METHODS QUALIFYING EXAM

August 2008

INSTRUCTIONS:

a.) DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.

b.) Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.

c.) Use only one side of each sheet of paper.

d.) Answer all the questions.

e.) Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.

f.) Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.

g.) There are 4 Problems in this exam.

Problem I.

For the following two experiments, provide the following information:

a. Type of Randomization, eg, CR, RB, LS Split Plot, etc;

b. Type of Treatment Structure, eg, crossed, nested, etc;

c. Identify each of the factors as being fixed or random;

d. Identify any covariates;

e. Describe the measurement units and the experimental units.

f. An ANOVA Table, Including : Sources of variation and Degrees of freedom.

Experiment 1: A laboratory study of the stress (psi) of titanium is to be designed involving laboratories in the United States and Germany. Three laboratories are randomly selected from the many laboratories within each of the two countries. Two temperatures(100, 200^0F), and four strain rates (1, 10, 100, 1000 sec^{-1}) are to be investigated. Two titanium specimens are randomly assigned to each lab-temperature-strain combination and a stress reading is made on each specimen.

Experiment 2: A textile specialist is investigating the variability in length of wool fibers from 4 breeds of sheep during each of the 2 harvesting seasons. For each breed, 8 ranches were randomly selected from a listing of ranches raising that breed of sheep. During each of the 2 harvesting seasons, a random sample of 5 sheep was selected at each of these ranches. The age of each of each sheep at the beginning of the study was recorded since the sheep ranged in age from 2-15 years. On each selected sheep, the wool length was determined at 4 randomly selected sites.

Problem II.

Sometimes a regression of Y on x can be reasonably represented by two intersecting straight lines, one being appropriate when $x \leq \gamma$ and the other when $x \geq \gamma$; thus

$$E[Y] = \alpha_1 + \beta_1 x \quad \text{when} \quad x \leq \gamma$$

$$E[Y] = \alpha_2 + \beta_2 x \quad \text{when} \quad x \geq \gamma$$

and

$$\alpha_1 + \beta_1 \gamma = \alpha_2 + \beta_2 \gamma = \theta.$$

For example, x may be an increasing function of time and at time t_c a treatment is applied that may possibly affect the slope of the regression line either immediately or after a time lag. We call $x = \gamma$ *the changeover point* and θ the *changeover value.*

1. Suppose we wish to fit the two-phase model

$$\begin{aligned} Y_{1i} &= \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \quad \text{for} \quad i = 1, 2, \ldots, n_1 \\ Y_{2i} &= \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \quad \text{for} \quad i = 1, 2, \ldots, n_2 \end{aligned}$$

by least squares, where

$$\alpha_1 + \beta_1 \gamma = \alpha_2 + \beta_2 \gamma = \theta$$

and

$$x_{11} < x_{12} < \cdots < x_{1n_1} < \gamma < x_{21} < x_{22} < \cdots < x_{2n_2},$$

with γ known. Derive a set of linear equations whose solution yields the least squares estimators of the unknown parameters. [Note: You do not need to explicitly solve the equations.]

2. Discuss briefly the complications that occur if γ is unknown. [Note: You can answer this part even if you are unsuccessful answering part 1.]

Problem III.

The Storm Prediction Center (an agency of NOAA) tracks the number and characteristics of tornadoes. In this problem we will consider primarily the variable **Killer_tornadoes**, which is the number of tornadoes with one or more deaths in a year. The summary statistics were found using `proc means`:

```
                    The MEANS Procedure

        Analysis Variable : Killer_tornadoes

  N          Mean         Std Dev        Minimum        Maximum

 48     2.0416667       2.9532408              0     13.0000000
```

