-Directions-

This exam includes 25 multiple-choice questions and three open-response questions that might be used as tie breakers. For questions 1 through 25 (the multiple-choice questions), mark your answer choice in the appropriate location on the sheet provided. After completing questions 1 through 25, answer each tie breaker question in sequential order (i.e., complete Question #1 first, then Question #2, and then Question #3 last). Be sure that your name is printed on each of the tie breaker questions. When time is called, you will be asked to turn in your multiple-choice question answer sheet and your written responses to the tie breaker questions.

1. The average amount of time individuals spent watching a video before stopping the video or navigating away from the video was recorded in seconds for n videos shared on social media. The number of times each video had been shared was also recorded. This data was used to construct a simple linear regression equation $\hat{y} = b_0 + b_1 x$, where x represents the number of times each video is shared and y represents the average amount of time individuals spend watching these videos (seconds). Summary statistics are provided in the table below.

	Number of Shares	Watch Time (seconds)
Mean	678	15
Standard Deviation	180	23

Given the summary statistics in the table, which of the following values could be the slope of the least squares regression line?

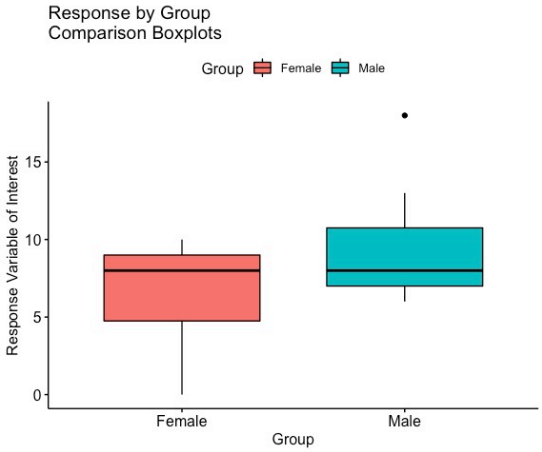
a. - 0.830
b. - 0.113
c. - 1.004
d. - 1.280

2. Consider a set of quantitative data with zero variance. Given this information, which of the following are true statements about the sample statistics for the dataset?

I. Mean = Mode
II. Minimum = Maximum = 0
III. Standard Deviation = 0
IV. Range = 0

a. All of the Above
b. III and IV only
c. II, III, and IV
d. I, III, and IV

3. The plot to the right provides a visual description of an observed variable of interest categorized by a grouping variable of sex. Determine the relationship between the mean and the median of the response for ***males***.



   a. Mean = Median
   b. Mean ≈ Median
   c. Mean < Median
   d. Mean > Median

4. Refer to the plot above. A researcher is interested in determining if the response variable of interest differs by sex. Given the researcher's goal, if a parametric test (e.g., a two-sample independent t-test) were conducted in this setting, which type of error (Type I or Type II) are we likely to be making? Please select the best answer of those provided below.

   a. A Type I Error is Likely
   b. A Type II Error is Likely
   c. Neither Error is Likely
   d. Cannot Be Determined

5. For a particular hypothesis test, the significance level is 0.01 and the statistical power is 0.80. What is the probability of making a Type II Error in this setting? Round to 3 decimal places.

   a. 0.010
   b. 0.800
   c. 0.200
   d. Cannot Be Determined

6. Twister is a game of physical skill that has been inducted into the American National Toy Hall of Fame. A spinner is used to determine where and how a player may move. The spinner consists of four different colors, which are each equally likely to come up in a single spin. Consider spins independent. When spinning 3 times in the game of Twister, what is the probability that any one of the four colors comes up all 3 times? Round to 3 decimal places.

   a. 0.016
   b. 0.063
   c. 0.250
   d. 0.422

For Questions 7-10, refer to the table, which summarizes conditional probabilities for the average customer Socioeconomic Status (SES) ***given*** primary store type of US stores. Establishments that accept only credit or debit cards are known as 'cashless'. Proponents of cashless stores believe they prevent theft and speed up service. However, these cashless stores unfairly disadvantage those individuals that do not have access to credit or debit cards.

Note that 9% of all US stores are cashless while the other 91% of US stores are traditional.

