STAT 218 – Midterm 2

Dr. Allison Theobold

May 17, 2020

Name: _____

Section (circle one): 8:10am 9:10am

Read and Sign the Following Statement:

I understand that give or receiving help on this exam is a violation of academic regulations and is punishable by a grade of **F** in this course. This includes looking at other students' exams and / or allowing other students, actively or passively, to see answers on my exam. This also includes revealing, actively or passively, any information about the exam to any member of Professor Theobold's STAT 218 class who has not yet taken the exam. The use of cell phones is strictly prohibited.

I pledge not to do any of these things.

Signed: _____

Instructions.

- Read and sign the honesty pledge at the top of this page. Your exam will not be graded unless the honesty pledge is signed!
- You may use a calculator. You **may not** use your phone or any device that connects to the internet as a calculator.
- Show all work as clearly as possible. Point totals are shown in brackets next to each part. Formulas without values entered do not count as work.
- All answers should be reported in decimal form, rounded to three decimal places.
- For multiple choice and multi-select problems, completely fill in the provided circle (multiple choice) or square (multi-select) for your desired answer choice(s). If you change an answer, be sure to completely erase your initial selection.
- You have 50 minutes to complete this exam, so budget your time wisely.

Golden Ticket

Scenario	One Quantitative Response	Two Quantitative Variables	Quant. Response and Categ. Explanatory
Type of plot	Dot plot, Histogram, Boxplot	Scatterplot	Faceted Histograms, Side-by-side Boxplots
Summary measure	Mean	Slope or Correlation	Difference in Means
Parameter notation	μ	Slope: β_1; Correlation: ρ	$\mu_1 - \mu_2$
Statistic notation	$\bar{x}$	Slope: b_1; Correlation: r	$\bar{x}_1 - \bar{x}_2$

Provided Formulas

$$IQR = Q3 - Q1$$

1.5 IQR Rule: above $Q3 + (1.5 \times IQR)$ or below $Q1 - (1.5 \times IQR)$

$$\hat{y} = b_0 + b_1 \times x$$

$$Residual = y - \hat{y}$$

t-based confidence interval: point estimate $\pm\, t^*_{df} \times SE$

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}$$

Q2[8 points] Is insomnia related to education level? Researchers at the University of Memphis, in Tennessee investigated this question in the *Journal of Abnormal Psychology* (February, 2005). Eight hundred (800) adults living in Tennessee were selected to participate in the study, using a random-digit telephone dialing procedure. Responses were collected from a total of 575 adults. Two variables measured were level of education and insomnia status (normal sleeper or chronic insomniac). The researchers discovered that the lower the level of education, the more likely the person was to have chronic insomnia.

(a)[2 points] What are the observational units in this study?

(b)[3 points] Fill in the blanks with one of the options provided in parentheses to identify and explain the study design.

This is a(n) (observational study / randomized experiment) _____ because (level of education / insomnia status) _____ (was / was not) _____ randomly (assigned / sampled) _____.

(c)[3 points] Which types of sampling bias may be present in this study? Select **all** that might be present, or if there is likely to be no bias present, select No bias.

(i) Selection bias

(ii) Response bias

(iii) Non-response bias

(iv) No bias

Q5[24 points] The *Journal of Food and Agriculture* contained an article titled "Influence of hydroponic and soil cultivation on quality and shelf life of ready-to-eat lamb's lettuce." In this article, researchers studied the effects of different hydroponic growing methods on the nitrate content of lettuce. In their study, the researchers randomly assigned 34 lettuce seedlings to one of three growing conditions: soil, hydroponic A, or hydroponic B. At the end of the growing period (60 days), nitrate measurements of the lettuce were taken (mg / kg).

Results from the study are presented in the table below.

Treatment Group	Mean Growth	Standard Deviation of Growth	Sample Size
Soil	3851	160.7	9
Hydroponic A	4652	81.08	12
Hydroponic B	3849	88.82	13

(a)[2 points] One of the researcher's main questions was to determine whether the growing method affects nitrate concentration in lettuce. Considering how this study was executed, can they address this question? *Briefly justify your answer.*

Below is an incomplete ANOVA table, summarizing the data. You may use this information for the subsequent problems.

term	df	sumsq	meansq	statistic	p.value
Growing Method	____	4998356	2499178	____	1.131e-18
Residuals	31	373482	12048	NA	NA

(b)[3 points] In the context of the research question and in plain English, what are the null and alternative hypotheses being tested in the ANOVA table above?

H_0**:**

H_A**:**

(c)[2 points] Rewrite the null hypothesis above to use notation for the *parameters* that are being tested.

H_0**:**

H_A**:**

(d)[2 points] The alternative hypothesis investigated in the ANOVA table above is:

$$H_A : \mu_{\text{Soil}} \neq \mu_{\text{Hydroponic A}} \neq \mu_{\text{Hydroponic B}}$$

True **False**

(e)[1 point] What are the degrees of freedom associated with `Growing Method`?

(f)[2 points] What is the value of the F-statistic?

(g)[2 points] The value of the F-statistic would be larger if the nitrate standard deviations were smaller for each group. Circle one.

True **False**

(h)[2 points] The value of the F-statistic would be larger if the nitrate means were more different across the groups. Circle one.

True **False**

(i)[1 point] Which distribution was used to obtain the p-value presented in the table? Circle one.

Simulated / Permuted Null Distribution **F-distribution**

(j)[4 points] Citing values from the ANOVA table to support your answer, what conclusions could be drawn regarding the hypotheses stated in (b)?

