STAT 13 Spring 2021

Final Exam: Part II

Name: ____________________________ SID:_____________________

Part II Multiple Choice Answer Sheet:

Please submit All your multiple choice answers on ccle week 11 link.
The computer will grade this part automatically.

Q01: _______ Q02: ________ Q03: _______ Q04: _______

Q05: _______ Q06: _______ Q07: _______ Q08: _______

Q09: _______ Q10: _______ Q11: _______ Q12: _______

Q13: _______ Q14: _______ Q15: _______ Q16: _______

Q17: _______ Q18: _______ Q19: _______ Q20: _______

Q21: _______ Q22: _______ Q23: _______ Q24: _______

Q25: _______ Q26: _______ Q27: _______ Q28: _______

Department of Statistics
Final Exam

STAT 13 June 9th, 2021

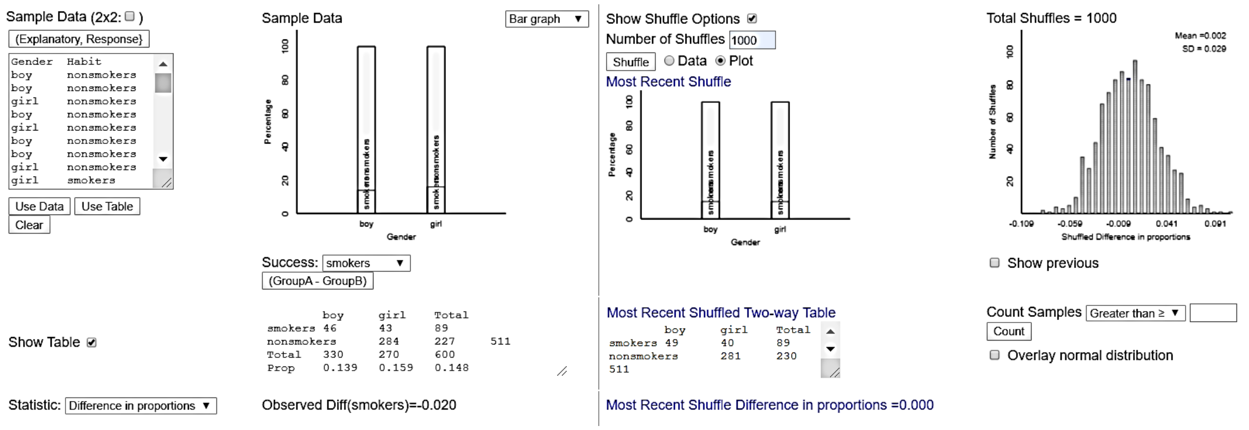
Name: ___________________________ SID:_______________ Section: ____

- If you work after you are told to stop, your exam may be invalidated.
- If you are caught cheating, you will be reported for academic misconduct.
- Make sure you answer the multiple choice part on the first page answer sheet.
- Short answers are best, and budget your time wisely.
- **Rounding to the closest 3 decimal places is acceptable in this exam.**

******* Good Luck *******

Part I: Show your work: 4 questions 14 points each: Total of 56 points.

Q1) Smoking Habit and Gender: A researcher was wonder whether the proportion of smokers among males is statistically different than the proportion of smokers among females. The following snapshot is a summary of the data and a simulated sampling distribution under the Null hypothesis.



Sample Data (2x2: □)

```
Gender  Habit
boy     nonsmokers
boy     nonsmokers
girl    nonsmokers
boy     nonsmokers
girl    nonsmokers
boy     nonsmokers
boy     nonsmokers
girl    nonsmokers
girl    smokers
```

Show Table ☑

Statistic: Difference in proportions ▼

Success: smokers ▼

```
           boy     girl    Total
smokers 46         43      89
nonsmokers         284     227      511
Total      330     270     600
Prop       0.139   0.159   0.148
```

Observed Diff(smokers)=-0.020

Show Shuffle Options ☑

Number of Shuffles 1000

○ Data ● Plot

Most Recent Shuffled Two-way Table

```
          boy     girl    Total
smokers 49        40      89
nonsmokers        281     230
511
```

Most Recent Shuffle Difference in proportions =0.000

Total Shuffles = 1000

Mean =0.002
SD = 0.029

□ Show previous

Count Samples Greater than ≥ ▼

□ Overlay normal distribution

a) Validate the assumptions before using the Normal Distribution as your approximation for the sampling distribution.

