 1

Your last answer was interpreted as follows: 1

Incorrect answer.

(d) True or False, X and Y are independent? False

Your last answer was interpreted as follows: **False**

Correct answer, well done.

(d) True or False: X and Y are uncorrelated: True

Your last answer was interpreted as follows: **True**

Correct answer, well done.

Your answer is partially correct.

$E[XY]$ = (-1 * $\frac{1}{4}$) + (-2 * $\frac{1}{4}$) + $\frac{1}{4}$ = $-\frac{1}{2}$

E[X] = 1*$\frac{1}{2}$ + 2 * $\frac{1}{4}$ = 1

E[Y] = (-1)*$\frac{3}{4}$ + $\frac{1}{4}$ = $-\frac{1}{2}$

X and Y are *independent* if P(X=x, Y=y) = P(X=x)P(Y=y) for each *x,y*.

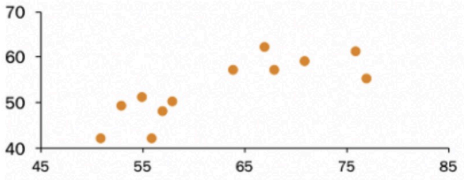
The Cov(X,Y) is $-\frac{1}{2} - 1 * -\frac{1}{2} = 0$. If the Cov(X,Y) is zero, then X and Y are not correlated; otherwise they are correlated.

Question **8**

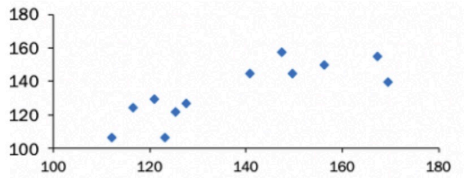
Correct

Mark 5.00 out of 5.00

Below is a scatterplot of random variables X and Y:



Below is a scatterplot of random variables W and Z:



Which of the following statements are true given the scatterplots above? SELECT ALL that apply.

Select one or more:

- [x] a. $Cor(X,Y) = Cor(W,Z)$ ✔
- [x] b. $Cov(X,Y) < Cov(W,Z)$ ✔
- [ ] c. $Cov(X,Y) = Cov(W,Z)$
- [x] d. $Cov(X,Y) > 0$ ✔
- [ ] e. $Cor(X,Y) < Cor(W,Z)$
- [ ] f. $Cor(W,Z) < 0$

Your answer is correct.

Question **9**

Correct

Mark 5.00 out of 5.00

You read a news headline that states that eating chocolate is positively correlated with a lower risk of heart disease.

True or False: From this information you can conclude that eating chocolate causes a lower risk of heart disease.

Select one:

- True
- False ✔

Question **10**

Correct

Mark 5.00 out of 5.00

The city of Boulder wants to hear from its homeowners on issues related to zoning laws. (For the purposes of this question, homeowners are individuals who own their home, instead of leasing or renting from someone else).

The City of Boulder has a list of all the homeowners' email addresses. They take the list of all homeowners' email addresses, shuffle it, and send a survey to every other email address. That is, from the shuffled list, they email the first, third, fifth, seventh, and so on. (You may assume that the shuffling is done uniformly at random, meaning that each email address has the same probability of landing in any particular position. You may also assume that the City of Boulder has the email address for every single homeowner, and that every single homeowner has a unique email address.)

In this sampling technique, the sampling frame is _____ the population of interest.

- a. Equal to ✔
- b. Smaller Than
- c. Greater Than

Your answer is correct.

Question **11**

Correct

Mark 5.00 out of 5.00

Choose any true completion for the following statement:

The law of large numbers states that ....

Select one:

- a. the difference between the estimated mean and true mean vanishes as n goes to infinity. ✔ Correct
- b. the sum of multiple samples will approximate the normal distribution.

Your answer is correct.

The law of large numbers is $\lim_{n\to\infty} P(|\bar{X}_n - \mu|) > \epsilon = 0$, or that the difference between the estimated mean and true mean vanishes as n goes to infinity.

Question **12**

Correct

Mark 3.00 out of 3.00

Select the statement that is a consequence of the Central Limit Theorem.

