Name:__________________ Section: __________

Statistics 311 **Final Exam** 18 March 2008

Please read the following instructions carefully. **Do not turn the page and start the exam until you are told to begin.**

You may have the formula card that comes with the textbook (or a copy of the same), and the key formula sheet posted on the course website. You may also have one sheet (double-sided) of 8.5 x 11 paper with notes (handwritten or typed). You may use a calculator, although be sure to show your work.

There are 2 multi-part short answer problems, 5 multiple choice problems, and 5 true/false problems worth a total of 100 points. Total points for each question are designated in parentheses at the beginning of the problem, with specific points for each part designated separately.

GOOD LUCK!

Name:__________________ Section: __________

Statistics 311 **Final Exam** 18 March 2008

Problem 1 (30 points): A company institutes an exercise break for its employees to see if this will improve job satisfaction, as measured by responses to a questionnaire that assesses worker's satisfaction (higher scores ➔ more satisfaction). The data based on **before and after** surveys of five randomly selected employees, are as follows:

Employee Number	Job Satisfaction Index			These columns are provided to help you calculate s_d (not required)	
	Before	After	d_i	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1	34	33	-1	-8	64
2	28	36	8	1	1
3	29	50	21	14	196
4	45	41	-4	-11	121
5	26	37	11	4	16

a) Complete the d_i column in the table. (4 points)

b) Compute $\bar{d}$ and s_d. (6 points)

$$\bar{d} = \frac{35}{5} = 7; \quad s_d = \sqrt{\frac{398}{5-1}} = \sqrt{\frac{398}{4}} = \sqrt{99.5} = 9.97$$

c) Write down appropriate null and alternative hypotheses to test the effectiveness of the exercise program. (4 points)

$H_0: \mu_d = 0$ vs. $H_1: \mu_d > 0$ where d = after − before NOTE: alternative could be in other direction if d = before - after

d) Construct the appropriate test statistic for your test in part c). (3 points)

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{7}{\frac{9.97}{\sqrt{5}}} = \frac{7}{4.46} = 1.57$$ NOTE: t = -1.57 if d = before - after

e) Using a 5% significance level, what is the critical value for this test? What is the *P*-value? What is your decision for this test? (4 points)

$t_{crit} = t_{.05(1),4} = 2.132$

$P - \text{value} = P(t_4 > 1.57) \Rightarrow 0.05 < P < 0.10$ using table

$P = 0.0958$ using TI-84

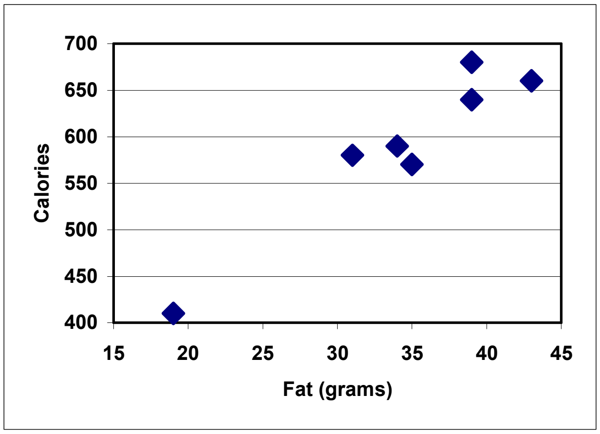
f) Using your test statistic and critical value (or *P*-value), state the conclusion from your hypothesis test in relation to the original question. (4 points)

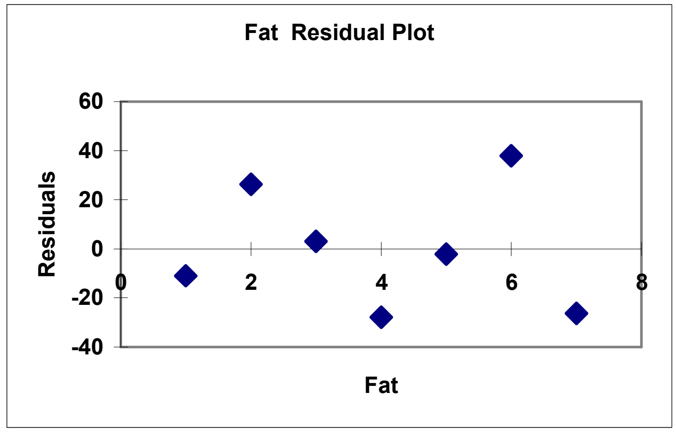
Fail to reject H_0. There is insufficient sample evidence to suggest that the exercise break improves job satisfaction at the 5% significance level.

g) Which type of statistical error do you risk committing based on the decision given in part e)? What does this potential error mean in the context of the problem? (5 points)

It is possible to have committed a Type II error—by failing the reject the null there is an unknown probability, β, that we made an incorrect decision and that the exercise break does indeed improve job satisfaction.

Problem 2 (40 points). Fast food is often considered unhealthy because much of it is high in both fat and calories. Can fat content be used to predict calories? To try to answer this question, seven fast-food hamburgers were analyzed for fat content and calories. A plot of the data, Excel regression output, and a residual plot are shown below. In addition, $\bar{x} = 34.286$ and $SS_{xx} = 365.429$.