	Socioeconomic Status		
Store Type	Low	Medium	High
Cashless Store	FIND ME	0.256	0.693
Traditional Store	0.405	0.291	0.304

7. What is the probability of one randomly selected US store having an average customer SES of 'Low' given the store is 'Cashless' (i.e., report the value for 'FIND ME')? Round to 3 decimal places.

    a. 0.595
    b. 0.051
    c. 0.567
    d. 0.090

8. What is the probability of one randomly selected US store having an average customer SES of 'Medium'? Round to 3 decimal places.

    a. 0.045
    b. 0.074
    c. 0.547
    d. 0.288

9. What is the probability of one randomly selected US store being 'Cashless' given they have an average customer SES of 'Medium'? Round to 3 decimal places.

    a. 0.256
    b. 0.080
    c. 0.023
    d. 0.042

10. Which of the following statistical procedures would be most appropriate to determine if customer SES (Low, Medium, High) is associated with store type? Please select the best answer of those provided below, assuming that any necessary statistical requirements hold.

    a. One-Way Analysis of Variance (ANOVA)

    b. Chi-Square test of independence

    c. Linear Correlation test

    d. Two-sample independent t-tests

11. Assume that 30 large simple random samples of the same size were taken from the same population and corresponding 90% confidence intervals for the population mean were constructed for each. How many of the 30 constructed confidence intervals would be expected to ***not*** contain the true value of the population mean?
    Please select the best answer of those provided below.

    a. 3

    b. 30

    c. 0

    d. None of the above

12. Using a representative sample of high school students, a 95% confidence interval (CI) was constructed for the population proportion of high school students that report feeling overwhelmed by their cell phone notifications during the school day. Given that the constructed 95% CI was (.673, .892), which of the following is ***false***?

    a. The null hypothesis would be rejected at the 5% significance level when testing the following: $H_0: p = .65$ vs. $H_1: p \neq .65$

    b. Approximately 78.3% of the high school students in the sample reported feeling overwhelmed by their cell phone notifications during the school day.

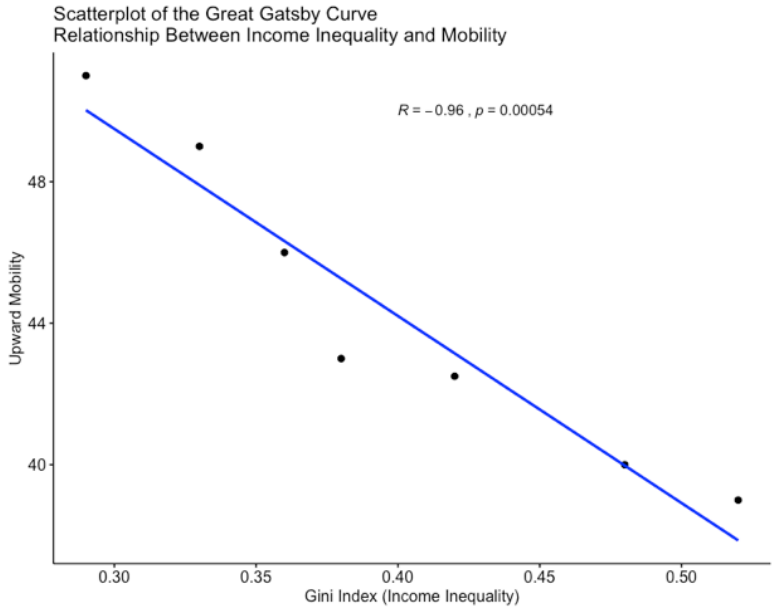
    c. The standard error of the sample proportion (i.e., the standard deviation of the sample proportion) is approximately .110.

    d. Using the same sample data, a 99% CI for the population proportion of high school students that report feeling overwhelmed by their cell phone notifications during the school day would be wider than the 95% CI.