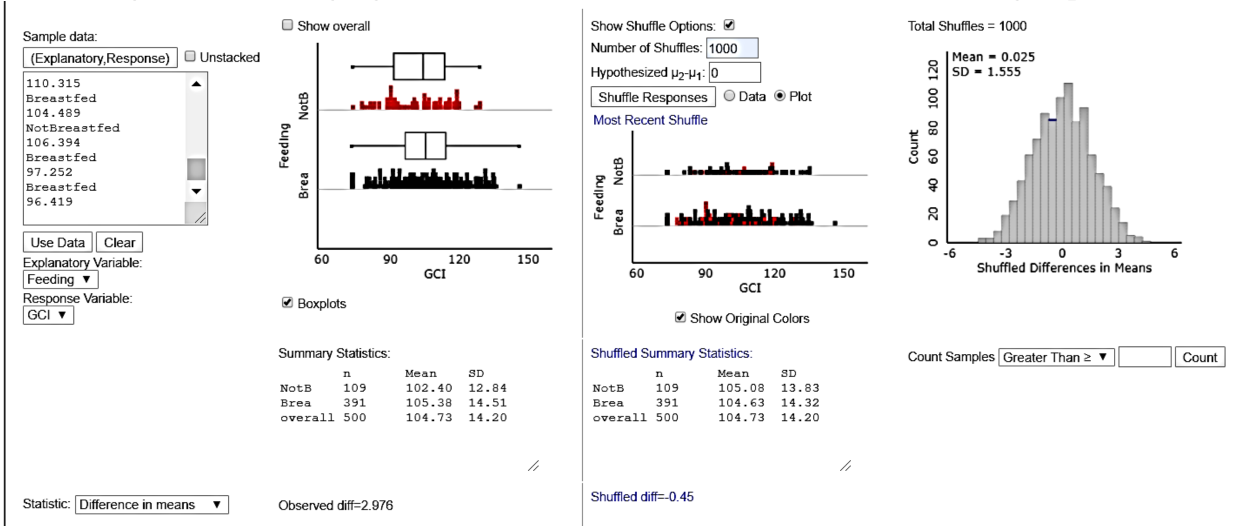
b) Construct a 95% confidence interval for the difference between the two population proportions.

c) State the Null and the alternative hypotheses based on the scenario given in the question, then test the claim and report your p-value. (you need $\hat{p}_{pooled}$ to test the hypothesis)

d) Based on your decision in part (c) what kind of error are you liable to?

e) Roughly speaking, would you get the same conclusion using the simulated sampling distribution? Explain.

Q2) **Breastfed vs. Not Breastfed and GCI:** A researcher studied whether and how children who were breastfed during infancy differed from those who weren't. Followed up at age of 4 and measured The General Cognitive Index (GCI) which is the overall present level of intellectual functioning. The following figure summarizes the differences between the two groups.



Sample data:

(Explanatory,Response) □ Unstacked

```
110.315
Breastfed
104.489
NotBreastfed
106.394
Breastfed
97.252
Breastfed
96.419
```

Use Data | Clear

Explanatory Variable: Feeding ▼

Response Variable: GCI ▼

Statistic: Difference in means ▼

□ Show overall

☑ Boxplots

Summary Statistics:

```
        n     Mean    SD
NotB    109   102.40  12.84
Brea    391   105.38  14.51
overall 500   104.73  14.20
```

Observed diff=2.976

Show Shuffle Options: ☑

Number of Shuffles: 1000

Hypothesized μ_2-μ_1: 0

Shuffle Responses ○ Data ● Plot

Most Recent Shuffle

☑ Show Original Colors

Shuffled Summary Statistics:

```
        n     Mean    SD
NotB    109   105.08  13.83
Brea    391   104.63  14.32
overall 500   104.73  14.20
```

Shuffled diff=-0.45

Total Shuffles = 1000

Mean = 0.025
SD = 1.555

Count Samples Greater Than ≥ ▼ Count

a) State the null and the alternative hypotheses based on this context.

b) Is it safe to assume equal variances here? If so, what is the value if S_{pooled}. Find the theory-based t-score (t observed) based on the given data summary. Is your t-score extreme? Report your P-value.

c) Based on the simulation plot above for the sampling distribution of the difference between the sample means, do you reject or fail to reject the null hypothesis? Why?

d) Construct a 99% confidence interval to estimate the difference between the averages $(since\ n1\ and\ n2\ are\ large\ use\ 2.575\ as\ your\ t_{\frac{\alpha}{2}}\ )$