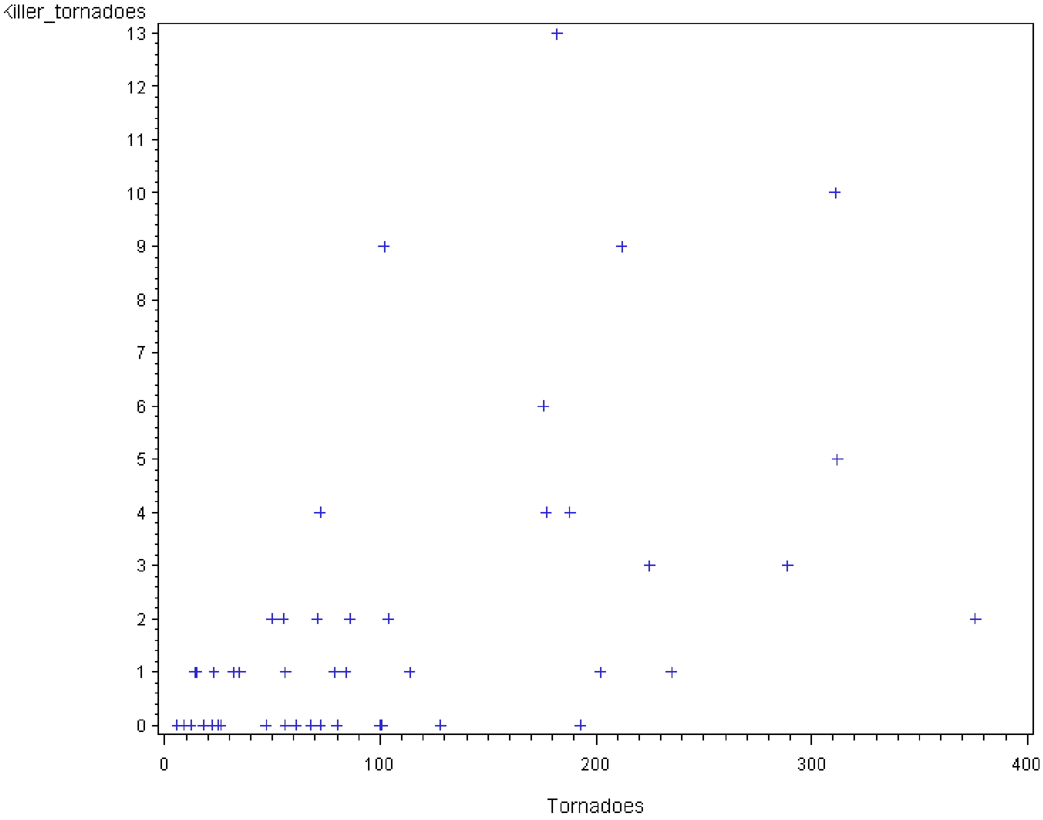
1. Explain why $\bar{X} \pm Z_{\alpha/2}\sqrt{\bar{X}/n}$ is a reasonable approximate $1-\alpha$ confidence set for λ, the mean of a Poisson distribution. Then assuming that the variable `Killer_tornadoes` has a Poisson distribution, obtain an approximate 95% confidence interval for the mean number of killer tornadoes.

