

## **Methods Qualifying Exam**

**January 2000**

### **Instructions:**

- 1)** 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.
- 2)** Please start your answer to each question on a separate sheet of paper.
- 3)** Answer all the questions.
- 4)** 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.
- 5)** 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.

## PROBLEM #1:

The tensile strength of a material is the ability that the material possesses to resist deformation when a force or a load is applied to it. A metallurgist conducts a study to evaluate the tensile strength of ductile iron strengthened at two different temperatures. She thinks that the lower temperature will yield the higher mean tensile strength. At each of the two temperatures,  $800^{\circ}C$  and  $1000^{\circ}C$ , 300 specimens of ductile iron were heat treated. The data consists of the tensile strengths from 300 specimens heated to  $800^{\circ}C$ :  $X_1, \dots, X_{300}$  which are iid with mean  $\mu_1$  and standard deviation  $\sigma_1$  and the tensile strengths from 300 specimens heated to  $1000^{\circ}C$ :  $Y_1, \dots, Y_{300}$  which are iid with mean  $\mu_2$  and standard deviation  $\sigma_2$ . Furthermore, the  $X$ 's and  $Y$ 's are independent.

- a. The metallurgist is interested in the null hypothesis  $H_0 : \mu_1 \leq \mu_2$  versus the alternative hypothesis  $H_1 : \mu_1 > \mu_2$ . Use the following steps to present the customary  $t$ -test of this null hypothesis based on  $X_1, \dots, X_{300}$  and  $Y_1, \dots, Y_{300}$ .
  1. Write down a general formula for the  $t$  test statistic commonly used for this hypothesis test.
  2. Write down the decision rule for this hypothesis test. Use  $\alpha = 0.05$ .
  3. State the necessary conditions needed for your procedure to be valid and how you would verify whether the conditions are satisfied in this experimental setting.
- b. For parts (b) and (c) of this question, you may assume that  $\sigma_1 = \sigma_2 = 1$  and that the sample sizes are large enough to invoke the central limit theorem if necessary.
  1. Calculate the power of your test for the following six values of the parameter:

$$\Delta = \frac{\mu_1 - \mu_2}{\sqrt{1/300 + 1/300}} = .5, 1.0, 1.5, 2.0, 2.5, 3$$

2. Use your results from (b.1) to sketch a power curve for your test. Be sure to label your axes clearly.
- c. The metallurgist in discussing your results from (a) and (b) states, "The power of the test when  $\Delta = 2.0$  is not large enough to meet industry standards. What needs to be done to increase it?" Answer the metallurgist's question, paying careful attention to: (i) your specific recommendation on how to increase the power; and (ii) explanation (based on the ideas from parts (a) and (b)) of **why** your recommendation will result in an increase in power.

- d. The 600 observations considered above represent the tensile strength obtained from the two levels of heat treatment. However, after the experiments were conducted, the metallurgist informs you that the heat treatment for the 300 specimens for each heat level were conducted in the following manner. The furnace used to heat treat the specimens could hold only 5 specimens at a time. Thus, a tray containing 5 randomly selected specimens was heated to the specified temperature for the prescribed length of time and then the tensile strength measurements were taken on the 5 specimens. The metallurgist states that there is some variation in the conditions within the furnace from one experimental run to the next. Thus, there may be a strong *positive* correlation between tensile strength readings for specimens on the same tray. Given this additional information, answer the following questions without carrying out any additional calculations.
1. How will this positive correlation within specimens affect the expectation of the variance estimator you used in part (a.1)?
  2. Suppose you did not adjust for the positive correlation within specimens and proceeded to use the ordinary *t*-test you proposed in part (a). Will the positive correlation in the data increase or decrease the numerical values of power you calculated for the test statistic in part (b)? Explain.
- e. In light of your answer to (d), the metallurgist states, “Using the *t*-test from (a) to test the research hypothesis is obviously flawed. What is an alternative approach to testing the research hypothesis?” Answer the metallurgist’s question by presenting a standard testing method that will account appropriately for the sampling design described in (d). Be sure to give clear, explicit statements of both your test statistic formula and your decision rule.

## PROBLEM #2:

Consider the usual (full rank) linear regression model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{Y}(nx1)$  is a vector of observed response variables,  $\mathbf{X}(nxp)$  is a matrix of fixed and observed explanatory variables and whose first column consists of ones,  $\boldsymbol{\beta}(px1)$  is a vector of unknown parameters and  $\boldsymbol{\epsilon}(nx1)$  is a vector of unobservable random variables assumed to have mean  $\mathbf{0}$  and variance-covariance matrix  $\sigma^2\mathbf{I}$ . Let  $\hat{\boldsymbol{\beta}}$  and  $s^2$  be the usual least squares estimators of  $\boldsymbol{\beta}$  and  $\sigma^2$ , respectively.

Suppose that  $\mathbf{x}_{n+1}^T = (1 \ x_{n+1,1} \ \dots \ x_{n+1,p-1})$  is a  $1 \times p$  row vector of explanatory variables for which no response variable has been observed. Define augmented matrices  $\mathbf{Y}_a((n+1)x1)$  and  $\mathbf{X}_a((n+1)x(p+1))$  as follows:

$$\mathbf{Y}_a = \begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} \quad \text{and} \quad X_a = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{x}_{n+1}^T & -1 \end{bmatrix}.$$

- (a) Prove that the  $(p+1)^{st}$  element of  $(\mathbf{X}_a^T \mathbf{X}_a)^{-1} \mathbf{X}_a^T \mathbf{Y}_a$  is equal to  $\hat{Y}_{x_{n+1}} \equiv \mathbf{x}_{n+1}^T \hat{\boldsymbol{\beta}}$ .

[Hint: The least squares criterion can be used to prove this result.]

- (b) Show the following result for partitioned matrices:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T) & (-\mathbf{F}\mathbf{E}^{-1}) \\ (-\mathbf{E}^{-1}\mathbf{F}^T) & (\mathbf{E}^{-1}) \end{bmatrix},$$

where  $\mathbf{E} = \mathbf{D} - \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}$  and  $\mathbf{F} = \mathbf{A}^{-1} \mathbf{B}$ .

- c) Use the matrix result of part b) to prove that the  $(p+1)^{st}$  diagonal element of the matrix  $s^2(\mathbf{X}_a^T \mathbf{X}_a)^{-1}$  is equal to the least squares estimate of the variance of the prediction error of  $\hat{Y}_{x_{n+1}}$ .

### **PROBLEM #3:**

A company that provides data input for billing operations is considering replacing the usual desktop computer/monitor/keyboard setup with notebook computers. Before doing so, they want to conduct a study to determine if the change will affect production (measured in number of entries per unit of time and in error rate=number of errors/total number of entries). They have asked you to help design a study to compare six possible configurations:

- T1. Usual desktop configuration with external keyboard, mouse, and monitor
  - T2. A notebook computer with external keyboard and mouse
  - T3. A notebook computer on raised blocks with external keyboard and mouse
  - T4. A notebook computer with external cordless keyboard and mouse
  - T5. A notebook computer on raised blocks with external cordless keyboard and mouse
  - T6. A notebook computer alone
- a. There are a large number of employees that are available for the study. However, the company would like to use as few as possible because while the employees are involved in the study they will be unavailable for their usual duties.
- 1. What information do you need in order to provide an estimate of the number of employees that should be used for the study?
  - 2. What type of design do you suggest? Describe briefly the randomization procedure that you would use. Outline the ANOVA for this design showing sources of variation and degrees of freedom. (Assuming the number of employees used in the study= $r$ ).
  - 3. Would you suggest inference procedures other than the usual ANOVA? If so, which procedures?
- b. Suppose it is not possible for each employee to be tested for more than four different configurations (treatments). What would you suggest? Does this place any restrictions on the number of employees used for the study? If so, what are the restrictions?
- c. The employees have a wide range of experience, ranging from those who have been performing the duties for seven years to those who have been on the job for only one year. It is thought there may be a relationship between the measures of production and the time an individual has been performing the duty. In addition, the typing skills (a measure of ability to use a keyboard) differs among the employees. It is possible to give each employee a standardized typing test to obtain a measure of his or her typing skill.
- Does this information change your suggested design in Part a.2? If so, how?

**PROBLEM #4:**

Let  $Y$  have a double exponential distribution, that is,  $Y$  has pdf and cdf as follows:

$$f(y) = \frac{1}{2\sigma} e^{-|y-\mu|/\sigma}, \quad F(y) = \begin{cases} \frac{1}{2}e^{(y-\mu)/\sigma}, & \text{for } y < \mu; \\ 1 - \frac{1}{2}e^{-(y-\mu)/\sigma}, & \text{for } y \geq \mu. \end{cases}$$

- a. Show that  $\mu$  and  $\sigma$  are location-scale parameters for the above cdf.
- b. If  $Y_1, \dots, Y_n$  are iid with cdf  $G(\cdot)$ , describe in detail a graphical procedure for evaluating whether  $G(\cdot)$  is a double exponential cdf.
- c. If  $Y_1, \dots, Y_n$  are iid with cdf  $G(\cdot)$ , describe in detail a statistical test of the hypothesis that  $G(\cdot)$  has a double exponential cdf.

## **Methods Qualifying Exam**

**August 2000**

### **Instructions:**

- 1)** 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.
- 2)** Please start your answer to each question on a separate sheet of paper.
- 3)** Answer all the questions.
- 4)** 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.
- 5)** 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.

## PROBLEM #1:

Weight gain in the first 3 months after birth is important for new born infants. A pediatrician wishes to test a new feeding formula to determine if it will cause greater weight gain in new born infants than the standard formula.

From her records she finds that the first 3 months weight gains of single birth infants on the standard formula have the following characteristics:

$$\mu_S = 15\text{oz.} \quad \text{and} \quad \sigma_S = 6\text{oz.}$$

On the other hand, the first 3 months weight gains of identical twins on the standard formula have the characteristics:

$$\mu_T = 12\text{oz.}, \quad \sigma_T = 6\text{oz.} \quad \text{variation between sets of twins, and}$$

$$\rho = .8 \quad (\text{correlation in weight gain of identical twins})$$

She wants to run a 3 month experiment on a group of infants to test the new formula versus the old formula. She has decided to use a 5% probability of Type I error, and wishes to be able to detect a 3 oz. increase in weight gain with 90% probability.

- (a.) What sample size must she use if she does the experiment with a CRD of single birth infants? State any assumptions you are making in this calculation.
- (b.) What sample size must she use if she does the experiment with a RCBD with pairs of identical infants? State any assumptions you are making in this calculation.
- (c.) Discuss the relative merits of the two experiments in terms of practicality and of her basic goal.

## PROBLEM #2:

I. Consider the model

$$\text{Model (1)} : Y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_{ij}$$

for  $i = 1, \dots, n$  and  $j = 1, 2$  where we have 2 replicates at each vector  $(1, X_i)$  of independent variables. Assume that  $\epsilon_{ij}$  are uncorrelated random variables with mean zero and variance  $\sigma^2$ . Consider also the model

$$\text{Model (2)} : \bar{Y}_{i\cdot} = \beta_0 + X_i \beta_1 + \bar{\epsilon}_i.$$

where  $\bar{Y}_{i\cdot}$  is the mean of the 2 replicates and  $\bar{\epsilon}_i$  is the corresponding error.

- (a.) Find the ordinary least squares estimator for  $(\beta_0, \beta_1)$  in Model (1) and compare it with the weighted least squares estimator of  $(\beta_0, \beta_1)$  in Model (2).
- (b.) Find a transformation in Model (2) [rescaled model from model (2)] such that the ordinary least squares estimates of  $(\beta_0, \beta_1)$  in this rescaled model coincide with the estimates you got from part (a.).

II. Let  $\theta$  ( $0 \leq \theta \leq 1$ ) denote the probability of pain at dose  $x$  of a certain drug. One model for  $\theta$  is to take

$$\theta = \int_{-\infty}^x f(u) du$$

where  $f(u)$  is a probability density function. If  $f(u)$  represents the extreme value distribution

$$f(u) = \beta \exp[(\alpha + \beta u) - \exp(\alpha + \beta u)], \quad -\infty < u < \infty.$$

- (a.) Find  $\log\{-\log(1 - \theta)\}$ .

- (b.) Let  $m_i$ ,  $i = 1, 2, \dots, n$  denote the number of patients exposed at dose level  $x_i$  and  $y_i$  denote the number of patients relieved from pain at that dose level. Show that for the logistic link

$$\theta = \frac{e^{\alpha+\beta x}}{(1 + e^{\alpha+\beta x})}$$

the maximum likelihood equations for estimating  $\alpha$  and  $\beta$  can be written as

$$\begin{aligned} \sum_{i=1}^n (y_i - m_i \hat{\theta}_i) &= 0, \\ \sum_{i=1}^n (y_i - m_i \hat{\theta}_i) x_i &= 0 \end{aligned}$$

where

$$\hat{\theta}_i = \frac{e^{\hat{\alpha}+\hat{\beta}x}}{(1 + e^{\hat{\alpha}+\hat{\beta}x})}.$$

### **PROBLEM #3:**

I. There are many private chemical laboratories in Texas that perform analyses of food products. One of the tests is the determination of the amount of cadmium in fish oil. Each laboratory has its own procedure for the assay, and the state regulatory agency wants to determine if a standard method would be more accurate.

To determine the adequacy of the lab procedures, the agency took a homogeneous supply of the fish oil (ten gallons), divided it into 2 five gallon portions, and enhanced one of the five gallon portions with additional cadmium. Fifteen laboratories were randomly selected from a list of hundreds of qualified laboratories, and each was sent a sample of the enhanced and unenhanced oils. Each lab was to split each sample into two portions and analyze the portions using their own method and the new, standard method. Therefore, each lab was to perform four analyses.

- (a.) Write the model for the data of this experiment.
- (b.) Determine the expected mean squares for the components of the AOV.
- (c.) The interaction between the lab and the analysis method could be important in the detection of whether the old methods that the labs were using were giving equivalent results. Can this design detect such an interaction? If not, how would you modify the design so that it could detect the interaction?

II. After conducting the F-tests for main effects and interactions in a CR factorial experiment involving two factors A and B, the experimenter wanted to further investigate the experimental data. For each of the following questions given on the next page, select a **ONE** letter from the list at the bottom of the page which is the best solution to each of the following situations. **Place your selection** in the space to the **left** of each situation.

**SITUATION:**

- ..... 1. The A\*B interaction is significant. The experimenter wants to compare the levels of factor A, which are unequally spaced numerical values.
- ..... 2. The A\*B interaction is not significant. The researcher selects several contrasts in the levels of a qualitative factor A based on the information observed in the experimental data and then wants to test whether the contrasts are significant.
- ..... 3. The A\*B interaction is not significant. The levels of both A and B are fixed qualitative levels. The experimenter wants to compare the levels of factor A.
- ..... 4. The A\*B interaction is significant. The levels of A are randomly selected, levels of B are fixed qualitative levels. The experimenter wants to perform a mean separation on the levels of A.
- ..... 5. The A\*B interaction is not significant. The levels of A are equally spaced numerical values but the levels of B are fixed and qualitative. The experimenter wants to evaluate trends in the levels of factor A.

**TECHNIQUE:**

- A. Either LSD or Tukey's multiple comparison procedure at each level of factor B.
- B. Either LSD or Tukey's multiple comparison procedure averaged over the levels of factor B.
- C. Fit orthogonal polynomial contrasts at each level of B.
- D. Fit orthogonal polynomial contrasts averaged over the levels of B.
- E. Nothing new is learned beyond the results of the F-tests from the AOV table.
- F. Comparison of means is not appropriate.
- G. Scheffe's technique
- H. Bonferroni F test for contrasts
- I. None of the above are appropriate

## **PROBLEM #4:**

A graduate student has approached you for help in designing her experiment and analyzing the data from the experiment. She has five “treatments” she wants to compare. The treatments are different combinations of setups for desktop computers. For each treatment she will measure a number of different variables, e.g., keying productivity measured in words per minute. Her experimental subjects are employees of a data entry firm. She believes there may be a difference in the response variables for male and female subjects, so will use some of each. She also believes there is a difference in the response variables between younger employees (19-25 years old) and older employees (more than 25 years old) so will use subjects from each of the two age groups. Each subject may be asked to perform the task more than once, i.e., with each of the treatments.

- (a.) Suggest an appropriate design for her experiment and give the sources of variation and degrees of freedom for the appropriate ANOVA. You may use a general symbol (e.g.,  $r$ ) for the number of employees or number of employees within a group.
- (b.) It is believed that the order in which an employee performs the task (i.e., the order of the treatments for an employee) may have an effect on the response. Would you modify the design you suggested in part a? If so, what would you suggest? Does this place any restriction on the number of employees used? Give the sources of variation and degrees of freedom for the appropriate ANOVA.
- (c.) Suppose the treatments are:
  - A. Desktop PC with 15” monitor with external keyboard and mouse
  - B. Desktop PC with 17” monitor with external keyboard and mouse
  - C. Notebook computer alone
  - D. Notebook computer with monitor blocks and external keyboard and mouse
  - E. Notebook computer with “Rock Solid” stand and external keyboard and mouse

Would you suggest any analysis (hypotheses tests) other than the basic ANOVA? If so, what analyses?

## **PROBLEM #5:**

For the following two experiments, provide the following information:

1. Type of Randomization, eg, CR, RB, LS, Split-plot, etc;
  2. Type of Treatment Structure, eg, single factor, crossed, nested, etc;
  3. Identify each of the factors as being fixed or random;
  4. Describe the experimental units.
  5. An ANOVA Table, Including : Sources of variation, Degrees of freedom, Expected mean squares, Denominator of the F-statistic for testing each relevant effect.
- A. A plant scientist wants to investigate plant uptake of nickel under four rates of sludge application. Three varieties of sweet corn were to be studied at each of the four rates of sludge application. For each of the varieties of corn, a total of 40 plants were planted in individual pots. After the plants were established, 10 pots of each variety were randomly selected to receive a rate of sludge. After a period of time, three leaves were randomly selected from each plant and the amount of nickel present in the leaf was determined. In addition to differences in uptake amount due to the different rates of sludge across the three varieties, there was interest in variation among plants treated alike as well as variation among leaves on the same plant.
- B. An industrial engineer is investigating the effect of four assembly methods on the assembly time for a color television component. Four operators are randomly selected for the study. Furthermore, the engineer knows that each assembly method produces such fatigue that the time required for the last assembly may be greater than the time required for the first, regardless of the method used in the assembly process. That is, a trend develops in the required assembly time. Each of the four operators uses all four assembly methods, however the order in which the methods of assembly were applied was different for each of the four operators.

## **Methods Qualifying Exam**

**January 2001**

### **Instructions:**

1. 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.
2. Please start your answer to each question on a separate sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## PROBLEM #1

Consider a set of observations  $Y_i, i = 1, 2, \dots, 400$ , which are assumed to be independent and identically distributed with a mean  $\mu$  and variance  $\sigma^2$ . NOTE that due to the large sample size involved here, you may use normal-distribution tables (distributed with this examination) for any probability calculations or decision rules required below.

1. Consider the null hypothesis  $H_0 : \mu = 12$  and the two-sided alternative hypothesis  $H_1 : \mu \neq 12$ . Use the following steps to present the customary  $t$ -test of this null hypothesis based on  $Y_1, \dots, Y_{400}$  and  $\alpha = 0.05$ .
  - (a) Write down a general formula for the  $t$  test statistic commonly used for this hypothesis test.
  - (b) Write down the decision rule for this hypothesis test. Use  $\alpha = 0.05$ .
2. In the context of the hypothesis test presented in (1.), give clear, explicit definitions of the following terms.
  - (a) Type I error
  - (b) Type II error
  - (c) Power of the test.
3. For parts (3.) and (4.) of this question, you may assume that  $\sigma$  is known and  $\sigma = 1$ .
  - (a) Calculate the power of your test for the following six values of the true parameter  $\mu$ : 11.9, 11.95, 11.975, 12.025, 12.05, 12.1.
  - (b) Use your results from (3.(a)) to sketch a power curve for your test. Be sure to label your axes clearly.
4. An agronomist reviews your work from (1.) through (3.) and objects, “The power you have for  $\mu = 12.025$  is lower than I was hoping to get. How can I increase it?” Answer your agronomist’s question, paying careful attention to:
  - (a) Your specific recommendation on how to increase the power; and
  - (b) Explanation (based on the ideas from parts (1.) through (3.)) of *why* your recommendation will result in an increase in power. Hint: Use the formula for the power function in answering this part of the question.
5. The 400 observations considered above represent the weight (in grams) of pecans (a type of nut). However, (unknown to you) the 400 pecans were actually collected from 10 trees, with 40 pecans picked from each tree. Also, your agronomist admits that within a given tree, pecan weights cannot be considered independent, and will have a strong *positive* correlation, due to common genetic and environmental factors. Given this additional information, answer the following questions without carrying out additional calculations.
  - (a) How will this positive correlation within trees affect the expectation of the variance estimator of the sample mean that you used in part (1.(a))?
  - (b) Suppose you ignored the positive correlation within trees and proceeded to use the  $t$ -test you proposed in part (1.). Will the actual values of the power of the  $t$ -test be larger or smaller than the values you calculated in part (3.)? Explain.
6. In light of your answer to part (5.), your agronomist says, “OK, I see that it’s wrong to use the  $t$ -test from (1.) to test our null hypothesis. What should I do instead?” Answer your agronomist’s question by presenting a standard testing method that will account appropriately for the nested design described in (5.). Be sure to give clear, explicit statements of both your test statistic formula and your decision rule.

## PROBLEM #2

Consider the simple first-order autocorrelated error regression model

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \dots, n,$$

where

$$\epsilon_i = \rho\epsilon_{i-1} + \delta_i, \quad |\rho| < 1$$

and  $\delta_i$ 's are i.i.d  $(0, \sigma^2)$  random variables for  $i \in \{\dots, -2, -1, 0, 1, 2, \dots\}$ .

- (i) Show that  $\epsilon_i = \rho^m \epsilon_{i-m} + \sum_{j=0}^{m-1} \rho^j \delta_{i-j}$  for any finite integer  $m$ .  
(You can use induction principle to prove it).
- (ii) Heuristically we see that  $\epsilon_i \rightarrow \sum_{j=0}^{\infty} \rho^j \delta_{i-j}$  in some sense as  $m \rightarrow \infty$  in (i). Then

$$\epsilon_i = \sum_{j=0}^{\infty} \rho^j \delta_{i-j} \tag{1}$$

with probability 1. Assuming that the interchange of the expectation and infinite summation operator is justified , use (1) to determine

- (a)  $E(\epsilon_i)$ ,
- (b)  $\text{var}(\epsilon_i)$ ,
- (c)  $\text{Covariance}(\epsilon_i, \epsilon_j)$ .
- (iii) Use the previous results to obtain the generalized least-squares estimate of  $\beta_0$  and its variance (you can assume  $n = 2$  for this problem).

Question 3.

A researcher has approached you for help with designing her study. The study is to compare measurements made on individuals while seated in an office chair. The plan is to have different individuals, randomly selected from a population of individuals, sit in the chairs. There are a number of response variables, the principal one being a measure of pressure on the chair seat. The researcher may use any number of individuals, but would like to keep the number reasonable. In each of the situations described below, individuals will be asked to do multiple "sittings," each under different conditions. It is felt that a reasonable limit on the number of sittings per individual is in the range 6 to 10.