Q3) Fish and cancer: In an article published in *Lancet* (2001), researchers shared their findings from a study where they followed 6,272 Swedish men for 30 years to see whether there was an association between the amount of fish in the diet and likelihood of prostate cancer. The results are presented in the following two-way table:

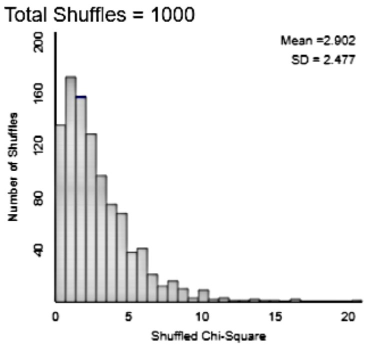
		Large	Moderate	Small	None	Total
Prostate cancer?	Yes	42	209	201	14	
	No	507	2,769	2,420	110	
	Total					

a) Create the expected cell count:

		Large	Moderate	Small	None	Total
Prostate cancer?	Yes					
	No					
	Total					

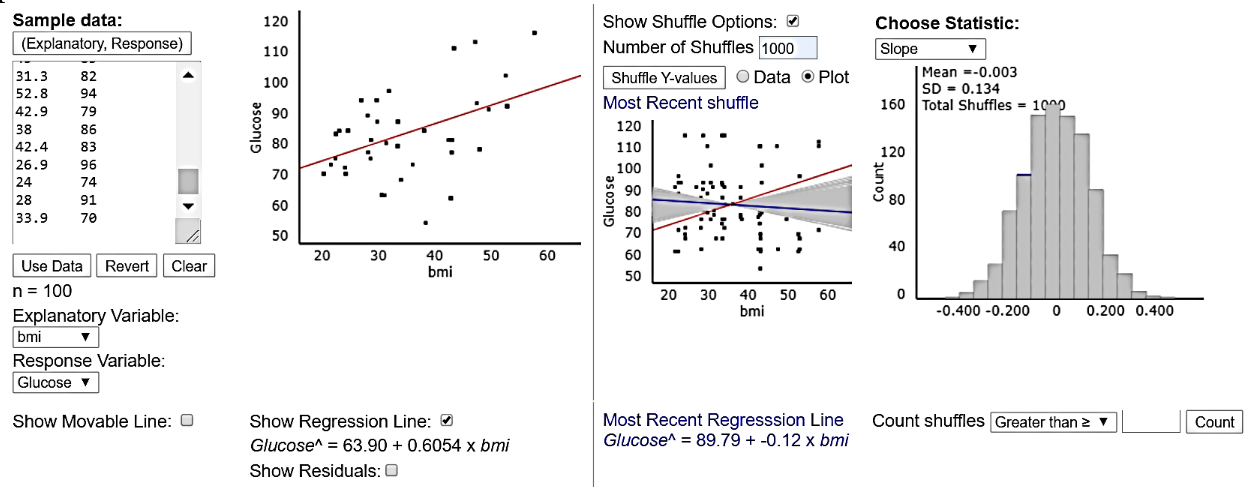
b) Calculate the χ^2 score based on your observed and expected tables.

c) State the null and the alternative hypotheses. What is the degree of freedom for this χ^2test?



d) Perform the test based on your result in part (b) and report your P-value. Do the test again using the sampling distribution given in the simulation figure above.

Q4) **Predicting Glucose Level from BMI Score:** The data plotted here represent the humans' BMI and their correspondent Glucose. Consider the following simulation for a simple linear regression using BMI as your independent variable. The results of the simple regression are provided below.



Sample data:

(Explanatory, Response)

31.3	82
52.8	94
42.9	79
38	86
42.4	83
26.9	96
24	74
28	91
33.9	70

n = 100

Explanatory Variable: bmi ▼

Response Variable: Glucose ▼

Show Movable Line: □

Show Regression Line: ☑

Glucose^ = 63.90 + 0.6054 x *bmi*

Show Residuals: □

Show Shuffle Options: ☑

Number of Shuffles 1000

○ Data ● Plot

Most Recent shuffle

Choose Statistic: Slope ▼

Mean =-0.003

SD = 0.134

Total Shuffles = 1000

Most Recent Regresssion Line

Glucose^ = 89.79 + -0.12 x *bmi*

Count shuffles Greater than ≥ ▼

A. What is the proper interpretation of the relation (the slope) between bmi and Glucose level here?

B. According to the simple linear regression above, if a person has a BMI of 30, what is the predicted Glucose level for that person?

