Stat 401 **FINAL EXAM** **Dec.20, 2006**
J.Millson

1. The mature heights of 5 tomato plants treated with Gro-Food once a week were 24,25,24,26,27 inches respectively and the mature heights of 5 different tomato plants treated with Gro-Food twice a week were 23,28.7,26,24,27.5 inches respectively. Assume that the distributions of growth are normal.

(a) Test whether the variances of the growths are equal using the two-sample F-test on your calculator. Make your decision on the basis of whether or not the resulting P-value for the F-test is large or small.

(b) Do a two-sample t-test to test whether there is a significant difference in the levels of growth using the two treatments (use the level $\alpha = .01$ and do a two-sided test). Use your answer to (a) to decide whether you should do a pooled or unpooled two-sample t-test. There will be five points for making the correct decision with the correct justification.

(20 points)

2. Let $X_1, X_2, \cdots, X_m$ be random sample from the space of random variable X with $N(\mu_1, \sigma^2)$ distribution and $Y_1, Y_2, \cdots, Y_n$ be a random sample from the space of a random variable Y with an $N(\mu_2, \sigma^2)$ distribution. Let $S_p^2 = \frac{m-1}{m+n-2}S_1^2 + \frac{n-1}{m+n-2}S_2^2$ be the pooled sample variance. In what follows you may assume

Theorem A. The random variable $T = (\,\overline{X} - \overline{Y} - (\mu_1 - \mu_2)\,)\,/(\,S_p\sqrt{1/m + 1/n}\,)$ has t-distribution with $m + n - 2$ degrees of freedom.

Recall that the upper-tailed pooled t-test for deciding

$$H_0 : \mu_1 = \mu_2$$
$$versus$$
$$H_a : \mu_1 > \mu_2$$

is given by the decision rule:

reject H_0 if

$$\overline{x} - \overline{y} \geq t_{\alpha,m+n-2}\ s_p\sqrt{1/m + 1/n}.$$

(i) Prove that the upper-tailed pooled t-test has significance level α.

(ii) Prove that the random interval

$$(\overline{X} - \overline{Y} - t_{\alpha,m+n-2}S_p\sqrt{1/m + 1/n},\ \infty\ )$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

(20 points)

3. There is a theory espoused by some baseball fans that the number of home runs a team hits is markedly effected by the altitude of the club's home park, the rational being that the air is thinner at the higher altitudes and the balls should go further. On the next page are the altitudes of the 6 American League East ballparks and the number of home runs each of those teams hit during the 1972 season.

Club	Altitude (ft)	Number of Home Runs
Cleveland	660	138
Milwaukee	635	81
Detroit	585	135
New York	55	90
Boston	21	130
Baltimore	20	84

(a) Find the least squares line corresponding the above data (y is the number of home runs and x is the altitude).

(b) Find r^2 the coefficient of determination.

(c) Taking into account your answer to (b), is the linear model a good one?

(d) Test H_0: The altitude makes no difference, against the *appropriate* alternative hypothesis at level $\alpha = .1$. Give a precise statement of H_0 (in terms of β_1 or ρ) instead of the above vague statement of H_0 and a precise statement of your alternative hypothesis (upper-tailed, lower-tailed or two-sided).

(e) Later, Denver (altitude 5000 ft) was added to the National League. On the basis of your answer to (a), predict how many home runs would be hit in the Denver ball park (Mile High Stadium). (Use the handout "Regression on your calculator" to paste the regression line into Y_1, then use the handout again to evaluate $Y_1(5000)$ or else use the formula you found in (a) for the least squares line and plug in 5000).

(20 points)

4. The point of this problem is to find the least square parabola $y = ax^2 + bx + c$ that best fits data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ in a special case. The formula for the least squares parabola is formula (13.10) in Section 13.3 of your text (k=2) but you shouldn't use this formula. *You can use your calculator to find the parabola.*

The following table gives the stopping distance D (in feet) of an automobile travelling at a speed of V (miles per hour) at the instant danger is sighted.

V	20	30	40	50	60	70
D	54	90	138	206	292	396

(a) Find the least squares parabola (enter the data , then go to $STAT \to CALC \to QuadReg$).

(b) Estimate D when V is 80 miles per hour (paste the regression PARABOLA into Y_1 and follow the instructions on the handout to evaluate it just as you did for the regression LINE.).

(20 points)

5. The goal of this problem is to prove the equality of the two t's of Chapter 12. The first t which we will call t_{old} is the t from the linear regression t-test. It is defined by the formula

$$t_{old} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}.$$

The second t which we will call t_{new} is the t from the test for correlation between two random variables whose joint distribution is bivariate normal. It is defined by the formula

$$t_{new} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

where the sample correlation r is defined by the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}.$$

Prove

$$t_{old} = t_{new}$$

by expressing both sides in terms of S_{xx}, S_{yy} and S_{xy}. You should get the same expression.

Here is how to do the problem.

Express t_{old} in terms of S_{xx}, S_{yy} and S_{xy} using the following formulas (you may assume they are true without proving them)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1}$$

$$SSE = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \tag{2}$$

$$s_{\hat{\beta}_1} = \sqrt{\frac{SSE}{(n-2)S_{xx}}} \tag{3}$$

Now express t_{new} in terms of S_{xx}, S_{yy} and S_{xy} by substituting the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

for r in the defining formula for t_{new} and simplify. The problem is high-school algebra - you don't need the expressions for S_{xx}, S_{yy} and S_{xy} in terms of the x_i's and the y_i's to do this problem.

(20 points)