Stat 5021, Final exam, Fall '15 Name _____

61 people took the exam and the high score was 93. There were 2 in the 90's, 9 in the 80's, 15 in the 70's, 8 in the 60's, 19 in the 50's, 4 in the 40's, 1 in the 30's and 3 in the 20's.

You may use one 8" by 11" formula sheet (both sides) but you **may not use** any electronic computing device. You **do not need** to reduce numerical formulas to their simplest form. Your answers may contain R commands.

1. Let $X_1, \ldots, X_{50}$ be iid each normal with a mean of 20 and a variance of 16.. For such a sample let W be the number of observations in the sample which are less than or equal to 22.5.

i) What is the expected value of W? Your answer may contain an R command.

Solution: Let $p = pnorm(22.5, 20, 4)$ then W is a binomial(50,p) random variable and $E(W) = 50 \times p$. ∎

ii) Write a mathematical expression for the probability of the event that $W \leq 40$, that is, at most 40 of the sample will be less than or equal to 22.5.

Solution:

$$P(W \leq 40) = \sum_{w=0}^{40} \binom{50}{w} p^w (1-p)^{n-w}$$

∎

2. Let X be a random variable which takes on the values 1, 2 and 4 with probabilities 5/10, 4/10 and 1/10. Find the variance of X.

Solution:

$$E(X) = \sum_x xp(x) = \frac{1 \times 5 + 2 \times 4 + 4 \times 1}{10} = 1.7$$

$$E(X^2) = \sum_x x^2 p(x) = \frac{1 \times 5 + 4 \times 4 + 16 \times 1}{10} = 3.7$$

Hence $Var(X) = E(X^2) - (E(X))^2 = 3.7 - (1.7)^2$. ∎

3. An experiment was conducted to see if in a certain population of individuals *biofeedback* exercises could help subjects to lower their blood pressure. The blood pressure measurements listed in the table represent readings before and after the biofeedback training of three subjects selected at random from the population. Let μ_B be the true average blood pressure for the population before the biofeedback training and μ_A the true population average after training. Find a 95% confidence interval for $\mu_B - \mu_A$.

Subject	1	2	3
Before	142	175	171
After	121	161	143

Solution: This is paired data. Let $D_i = X_{B,i} - X_{A,i}$ then the observed values for the D_i's are 142-121=21, 175-161=14 and 171-143 = 28. Since $t_{2,.025} = 4.303$ the answer is

$$21 \quad \overline{+} \quad 4.303 \frac{7}{\sqrt{3}}$$

∎

4. Criminologist have long debated whether there is a relationship between weather conditions and the incidence of violent crime. Data from one study are given in the table to the right. Find the value of the test statistics for testing the null hypothesis that the number of crimes is constant across the seasons and explain how its p-value is computed.

Season	Winter	Spring	Summer	Fall
Number of crimes	312	468	481	339

Solution: We use the chi-squared test here where we calculate $T = \sum_{i=1}^{4}(Obs_i - E_i)^2/E_i$. The Obs_i's are given in the table and $E_i = 400$. The p-value is equal to the probability that a chi-squared distribution with 3 degrees of freedom exceeds the computed value of T. ∎

5. Let $X_1, \ldots, X_n$ be i.i.d. each Normal(μ,36). Suppose we are testing $H : \mu = 20$ against $K : \mu > 20$ at level $\alpha = .05$.

i) If $n = 20$ and $\mu = 23$ give a mathematical expression for the probability of making the type II error.

ii) If $\mu = 23$ find a mathematical equation involving n that allows us to determine how large must n be so that the probability of making the type II error is 0.15?

Solution: i) Since $qnorm(.95, 0, 1) = 1.645$ we will reject $H : \mu = 20$ if and only if

$$\frac{\overline{X} - 20}{\sqrt{36/20}} > 1.645 \Longleftrightarrow \overline{X} > 22.207$$

So the probability of the type II error when $\mu = 23$ is

$$P(N(23, 36/20) < 22.207) = pnorm(22.207, 23, \sqrt{36/20}) = 0.2776$$

ii) If $\mu = 23$ how large must n be so that the probability of making the type II error is .15?

We want to find n such that

$$\begin{aligned} 0.15 &= P\big(N(23, 36/n) < 20 + 1.645\frac{6}{\sqrt{n}}\big) \\ &= P\big(N(0,1) < -3\frac{\sqrt{n}}{6} + 1.645\big) \end{aligned}$$

So n must satisfy

$$\begin{aligned} -\frac{\sqrt{n}}{2} + 1.645 &= -1.037 \\ \sqrt{n} &= 2(1.037 + 1.645) \\ n &= 28.77 \end{aligned}$$

and hence $n = 29$ is the answer.

∎

6. A experiment was conducted to study the effect of diet on the time it takes for blood to coagulate. For four different diets four observations of Y, the "blood coagulation time" was collected. The results were

```
diet  1  1  1  1  2  2  2  2  3  3  3  3  4  4  4  4
Y     62 60 63 59 63 67 71 67 68 66 71 67 57 62 60 61
```

The oneway analysis of variance was done on Y using the model $Y_{ij} = \mu + \alpha_i + Z_{ij}$. Recall that $\mu + \alpha_i = \mu_i$ is the average coagulation time for the ith diet. The results were

```
           Df    Sum of Sq      Mean Sq
diet        3       200         66.66667
Residuals 12        70          5.83333
```

i) At level $\alpha = .05$ test H: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ against K: At least one is not zero.

ii) Find a 95% confidence interval for $\mu_1 - \mu_3$.