- a. The distribution of the sample means drawn from an arbitrary probability distribution F, given that the samples are independent and identically distributed (IID) and the sample size is sufficiently large, will approximate a normal distribution. ✔
- b. The histogram of the individual data points drawn from an arbitrary probability distribution F will have a normal distribution.

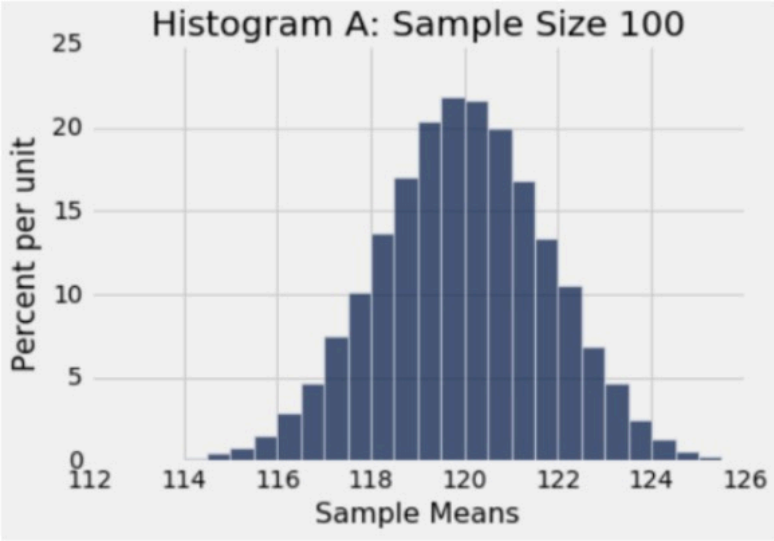
Your answer is correct.

Question **13**

Correct

Mark 5.00 out of 5.00

A population consists of more than half a million people. Histogram A is an empirical histogram of the mean weight (in pounds) of a random sample of 100 people drawn with replacement from the population, based on 25,000 repetitions of the sampling process.



The Standard Deviation of the 25,000 sample means used to construct Histogram A is closest to (pick ONE option):

- a. 1 pound
- b. 10 pounds
- c. 20 pounds
- d. 4 pounds
- e. 2 pounds ✔
- f. 3 pounds

Your answer is correct.

Question **14**

Incorrect

Mark 0.00 out of 5.00

Given an empirical cumulative distribution function (CDF) for a discrete probability distribution below, provide the probability mass function (PMF).

x	CDF	PMF	
1	0.15	0.85	Your last answer was interpreted as follows: 0.85 Incorrect answer.
2	0.27	0.73	Your last answer was interpreted as follows: 0.73 Incorrect answer.
3	0.35	0.65	Your last answer was interpreted as follows: 0.65 Incorrect answer.
4	0.49	0.51	Your last answer was interpreted as follows: 0.51 Incorrect answer.
5	1	0	Your last answer was interpreted as follows: 0 Incorrect answer.

Incorrect answer.

Question **15**

Complete

Mark 5.00 out of 5.00

Consider a population in which a proportion p of individuals are called "successes" (or 1, if you prefer) and the remaining proportion are rudely called "failures" (or 0).

If you draw a sample of size n (where n is some positive integer) at random with replacement from the population, then the number of successes is a random variable that follows the binomial distribution.

Suppose you sample 100 times at random with replacement from a population in which 26% of the individuals are successes. What is the probability that the sample has 20 successes?

Note: You don't need to simplify to a decimal.

If you use permutations or combinations, you can leave notation such as C(10, 6) or P(3, 1) in your answer.

To solve this problem, we can use the binomial theorem, namely:

$$P(X = 20) = \binom{100}{20} 0.26^{20}(1 - 0.26)^{100-20}$$
$$= \binom{100}{20} 0.26^{20}(0.74)^{80}$$

Where in this case n = 100, k = 20, and p = 0.26.

Comment:

Question **16**

Complete

Mark 4.00 out of 5.00

For both parts of this question, **determine if you have enough information to answer. If so, give the answer rounded to the nearest hundredth. Show all steps for your work. If you don't have enough information to answer, write what additional information, you would need to be able to answer.**

From scipy.stats.norm documentation:

scipy.stats.norm

scipy.stats.norm = *<scipy.stats._continuous_distns.norm_gen object>* [source]

A normal continuous random variable.

