

1. (8 pts; 2 pts each) Multiple choice: Circle the best answer.

a) This is very unrealistic, but suppose that for every student in this class, the sum of their exam1 score and their exam2 score equals exactly 80. What would the numerical value of the correlation coefficient between exam1 score and exam2 score be?

The correlation coefficient equals -1, because the points would fall on a perfectly straight line with negative slope. (Because every student has the same total for the two exams combined, those who did well on one exam must have done poorly on the other.)

b) It can be shown that the *sum* of the residuals from a least squares regression line equals zero. What does this fact imply about the *mean* and *median* of the residuals?

The mean must equal zero (because the mean is the sum divided by the number of observations). But the median does not necessarily equal zero, as the following small example illustrates: -4, -2, 1, 2, 3.

c) In the study comparing four popular diets, the ANOVA *F*-test produced a *p*-value of 0.659. If we were to examine Tukey multiple comparison intervals for comparing pairs of diets, what would we most likely find?

All intervals would contain 0, because the large *p*-value indicates that none of the group means differ significantly from each other..

d) Recall that we simulated hundreds of repetitions of a random draft lottery, to investigate how random the famous 1970 draft lottery really was. How did the simulation turn out, and what did we conclude?

We *never* got a correlation as large as the actual result, so we concluded that the draft lottery was probably *not* random.

2. (14 pts) Researchers studied heart rates after engaging in physical exercise for adults who were also classified according to whether and how much they smoke. Data were collected to investigate whether there are differences in mean heart rates among various smoking classifications (heavy, light, moderate, non-smoker). The Minitab output below (with some entries edited out) pertains to an ANOVA analysis addressing this issue:

Source	DF	SS	MS	F	P
smoking status	3	1464.1	_____	_____	0.004
Error	20	1604.8	80.2		
Total	___	3069.0			

Level	N	Mean	StDev
heavy	6	81.667	10.821
light	6	63.167	8.208
moderate	6	71.667	9.201
non	6	62.333	7.202

a) (1 pt) How many subjects participated in this study?

24

b) (2 pt) State the appropriate null hypothesis in symbols.

$H_0: \mu_{\text{heavy}} = \mu_{\text{moderate}} = \mu_{\text{light}} = \mu_{\text{non}}$
(where μ_i represents the population mean heart rate in smoking level i)

c) (2 pts) Comment on whether the technical condition about standard deviations is satisfied.

Yes, because the ratio of the largest sample standard deviation to the smallest one is less than 2:
 $10.821 / 7.202 \approx 1.50$.

d) (3 pts) Determine the value of the F -statistic.

$MS(\text{smoking status}) = 1464.1 / 3 \approx 488.033$

$F = MS(\text{between}) / MS(\text{within}) = MS(\text{smoking status}) / MS(\text{error}) = 488.033 / 80.2 \approx 6.085$

e) (3 pts) Summarize the conclusion that you would draw from the ANOVA F -test.

Because the p -value is so small (.004), the sample data provide very strong evidence that there is a difference in population mean heart rate among these four smoking groups. At least one of the group's mean heart rate differs significantly from another group's.

f) (3 pts) Using the Tukey multiple comparisons procedure reveals that at the .05 level, the pairs of groups that differ significantly are (heavy, light) and (heavy, non). Represent this finding with appropriate underlining of the group means below:

Non	Light	Moderate	Heavy
62.333	63.167	71.667	81.667

The top line indicates that the non-smoking group differs significantly from the heavy smoking group but not from the other two groups. The bottom line indicates that the heavy smoking group differs significantly from both the light- and non-smoking groups but not from the moderate groups.

3. (8 pts) In a recent study, researchers investigated possible biochemical mechanisms that could be involved in the early stages of romantic love. They measure plasma level of neurotrophins for a sample of 58 subjects who had recently fallen in love. They also asked each subject to rate his/her level of passionate love feelings on a numerical scale. Researchers calculated the correlation coefficient between the level of passionate love and plasma level of neurotrophins to be $r = 0.34$.

Conduct the appropriate test of whether this sample provides strong evidence (at the $\alpha = .05$ level) of a positive correlation between these variables in the population. Report the relevant hypotheses, test statistic, p-value, test decision, and conclusion.

