STT 231 – Exam 3 Review

Your third exam will have 9-12 short answer questions covering the entire semester of content, but highly emphasizing lectures 4.2 – 5.3; you'll have 120 minutes to complete Exam 3, although it is designed to take, on average, roughly 70 minutes. The following pages contain sample exercises that are similar in format and scope to those you can expect to see on Exam 3. Many of them are exercises that appeared on Exam 3 in previous semesters. In addition to this review guide, I recommend using past WebWorK assignments, At-Home Reading exercises, and recitation activities as study tools. Moreover, make use of the office hours your instructor and TA offer!

---

1. **Railways & Housing Values ~** In the 1800s, an extensive system of railroads connected towns in New England but as automobile use spread most of the train tracks were disassembled. In recent years, many cities have converted the unused railroad beds into "rail trails" for citizens to use for walking and biking. In one such town, researchers collected information on 104 homes and classified them as either "Closer" or "Farther Away" from the rail trail and then calculated the percentage change in estimated sale price for each home between the years 1998 and 2014.

	Closer	Farther Away
Mean	48.0	38.6
SD	25.4	32.5
Sample size	40	64

$S_p = 29.9845$

a. Conduct a hypothesis test to determine if the mean percent change in estimated sale price is different for the "Closer" and "Farther Away" groups of homes.

H_0: $\mu_1 = \mu_2$ vs H_a: $\mu_1 \neq \mu_2$

Observed test statistic: 1.6455 *p*-value: 0.1036 and est. effect size: 0.3135

b. In order for this test to be valid, certain conditions must be met. Which of the following *is **not*** one of these conditions? **Select *all* the incorrect statements.**

X The observed 104 observations must be normally distributed.
______ The two samples must be drawn separately from two independent, approximately normal populations.
X The standard deviations of the two populations must be approximately equal.
X We must be able to expect at least 10 successes and 10 failures in each sample.

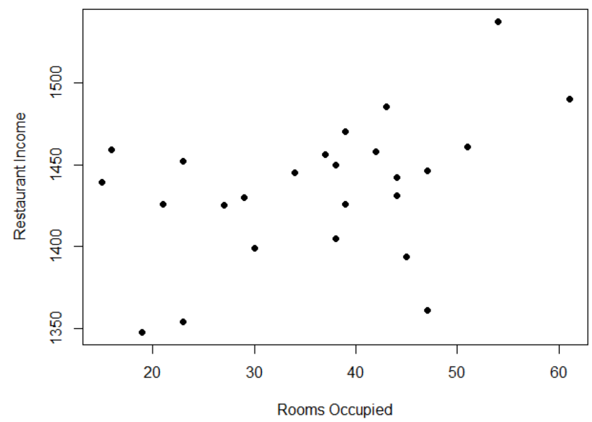
c. Another researcher decides to replicate this study, using data from a different, but similar town. Conducting the same test with a sample size of 50 "Closer" and 60 "Farther Away" homes, the researcher calculates a p-value of 0.053 and an effect size of 0.28. Does the second researcher's study confirm or conflict with the findings of the first researcher.

Circle one: ***Confirm*** ***Conflict***

because:

i. The p-value is smaller in the second study.
ii. The effect sizes are similar in both studies.
iii. The sample sizes are not the same.
iv. The effect size is smaller in the second study.
v. The p-value is larger in the second study.

2. Occupancy vs Revenue ~ A suburban hotel derives its revenue from its hotel and restaurant operations. The owners are interested in the relationship between the number of rooms occupied on a nightly basis and the revenue per day in the restaurant. They collect a sample of 25 days (Monday through Thursday) from last year and record the restaurant income and the number of rooms occupied. Summary results and a scatterplot of the data are shown below.



Rooms Occupied: mean = 36.42, sd = 12.245
Restaurant Income: mean = 1435.56, sd = 42.795
r = .423

a. Using the scatter plot, describe the relationship between rooms occupied and restaurant income.