C. A person has the following measures:

bmi	Glucose
43.3	113

Using the regression model, what is the residual (error) for this person's Glucose level?

d. Given the following descriptive statistics for this data:

```
> summary(BMIG.New$bmi)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.20   28.07   33.40   35.71   43.00   57.60
> summary(BMIG.New$Glucose)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 56.00   77.00   84.00   85.52   94.00  118.00
> sd(BMIG.New$bmi)
[1] 10.13809
> sd(BMIG.New$Glucose)
  [1] 13.97218
```

Find the linear correlation coefficient r.

e. Find the Coefficient of determination R^2 and interpret it according to the context.

PART B: Please circle the correct answer in the following 28 multiple choice questions. (2 points each). Total of 56 points.

1. When the p-value is very large, we may conclude that

a) The null hypothesis has been proven to be true
b) There is evidence for the null hypothesis
c) There is evidence against the alternative hypothesis
d) There is little to no evidence for the alternative hypothesis
e) The alternative hypothesis has been proven to be false

2. Researchers surveyed 1,000 randomly (SRS) selected adults in the U.S. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling they typically collect in a week. Please select the best interpretation of this result.

A. We cannot conclude whether earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.
B. This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.
C. This result indicates that earning more money influences people to recycle more than people who earn less money.
D. None of the above.

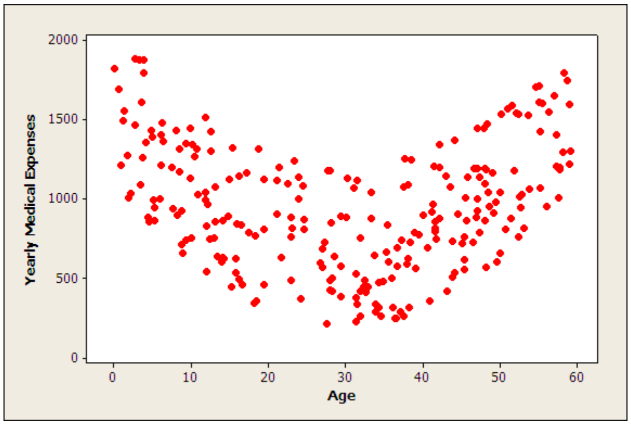
3. According to the US Census Bureau, in 2004 the median household income in the United States was \$43,389 and the mean household income was \$60,528. Based on these two numbers would you say the distribution of household income was:

a. Symmetric.
b. Skewed to the right.
c. Skewed to the left.
d. Bimodal.

4. A student participates in a Coke versus Pepsi taste test. She correctly identifies which soda is which four times out of six tries. She claims that this proves that she can reliably tell the difference between the two soft drinks. You have studied statistics and you want to determine the probability of anyone getting at least four right out of six tries just by chance alone. Which of the following would provide an accurate estimate of that probability?

A. Have the student repeat this experiment many times and calculate the percentage time she correctly distinguishes between the brands.
B. Simulate this on the computer with a 50% chance of guessing the correct soft drink on each try, and calculate the percent of times there are four or more correct guesses out of six trials.
C. Repeat this experiment with a very large sample of people and calculate the percentage of people who make four correct guesses out of six tries.
D. All of the methods listed above would provide an accurate estimate of the probability.

5. The graph below shows a scatter plot of medical expenses in the past year by age for a sample of Americans.



Which *one* of the following is a true statement about the data shown in the graph?

A. The correlation must be close to one because there is a strong relationship between age and medical expenses.

B. Using correlation on the data shown in the graph above is not appropriate because the relationship shown in the graph is not linear.

C. Neither A nor B is a true statement.

6. Which one of the following statements best explains the relationship between residuals and correlation?

a. As the correlation gets closer to 1 the residuals get smaller.

b. As the correlation gets closer to 1 the residuals get larger.

c. As the correlation gets closer to 1, the residuals do not change.

7. A recent research study randomly divided participants into groups who were given different levels of Vitamin E to take daily. One group received only a placebo pill. The research study followed the participants for eight years to see how many developed a particular type of cancer during that time period. Which of the following responses gives the best explanation as to the purpose of randomization in this study?

a. To increase the accuracy of the research results.

b. To ensure that all potential cancer patients had an equal chance of being selected for the study.

c. To reduce the amount of sampling error.

d. To produce treatment groups with similar characteristics.

e. To prevent skewness in the results.