The hypotheses are $H_0: \rho = 0$ vs. $H_a: \rho > 0$, where ρ represents the correlation coefficient between level of passionate love and plasma level of neurotrophins in the population of all people who have recently fallen in love. (It is equivalent to state the hypotheses in terms of the population slope coefficient: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 > 0$,

The test statistic is: $t = r \sqrt{\frac{n-2}{1-r^2}} = 0.34 \sqrt{\frac{58-2}{1-(0.34)^2}} \approx 2.71$.

The p-value is found from the right tail of a t-distribution with 56 degrees of freedom. Looking in Table T reveals that $p\text{-value} < .005$.

Because $p\text{-value} < .05$, we reject the null hypothesis at the $\alpha = .05$ significance level.

The sample data provide very strong evidence that there is a positive correlation between level of passionate love and plasma level of neurotrophins in the population of all people who have recently fallen in love.

4. (20 pts) Between the months of September 1990 and May 1997, a statistics teacher gathered data on the average temperature for that month (in degrees Fahrenheit) and the amount of gas usage in his home for that month (in units called therms). Summary statistics for these variables follow:

Variable	N	Mean	SE Mean	StDev	Median	IQR
avg temp	71	46.35	1.80	15.16	45.00	26.00
gas usage per day	71	5.311	0.420	3.538	5.000	6.600
Pearson correlation of avg temp and gas usage per day = -0.930						

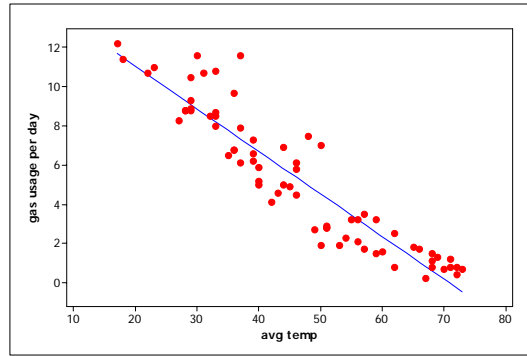
a) (4 pts) Determine and report the equation of the least squares line for predicting a month's gas usage per day based on its average temperature.

The slope coefficient is: $b_1 = r \frac{s_y}{s_x} = (-0.930) \times (3.538) / (15.16) \approx -0.217$

The intercept coefficient is: $b_0 = \bar{y} - b_1 \bar{x} = 5.311 - (-0.217) \times 46.35 \approx 15.369$

The equation of the least squares line is therefore:
 predicted gas usage per day = $15.369 - 0.217 \times \text{distance}$.

b) (1 pt) In the following scatterplot with the least squares line superimposed, circle the point corresponding to the month with the largest positive residual:



The largest positive residual corresponds to the point farthest above the least squares line.

c) (3 pts) Calculate the value of r^2 , and write a sentence interpreting what this value means.

The value of $r^2 = (-0.930)^2 \approx 0.865$, which means that 86.5% of the variability in gas usage per day across these months is explained by knowing the average temperature for the month.

d) (2 pts) Predict the gas usage per day for a month in which the average temperature is 50 degrees Fahrenheit.

This prediction is: predicted gas usage per day = $15.369 - 0.217 \times 50 \approx 4.519$ therms per day.

e) (2 pts) A significance test or confidence interval for the slope coefficient will be based on how many degrees of freedom?

The degrees of freedom is $n - 2 = 71 - 2 = 69$.

f) (3 pts) Minitab reports the standard error of the sample slope coefficient to be 0.01036. Use this information to produce a 95% confidence interval for the population slope coefficient.

This 95% confidence interval is: $b_1 \pm t^* \times SE(b_1)$, which is $-0.217 \pm 2.000 \times 0.01036$, which is $-0.217 \pm .021$, which is the interval $(-0.238, -0.196)$.

g) (3 pts) Write a sentence interpreting this interval, including what the slope coefficient means.

We are 95% confident that the predicted *decrease* in gas usage for each additional degree (F) of temperature is between 0.238 therms/degree and 0.196 therms/degree.

h) (2 pts) At what temperature will a prediction interval for gas usage per day be narrowest?

The prediction interval is narrowest at the mean temperature: 46.35 degrees Fahrenheit.