Stat 5021, Final Exam, Spring '15 Name _____

You may use one 8" by 11" formula sheet (both sides) but you **may not use** any electronic computing device. There is no need to reduce numerical formulas to their simplest form. Your answers may contain R commands for finding quantiles of the appropriate distribution.

Thirty six people took the exam. The high score was 95. There were 3 in the 90's, 4 in the 80's, 8 in the 70's, 6 in the 60's, 6 in the 50's, 5 in the 40's, and 4 in the 30's.

1. Suppose the diameter of ball bearings produce by a factory follow a normal distribution with a mean of 10 centimeters and a variance of .04 centimeters. Assume a random sample of seven such ball bearings was collected.

i) What is the excepted number of sampled bearings whose diameters fall between 9.8 and 10.1?

ii) Give a **mathematical expression** (no R commands here) for the probability that at least five of the sample bearings have a diameter between 9.8 and 10.1.

Solution:

i)

$$P(9.8 < N(10, .04) < 10.1) = pnorm(10.1, 10, .2) - pnorm(9.8, 10, .2) = 0.53$$

So $7(.53) = 3.71$ is the expected number.

ii) $\sum_{x=5}^{7} \binom{7}{x} .53^x .47^{7-x}$

∎

2. i) In a survey of 95 randomly selected Minnesota residents 65 stated that they had traveled outside the state in the last six months. Give a **mathematical expression** for the 99% confidence interval for the true proportion p_1 of Minnesota residents who have traveled outside the state in the last six months.

Solution:

$$\frac{65}{95} \pm 2.58\sqrt{\frac{65}{95}\frac{30}{95}\frac{1}{95}}$$

since $qnorm(.995) = 2.575829$ ∎

ii) In a similar survey of Florida residents 42 of 85 surveyed state that they had traveled outside the state in the last six months. If p_2 is the true proportion of Florida residents have traveled outside the state in the last six months. give a **mathematical ex[ression** for the 99% confidence interval for $p_1 - p_2$.

Solution:

$$\left(\frac{65}{95} - \frac{42}{85}\right) \pm 2.58\sqrt{\frac{65}{95}\frac{30}{95}\frac{1}{95} + \frac{42}{85}\frac{43}{85}\frac{1}{85}}$$

∎

3. A random sample of 200 adults were asked to indicated their agreement or disagreement with the following statement. The government needs to be able to scan Internet messages to be able to prevent fraud and other crimes. They were also asked to identify themselves as high or low users of the Internet. Their responses are given below.

	agree	disagree
low usage	30	50
high usage	40	80

Give a **mathematical expression** to test whether a individual's attitude towards the statement is independent of their usage at level $\alpha = .01$.

Solution:

Under the hypothesis of independence $200(80/200)(70/200) = 28$ is the expected number of low users who agree. In similar fashion we can find the other expected numbers and find that

$$\sum \frac{(O-E)^2}{E} = \frac{4}{28} + \frac{4}{52} + \frac{4}{42} + \frac{4}{78} < 6.64$$

because $qgamma(.99, .5, .5) = 6.634897$ or $P(\chi_1^2 > 6.64) = .01$ and so we would accept the hypothesis of independence.

∎

4.An engineer is interested in the compressive strength of concrete cylinders. He wishes to compare three different methods of the capping treatment and two different types of concrete. This is a 3 by 2 factorial experiment with two replications for each combination. The data is given below along with the usual anova table and a plot of the means for each combination.

```
   method type strength
 1      I    A      613
 2     II    A      648
 3    III    A      585
 4      I    B      612
 5     II    B      575
 6    III    B      629
 7      I    A      603
 8     II    A      638
 9    III    A      595
10      I    B      619
11     II    B      568
12    III    B      621
```

```
             Df Sum of Sq  Mean Sq  F Value     Pr(F)
     method  2    51.167   25.583  0.66450 0.5486796
       type  1   280.333  280.333  7.28139 0.0356516
method:type  2  6113.167 3056.583 79.39177 0.0000483
  Residuals  6   231.000   38.500
```

```
sapply(split(y,interaction(method,type)),mean)
  1.1   2.1   3.1   1.2   2.2   3.2
608.0 643.0 590.0 615.5 571.5 625.0
```

i) At level $\alpha = .01$ test the null hypothesis that there is no interaction against the alternative that interaction is present.

ii) Find a 99% confidence interval for $\mu_{B,II} - (\mu_{B,I} + \mu_{B,III})/2$ where $\mu_{B,II}$ is the true average strength of a cylinder made from concrete B with capping method II.

iii) What conclusions are suggested by the above output?