8. Jean lives about 10 miles from the college where she plans to attend a 10-week summer class. There are two main routes she can take to the school, one through the city and one through the countryside. The city route is shorter in miles, but has more stoplights. The country route is longer in miles, but has only a few stop signs and stoplights. Jean sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the following data for 5 days of travel on each route.

Country Route: 17, 15, 17, 16, 18

City Route: 18, 13, 20, 10, 16

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early because that would increase her parking fees. Based on the data gathered, which route would you advise her to choose?

A. The Country Route, because the times are consistently between 15 and 18 minutes.
B. The City Route, because she can get there in 10 minutes on a good day and the average time is less than for the Country Route.
C. Because the times on the two routes have so much overlap, neither route is better than the other. She might as well flip a coin.

9. Suppose a test of significance was correctly conducted and showed no statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?

A. The researcher must not be interpreting the results correctly; there should be a significant difference.
B. The sample size may be too small to detect a statistically significant difference.
C. It must be true that the herbicide does not cause higher levels of the enzyme.

10. Suppose a test of significance was correctly conducted and showed a statistically significant difference in average enzyme level between the fish that were exposed to the herbicide and those that were not. What conclusion can the graduate student draw from these results?

a. There is evidence of association, but no causal effect of herbicide on enzyme levels.
b. The sample size is too small to draw a valid conclusion.
c. He has proven that the herbicide causes higher levels of the enzyme.
d. There is evidence that the herbicide causes higher levels of the enzyme for these fish.

11. Two groups are conducting surveys about habits of Hope College students. Both groups ask if students drink coffee regularly. They both calculate 95% confidence intervals for the proportion of all students at Hope that drink coffee regularly. One group has a sample of 100 students and the other group has a sample of 200 students.

a) We would expect the group with the larger sample to have a narrower interval.
b) We would expect the group with the smaller sample to have a narrower interval.
c) Since they are estimating the same population proportion, we expect both groups to get intervals of the same width.
d) If you pick this answer you will get this question wrong.

Q12-Q15: You want to investigate a claim that the proportions of men and women that are left-handed differ. You take a random sample of men and women (in your community) and find the proportion of each gender that are left-handed. (Questions 12-15 are based on this study.)

12. If the difference in the proportions (of those that are left-handed) between the two samples gives a small p-value (less than 0.05), which one of the following is the best interpretation?

a) It would be very unlikely to obtain the observed sample results if there is really **no** difference between the proportions of men and women in your community that are left-handed.

b) It would be very unlikely to obtain the observed sample results if there is really a difference between the proportions of men and women in your community that are left-handed.
c) It would **not** be very unlikely to obtain the observed sample results if there is really **no** difference between the proportions of men and women in your community that are left-handed.
d) The probability is very small that there is a difference between the proportions of men and women in your community that are left-handed.

13. If the difference in the proportions (of those that are left-handed) between the two samples gives a large p-value (greater than 0.05), which of the following is the best conclusion to draw?

A. You have found strong evidence that there is **no** difference between the proportions of men and women in your community that are left-handed.
B. You have found strong evidence that there is a difference between the proportions of men and women in your community that are left-handed.
C. You have **not** found enough evidence to conclude there is a difference between the proportions of men and women in your community that are left-handed.
D. The same proportion of men and women are left-handed.

14. Suppose that four different studies are conducted on this issue with the following results.

- Study A: 10 of 100 women (10%) are left-handed, compared to 15 of 100 men (15%).
- Study B: 10 of 100 women (10%) are left-handed, compared to 20 of 100 men (20%).
- Study C: 100 of 1000 women (10%) are left-handed, compared to 150 of 1000 men (15%).
- Study D: 100 of 1000 women (10%) are left-handed, compared to 200 of 1000 men (20%).

Which study provides the strongest evidence that there is a difference between men and women on this issue?

A. Study A
B. Study B
C. Study C
D. Study D

15. Suppose a small p-value (less than 0.05) is found when you run your test to determine if there is a difference between men and women on this issue. Which of the following 95% confidence intervals for the difference in proportions would correspond to this small p-value?

A. 0.03 to 0.13
B. -0.03 to 0.10
C. -0.13 to 0.03
D. -1.13 to 1.13

16. Which of the following is true of all randomized experiments?

A. The researchers randomly assign the response variable to the subjects.
B. The researchers randomly assign the explanatory variable to the subjects.
C. The researchers randomly assign both the explanatory and response variable to the subjects.
D. The researchers randomly assign neither the explanatory nor the response variable to the subjects.