- a. There are fifteen (15) different chairs to be compared. A simple experimental design would be to have each individual sit in each chair. However, this would mean that each individual will have to do 15 sittings. One approach would be to consider the individuals as "blocks" and the chairs as "treatments, and use a Balanced Incomplete Block Design. Balanced Incomplete Block Designs have the parameters  $t$ ,  $b$ ,  $r$ ,  $k$ , and  $\lambda$ . Define each of these parameters and give the values for a suitable BIBD. How many individuals are needed? Discuss briefly the randomization process that you would use.
- b. Unfortunately, the researcher changes her mind. She now wants to compare only eight different chairs. However, for each chair she wants to compare the chair with and without armrests (Factor A) and for two different backrest angles (Factor B). Hence, there are now 32 different "treatments" (8 chairs x 2 levels of A x 2 levels of B). Again, it is believed that for each individual to have 32 sittings is excessive. Again, considering individuals as blocks, suggest an appropriate design. (Hint: You can think of the 8 chairs as corresponding to combinations of factors C, D, and E, each at 2 levels.) Although a comparison of the chairs is of interest, the major interest is in the main effects for armrests (Factor A) and backrest angle (Factor B) and their possible interaction. How many individuals are needed? Show at least a portion of the design and discuss briefly the randomization process you would use. Remember, no individual should have more than ten sittings.
- c. Still another change. There are now only six chairs, but the number of backrest angles is changed to three levels (factor B has 3 levels). Each chair will still be used with and without armrests (factor A still has 2 levels). Hence there are 6 "treatments" for each chair. For each individual to sit in each chair under each "treatment" would require 36 sittings for each individual. Again, this believed to be unreasonable. The researcher would like each individual to sit in each chair at least once and under each "treatment" at least once. Suggest an appropriate design. For simplicity you may use A, B, C, D, E, F to represent the six "treatments," i.e., the Armrest by Backrest combinations. How many individuals are needed? Show at least a portion of the design and discuss briefly the randomization process you would use

## **METHODS QUALIFYING EXAM**

**AUGUST 2001**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## PROBLEM #1

In a study of the effectiveness of several new types insecticides, an entomologist designs an experiment to evaluate the relative effectiveness of the five insecticides ( $I_1, I_2, I_3, I_4, I_5$ ) on brown spotted ticks. From previous studies with similar types of insecticides, she anticipates that under normal conditions, a brown spotted tick exposed to any one of the five insecticides would have a chance of surviving between 30% and 70%.

Five thousand (5,000) very similar brown spotted ticks were partitioned among ten containers, with 500 brown spotted ticks assigned to each container. After an assimilation period, two of the ten containers were randomly selected to receive insecticide  $I_1$ , two other containers were randomly selected to receive insecticide  $I_2$ , two other containers were randomly selected to receive insecticide  $I_3$ , two other containers were randomly selected to receive insecticide  $I_4$ , and the final two containers were assigned to receive insecticide  $I_5$ . At the end of a 24 hour exposure period, the entomologist examined each of the 5,000 brown spotted ticks and classified the tick as “dead” or “not dead”. The response of interest to the entomologist was  $y_{ij}$ , the number of “dead” brown spotted ticks (out of the 500) in container  $j$  assigned to insecticide,  $I_i$ ;  $i = 1, 2, 3, 4, 5$  and  $j = 1, 2$ .

- (1) Provide an ANOVA table for this experiment, including sources of variation and degrees of freedom. Indicate for each source of variation whether the source is a fixed or random effect. Find the expected mean square for each source of variation in your ANOVA table. Indicate the appropriate denominator of the  $F$  statistic for each relevant test.
- (2) The customary ANOVA methodology appropriate for this problem provides a test of the null hypothesis that the five insecticides were equally effective. If the test statistic rejects the null hypothesis, explain briefly how you would identify *which* of the five insecticides differed in their effectiveness in controlling brown spotted ticks. Be as specific as possible, and be sure to explain why you believe that your recommended method is appropriate for this problem.
- (3) Identify any potential problems with using the type of analysis given in Part (1) and how you would resolve these problems.

- (4) After discussing your proposed analysis, the entomologist states, “I think I can just carry out a standard chi-square test of the 5,000 brown spotted ticks classified by a 2x5 contingency table, with the five columns corresponding to the five insecticides and the two rows corresponding to the dead/not dead classification.” Do you agree? Explain why or why not.
- (5) After the experiment was completed, the entomologist asks, “I plan on conducting out a similar experiment next year on the green spotted tick, again involving 5,000 ticks and five different insecticides. How can I change my experimental design to be more efficient?”

Answer the entomologist’s question, taking care to define clearly any notation or special terms you use. In addition, *explain why* you believe your design will be more efficient than the design used in Part (1).

## PROBLEM #2

You are given independent observations  $(Y_{i1}, Y_{i2})$ ,  $i = 1, \dots, N$  on  $N$  individuals.  $Y_{i1}$  denotes a pre-treatment score and  $Y_{i2}$  denotes the post-treatment score on the  $i$ th individual, with  $E(Y_{ij}) = \tau_j$ ,  $\text{Var}(Y_{ij}) = \sigma^2$ ,  $j = 1, 2$ , and  $\text{Corr}(Y_{i1}, Y_{i2}) = \rho$  for  $i = 1, \dots, N$ . (Assume  $\rho$  is known.)

- (i) Set this up in a linear model framework, stating all the underlying assumptions and writing explicitly the variance-covariance matrix of the error.
- (ii) Again set this up in a linear model framework based on the difference of the responses  $(Y_{i1}, Y_{i2})$ ,  $i = 1, \dots, N$  rather than the original responses. It means your new response is  $Z_i = Y_{i1} - Y_{i2}$ ,  $i = 1, \dots, N$ . Write the new variance-covariance matrix of the error.
- (iii) Find the unbiased estimators of the model parameters  $\tau_1 - \tau_2$  and  $\sigma^2$  based on the model in (ii).
- (iv) Suppose we assume a bivariate-normal distribution for  $(Y_{i1}, Y_{i2})$ ,  $i = 1, \dots, N$  then obtain the maximum-likelihood estimators of  $\tau_1 - \tau_2$  and  $\sigma^2$ .
- (v) Now if you assign uniform prior distributions for  $\tau_1 - \tau_2$  and the prior distribution for  $\sigma^2$  is  $\pi(\sigma^2) = 1/\sigma^2$ ,  $\sigma^2 > 0$ , then find the conditional posterior distributions of  $\tau_1 - \tau_2$  (conditioned on  $\sigma^2$ )
- (vi) Hence find the posterior probability that the difference of  $\tau_1$  and  $\tau_2$  will be positive (conditioned on  $\sigma^2$ ).

[Hint: If  $(X, Y)$  follows a bivariate normal distribution with means  $\mu_1$ ,  $\mu_2$ , variances  $\sigma_1^2$ ,  $\sigma_2^2$  and correlation coefficient  $\rho$  then the density function is

$$f(X, Y) = \left(2\pi\sigma_1^2\sqrt{1-\rho^2}\right)^{-1} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2} \left[(x - \mu_1^2) - 2\rho(x - \mu_1)(y - \mu_2) + (y - \mu_2)^2\right]\right\}$$

### **PROBLEM #3**

Three different experimental situations are described below. Provide the information requested for each.

- A. A study is conducted to examine the effects of three different room temperatures and absence or presence of background noise on the performance of students taking an exam. A total of 48 students are available for the study, with the midterm exam score available for each student. Eight students are to be assigned to each of the six temperature by background noise combinations.
- (i) How would you suggest the students for each temperature by background noise combination be chosen?
  - (ii) The score on the exam, taken under the experimental conditions, is recorded for each student. Give the sources and degrees of freedom for the appropriate ANOVA table. Where appropriate, indicate the numerator and denominator mean squares for testing significance of an effect.
  - (iii) The exam consists of a multiple choice part, a short answer part, a "work a problem" part, and an essay question. Instead of a single exam score for each student, the scores on the individual parts of the exam are recorded separately. Does this change your ANOVA table? If so, provided the correct ANOVA table and F-ratios.

B. A research assistant to the president of a university collected data for a random sample on  $n = 100$  full time tenure track faculty members. The variables, and their values, are:

SALARY = nine month salary equivalent (in Dollars)

AGE = age, in years

RANK = Academic Rank = 1 for Professor

2 for Associate Professor

3 for Assistant Professor

4 for Instructor, Lecturer, etc.

TENURE = Tenure Status = 0 if not tenured and not tenure track

1 if not tenured, but tenure track

2 if tenured

DEGREE = Final Degree = 1 if bachelors

2 if masters

3 if postmaster, but not doctorate

4 if doctorate

SEX = sex of faculty member = 0 if male

1 if female

TIME = Length of time (in years) since initial appointment to faculty

at the university

The president has asked her assistant to determine if there is evidence of discrimination in salaries based on the sex of the faculty member. The research assistant has asked your help.

- (i) If you were to fit a model using SALARY as the response variable (the "Y"), which of the other variables would you consider to be class variables (i.e., ANOVA type variables) and which of the variables would you consider to be covariates (i.e., regression type variables)?
- (ii) The president asks if there is evidence the change in average salary for each additional year since initial appointment is not the same for males and females. How would you determine if there is evidence of this, i.e., what term or terms would you add to the model?
- (iii) Generally, faculty members at the rank of Associate Professor and Professor have tenure, faculty members at the rank of Assistant Professor do not have tenure but are tenure track, and faculty members at the rank of Instructor, Lecturer, etc., do not have tenure and are not tenure track. Could this create any difficulties in fitting a model with all the variables mentioned above included? If so, what is the problem?

C. An analysis of covariance was used to analyze data from an experiment on ten treatments in a Randomized Complete Block Design with five blocks and a covariate, X. Various models were fit to the data. The Error sums of squares (Residual sums of squares) for different models are shown below. The terms in the model are indicated by M = overall mean, BLK = class variable or dummy variables for blocks, TRT = class variable or dummy variables for treatments, and X for the covariate.

Model Includes	SS Error	Error DF
<hr/>		
M	393.500	49
M, BLK	332.300	45
M, TRT	321.200	40
M, BLK, TRT	259.000	36
M, BLK, TRT, X	123.120	35
M, BLK, TRT, X(TRT)	82.226	26
M, BLK, TRT, X, X*TRT	82.226	26
M, BLK, X	197.599	44
M, TRT, X	139.928	39
M, X	214.637	48

- (i) Give the value of the F-ratio to test if the slopes for the ten lines are equal.
- (ii) Assuming a common slope, test that the adjusted treatment means are equal.

**METHODS QUALIFYING EXAM**

**JANUARY 2002**

**INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## QUESTION 1

A crop scientist has asked your help in analyzing her data on plant growth for three different varieties of cotton and four different row spacings (4", 8", 12", and 16"). Each row spacing was randomly assigned to two plots. Each plot was divided into three subplots, with a variety randomly assigned to each subplot.

(A) One of the response variables is the total yield from each subplot.

- i. Give an appropriate model for analyzing these data. For each term, indicate if the term is a fixed effect or random effect.
- ii. Give the sources and degrees of freedom for the appropriate ANOVA table. Where appropriate, identify the denominator mean square for the F-test.
- iii. Are there any additional test procedures you would recommend? If yes, identify the procedures.

(B) Suppose instead of measuring the yield for the entire subplot, she measured the yield for randomly selected plants within each subplot.

- i. If she measured six plants within each subplot, what changes would you make to the analyses in part (A)?
- ii. Suppose some subplots have measurements for six plants, others for five plants, others for four plants and others for only two plants. Does this change your recommendations? Does this create any problems for the analyzes? If yes, what are the problems?

(C) Suppose she has six plants for each subplot. In addition to the yield for each plant, she has recorded the plant height. She believes the yield for an individual plant may be linearly related to the plant height.

- i. How would you incorporate plant height into your analysis?
- ii. How would you determine if it is reasonable to assume the linear relationship between plant height and plant yield is the same, except for perhaps the intercept, for all varieties?

## QUESTION 2

- (A) Let  $\theta$  ( $0 \leq \theta \leq 1$ ) denote the probability of pain relief at dose  $x$  of a certain drug. One model for  $\theta$  is to take  $\theta = \int_{-\infty}^x f(u)du$  where  $f(u)$  is a probability density function. If  $f(u)$  represents the extreme value distribution

$$f(u) = \beta \exp[(\alpha + \beta u) - \exp(\alpha + \beta u)], \quad -\infty < u < \infty$$

then find  $\log\{-\log(1 - \theta)\}$ .

- (B) Let  $m_i$ ,  $i = 1, 2, \dots, n$  denote the number of patients exposed at dose level  $x_i$ , and  $y_i$  denote the number of patients relieved from pain at that dose level. Suppose you model the probability of pain relief  $\theta_i$  at the dose level  $x_i$  by a logistic model with dose level ( $x$ ) as the only explanatory variable with intercept parameter  $\alpha$  and the slope parameter  $\beta$ .

In the logistic model we use logistic link function so that  $\theta = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ .

- i. Write the model with a graphical presentation.
- ii. Find the maximum likelihood equations for estimating  $\alpha$  and  $\beta$ .

### QUESTION 3

A company designs a study to evaluate two methods ( $M_1, M_2$ ) for converting recycled automobile tires into surfaces for tennis courts. The company wants to compare the average surface traction of the material produced by the two processes. Method  $M_1$  is the conventional method of conversion and Method  $M_2$  is a new method which is more expensive in its conversion of the tires. The company wants to determine if  $M_2$  produces a surface having a higher average traction rating than  $M_1$ . Since  $M_2$  is more expensive, the mean traction of  $M_2$  must be at least 5 units larger than the mean for  $M_1$  in order for it to be considered economically feasible. The company decides to take a random sample of material on 50 consecutive days of production from each of the two methods. The data consists of the surface traction measurements of the 50 samples from

$M_1 : X_1, \dots, X_{50}$  with process mean  $\mu_1$  and process standard deviation  $\sigma_1$

and the surface traction measurements from 50 specimens from

$M_2 : Y_1, \dots, Y_{50}$  with process mean  $\mu_2$  and process standard deviation  $\sigma_2$ .

(A) The company is interested in the research hypothesis  $H_1 : \mu_1 + 5 < \mu_2$ .

- i. Write down a general formula for the  $t$  test statistic for this hypothesis test (It is presumed that  $M_2$  will produce a product having a more consistent surface traction than  $M_1$ .)
- ii. Write down the decision rule for this hypothesis test. Use  $\alpha = 0.05$ .
- iii. State the necessary conditions needed for your procedure to be valid and how you would verify whether the conditions are satisfied in this experimental setting.

(B) In the context of the hypothesis test presented in (A), give clear, explicit definitions of the following terms, **Make Sure to Frame Your Definitions in Terms of This Specific Problem**

- i. Type I error
- ii. Type II error
- iii. Power of the test.

(C) For parts (C) and (D) of this question, you may assume that  $\sigma_1 = 3$  and  $\sigma_2 = 1$  and that the sample sizes are large enough to invoke the central limit theorem if necessary.

- i. Calculate the power of your test for the following six values of the parameter:

$$\mu_2 - \mu_1 = 4.5, 5.0, 5.5, 6.0, 6.5, 7$$

- ii. Use your results from (C.i) to sketch a power curve for your test. Be sure to label your axes clearly.

(D) The company's engineer examines your results from (A) through (C) states, "The power of the test when  $\mu_2 - \mu_1 = 5.5$  is not large enough. Determine the minimum sample size necessary to achieve a power of at least 0.90 when  $\mu_2 - \mu_1 \geq 5.5$ .

(E) The 50 observations from each process represent the surface traction obtained from a single batch of production. Thus, there may be a strong *positive* correlation within the 50 surface traction measurements from a given process. However, the measurements between the two processes remain independent. Given this additional information, answer the following questions.

- i. If the correlations between all pairs of daily measurements from method  $M_1$  are equal to  $\rho_1 > 0$  and the correlations between all pairs of daily measurements from method  $M_2$  are equal to  $\rho_2 > 0$ , how does this positive correlation between the daily measurements affect the estimated standard error of  $\hat{\mu}_1 - \hat{\mu}_2$ ? Justify your answer mathematically.
- ii. Suppose you **did not** adjust for the positive correlation between the daily measurements and proceeded to use the test you proposed in part (A). Will the positive correlation in the data increase or decrease the numerical values of power you calculated for the test statistic in part (C)? Explain.
- iii. Suppose you **did not** adjust for the positive correlation between the daily measurements and proceeded to obtain a 95% C.I. for  $\mu_2$  using procedures for independent random samples. What is the effect of the positive correlation in the data on the level of confidence of your C.I.? What is the effect of the positive correlation in the data on the width of your C.I.?
- iv. Suppose that it is known that  $\rho_1 = \rho_2 = .9$  and you use this information to adjust your test statistic to account for the positive correlation.

Taking into account the effect of the positive correlation on the standard error of  $\hat{\mu}_1 - \hat{\mu}_2$  and assuming that  $\sigma_1 = 3$ ,  $\sigma_2 = 1$ ,  $\rho_1 = \rho_2 = .9$  and that the sample sizes are large enough to invoke the central limit theorem:

- i. Calculate the power of an  $\alpha = 0.05$  test for the following six values of the parameter:

$$\mu_2 - \mu_1 = \quad 4.5, \quad 5.0, \quad 5.5, \quad 6.0, \quad 6.5, \quad 7$$

- ii. Use your results to sketch a power curve for the adjusted test. Compare this curve to the curve from (C.ii).

## QUESTION 4

(A) A data set contained 42 observations. Each observation consisted of a value of a response variable, Y, and four predictor variables: X1,X2,X3, and X4. The table shown below gives the Error Sum of Squares (SSE) for each of the possible models of from 1 to 4 of the predictor variables. (An X indicates that the predictor variable is in the model. All models include an intercept term.)

Model	X1	X2	X3	X4	SSE
1	X				1062.92
2		X			2431.72
3			X		1941.64
4				X	3530.88
5	X	X			1055.03
6	X		X		968.99
7	X			X	1020.51
8		X	X		1923.71
9		X		X	2322.45
10			X	X	1931.20
11	X	X	X		896.57
12	X	X		X	1009.58
13	X		X	X	958.85
14		X	X	X	1853.74
15	X	X	X	X	884.51

- i. Use the information in the above table to calculate the numerator and denominator sum of squares for the F-statistics to test the null hypotheses listed below:

$$H_o : \beta_3 = 0 | \beta_o, \beta_1, \beta_2, \beta_4 \neq 0$$

$$H_o : \beta_1 = 0, \beta_4 = 0 | \beta_o, \beta_2, \beta_3 \neq 0$$

$$H_o : \beta_2 = 0 | \beta_o, \beta_1, \beta_4 \neq 0, X_3 \text{ not in the model}$$

- ii. Suppose you want to test  $H_o : \beta_3 = 3, \beta_2 = -2\beta_4 | \beta_o, \beta_1 \neq 0$ . Indicate the appropriate numerator and denominator mean square for the F-statistic. If the appropriate SSE is not in the table, explain how you would find the appropriate SSE.

(B) The ANOVA table and information about the coefficients for a data set with seven predictor variables are shown below. Use this information to answer the questions following the tables.

Model	Sum Squares	df	Mean Square	F	Significance
Regression	31210146	7	4458592.252	20.187	.000
Residual	7509549	34	220869.100		
Total	38719695	41			

Term	Unstandarized Coefficients		Stand.Coefficient $\hat{\beta}'_i$	t	Significance	Collinearity Sta	
	$\hat{\beta}_i$	Std. Error				Tolerance	
Constant	-151.869	348.256		-.436	.666		
X1	-4.24E-05	.000		-.059	-.314	.755	.160
X2	6.321	.972		.678	6.501	.000	.524
X3	.622	.545		.267	1.141	.262	.104
X4	-.884	14.911		-.015	-.059	.953	.084
X5	-1.848	2.563		-.090	-.721	.476	.365
X6	1.268E-03	.007		.036	.195	.847	.166
X7	17.875	10.720		.179	1.668	.105	.497

- From the information in the above table, can you determine that both X4 and X6 may be deleted from the model, using a significance level of 0.10? If not, why not?
- If BACKWARD variable selection is to be used, with significance level 0.10, which variable would be first deleted? Why?
- Is there an indication of multicollinearity? If so, which variables do you suspect? Support your answer with results shown in the tables.

**METHODS QUALIFYING EXAM**

**AUGUST 2002**

**INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## **PROBLEM #1**

You have been asked to help a researcher design her study of the effect of temperatures on the germination of tomato seeds. The experiment is a two-factor factorial, with combinations of four temperatures (Factor T) and two different types of seeds (Factor S) i.e., eight Atreatment combinations@. A fixed number of seeds will be placed in a growth chamber, with the temperature of the growth chamber set at one of the four temperatures. There is only one growth chamber available for the study. After one hour (I know this is too short a time, but humor me), the percentage of seeds that have germinated is determined. Warning: The descriptions below may include more factors than necessary for the appropriate ANOVA.

Various study conditions are described below. For each description (i) state briefly how you would assign the different levels of factors T and S, (ii) provide the appropriate ANOVA table (sources and degrees of freedom), and (iii) indicate the denominators of the F-tests for testing main effects and interactions for the factors T and S.

- (A) For each of six days (Factor D), there are eight time periods (Factor P) available. Seeds of only one type can be in the growth chamber at any one time. The temperature of the chamber may be changed each time period, if necessary. (Total of 48 observations)
- (B) Suppose it is inconvenient to change the temperature, so it is preferred that both types of seeds be used at a temperature setting before changing to a different setting. As before, only one type of seed can be in the chamber at any one time. There are still six days (Factor D) with eight time periods (Factor P) per day. (Total of 48 observations)
- (C) Another change. There may only be one temperature per day for the chamber, and there are only two periods (Factor P) per day, but now the study will be conducted over 24 days (Factor D). Only one type of seed may be in the chamber at any one time. (Total of 48 observations)

## PROBLEM #2

The American Red Cross analyzes blood samples from  $n = 100$  randomly selected donors in Texas. The samples were then analyzed and produced a measurement,

$$X_i = \text{Serum lead level (in micrograms per deciliter) for donor } i.$$

From past experience, the distribution of  $X_i$  is a lognormal distribution. Thus, a statistician defined the transformed variable  $Y_i = \ln(X_i)$ . For purposes of our analysis, we will assume that

$$Y_i = \mu + e_i \quad (1)$$

where  $\mu$  is a fixed mean parameter for blood donors in Texas and  $e_i, i = 1, \dots, 100$  are independent and identically distributed normal random variables with mean zero and variance  $\sigma^2$ . For our sample of 100 donors, some summary statistics for the  $Y_i$  values are the sample mean  $\bar{y} = 1.36$  and the sample variance  $s_y^2 = 1.44$ .

- (A) Use model (1) and the information above to compute a 95% confidence interval for  $\mu$ .
- (B) Now consider a new observation  $Y_{n+1}$  that also satisfies model (1). Compute a 95% *prediction interval* for  $Y_{n+1}$ , based on the old data  $Y_i, i = 1, \dots, 100$ .
- (C) Give a customary interpretation of your prediction interval in (B), paying special attention to: (i) what is fixed; (ii) what is random; and (iii) to what probability does the term “95%” refer?
- (D) Explain how the interpretation of your prediction interval from (C) differs from the customary interpretation of the confidence interval in (A).
- (E) Recall that our transformed observations  $Y_i = \ln(X_i)$  were recorded on a log-transformed scale. Use the results from the preceding steps to compute a 95% prediction interval for a new observation  $X_{n+1}$  on the *original* (untransformed) scale.
- (F) Explain why the procedure you used in (E) to obtain the 95% prediction interval for  $X_{n+1}$  would not yield an **exact** 95% confidence interval for  $E(X_i)$ .
- (G) Give a justification for using the procedure in (E) to obtain an **approximate** 95% confidence interval for  $E(X_i)$ .
- (H) The prediction interval in part (B) is based on the assumption that the  $Y_i$  values are normally distributed. List *two* specific methods that you could use to check this assumption based on the transformed observations  $Y_i$ . For *each* of these two methods, give explicit rules (omitting numerical critical values) you would follow to decide whether the observations are consistent with the normal-distribution assumption.

### PROBLEM #3

A study was conducted to understand features of growth curves for boys and girls between age 7 – 12.  $n_1$  boys and  $n_2$  girls are randomly selected into the study. The following variables are considered:

$Y$ : height in inches.

$X$ : age.

$Z$ : a binary indicator; 0 for boys and 1 for girls.

It is widely accepted that the best parametric model for growth curves is a quadratic polynomial model. Answer the following questions.

- (A) Let  $\beta$ 's denote the regression parameters in the mean function. Please write down one general regression model that could be used to describe all boys and girls growth data.

- (B) For a regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,$$

we let  $\text{RSS}(X_1, X_2)$  denote the sum of squared residuals,  $\sum_{i=1}^n (Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$ , where  $(\hat{\beta}_1, \hat{\beta}_2)$  are the least square estimate of the  $\beta$ 's. For the following tests, please provide, (i) a model which could be a sub (reduced) model of the model in (A); (ii)  $H_0$  and  $H_A$ ; (iii) Test statistic; (iv) the proper test procedure which includes the correct specification of the degree of freedoms (if needed).

- (a) There are 2 different growth curves for boys and girls.

- (b) It is widely believed that the growth of boys and girls have the same starting point. Based on this believe, please test there are 2 different growth curves.

