Types of multiple choice questions:

1. Do heavier cars use more gasoline? To answer this question, a researcher randomly selected 15 cars. He collected their weight (in hundreds of pounds) and the mileage (MPG) for each car. From a scatterplot made with the data, a linear model seemed appropriate. Which of the following descriptions of the value of the slope is the correct description?

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.44	6.275	6.445	0
Weight	-0.521	0.164	-3.182	0.0072

   a) We cannot interpret the slope because we cannot have a negative weight of a car.
   b) We estimate the mileage to decrease by 0.521 miles per gallon when the weight of a car increases by 1 pound.
   c) We estimate the mileage to decrease by 52.1 miles per gallon when the weight of a car increases by 100 pounds.
   d) We estimate the mileage to decrease by 0.521 miles per gallon when the weight of a car increases by 100 pounds.

2. Suppose you are testing the null hypothesis that the slope of the regression line is zero versus the alternative hypothesis that the slope is different than zero. Would a very small *P*-value (i.e., less than 0.0001) indicate a strong relationship between the explanatory variable and the response variable?
   a) Yes, because the *P*-value will give the strength of the association between the explanatory variable and the response variable.
   b) Yes, because if the *P*-value is small, then R^2 is large.
   c) No, because there could be a lot of scatter about the regression line, indicating a weak relationship between the explanatory variable and the response variable.
   d) No, because a large *P*-value would indicate a strong relationship between the explanatory variable and the response variable.

3. In a simple linear regression model, the deviations, ϵ_i, are assumed to be _____.
   a) $N(0,1)$
   b) $N(0,\sigma^2)$
   c) $t(1)$
   d) $t(0)$

4. When trying to explain the relationship between two quantitative variables, it would be best to use a _____.
   a) scatterplot
   b) density curve
   c) boxplot
   d) histogram

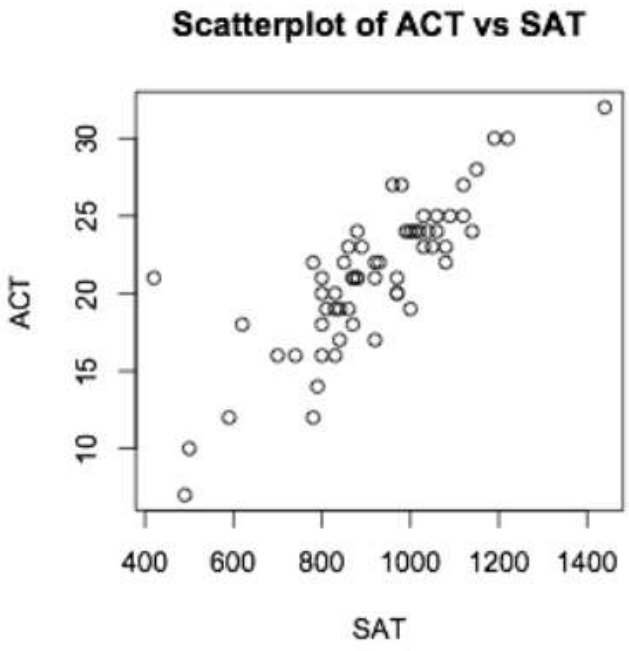
6. Does the major of a Purdue student have a significant effect on how much on average the student spends on ice cream during the year? What is the most appropriate statistical technique that should be used to analysis this?
   a) 1-sample
   b) 2-sample independent
   c) 2-sample paired
   d) ANOVA
   e) linear regression

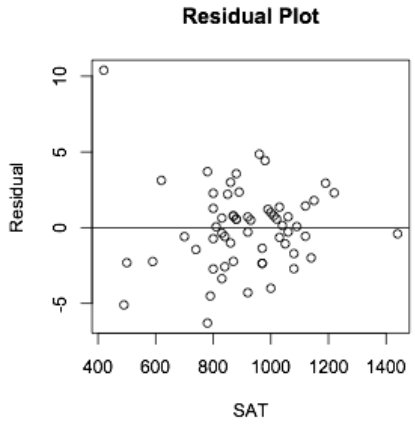
7. What is the average price of ice cream in West Lafayette? What is the most appropriate statistical technique that should be used to analysis this?
   a) 1-sample
   b) 2-sample independent
   c) 2-sample paired
   d) ANOVA
   e) linear regression