Relationship appears linear, moderately strong, and positive.

b. Use the summary statistics to calculate the equation of the ordinary least-squares regression line between $x =$ *rooms occupied* and $y =$ *restaurant income*.

Rooms Occupied: mean = 36.42, sd = 12.245
Restaurant Income: mean = 1435.56, sd = 42.795, r = .423

$b_1 = \frac{42.795}{12.245}(0.423) = 1.4783$

$b_0 = 1435.56 - 1.4783(36.42)$

Equation: $\hat{y} = 1381.72 + 1.4783x$

c. Complete the following sentence to interpret the slope of the regression line:

An increase of 1 in Rooms Occupied corresponds to a/an increase of 1.4783 in Restaurant Income.

d. Last Tuesday, 23 rooms were occupied and the restaurant recorded an income of $1,452. Calculate the predicted income for the restaurant along with its residual.

$\hat{y} = 1381.72 + 1.4783(23) = 1415.7209$

$e = y - \hat{y} = 36.2791$

Predicted value: 1415.7209 and residual: 36.2791

3. Living Area and Sale Price ~ A random sample of 149 homes that were recently sold in the Old Town neighborhood of Ames, Iowa was selected from county records. We can run a linear regression command with statistical software using *living_area* as the explanatory variable and *sale_price* as the response variable. The software output is shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32477.993   7466.244    4.35 2.53e-05 ***
living_area    41.278      3.298   12.52  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26690 on 147 degrees of freedom
Multiple R-squared:  0.5159,    Adjusted R-squared:  0.5126
F-statistic: 156.7 on 1 and 147 DF,  p-value: < 2.2e-16
```

$n = 149$ so

$df = 147$

a. Write the equation for the regression line for predicting sale price from living area.

Equation: $\hat{y} = 32477.993 + 41.278x$

b. Which of the following is the correlation between *living_area* and *sale_price*?

Circle one: ***0.5159*** ***-0.5159*** ***0.7183*** ***-0.7183***

c. Do the data provide strong evidence (α = 0.05) that sale price is associated with the living area? Conduct a hypothesis test using the information given in the R output.

H_0: $\beta_1 = 0$ **vs** H_a: $\beta_1 \neq 0$

Give the test statistic value that would be used to test these hypotheses, along with its corresponding p-value:

Test statistic: 12.52 ***p-value:*** $2e^{-16}$ (i.e., basically 0)

d. Calculate a 95% confidence interval for the slope, β_1. Show all work.

$41.278 \pm 1.9762(3.298)$

Confidence Interval: (34.7605 , 47.7955)

e. Complete the following sentence:

51.59 % of the variation in sale price can be explained by its linear relationship with living area.

4. **Air pollution ~** A chemical plant is required to maintain sulfur dioxide levels in the working environment atmosphere at an average level of no more than 0.125 parts per million (ppm). Safety engineers measure the levels at a randomly chosen 10 intervals each week. If the sample mean sulfur dioxide level is more than 0.15 ppm, the safety protocols say that the plant will be evacuated while the air is scrubbed and machines are adjusted. The standard deviation of the measurements is known to be 0.04 ppm, and the testing scenario tests the hypotheses $H_0: \mu = 0.125\ vs.\ H_a: \mu > 0.125$.

a. What are the consequences of making a Type I error in this situation?
   i. The plant will not be evacuated when it needs to be.
   ii. The plant will be evacuated when it needs to be.
   iii. The plant will not be evacuated when it does not need to be.
   iv. The plant will be evacuated when it does not need to be.

b. What are the consequences of making a Type II error in this situation?
   i. The plant will not be evacuated when it needs to be.
   ii. The plant will be evacuated when it needs to be.
   iii. The plant will not be evacuated when it does not need to be.
   iv. The plant will be evacuated when it does not need to be.

c. Using this decision rule, what is the chance the safety engineers will make a Type I error?

