Name: ________________________

Score: ______ / ______

Mid-term exam

Part 1: True-or-False

1

Suppose we want to investigate if there is a difference in the gender proportions between Math majors and Sociology majors at Duke. It is reasonable to set up our null and alternative hypotheses in the following manner:

\(H_0\): the proportion of female students in the Math department is equal to that in the Sociology department.

\(H_A\): the proportion of female students in the Math department is NOT equal to that in the Sociology department.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: True

2

Box plots can help us identify extreme values and potential outliers, but histograms can not.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: False

3

The zip code of Duke campus is 27708. A "zip code" can be considered as a numerical variable since it consists of digits only.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: False

4

Suppose we want to test a hypothesis using a dataset. Prior to analyzing the data, we set the significance level at \(\alpha = 0.05\), and after analyzing the data we get a p-value of 0.04. Then the null hypothesis must be false.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: False

5

If a given value (for example, the actual population mean of household income) is within a 90% confidence interval, it will definitely also be within a 95% confidence interval calculated based on the same sample.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: True

6

A scientist wanted to learn the effect of a new drug. She randomly selected 50 patients from a hospital where this new drug is tried out, and also randomly selected another 50 patients from a different hospital where a traditional treatment was adopted. She followed up with all patients after 6 months to observe their recovery, and compared the outcomes between patients in the two hospitals. Her analysis WILL reveal a causal relationship between the new drug and patient recovery.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: False

7

For data with a right-skewed distribution, the sample mean is definitely larger than the median.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: True

8

The average percentage of Stephen Curry making a 3-point shot during the 20-21 season was 0.429. (That is, on average, 42.9% of his 3-point shot attempts were successful.) Thus, the probability of him making three 3-point shots in a row should be $0.429 ^ 3 \approx 0.079$.

○ True

○ False

Answer Point Value: 4.0 points
Answer Key: False

Part 2: Fill in the blank

9

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

A political scientist randomly surveyed 100 Durham voters on whether or not they voted for Joe Biden in the 2020 presidential election. The results showed that 80 of them did. She used her survey outcomes to construct a 95% confidence interval for the proportion among all Durham voters who voted for Biden. The 95% confidence interval would be (____,____).

(Note: round to 2 digits after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 0.70|0.74, 0.85|0.89

10

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

To estimate the proportion of Duke students satisfied with their living conditions, Alice took a random sample of 100 students. Upon calculation, she found out that the standard error of her estimate is 0.1. To reduce the standard error down to 0.05, she has to take a larger sample, with sample size increased to ____ at least.

Answer Point Value: 4.0 points
Answer Key: 400

11

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

Jack is a knife throwing performer and he can successfully hit the target 80% of the time. Assume that each of his knife-throw attempts has the same success rate (80%) and all his knife-throws are independent. The probability that 3 out of 5 knives that he throws hit the target is ____.

(Note: round to 3 digits after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 0.180|0.220

12

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

Load the Boston housing dataset included in the "MASS" package in R using the following two commands: (you can do it in your RStudio container or your local R environment)
library(MASS)
data("Boston")
Check out the dataset and fill in the blanks: there are ____ observations and ____ variables in total.

Answer Point Value: 4.0 points
Answer Key: 506, 14

13

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

The height of US women aged 20 or above has a mean of 63.7 inches, with a standard deviation of 2.5 inches. Assume that the height of a US woman (aged 20 or above) follows a normal distribution. Then the probability that a woman is taller than 68.7 inches is ____.
(Note: round to 3 digits after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 0.021|0.024

14

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation).
NOTE: For scientific notation, a period MUST be used as the decimal point marker.

Tom buys lunch from a sandwich shop every day. He typically has a sandwich and a side. The price of sandwiches has a mean of $5 and a standard deviation of $1.6, while the price of sides has a mean of $3 and a standard deviation of $1.2. The prices of a sandwich and a side are independent. Then Tom's average cost of a lunch is $____ with a standard deviation of $____.
(NOTE: round to 1 digit after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 8.0, 1.8|2.2

15

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

On average, 3 major earthquakes (i.e., earthquakes with a magnitude of 5 or higher) occur in Japan during a week. Let's assume that the number of earthquakes in Japan follow a Poisson distribution, then the probability of 4 major earthquakes taking place across Japan during a week is ____.