Excel SUMMARY OUTPUT

Regression Statistics	
R Square	**0.923**
Standard Error	**27.334**
Observations	**7**

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	**210.954**	**50.101**	**4.211**	**0.0084**	**82.164**	**339.744**
Fat	**11.056**	**1.430**	**7.732**	**0.0006**	**7.380**	**14.731**

Note: Excel's reported *P*-values are for two-tailed hypotheses.

a) Based on the information provided above, list two things that make you think that using simple linear regression as a model might be a reasonable approach. (2 points)

1) The scatter plot shows a reasonably increasing linear relationship between fat content and calories.
2) r = 0.96, confirming the presence of a strong positive linear association

b) Write down the equation for the regression line. (3 points)

$\hat{y} = 11.056x + 210.954$

c) Interpret the estimated slope value in the context of the problem. (3 points)

For each one gram increase in fat content, a burger will have about 11 more calories, on average.

d) Does the estimated y-intercept have any practical meaning in the context of the problem? **Briefly explain.** (3 points)
It is possible to have fat free food (0 g fat), although I don't think that would be the case for red-meat burgers. If you could make a fat free burger, the y-intercept value of 211 calories would represent the calories due to things other than fat—carbs, protein, etc.

e) What is the 95% confidence interval for the regression equation slope parameter? ***Interpret the CI in the context of the problem.*** (3 points)
From the Excel output the 95% CI for the population slope parameter is (7.38, 14.73). We are 95% confident that the average increase in calories for each one gram increase in fat will be between 7.8 and 14.7 calories.

f) What is the margin of error, E, that would be used for calculating a 90% confidence interval for the regression equation slope parameter? Would the 90% CI be wider or narrower than the 95% interval from part e)? ***Briefly explain.*** (4 points)
$E = t_{\alpha/2, df} SE_{b_1} = t_{0.10(2),5} SE_{b_1} = 2.015(1.43) = 2.88$
The 90% CI will be narrower than the 95% CI—if we are willing to be less confident, we can form a narrower interval.

g) Write down the null and alternative hypotheses to test that the slope is greater than zero. (3 points)

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 > 0$

h) What is the t-score that is reported by Excel to test the hypotheses you specified in part g)? (2 points)
$t = 7.732$

i) What is the P-value for the test you specified in part g)? (2 points)
$P = 0.0006/2 = 0.0003$

j) Based on the Excel output what do you conclude, in the context of the problem, for the test specified in part g)? Use $\alpha = 0.01$. (3 points)

Reject the null. Conclude that the slope parameter is statistically greater than zero—there is a positive increase in calories as fat content increases.

k) What is the expected value for average number of calories when a burger has 25 grams of fat? (2 points)
$\hat{y} = 11.056(25) + 210.945 = 276.4 + 210.945 = 487.35$ calories

l) Calculate a 98% prediction interval for the calories in a new randomly selected burger, given that the burger has 25 grams of fat. ***Interpret the interval in the context of the problem.*** (6 points)

$$\hat{y} \pm t_{0.02(2),5} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_{xx}}} \Rightarrow 487.35 \pm 3.365(27.334)\sqrt{1 + \frac{1}{7} + \frac{(25 - 34.286)^2}{365.429}}$$

$$\Rightarrow 487.35 \pm 91.979\sqrt{1.3788} \Rightarrow 487.35 \pm 91.979(1.174)$$

$$\Rightarrow 487.35 \pm 108.005 \Rightarrow (379.345, 595.355)$$

There is a 98% chance that the new randomly selected burger will have between 379 and 595 calories, given that the burger contains 25 grams of fat.

m) In the context of the problem, what does the R Square value reported in the Excel output mean? ***Limit your response to one sentence.*** (2 points)

About 92.3% of the variation in calories of burgers can be explained by knowing the fat content of the burger.

n) Take a look at the fat residual plot. ***In one sentence***, what does this plot tell in terms of the model assumptions regarding the residuals? (2 points)
The fat residual plot shows that the residuals are centered around zero with a similar pattern above and below the zero line, which is consistent with the assumption that ε iid $N(0, \sigma^2)$.

Problem 3 (15 points total, 3 points each): Indicate whether each of the following statements is True or False.

TRUE A confidence interval for the average value of y at a given value $x = x_0$, formed as part of a linear regression analysis, will always be narrower than the corresponding prediction interval estimate for a new individual value of y at the same value $x = x_0$.

TRUE One assumption for simple linear regression is that the epsilons are normally distributed with constant variance.

FALSE Power is equal to $1 - \alpha$, where α is the probability of a Type I error .

TRUE The level of significance associated with a significance test is the probability of committing a Type I error.

FALSE A statistical hypothesis test is always conducted under the assumption that the alternative hypothesis is true.

Problem 4 (15 points total, 3 points each): Answer the following multiple choice questions by circling the letter with the correct response:

a) We have created a 95% confidence interval for μ with the result ($10 < \mu < 15$). What conclusion will we make if we test $H_0: \mu = 16$ vs. $H_1: \mu \neq 16$ at $\alpha = 0.05$?

A. Reject H_0 in favor of H_1.

B. Accept H_0 in favor of H_1.

C. Fail to reject H_0.

D. We cannot tell what our decision will be with the information given.

b) The value that separates a rejection region from an acceptance region is called a __________.

A. parameter

B. confidence coefficient

C. significance level

D. critical value

c) A hypothesis test is used to prevent a machine from under-filling or overfilling quart bottles of beer. On the basis of a sample, the null hypothesis is rejected and the machine is shut down for inspection. A thorough examination reveals there is nothing wrong with the filling machine. From a statistical point of view:

A. A correct decision was made.

B. A Type I and Type II error were made.

C. A Type I error was made.

D. A Type II error was made.

d) The larger the *P*-value, the

A. weaker the evidence against the alternative hypothesis.

B. stronger the evidence for the null hypothesis.

C. stronger the evidence against the null hypothesis.

D. None of the above.

e) We never conclude "Accept H_0" in a test of hypothesis because _____________.

A. α is the probability of a Type I error.

B. the rejection region is not known.

C. the *P*-value is not small enough.

D. $\beta = P$(Type II error) is not known.