Name: ______________________ Lab Section/TA name: __________

Topics to be considered for final exam:

- Module 6: Hypothesis Testing
- Module 7: Simple Linear Regression
- Module 8: Multiple Regression
- Big picture ideas:
  - The difference between a population and sample
  - The difference between a parameter and a statistic
  - How are confidence intervals and hypothesis tests similar and different
  - Notation involving probabilities on a normal distribution and t distribution
  - Notation that has been used throughout the semester, including:
    * $\mu, \bar{x}, \sigma, s, n, z, t$
    * $H_0, H_A, r, R^2, \beta_0, \beta_1, \beta_2, \epsilon, b_0, b_1, b_2$
  - Definitions of key words that have been used throughout the semester, including:
    * mean, standard deviation, variance, confounding variables, percentile, distribution, sampling distribution, standard error, confidence intervals, margin of error
    * null and alternative hypothesis, p-value, Type I and Type II errors, correlation, response variable, predictor variable, intercept, slope, R squared

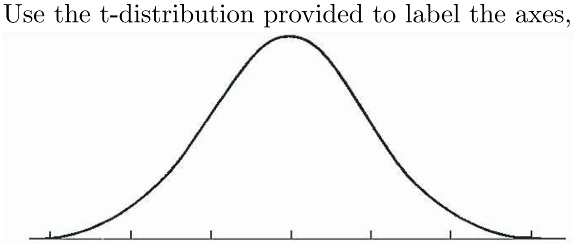
This practice exam will focus on modules 6,7,8 but you should review the big picture ideas on your own and be prepared for a couple of questions on those topics.

1. You recently got a complaint from a customer who said they were on hold with customer service for too long. You want to investigate the average time a caller spends on hold. The head of customer service claims that the average wait time is only 5 minutes, however you suspect it might be longer than that.

   (a) State the null and alternative hypothesis being tested.

   (b) You call in at 36 random times and the average time on hold is 6.1 minutes with a standard deviation of 3 minutes. Give the formula for the t-statistic and calculate the numerical result.

   (c) Use the t-distribution provided to label the axes, your t-statistic, and the p-value desired.

   (d) If you know that $P(t < 2.2) = .9827$ when degrees of freedom is 35, find the p-value desired.

   (e) Explain what the p-value represents in words and in the context of this problem.

   (f) Based on your hypothesis test what would you conclude? Do people spend on average more than 5 minutes on hold?

2. In September 2018 Nike launched a new ad campaign that caused some controversy. You want to see if tweeting in support of nike would change your number of twitter followers. You randomly select 30 people who have between 200 and 1000 twitter followers and who tweeted about Nike. You record the number of followers they had before they tweeted in support of Nike and after they tweeted. The excel output below shows the result of your test.

t-Test: Paired Two Sample for Means		
	After	*Before*
Mean	601.131959	595
Variance	65338.73543	54046.55172
Observations	30	30
Pearson Correlation	0.992176488	
Hypothesized Mean Difference	0	
df	29	
t Stat	0.877467938	
P(T<=t) one-tail	0.193722135	
t Critical one-tail	1.699127027	
P(T<=t) two-tail	0.38744427	
t Critical two-tail	2.045229642	

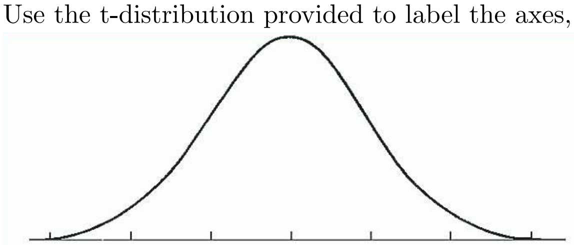
(a) Is this data paired or not paired?

(b) State the null and alternative hypothesis being tested.

(c) Give the formula for the t-statistic and give the numerical result.

(d) Give the p-value for this hypothesis test.

(e) Use the t-distribution provided to label the axes, your t-statistic, and the p-value.



(f) In the sample, did people on average tend to gain or lose followers after tweeting in support of Nike?

(g) Based on the hypothesis test what should you conclude about the difference in means? Does tweeting in support of Nike change your number of twitter followers significantly?

(h) Based on the p-value of your test, would you expect the confidence interval for the difference in means to include zero? Explain why or why not.