(Note: round to 3 digits after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 0.155|0.180

16

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

A rapid test for HIV has a 99% accuracy on HIV-positive cases and a 95% accuracy on HIV-negative cases. (That is, it returns a positive result 99% of the time for someone with HIV, and a negative result 95% the time for someone without it.)

The HIV prevalence of a certain country is 6%. A randomly selected person from this country gets this rapid test and receives a positive result. The probability that he actually has HIV is ____%.

(Note: round to 1 digit after the decimal point.)

Answer Point Value: 4.0 points
Answer Key: 54.5|57.0

17

Accepted characters: numbers, decimal point markers, sign indicators (-), spaces (e.g., as thousands separator, 5 000), "E" or "e" (used in scientific notation). **NOTE:** For scientific notation, a period MUST be used as the decimal point marker.

Among all customers in a bar, 80% would order alcoholic drinks, 69% would order snacks, and 55% would order both drinks and snacks. Then ____% of all customers in this bar would order either drinks or snacks.

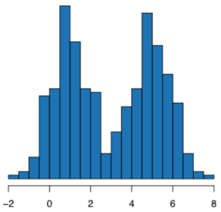
Answer Point Value: 4.0 points
Answer Key: 94

Part 3: Multiple choices

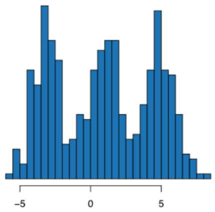
18

You are given 4 datasets with their distributions visualized in the 4 histograms (a), (b), (d), and (d) (see the attached image). Please match them with the correct descriptions of these data distributions

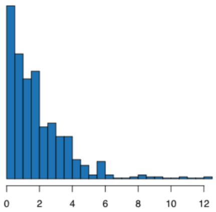
Attachments



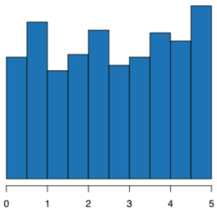
(a)



(b)



(c)



(d)

1. Bimodal
2. Multi-modal
3. Right-skewed
4. Uniform

A. Plot (b)

B. Plot (d)

C. Plot (c)

D. Plot (a)

Answer Point Value: 4.0 points

Answer Key: 1:D, 2:A, 3:C, 4:B

19

The "iris" data contain 50 observations of 3 species of iris flowers, including the sepal length and width and petal length and width for each flower.
You may use the following command to load the data in R:
data("iris")
Check out the iris dataset. Which type of plot would be the most useful in visualizing the relationship between the species and petal length of these flowers?

○ A.
side-by-side box plot

○ B.
side-by-side bar plot

○ C.
histogram

○ D.
dot plot

Answer Point Value: 4.0 points
Answer Key: A

20

About conditions for applying the Central Limit Theorem when estimating single proportions, which of the following statements is true?

○ A. The size sample is considered as sufficiently large if one of np and n(1-p) is larger than 10.

○ B. We can use the sample proportion as an approximation of the true proportion to check the success-failure condition.

○ C. If sample size n is larger than 1000, then we can definitely assume that the sample proportion approximately follows a normal distribution.

○ D. The observations in the data can be dependent.

Answer Point Value: 4.0 points
Answer Key: B

21

Kim wants to test if a coin is fair. She conducts hypothesis testing where the null hypothesis (\(H_0\)) is "the coin is fair". After tossing the coin many times, she decides not to reject the hypothesis that the coin is indeed fair. What decision error could she have made?

○ A. Neither type 1 nor type 2 errors.

○ B. Both type 1 and type 2 errors.

○ C. Type 1 error.

○ D. Type 2 error.

Answer Point Value: 4.0 points
Answer Key: D

22

Which of the following statements about z-scores is/are true?

□ A. Larger z-scores are always better.

□ B. The z-score for an observation that is equal to the mean is 0.