13. Ideal kite flying wind speeds for delta and diamond shaped kites (which are most common for the non-serious kite flyer) are normally distributed, with a mean of 10.5 miles per hour (mph) and a variance of 5.29. A particular non-serious kite flyer is trying to fly a kite on a day with wind speeds corresponding to a z-score of -1.9. What is the wind speed on this day in terms of mph? Round to 3 decimal places.

    a. 0.449
    b. 14.870
    c. 6.130
    d. 20.551

14. The scatterplot below is a visualization of the 'Great Gatsby Curve' i.e. the relationship between income inequality and upward social mobility. The sample linear correlation coefficient and respective p-value for a test of linear correlation are also shown. What can we conclude from the results? Please select the best answer of those provided below.

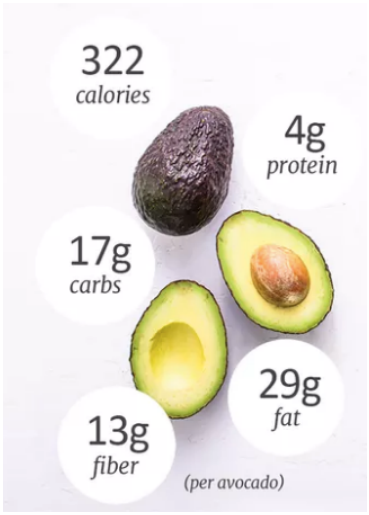


    a. Increases in income inequality cause decreases in upward mobility.
    b. Increases in income inequality cause increases in upward mobility.
    c. Increases in income inequality are associated with decreases in upward mobility.
    d. Increases in income inequality are associated with increases in upward mobility.

15. RateMyProfessors.com is the largest online destination for professor ratings. The website relies on user-generated student comments about university professors in order to create professor ratings. Respondents are not randomly selected but, rather, must self-select and access the internet to participate in the survey. Determine which of the following statements regarding selection bias may result from such sampling. Please select the best answer of those provided below.

    a. Undercoverage is possible in this setting because students without consistent and reliable access to the internet are not adequately represented in the sample.
    b. Nonresponse bias is possible in this setting because students that respond might differ in meaningful ways from nonrespondents.
    c. Voluntary response bias is possible in this setting because respondents are self-selected volunteers and, consequently, individuals with strong opinions are overrepresented.
    d. All of the above are possible examples of selection bias given the sampling process utilized by RateMyProfessors.com.

16. Avocados are a super food! They are in fact super delicious! Assume that the amount of fiber in all avocados is normally distributed with a mean of 13 grams and a standard deviation of 2.9 grams. Given that you eat one randomly selected avocado each day for ten weeks (i.e., 70 randomly selected avocados), what is the probability that the mean amount of daily fiber you get from avocados during these ten weeks is greater than 14 grams? Round to 3 decimal places.

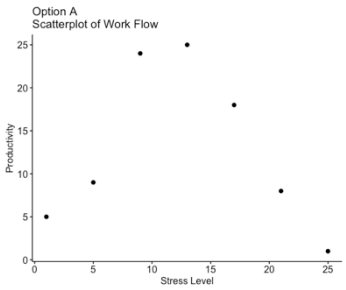


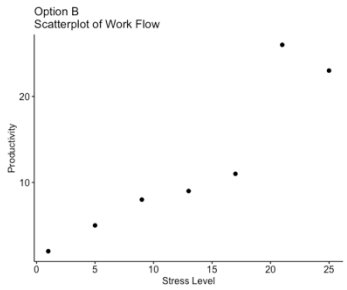
    a. 0.002
    b. 0.365
    c. 0.998
    d. 0.635

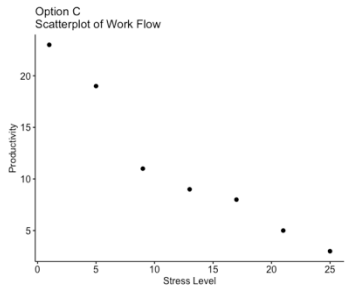
17. Consider ν to be a new sample statistic for measuring skewness defined as $\bar{X}\,/Q_2$ where $\bar{X}$ represents the sample mean and Q_2 represents the second quartile of the sample data. If a sample data distribution contains only positive observations (i.e., all observations are greater than 0), which of the following conditions would indicate a negative skew?