- (C) Several assumptions are needed for the test(s) in (B). Please state them. Please also make comments on would large  $n_1$  and  $n_2$  allow each of the assumptions stated be eliminated.

#### **PROBLEM #4**

You have been asked to help with the analysis of data from a study using a 3x4 factorial arrangement of four different growth stimulators at three different dose levels on the growth of a particular type of animal. There were three different dosage rates (say R1, R2, and R3) are equally spaced. Each dosage rate by stimulator type combination was feed to six randomly selected animals. The response variable is a measure of growth for a fixed period of time. (Total of 72 observations)

- (A) What analysis would you recommend in addition to the usual ANOVA?
- (B) Whoops, you learn that  $R1 = 0$ , i.e., an absence of any stimulator. Does this change your analysis? If so, how?

**METHODS QUALIFYING EXAM**

**JANUARY 2003**

**INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## QUESTION 1

A study was conducted to compare different methods of reducing the concentration of cedar flies (insects that inhabit cedar trees). The reduction methods consisted of two types of chaining (cutting the trees down) along with burning all the trees, grass and leaf litter. (Cedar flies overwinter in the leaf litter.) The study design had the following "factors":

1. There were four different geographical regions containing cedar trees.
2. Within each region, there were two techniques for cutting the trees down (smooth chaining and elevated chaining) plus a control (no chaining) assigned at random to one of three groves. The chaining was done prior to collecting the data in the first year.
3. The study was conducted over a five-year period. The treatment assigned to a grove was the same every year. A sample of cedar flies was taken at fixed periods during the summer months and the average number of flies was recorded for each treatment (that is, one assessment of fly concentration for each grove for each year).
4. The time frame for years was:
  - Year 1: In the spring the treatments (smooth chaining, elevated chaining, and no chaining) were applied to the appropriate groves. Data were collected during the summer.
  - Year 2: No additional treatment of the groves. Data was collected during the summer.
  - Year 3: No additional treatment of the groves. Data was collected during the summer.
  - Year 4: For the sites with chaining, the litter was burned in the spring. Nothing was done to the control sites. Data was collected during the summer.
  - Year 5: No additional treatment of the groves. Data was collected during the summer.

There are a total of 60 data values.

- (A) Give an appropriate ANOVA table for these data, showing sources of variation, degrees of freedom, and expected mean squares.
- (B) What hypotheses would you test? Show the appropriate F-ratios for each hypothesis using the names of the mean squares.
- (C) Are there any other tests and/or estimates you would conduct in addition to these F-tests? If so, describe them.

## **QUESTION 2**

In both analysis of variance and regression analyses when the sample sizes are relatively small, the model conditions become crucial in obtaining valid inferences. In most situations, the residuals are used to assess whether these conditions hold for a given data set. Describe methods for using the residuals to assess the validity of each of the following conditions.

- (A) Normality
- (B) Independence
- (C) Discuss three different procedures for testing the equality of variance. Give the advantages and disadvantages of each procedure.

### QUESTION 3

Suppose 250 independent pairs,  $(X_1, Y_1), \dots, (X_{250}, Y_{250})$ , are randomly selected from a population having an underlying quadratic relationship,  $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ , where  $\epsilon_i$ 's are i.i.d. random variables with mean 0 and variance  $\sigma^2$ . The researcher suspects a nonlinear relationship between  $X$  and  $Y$  and decides to fit a cubic polynomial regression model to the data. Let  $\mathbf{X}_{[3]}$  and  $\mathbf{H}_{[3]}$  denote the design matrix and the "hat matrix" for this cubic polynomial regression. Also, let  $\mathbf{1}$ ,  $\underline{\mathbf{x}}$ ,  $\underline{\mathbf{x}^2}$ , and  $\underline{\mathbf{x}^3}$  denote the columns in  $\mathbf{X}_{[3]}$ .

- (A) Show that  $(\mathbf{X}_{[3]}^t \mathbf{X}_{[3]})^{-1} \mathbf{X}_{[3]} \mathbf{1} = (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})^t$  and that  $(\mathbf{X}_{[3]}^t \mathbf{X}_{[3]})^{-1} \mathbf{X}_{[3]} \underline{\mathbf{x}^2} = (\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0})^t$
- (B) Suppose one goal of the study is to estimate the mean of  $Y$  when  $X = x_o$  for a given  $x_o$ . Is it possible to obtain an unbiased estimate of the conditional mean even though the wrong model has been fitted to the data? Use (A) to explain your answer.
- (C) Suppose one also wants to provide a prediction interval for a possible new observation  $Y^*$  when  $X = x_o$  using the fitted cubic model. Would the mis-specification of the model play a role in this prediction? Are there any other issues that one should consider before constructing such an interval?

#### QUESTION 4

Three different alloys were prepared with four separate castings for each alloy. Two bars from each casting were tested for tensile strength. The data are tensile strengths of the individual bars. The data is given below:

Alloys	Castings				Mean
	1	2	3	4	
A	13.2	15.2	14.8	14.6	14.7
	15.5	15.0	14.2	15.1	
B	17.1	16.5	16.1	17.4	16.7
	16.7	17.3	15.4	16.8	
C	14.1	13.2	14.5	13.8	14.1
	14.8	13.9	14.7	13.5	

- (A) Write a linear model for the experiment. Make sure to identify which terms are fixed and which terms are random.
- (B) Complete the following ANOVA table:

Source	DF	SS	MS	Expected Mean Squared
Alloys		29.38		
Castings		4.68		
Bars		4.32		

- (C) Is there a significant difference in the mean tensile strength of the three alloys?
- (D) Compute 95% confidence intervals for the mean tensile strength of the three alloys.
- (E) Proportionally allocate the variance in tensile strengths to its components.

## **METHODS QUALIFYING EXAM**

**AUGUST 2003**

### **INSTRUCTIONS:**

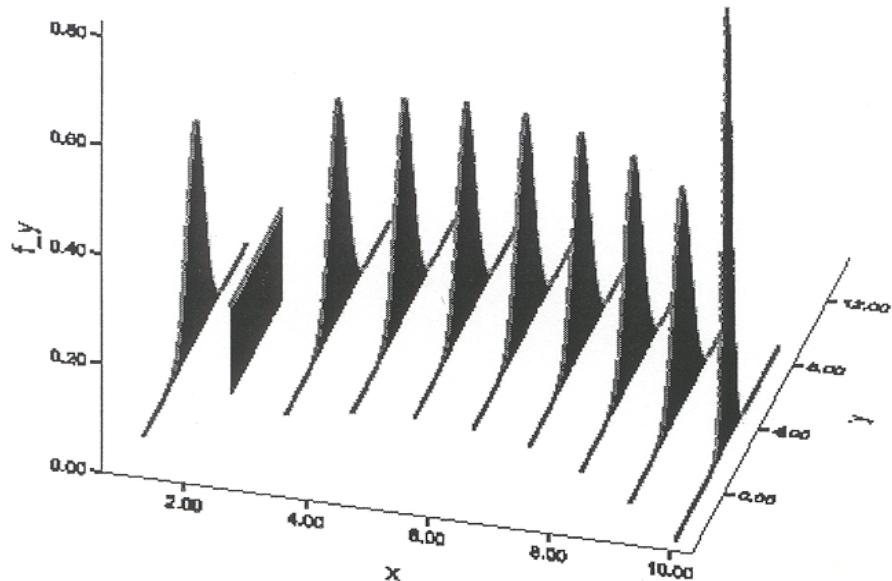
1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.
5. 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.

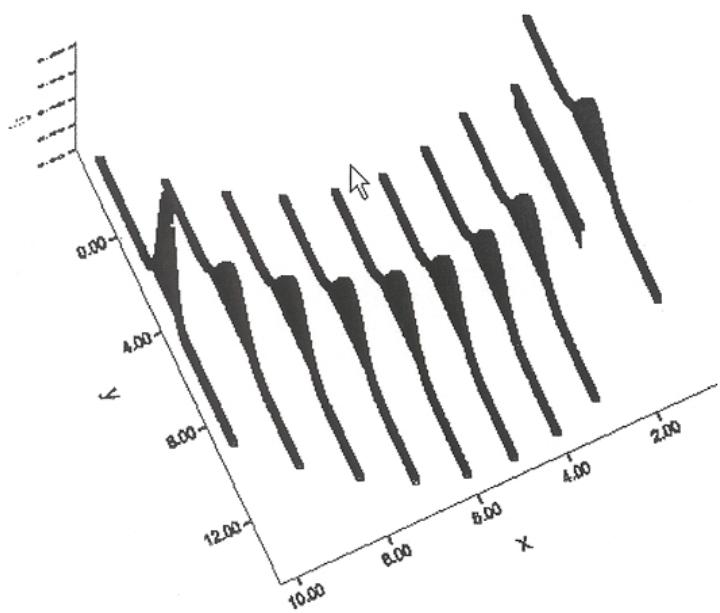
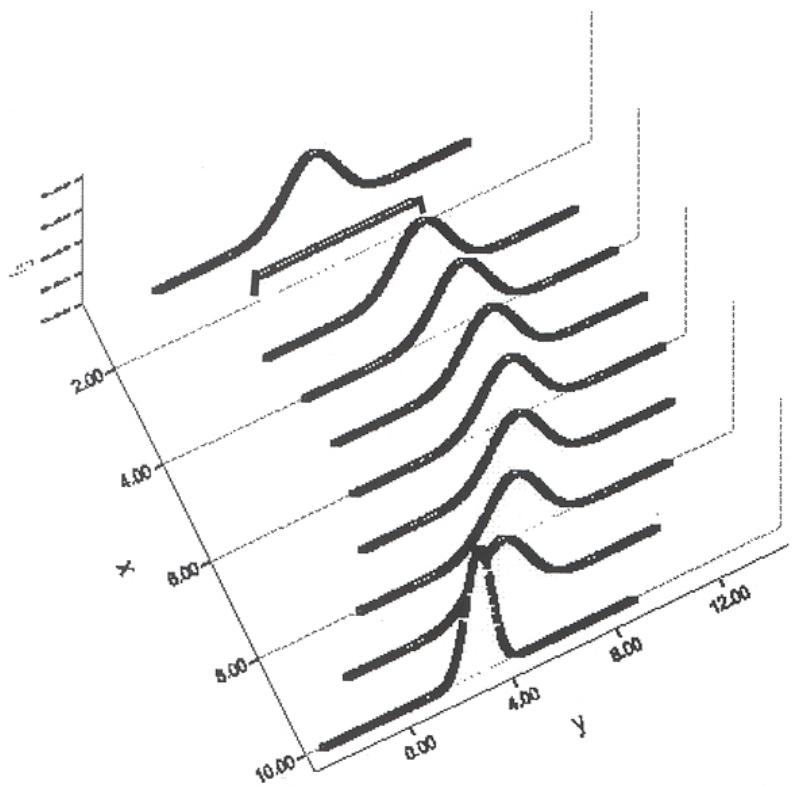
### QUESTION #1:

We have observed  $y$  = response (change in blood pressure) and  $x$  = dosage level of a drug. We assume a polynomial relationship between  $y$  and  $x$ .

The three graphs are all the same; but have been rotated to give additional views.

1. What is the appropriate model?
2. What are the usual regression assumptions?
3. Answer the following (in detail where appropriate):
  - a. Sketch  $E(y)$
  - b. Based on the graphs, make comments about the assumptions. Do they appear to be satisfied or violated?
  - c. How many populations are represented by the graphs?
  - d. For  $x = 10$ , sketch the pdf of  $y$ . Label the axis, approximate min and max values such that  $f_y > 0$ , indicate the mean and standard deviation.
  - e. list all of the parameters in the model
  - f. What is the effect of non-normal errors on hypothesis testing?
  - g. What is the effect of non-constant variance on hypothesis testing?





## QUESTION #2:

A company is interested in the ability of a machine to consistently place electrical wire on a coil. There are three varieties of machines available: hand operated(HO), partially computer operated(PCO), and completely automated(CA). Three machines of each type are randomly selected from their suppliers for use in the study. The wire placed on the coils comes in one of three thicknesses: .02mm, .04mm, or .06mm. Each of the machines assembles two coils of each of the three wire thicknesses. Each wound coil is then measured for the uniformity of windings at a middle position on the coil. These measurements are given in the following table.

		VARIETY OF MACHINE								
		HO			PCO			CA		
MACHINE ID		1	2	3	4	5	6	7	8	9
THICKNESS	.02mm	12.30	13.46	12.35	13.01	13.46	13.15	5.47	5.75	6.24
		12.59	14.00	12.06	12.63	13.92	13.20	5.96	5.68	6.15
	.04mm	13.16	13.29	12.50	12.74	13.84	13.46	5.73	5.60	5.92
		13.00	13.62	12.39	12.68	13.75	13.57	5.64	5.65	5.64
	.06mm	12.87	13.46	12.73	12.47	13.62	13.36	5.01	5.80	6.19
		12.92	13.82	12.15	12.15	13.28	13.42	5.62	5.71	6.23

- A. Write a linear model for the above experiment. Make sure to identify the terms in your model with respect to distributional properties or restrictions on population parameters.
- B. Complete the following AOV table.

SOURCE	DF	MS	EMS
THICKNESS		0.1263	
VARIETY		6.4424	
THICKNESS*VARIETY		0.0594	
MACHINE(VARIETY)		0.6702	
THICKNESS*MACHINE(VARIETY)		0.0919	
ERROR		0.0404	

- C. Using the numeric values of the MS's given above and your EMS's, provide the following information:
- Estimate the variance in the uniformity of windings of a randomly selected coil wound with .04mm wire using a CA winding machine.
  - Proportionally allocate the variance of the uniformity in windings of a randomly selected coil to the various variance components.
  - An estimate of the standard error of the estimated mean uniformity of windings from a CA winding machine.
  - An estimate of the standard error of the estimated difference in the mean uniformity of windings between HO and CA winding machines.
  - An estimate of the standard error of the estimated difference in the mean uniformity of windings of coils using wire of thickness .02mm and .06mm assembled using a CA winding machine.

### QUESTION #3:

The data set consists of the effects of five different drugs on lab rats. A response variable, Y, is measured for each rat. It was believed the response variable for each rat may have a linear relationship with the rat's initial weight, W. The ANOVA tables for five different analyses of these data are shown below and on the next pages. Use them to answer the following questions. For each part, A., B., and C., state the appropriate  $H_0$  and  $H_a$ , the model used (for example Model I), the value of the test statistic, and your conclusion.

- It is known there is a linear relationship between the response Y and the initial weight W. Is it reasonable to assume the slope of the response is the same for all five drugs?
- Perhaps counter to what you found in A., assume a common slope for the five drugs. Is there significant evidence the mean responses for the five drugs, after adjusting for initial weight, are not all equal?
- The SPSS printout for the least squares means (adjusted means) for Model V are shown. Which pairs of means are significantly different at a 0.05 level of significance?

### ANALYSES FOR QUESTION 3

In all five models, the variable DRUG has the "values" 1, 2, 3, 4, and 5. The variable W is a continuous variable. (The analyses for the five models is continued on the next pages.)

MODEL I      Y = DRUG

#### Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.733a	4	.183	10.657	.000
Intercept	8.567	1	8.567	498.346	.000
DRUG	.733	4	.183	10.657	.000
Error	.258	15	.1719E-02		
Total	9.558	20			
Corrected Total	.991	19			

a. R Squared = .740 (Adjusted R Squared = .670)

MODEL II      Y = W

#### Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.461a	1	.461	15.662	.001
Intercept	7.894E-02	1	7.894E-02	2.682	.119
W	.461	1	.461	15.662	.001
Error	.530	18	.2943E-02		
Total	9.558	20			
Corrected Total	.991	19			

MODEL III    Y = DRUG    DRUG\*W

**Tests of Between-Subjects Effects**

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.874 <sup>a</sup>	9	9.716E-02	8.358	.001
Intercept	8.482E-03	1	8.482E-03	.730	.413
DRUG	5.704E-02	4	1.426E-02	1.227	.359
DRUG * W	1.070E-01	5	2.140E-02	1.840	.193
Error	.116	10	1.163E-02		
Total	9.558	20			
Corrected Total	.991	19			

a. R Squared = .883 (Adjusted R Squared = .777)

MODEL IV    Y = DRUG    W    DRUG\*W

**Tests of Between-Subjects Effects**

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.874 <sup>a</sup>	9	9.716E-02	8.358	.001
Intercept	8.482E-03	1	8.482E-03	.730	.413
DRUG	5.704E-02	4	1.426E-02	1.227	.359
W	6.015E-02	1	6.015E-02	5.174	.046
DRUG * W	4.686E-02	4	1.172E-02	1.008	.448
Error	.116	10	1.163E-02		
Total	9.558	20			
Corrected Total	.991	19			

a. R Squared = .883 (Adjusted R Squared = .777)

MODEL V      Y = DRUG      W

### Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.828 <sup>a</sup>	5	.166	14.206	.000
Intercept	7.138E-04	1	7.138E-04	.061	.808
DRUG	.367	4	9.166E-02	7.867	.002
W	9.476E-02	1	9.476E-02	8.133	.013
Error	.163	14	1.165E-02		
Total	9.558	20			
Corrected Total	.991	19			

a. R Squared = .835 (Adjusted R Squared = .777)

Least Square means for Drugs 1-4 using MODEL V

### Pairwise Comparisons

Dependent Variable: Y

(I) MEDICINE	(J) MEDICINE	Mean Difference (I-J)	Std. Error	Sig. <sup>a</sup>	95% Confidence Interval for Difference <sup>a</sup>	
					Lower Bound	Upper Bound
1.00	2.00	1.548*	.567	.008	.414	2.682
	3.00	1.868*	.579	.002	.710	3.027
	4.00	2.986*	.576	.000	1.833	4.139
2.00	1.00	-1.548*	.567	.008	-2.682	-.414
	3.00	.320	.575	.579	-.830	1.470
	4.00	1.438*	.573	.015	.292	2.584
3.00	1.00	-1.868*	.579	.002	-3.027	-.710
	2.00	-.320	.575	.579	-1.470	.830
	4.00	1.118	.566	.053	-1.585E-02	2.251
4.00	1.00	-2.986*	.576	.000	-4.139	-1.833
	2.00	-1.438*	.573	.015	-2.584	-.292
	3.00	-1.118	.566	.053	-2.251	1.585E-02

Based on estimated marginal means

\*. The mean difference is significant at the .05 level.

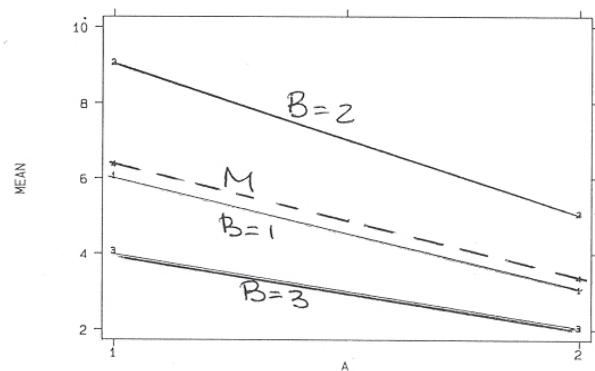
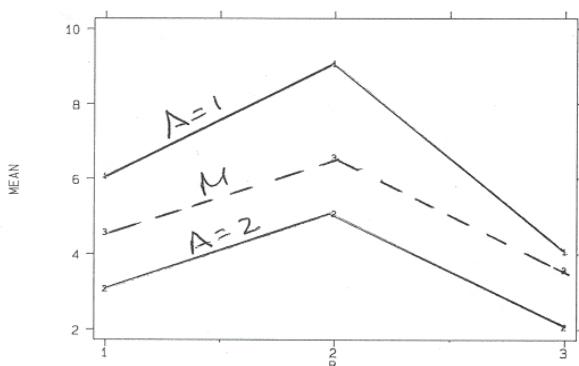
a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

### QUESTION #4:

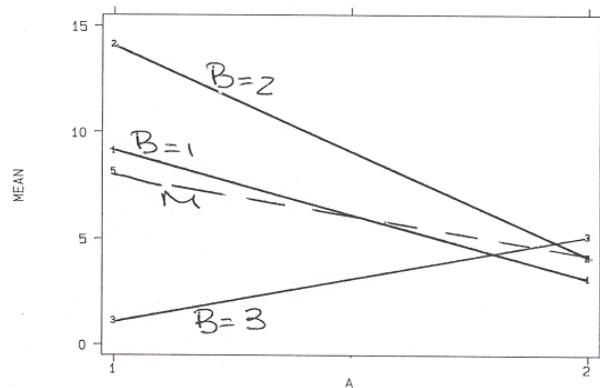
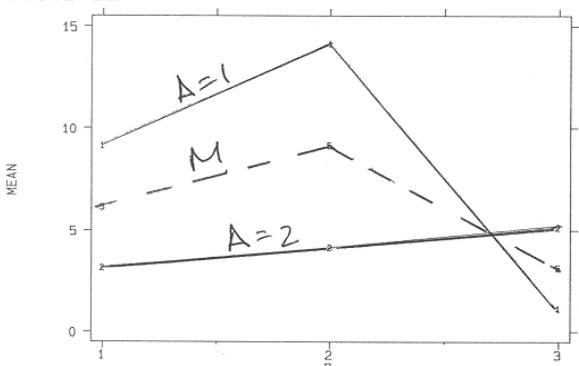
Short discussion questions. Provide short, but sufficient, answers to the following questions.

- A. Consider a two factor experiment with the two factors, A at 2 levels and B at 3 levels. Three pairs of plots of simple effects and the main effects (dashed line labeled M) are shown below. For each pair of plots indicate which effects (A main effect, B main effect, AB interaction) you would judge to be significant. Remember, there may be some variation due to random variation.

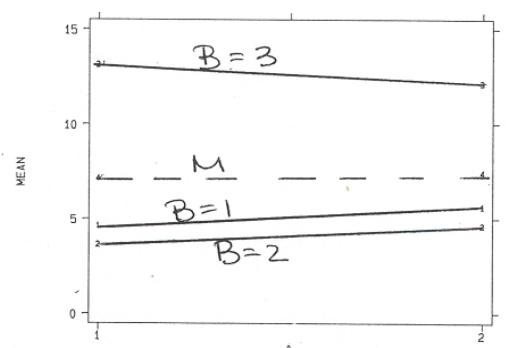
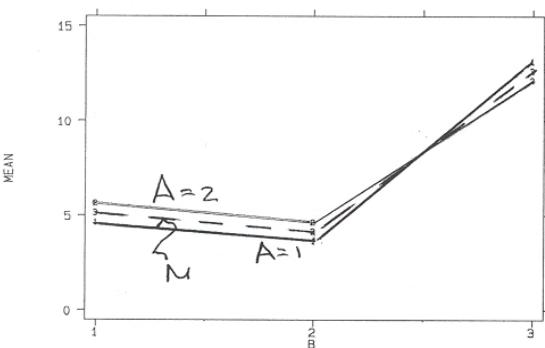
PLOTS I



PLOTS II



PLOTS III



- B. Suppose you have an experiment with six treatment groups. Different descriptions of these treatments groups are described below. The F-ratio for treatments in the ANOVA is significant. For each indicate the type of analysis of means (pairwise comparisons, contrasts, etc.) you believe most appropriate. Give a brief description of the analysis and explain why you choose the analysis.
- a. The six groups are different brands of cereal, two made by General Mills and four made by Post.
  - b. The six groups correspond to two classes taught by Full Professors, two classes taught by Associate Professors, one class taught by an Assistant Professor, and one class taught by a Graduate Assistant.
  - c. The groups are for a study to determine if there are any differences in the mean moisture content of product from six different production methods for producing the product.
  - d. The groups are six methods for measuring the amount of protein in samples of breakfast cereals. One of the groups is the current method. The other methods were proposed methods which are cheaper and easier than the current method. Any of the proposed methods may replace the current method if it is shown that the average response for the samples used in the ANOVA is not significantly different from the mean of the current method
  - e. The six groups are drugs to treat chronic headaches. The measurement is time to relief. One of the groups is a standard drug and the goal of the study is to determine drugs that give faster relief (shorter time to relief) than the standard.
  - f. The six groups correspond to pressures of 5, 10, 15, 20, 25, and 30 psi.
- C. Research scientists of a paint company studied the weathering ability of two new paints when applied to soft and hard wood boards. Two soft woods (pine and poplar) and two hard woods (oak and maple) were used. Three boards of each type were painted with paint A, three boards of each type were painted with paint B. However, the pine boards for Paint A were lost so only the treatment means shown below were available. MSE and its degrees of freedom are:
- | Paint | B    | A      | B      | A     | B     | A   | B   | MSE    |
|-------|------|--------|--------|-------|-------|-----|-----|--------|
| Wood  | Pine | Poplar | Poplar | Maple | Maple | Oak | Oak | = 0.30 |
| Mean  | 7.4  | 8.6    | 8.2    | 9.0   | 8.4   | 7.6 | 7.4 |        |
- a. Give the appropriate sources and degrees of freedom for the ANOVA table.
  - b. The scientists use a contrast with coefficients (-3, 4, -3, 4, -3, 4, -3).
    - i. What  $H_0$  and  $H_a$  (in words) is this contrast testing?
    - ii. Provide the values of the contrast, test statistic, and other quantities necessary for testing the hypotheses.
  - c. Suggest another appropriate contrast. State (in words) what this contrast is testing and provide the coefficients.