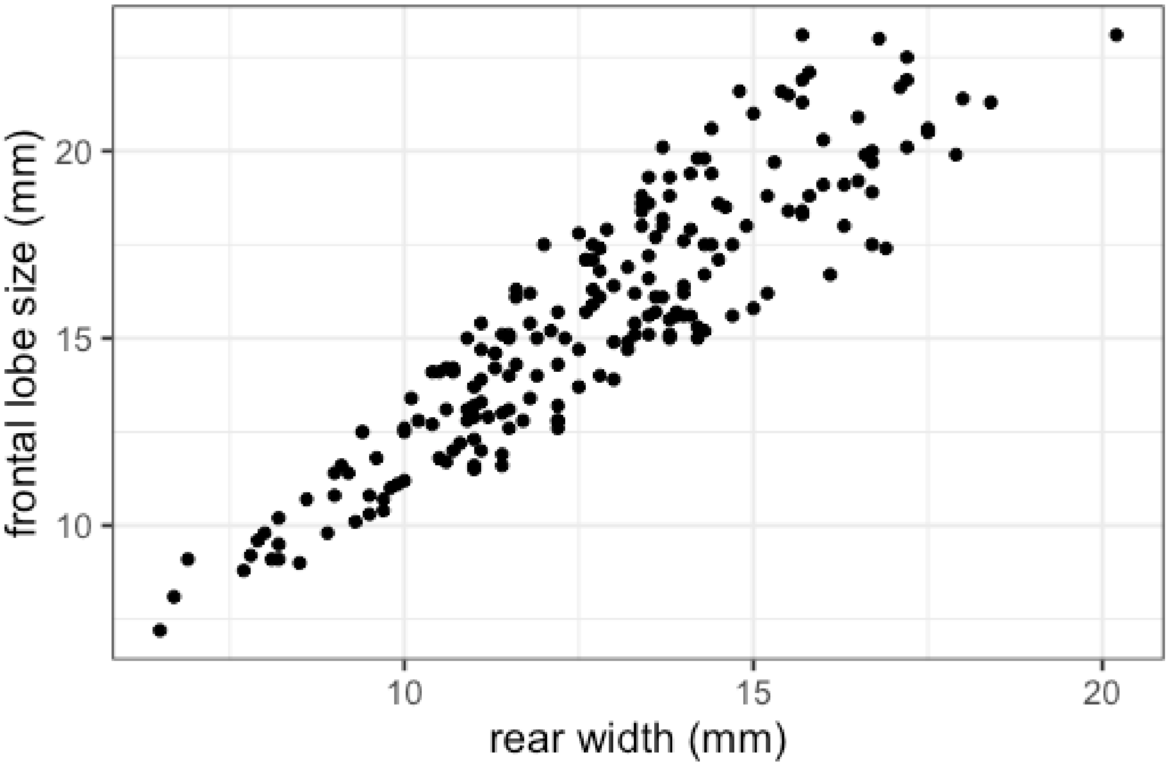
□ C. If a z-score is 2 that means that the observation is two times the value of the mean.

□ D. If a z-score is negative that means that the observation is less than mean.

Answer Point Value: 4.0 points
Answer Key: B,D

The attached plot (see attachment) is a scatterplot that visualizes the relationship between the frontal lobe size and rear width of 200 Leptograpsus crabs in west Australia. Which of the following statements are correct? (Select ALL that are correct.)

Attachments



□ A. The relationship between frontal lobe sizes and rear widths appears linear.

□ B. Frontal lobe sizes and rear widths are independent variables.

□ C. The relationship between frontal lobe sizes and rear widths is nonlinear.

□ D. Frontal lobe sizes and rear widths are strongly associated.

Answer Point Value: 4.0 points
Answer Key: A,D

24

A political scientist is interested in the effect of economic development on social equality. She wants to use a sample of 50 countries evenly represented among the Americas, Europe, Asia, and Africa to conduct her analysis. What type of study and strategy should she use to ensure that countries are selected from each region of the world?

○ A. Observational study, with cluster sampling.

○ B. Experiment, with random assignment.

○ C. Observational study, with stratified sampling.

○ D. Experiment, with blocking.

○ E. Observational study, with simple random sampling.

Answer Point Value: 4.0 points
Answer Key: C

25

A comprehensive survey conducted on Duke students show that the true proportion of all Duke students who have taken at least one Statistics course is 0.4. You survey 60 students in your dorm and record that the proportion of students who have taken Statistics courses is 0.25. The proportion of all students at this college who have taken Statistics courses is a ________ and the proportion of students who have taken Statistics courses in your dorm is a ________.

○ A. population; sample.

○ B. parameter; statistic.

○ C. None of the other options is correct.

○ D. measure of central tendency; measure of variability.

○ E. statistic; parameter.

Answer Point Value: 4.0 points
Answer Key: B