    a. $\nu < 0$
    b. $\nu = 1$
    c. $\nu > 1$
    d. $\nu < 1$

18. Whether at the office or in the classroom, productivity is low when an individual experiences very little stress. Productivity is also low when an individual experiences too much stress. The highest levels of productivity require moderate stress. Based upon the description above, determine which of the following plots most likely represents the relationship between stress and productivity. Please select the best answer of those provided below.







a. Option A Plot
b. Option B Plot
c. Option C Plot
d. None of the above

19. Of 435 total house representatives in the 116th Congress, 102 are female. Three members of the house will be randomly selected to form a relay team for a congressional field day. What is the probability that the relay team will include only ***males***? Round to 3 decimal places.

a. 0.013
b. 0.448
c. 0.234
d. 0.703

20. Refer to the discrete probability distribution provided in the table below.

$X = x$	1	2	3	4	5
$P(X = x)$	0.125	0.154	0.630	?	?

Find the probability that X is greater than or equal to 2. Round to 3 decimal places.

a. 0.091
b. 0.279
c. 0.875
d. Cannot Be Determined

For Questions 21–23, refer to the setting and data provided below.

Researchers have hypothesized that colder temperatures result in infants taking longer to learn to crawl because, dressed in heavier clothing, their movement is more restricted than it would be with warmer temperatures. As this primarily impacts infants born during winter months, infants born during the winter months were randomly selected from Northern states and from Southern states to participate in an additional study investigating the impact location has on crawling time.

This additional study was conducted to determine if the average time to begin crawling (weeks) for infants born in Northern states is ***longer*** than it is for infants born in Southern states.

The crawling times for 5 randomly selected infants in Northern states and 5 randomly selected infants in Southern states were recorded in weeks and provided in the table below. In addition, a variable labeled 'Differences' is provided, which is the crawling time in weeks of the Northern state observation minus the crawling time in weeks of the Southern state observation.

Northern	34	38	34	33	30
Southern	26	35	31	25	29
Differences	8	3	3	8	1

21. Which of the following statistical procedures would be most appropriate to test the claim that average time to begin crawling (weeks) for infants born in Northern states is longer than it is for infants born in Southern states? Assume that any necessary normality requirements hold.

    a. One-tailed two-sample paired/dependent t-test of means
    b. One-tailed two-sample independent t-test of means
    c. Two-tailed two-sample paired/dependent t-test of means
    d. Two-tailed two-sample independent t-test of means

22. Referring to the setting and data provided above, what is the appropriate p-value for testing the claim that average time to begin crawling (weeks) for infants born in Northern states is longer than it is for infants born in Southern states? Round to 3 decimal places.

    a. 0.033
    b. 0.016
    c. 0.075
    d. 0.037

23. Refer to Questions 21-22. Using a 0.10 significance level, which of the following is the most appropriate conclusion for the hypothesis test given the results?

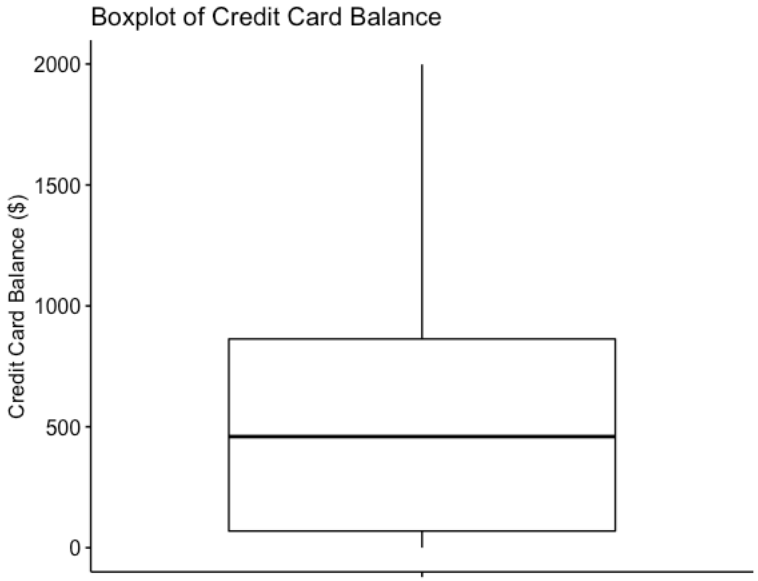
    a. Reject the null hypothesis; there is sufficient evidence to suggest that the average time to begin crawling (weeks) for infants born in Northern states is longer than it is for infants born in Southern states.
    b. Reject the null hypothesis; there is not sufficient evidence to suggest that the average time to begin crawling (weeks) for infants born in Northern states is longer than it is for infants born in Southern states.
    c. Fail to reject the null hypothesis; there is sufficient evidence to suggest that the average time to begin crawling (weeks) for infants born in Northern states is longer than it is for infants born in Southern states.
    d. Accept the null hypothesis; there is sufficient evidence to suggest that the average time to begin crawling (weeks) for infants born in Northern states is the same as it is for infants born in Southern states.

