Fifty two people took the exam. The high score was 91. There were 2 in th 90's, 10 in the 80's, 11 in the 70's, 13 in the 60's, 10 in the 50's, 3 in the 40's, 2 in the 30's and 1 in the 20's.

1. Suppose we have three urns, the first contains 2 blue and 5 white balls, the second 7 blue and 4 white and the third 12 blue and 4 white. Consider the experiment where we independently draw one ball at random from each urn.

i) Find the probability that a blue ball selected is from the second urn and white balls from the other two.

ii) Find the probability that at least one of the three balls selected is blue.

Solution:

i)
$$\frac{5}{7}\frac{7}{11}\frac{4}{16}$$

ii)
$$1 - \frac{5}{7}\frac{4}{11}\frac{4}{16}$$

∎

2. Let the joint distribution of X and Y be given in the table below.

		Y			
		1	2	3	4
X	0	.15	.10	.10	.20
	1	.05	.25	.10	.05

Find the conditional probability of the event $Y > 2$ given the event that $X = 0$.

Solution:

$$P(Y > 2 | X = 0) = \frac{P(Y > 0 \text{ and } X = 0)}{P(x = 0)} = \frac{.10 + .20}{.15 + .10 + .10 + .20}$$

∎

3. In a balloon flight near the north pole the numbers of positive and negative particles in cosmic rays were counted and further each particle was categorized as either low energy or high energy.

	positive	negative
low energy	40	70
high energy	20	120

Test whether a particle being positive or negative is independent of its energy level at level $\alpha = .05$.

Solution:

Under the hypothesis of independence the expected number of observations in the low energy and positive cell is $250(110/250)(60/250) = 26.4$. The expected numbers for the other three cells are found in the same way. Then since

$$\sum \frac{(O - E)^2}{E} = \frac{(40 - 26.4)^2}{26.4} + \frac{(70 - 83.6)^2}{83.6} + \frac{(20 - 33.6)^2}{33.6} + \frac{(120 - 106.4)^2}{106.4}$$

is greater than $3.84 = \chi_{1,.05}$ we reject the hypothesis of independence.

∎

4. A plant purifies its liquid wastes and discharges the water into a local river. For four different working days an EPA inspector collected water specimens just down stream from the plant and upstream from the plant. For each sample the inspector measured the amount of a certain pollutant present. The data are given in the table below. Let μ_a and μ_b be the true average amount of pollutant above and below the plant. Find a 99% confidence interval for $\mu_b - \mu_a$. You may assume that the observations are normally distributed.

location	1	2	3	4
below	42	40	52	46
above	36	48	42	38

Solution:
This is paired data so let $X_i = X_{b,i} - X_{a,i}$ and we have as the paired differences 6, -8, 10 and 8. This set of four numbers have a mean of 4 and a sample variance of $200/3$. Since $t_{3,.995} = 5.84$ the answer is

$$4 \mp 5.84 \frac{8.16}{\sqrt{4}}$$

∎

5. Let X be binomial$(30, p)$. Suppose for testing $H : p = 0.25$ against $K : p > 0.25$ we decide to reject the null hypothesis H if and only if $x = 13, 14, \ldots, 30$. For both parts of this question your answer should include both a **mathematical expression** and a **R command**.
i) Find the probability of making the Type I error with this test.
ii) Find the probability of making the Type II error with this test when $p = 0.6$.
Solution:
i)

$$\sum_{x=13}^{30} \binom{n}{x} 0.25^x 0.75^{30-x} = 1 - pbinom(12, 30, .25) = 0.02159364$$

ii)

$$\sum_{x=0}^{12} \binom{n}{x} 0.6^x 0.4^{30-x} = pbinom(11, 30, .6) = 0.008301584$$

∎

6. Consider a metallurgical project that involved the study of the tempering response of a certain grade of steel. Slugs of this steel were preprocessed to reasonably uniform hardness which was measured and recorded. The slugs were then tempered at various temperatures for various lengths of time. The hardness was then measured and the the change in hardness, Y, computed. (Note a negative value for Y means that the tempering process has made the steel harder.)