## **METHODS QUALIFYING EXAM**

**JANUARY 2004**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

### PROBLEM #1:

An experiment was designed to compare three different methods of assessing the knowledge obtained by students in an undergraduate statistics course:

- Method 1: Multiple choice questions
- Method 2: Student provides detailed solutions to problems
- Method 3: Individual oral examinations

To conduct the experiment, four sections of STAT 30x will be randomly selected. The four sections are taught by four different instructors. Six students will be randomly selected from each of four different sections of STAT 30x. Each student will take all three exams (a total of 72 observations). The researcher is interested in the difference in average scores on the three exams and whether the size of the differences between average scores is consistent across the various sections of STAT 30x.

- (A) Suppose the order in which the three exams are taken is randomly determined for each individual student. Display an appropriate ANOVA table for this experiment with sources of variation, degrees of freedom, expected mean squares, and the F-ratio for testing all relevant hypotheses.
- (B) There is concern that there may be an effect based on whether a test is taken during the first, second, or third testing period. Hence, you want to ensure that each test appears in each testing period. Will this change the design? If your answer is yes,
- (i) In the following display, give an example of an appropriate assignment of the three tests ( $M_1, M_2, M_3$ ) to the testing periods for the students.

Student	Instructor 1			Instructor 2			Instructor 3			Instructor 4		
	Period			Period			Period			Period		
	1	2	3	1	2	3	1	2	3	1	2	3
1												
2												
3												
4												
5												
6												

- (ii) Display an appropriate ANOVA table for this experiment listing just the sources of variation and degrees of freedom. You do not need to determine the expected mean squares for this design.

**PROBLEM #2:**

A programmer claims that  $U_1, \dots, U_{50}$ , is a random sample of size 50 from a uniform (0,1) distribution.

- (A) Describe a graphical method to evaluate the programmer's claim. Be sure to label your axes.
- (B) Describe a test of hypotheses to evaluate the programmer's claim.
- (C) If  $U_1, \dots, U_{50}$  were determined to be in fact a random sample from a uniform on (0,1) distribution, show how  $U_1, \dots, U_{50}$  could be used to generate a random sample of 50 observations from a distribution having cdf given by

$$F(y) = 1 - \exp(-(y - \theta)/\beta) \text{ if } y \geq \theta,$$

where  $\theta$  and  $\beta$  are **known** constants.

- (D) Suppose we have  $Y_1, \dots, Y_n$  is a random sample from a population having cdf,  $F(y)$ . Describe a graphical method to evaluate whether  $F(y)$  has the form:

$$F(y) = 1 - \exp(-(y - \theta)/\beta) \text{ if } y \geq \theta,$$

where  $\theta$  and  $\beta$  are **unknown** constants. How can the graphical method be used to yield estimates of  $\theta$  and  $\beta$ ?

### PROBLEM # 3

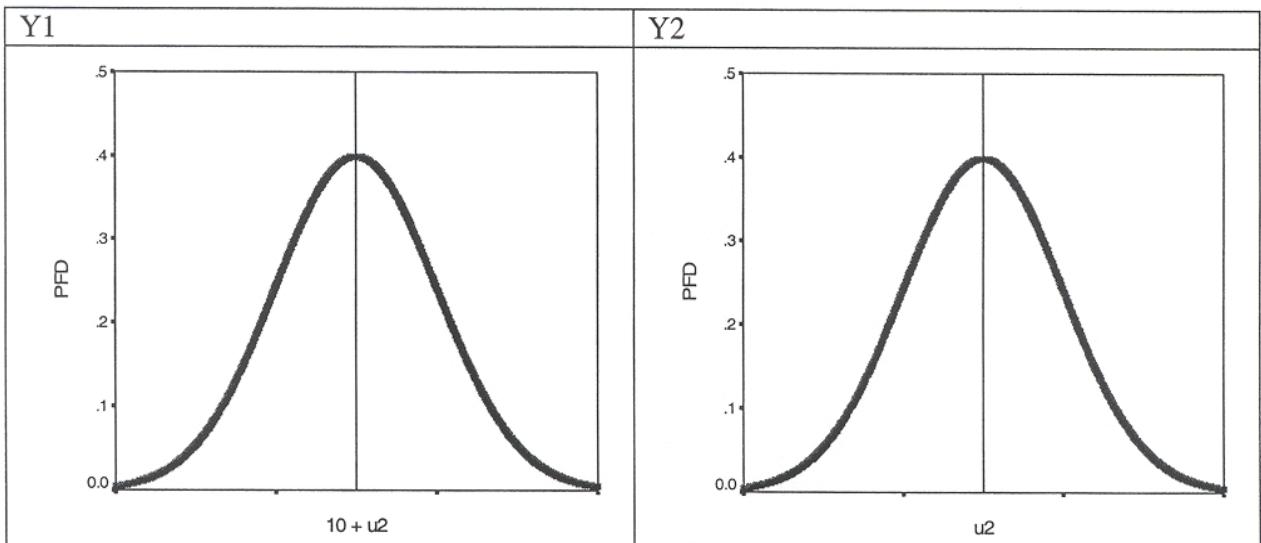
One of the purposes of this exam is to test you on your knowledge. Another purpose is to have you synthesize (bring together) a number of different concepts into a unified approach. Such are the purposes of this problem. There will be four parts to this problem. The following example will illustrate the types of questions for the other four parts.

Example. Suppose we have sample from two normal populations.  $Y_1 \sim N(\mu_1, \sigma_1^2)$  and  $Y_2 \sim N(\mu_2, \sigma_2^2)$ . Also suppose we know that  $\mu_1 = 10 + \mu_2$ . Answer the following questions?

- 1) How many population parameters are there?
- 2) How many parameters do we have to estimate assuming equal variances?
- 3) Sketch the distribution of  $Y_1$  and  $Y_2$ , using what you know.
- 4) What have we learned?

Answers:

- 1) there are 4 population parameters -  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ .
- 2) Since  $\mu_1 = 10 + \mu_2$  and  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  there are only 2 parameters to estimate.
- 3) See below.



- 4) By knowing that  $\mu_1 = 10 + \mu_2$ , we do not have to sample both populations. Instead of three unknown parameters we only have 2.

The actual problem: Note you should not do any calculations – i.e. means variance etc.

Part 1. A regression Problem

Make all of the usual assumptions.

Given the data in Table 1, do the following:

- 1) Write the appropriate model.
- 2) How many populations have been sampled? \_\_\_\_\_
- 3) How many population parameters are there? \_\_\_\_\_
- 4) How many population parameters do we have to estimate? \_\_\_\_\_
- 5) The name of the parameter that regression analysis is most interested in is \_\_\_\_\_ (not the variance)?
- 6) Graph the data
- 7) Put in the least squares line (guess at it)
- 8) Sketch the distribution of Y when X = 5.

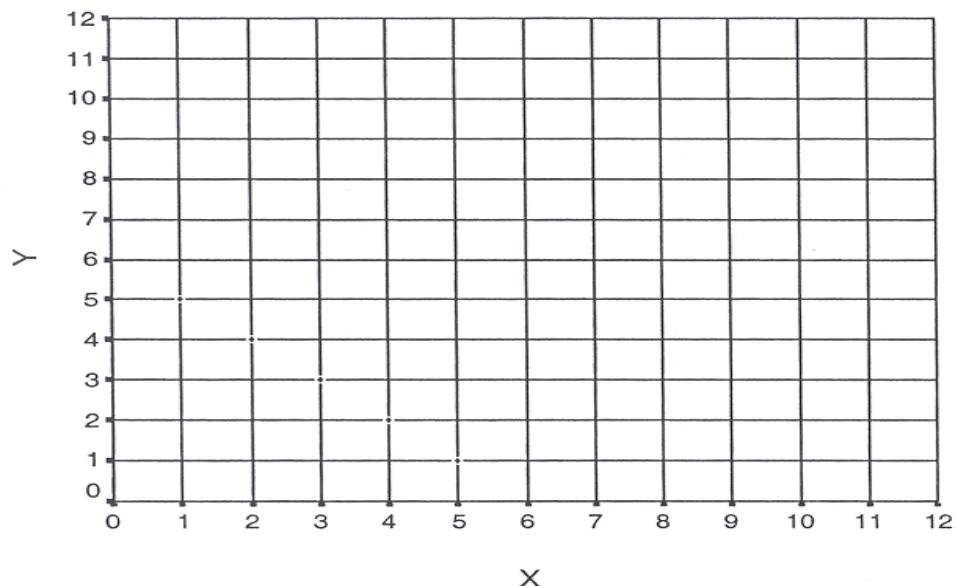


Table 1.

Y	X
1	1
2	3
4	5
5	5
5	8
4	9
7	11

## Part 2. An Analysis of Covariance Problem

Make all of the usual assumptions , and assume unequal slopes.

Given the data in Table 2, do the following:

- 1) Write the appropriate model. Do not use  $\mu + \alpha_i$  as the intercepts. Simply use  $\alpha_i$ .
- 2) How many populations have been sampled? \_\_\_\_\_
- 3) How many population parameters are there? \_\_\_\_\_
- 4) How many population parameters do we have to estimate? \_\_\_\_\_
- 5) The names of the parameters that covariance analysis is most interested in are \_\_\_\_\_ and \_\_\_\_\_ (not the variance)?
- 6) Graph the data
- 7) Put in the least squares lines (guess at them)
- 8) Sketch the distribution of Y when X = 5.

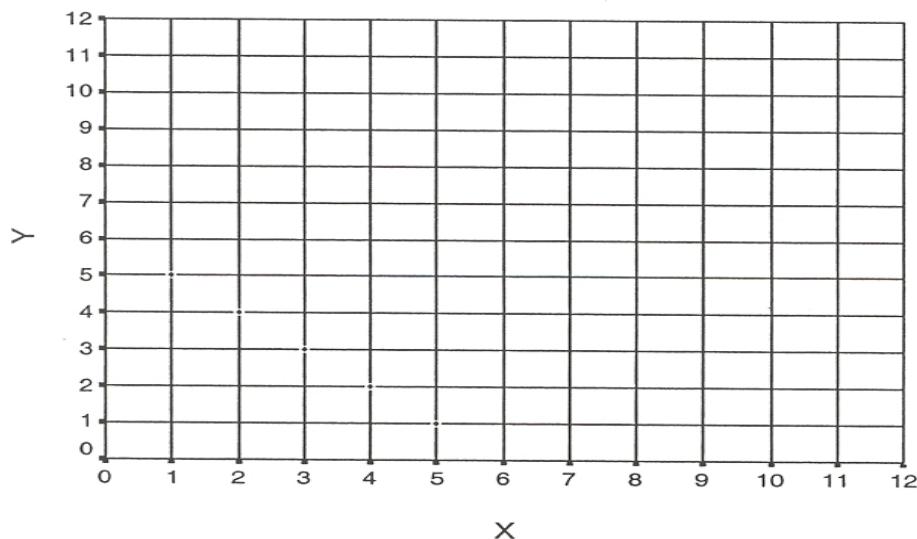


Table 2.

Y	Treatment	X
1	0	1
2	0	3
4	0	5
5	0	5
5	0	8
4	0	9
7	0	11
9	1	2
8	1	4
7	1	7
6	1	7
5	1	10
4	1	12

### Part 3. An Analysis of Variance Problem

Make all of the usual assumptions.

Given the data in Table 3, do the following:

Make this look like an analysis of covariance by attaching  $x = 0$  to each  $y$ .

- 1) Write the appropriate model. Do not use  $\mu + \alpha_i$  as the intercepts. Simply use  $\alpha_i$ .
- 2) How many populations have been sampled? \_\_\_\_\_
- 3) How many population parameters are there? \_\_\_\_\_
- 4) How many population parameters do we have to estimate? \_\_\_\_\_
- 5) The names of the parameters that analysis of variance is most interested in are \_\_\_\_\_ or \_\_\_\_\_ (not the variance)?
- 6) Graph the data
- 7) Put in the least squares lines (guess at them)
- 8) Sketch the distribution of  $Y$  when  $X = 0$ .

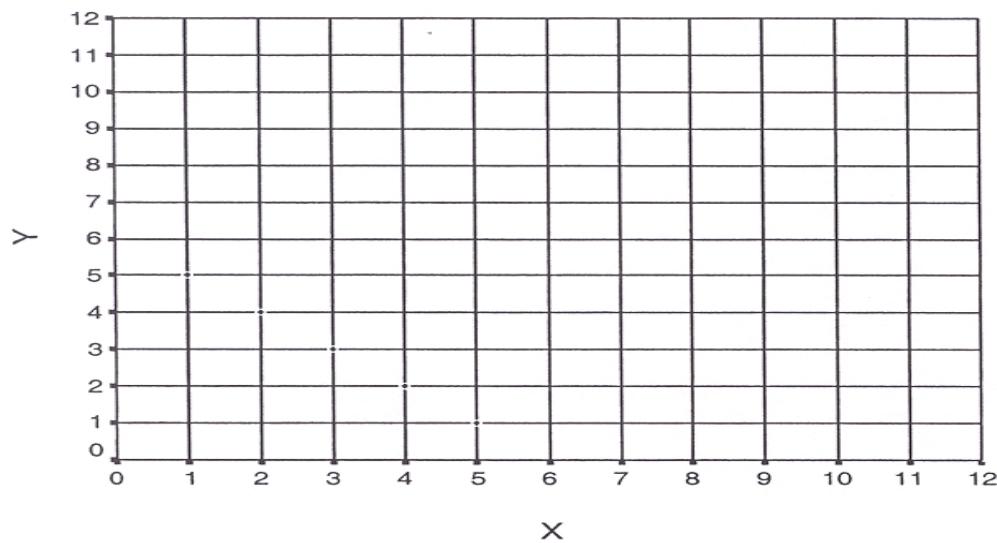


Table 3.

Y	Treatment
2	0
2	0
4	0
5	1
5	1
6	1
7	1
8	1
9	2
10	2
11	2
11	2
12	2

## Part 4.

Hopefully, you have seen that you can display regression, ANCOVA and ANOVA all on a similar type Y by X graph. What have you learned by doing the above parts? What is the commonality of the three analyses? Comment on the number of samples per population needed by each approach.

**METHODS QUALIFYING EXAM**

**AUGUST 2004**

**INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

## PROBLEM I.

The tensile strength of a material is the ability that the material possesses to resist deformation when a force or a load is applied to it. A metallurgist conducts a study to evaluate the tensile strength of ductile iron strengthened at two different temperatures. She thinks that the lower temperature will yield the higher mean tensile strength. At each of the two temperatures,  $800^{\circ}C$  and  $1000^{\circ}C$ , 300 specimens of ductile iron were heat treated. The data consists of the tensile strengths from 300 specimens heated to  $800^{\circ}C$ :  $X_1, \dots, X_{300}$  which are iid with mean  $\mu_1$  and standard deviation  $\sigma_1$  and the tensile strengths from 300 specimens heated to  $1000^{\circ}C$ :  $Y_1, \dots, Y_{300}$  which are iid with mean  $\mu_2$  and standard deviation  $\sigma_2$ . Furthermore, the  $X$ 's and  $Y$ 's are independent.

- (1) The metallurgist is interested in the null hypothesis  $H_0 : \mu_1 \leq \mu_2$  versus the alternative hypothesis  $H_1 : \mu_1 > \mu_2$ . Use the following steps to present the customary  $t$ -test of this null hypothesis based on  $X_1, \dots, X_{300}$  and  $Y_1, \dots, Y_{300}$ .
  - i. Write down a general formula for the  $t$  test statistic commonly used for this hypothesis test.
  - ii. Write down the decision rule for this hypothesis test. Use  $\alpha = 0.05$ .
  - iii. State the necessary conditions needed for your procedure to be valid and how you would verify whether the conditions in are satisfied in this experimental setting.

For parts (2), (3), and (4) of this question, you may assume that  $\sigma_1 = \sigma_2 = 1$  and that the sample sizes are large enough to invoke the central limit theorem if necessary.

- (2) Calculate the power of your test for the following six values of the parameter:

$$\Delta = \frac{\mu_1 - \mu_2}{\sqrt{1/300 + 1/300}} = .5, 1.0, 1.5, 2.0, 2.5, 3$$

- (3) Use your power calculations in (2) to sketch a power curve for your test. Be sure to label your axes clearly.
- (4) The metallurgist in discussing your results from parts (1) through (3) states, "The power of the test when  $\Delta = 1.0$  is not large enough to meet industry standards. What needs to be done to increase it?" Answer the metallurgist's question, paying careful attention to: (i) your specific recommendation on how to increase the power; and (ii) explanation (based on the ideas from parts (1) through (3)) of why your recommendation will result in an increase in power.
- (5) The 600 observations considered above represent the tensile strength obtained from the two levels of heat treatment. However, after the experiments were conducted, the metallurgist informs you that the heat treatment for the 300 specimens for each heat level were conducted in the following manner. The furnace used to heat treat the specimens could hold only 5 specimens at a time. Thus, a tray containing 5 randomly selected specimens was heated to the specified temperature for the prescribed length of time and then the tensile strength measurements were taken on the 5 specimens. The metallurgist states that there is some variation in the temperature from one experimental run to the next. Thus, there may be a strong *positive* correlation between tensile strength readings for specimens on the same tray. Given this additional information, answer the following questions without carrying out any additional calculations.
  - i. How will this positive correlation within specimens affect the expectation of the variance estimator you used in part (1)?
  - ii. Suppose you did not adjust for the positive correlation within specimens and proceeded to use the ordinary  $t$ -test you proposed in part (1). Will the positive correlation in the data increase or decrease the numerical values of power you calculated for the test statistic in part (2)? Explain.
- (6) In light of your answer to (5), the metallurgist states, "Using the  $t$ -test from (1) to test the research hypothesis is obviously flawed. What is an alternative approach to testing the research hypothesis?" Answer the metallurgist's question by presenting a standard testing method that will account appropriately for the sampling design described in (5). Be sure to give clear, explicit statements of both your test statistic formula and your decision rule.

## **PROBLEM II.**

A colleague in meat science has approached you for help with an experiment she conducted. The experiment consisted of asking a sample of consumers to taste five different recipes for meat loaf. When a consumer tastes a sample, he or she will give scores to several characteristics of the meat loaf. These scores are combined into a single overall rating, called TASTE. Hence, there is one response variable for each consumer for each meat loaf product. The meat tasting literature indicates that in this type of study some consumers tend to give all recipes low scores, while other consumers tend to give all recipes high scores.

- (1) There are at least two possible experimental designs:

Design A: For a random sample of 100 consumers, 20 would be randomly assigned to taste Recipe R1, 20 randomly assigned to taste Recipe R2, and so on.

Design B: A random sample of 20 consumers would taste all five recipes. The recipes would be presented to the 20 consumers in random order.

Both designs result in a total of 100 data values, twenty for each recipe. Which design would you recommend? Provide an explanation of your choice of design.

- (2) The meat scientist selects Design B (perhaps contrary to your advice). She runs the experiment and collects the 100 data values. She asks if the correct model statement for PROC GLM in SAS is given by

**MODEL TASTE = CONSUMER RECIPE CONSUMER\*RECIPE;**

What would you tell her? Explain your answer.

- (3) When asked if there was any problems in running the experiment, she replies that one recipe smelled so bad she had eliminated it from the experiment. Is this a problem for conducting the proper analysis? Why or why not?
- (4) She also mentions that several consumers were unable to complete the tasting of all four remaining recipes. For some consumers she has only data for only one, two, or three recipes. She used the same model statement as given above in SAS. Are the results from the SAS output correct? Explain your answer.
- (5) Although the subjects use water to rinse their mouths between tasting the different recipes, there is some concern that a lingering flavor from tasting one recipe may influence the response obtained from the tasting of another recipe. Based on this information, would you suggest a different experimental design than the design you selected in (1)? If yes, give a brief description of the design. You do not need to be limited to having twenty observations per recipe, and you may assume that each consumer will complete his or her tasting assignment for your design.

### PROBLEM III.

This problem is about classical multiple linear regression; i.e.  $Y_{nx1} = X_{n \times p} \beta_{px1} + e_{nx1}$ . Here  $p = 5$  and miles per gallon is the dependent variable. The following table is part of the data:

mpg	engine	horse	weight	accel
10	360	215	4615	15
10	307	200	4376	15
11	318	210	4382	16
11	429	208	4633	15
11	400	150	4997	15
11	350	180	3664	16
12	383	180	4955	15
12	350	160	4456	15
12	429	198	4952	15
12	455	225	4951	15
12	400	167	4906	15
12	350	180	4499	15
13	400	170	4746	16
13	400	175	5140	15
13	350	165	4274	16
13	360	155	4602	15

Engine = Engine Displacement (cu. inches)

Horse = Horsepower

Weight = Vehicle Weight (lbs.)

Accel = Time to Accelerate from 0 to 60 mph (sec)

Assume that there is an intercept in the model. Answer the following questions:

- 1) Explain, in layperson's terms, what is meant by multicollinearity.
- 2) What does the determinant of  $X'X$  have to do with multicollinearity?
- 3) Given Tables 1 and 2 below, do you think that there is multicollinearity. Explain your answer?
- 4) Given Table 2 below, under the Column Dimension – see row 5, both the Constant Term and Time to Accelerate have Variance Proportion = 1 and the Condition Index = 168.934. Explain what that means?
- 5) Given Table 2 below, under the Column Dimension – see row 4, both the Engine and Weight have high Variance Proportions and the Condition Index = 33.48. Explain what that means?
- 6) Given Tables 3 and 4 below, do you think that there is multicollinearity. Explain your answer?

- 7) Given Table 4 below, under the Column Dimension – see row 5, neither the Constant Term and Time to Accelerate have Variance Proportion = 1 and the Condition Index = 7.235. Explain what that means?
- 8) Given Table 4 below, under the Column Dimension – see row 5, both the Engine and Weight have high Variance Proportions and the Condition Index = 7.235. Explain what that means?
- 9) Clearly Tables 1 & 2 are different from Tables 3 & 4. Which set of tables would you recommend to your client?

Table 1.

Model	Coefficients <sup>a</sup>						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	31.871	11.712	2.721	.007		
	Engine Displacement (cu. inches)	-.026	.008	-.393	-3.378	.001	.083 12.008
	Horsepower	.011	.013	.069	.882	.379	.183 5.452
	Vehicle Weight (lbs.)	-.004	.001	-.538	-5.585	.000	.121 8.241
	Time to Accelerate from 0 to 60 mph (sec)	.509	.751	.023	.678	.498	.986 1.015

a. Dependent Variable: Miles per Gallon

Table 2

Model	Dimension	Eigenvalue	Condition Index	Collinearity Diagnostics <sup>b</sup>				
				Variance Proportions				
				(Constant)	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
1	1	4.863	1.000	.00	.00	.00	.00	.00
	2	.119	6.386	.00	.03	.02	.00	.00
	3	.013	19.525	.00	.16	.93	.08	.00
	4	.004	33.480	.00	.81	.05	.92	.00
	5	.000	168.934	1.00	.00	.00	.00	1.00

a. Dependent Variable: Miles per Gallon

Table 3. Same data set except that Independent Variables have been Standardized

Model		Coefficients <sup>a</sup>						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	20.219	.215		93.972	.000		
	Zscore: Engine Displacement (cu. inches)	-2.519	.746	-.393	-3.378	.001	.083	12.008
	Zscore: Horsepower	.444	.503	.069	.882	.379	.183	5.452
	Zscore: Vehicle Weight (lbs.)	-3.441	.616	-.538	-5.585	.000	.121	8.241
	Zscore: Time to Accelerate from 0 to 60	.148	.218	.023	.678	.498	.986	1.015

a. Dependent Variable: Miles per Gallon

Table 4. Same data set except that Independent Variables have been Standardized

Model	Dimension	Eigenvalue	Condition Index	Collinearity Diagnostics <sup>a</sup>				
				(Constant)	Variance Proportions			
					Zscore: Engine Displacement (cu. inches)	Zscore: Horsepower	Zscore: Vehicle Weight (lbs.)	
1	1	2.815	1.000	.00	.01	.02	.01	.00
	2	1.005	1.674	.82	.00	.00	.00	.17
	3	.979	1.696	.18	.00	.00	.00	.82
	4	.148	4.358	.00	.02	.76	.30	.00
	5	.054	7.235	.00	.97	.22	.69	.00

a. Dependent Variable: Miles per Gallon

**METHODS QUALIFYING EXAM**

**JANUARY 2005**

**INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

**PROBLEM #1:**

Three different experimental situations are described below. Provide the information requested for each.