24. Assume for two events, A and B: $P(A) = 0.22$, $P(B) = 0.78$. Assume the events A and B are independent. Find $P(B|A)$. Round to 3 decimal places.

    a. 0.220
    b. 0.780
    c. 0.172
    d. 0.828

25. The graph to the right represents the credit card balance in US dollars ($) for a sample of surveyed adults. How many adults were surveyed?

    a. 2000
    b. 500
    c. 1
    d. Cannot Be Determined



-Tie Breaker Question 1-

Assume a researcher wishes to compare five groups.

If no multiple comparison procedure is applied to each of the resulting pairwise tests, which are each conducted at the $\alpha = 0.05$ significance level, what is the probability of the researcher making at least one Type I Error? Round to 3 decimal places.

You must provide reasoning for your answer.

-Tie Breaker Question 2-

College is expensive! This disproportionately impacts students that do not have the financial resources required for a college education. However, many schools provide gift aid for lower income students. Gift aid, unlike a loan, does not need to be paid back. It is hypothesized that family income (in thousands of US dollars) is related to the amount of gift aid provided to a student (in thousands of US dollars).

A simple random sample of 50 college students was used to construct the simple linear regression equation $\hat{y} = 24.319 - 0.043x$ where x represents the family income (in thousands of US dollars) and y represents the amount of gift aid provided to a student (in thousands of US dollars).

One student has a reported family income of 58.000 (in thousands of US dollars) and a residual of -15.291. What was the ***observed*** amount of gift aid provided to the student (in thousands of US dollars)? Round to 3 decimal places. You must provide reasoning for your answer.

-Tie Breaker Question 3-

Malaria is a severe illness that preventative measures, such as using mosquito nets nightly, can help combat. Refer to the table below, which relates the education level of a mother/guardian in a household (10 years or more vs. Less than 10 years) to the use of a mosquito net during the previous night (Used net vs. Did not use net). The goal of the study was to determine if there was a statistically significant association between education level and mosquito net usage.

- Tie Breaker 3 -	Mosquito Net Usage	
Education Level	Used Net	Did Not Use Net
≥ 10 years	463	309
< 10 years	324	356

Do the data indicate an association between education level and mosquito net usage? Conduct an appropriate hypothesis test to answer this question using a 0.05 significance level. Provide the hypotheses, test statistic(s), p-value(s), and a formal conclusion.

Multiple Choice Key

1. b
2. d
3. d
4. a
5. c
6. b
7. b
8. d
9. b
10. b
11. a
12. c
13. c
14. c
15. d
16. a
17. d
18. a
19. b
20. c
21. b
22. d
23. a
24. b
25. d

-Tie Breaker Question 1-

Assume a researcher wishes to compare five groups.

If no multiple comparison procedure is applied to each of the resulting pairwise tests, which are each conducted at the $\alpha = 0.05$ significance level, what is the probability of the researcher making at least one Type I Error? Round to 3 decimal places.

You must provide reasoning for your answer.

Solution: $\text{P(At Least One Type I Error)} \approx 0.401$

**** Steps***

1. Notice that with five groups there are $\binom{5}{2} = 10$ pairwise comparisons needed

2. The probability of making a Type I Error on one pairwise comparison is: $\alpha = 0.05$

3. So, the probability of NOT making a Type I Error on one pairwise comparison is:

$$1 - \alpha = 1 - 0.05 = 0.95$$

4. The probability of making NO Type I Errors on all ten pairwise comparisons is:

$$(1 - \alpha)^{10} = (1 - 0.05)^{10} = (0.95)^{10} \approx 0.599$$

5. So, the probability of making at least one Type I Error on all ten pairwise comparisons is:

$$\text{P(At Least One Type I Error)} = 1 - \text{P(No Type I Errors)} \approx 1 - 0.599 \approx 0.401$$

Rubric: 0 pts to 5 pts Possible

1 point for each completely correct step and 0 points for each incorrect step on the problem steps labeled #1-5 above.