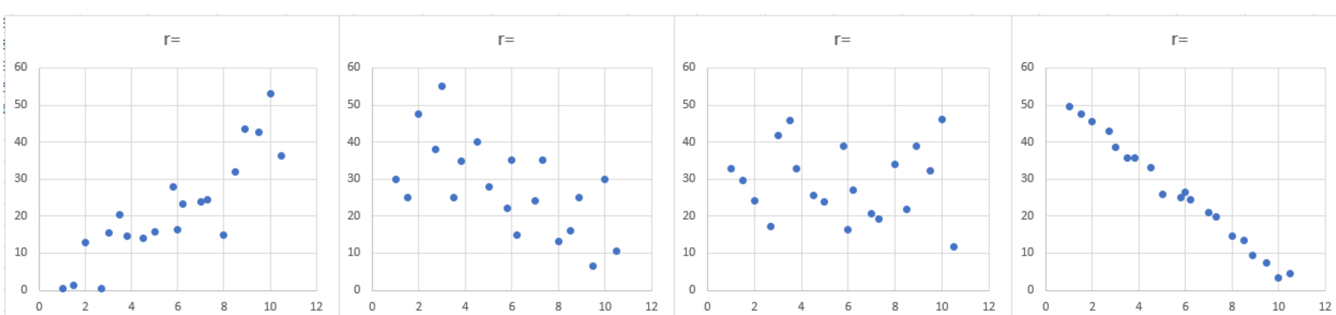
3. Put the following t-statistics in order from smallest p-value to largest p-value, assuming the same degrees of freedom for all and a two sided test.

t = 3.2 t = -.5 t=1.4 t = -1.3

smallest p-value __________ __________ __________ __________ largest p-value

4. Fill in each graph with it's correct correlation. Not all values will be used.

r=-1.4 r=-.95 r=-.6 r=0 r=.75 r=1.1



The next two questions will use the following information. Insurance companies use personal information to predict the medical costs you will need covered during a year in order to set policy prices and make decisions. The following variables are measured for a sample size of 1338 people.

- Charges: Individual medical costs billed by health insurance
- Age : age of insurance policy holder (between 18 and 65)
- BMI: Body mass index, a measurement of height and weight
- Children: Number of children covered by health insurance plan
- Smoker: 1 if the person is a smoker, 0 if the person is not a smoker

5. First let's use only Age to predict Charges.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3165.885	937.1494651	3.37821	0.00075	1327.4403	5004.3297
age	257.72262	22.50238929	11.4531	4.9E-29		

(a) Give the theoretical model you are estimating.

(b) Based on the excel output give the line of best fit.

(c) Interpret the slope for age in context of this problem.

(d) Perform a hypothesis test to check if the slope of the line is equal to zero.

H_0 :

H_A :

$t =$

$p - value =$

conclusion:

(e) Give the confidence interval for the slope of age.

95% confidence interval for ________ : (________ , ________)

(f) Predict the amount in charges to insurance when the person's age is 60 years old.

(g) There is a 60 year old in the data set whose charges were $12147. What is the residual error?

6. Next, let's use all the variables we have in the model as predictors: Age, BMI, Children, and Smoker.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-12102.769	941.9839411	-12.8482	1.1E-35	-13950.7	-10254.837
age	257.84951	11.89638633	21.6746	1.7E-89	234.51183	281.18719
bmi	321.8514	27.37763213	11.756	2E-30	268.14346	375.55934
children	473.50232	137.7916715	3.43636	0.00061	203.19016	743.81447
smoker	23811.4	411.2197148	57.9043	0	23004.692	24618.108

(a) Give the theoretical equation of the model in this situation.

(b) Give the estimated line of best fit.

(c) Interpret the slope for children.

(d) Does being a smoker significantly affect the amount your charge for insurance, assuming all other predictors are constant? Explain why or why not.

(e) Use the model to predict the amount in charges to insurance for a 40 year old nonsmoker with 2 children and BMI of 25.

(f) The model with age only had an R squared value of .09 and the model with all the predictors had an R squared value of .75. Which model is better? Explain why.

(g) Let's say this insurance company covered 4 states, Washington, Oregon, Idaho, and Montana. You want to add the state the customer was from to this model.

i. Explain how you would put the customers' state into the data set so that you could use it in your model.

ii. Give the new population model you would be using.

7. Explain what extrapolation is and why it is something we should avoid.