Solution:

i) From the anova table the p-value for the hypothesis of no interaction is 0.0000483 so we reject the null hypothesis of no interaction.

ii) Note

$$\overline{Y}_{B,II} - (\overline{Y}_{B,I} + \overline{Y}_{B,III})/2 \sim N(\mu_{B,II} - (\mu_{B,I} + \mu_{B,III})/2, \frac{\sigma^2}{2}(1 + 1/4 + 1/4))$$

Since $qt(.995, 6) = 3.71$ we have

$$\frac{575+568}{2} - \frac{(612+619)/2+(629+621)/2}{2} \quad \mp \quad 3.71\sqrt{\frac{38.5}{2}(1+1/4+1/4)}$$

iii) Because of the presence on interaction the test for the main effects do not mean much here. Looking at the plots of the sample means the most striking inference is that method II seems to work much better on concrete of Type A than on type B.

∎

5. A manager was interested in developing a model that would allow her to predict predict the yearly income for her sale persons. The variables were

$Y =$ this years income in thousand of dollars

$X_1 =$ months on the job

$X_2 =$ years of education

$X_3 =$ last years income in thousand of dollars

The regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + Z \tag{1}$$

was fit to the data for twenty sales persons. The results along with three anova tables for the full model and the predict command are given at the end of the exam If a question cannot be answered with the output at hand specify what additional output is needed.

i) Find a 95% confidence interval for the expected value of Y when $X_1 = 54, X_2 = 18$ and $X_3 = 130$.

Solution:

$$138.47 \pm 2.131 \times 5.09$$

since $qt(.975, 15) = 2.13145$ ∎

ii) In the full model at level $\alpha = .05$ test $H : \beta_1 = 0$ against $K : \beta_1 < 0$.

Solution: Since $qt(.05, 15) = -1.75305$ and

$$-1.093 = \frac{-3.1026 - 0}{2.8393} > -1.753$$

we accept H. ∎

iii) In the full model, at level $\alpha = .05$, test $H : \beta_2 = \beta_3 = 0$ against K : At least one is not zero.

Solution: Cannot do; you need an Anova table where $X2$ and $X3$ are the last two independent variables.

∎

iv) Among the three models $Y = \beta_0 + \beta_i X_i + Z$ for $i = 1, 2$ and 3 which will have the largest R^2? Find this value.

Solution: Note from the first anova table

$$SSReg(\beta_1, \beta_2, \beta_3, \beta_4 | \beta_0) = 2580.88 + 1332.23 + 132.46 + 275.52 = 4321.09$$

and so we have

$$TCSS = 4321.09 + 1534.66 = 5855.75$$

Again from the first anova table we have that for the model $Y = \beta_0 + \beta_1 X_1 + Z$ the value of R^2 is

$$\frac{SSReg(\beta_1|\beta_0)}{TCSS} = \frac{2580.88}{5855.75} = .44$$

which is larger than R^2 for the other two models which we can find from the second and third anova tables.

∎

v) In the full model, at level $\alpha = .05$, test $H : \beta_2 = \beta_4 = 0$ against K : At least one is not zero.

Solution: Since

$$\frac{(SSReg(\beta_2|\beta_0, \beta_3, \beta_1) + SSReg(\beta_4|\beta_0, \beta_3, \beta_1, \beta_2))/2}{RSS/15} = \frac{(1267.05 + 275.52)/2}{102.31} = 7.539 > 3.68$$

we reject H because $qf(.95, 2, 15) = 3.68$

∎

vi) In the model $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + Z$, at level $\alpha = .05$, test $H : \beta_3 = 0$ against $K : \beta_3 \neq 0$.