normalcdf$(0.15, 10^{\wedge}10, 0.125, 0.04/\sqrt{10})$

Final answer: 0.0241

d. Using the same decision rule, what is the power of this test if the mean pollution level is 0.14 ppm?

normalcdf$(0.15, 10^{\wedge}10, 0.14, 0.04/\sqrt{10})$

Final answer: 0.2160

---

5. **Home shopping ~** A marketing agency wishes to determine the proportion of families in the Chicago metropolitan area who have ever watched a televised home shopping program. A random sample of 425 adults was selected, of whom 220 indicated that they had watched such a program. The agency uses this information to calculate a confidence interval of (0.4778, 0.5575).

a. What level of confidence was used to construct this interval?
   i. 90%
   ii. 95%
   iii. 99%
   iv. Something else

$\hat{p} = 220/425 = 0.5176$

CI: $0.5176 \pm z^* \sqrt{\frac{.5176(.4824)}{425}}$ $\rightsquigarrow$ $z^* = 1.645$

b. The estimate standard error for this sampling situation is 0.0242. What is the correct interpretation of the estimated standard error?

   i. In repeated samples, we would expect our estimate of $\hat{p}$ to be equal to the population parameter value of p approximately 2.42% of the time.
   ii. We have strong evidence that our confidence interval is valid.
   iii. In repeated samples, we would expect our estimate to be, on average, approximately 0.0242 away from the actual population parameter.
   iv. Any estimate of the sample proportion will be wrong 2.42% of the time.

---

6. Traveling distances between cities ~ The U.S. Department of Transportation provides the number of miles that residents of the 75 largest metropolitan areas travel per day in a car. Suppose that for a simple random sample of 50 Buffalo residents the mean is 22.5 miles a day and the standard deviation is 8.4 miles a day, and for an independent simple random sample of 50 Boston residents the mean is 18.6 miles a day and the standard deviation is 7.4 miles a day.

a. A researcher wants to test to see if there is a significant difference in mean travel times between the two cities. The researcher should use…

   i. Two-sample z-test for the difference in population proportions
   ii. Two-sample t-test for the difference in independent means
   iii. χ^2-test of independence

b. Express the research question in terms of two competing hypotheses:

H_0: $\mu_1 = \mu_2$ vs H_a: $\mu_1 \neq \mu_2$

c. Conduct a test of these two hypotheses. Calculate a test statistic and a p-value. Show all work and/or calculator input/output.

$S_p = 7.9158$

Observed test statistic: 2.4634 *p*-value: 0.0155 and est. effect size: 0.4927

d. Given our *p*-value, how much evidence do we have that the null hypothesis model is not a good fit for our observed result?

   i. Little evidence
   ii. Some evidence
   iii. Strong evidence
   iv. Very strong evidence
   v. Extremely strong evidence

e. Construct a 95% confidence interval for the difference between the two population means. Show all work, including any calculator input.

$22.5 - 18.6 \pm 1.9849\sqrt{\frac{8.4^2}{50} + \frac{7.4^2}{50}}$

$(0.7576, 7.0424)$

Confidence interval: (__________ , __________)

7. **Cereal boxes ~** A large box of corn flakes claims to contain 510 grams of cereal. Since cereal boxes must contain at least as much product as their packaging claims, the machine that fills the boxes is set to put 513 grams in each box. The machine has a known standard deviation of 3 grams and the distribution of fills is known to be normal. At random intervals throughout the day, workers sample 5 boxes and weigh the cereal in each box. If the average is less than 511 grams, the machine is shut down and adjusted.

a. How often will the workers make a Type I error with this decision rule and the hypotheses: $H_0: \mu = 513\ vs.\ H_A: \mu < 513$?

normalcdf(-10^10, 511, 513, 3/√5)

Final answer: 0.0680

