CSPB 3022 - Kim - Introduction to Data Science Algorithms

Question **1**

Not complete

Marked out of 5.00

Flag question

Data were collected concerning 72 randomly selected advertisements in the The Guardian (April 6, 1992). The data are salaries corresponding to two kinds of occupations (n = m = 72): (1) creative, media, and marketing and (2) education. The sample mean and sample variance of the two datasets are, respectively:

(1) $\bar{x}_{72} = 17410$ and $s_x^2 = 41258741$

(2) $\bar{y}_{72} = 19818$ and $s_y^2 = 50744521$

Assume that both datasets are normal distributions with expectation μ_x and μ_y.

Test $H_0 : \mu_x = \mu_y$ at the level $\alpha = 0.05$ under the assumption of *equal variances*. The right critical value $T_{n+m=2} = T_{142,1-0.05/2}$ is 1.97681035456.

Compute S_p^2, the pooled variance:

Compute the test statistic:

Should we reject H0 in favor of H1 ?

(No answer given)

Check

Question **2**

Not complete

Marked out of 3.00

Of the voters in Florida, a proportion *p* will vote for candidate *G*, and a proportion *1 − p* will vote for candidate *B*. In an election poll a number of v... are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for *G*." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables $X_1, X_2, \dots$ are independent and have a Bernoulli

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Question 2
Not complete
Marked out of 3.00
Flag question

Of the voters in Florida, a proportion *p* will vote for candidate *G*, and a proportion *1 − p* will vote for candidate *B*. In an election poll a number of voters are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for *G*." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables $X_1, X_2, \dots$ are independent and have a Bernoulli distribution with parameter *p*.

Suppose we use $\hat{X}_n$ to predict *p*. According to Chebyshev's inequality, how large should *n* be (how many people should be interviewed) such that the probability that $\hat{X}_n$ is within 0.05 of the "true" *p* is at least 0.85?

In other words $P(|\hat{X}_n - p| < 0.05) \geq 0.85$, or, equivilently $P(|\hat{X}_n - p| \geq 0.05) \leq 0.15$.

Hint: solve this first for p = 1/2, and use the fact that the variance for the bernoulli trial p(1 − p) ≤ 1/4 for all 0 ≤p ≤1.

What is the smallest integer *n* given the variance assumption above:

Check

Question 3
Not complete
Marked out of 5.00
Flag question

Let $X_1, X_2, \dots, X_{121}$ be independent identically distributed random variables, each with expected value $\mu = E[X_i] = 3$, and variance $\sigma^2 = Var(X_i) = 4$.

Approximate $P(X_1 + X_2 + \dots + X_{121} > 302.5)$, using the central limit theorem.

Check

Question 4
Not complete
Marked out of 1.00
Flag question

If the average IQ is 100, Markov's inequality shows that no more than 1/2 of the population can have an IQ more than 200.

Select one:
- True
- False

Check

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Question **5**

Not complete

Marked out of 5.00

Flag question

Of the voters in Florida, a proportion p will vote for candidate G, and a proportion $1 - p$ will vote for candidate B. In an election poll a number of voters are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for G." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables $X_1, X_2, \dots$ are independent and have a Bernoulli distribution with parameter p.

If p > 1/2 candidate G wins; if $X_n > 1/2$ you predict that G will win. Find an n (as small as you can) such that the probability that you predict correctly is at least 0.9, if in fact p = 0.6.

In other words $P(\bar{X}_n \leq 0.5) \leq P(|\bar{X}_n - 0.6| \leq 0.2) \geq 0.9$, or, equivalently $P(|\bar{X}_n - 0.6| \geq 0.2) \leq 0.1$.

Hint: Use Chebyshev's inequality with an estimate that $p = 0.6$.

What is the smallest integer n given the variance assumption above:

Check

Question **6**

Not complete

Marked out of 1.00

Flag question

Assume you've been given a numpy vector **x, y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]
...	...	..

Compute the $E[XY]$ for this numerical distribution using Python.

For example:

Test	Result

`# x = np.array([0, 0, 1, 1])`
`# y = np.array([0, -1, 0, -1])`
`# p = np.array([1/6, 3/6, 1/6, 1/6])`
`if abs(exy - correct_exy) < 0.001: print('Correct')`
	`Correct`

Answer: (penalty regime: 10, 20, 30, ... %)

Reset answer

Ace editor not ready. Perhaps reload page?