17. Which of the following allows us to determine cause and effect?

A. Random assignment to groups in an experiment.
B. Random sampling from a population.
C. Randomization in developing a null distribution.
D. Randomly determining which test of significance should be performed.

18. Which of the following is the primary purpose of randomly assigning subjects to treatments in an experiment?

a) To simulate what would happen in the long run
b) To give each subject a 50-50 chance of obtaining a successful outcome
c) To produce a representative sample so results can be generalized to a larger population
d) To produce similar (experimental) groups so any differences in the response variable can be attributed to the explanatory variable

Questions 19-21: Use the following scenario. Suppose you want to see if there are gender differences in exercising at Hope College. A random sample of 150 Hope College students are asked three questions:

- What is your gender? (Male/Female)
- Do you exercise regularly? (Yes/No)
- If you exercise regularly, how many minutes was your last workout?

19. First you want to find out if the proportion of all Hope College females who exercise regularly is different from the proportion of all Hope College males that exercise regularly. The correct set of hypotheses is

a) H_0: $\pi_{\text{female}} = \pi_{\text{male}}$ H_a: $\pi_{\text{female}} \neq \pi_{\text{male}}$
b) H_0: $\mu_{\text{female}} = \mu_{\text{male}}$ H_a: $\mu_{\text{female}} \neq \mu_{\text{male}}$
c) H_0: $\hat{p}_{\text{female}} = \hat{p}_{\text{male}}$ H_a: $\hat{p}_{\text{female}} \neq \hat{p}_{\text{male}}$
d) H_0: $\bar{x}_{\text{female}} = \bar{x}_{\text{male}}$ H_a: $\bar{x}_{\text{female}} \neq \bar{x}_{\text{male}}$

20. The p-value for the test in the previous question was found to be 0.035. Which **99%** confidence interval for the difference in population proportions makes sense?

A. (-0.026, 0.244) B. (0.026, 0.244)
C. (-0.244, -0.026) D. None of these make sense since none are centered on 0.035.

21. You want to find if females exercise more on average than do males? The correct hypotheses for this test are:

a) H_0: $\pi_{\text{female}} = \pi_{\text{male}}$ H_a: $\pi_{\text{female}} > \pi_{\text{male}}$
b) H_0: $\mu_{\text{female}} = \mu_{\text{male}}$ H_a: $\mu_{\text{female}} > \mu_{\text{male}}$
c) H_0: $\hat{p}_{\text{female}} = \hat{p}_{\text{male}}$ H_a: $\hat{p}_{\text{female}} > \hat{p}_{\text{male}}$
d) H_0: $\bar{x}_{\text{female}} = \bar{x}_{\text{male}}$ H_a: $\bar{x}_{\text{female}} > \bar{x}_{\text{male}}$

Q22 - Q24: CIRCLE THE CORRECT ANSWER {(A) increases or (B) decreases} for each of the following statements about strength of evidence.

22. As the sample size decreases, strength of evidence: **A) increases B) decreases.**
23. As the means get farther apart, strength of evidence: **A) increases B) decreases.**
24. As MAD statistic gets smaller, strength of evidence: **A) increases B) decreases.**

Q25- Q28: Use the following information for questions 25 -28:
The general social survey (GSS) is a yearly survey of adult US citizens on a variety of social issues. Over 5300 different questions have been asked over the nearly 40 years of administering the survey, with many of the questions, the same from year to year. Here are some of the questions asked:

- Gender (Male/Female)
- Age (years)
- Highest Degree Earned (None, High School Diploma, Associate's level, Bachelor's level, Post-bachelor's [MS, PhD, MD, etc.])
- Political Party (Republican, Democrat, Other)
- Household yearly income ($'s)

For each of the following questions, indicate the most appropriate analysis to be used to test the research hypothesis. Choose one of the following strategies: (repeated answers is allowed)
A) Independent Samples T-Test (comparing two group means)
B) Correlation (Linear Regression)
C) Chi-squared Test (comparing two or more proportions)
D) One proportion test.
E) Comparing a Mean Difference (paired data)

25. A relationship between Gender and Highest Degree Earned__________________

26. A relationship between Age and Income________________________

27. A relationship between Age (as Young or old) and Highest Degree Earned____________

28. A relationship between Political Party Affiliation (Republican, Democrat only) and Income___________________