b. Suppose the workers are dissatisfied by the Type I error rate in (a). A portion of the group argues the decision rule should be changed such that workers will only shut down the factory if the average is less than 509 grams. Another group believes they should keep the same decision rule, but sample 10 boxes for each quality control check, rather than only 5. Which option will result in the 'better' Type I error rate?

Option 1: normalcdf(-10^10, 509, 513, 3/√5) = 0.0014

Option 2: normalcdf(-10^10, 511, 513, 3/√10) = 0.0175

Final answer: Option 1

c. Using the same decision rule, what is the power of the hypothesis test if the machine is actually filling boxes with an average of 510 grams? That is, calculate the chance of observing a sample mean less than 511 if the true mean is 510.

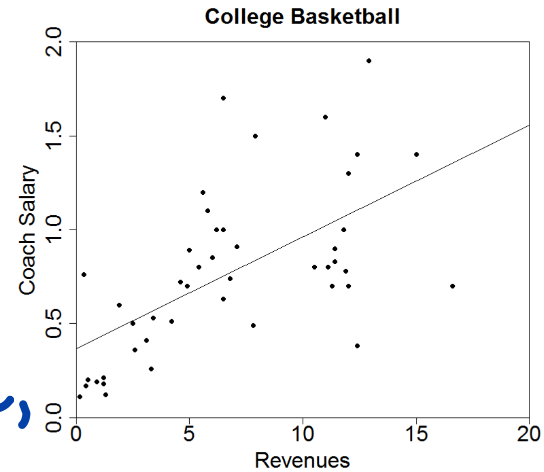
normalcdf(-10^10, 511, 510, 3/√5)

Final answer: 0.7720

d. What is a Type II error in this testing situation?

i. The factory shuts down the machine when it should not.
ii. The factory does not shut down the machine when it should.
iii. The factory shuts down the machine when it should.
iv. The factory does not shut down the machine when it should not.

8. Basketball Salaries vs. Team Revenues ~ College basketball is big business, with coaches' salaries, revenues, and expenses in millions of dollars. A sample of 45 college basketball programs was selected. The data is summarized below. Both revenues and salaries are reported in millions of dollars.



Revenues: $\bar{x}$ = 6.74, s_x = 4.52;
Coach Salary: $\bar{y}$ = 0.77, s_y = 0.44
$r = 0.61$

a. Using the scatter plot, describe the relationship between revenues and coach salary for these 45 colleges.

Plot shows a moderate, positive, linear association b/t revenues & salary.

b. Use the summary statistics above to calculate the regression line $\hat{y} = b_0 + b_1 x$ between Revenues and Coach Salary using Revenues as the independent variable.

$b_1 = 0.0594 \quad b_0 = 0.3696$

c. Interpret the slope in the context of the problem. Be specific.

We predict a typical coach will receive a $59,400 increase in salary for every $1 million increase in revenue.

d. Interpret the intercept in the context of the problem. Be specific. Is the model valid all the way to x=0?

A team that brings in $0 in revenue will pay its coach $369,600, which seems unrealistic. Model is not valid at x=0!

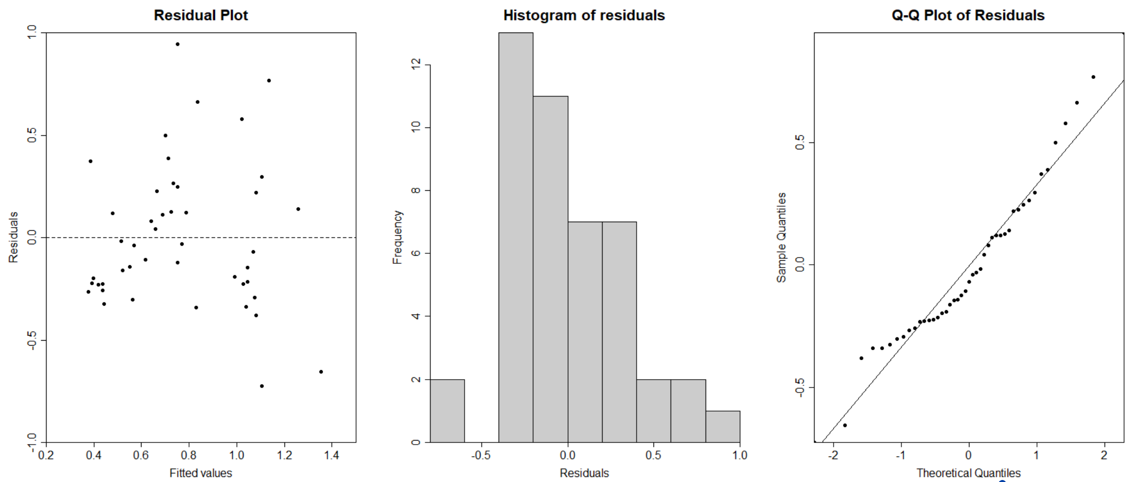
e. Use the regression line to predict the coach salary for a college with revenues of $11 million.

$\hat{y} = 1.023$

f. In the data set, Michigan State has a recorded revenue of $11 million and a coach salary of $1.6 million. Calculate the residual for this college and explain its meaning.

$e = y - \hat{y} = 0.577$

g. Do the data meet the conditions required for fitting a least squares line? Use the following residual plots to explain your answer this question.



The res. vs. fitted plot shows random scatter & QQ plot appears normal. Our data meet OLS conditions!

---

9. Boomburgs ~ Many suburbs in California have grown into large cities, but are often not thought of as cities because they are in the shadow of a major city. Urban experts recognize suburbs having either a large population or having a substantial growth spurt as suburban cities. These suburban cities often have more homes than jobs and are sometimes called *boomburbs*. Fourteen suburban cities were sampled, and their age as a suburban city, the initial population of the suburban city when it was first recognized as a city, and the population size of the city in the year 2000 were recorded. To answer the question, "Can the size of the city when it was first recognized as a city predict the population of the city in 2000?" I ran a linear regression command in R for using initial population as the explanatory variable and population in 2000 as the response variable. Summary output from R is shown below.