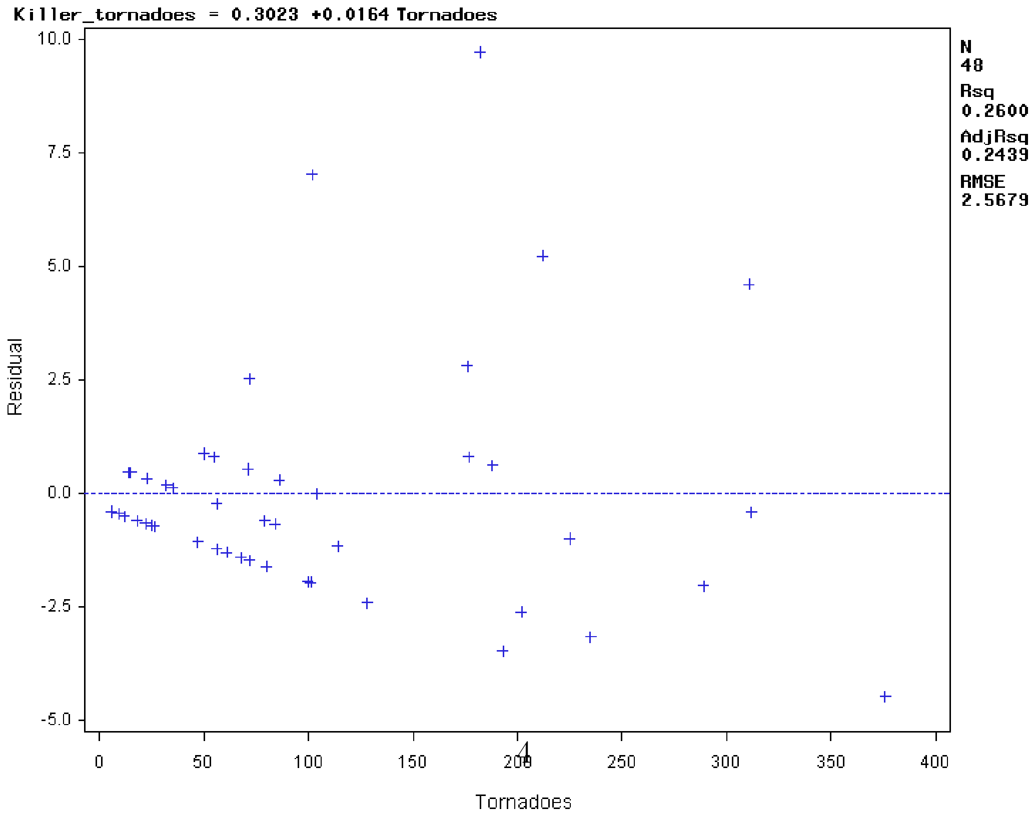
2. Is the assumption of a Poisson distribution reasonable for the variable `Killer_tornadoes`? Explain why or why not based only on the summary statistics above.

3. A chi-squared goodness-of-fit test was carried out by using the Poisson distribution with the estimated mean $\hat{\mu} = 2.042$ to specify the cell probabilities for $x = 0, 1, 2, 3, \geq 4$. The test statistic was computed to be $X^2 = 32.35$. What does this tell you about the assumption of the Poisson distribution for the variable `Killer_tornadoes`? Explain your reasoning.

4. Suppose that the researcher decides to predict the number of killer tornadoes using `Tornadoes`, the number of tornadoes in a year, as a predictor. A simple linear regression model was fit with `Killer_tornadoes` as the response and `Tornadoes` as the predictor. Some SAS output is given below along with a scatter plot of the data and a plot of residuals versus the predictor. Discuss the appropriateness of using this prediction model for the number of killer tornadoes.

```
                          Analysis of Variance

                                 Sum of           Mean
Source                 DF       Squares         Square    F Value    Pr > F

Model                   1     106.59133      106.59133      16.16    0.0002
Error                  46     303.32533        6.59403
Corrected Total        47     409.91667


             Root MSE              2.56788    R-Square     0.2600
             Dependent Mean        2.04167    Adj R-Sq     0.2439
             Coeff Var           125.77392


                         Parameter Estimates

                         Parameter       Standard
 Variable        DF       Estimate          Error    t Value    Pr > |t|

 Intercept        1        0.30234        0.56967       0.53      0.5982
 Tornadoes        1        0.01641        0.00408       4.02      0.0002
```


Killer_tornadoes
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
100
200
300
400
Tornadoes


Killer_tornadoes = 0.3023 +0.0164 Tornadoes
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
Residual
0
50
100
150
200
250
300
350
400
Tornadoes
N
48
Rsq
0.2600
AdjRsq
0.2439
RMSE
2.5679

Problem IV.

Consider regression though the origin (i.e., straight line regression with population intercept known to be zero) with a predictor x_i which only takes positive values and is such that $Var(e_i|x_i) = x_i\sigma^2$. The corresponding regression model is

$$Y_i = \beta x_i + e_i \qquad (i = 1, \ldots, n)$$

1. Find an explicit expression for the weighted least squares estimator of β.
2. Show that the weighted least squares estimator of β is unbiased.
3. Find an explicit expression for the variance of the weighted least squares estimator of β.

Some Chi-Squared Percentiles

	Right-Tail Probability			
df	0.100	0.050	0.025	0.010
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21

Some Normal Percentiles

Right-Tail Probability			
0.100	0.050	0.025	0.010
1.282	1.645	1.960	2.326