- A. A study is conducted to examine the effects of three different room temperatures and absence or presence of background noise on the performance of students taking an exam. A total of 48 students are available for the study, with the midterm exam score available for each student. Eight students are to be assigned to each of the six temperature by background noise combinations.
- (i) How would you suggest the students for each temperature by background noise combination be chosen?
  - (ii) The score on the exam, taken under the experimental conditions, is recorded for each student. Give the sources and degrees of freedom for the appropriate ANOVA table. Where appropriate, indicate the numerator and denominator mean squares for testing significance of an effect.
  - (iii) The exam consists of a multiple choice part, a short answer part, a "work a problem" part, and an essay question. Instead of a single exam score for each student, the scores on the individual parts of the exam are recorded separately. Does this change your ANOVA table? If so, provided the correct ANOVA table and F-ratios.

B. A research assistant to the president of a university collected data for a random sample on  $n = 100$  full time tenure track faculty members. The variables, and their values, are:

SALARY = nine month salary equivalent (in Dollars)

AGE = age, in years

RANK = Academic Rank = 1 for Professor

2 for Associate Professor

3 for Assistant Professor

4 for Instructor, Lecturer, etc.

TENURE = Tenure Status = 0 if not tenured and not tenure track

1 if not tenured, but tenure track

2 if tenured

DEGREE = Final Degree = 1 if bachelors

2 if masters

3 if postmaster, but not doctorate

4 if doctorate

SEX = sex of faculty member = 0 if male

1 if female

TIME = Length of time (in years) since initial appointment to faculty

at the university

The president has asked her assistant to determine if there is evidence of discrimination in salaries based on the sex of the faculty member. The research assistant has asked your help.

- (i) If you were to fit a model using SALARY as the response variable (the "Y"), which of the other variables would you consider to be class variables (i.e., ANOVA type variables) and which of the variables would you consider to be covariates (i.e., regression type variables)?
- (ii) The president asks if there is evidence the change in average salary for each additional year after initial appointment is not the same for males and females. How would you determine if there is evidence of this, i.e., what term or terms would you add to the model?
- (iii) Generally, faculty members at the rank of Associate Professor and Professor have tenure, faculty members at the rank of Assistant Professor do not have tenure but are tenure track, and faculty members at the rank of Instructor, Lecturer, etc., do not have tenure and are not tenure track. Could this create any difficulties in fitting a model with all the variables mentioned above included? If so, what is the problem?

C. An analysis of covariance was used to analyze data from an experiment on ten treatments in a Randomized Complete Block Design with five blocks and a covariate, X. Various models were fit to the data. The Error sums of squares (Residual sums of squares) for different models are shown below. The terms in the model are indicated by M = overall mean, BLK = class variable or dummy variables for blocks, TRT = class variable or dummy variables for treatments, and X for the covariate.

Model Includes	SS Error	Error DF
M	393.500	49
M, BLK	332.300	45
M, TRT	321.200	40
M, BLK, TRT	259.000	36
M, BLK, TRT, X	123.120	35
M, BLK, TRT, X(TRT)	82.226	26
M, BLK, TRT, X, X*TRT	82.226	26
M, BLK, X	197.599	44
M, TRT, X	139.928	39
M, X	214.637	48

- (i) Give the value of the F-ratio to test if the slopes for the ten lines are equal.
- (ii) Assuming a common slope, give the value of the F-ratio to test that the adjusted treatment means are equal.

**PROBLEM #2:** A programmer claims that  $Y_1, \dots, Y_{50}$ , is a random sample of size 50 from a distribution having cdf given by

$$F(y) = 1 - \exp(-(y - \theta)/\beta) \text{ if } y \geq \theta,$$

where  $\theta$  and  $\beta$  are known constants.

- (A) Describe a graphical method to evaluate the programmer's claim. Be sure to label your axes.
- (B) Describe a test of hypotheses to evaluate the programmer's claim.
- (C) Suppose  $\theta$  and  $\beta$  are unknown constants. Describe a graphical method to yield rough estimates of  $\theta$  and  $\beta$ ?
- (D) Explain why the Chi-squared Goodness-of-Fit would not be a good choice for your test of hypotheses in part (B)?
- (E) Under what circumstances, is the Chi-squared Goodness-of-Fit a good choice for a test statistic for testing the fit of a specified model?

**PROBLEM #3:**

These are regression problems. Assume all of the usual assumptions are met unless told otherwise. Please give complete answers to all questions.

- A. Your client comes to you with the following output and claims: "I am very happy that vehicle weight is more significant than horsepower since the sig value (p value) for weight is less than the sig value for horsepower."

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.8222 <sup>a</sup>	.675	.674	4.459

- a. Predictors: (Constant), Vehicle Weight (lbs.), Horsepower

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16085.855	2	8042.928	404.583	.000 <sup>a</sup>
	Residual	7733.138	389	19.880		
	Total	23818.993	391			

- a. Predictors: (Constant), Vehicle Weight (lbs.), Horsepower  
 b. Dependent Variable: Miles per Gallon

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Beta	t	Sig.
		B	Std. Error			
1	(Constant)	44.777	.825		54	.000
	Horsepower	-.061	.011	-.299	-5	.00000016227
	Vehicle Weight (lbs.)	-.005	.001	-.551	-10	.00000000000

- a. Dependent Variable: Miles per Gallon

In responding to her comment, provide in detail

- 1) the usual assumption for a regression analysis
- 2) the model being used.
- 3) the hypotheses being tested by Horsepower and VehicleWeight.
- 4) the reasons why, if you agree with her statements; or give the reasons why you disagree with her statements.

- B. Another client comes with the following output and asks: "What is my prediction equation?

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	396.985	3	132.328	21.516	.000 <sup>a</sup>
	Residual	98.405	16	6.150		
	Total	495.390	19			

a. Predictors: (Constant), x3, x2, x1

b. Dependent Variable: y

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Beta	t	Sig.
	B	Std. Error			
1	(Constant)	117.085	99.782	1.173	.258
	x1	4.334	3.016	1.437	.170
	x2	-2.857	2.582	-2.929	.285
	x3	-2.186	1.595	-1.561	.190

a. Dependent Variable: y

In answering his question, give:

- 1) the model used
- 2) the hypotheses being tested by x1, x2, and x3
- 3) the prediction equation
- 4) any other comments you might want to make to clarify what you are providing the client

**METHODS QUALIFYING EXAM**

**AUGUST 2005**

**INSTRUCTIONS:**

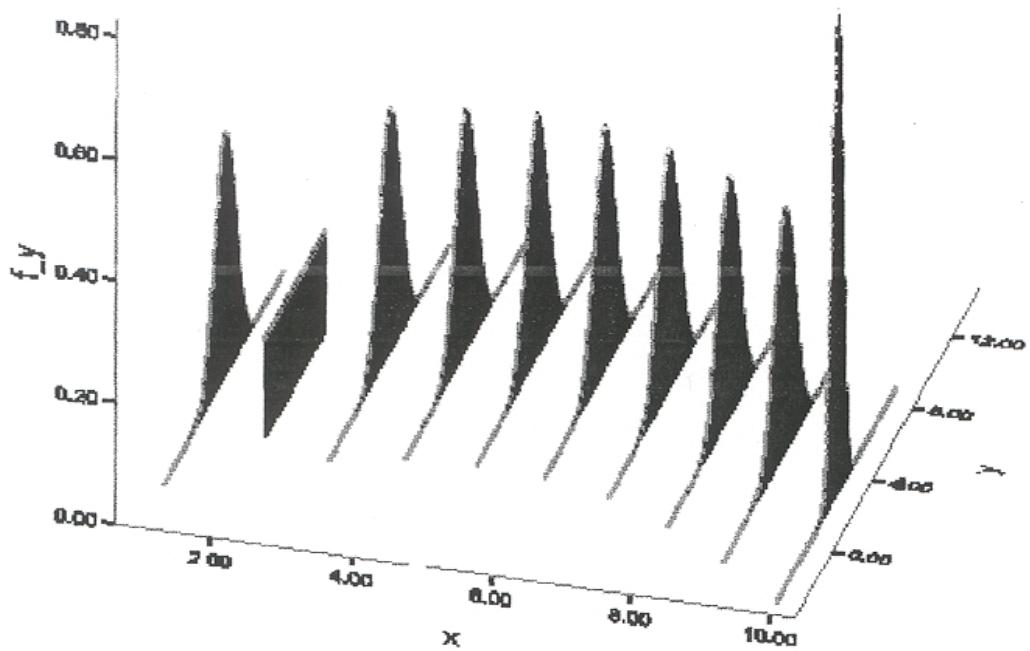
1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.
5. 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.

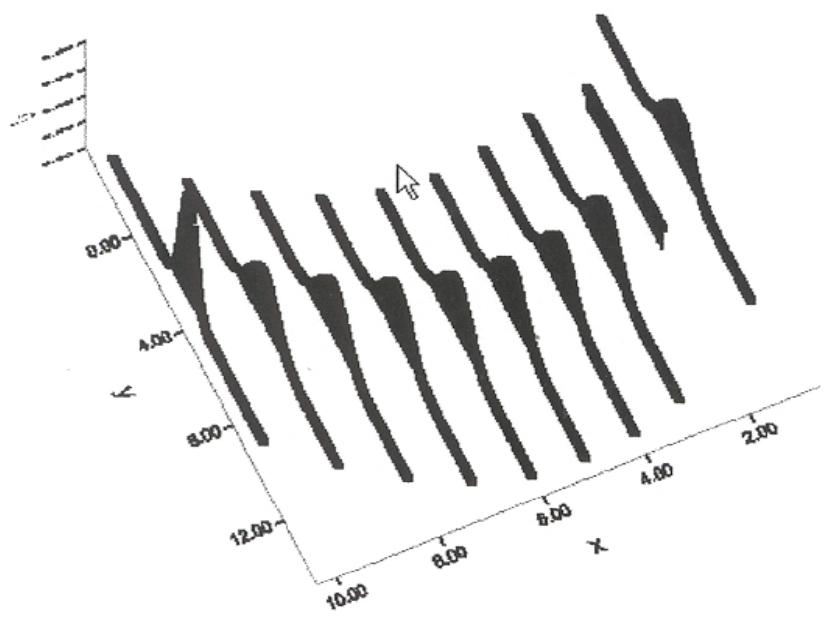
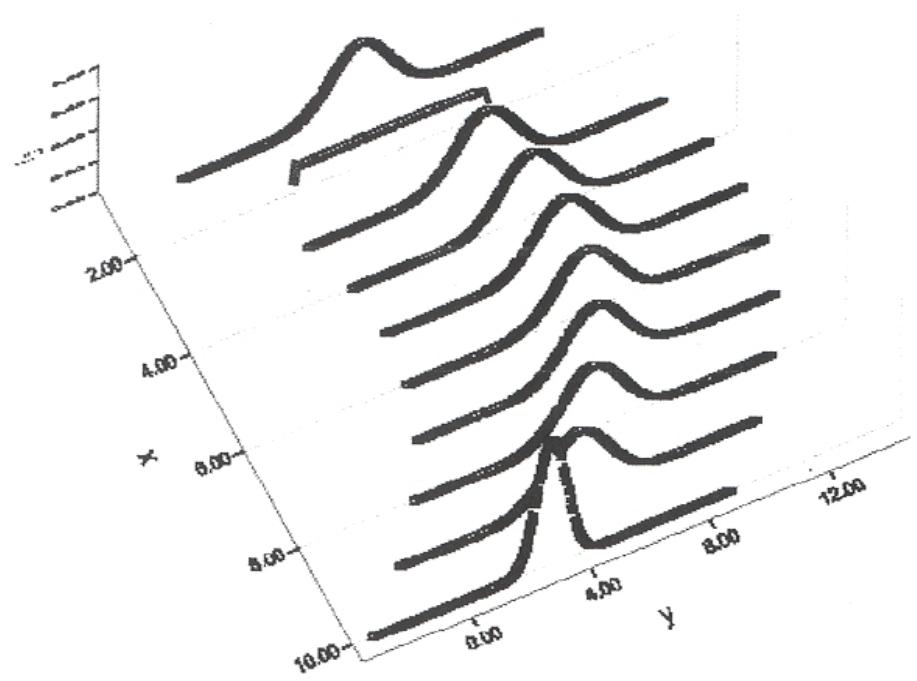
Problem 1. Regression Problem:

We have observed  $y$  = response (change in blood pressure) and  $x$  = dosage level of a drug. We assume a polynomial relationship between  $y$  and  $x$ .

The three graphs are all the same; but have been rotated to give additional views.

1. What is the appropriate model?
2. What are the usual regression assumptions?
3. Answer the following (in detail where appropriate):
  - a. Sketch  $E(y)$
  - b. Based on the graphs, make comments about the assumptions. Do they appear to be satisfied or violated?
  - c. How many populations are represented by the graphs?
  - d. For  $x = 10$ , sketch the pdf of  $y$ . Label the axis, approximate min and max values such that  $f_y > 0$ , indicate the mean and standard deviation.
  - e. List all of the parameters in the model
  - f. What is the effect of non-normal errors on hypothesis testing?
  - g. What is the effect of non-constant variance on hypothesis testing?





**Problem 2.**

This problem investigates the effect of correlated errors on least squares estimators. In order to keep the algebraic complexity to a minimum, we consider the simple linear (no-intercept) regression model:

$$Y_t = \beta X_t + u_t ,$$

where we assume that the disturbance  $u_t$  follows the first-order autoregressive scheme:

$$u_t = \rho u_{t-1} + \varepsilon_t ,$$

where  $|\rho| < 1$  and  $\varepsilon_t$  satisfies the assumptions:

$$\left. \begin{array}{l} E(\varepsilon_t) = 0 \\ E(\varepsilon_t \varepsilon_{t+s}) = \sigma_\varepsilon^2 \quad \text{if } s = 0 \\ = 0 \quad \text{if } s \neq 0 \end{array} \right\} \text{for all } t .$$

Assume that the independent variables,  $X_t$ , are random variables that we treat as “fixed” by conditioning on their observed values.

- a) Prove heuristically (i.e., show the algebra, but neither prove convergence nor justify interchange of limits rigorously) that:

$$u_t = \sum_{r=0}^{\infty} \rho^r \varepsilon_{t-r} ,$$

and, therefore, that:

$$E(u_t) = 0 .$$

- b) Using the representation for  $u_t$  given in part a), show (heuristically, as defined in part a)) that:

$$E(u_t u_{t-s}) = \rho^s \sigma_u^2 ,$$

where:

$$\sigma_u^2 = \frac{\sigma_\varepsilon^2}{1 - \rho^2} .$$

- c) Show that the least squares estimator (LSE) of  $\beta$  under this model is: (You may use without proof the matrix formula  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ , but define clearly the elements of  $\beta$ ,  $X$  and  $Y$ .)

$$\hat{\beta} = \frac{\sum_{t=1}^n X_t Y_t}{\sum_{t=1}^n X_t^2} .$$

- d) Conditioning on the  $X_t$ 's (i.e. treating them as "fixed"), show that:

$$Var(\hat{\beta}) = \left( \frac{\sigma_u^2}{\sum_{t=1}^n X_t^2} \right) \times \left( 1 + 2\rho \frac{\sum_{t=1}^{n-1} X_t X_{t+1}}{\sum_{t=1}^n X_t^2} + 2\rho^2 \frac{\sum_{t=1}^{n-2} X_t X_{t+2}}{\sum_{t=1}^n X_t^2} + \dots + 2\rho^{n-1} \frac{X_1 X_n}{\sum_{t=1}^n X_t^2} \right) .$$

- e) As a further algebraic simplification, suppose that the  $X_t$ 's also follow a first-order autoregressive scheme with parameter  $\rho$  identical to that of the  $u_t$ 's, so that for large  $n$ :

$$\frac{\sum_{t=1}^{n-s} X_t X_{t+s}}{\sum_{t=1}^n X_t^2} \doteq \rho^s .$$

Using this approximation (without proving it), show that:

$$Var(\hat{\beta}) \doteq \left( \frac{\sigma_u^2}{\sum_{t=1}^n X_t^2} \right) \times \left( \frac{1+\rho^2}{1-\rho^2} \right) .$$

Evaluate this approximate expression for the variance of the least LSE of  $\beta$  when  $\rho = 0.8$ . Comment on this effect of correlated errors *under this set of assumptions* on the LSE of the regression coefficient,  $\beta$ .

- f) Under the same assumptions used in part e) to derive the approximate expression for  $Var(\hat{\beta})$ , it can be shown that (but you need not show this):

$$E(s^2) \doteq \left( \frac{\sigma_u^2}{n-1} \right) \left( n - \frac{1+\rho^2}{1-\rho^2} \right) ,$$

where  $s^2$  is the usual LSE of variance, i.e.:

$$s^2 = \left( \frac{1}{n-1} \right) \sum_{t=1}^n (Y_t - \hat{\beta} X_t)^2 .$$

Evaluate this approximate expression for the expected value of the least LSE of variance when  $\rho = 0.8$ . Comment on this effect of correlated errors *under this set of assumptions* on the LSE of variance.

- g) Taken together, comment on the results of parts e) and f) as they pertain to least squares inferences concerning  $\beta$  *under this set of assumptions*.

Problem 3.

A graduate student in a Statistics Department has extended their advisor's robust estimation method. The student is in the process of putting a research proposal together. The advisor wants the student to compare the modified method to some well-known methods in a simulation experiment. The advisor tells the student that they are expected to use a well-designed experiment. The advisor's method takes about 1 minute to calculate on a sample size of about 100 and increases linearly with sample size. The method works on univariate Y data having model  $Y = X + \sigma\epsilon$ . Here the  $X$  and  $\epsilon$  are independent random variables. The parameter  $\sigma$  is measurement error.

The advisor tells the student that the following factors need to be included in the study. Those factors are sample size, error distributions (at least the normal, a heavy tailed distribution, and an asymmetric distribution), different levels of measurement error variance and the advisor's guess is that the MSE of all estimators for this problem will increase as a quadratic function of error variance, and will depend on the distribution of the true unobserved X values (use at least 3 or more types). The student is to compare their new method to three standard robust estimation methods, one of which is the sample median.

Assume that you are to design the experiment.

1. List at least two important considerations in choosing the levels of each of the following factors:
  - a. sample sizes
  - b. error distributions
  - c. measurement error variance
  - d. distribution for the underlying  $X$ .
2. What general class of experiment would you choose and why? That is, would you choose a fractional factorial, repeated measure, D-optimal design, etc.
3. How many sample sizes might you choose and why?
4. How many error variances would you choose and why?
5. How would you choose the number of replicates?
6. Once the experiment is completed, what method(s) do you expect to use to compare the results, and why?

Problem 4.

It is desired to estimate the weight of each of 4 objects on a balance. Let  $\beta_1, \beta_2, \beta_3, \beta_4$  be their actual weights. Weight measurements are subject to additive error, so that for any given object with true weight  $\beta$ , its measured weight  $Y$  is a random variable satisfying

$$Y = \beta + \epsilon$$

where  $E(\epsilon) = 0$  and  $Var(\epsilon) = \delta^2$ . Suppose that each object is weighed separately once, the first and second objects are weighed together once and the third and fourth objects are weighed together once as well, yielding measurements  $Y_1, Y_2, \dots, Y_6$  satisfying

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_1 + \beta_2 \\ \beta_3 + \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

where  $E(\epsilon_j) = 0$  and  $Var(\epsilon_j) = \delta^2$  and  $Cov(\epsilon_i, \epsilon_j) = 0$  for  $i \neq j$ .

- (a) Find the best linear unbiased estimator of  $\beta = (\beta_1, \dots, \beta_4)$ .
- (b) For the estimators  $\hat{\beta}_j, j = 1, \dots, 4$  of part (a) calculate  $Var(\hat{\beta}_j)$ .
- (c) Suppose that you were told that  $\beta_1 - \beta_2 = m_1$  and  $\beta_3 - \beta_4 = m_2$  where  $m_1$  and  $m_2$  are known quantities. Provide a linear unbiased estimator of  $\beta = (\beta_1, \dots, \beta_4)$  that incorporates this additional knowledge. What is the variance of the new estimators of  $\beta_j, j = 1, \dots, 4$ ?

## **METHODS QUALIFYING EXAM**

**JANUARY 2006**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

**Problem 1**

The National Institute for Standards and Technology conducted a study to develop standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and a section 3 mm in diameter was taken from the filter and mounted on a transmission electron microscope. An operator counted the number of asbestos fibers on the section. This procedure was repeated for 200 such samples. The 200 sections yielded the following counts: (the researcher no longer had the original counts – just the following grouped data and the mean of the 200 counts,  $\bar{Y} = (1/200)\sum_{i=1}^{200} Y_i = 4940/200 = 27.7$ )

	Grouped Counts							
	0–10	11–15	16–20	21–24	25–27	28–30	$\geq 31$	Total
$O_i$	2	1	36	52	50	39	20	200
$E_i$	0.12	2.43	34.62	57.51	45.36	32.62	27.34	200
$\frac{(O_i - E_i)^2}{E_i}$	30.25	0.85	0.06	0.53	0.48	1.25	1.97	35.39

- a) The consulting statistician computed the chi-square goodness of fit and obtained a  $p$ -value  $< 0.001$ , and then stated that the Poisson model provided an inadequate fit to the data. Do you agree with his results? If not, correct his computations and reassess the fit of the Poisson model.
- b) Assuming that the Poisson model provided a reasonable fit to the counts, construct a 95% confidence interval for the average number of asbestos fibers per 3 mm diameter area. (Hint: If  $Y_1, \dots, Y_n$  are iid  $\text{Poisson}(\lambda)$  random variables, then by the central limit theorem, the distribution of

$$\frac{\sqrt{n}(\bar{Y} - \lambda)}{\sqrt{\lambda}}$$

is approximately  $N(0, 1)$  for large  $n$ .

- c) Under the assumption that the Poisson model provides an adequate representation of the distribution of asbestos fibers, is there sufficient evidence ( $\alpha = 0.05$ ) that the average number of asbestos fibers per 3 mm diameter area is greater than 27?

(continued)

- d) What is the power of the test developed in c) if the true average number of asbestos fibers per  $3 \text{ mm}$  diameter area equals 28?
- e) Determine the sample size such that the test developed in b) will have a power of at least 0.90 when  $\lambda = 28$ .

**Problem 2**

This question tests your knowledge of fractional factorial designs.

- a) For a resolution III design, main effects are confounded only with 3<sup>rd</sup> and higher order interactions. True or False?
- b) For a resolution IV design, second order interactions are confounded with other 2<sup>nd</sup> order interactions. True or False?
- c) A  $2^{n-p}$  design that has resolution III requires how many generators?
- d) What resolution are Plackett-Burman designs, and under what class of designs do they fall?
- e) For an experiment that has eight (8) binary factors, what is the highest resolution you can get from a  $2^{(8-4)}$  design and how many generators do you need?