```
Call:
lm(formula = X2000Pop ~ InitialPop)

Residuals:
   Min     1Q Median     3Q    Max
-34260 -13378  -6444   8545  61361

Coefficients:
                Estimate  Std. Error t value     Pr(>|t|)
(Intercept) 129954.8611  11481.5113  11.319 0.0000000925 ***
InitialPop       0.2762      0.2456   1.125        0.283
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27620 on 12 degrees of freedom
Multiple R-squared:  0.09538, Adjusted R-squared:  0.01999
F-statistic: 1.265 on 1 and 12 DF,  p-value: 0.2827
```

c. What is the reported standard error for the regression model?

Final answer: 27620

d. Do the data provide evidence that the initial population is associated with population in 2000? Conduct a hypothesis test using the information given in the R output.

H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$

test statistic: 1.125 and p-value: 0.283

e. Based on the results of the hypothesis test, how much evidence do we have that there is a linear relationship between the size of the city when it was first recognized as a city and the population of the city in 2000?

i. Little evidence
ii. Some evidence
iii. Strong evidence
iv. Very strong evidence
v. Extremely strong evidence

f. What percentage of the variation in the population of a city in 2000 can be explained the linear relationship with the size of the city when it was first recognized as a city?

Final answer: 9.538%

g. Calculate a 99% confidence interval for the slope, β_1.

df=12, so

$0.2762 \pm 3.0545(0.2456)$

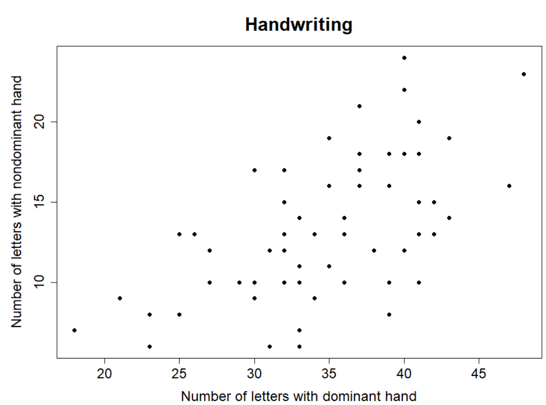
$t^* = \text{invT}(0.005, 12) = 3.0545$

Confidence interval: (-0.4740 **,** 1.0264 **)**

---

10. Handwriting ~ A sample of 63 students wrote as many capital letters of the alphabet in order as they could in 15 seconds using their dominant hand, and then they repeated this task using their nondominant hand. A summary of the data is given by:

Dominant hand: mean = 34.2, standard deviation = 6.25
Nondominant hand: mean = 13.14, standard deviation = 4.26
The correlation is $r = 0.597$



a. The relationship between the number of letters written with the dominant and nondominant hand is

(circle one) **positive** **negative**

(circle one) **linear** **nonlinear**

b. Use the summary statistics to find the slope and y-intercept of the regression line for predicting the number of letters written with the nondominant hand from the number written with the dominant hand. **Show your work.**

$b_1 =$ 0.4069 $\qquad$ b_0: -0.7765

c. Fill in the blanks to **interpret the slope** of the estimated regression line: *(put a number in each blank)*

Each additional 1 letter written with the dominant hand is associated with

(circle one) **an increase** **a decrease** of 0.4069 in the number of letters written with the nondominant hand.

d. What is the R^2 value for this regression model? (No work required.)

Final answer: 0.3564

e. Interpret the R^2 value by completing the sentence below:

35.64 % of the variation in # of letters written by non-dom hand can be explained by the linear relationship with # of letters written by dom. hand.

---

11. Oil pipes ~ Oil field pipes are internally coated in order to prevent corrosion. Researchers investigated the influence that coating may have on the surface roughness of oil field pipes. A random sample of 20 sections of coated interior pipe was scanned with a probe and the standard deviation of the measurements was 0.52 micrometers.