The location (`loc`) keyword specifies the mean. The scale (`scale`) keyword specifies the standard deviation.

As an instance of the `rv_continuous` class, `norm` object inherits from it a collection of generic methods (see below for the full list), and completes them with details specific for this particular distribution.

pdf(x, loc=0, scale=1)	Probability density function.
cdf(x, loc=0, scale=1)	Cumulative distribution function.

Potentially Useful Output from Python:

```
stats.norm.pdf(116, 114, 12) = 0.03
stats.norm.cdf(116, 114, 12) = 0.57
stats.norm.pdf(116, 114, 2) = 0.12
stats.norm.cdf(116, 114, 2) = 0.84

stats.norm.pdf(111, 114, 12) = 0.03
stats.norm.cdf(111, 114, 12) = 0.40
stats.norm.pdf(111, 114, 2) = 0.06
stats.norm.cdf(111, 114, 2) = 0.07
```

In a particular area of the Rocky Mountains, there are 10,000 trees. The mean of this population of trees is 114 feet with a standard deviation of 12 feet.

a) What is the probability that a randomly sampled tree from this region is taller than 116 feet?

b) A simple random sample of 36 trees from the area is measured. What is the probability that the mean height of this sample will be between 111 and 116 feet?

Part A

We are not told anything about the distribution of this data so we **do not** have enough information to calculate the probability in this case.

Part B

We are told that the population mean in this survey is 114 ft. The standard deviation is 12 ft. To answer this question, we need to calculate the area under the curve between 111 and 116 feet. We cannot calculate this integral analytically so we will use a cumulative density function (the area under the curve up to a specific limit) for each end of our interval, and then find the difference between the largest and smallest end points. In python this would be

stats.norm.cdf(116,114,12) - stats.norm.cdf(111,114,12)

Which would evaluate to a probability of

$$P = 0.57 - 0.40 = 0.17.$$

Giving us a probability of 0.17 for 36 trees to be between 111 and 116 feet.

Comment:

In part b we are discussing the distribution of a statistic (sample mean), which we know by the central limit theorem to be normal distribution.

We can leverage the CLT because:

- n = 36, which is greater than 30, and thus considered large
- n = 36, which is less than 10% of the 10000 population, and thus we can assume the sample is IID

To find the probability of the mean height being between 111 and 116, we must subtract the cdf of the sample mean distribution @ 111 from the cdf of the sample mean distribution at 116, as the cdf measures total area under the curve up to the value indicated. The normal distribution for the sample mean has mean 114 (same as population) and a standard deviation of sigma/sqrt(n) = 12/sqrt(36) = 12/6 = 2

p(sample mean <111) = stats.norm.cdf(111, 114, 2) = 0.07

p(sample mean <116) = stats.norm.cdf(116, 114, 2) = 0.84

p(111 < sample mean < 116) = 0.84 - 0.07 = **0.77**

Question **17**

Complete

Mark 5.00 out of 5.00

If you conduct a **simple random sample** where each case in the population has an equal chance of being included, and there is no implied connection between the cases in the sample, can you still have problems with **bias**?

Explain your answer.

Yes, you can always have problems with bias. This is because no matter the sampling technique, there is always a bias of response bias. Meaning, you can conduct the perfect kind of sampling and someone could just lie to you and this would then in turn cause your sampling to have bias. Participants in the sample could also choose to not respond, so we could also have non response bias as well. No matter your sampling technique, there will always be bias.

Comment:
Good answer!

Question **18**

Complete

Mark 5.00 out of 5.00

Ex). Major earthquakes (magnitude 8.0+) occur once every 500 years in California (according to historical data from USGS, 2015)

What is the probability of zero major earthquakes in California next year?

We know: $500 \ \frac{\text{years}}{\text{earthquake}}$

$0.002 \frac{\text{earthquakes}}{\text{year}}$

$1 \ \frac{\text{earthquakes}}{500 \text{ years}}$

Explain a strategy you can use to solve this problem.

Show/explain all steps of your work including the setup of the formula and substitution of any parameters. You don't need to calculate a decimal value.

The strategy you use may not make use of the following, but it may be helpful to know the CDF of the Exponential Distribution is as follows:

CDF	$1 - e^{-\lambda x}$

Instead of entering symbols, it is alright to type out words like lambda.