Solution: Recall from part iv) $TCSS = 5855.75$. Let RSS^* be the residual sum of squares for this model. Then from the third anova table we have

$$\begin{aligned} TCSS &= SSReg(\beta_2|\beta_0) + SSReg(\beta_3|\beta_0, \beta_2) + RSS^* \\ 5855.75 &= 1335.51 + 1086.36 + RSS^* \end{aligned}$$

and we have $RSS^* = 3433.88$. Since it will have 20-3-1=16 degrees of freedom we see that

$$\frac{1086.36/1}{3433.88/16} = 5.062$$

and so we reject H because $qf(.95, 1, 16) = 4.45$.

∎

vii) If you could choose just one independent variable from the four that appear in the full model in equation 1 to use to predict Y in a simple linear regression model which one would it be? Why? How useful is this variable in explaining Y in the full model? Briefly explain.

Solution: In the full model the t tests tell us that amoung the four variables the fourth, $X_1 \times X_2$ contributes the most explanation given the other three are in the model however its t-value $= 0.122$ is not particularly small. On the other hand we see from the last anova table that

$$SSReg(\beta_4|\beta_0) = 3997.6$$

and so $R^2 = 3997.6/5855.75 = .68$ for the simple linear regression model with $X_1 \times X_2$ is the largest amoung the four possible one variable models. (See part iv).) Moreover its R^2 of .68 is close to .74 the R^2 value for the full model and so this simple model does almost as well as the full model. It does not appear so important in the full model because X_1 and X_2 are present. ∎

Output for problem 5

```
> Y
 [1] 103 118 121 109  95 129 108 148  98  93  96  95  86 124  90 117 138 108  96
[20] 133
> X1
 [1] 51 48 55 50 50 49 45 54 40 33 43 39 47 50 33 45 57 37 42 47
> X2
 [1] 14 14 13 14 14 16 16 18 18 15 13 19 12 16 15 19 19 15 15 18
> X3
 [1] 133 128 135 115 118 131 114 125 122 117 109 112 109 107 117 110 135 118 128
[20] 126

>prob3.lm <- lm(formula = Y ~ X1 + X2 + X3 + X1 * X2)
>summary(prob3.lm)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9283  -8.3197   0.1343   8.9662  11.8035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.7439   140.3360   1.188    0.253
X1           -3.1026     2.8393  -1.093    0.292
X2          -10.1122     8.5143  -1.188    0.253
X3            0.2852     0.2849   1.001    0.333
X1*X2         0.2924     0.1782   1.641    0.122

Residual standard error: 10.11 on 15 degrees of freedom
Multiple R-Squared: 0.7379,
F-statistic: 10.56 on 4 and 15 degress of freedom, p-value: 0.0002841
```

```
>predict(prob3.lm,data.frame(X1=44,X2=15,X3=120),se.fit=T)
$fit=105.7500, $se.fit=2.488824 and $df=15

> predict(prob3.lm,data.frame(X1=54,X2=18,X3=130),se.fit=T)
$fit=138.4665, $se.fit=5.092558 and $df=15

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 2580.88 2580.88 25.2259 0.0001516 ***
X2         1 1332.23 1332.23 13.0214 0.0025802 **
X3         1  132.46  132.46  1.2947 0.2730410
X1*X2      1  275.52  275.52  2.6929 0.1215888
Residuals 15 1534.66  102.31

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
X3         1 1240.34 1240.34 12.1233 0.003346 **
X1         1 1538.17 1538.17 15.0343 0.001488 **
X2         1 1267.05 1267.05 12.3843 0.003099 **
X1*X2      1  275.52  275.52  2.6929 0.121589
Residuals 15 1534.66  102.31

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
X2         1 1335.51 1335.51 13.0534 0.002557 **
X3         1 1086.36 1086.36 10.6183 0.005290 **
X1         1 1623.70 1623.70 15.8702 0.001198 **
X2*X1      1  275.52  275.52  2.6929 0.121589
Residuals 15 1534.66  102.31

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1*X2      1 3997.6  3997.6 39.0733 1.553e-05 ***
X3         1  161.0   161.0  1.5741    0.2288
X2         1   40.2    40.2  0.3933    0.5400
X1         1  122.2   122.2  1.1941    0.2917
Residuals 15 1534.7   102.3
```