a. If the researchers want to construct a 95% confidence interval with a margin of error of no more than 0.15 micrometer, how large of a sample must they take?

$$n = \left(\frac{st^*}{ME}\right)^2 = \left(\frac{0.52(2.093)}{0.15}\right)^2$$

Final answer: 53

b. If the researchers increase their sample size and calculate the confidence interval to be (1.747, 2.023) micrometers, what was their sample mean?

$$\bar{x} = \frac{1.747 + 2.023}{2}$$

Final answer: 1.885

12. Name that scenario ~ One important aspect in statistics is to understand which statistical methods or procedures are appropriate to use to address the research problem or question of interest. For each description of a research question below, clearly select and write the letter corresponding to the statistical analysis technique most appropriate for addressing that research question.

__H__ 1. A researcher at the National Highway Traffic Safety Administration (NHTS) plans to obtain a random sample of 120 drivers from across the US and will record for each driver whether or not they usually use their seat belt and the region where they live (North, South, East, or West). The researcher wishes to assess if there is an association between seat belt usage and region.

__A__ 2. A new sprinkler system is being installed at a large office complex. The manufacturer of the new system claims the mean activation time is less than 25 seconds. A set of fire alarm/sprinkler system tests were performed and the activation time (in seconds) for reach test were recorded to assess the manufacturer's claim.

__F__ 3. A medical researcher conjectured that smoking might result in wrinkled skin around the eyes. She will randomly sample 150 smokes and 250 nonsmokers and record whether or not they have prominent wrinkles around their eyes to assess if there is a *higher* incidence of wrinkles for the smoking population as compared to the nonsmoking population.

__D__ 4. Plants emit gas that trigger the ripening of fruit, attract pollinators, and cue other physiological responses. Researchers were interested in whether plants with more mass emitted more of these gasses (i.e., that the amount of gas a plant emitted in hundreds of nanograms could be predicted by the plant's weight in grams). They collected data on the weight and volatile compounds emitted for 11 potato plants.

__H__ 5. Is bone marrow density (BMD) related to clinical depression? Researchers conducted an observational study on approximately 2,000 Hong Kong men aged 65 to 92 years old. Each participant in the study was classified as either 'depressed' or 'not depressed' and as either 'osteoporotic,' 'Low BMD' or 'Normal BMD.' The researchers hope to use the collected data to assess whether BMD and depression share an association.

__F__ 6. Approximately 450,000 vasectomies are performed each year in the US. In this surgical procedure for contraception, the tube carrying sperm from the testicles is cut and tied. Several studies have been conducted to analyze the relationship between vasectomies and prostate cancer. The results of one such study found that of 21,300 men who had not had the procedure, 69 were found to have prostate cancer; of 22,000 men who had had a vasectomy, 113 were found to have prostate cancer. The researchers hope to assess if having a vasectomy increases the risk of prostate cancer.