This example is a perfect candidate for using a Poisson distribution. In this case, using the Poisson distribution, $\lambda = 0.002$ earthquakes per year and $k = 0$ earthquakes in the following year. Using the Poisson distribution the probability that zero earthquakes occur in the following year is:

$$P(X = 0) = \frac{0.002^0 \cdot e^{-0.002}}{0!} = \frac{1 \cdot e^{-0.002}}{1} = e^{-0.002} = 0.9980019987 \approx 0.9980.$$

So, there is roughly a 99.80% chance that no major earthquakes occur in California in the next year.

Comment:

Information

Use the following for questions 19 - 22.

In the United States, 31% of adults report being online almost constantly. A team of scientists took **a simple random sample of 100 adults who live in Denver and found that 37% reported being online almost constantly.**

One member of the data science team says, "In the total population of Denver adults, the percent who are online almost constantly is more than the national percentage."

Another member of the team says, "No, any difference between the national percentage and the percentage that we observed in our random sample from Denver is just by chance."

In order to decide between these two positions, the data scientists decide to conduct a test of hypotheses at a significance level of 0.05.

Question **19**

Correct

Mark 3.00 out of 3.00

Referring back to the above prompt about the online behavior of adults in the United States and the data science team's findings from the Denver sample.

Which ONE of the following is the most appropriate null hypothesis to test the statements of the data scientists above?

a. The sample of adults from Denver is like draws at random with replacement from individuals of whom **37%** are labeled "online almost constantly."

b. The sample of adults from Denver is like draws at random with replacement from individuals of whom **less than 31%** are labeled "online almost constantly."

c. " The sample of adults from Denver is like draws at random with replacement from individuals of whom **more than 31%** are labeled "online almost constantly."

d. In the sample of 100 Denver adults, the percent who are online almost constantly is 37%.

e. The sample of adults from Denver is like draws at random with replacement from individuals of whom **31%** are labeled "online almost constantly. ✔

Your answer is correct.

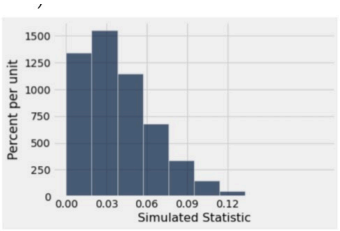
Question **20**

Correct

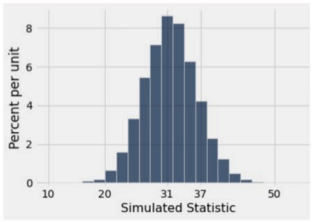
Mark 3.00 out of 3.00

Referring back to the above prompt about the online behavior of adults in the United States and the data science team's findings from the Denver sample.

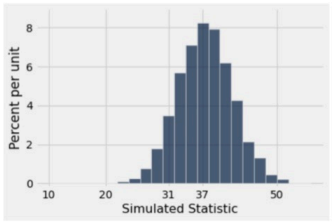
In order to decide between their two hypotheses, the data scientists have picked an appropriate test statistic and simulated it 10,000 times under appropriate conditions. One of the graphs below is the histogram of their simulated values. Which one is it?



Testing Option A



Testing Option B



Testing Option C

(Note that in each graph, some relevant values are labeled on the horizontal axis).

- a. A
- b. B ✔
- c. C

Your answer is correct.

Question **21**

Complete

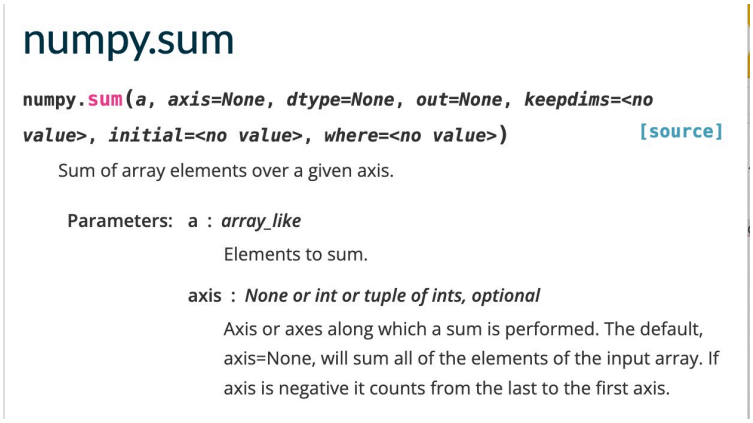
Mark 1.50 out of 3.00

Referring back to the above prompt about the online behavior of adults in the United States and the data science team's findings from the Denver sample.