If the correct final solution is provided but a step (#1-5) is skipped, 5 points should be awarded as long as sufficient work was shown or sufficient reasoning was provided for the answer.

-Tie Breaker Question 2-

ABBREVIATED Question: A simple random sample of 50 college students was used to construct the simple linear regression equation $\hat{y} = 24.319 - 0.043x$ where x represents the family income (in thousands of US dollars) and y represents the amount of gift aid provided to a student (in thousands of US dollars).

One student has a reported family income of 58.000 (in thousands of US dollars) and a residual of -15.291. What was the ***observed*** amount of gift aid provided to the student (in thousands of US dollars)? Round to 3 decimal places. You must provide reasoning for your answer.

Solution: $\text{Gift Aid}_{\text{Observed}} = y \approx 6.534$ or \$6,534

**** Steps***

1. Notice that a residual is equal to the observed value minus the predicted value i.e.

$$\text{Residual} = \text{Gift Aid}_{\text{Observed}} - \text{Gift Aid}_{\text{Predicted}} = y - \hat{y}$$

2. Thus, we know that $-15.291 = \text{Gift Aid}_{\text{Observed}} - \text{Gift Aid}_{\text{Predicted}} = y - \hat{y}$

   We are looking for $\text{Gift Aid}_{\text{Observed}}$

   At first it looks like this would be impossible to find until we realize we can calculate the

   predicted amount of gift aid provided i.e. $\text{Gift Aid}_{\text{Predicted}}$ from the SLR equation as

$$\text{Gift Aid}_{\text{Predicted}} = \hat{y} = 24.319 - 0.043 * 58.00 \approx 21.825$$

3. So, now, we have the following:

$$-15.291 = \text{Gift Aid}_{\text{Observed}} - 21.825$$

   And solving for $\text{Gift Aid}_{\text{Observed}}$, we find that

$$\text{Gift Aid}_{\text{Observed}} = y \approx 6.534 \text{ or } \$6{,}534$$

Rubric: 0 pts to 3 pts Possible

1 point for each completely correct step and 0 points for each incorrect step on the problem steps labeled #1-3 above i.e., one point is possible for each of the following: (1) correct definition of residual, (2) correct calculation of predicted gift aid from SLR equation, (3) correct calculation of observed gift aid

If the correct final solution is provided but a step (#1-3) is skipped, 3 points should be awarded as long as sufficient work was shown or sufficient reasoning was provided for the answer.

-Tie Breaker Question 3-

ABBREVIATION Question:

- Tie Breaker 3 -	Mosquito Net Usage	
Education Level	Used Net	Did Not Use Net
≥ 10 years	463	309
< 10 years	324	356

Do the data indicate an association between the education level and mosquito net usage? Conduct an appropriate hypothesis test to answer this question using a 0.05 significance level. Provide the hypotheses, test statistic(s), p-value(s), and a formal conclusion.

Solution:

- **Hypotheses**
  $\begin{cases} H_0: \text{Education Level and Mosquito Net Usage are Independent (there is no association)} \\ H_1: \text{Education Level and Mosquito Net Usage are Dependent (there is association)} \end{cases}$

- **Test Statistic**
  $\chi^2 = 22.132,\ df = 1$

- **P-Value**
  $p - value < 0.0001$

- **Formal Conclusion at $\alpha = 0.05$ (in terms of H_0)**
  Reject the null hypothesis at the 5% significance level. There is sufficient evidence to support the claim that education level and mosquito net usage are associated/dependent.

Rubric: 0 pts to 4 pts Possible

1 point for each completely correct answer and 0 points for each incorrect answer of the following tie breaker components: (1) hypotheses, (2) test statistic, (3) p-value, and (4) formal conclusion