__C__ 7. Previous research has suggested that changes in the activity of dopamine, a neurotransmitter in the brain, may be a causative factor for schizophrenia. To assess this possibility, researchers plan to measure the enzyme dopamine beta-hydroxylase (in nanomoles per milliliter-hour per milligram) among 25 schizophrenic patients and compare the data between those who had been classified as psychotic by hospital staff to those who have not.

quantitative

A. 1-sample t-test for a population mean

C. 2-sample t-test for the comparison of two population means

D. Linear regression model

F. 2-sample z-test for the comparison of two population proportions

G. 1-sample z-test for a population proportion

H. χ^2 test of independence

J. χ^2 test of goodness of fit

A 8. In 2010, the Federal Highway Administration reported that the average car was driven 12,400 miles. Researchers believe that changes in housing markets has increased the commute for many Americans, and that this average annual mileage may have changed in the past nine years. They plan to randomly sample 500 cars and record the number of miles they drive between Jan. 1, 2020 and Dec. 31, 2020, to help assess whether their suspicion is well-founded.	

13. Neurosurgery operative times ~ Several neurosurgeons wanted to determine whether a dynamic system (Z-plate) reduced the operative time relative to a static system (ALPS plate). Two professors from Arizona State University collected the data below, which gives the operative times, in minutes, for the two systems. Summary data is offered below:

$S_p = 74.7938$

Method	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs	$\bar{x}$	s	n
Dynamic	370	360	510	445	295	315	490	345	450	505	335	280	325	500	**394.6**	**84.7**	**14**
Static	430	445	455	455	490	535	NA	NA	NA	NA	NA	NA	NA	NA	**468.3**	**38.2**	**6**

a. Conduct a hypothesis test to determine if the mean operative time is different for the two procedures.

H_0: $\mu_1 = \mu_2$ vs H_a: $\mu_1 \neq \mu_2$

Observed test statistic: -2.6812 p-value: 0.0153 and est. effect size: -0.9860

b. Use your findings in (a) to complete the following sentence:

Based on your results from (a), there is ***(circle one)***

little ***some*** ***strong*** ***very strong*** ***extremely strong***

evidence that the null model is not compatible with our sample results. Our sample results indicate the estimated effect size is ***(circle one)***

small. ***small to moderate.*** ***moderate to large.*** ***large.***

c. The table below offers the results of three different follow-up studies. Which among them most *confirms (or replicates)* the findings of the study in (a)? Provide brief justification of your choice.

Study A has most similar effect size!

$S_p = 62.2722$ $S_p = 62.5127$ $S_p = 79.3632$

Study A	$\bar{x}$	s	n
Dynamic	389.2	89.1	70
Static	470.1	41.3	130

Study B	$\bar{x}$	s	n
Dynamic	389.2	41.3	22
Static	475.1	88.1	13

Study C	$\bar{x}$	s	n
Dynamic	480.1	89.1	15
Static	386.3	41.3	6

Study A: $t = -7.1920$, p-val: 0, $\hat{d} = -1.2991$

Study B: $t = -3.307$, p-val: 0.0047, $\hat{d} = -1.3741$

Study C: $t = 3.2886$, p-val: 0.0040, $\hat{d} = 1.1819$