Solution:

i) Since

$$\frac{66.67}{5.83} > 3.49 = f_{3,12,.05}$$

we reject H.

ii) Since $\hat{\mu}_1 = (62 + 60 + 63 + 59)/4 = 61$ and $\hat{\mu}_3 = 68$ and $t_{12,.025} = 2.179$ the interval is

$$61 - 68 \quad \mp \quad \sqrt{5.83}\sqrt{1/4 + 1/4}\ 2.179$$

∎

7. See the attached sheet with the output for this problem.

i) In the full model what does the regression equation predict for the response for a rat which receives the new substance with the log of the dose level equal to 1.3.

Solution:

$$\hat{y} = 60.00 + 70.48 \times 0 + 45.83 \times 1.3 - 26.8 \times 1$$

∎

ii) In the full model at level $\alpha = .05$ test $H : \beta_3 = 0$ against $K : \beta_3 \neq 0$.

Solution: Since $t_{11,.025} = 2.2$ and since from the output we have $|-1.077| < 2.2$ we accept $H\beta_3 = 0$. ∎

iii) In the full model at level $\alpha = .05$ test $H : \beta_1 = \beta_2$ against $K : \beta_1 \neq \beta_2$.

Solution: Since

$$\frac{(RSS(Reduced) - RSS(Full))/1}{RSS(Full)/11} = \frac{(925.89 - 872.19)/1}{79.29} < 4.84 = f_{1,11;.05}$$

we accept H. ∎

iv) Consider the smaller model. Note that the p-value for the t statistic associated with X is quite small while in the anova table the p-value for the F statistic for X is relatively big. Explain.

Solution: From the plot and the above calculations the most reasonable model seems to be to assume different intercepts but the same slope for the standard and the new. This explains the t-value = .00229 in the smaller model. If you allow for different intercepts X explains a lot. In the anova table for out1 X comes before $X3$ so it is not allowing for different intercepts and its p-value is much larger. On the other hand suppose you assume that both the slopes and intercepts are equal and you fit one simple linear regression to all the data. From the plot you can see that the least squares would be roughly parallel to the x axis. So without the two intercepts in the model X does not explain much. ∎

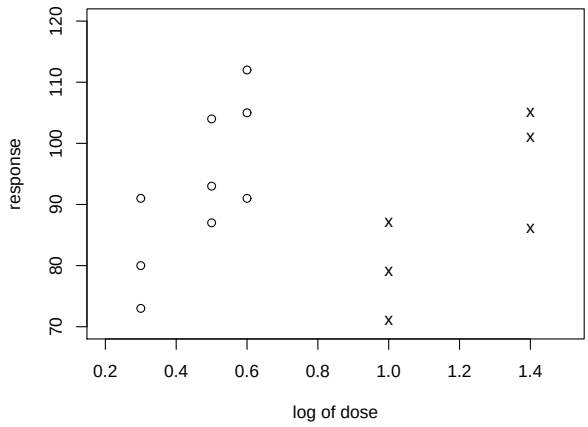
Problem 7. Bioassay

An important problem in biological assay is to compare the potency of a new substance (vitamins, hormones, etc) with that of a standard one. The experimental units, usually animals, are randomly divided into two groups. The first group receives various doses of the standard while the second receives various doses of the new substance. In the regression model it is the log of the dose levels which the independent variables.

Consider an experiment where three groups of rats of three each received 0.3, 0.5 and 0.6 levels of the standard and two groups of three each received 1.0 and 1.4 levels of the new substance and a response Y was measured for each rat. We will first consider a regression model which includes all the data in one model and assumes that in each case the response is a linear function of log(dose) but which allows for different slopes and different intercepts for the standard and new substance. To this end we set

```
X1<-c(.3,.3,.3,.5,.5,.5,.6,.6,.6,0,0,0,0,0,0)
X2<-c(0,0,0,0,0,0,0,0,0,1,1,1,1.4,1.4,1.4)
X3<-c(0,0,0,0,0,0,0,0,0,1,1,1,1,1,1)
Y<-c(73,91,80,87,104,93,105,91,112,79,87,71,101,86,105)
```

Below is a plot of the data where 'o' indicates the response for a standard response and 'x' for a response for the new substance.



First we fitted the model

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + Z$$

The experimenter also wanted to consider the model with $\beta_1 = \beta_2$ so the model

$$Y = \beta_0 + \gamma X + \beta_3 X3 + Z$$

was also fitted where $X = X1 + X2$. The computer output follows.

```
> out_lm(Y~X1+X2+X3)
> summary(out)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    60.00      11.50   5.219 0.000286 ***
X1             70.48      23.80   2.961 0.012943 *
X2             45.83      18.18   2.522 0.028394 *
X3            -26.83      24.92  -1.077 0.304664
---

Residual standard error: 8.904 on 11 degrees of freedom
Multiple R-Squared: 0.5947,
F-statistic:  5.38 on 3 and 11 degrees of freedom, p-value: 0.01591

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 342.90  342.90  4.3246 0.061743 .
X2         1 844.99  844.99 10.6570 0.007538 **
X3         1  91.92   91.92  1.1593 0.304664
Residuals 11 872.19   79.29

> out1_lm(Y~X+X3)
> summary(out1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.263      7.266   9.258 8.19e-07 ***
X             54.912     14.249   3.854  0.00229 **
X3           -44.991     11.429  -3.937  0.00198 **
---

Residual standard error: 8.784 on 12 degrees of freedom
Multiple R-Squared: 0.5698,
F-statistic: 7.946 on 2 and 12 degrees of freedom, p-value: 0.006343

> anova(out1)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
X          1   30.47   30.47  0.3949 0.541531
X3         1 1195.65 1195.65 15.4963 0.001975 **
Residuals 12  925.89   77.16
```