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question 7

Not complete

Marked out of 5.00

Flag question

Choose any true completion for the following statement:

The law of large numbers states that ....

Select one:

a. the sum of multiple samples will approximate the normal distribution.

b. the difference between the estimated mean and true mean vanishes as n goes to infinity.

Check

Question 8

Not yet answered

Marked out of 3.00

Flag question

At a London underground station, the number of women was counted in each of 100 queues of length 10. In this way a dataset $x_1, x_2, \ldots, x_{100}$ was obtained, where x_i denotes the observed number of women in the i^{th} queue. The dataset is summarized in the following table and lists the number of queues with 0 women, 1 woman, 2 women, etc.

Count	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	3	4	23	25	19	18	5	1	1	0

In the statistical model for this dataset, we assume that the observed counts are a realization of a random sample $X_1, X_2, \ldots, X_{100}$.

1. Assume that people line up in such a way that a man or woman in a certain position is independent of the other positions, and that in each position one has a woman with equal probability. What is an appropriate choice for the model distribution?
2. Use the table to find an estimate for the parameter(s) of the model distribution chosen in part 1.

Quiz navigation

Finish attempt ...

1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Question **9**

Not complete

Marked out of 5.00

Flag question

Suppose that we are comparing two samples of size M and N from distributions with the *same* (but unknown) variance $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Which of the following are true? Wrong answers count against you, so don't guess.

Select one or more:

- a. We can use the student T-distribution if the samples are from normal distributions.
- b. We can always use the student T-distribution to determine the critical values for for the comparison of the test statistic.
- c. We should use the non-pooled variance $S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$
- d. We should pool the sample variance using the pooled variance $S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)$
- e. We should use the pooled studentized mean $\frac{(\bar{X}_n-\bar{Y}_m)-(\mu_1-\mu_2)}{S_p}$ as the test statistic in the hypothesis test.

Check

Question **10**

Not complete

Marked out of 4.00

Flag question

When you are computing the confidence interval for [ ] samples from [ ] populations, you should use the student T-distribution to compute the interval.

When you are computing the confidence interval for [ ] samples from [ ] , you should use the bootstrap method to determine the interval.

It is acceptable to substitute the [ ] distribution for large n when computing the confidence interval.

| many | student-t | normal | non-normal | a few |

Check

Question **11**

Assuming no income is negative, Markov's inequality shows that no more than 1/5 of the population can have less than 1/5 the average income.

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question 11
Not complete
Marked out of 1.00
Flag question

Assuming no income is negative, Markov's inequality shows that no more than 1/5 of the population can have less than 1/5 the average income.

Select one:
- True
- False

Check

Question 12
Not complete
Marked out of 5.00
Flag question

A bottling machine is known to fill wine bottles with amounts that follow an $N(\mu, \sigma^2)$ distribution with $\sigma = 4$ (ml). In a sample of 64 bottles, $\bar{x} = 740$ (ml) was found. Construct a 95% confidence interval for the unknown expectation μ.

You may find it useful to know the following values for the inverse of the standard normal, ϕ^{-1}:

- $\phi^{-1}(0.925) = 1.44$
- $\phi^{-1}(0.95) = 1.64$
- $\phi^{-1}(0.975) = 1.95$

The 95% confidence interval is ______ ± ______

Check

Question 13
Not yet answered
Marked out of 5.00
Flag question

During the Second World War, London was hit by numerous flying bombs. The following data are from an area in South London of 36 square kilometers. The area was divided into 576 squares with sides of length 1/4 kilometer. For each of the 576 squares the number of hits was recorded. In this way we obtain a dataset $x_1, x_2, \ldots, x_{576}$, where xi denotes the number of hits in the ith square. The data are summarized in the following table which lists the number of squares with no hits, 1 hit, 2 hits, etc.

Number of hits	0	1	2	3	4	5	6	7
Number of Squares	229	211	93	35	7	0	0	1

An interesting question is whether London was hit in a completely random manner. In that case a Poisson distribution should fit the data.

1. If we model the dataset as the realization of a random sample from a Poisson distribution with parameter μ, then what would you choose as an estima ?
2. Check the fit with a Poisson distribution by comparing some of the ob- served relative frequencies of 0's, 1's, 2's, etc., with the corresponding probabilities the Poisson distribution with μ estimated as in part 1.