### Problem 3

Assume the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

with

$$\varepsilon_i \text{'s iid } N(0, \sigma^2).$$

Recall that:

- 1) If  $w$  is  $N(\mu, \sigma^2)$  and  $q$  is  $\chi^2(p)$  and  $w$  and  $q$  are independent, then  $t = \frac{(w/\sigma)}{\sqrt{q/p}}$  is distributed as  $t(p, \delta)$ , the non-central  $t$  with  $p$  degrees of freedom and non-centrality parameter  $\delta = \mu/\sigma$ .
- 2)  $\hat{\beta}_1$  is  $N\left[\beta_1, \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2\right]$ .
- 3)  $(n-2)s^2 / \sigma^2$  is  $\chi^2(n-2)$ .
- 4)  $\hat{\beta}_1$  and  $s^2$  are independent.

Use the above to develop a test for:

$$\begin{aligned} H_0 &: \beta_1 = c \\ H_A &: \beta_1 \neq c \end{aligned}$$

Give details and the critical region.

**Problem 4**

Let  $a_1 + b_1(x - \bar{x}_1)$  and  $a_2 + b_2(x - \bar{x}_2)$  be two regression lines estimated from independent samples of sizes  $n_1$  and  $n_2$ . Further let  $S_{yy}^{(1)}$ ,  $S_{xy}^{(1)}$  and  $S_{xx}^{(1)}$  be the corrected sum of squares and products for the first sample, with superscript (2) replacing (1) to represent the analogous quantities from the second sample. We wish to find a confidence interval for the value  $\xi$  of  $x$  at which the true regression functions intersect. Assume that the usual regression assumptions apply and define the quantity  $z$  to be:

$$z = a_1 + b_1(\xi - \bar{x}_1) - a_2 - b_2(\xi - \bar{x}_2)$$

- a) Show that  $E(z) = 0$ .
- b) Show that  $Var(z) = c\sigma^2$ , where

$$c = \frac{1}{n_1} + \frac{(\xi - \bar{x}_1)^2}{S_{xx}^{(1)}} + \frac{1}{n_2} + \frac{(\xi - \bar{x}_2)^2}{S_{xx}^{(2)}}.$$

- c) Show that  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ , where

$$\hat{\sigma}^2(n_1 + n_2 - 4) = S_{yy}^{(1)} - \frac{\left[S_{xy}^{(1)}\right]^2}{S_{xx}^{(1)}} + S_{yy}^{(2)} - \frac{\left[S_{xy}^{(2)}\right]^2}{S_{xx}^{(2)}}.$$

- d) Using the quantities  $z$  and  $c\hat{\sigma}^2$  defined above, form a quadratic (in  $\xi$ ) equation whose solution yields a  $100(1-\alpha)$  percent confidence region for  $\xi$ . (You do not need to solve this equation, rather just set up the equation that can be solved to form the confidence interval.) Carefully define any notation you introduce and justify any additional distributional results that you use to construct your confidence interval.

## **METHODS QUALIFYING EXAM**

**August 2006**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

### **Problem I.**

A researcher is developing a commercial shrimp farming operation. She has sought your help in designing and analyzing a study to investigate the influence of three factors on the growth rate of shrimp raised in aquaria. The three factors are:

T = Water Temperature ( $25^{\circ}C$ ,  $35^{\circ}C$ )

S = Water Salinity (10%, 25%, 40%)

D = Density of shrimp in the aquarium (2 shrimp/liter, 4 shrimp/liter)

The response variable is the four week weight gain on a per shrimp basis.

- A. Two possible experimental designs are listed below. For each design, discuss the advantages and disadvantages of the design. In addition, give a brief description of how you would assign the levels of the three factors, or combinations of factor levels to the experimental material. Suppose there are 36 aquaria available for the study. If it helps, you can sketch the experimental layout.
- D1. Each aquarium can be partitioned into two sections, but the water in the two sections is common to both sections (i.e., the water in one section circulates through the entire aquarium).
- D2. It is not possible to partition the aquaria into sections.

**For the following questions, assume that the experiment is run using Design D2:**

- B. The researcher asks if 36 aquaria will provide sufficient data points. How would you try to answer her question? Is there additional information needed? If so, what information?
- C. Write a model for the experiment and provide an AOV table with just the source of variation and degrees of freedom (**do not calculate any sum of squares**).
- D. The cell means for the experiment are given on the next page. Sketch the profile plot showing overall S\*D interaction and then sketch the profile plots showing the S\*D interactions separately for each level of Water Temperature.
- E. Using the sketches and assuming that the experimental data yielded  $MSE = 3280$ , which of the following main effects and interactions do you think are significant (**do not calculate any sum of squares**)? Provide justifications for your answers.
- i. T\*S\*D
  - ii. S\*D
  - iii. S

### Tables of Means

TEMP	DEN	SAL	$\bar{y}_{ijk.}$	$\bar{y}_{ij..}$
$25^{\circ}C$	2	10%	70	
		25%	465	
		40%	359	298
$25^{\circ}C$	4	10%	72	
		25%	333	
		40%	252	219
$35^{\circ}C$	2	10%	408	
		25%	276	
		40%	243	309
$35^{\circ}C$	4	10%	330	
		25%	312	
		40%	231	291

SAL	TEMP		$\bar{y}_{..k.}$
	$25^{\circ}C$	$35^{\circ}C$	
10%	71	369	220
25%	399	294	346.5
40%	305.5	237	271.25
$\bar{y}_{i...}$	258.5	300	

SAL	DEN		$\bar{y}_{..k.}$
	2	4	
10%	239	201	220
25%	370.5	322.5	346.5
40%	301	241.5	271.25
$\bar{y}_{j..}$	303.5	255	

**Problem 2**

- (a) A drug is tested at four equally spaced doses on a total of 20 patients, with 5 patients in each dose group. The response of interest for the  $i^{\text{th}}$  patient at the  $j^{\text{th}}$  dose is  $Y_{ij}$ ,  $i = 1, \dots, 5$ ;  $j = 1, \dots, 4$ . The responses  $Y_{ij}$  are independent and normally distributed with common variance  $\sigma^2$ . Use a sum of squares partition in ANOVA to test for linear, quadratic and cubic trends. (HINT: The relevant orthogonal polynomials lead to these contrast vectors:  $(-3, -1, 1, 3)$ ,  $(1, -1, -1, 1)$  and  $(-1, 3, -3, 1)$ ).
- (b) We can also approach this problem using ordinary least squares regression. What design matrix would you use to estimate regression parameters associated with linear, quadratic and cubic trends? Describe a test statistic and its null distribution to test for linear trends in a model that includes an intercept and covariates for linear, quadratic and cubic trends.
- (c) Suppose we have two drugs, A and B, tested at the same doses as in part (a) and we have two sets of responses,  $Y_{ij}$  and  $Z_{ij}$ ,  $i = 1, \dots, 5$ ;  $j = 1, \dots, 4$ , corresponding to the two drugs. Assume the  $Y_{ij}$  and  $Z_{ij}$  are independently and normally distributed with common variance  $\sigma^2$ . Describe a test statistic and its null distribution for testing the null hypothesis that the effects of the two drugs follow the same linear trend.

### Problem 3

A book<sup>1</sup> on robust statistical methods published in June 2006 considers regression models for a data set taken from Jalali-Heravi and Knouz (2002, Electronic Journal of Molecular Design, 1, 410-417). The aim of the modeling is to predict a physical property of chemical compounds called the Krafft point based on four potential predictor variables using a data set of size  $n=32$ . According to Maronna, Martin and Yohai (2006, p. 380):

The Krafft point is an important physical characteristic of the compounds called surfactants, establishing the minimum temperature at which a surfactant can be used.

The authors of the original paper sought to find a regression model to predict:

$$Y = \text{Krafft Point (KPOINT)}$$

from

$x_1$  = Randic Index (RA)

$x_2$  = Heat of formation (HEAT)

$x_3$  = Reciprocal of volume of the tail of the molecule (VTINV)

$x_4$  = Reciprocal of Dipole Moment (DIPINV)

The first model considered by Jalali-Heravi and Knouz (2002) was

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \quad (1)$$

Output from model (1) appears on the following pages.

- a) Decide whether (1) is a valid model. Give reasons to support your answer.
- b) The plots of standardized residuals against RA and VTINV produce curved patterns. Describe what, if anything can be learnt about model (1) from these plots. Give a reason to support your answer.
- c) Jalali-Heravi and Knouz (2002) give "four criteria of correlation coefficient ( $r$ ), standard deviation ( $s$ ), F value for the statistical significance of the model and the ratio of the number of observations to the number of descriptors in the equation" for choosing between competing regression models. Provide a detailed critique of this suggestion.

<sup>1</sup>Maronna, R.A., Martin, R.D. & Yohai, V.I. (2006) *Robust Statistics*. Wiley, New York

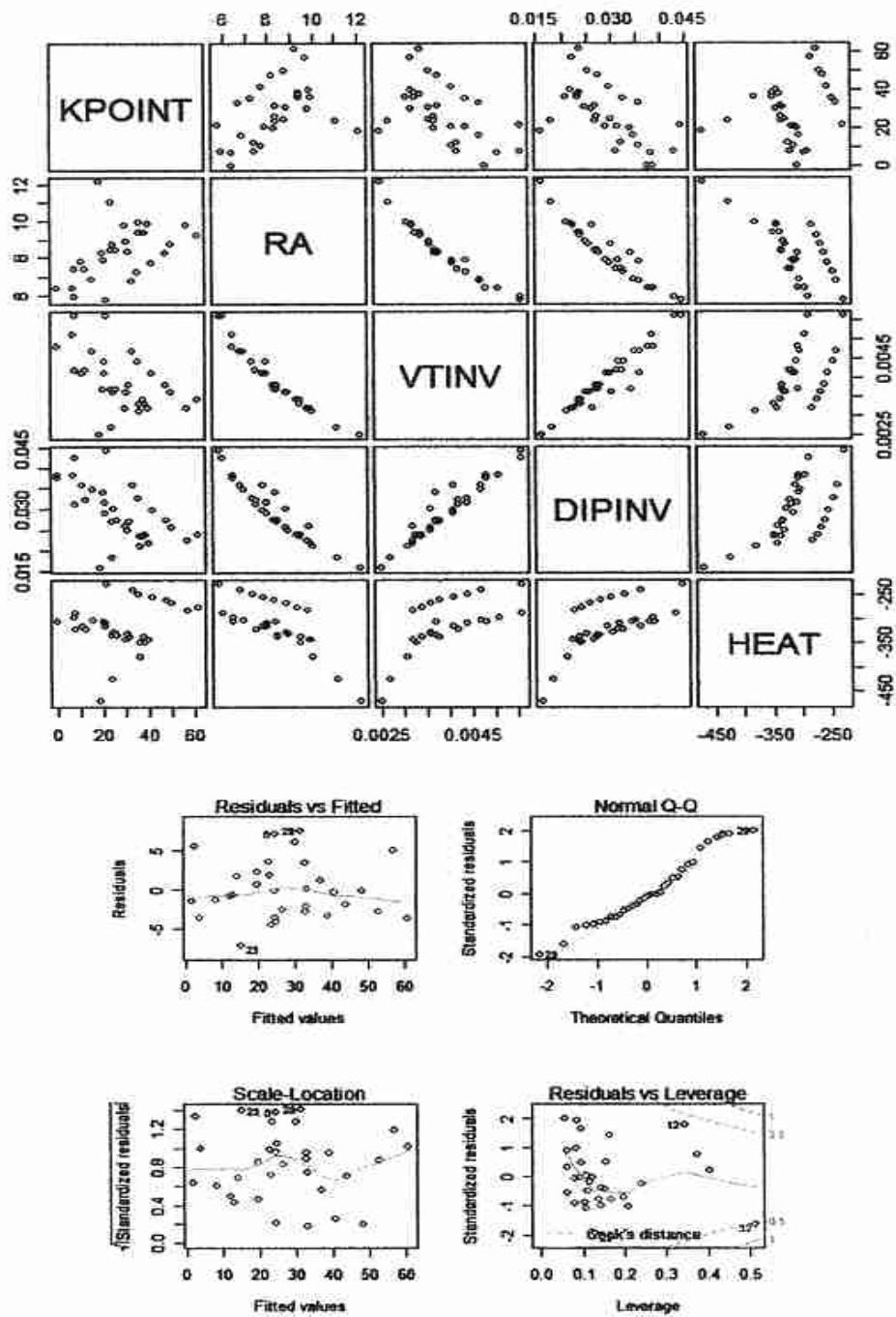
The second model considered by Jalali-Heravi and Knouz (2002) was:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + e \quad (2)$$

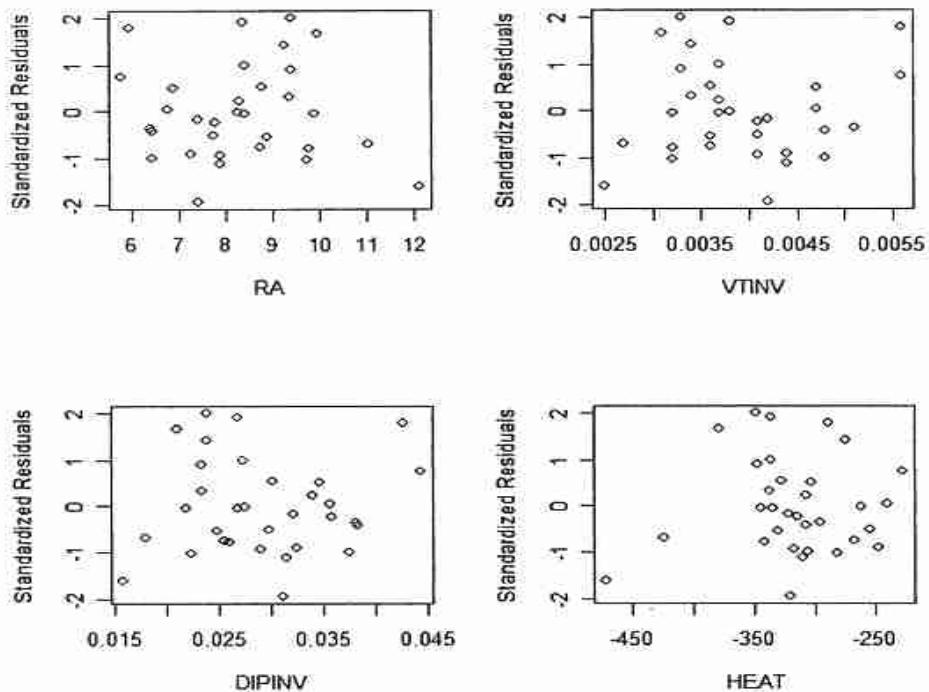
Output from model (2) appears on the following pages.

- d) Decide whether (2) is a valid model. Give reasons to support your answer.
- e) Jalali-Heravi and Knouz (2002) "believe that model (2) is superior to model (1)" since it produces a lower value of AIC, and since VTINV shows a "high correlation with" RA and DIPINV" and they recommend this model or use in practice. Do you agree with their conclusion and recommendation? Give reasons to support your answer.
- f) A statistics professor from Texas A&M University discovered while reading the paper by Jalali-Heravi and Knouz (2002) that there are 4 groups in the data set corresponding to four different surfactants. The last page contains a scatter plot matrix of the data with the four different groups marked by different plotting symbols. Also given in each plot are the least squares fits for each group. Explain carefully the steps you would take to obtain a final model allowing for the different groups.

## Output from model (1)



### Output from model (1)



### Output from R

```

Call:
lm(formula = KPOINT ~ RA + VTINV + DIPINV + HEAT)

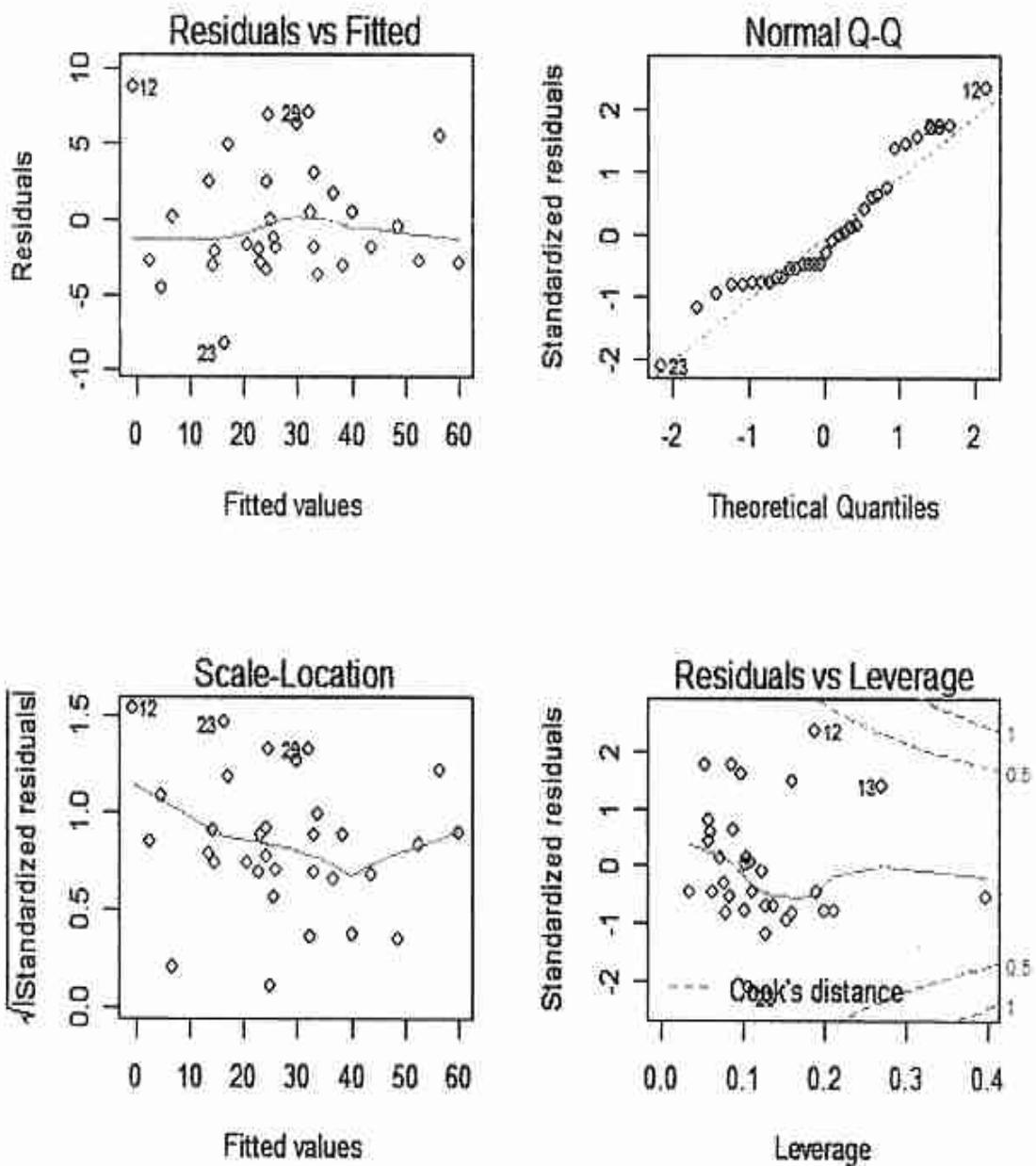
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.031e+01 3.368e+01 2.088 0.046369 *
RA          1.047e+01 2.418e+00 4.331 0.000184 ***
VTINV       9.038e+03 4.409e+03 2.050 0.050217 .
DIPINV      -1.826e+03 3.765e+02 -4.850 4.56e-05 ***
HEAT        3.550e-01 2.176e-02 16.312 1.66e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.919 on 27 degrees of freedom
Multiple R-Squared:  0.9446,    Adjusted R-squared:  0.9363
F-statistic: 115 on 4 and 27 DF,  p-value: < 2.2e-16

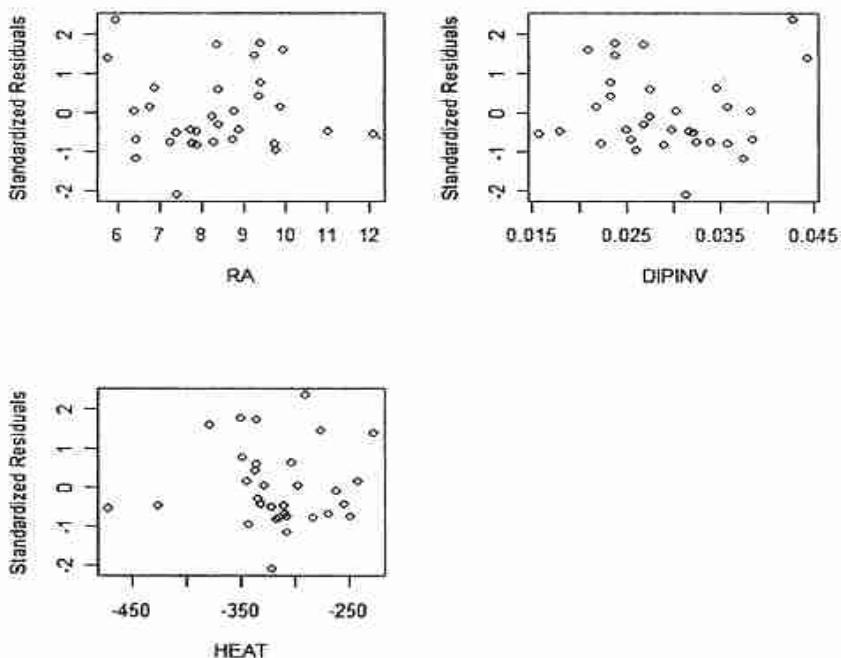
vif(m1)
      RA       VTINV      DIPINV      HEAT
25.792770 22.834190 13.621363  2.389645