There were four different lengths of time 5, 50, 150 and 500 minutes and four different temperatures 800, 900, 1000 and 1100 degrees Fahrenheit. There were two independent measurements taken at each of the $4 \times 4 = 16$ possible combinations. First the quadratic model

$$Y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 + \beta_3 (\ln(X_1))^2 + \beta_4 X_2^2 + \beta_5 X_2 \ln(X_1) + Z$$

was fit to the data where X_1 was time and X_2 was temperature.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.932e+01  2.684e+01   1.838 0.077588 .
```

```
ln(X1)         9.363e+00   1.747e+00    5.359 1.31e-05 ***
X2            -1.235e-01   5.586e-02   -2.211 0.036010 *
ln(X1)sq      -5.252e-01   1.230e-01   -4.269 0.000231 ***
X2sq           6.875e-05   2.917e-05    2.357 0.026249 *
ln(X1):X2     -6.533e-03   1.538e-03   -4.247 0.000245 ***
---
Signif. codes:  0 '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 1.65 on 26 degrees of freedom
Multiple R-Squared: 0.8304,      Adjusted R-squared: 0.7978
F-statistic: 25.46 on 5 and 26 degrees of freedom,      p-value: 3.002e-09

Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value      Pr(>F)
ln(X1)      1  68.813  68.813 25.2693 3.131e-05 ***
X2          1 164.025 164.025 60.2332 3.109e-08 ***
ln(X1)sq    1  49.620  49.620 18.2216 0.0002313 ***
X2sq        1  15.125  15.125  5.5542 0.0262487 *
ln(X1):X2   1  49.115  49.115 18.0359 0.0002450 ***
Residuals  26  70.802   2.723
```

   Next the standard two way anova model was fit to these data. Below is the anova table, the individual cell means along with a plot of the cell means.

```
> hard.aov_aov(y~Time*Temp)
> summary(hard.aov)
           Df  Sum Sq Mean Sq F value     Pr(>F)
Time        3 119.250  39.750 11.3571  0.000307 ***
Temp        3 182.750  60.917 17.4048 2.721e-05 ***
Time:Temp   9  59.500   6.611  1.8889  0.127969
Residuals  16  56.000   3.500

> sapply(split(y,interaction(Time,Temp)),mean)
  1.1   2.1   3.1   4.1
 -0.5   3.5   3.0  -1.0

  1.2   2.2   3.2   4.2
 -2.5  -2.0  -1.5  -5.0

  1.3   2.3   3.3   4.3
 -1.0  -2.5  -4.5  -7.5

  1.4   2.4   3.4   4.4
 -1.5  -3.0  -6.0 -10.0
```

3

Use the information provided to answer the following questions. If a question cannot be answer because of insufficient information specify what additional computer output is needed to answer the question.

i) What does the regression model predict for Y when $X_1 = 50$ and $X_2 = 950$?

ii) In the regression model at level $\alpha = .01$ test $H : \beta_1 = 0$ against $K : \beta_1 \neq 0$.

iii) In the regression model at level $\alpha = .01$ test $H : \beta_4 = \beta_5 = 0$ against $K :$ At least one not zero.

iv) In the regression model at level $\alpha = .01$ test $H : \beta_2 = \beta_3 = 0$ against $K :$ At least one not zero.

v) Let $\mu_{5,1000}$ be the average response of Y when $X_1 = 5$ and $X_2 = 1000$. Find a 99% confidence interval for $\mu_{5,1000}$.

vi) Find a 99% confidence interval for $\mu_{5,1000} - \mu_{500,1000}$.

vii) Find the lack of fit sum of squares for the regression model.

viii) Briefly summarize what is to be learned from this experiment. In particular how is the hardness of steel affected by the length of time and temperature of the tempering process.

Solution:

i)

$$\hat{y} = 4.932 + 9.363 \times \ln(50) - .1235 \times 950 - .5252 \times (\ln(50))^2$$
$$+ .00006875 \times (950)^2 - .006533 \times \ln(50) \times 950$$

ii) From the output we see the t-value for $\hat{\beta}_1$ is 5.359 which is greater that $t_{26,.005} = 2.779$ so we reject H.

iii) Note $f_{2,26;.01} = 5.53$. Then from the anova table we have that

$$\frac{(15.125 + 49.115)/2}{70.802/26} > 5.53$$

and so we reject H.

iv) Cannot do. We need an anova table where these two variables come last.

v) Since $t_{16,.01} = 2.92$ the interval is

$$-1 \mp 2.92 \times \frac{\sqrt{3.5}}{\sqrt{2}}$$

Note in principle you could also use the quadratic regression model to get this interval although I did not give you the necessary output. However this should be better since this model should give us a better estimate of the error variance.

vi)
$$(-1 - (-7.5)) \mp 2.92 \times \sqrt{3.5} \times \sqrt{1/2 + 1/2}$$

vii) This will be the residual sum of squares from the regression model minus the within sum of squares (or the residual sum of squares) from the two way anova model, that is, $70.802 - 56 = 14.802$.

viii) From the second anova table we see that both factors Time and Temp are significant. Note that there is some interaction because the first two or lower temperatures appear to act differently than the two higher temperatures. This is also reflected in the anova table where the p-value of the F value for the interaction term is highly significant. To get the hardest steel we need to cure the slugs at the highest temperature and the longer we temper them the harder the steel seems to get. At the two lower temperatures the steel appears to start to harden if it is tempered long enough.

■