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

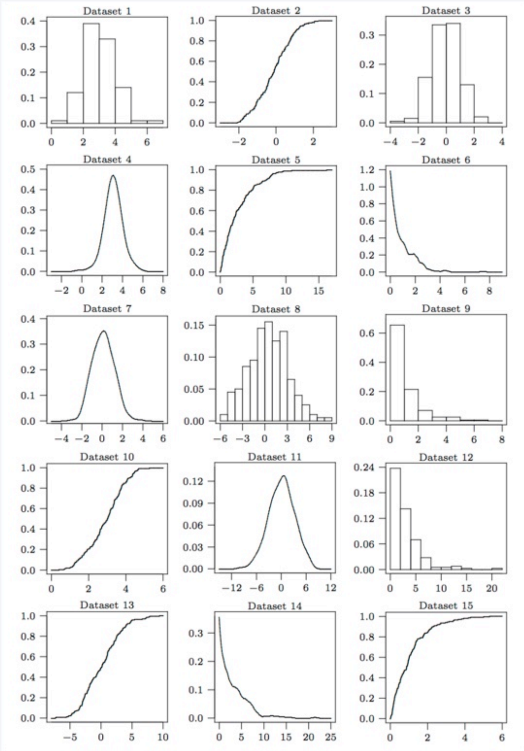
Question **14**

Not complete

Marked out of 5.00

Flag question

The figure displays several histograms, kernel density estimates, and empirical distribution functions. It is known that all figures correspond to datasets of size 200 that are generated from normal distributions N(0,1), N (0, 9), and N (3, 1), and from exponential distributions Exp (1) and Exp (1/3). Report for each figure from which distribution the dataset has been generated.



Drag the appropriate distributions unto the diagram below.

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Not complete

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Question 15
Not complete
Marked out of 5.00
Flag question

You are trying to determine the melting point of a new material, of which you have a large number of samples. For each sample that you measure you find a value close to the actual melting point c but corrupted with a measurement error. We model this with random variables $M_i = c + U_i$ where M_i is the measured value in degree Kelvin, and U_i is the occurring

random error. It is known that $E[U_i] = 0$ and $Var(U_i) = 1$, for each i, and that we may consider the random variables $M_1, M_2, \ldots$ independent. According to Chebyshev's inequality, how many samples do you need to measure to be 85.0% sure that the average of the measurements is within 0.4 of c?

Check

Question 16
Not complete
Marked out of 5.00
Flag question

Select all statements that are consequences of the central limit theorem

Select one or more:

- a. The histogram of samples drawn from an arbitrary probability distribution F will have a normal distribution.
- b. The means of samples drawn from an arbitrary probability distribution F will have a normal distribution.

Check

Question 17
Not complete
Marked out of 5.00
Flag question

In a hypothesis test such as the 2-sample t-test, a ______ t-value leads to a ______ p-value.

larger | smaller

Check

Question 18
Not complete
Marked out of 3.00
Flag question

Assuming no income is negative, Markov's inequality shows that no more than 1/5 of the population can have more than 5 times the average income.

Select one:
- True
- False

Check

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question 19

Not yet answered

Marked out of 1.00

Flag question

You are trying to determine the melting point of a new material, of which you have a large number of samples.

For each sample that you measure you find a value close to the actual melting point c but corrupted with a measurement error. We model this with random variables $M_i = c + U_i$ where M_i is the measured value in degree Kelvin, and U_i is the occurring random error.

It is known that $E[U_i] = 0$ and $Var(U_i) = 3$, for each i, and that we may consider the random variables $M_1, M_2, \ldots$ independent.

According to Chebyshev's inequality, how many samples do you need to measure to be 95% sure that the average of the measurements is within 0.5 of c?

Show your work as you solve the problem. You should first pose the question in an appropriate form and then solve the problem. You may use the equation editor (the little calculator symbol in the menu) to enter equations or write close approximations using text. If your solution involves a complicated fraction, you do not need to simplify it.

Quiz navigation

Finish attempt ...

1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question 20

Not yet answered

Marked out of 3.00

Flag question

The chest circumference of 5732 Scottish soldiers are shown in the histogram below, which suggests modeling the data as the realization of a random sample from a normal distribution.