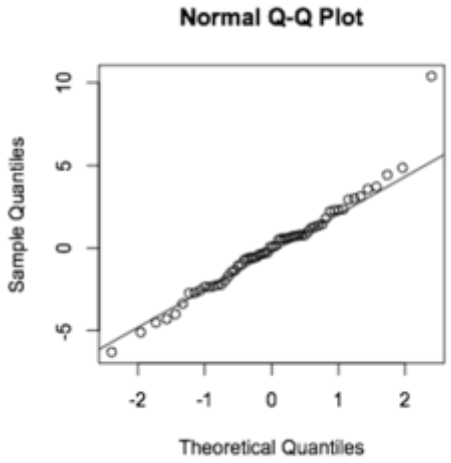
8. Can we predict the ice cream sales (in dollars) from the outdoor temperature (in degrees)? What is the most appropriate statistical technique that should be used to analysis this?
   a) 1-sample
   b) 2-sample independent
   c) 2-sample paired
   d) ANOVA
   e) linear regression

9. At each store, we check the price for both soft-serve ice cream and hard-packed ice cream and want to know if there is a significant difference in the price between the kinds of ice cream. What is the most appropriate statistical technique that should be used to analysis this?
   a) 1-sample
   b) 2-sample independent
   c) 2-sample paired
   d) ANOVA
   e) linear regression

1. (38 points) The SAT and the ACT are the two major standardized tests that colleges use to evaluate candidates. Most students take just one of these tests. However, some students take both. Consider the scores of 60 students who took both tests. A scatterplot, residual plot, QQplot of the residuals, and regression output are shown below.







Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**1.63**	**1.84**	**xxxxx**	**0.382**
SAT	**xxxxx**	**xxxx**	**xxxxx**	**0.000**

S_{XX} = 3797.04 **S_{YY} = 3.208** **S_{XY} = 90.14**

1. (Cont'd)

a) (6 points) What are the conditions or assumptions for using linear regression analysis? For each of the assumptions, indicate the graph(s) or equations that would be used to verify them.

1	The data must be a **SRS.**	0.5	no plot
1	The relationship between x and y must be **linear**	0.5	scatterplot, residual plot
1	The residuals have **Normally** distributed	0.5	QQ plot, histogram
1	The residuals have to have the **same variance**	0.5	scatterplot, residual plot

b) (6 points) Using the information provided, write the equation of the fitted regression line.

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{90.14}{3797.04} = 0.0237$$

ACT = 1.63 + 0.0237*SAT (2 points each for intercept and slope and equation of line)

c) (4 points) What is the value of the correlation coefficient r?

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{90.14}{\sqrt{(3797.04)(3.208)}} = 0.82$$

(2 pts equation, 2 pts work)

d) (2 points) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity?

$r^2 = (0.82)^2 = 0.667$ (1 pt for realizing that you are supposed to calculate r^2, 1 pt. work)

e) (10 points) Find the 95% confidence interval for the slope parameter. Make sure to interpret your interval. (Assume that the conditions for the inference are satisfied.)

1	$b_1 \pm t_{\frac{\alpha}{2},n-1}\, SE(b_1)$
2	DF = n - 2 = 60 - 2 = 58
2	t = 2.0086 (using df = 50)
2	0.0237 ± 2.0086*__0.00199__ $= 0.0237 \pm 0.003988$
1	(0.0197, 0.0277) or (0.020, 0.028)
2	We are 95% confident that the population (true) slope is between 0.020 and 0.028.

$$MSE = \frac{SST - SSR}{dfe} = \frac{S_{YY} - b_1 S_{XY}}{n-2} = \frac{3.288 - 0.0237 \cdot 90.14}{58} = \frac{1.152}{58} = 0.00199$$

f) (10 points) Is there an association between ACT score and SAT score? If so, does a high score on the SAT test cause a high score on the ACT test? Please explain your answer for both parts.

3	Yes, there is an association between ACT score and SAT score because 0 is not in the confidence interval for the slope.
2	From the scatter plot and part (c), there is an obvious positive association between ACT score and SAT score.
3	However, a high score on the SAT test does not cause a high score on the ACT test.
2	Strong correlation does not imply causation. Student's knowledge and learning is the lurking variable.