```

## Output from model (2)



## Output from model (2)



## Output from R

Call:

```
lm(formula = KPOINT ~ RA + DIPINV + HEAT)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.304	-2.762	-1.491	2.408	8.775

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.214e+02	2.388e+01	5.086	2.19e-05 ***
RA	7.147e+00	1.893e+00	3.776	0.000764 ***
DIPINV	-1.508e+03	3.621e+02	-4.165	0.000270 ***
HEAT	3.465e-01	2.255e-02	15.364	3.58e-15 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.137 on 28 degrees of freedom

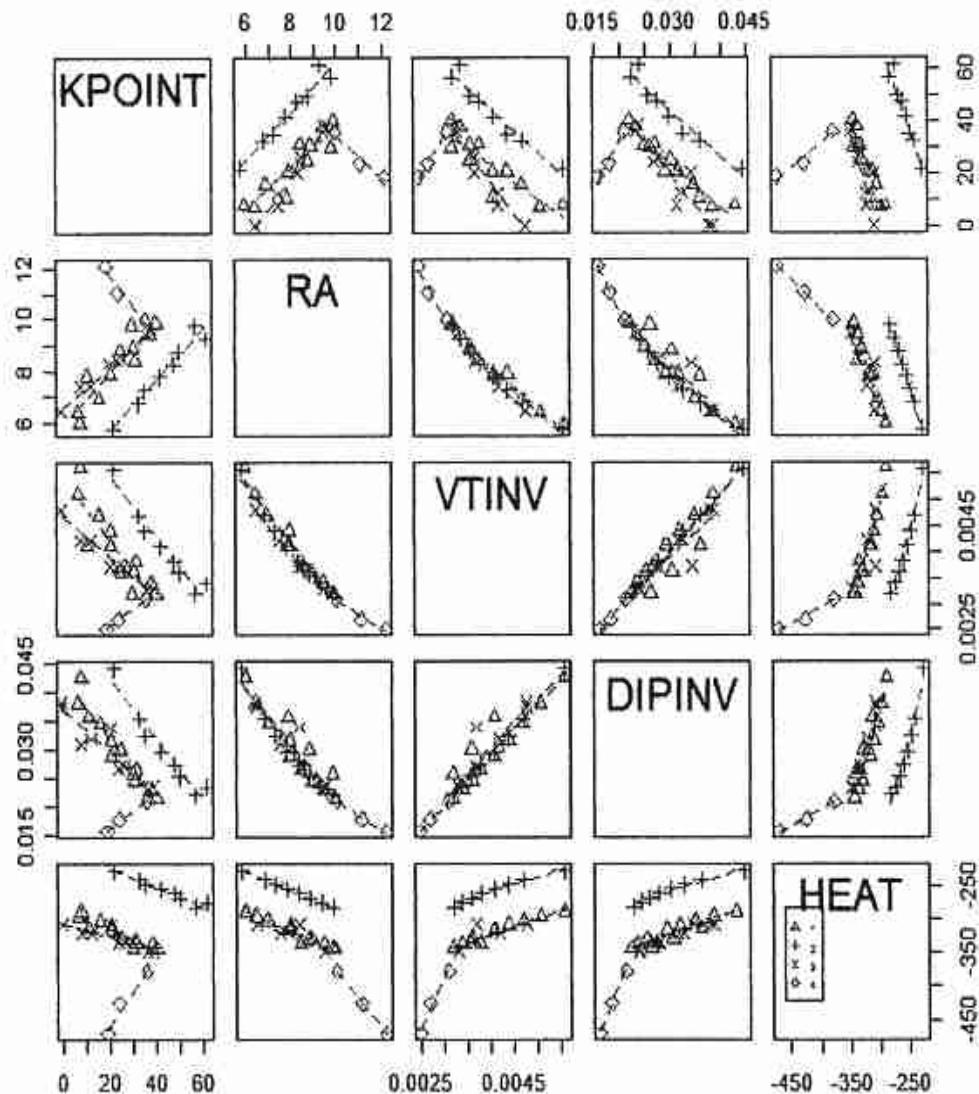
Multiple R-Squared: 0.9359, Adjusted R-squared: 0.9291

F-statistic: 136.4 on 3 and 28 DF, p-value: < 2.2e-16

vif(m2)

RA	DIPINV	HEAT
14.178611	11.304448	2.302705

Scatter plot matrix of the data with the four different groups marked by different plotting symbols



#### **Problem 4**

Background:

The following questions have been posted to a discussion board. Assume that you are the statistical consultant. Please provide answers to the client's questions.

A.

*Hello everybody.*

*I am doing a regression analysis. I made the K-S test on the dependent variable and saw that it is not normally distributed (Asymp. Sig. = 0.000). Is there a data transformation or some magical incantation that exists somewhere to transform this data into a normally distributed set?*

*Bye,*

*Marc.*

B.

*Hi all:*

*I have question (a couple maybe) about the GLM procedure. I have run a simply two-way ANOVA, but the cell sizes are unequal (i.e., I have an unbalanced design). When I run the basic procedure (and ask for descriptive statistics) I get weighted and unweighted marginal means. This means are fairly different.*

*My first question is: to which of these means do the statistical tests for the main effects in the ANOVA Summary table correspond to, weighted or unweighted?*

*My second question is: how do I get an ANOVA summary table for the "other" set of means given that the default is either weighted or unweighted?*

*Kris*

(Note: To answer Kris, give the model using the “cell means” model and give the null hypothesis for testing “main effect A.”)

C.

Dear All,

Here is my question:

I have three Groups (Experimental 1, Experimental 2, Control); a dependent variable Y, and a covariate X in my data file. I am running GLM treating Group as a fixed effect and using X as the covariate. I got the following output. It seems to me that since the p-value of X is greater than .05, there is no effect to the covariate X. Is this correct? If not, what is being tested by the term X?

Tests of Between-Subjects Effects

Dependent Variable: y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2142.958 <sup>a</sup>	5	428.592	587.859	.000
Intercept	10.840	1	10.840	14.869	.002
group	1.594	2	.797	1.093	.362
x	.182	1	.182	.249	.625
group * x	131.341	2	65.671	90.074	.000
Error	10.207	14	.729		
Total	2257.412	20			
Corrected Total	2153.165	19			

a. R Squared = .995 (Adjusted R Squared = .994)

Best,

Enis

D.

Hi Group,

My boss asked me to explain what is meant by “testing for equal variances” in regression. He learned that “equal variances” was an important assumption in regression analysis. I told him that we were testing a null hypothesis which is:

$H_0: \text{variance} = \text{a constant}$  (this value comes from the MSE from the ANOVA table).

He said “No Way. The term “testing for equal variances” means more than one variance.” Am I correct and, if not, he wants to know how many variances we are testing?

Please help.

Mary

## **METHODS QUALIFYING EXAM**

**January 2007**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

### Problem I.

A researcher was interested in studying the effect of the distance of an object from the eye on the eye focus time. There are  $t$  different distances of interest. The researcher has  $r$  subjects available for the experiment. Because there may be differences among subjects, she decides to conduct the experiment in a randomized block design, that is, let  $y_{ij}$  denote the focus time from the  $j$ th subject using the  $i$ th distance. The following model was used for this data: For  $i = 1, \dots, t$  and  $j = 1, \dots, r$

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}$$

where  $\mu, \tau_1, \dots, \tau_t$  are unknown parameters;  $b_1, \dots, b_r$  are iid  $N(0, \sigma_b^2)$  random variables;  $e_{ij}(i = 1, \dots, t; j = 1, \dots, r)$  are iid  $N(0, \sigma_e^2)$  random variables; all  $b_j$  and  $e_{ij}$  are jointly independent; and  $\sigma_b^2$  and  $\sigma_e^2$  are unknown positive variances.

- a) In the following AOV table, fill in the formulas for the missing elements in the df and SS columns:

Source of Var.	D.F.	Sum of Squares
Subject	_____	_____
Distance	_____	_____
Error	_____	_____
Total	n-1	$\sum_{i=1}^t \sum_{j=1}^r (y_{ij} - \bar{y}_{..})^2$

- b) What is the expected MS for the Block term?  
 c) Using your expression from part b), provide an estimator of  $\sigma_b^2$ .  
 d) Provide an expression for the test statistic for testing  $H_0 : \tau_1 = \dots = \tau_t$ .  
 e) The researcher in fact was also interested in the impact of Age of the subject on the eye focus time. The age variable was divided into  $m$  groups. Let  $y_{ijk}$  denote the focus time from the  $j$ th subject from the  $k$ th age group using the  $i$ th distance. The following model was used for this data: For  $i = 1, \dots, t$ ;  $j = 1, \dots, r$ ; and  $k = 1, \dots, m$ ;

$$y_{ijk} = \mu_{ik} + b_j + e_{ijk}$$

where  $\mu_{ik}(i = 1, \dots, t; k = 1, \dots, m)$  are unknown parameters;  $b_1, \dots, b_r$  are iid  $N(0, \sigma_b^2)$  random variables;  $e_{ijk}(i = 1, \dots, t; j = 1, \dots, r)$  are iid  $N(0, \sigma_e^2)$  random variables; all  $b_j$  and  $e_{ijk}$  are jointly independent; and  $\sigma_b^2$  and  $\sigma_e^2$  are unknown positive variances. Under the above model, answer the following questions:

- 1) Provide an expression for the correlation between the eye focus times for a subject in Age Group 1 viewing the object at Distances 2 and 3.
- 2) The researcher was very interested in determining if there was an interaction between Age and Distance. In terms of the model parameters given above, state the null hypothesis for the test of interaction between Age and Distance.

## PROBLEM II.

The American Red Cross does a detailed analysis every month on 100 randomly selected blood samples from donors in Texas. The samples produced a measurement,

$$X_i = \text{Serum lead level (in micrograms per deciliter) for sample } i.$$

From past experience, the distribution of  $X_i$  is known to be lognormal. Thus, a statistician defined the transformed variable  $Y_i = \ln(X_i)$ . For purposes of our analysis, we will assume that

$$Y_i = \mu + e_i \tag{1}$$

where  $\mu$  is a fixed mean parameter for blood donors in Texas and  $e_i, i = 1, \dots, 100$  are independent and identically distributed  $N(0, \sigma_e^2)$  random variables.

- a) Using the above model, provide an expression for a 95% confidence interval for  $\mu$ .
- b) Now consider a new observation  $Y_{n+1}$  that also satisfies model (1). Provide an expression for a 95% *prediction interval* for  $Y_{n+1}$ , based on the old data  $Y_i, i = 1, \dots, 100$ .
- c) Give a customary interpretation of your prediction interval in b), paying special attention to: i) what is fixed; ii) what is random; and iii) to what probability does the term “95%” refer?
- d) Explain how the interpretation of your prediction interval from c) differs from the customary interpretation of the confidence interval in a).
- e) Recall that our transformed observations  $Y_i = \ln(X_i)$  were recorded on a log-transformed scale. Using the results from the preceding steps to provide an expression for a 95% prediction interval for a new observation  $X_{n+1}$  on the *original* (untransformed) scale.
- f) Explain why the procedure you used in e) to obtain the 95% prediction interval for  $X_{n+1}$  would not yield an **exact** 95% confidence interval for  $E(X_i)$ .
- g) Describe an alternative method for finding a 95% confidence interval for  $E(X_i)$ .
- h) Suppose the Red Cross wants to compare the mean serum lead levels for the previous 12 months.
  - 1) How would you adjust the individual monthly C.I.s in order to obtain a simultaneous coverage probability of 95% for the 12 monthly C.I.s?
  - 2) What would be the impact on the average widths of the monthly C.I.s?
  - 3) Describe how the simultaneous monthly C.I.s could be used to determine which pairs of months had different means.
- i) The prediction interval in part b) is based on the assumption that the  $Y_i$  values are normally distributed. List *two* specific methods that you could use to check this assumption based on the transformed observations. For *each* of these two methods, give explicit rules (omitting numerical critical values) you would follow to decide whether the observations are consistent with the normal-distribution assumption.

### Problem III

A scientist wants to study the effect of jazz music on intelligence in rats. She uses 4 female rats from a single birth episode. Two of the rats are raised in a quiet laboratory and the other two rats are raised under the same conditions, except that they listen to Thelonious Monk on the piano for 6 hours each day. At maturity, each rat runs through a maze, the times (in minutes) for the rats to run the maze are  $y_1 = 5$  and  $y_2 = 7$  (quiet), and  $y_3 = 1$  and  $y_4 = 9$  (Monk piano). The scientist proposed the following model for analyzing the data:

$$Y = X\beta + \epsilon,$$

where the elements in  $Y$  are the times,  $\epsilon$  is a random vector with  $E(\epsilon) = 0$  and  $Var(\epsilon) = \sigma^2 I$ ,  $\beta = (\beta_0, \beta_1)^T$  and

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Give numerical answers to the following questions whenever possible.

- a) Explain in one sentence each interpretations of the parameters  $\beta_0$  and  $\beta_1$ .
- b) Determine the least squares estimate of  $\beta$ ,  $\hat{\beta}$ .
- c) Determine  $Var(\hat{\beta})$ .
- d) Determine the “hat” matrix,  $H$ .
- e) Give an unbiased estimate of  $\beta_0 + \beta_1$ .
- f) What additional assumptions are needed so the estimate in e) is the best linear unbiased estimate? Why?
- g) Suppose there is a fifth female rat from the same litter and suppose that she was raised in a quiet lab. Predict her time,  $\hat{y}_5$ , to run the maze on her first try at maturity.
- h) Estimate  $\sigma^2$  unbiasedly using the original four observations.
- i) Suppose you had only the 2 rats in the quiet group, how would you then estimate  $\sigma^2$  using these two observations?
- j) Based on the answers to h) and i) what assumption(s) may be violated?
- k) Estimate the variance of the estimate in e) based on the original four observations.
- l) Estimate  $Var(\hat{y}_5 - y_5)$  where  $\hat{y}_5$  is defined in g).
- m) After correcting for the mean, what proportion of the total variability in  $y$  is explained by the model?

## **PROBLEM IV.**

Below is a message from Nicki to the SPSS discussion board asking for help. Following her message are suggestions from an anonymous respondent, Stephen.

After reading Nicki's questions and Stephen's suggestions, answer the following questions a), b), and c). Give reasons for all of your answers. Be very critical of Stephen's suggestions. Do they make sense? Would you recommend doing what he suggests?

- a) For each of Stephen's suggestions i), ii), and iii); do you agree or disagree with Stephen's point? Give a brief discussion of why you agree or disagree.
- b) Explain to Nicki what it means to have a significant interaction term between the fixed factor and the continuous covariate.
  - 1) give the model in terms she can understand;
  - 2) sketch a graph to explain what a significant interaction means;
  - 3) explain to her in terms of the model what a significant interaction means.
- c) Explain what is being tested by the pairwise comparisons. Answer in terms of the model.

### **Nicki's Questions:**

I ran an ANCOVA model which yielded a significant interaction between a fixed factor (having four levels) and a continuous covariate. I am interested in investigating the interaction further. I have the additional problem that the pairwise comparisons were not significant. So I am wondering if there are any other statistical methods to investigate an interaction between a continuous covariate and a factor.

### **Stephen's Suggestions:**

- i) One of the core assumptions of ANCOVA is that the relationship between the covariate and the dependent variable is the same for both groups. When you have a significant group by covariate interaction, then this assumption is not met.
- ii) As a result, I would not use ANCOVA to analyze these data. I would use a regression framework instead, with the following predictors: 1) the continuous variable; 2) dummy codes for the categorical; 3) a cross-product, or interaction term, between the continuous variable and each of the dummy codes.
- iii) You can use the regression coefficients to compute values of the dependent variable for each group at one standard deviation above and below the mean on the continuous variable. These points will allow you to plot the slope and intercept of the continuous variable for each group.

## **METHODS QUALIFYING EXAM**

**August 2007**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

### Problem 1

A biologist conducted an experiment to investigate the impact of a photosynthesis-inhibiting herbicide on the plankton level in ponds. Water was obtained from three ponds. From each of the three ponds, four containers holding eight (8) gallons of well-mixed pond water were obtained. The four containers were randomly assigned to be dosed with one of the following rates of herbicide: 0, 0.1, 0.5, 1.0 mg/liter. The large containers were divided into eight 1-gallon glass jugs filled with the well-mixed, dosed pond water, sealed and suspended in the pond just below the water surface. Bottles were given labels so that two bottles of each dose could be removed at the end of the first day, at the end of the first week, the second week, and finally the third week. Rotifers are a major element of the plankton food chain in this pond. The number of rotifers present in the bottle was counted immediately after the bottle was removed from the pond. The counts in numbers of rotifers/liter are given.

		Pond 1				Pond 2				Pond 3			
		Week				Week				Week			
Dose	Bottle	0	1	2	3	0	1	2	3	0	1	2	3
0	1	6718	5166	3815	2340	6222	4656	3351	1804	7081	5466	4151	2604
	2	6392	5039	3382	2881	5891	4527	2828	2318	6629	5393	3628	3118
0.1	1	5832	5233	4373	2453	5323	4722	3837	1935	6123	5533	4637	2735
	2	5569	4350	3333	2924	5071	3849	2833	2442	5896	4605	3633	3242
0.5	1	5924	3953	4912	2575	5423	3435	4421	2057	6242	4235	5221	2857
	2	6715	4295	4427	4211	6205	3759	3972	3711	7051	4559	4772	4511
1.0	1	6821	4849	4900	4451	6316	4394	4400	3915	7112	5194	5200	4715
	2	5871	4763	4067	4552	5368	4236	3576	4025	6117	5036	4376	4825

Answer each of the following questions.

- Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, SUBSAMPLING, etc.
- Type of Treatment Structure, for example, Single Factor, Crossed, Nested, etc.
- Identify each of the factors as being Fixed or Random.
- Describe the Experimental Units and/or Measurement Units.
- Provide a partial AOV Table with Source of Variation, Degrees of Freedom, and Expected Mean Squares.
- Let  $y_{ijkl}$  be the count from the  $l^{\text{th}}$  bottle taken on the  $k^{\text{th}}$  week from the  $i^{\text{th}}$  dose from the  $j^{\text{th}}$  pond. Express the AOV-MOM estimate of the variance of  $(\bar{y}_{1...} - \bar{y}_{2...})$  in terms of the means squares from the AOV given in part e).
- Describe how you would obtain an estimate of the degrees of freedom of your estimate in f).
- The researcher wants to compare the mean counts for the four doses.
  - Under what relationship between the factors dose and week would these comparisons be valid?
  - Provide the form of Tukey's HSD for making these comparisons. Make sure to identify all the terms in your formula.

### Problem 2

The rate ( $R$ ) of a metabolic reaction in humans is thought to be related to body temperature ( $t$ ) by the following equation:

$$R_i = \beta_0 + \beta_1 t_i + \varepsilon_i \quad (\text{I}),$$

where  $\beta_0$  and  $\beta_1$  are parameters,  $t_i$  is the body temperature for the  $i^{\text{th}}$  individual observed, and  $\varepsilon_i$  are assumed to be independent  $N(0, \sigma^2)$  random variables.

At the standard body temperature of 98.6 °F, thousands of observations have been made. Therefore, the expected value of  $R$  at  $t^* = 98.6$  °F is assumed to be known as  $R^* = 73.8$  moles per minute, and the variance of  $R$  is known to be  $(0.08 \text{ moles/minute})^2$ . We observe the rate of reaction in 14 unhealthy individuals (i.e., individuals with elevated or depressed body temperature.) The following data are reported:

Temperature	Number of Individuals Observed	Mean Reaction Rate
103.8	2	84.2
100.1	3	75.9
97.4	4	71.3
96.9	5	70.5

- a. Show that it suffices to consider the model:

$$Y_i = \beta x_i + \varepsilon_i \quad (\text{II})$$

where  $Y_i = R_i - R^*$ ,  $x_i = t_i - t^*$ , and  $\beta = \beta_1$ .

- b. Calculate the least squares estimate of  $\beta$ .
- c. Test the hypothesis that  $\beta = 0$  against the two sided alternative at the .01 level. Give the test statistic and the rejection rule (in terms of a commonly tabulated distribution.)
- d. Test model (II) for lack of fit to the data at the .01 level. List the test statistic and the rejection rule (in terms of a commonly tabulated distribution.)
- e. Can it occur that we reject the null hypothesis in both parts c) and d)? Explain briefly.
- f. Do you suspect any of the observations are high leverage points? Why or why not?

### Problem 3

- a. A statistics professor at Texas A&M University is involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots which each show an estimated regression line and associated individual 95% confidence intervals for the regression function at each  $x$  value, as well as the observed data. A referee has asked the following question:

"I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures."

Prepare a response to this statement.

- b. The Sunday April 15, 2007 issue of the *Houston Chronicle* included a section devoted to real estate prices in Houston. In particular, data are presented on the 2006 median price per square foot for 1922 subdivisions. Interest centers on developing a regression model to predict

$Y_i$  = 2006 median price per square foot

from

$x_{1i}$  = %NewHomes (i.e., of the houses that sold in 2006, the percentage that were built in 2005 or 2006)

$x_{2i}$  = %Foreclosures (i.e., of the houses that sold in 2006, the percentage that were identified as foreclosures)

for the  $i = 1, \dots, 1922$  subdivisions.

The first model considered was

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (1)$$

Model (1) was fit using weighted least squares with weights

$$w_i = n_i$$