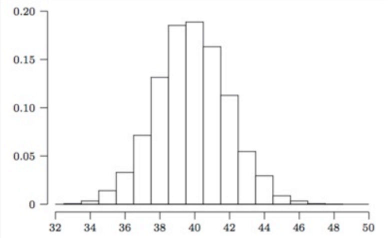
Question 20

Not yet answered

Marked out of 3.00

Flag question

The chest circumference of 5732 Scottish soldiers are shown in the histogram below, which suggests modeling the data as the realization of a random sample from a normal distribution.



1. Suppose that for the dataset $\sum x_i = 228377.2$ and $\sum x_i^2 = 9124064$. What would you choose as estimates for the parameters μ and σ of the N(μ,σ^2) distribution?
2. Give an estimate for the probability that a Scottish soldier has a chest circumference between 38.5 and 42.5 inches.

Question **21**

Not complete

Marked out of 1.00

Flag question

Assume you've been given a numpy vector **x**, **y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]
...	...	..

Compute the $E[XY]$ for this numerical distribution using Python.

For example:

Test	Result

`# x = np.array([0, 0, 1, 1])`
`# y = np.array([0, -1, 0, -1])`
`# p = np.array([1/6, 3/6, 1/6, 1/6])`
`if abs(exy - correct_exy) < 0.001: print('Correct')`
	`Correct`

Answer: (penalty regime: 10, 20, 30, ... %)

Reset answer

Ace editor not ready. Perhaps reload page?
Falling back to raw text area.

```
import numpy as np
#
# Enter your solution based on 'x', 'y' and 'p' here.
# You can use more variables but you must assign your
# solution to 'exy'
#
exy = ...
```

Precheck Check

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30
Not complete

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Question 22
Not complete
Marked out of 3.00
Flag question

Both Chebyshev's inequality and the normal approximation enabled by the Central Limit Theorem (CLT) can be used to estimate the probability of an event occuring. This problem concerns the CLT.

Of the voters in Florida, a proportion p will vote for candidate G, and a proportion $1 - p$ will vote for candidate B. In an election poll a number of voters are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for G." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables $X_1, X_2, \ldots$ are independent and have a Bernoulli distribution with parameter p.

Let $S_n = X_1 + X_2 + \ldots + X_n$. Suppose we use $E[S_n/n] = \bar{X}_n$ to predict p.

Because S_n is the sum of a number of random variables, the central limit theorem indicates that the resulting sum will, in the limit, approach the standard normal distribution $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ or if all the X_i are identical, $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$.

How large should n be (how many people should be interviewed) such that the probability that $\bar{X}_n$ is within 0.05 of the "true" p is at least 0.1?

In other words $P(|\bar{X}_n - p| < 0.05) \geq 0.9$, or, equivilently $P(|\bar{X}_n - p| \geq 0.05) \leq 0.1$.

Hint:

1. Because you are trying to estimate p and you don't know $\mu = p$ and also don't know σ, the best we can do is obtain an upper bound by assuming p = 1/2 to set an upper bound on the variance for the bernoulli trial p(1 - p) ≤ 1/4 for all 0 ≤p ≤1.
2. You can use **quantile_normal(p, 0, 1)** to compute ϕ^{-1} or $Z_n^{-1}(p)$, the inverse CDF of the standard normal (or look it up in a standard normal table or use Python's scipy.stats.norm.ppf function).

What is the smallest integer n given the variance assumption above:

Check

Question 23
Not complete
Marked out of 5.00
Flag question

The Cleveland Casting Plant is a large highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. One process variable of interest to Cleveland Casting is the pouring temperature of molten iron. The pouring temperatures (in degrees Fahrenheit) of ten crankshafts are given below. The mean of that data is 2558.7 and the variance is 517.34. The target setting for the pouring temperature is set at 2550 degrees. One wants to conduct a test at level α = 0.01 to determine whether the pouring temperature differs from the target setting.

2543, 2541, 2544, 2620, 2560

2559, 2562, 2553, 2552, 2553

You formulate $H_0 : \mu = 2550$ and alternate $H_1 : \mu \neq 2550$. Compute the t-statistic and then determine if you accept or reject your null hypothesis. The right critical value $T_{9,1-0.01/2}$ is 3.24983553783.

What is the value of the t-test statistic? Enter your answer as a numeric value.

Should we reject H_0 in favor of $H_1 : \mu \neq 2550$? (No answer given)

Check

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question **24**

Not complete

Marked out of 1.00

Flag question

Assume you've been given a numpy vector **x, y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]
...	...	..