(k)[3 points] The table below presents all comparisons of the different soil treatments. What value of α should the researchers use to determine which, if any, of these tests produced "significant" results, so that the overall Type I error rate for these tests is less than 5%?

Group 1	Group 2	p.value
Hydroponic B	Hydroponic A	< 0.000001
Soil	Hydroponic A	< 0.000001
Soil	Hydroponic B	0.9779

Extra Credit[3 points] Based on the comparisons above, sketch what you would expect the side-by-side boxplots to look like. Be sure to label your x-axis and y-axis!

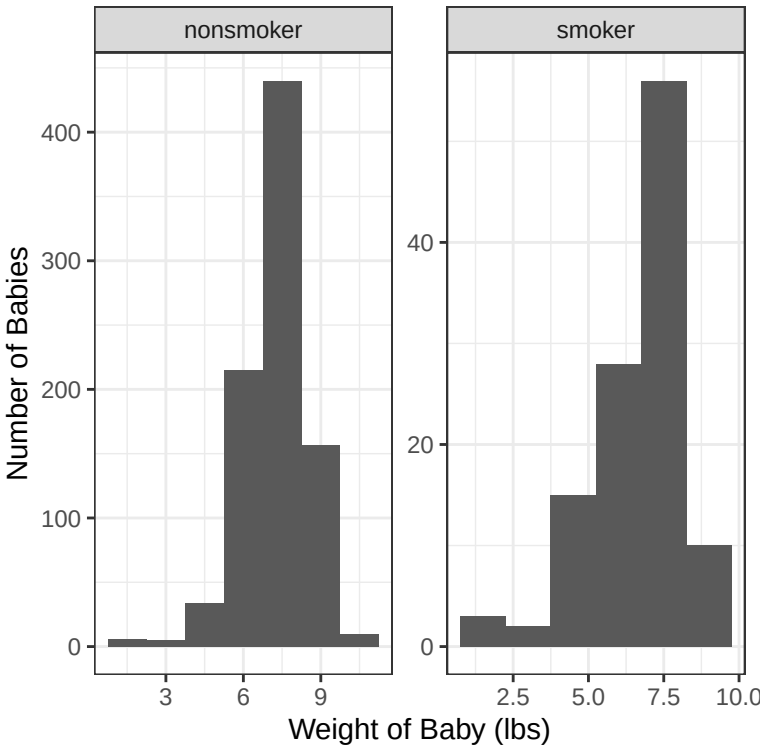
Q3[24 points] In 2004, the state of North Carolina released to the public a large dataset containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This analysis will focus on a random sample of 1,000 observations from the published dataset.

(a)[3 points] Hospital administration at Duke University Hospital are interested in the difference in the mean baby birth weight between mothers who do not smoke and mothers do smoke. Using the table below, report the observed statistic for this comparison. Indicate in your answer what notation should be used for this statistic.

Answer:

Notation:

habit	min	Q1	median	Q3	max	mean	sd	n	missing
nonsmoker	1.19	6.57	7.35	8.06	10.42	7.27	1.233	867	0
smoker	0.75	5.953	7.03	7.805	9.25	6.677	1.597	114	0



(b)[4 points] These hospital administrators are interested in *estimating* the true difference in mean birth weight between mothers who smoke and mothers who do not smoke. The administrators learned in their Statistics class how obtain a confidence interval for a difference in means using a t-distribution. Using the plots above evaluate whether it would be appropriate for the administrators to use a t-distribution to obtain a confidence interval for the true difference in means.

(c)[5 points] The administrators contacted the Department of Statistics at Duke and requested a consultation. The Statistician they spoke with suggested they use bootstrapping instead of a t-distribution to obtain a confidence interval. Fill in the blanks below to explain to the administrators how one bootstrap (re)sample is found.

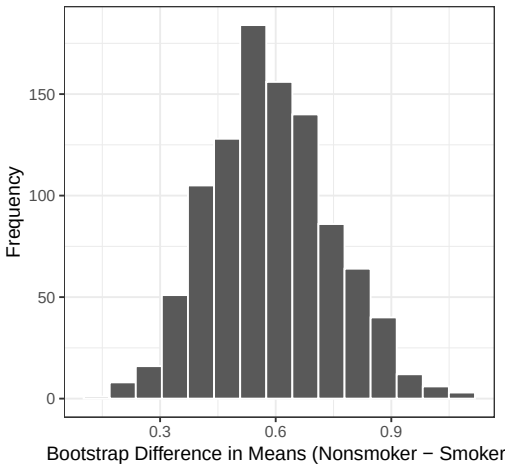
On _____ total cards, write _____ on the cards.

Generate a new sample by:

Calculate and plot the _____ from each bootstrap (re)sample.

(d)[3 points] The Statistician carried out the above process to find 1000 bootstrap resamples using the order (`nonsmoker - smoker`) and obtained the bootstrap distribution below.

Where is the bootstrap distribution centered? Why is the distribution centered at that value?



(e)[3 points] The table below presents percentiles for the bootstrap distribution shown above. Circle the two values which will construct a **99%** confidence interval.

Quantile	Value
0.5%	0.1953
1%	0.2485
2.5%	0.3099
5%	0.3387
90%	0.8016
95%	0.8623
97.5%	0.8982
99%	0.9515
99.5%	1.005

(f)[4 points] Interpret the 99% confidence interval found in part (e) **in the context of this study**.

(g)[2 points] Based on your confidence interval in (e), which of the following is the most likely p-value for a two-sided hypothesis test? Circle one.

(i) 0.20 (ii) 0.10 (iii) 0.05 (iv) 0.01 (v) 0.005