The 10,000 simulated values of the data scientists' test statistic are stored in a numpy array called sim_stat. Write a Python expression that evaluates to the empirical p-value of the test.

```
import numpy as np
empirical_p = ____________________________________________________________________
```

The following NumPy documentation is provided to assist you in answering the question. Please note that you are not required to use both images; you may refer to either one or both as you find necessary to support your answers.



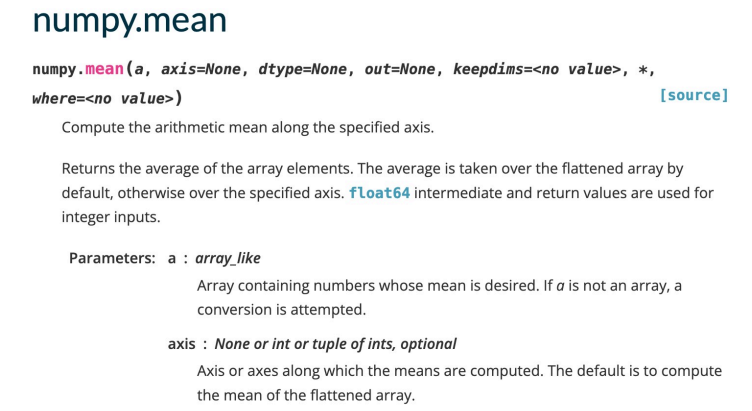
numpy.sum

numpy.sum(*a, axis=None, dtype=None, out=None, keepdims=<no value>, initial=<no value>, where=<no value>*) [source]

Sum of array elements over a given axis.

Parameters: **a** : *array_like*
Elements to sum.

axis : *None or int or tuple of ints, optional*
Axis or axes along which a sum is performed. The default, axis=None, will sum all of the elements of the input array. If axis is negative it counts from the last to the first axis.



numpy.mean

numpy.mean(*a, axis=None, dtype=None, out=None, keepdims=<no value>, *, where=<no value>*) [source]

Compute the arithmetic mean along the specified axis.

Returns the average of the array elements. The average is taken over the flattened array by default, otherwise over the specified axis. `float64` intermediate and return values are used for integer inputs.

Parameters: **a** : *array_like*
Array containing numbers whose mean is desired. If *a* is not an array, a conversion is attempted.

axis : *None or int or tuple of ints, optional*
Axis or axes along which the means are computed. The default is to compute the mean of the flattened array.

We need to sum the values against the observed results under the assumption that our null hypothesis is correct. So, essentially we need to sum all values that are less than or equal to the observed result:

empirical_p = numpy.sum(sim_stat <= 37)

Here, we are assuming that the percentages are out of 100 and not out of 1. If the simulated statistic were out of 1 we would use

empirical_p = numpy.sum(sim_stat <= 0.37)

Comment:

Could use np.sum(sim_stat >= 37) / 10000 OR (sim_stat >= 37).mean()

Should be >=

Didn't divide by 10000 (or len(sim_start)).

Question **22**

Correct

Mark 3.00 out of 3.00

Referring back to the above prompt about the online behavior of adults in the United States and the data science team's findings from the Denver sample.

Suppose you calculate an empirical p-value of 0.118. Based on the significance level chosen by the data scientists, which of the following statements is a valid conclusion of this hypothesis test?

- a. We don't have enough evidence to reject the claim that any difference (between the national percentage of adults who are online almost constantly and the percentage that we observed in our random sample from Denver) is just by chance. ✔
- b. This test proves that in the population of all Denver adults, the percent who are online almost constantly is different than the national percent.
- c. This test proves that in the population of all Denver adults, the percent who are online almost constantly is the same as the national percent.
- d. This test proves that in the population of all Denver adults, the percent who are online almost constantly is greater than the national percentage.
- e. Being a Denver resident causes adults to be online more than the national average.

Your answer is correct.