Compute the $E[X + Y]$ for this numerical distribution using Python.

For example:

Test	Result

`# x = np.array([0, 0, 1, 1])`
`# y = np.array([0, -1, 0, -1])`
`# p = np.array([1/6, 3/6, 1/6, 1/6])`
`if abs(explusy - correct_explusy) < 0.001: print('Correct')`
	Correct

Answer: (penalty regime: 10, 20, 30, ... %)

Reset answer

Ace editor not ready. Perhaps reload page?
Falling back to raw text area.

```
import numpy as np
#
# Enter your solution based on 'x', 'y' and 'p' here.
# You can use more variables but you must assign your
# solution to 'explusy'
#
explusy = ...
```

Precheck Check

Question **25**

Not complete

Marked out of 1.00

Flag question

Assume you've been given a numpy vector **x, y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]

Question **25**

Not complete

Marked out of 1.00

Flag question

Assume you've been given a numpy vector **x, y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]
...	...	..

Compute the **correlation coefficient** ρ for this distribution. You can use numpy functions like **np.sum** and **np.sqrt**.

For example:

Test	Result

`# x = np.array([0, 0, 1, 1])`
`# y = np.array([0, -1, 0, -1])`
`# p = np.array([1/6, 3/6, 1/6, 1/6])`
`if abs(correlation - correct_correlation) < 0.001: print('Correct')`
	`Correct`

Answer: (penalty regime: 10, 20, 30, ... %)

Reset answer

Ace editor not ready. Perhaps reload page?
Falling back to raw text area.

```
import numpy as np
#
# Enter your solution based on 'x', 'y' and 'p' here.
# You can use more variables but you must assign your
# solution to 'correlation'
#
correlation = ...
```

Precheck Check

Question **26**

Not complete

Marked out of 2.00

Flag question

A confidence interval for $\bar{X}_n$ will always contain the mean μ.

Select one:

True

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Question 26
Not complete
Marked out of 2.00
Flag question

A confidence interval for $\bar{X}_n$ will always contain the mean μ.

Select one:
- True
- False

Check

Question 27
Not complete
Marked out of 5.00
Flag question

You are given a dataset that may be considered a realization of a normal random sample.

The size of the dataset is $16.$, the average is 35.4 and the sample standard deviation is 3.34.

Construct a 95% confidence interval for the unknown expectation μ. The value of $\phi^{-1}(0.975) = 1.96$ and $t_{0.975,16-1} = 2.14$.

Complete the statement:

The 95% confidence interval is ____ ± ____

Check

Question 28
Not complete
Marked out of 1.00
Flag question

Assume you've been given a numpy vector **x, y** and **p**, to represent a probability distribution in a column form as discussed in lecture. E.g.

x	y	p
x[0]	y[0]	p[0]
x[1]	y[1]	p[1]
...	...	..

Compute the **Cov(X,Y)** for this distribution.

For example:

Test	Result

`# x = np.array([0, 0, 1, 1])`
`# y = np.array([0, -1, 0, -1])`
`# p = np.array([1/6, 3/6, 1/6, 1/6])`
`if abs(cov - correct_cov) < 0.001: print('Correct')`
	`Correct`

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

SECTIONS
1
2
3
4
5
6
7
8
9
10
11
12

Precheck Check

Question 29
Not complete
Marked out of 5.00
Flag question

We perform a t-test for the null hypothesis $H_0 : \mu = 11$ by means of a dataset consisting of n = 16 elements with sample mean 12 and sample variance 1. We use significance level 0.05.

The right critical value $T_{15,1-0.05/2}$ is 2.13144938267. The critical value $T_{15,1-0.05}$ is 1.75305022179.

What is the value of the t-test statistic? Enter your answer as a numeric value.

Should we reject H_0 in favor of $H_1 : \mu \neq 11$? (No answer given)

Should we reject H_0 in favor of $H_1 : \mu > 11$? (No answer given)

Check

Question 30
Not complete
Marked out of 1.00
Flag question

Assuming no income is negative, Markov's inequality shows that no more than 1/5 of the population can have more than 5 times the average income.

Select one:
- True
- False

Check

Quiz navigation
Finish attempt ...
1 2 3 4 5 6
7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 23 24
25 26 27 28 29 30

Finish attempt ...