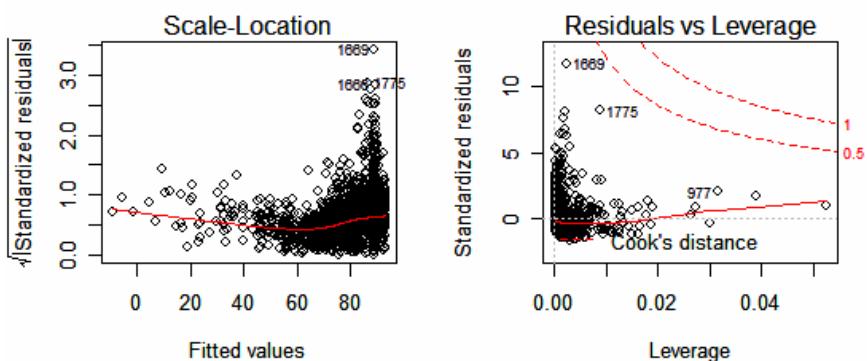
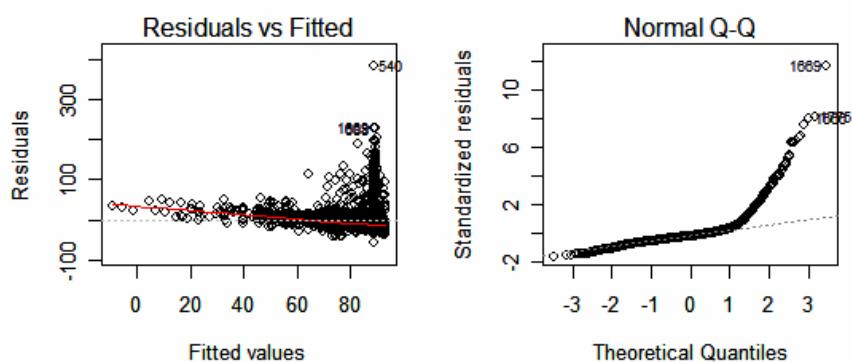
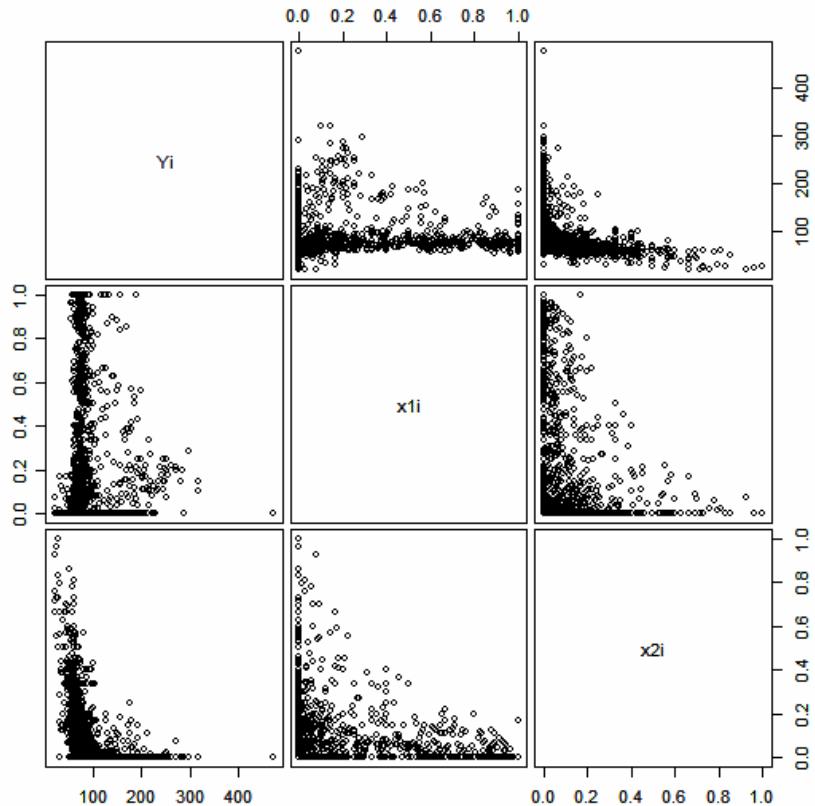
where

$n_i$  = the number of homes sold in subdivision  $i$  in 2006.

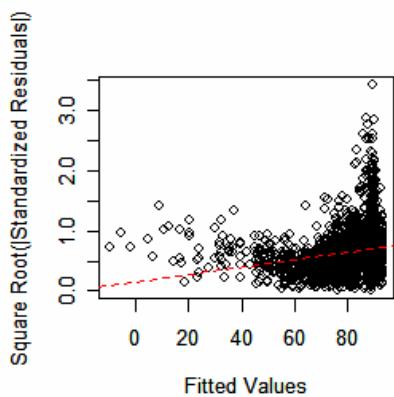
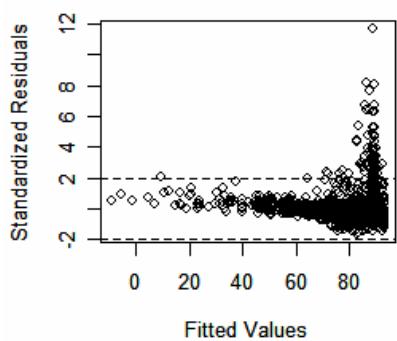
Output from model (1) appears on the following pages.

- i. Explain why it is necessary to use weighted least squares to fit model (1) and why  $w_i = n_i$  is the appropriate choice for the weights.
- ii. Explain why (1) is not a valid regression model.
- iii. Describe what steps you would take to obtain a valid regression model.

*Output from model (1)*



(continued)



#### Problem 4

A running shoe company produces a new model of running shoe that includes a harder material for the insert that corrects for overpronation. Two studies were carried out to see if the proportion of runners with heel tenderness is affected by the new shoe. Researchers asked the runners whether they experienced occasional heel tenderness.

- a. Two random samples of 87 runners were assigned the two types of running shoes. The first sample of runners was given ordinary running shoes and the second sample was given the shoes with the new insert. At the end of the study, each runner was asked whether he or she experienced occasional heel pain. Write out the hypotheses, formula for the test statistic, and the rejection region (in terms of a commonly tabulated distribution) to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes.

Shoe Type	Heel Pain	
	Yes	No
Ordinary	63	24
New	53	34

- b. Now suppose that two random samples of eight (8) runners were assigned the two types of running shoes. The first sample of runners was given ordinary running shoes and the second sample was given the shoes with the new insert. At the end of the study, each runner was asked whether he or she experienced occasional heel pain. Discuss the difficulty in using the data in the following table to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes. What procedure should you use?

Shoe Type	Heel Pain	
	Yes	No
Ordinary	6	2
New	5	3

- c. Now suppose a group of 87 runners uses both types of shoes. The runner indicated whether heel tenderness was experienced after using an ordinary shoe and the same runner was asked whether heel tenderness was experienced after using the new shoe. Explain why the method used in part a) cannot be used to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes.

Shoes	New Shoes	
	Yes	No
Yes	48	15
No	5	19

- d. Suppose we are back in the situation of part a). Define the variables  $Y = 1$  if “yes” and  $Y = 0$  if “no” and  $x = 1$  if new type and  $x = 0$  if ordinary type. Consider the logistic regression model

$$\text{logit}(\pi(x)) = \alpha + \beta x.$$

Explain how to test whether the proportion of runners experiencing heel pain differs using the logistic regression model. Write out the hypotheses, formula for the test statistic, and the rejection region (in terms of a commonly tabulated distribution.)

## **METHODS QUALIFYING EXAM**

**January 2008**

### **INSTRUCTIONS:**

1. 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.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. 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.
5. 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.

### **Problem I.**

Weight gain in the first 3 months after birth is important for new born infants. A pediatrician wishes to test a new feeding formula to determine if it will cause greater weight gain in new born infants than the standard formula.

From her records she finds that the first 3 months weight gains of single birth infants on the standard formula have the following characteristics:

$$\mu_S = 15\text{oz.} \quad \text{and} \quad \sigma_S = 6\text{oz.}$$

On the other hand, the first 3 months individual weight gains of identical twins on the standard formula have the following characteristics:

$$\mu_T = 12\text{oz.}, \quad \sigma_T = 6 \text{ oz. , with}$$

$$\rho = .8 \text{ (correlation in individual weight gains of identical twins)}$$

She wants to run a 3 month experiment on a group of infants to test the new formula versus the standard formula. She has decided to use a 5% probability of Type I error, and wishes to be able to detect a 3 oz. increase in weight gain with 90% probability.

- (a) Suppose the researcher conducts the experiment as a Completely Randomized Design of single birth infants with two treatments, standard and new. What is the required sample size? State any assumptions you are making in this calculation and show a detailed justification for your sample size.
- (b) Suppose the researcher conducts the experiment as a Paired Design with two treatments, standard and new, randomly assigned within each set of twins. What is the required sample size? State any assumptions you are making in this calculation and show a detailed justification for your sample size.
- (c) Discuss the relative merits of the two experiments in terms of practicality and the researcher's basic goal for the experiment.

## PROBLEM II.

A tire manufacturer wants to study the relationship between tread density and traction. Four different tread densities are considered and for each density there are three tires tested. The data are summarized below. Let  $(Y_{ji})$  be the traction measured for the  $i$ 'th tire having tread density  $x_j$ .)

$x_j$ (treads/inch)	$\bar{Y}_j = \sum_{i=1}^3 Y_{ji}/3$	$\sum_{i=1}^3 (Y_{ji} - \bar{Y}_j)^2$
0.0	2.0	0.38
1.0	4.0	23.6
2.0	3.8	0.14
3.0	3.0	13.38

A simple linear regression relating traction to tread density was fit yielding  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  with  $\hat{\beta}_0 = 2.78$  and  $\hat{\beta}_1 = 0.28$ .

- (a) Could these least-squares estimates be obtained using *only* the information in the above table. Why or why not?
- (b) Calculate the expected traction at each tread density assuming that the simple linear regression model is correct.
- (c) Based on a plot of  $(x_j, \bar{Y}_j)$  do you feel the simple linear regression is appropriate? Why or why not?
- (d) Do the errors appear to be homoscedastic? (Yes or No). If not, what transformation should be applied so that homoscedasticity would be a better assumption. Give the transformed  $\tilde{Y}$  and  $\tilde{X}$  as functions of  $Y$  and  $X$ .
- (e) Consider the model that uses  $\bar{Y}_j$  to estimate mean traction for tread density  $x_j$ ,  $j = 1, 2, 3, 4$ . Are these estimates unbiased if the simple linear regression model holds?
- (f) Are the estimates in (e) unbiased if the simple linear regression model does not hold?
- (g) How are the estimates in (e) typically inferior to those in (b) to estimate mean traction for tread density  $x_j$  when the simple linear regression model does hold? Explain briefly.

### **PROBLEM III.**

A poultry researcher is growing small chickens in an experiment conducted in an enclosed chicken house. The researcher has 12 pens available and has 12 different sources of protein that she wants to evaluate. There are 20 chickens in each of the pens and they are **pen fed**, that is, the 20 chickens share a common feeding trough. Therefore, the 12 sources of protein are randomly assigned to the 12 pens with 1 source for each pen. Some of the response variables are measured on a **pen** basis, for example, **feed conversion**, the amount of feed needed for an increase of 1 kg in body weight. Other response variables are measured on the individual chicken, for example, **average daily weight gain** (ADWF) and **percent body fat** (PDF).

1. If this experiment is not repeated elsewhere (either in time or space) is it a valid experiment? Why or why not?
  - (a) If valid, explain how you would analyze the data.
  - (b) If invalid, explain why it is invalid.
2. If the experiment was repeated in time, that is, a similar experiment was conducted 3 months later, explain how you would analyze the data. Include in your explanation, models, anova table, expected mean squares, testing procedures, and any other pertinent information.

## PROBLEM IV.

### Part (A)

Consider the simple linear regression model:  $Y = \beta_0 + \beta_1 x + e$ . Analysis of Variance can be used to test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0.$$

Analysis of Variance is based on the following three statistics:

The Total Corrected Sum of Squares of the  $Y$ s:

$$SST = SYY = \sum_{i=1}^n (y_i - \bar{y})^2,$$

The Residual Sum of Squares:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

The Regression Sum of Squares:

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

where  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ ,  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ ,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

In this questions, we will show that  $SST = SS_{reg} + RSS$ . To do this we will show that  $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ .

- (a) Show that  $(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$ .
- (b) Show that  $(\hat{y}_i - \bar{y}) = \hat{\beta}_1(x_i - \bar{x})$ .
- (c) Utilizing the fact that  $\hat{\beta}_1 = \frac{SXY}{SXX}$ , show that  $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ .

### Part (B)

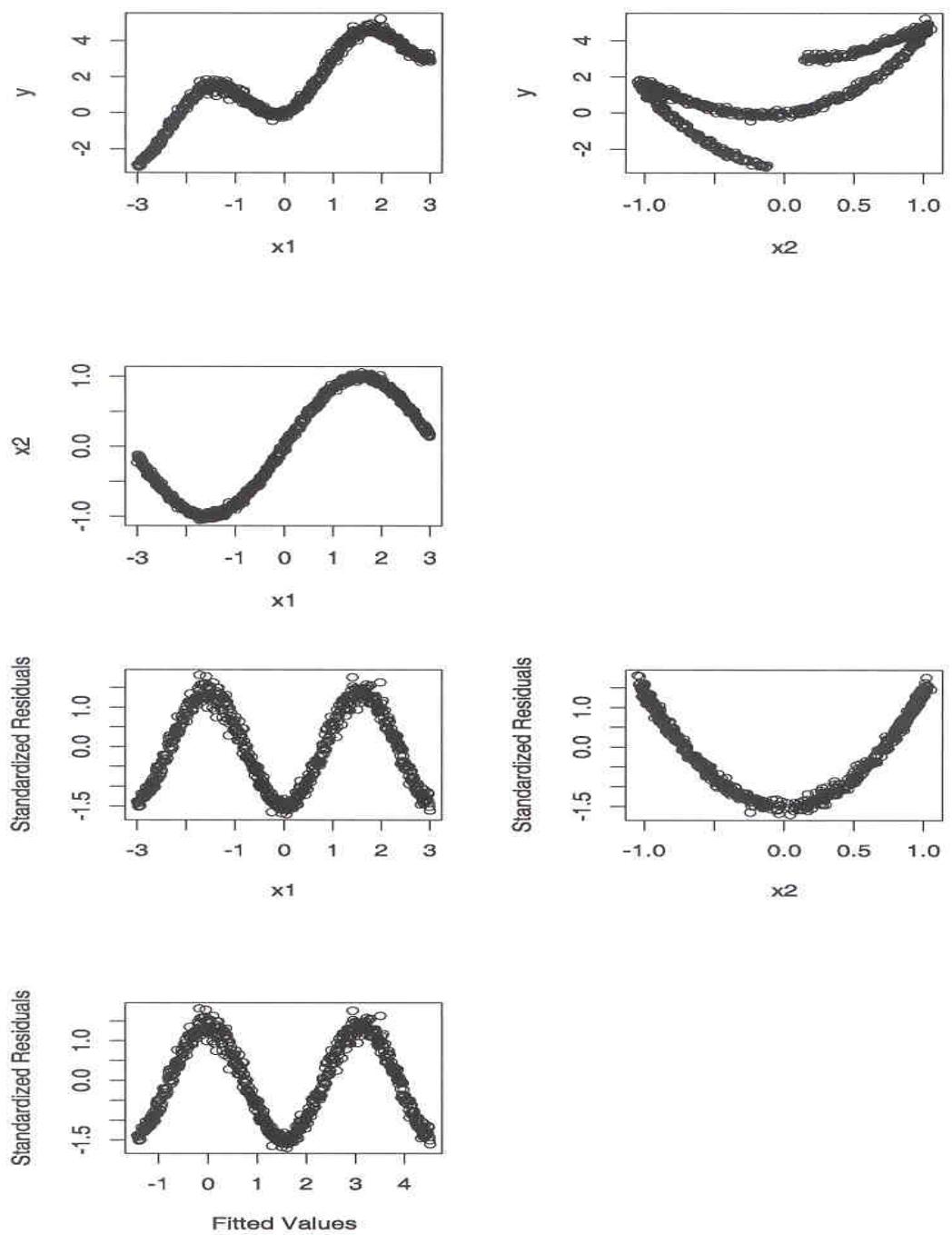
This question deals with a generated multiple regression data set with outcome variable  $Y$  and predictors  $x_1$  and  $x_2$ . There are a total of  $n = 601$  cases. One aim is to develop a valid model for  $Y$  based on  $x_1$  and  $x_2$ . The first model fit to the data was

$$Y_i = \beta_0 + \beta_1 x_{1i} + x_{2i} + e_i \quad (1)$$

Plots associated with Model (1) appear on the following page.

- a. Decide whether (1) is a valid model. Give reasons to support your answer.
- b. Decide whether the plots of standardized residuals provide any direct information on how model (1) is misspecified. Give a reason to support your answer.
- c. Describe what steps you would take to obtain a valid regression model.

*Output from model (I)*



# **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<sup>0</sup>F), and four strain rates (1, 10, 100, 1000 sec<sup>-1</sup>) 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

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

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

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  $\theta$  the *changeover value*.

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

$$Y_{1i} = \alpha_1 + \beta_1 x_{1i} + \epsilon_{1i} \quad \text{for } i = 1, 2, \dots, n_1$$

$$Y_{2i} = \alpha_2 + \beta_2 x_{2i} + \epsilon_{2i} \quad \text{for } i = 1, 2, \dots, n_2$$

by least squares, where

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

and

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

with  $\gamma$  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  $\gamma$  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**:

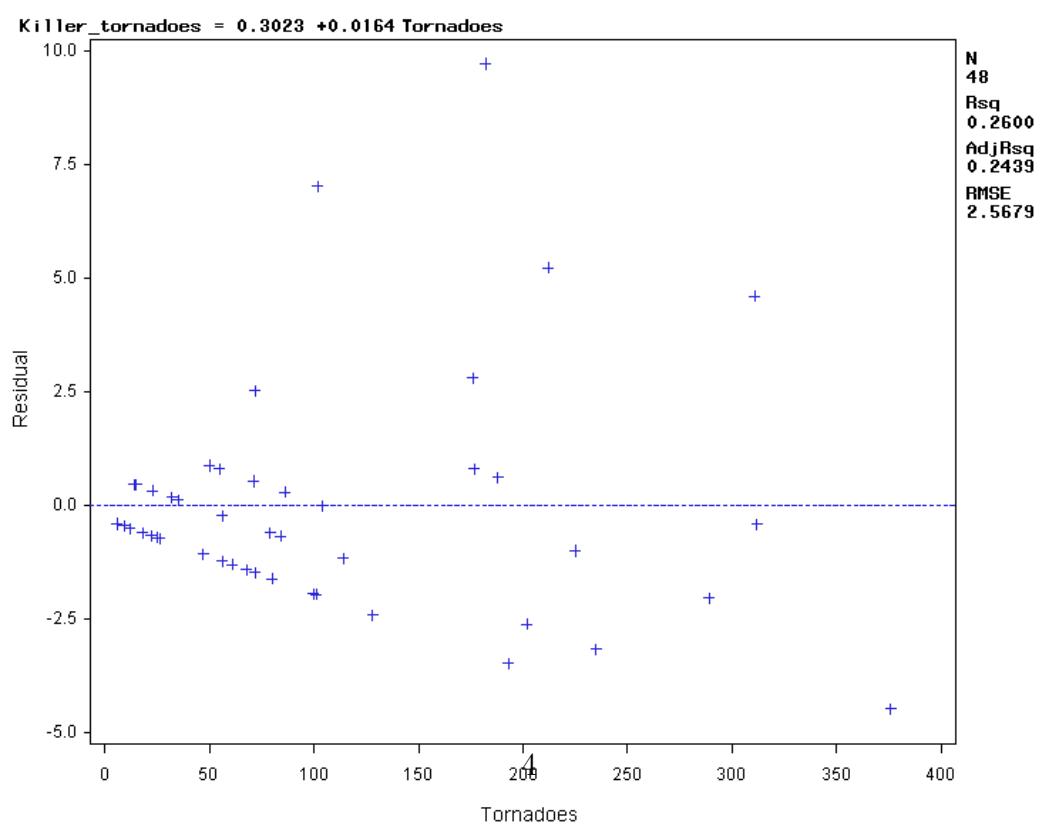
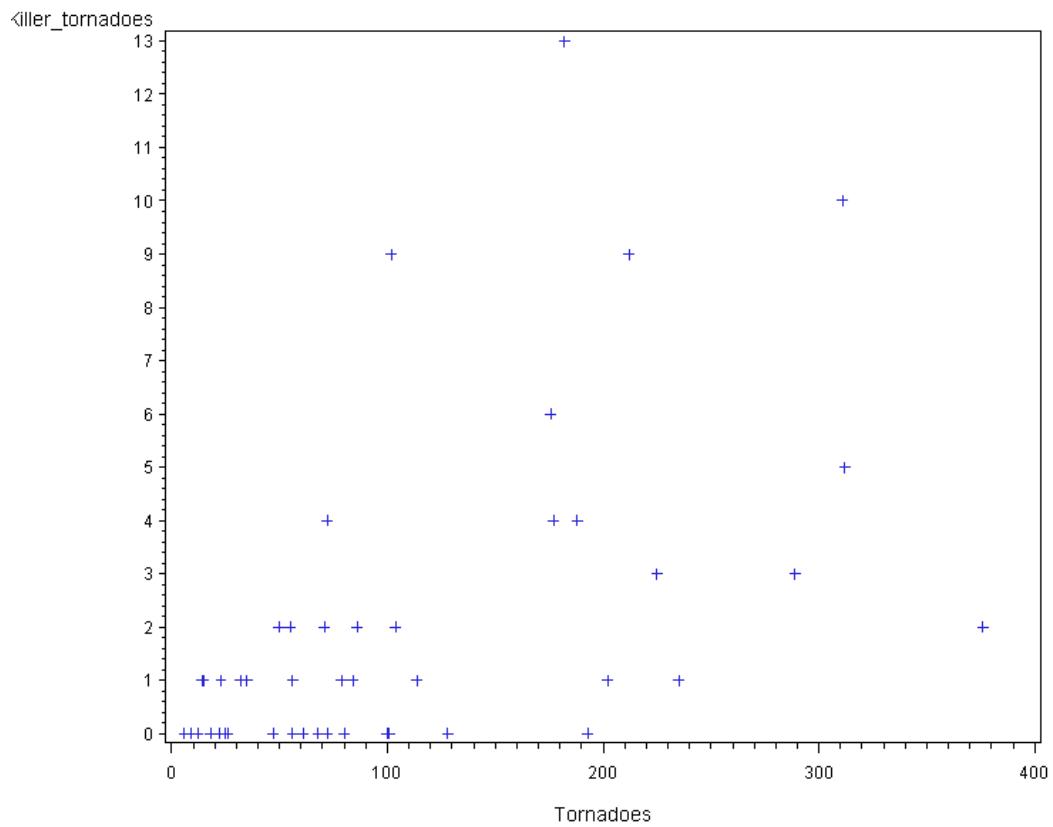
The MEANS Procedure				
Analysis Variable : Killer_tornadoes				
N	Mean	Std Dev	Minimum	Maximum
48	2.0416667	2.9532408	0	13.0000000

1. Explain why  $\bar{X} \pm Z_{\alpha/2} \sqrt{\bar{X}/n}$  is a reasonable approximate  $1 - \alpha$  confidence set for  $\lambda$ , 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					
Source	DF	Sum of Squares	Mean 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					
Variable	DF	Parameter Estimate	Standard 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



**Problem IV.**

Consider regression through 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

$$Y_i = \beta x_i + e_i \quad (i = 1, \dots, n)$$

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

## Some Chi-Squared Percentiles

df	Right-Tail Probability			
	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	

# **METHODS QUALIFYING EXAM**

**January 2009**

## **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.) Answer all the questions.
- d.) 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.
- e.) 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.
- f.) You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.

### Problem I.

Seven mice are subjected to a treatment and researchers are interested in the survival in days of these mice. Here are the survival times in days of a random sample of 7 mice.

16, 23, 38, 94, 99, 141, 197

- a.) Give a 95 percent confidence interval (CI) for mean survival of all mice subjected to the treatment using the assumption that  $t = \sqrt{n}(\bar{x} - \mu)/s$  has a t-distribution. [the 95th percentile of the t-distribution with 6 d.f. is 1.943, the 97.5th percentile is 2.447].
- b.) The researchers feel the interval in a) is too imprecise, approximately how many mice would they need in a new experiment to get an interval of width=30 days?
- c.) The researchers suspect that a) is not appropriate. It now assumed that the distribution of survival times is exponential, that is  $f(x) = \lambda \exp(-\lambda x)$  for some  $\lambda$ . Explain how to simulate to find the distribution of t. In particular, explain precisely how we find  $t_\alpha$ , the  $\alpha$ th percentile of the distribution of t for any  $\alpha \in (0, 1)$ ?
- d.) If from this simulation,  $t_{.025} = -3.75$ ,  $t_{.05} = -2.05$ ,  $t_{.95} = 1.45$  and  $t_{.975} = 1.75$ . What is the appropriate confidence interval for mean survival?

## Problem II.

An entomologist was studying the effect of two insecticides (labelled insecticides A and B) on four species of ants (labelled species 1, 2, 3 and 4). The entomologist anticipated that approximately 50% of each species would survive twenty-four hours' exposure to either of the insecticides.

A total of 2000 ants were obtained from each of the four species. For species 1, the ants were randomly divided into groups of 200. Each group was placed in a separate flask, resulting in a total of ten flasks filled with 200 ants each. Five of the ten flasks were randomly selected and then exposed to insecticide A. The other five flasks were exposed to insecticide B.

The same general procedure also was applied independently to each of species 2, 3 and 4.

After twenty-four hours of exposure to the selected pesticide, the entomologist examined each flask and recorded

$$y_{ijk} = \text{Number of ants (out of 200) that are still alive after}\\ 24 \text{ hours from species } i, \text{ insecticide } j, \text{ flask } k$$

- a.) Our entomologist provides you with the 40 values  $y_{ijk}$ ,  $i = 1, 2, 3, 4$ ;  $j = 1, 2$ ;  $k = 1, 2, 3, 4, 5$ . Write down a model for  $y_{ijk}$  associated with the above description of the experiment.
- b.) Display an ANOVA table for this experiment, showing sources of variation and degrees of freedom. Indicate for each source whether the source is a fixed or random effect. Find the expected mean square for each of source of variation in your ANOVA table. Indicate the appropriate denominator of the  $F$  statistic for each relevant test.
- c.) Give a brief interpretation of the *interaction* term in the model in part b). Be sure to include: i) an algebraic explanation in terms of cell means; and ii) a graphical illustration of two cases involving “zero interaction” and “nonzero interaction,” respectively.
- d.) Describe briefly two important diagnostic checks you would want to carry out before reporting analysis results from a) and b).
- e.) Now our entomologist wants to carry out a similar study involving the same two insecticides. The design is identical to the design described above, except that instead of having four species of ants, the entomologist will use four species of termites. Based on previous studies, the entomologist anticipates that about 0.1% of the termites will survive 24 hours' exposure to either insecticide, while the other 99.9% of the termites will not. Given this new information about this new experiment, would you consider it appropriate to use the methods in a) and b) to analyze the new termite data? Explain why or why not.

### **Problem III.**

A study was carried out on how general daily stress level (low or high) affects one's opinions (favorable or unfavorable) of a proposed new health policy. Interviews were made of random samples of subjects from both rural and urban areas where each was classified according to stress level and opinion of the proposed new health policy. Use the accompanying SAS output to help you answer this question.

- a.) Was the residence area associated with either of the two variables of primary interest, daily stress level and opinion on the proposed new health policy? Explain your reasoning.
- b.) Carry out a test of independence of the opinion on health policy with stress level, ignoring the area. If there is association, describe the nature of the association.
- c.) For each area (rural and urban), carry out a test of independence of the opinion on health policy with stress level. If there is association, describe the nature of the association.
- d.) Describe the difference in the results for b.) and c.). Explain how they can differ.
- e.) Describe how a similar phenomenon can occur in regression analysis where one fits a model relating two continuous variables for two treatment groups. Give a concrete example.

The FREQ Procedure

Table of stress by opinion

stress      opinion

	Frequency	favor	unfavor	Total
low		103		147   250
high		133		147   280
Total		236	294	530
		44.53	55.47	100.00

Statistics for Table of stress by opinion

Statistic	DF	Value	Prob
Chi-Square	1	2.1222	0.1452

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	0.7744	0.5489 1.0926
Sample Size = 530		

Table of residence by stress

residence      stress

	Frequency	low	high	Total
urban		60	220	280
rural		190	60	250
Total		250	280	530

Statistics for Table of residence by stress

Statistic	DF	Value	Prob
Chi-Square	1	157.8362	<.0001
Sample Size =	530		

Table of residence by opinion

residence      opinion

	Frequency	favor	unfavor	Total
urban		174	106	280
rural		62	188	250
Total		236	294	530

Statistics for Table of residence by opinion

Statistic	DF	Value	Prob
Chi-Square	1	74.5641	<.0001
Sample Size =	530		

Table 1 of stress by opinion  
Controlling for residence=urban

stress	opinion	Statistics for Table 1 of stress by opinion		
			Controlling for residence=urban	
Frequency	favor	unfavor	Total	
low	48	12	60	
high	126	94	220	
Total	174	106	280	

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	2.9841	1.5018      5.9297

Sample Size = 280

Table 2 of stress by opinion  
Controlling for residence=rural

stress	opinion	Statistics for Table 2 of stress by opinion		
			Controlling for residence=rural	
Frequency	favor	unfavor	Total	
low	55	135	190	
high	7	53	60	
Total	62	188	250	

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	3.0847	1.3207      7.2045

Sample Size = 250

Summary Statistics for stress by opinion  
Controlling for residence

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	17.5785	<.0001
Estimates of the Common Relative Risk (Row1/Row2)				
Type of Study	Method		Value	95% Confidence Limits
Case-Control (Odds Ratio)	Mantel-Haenszel Logit		3.0255 3.0235	1.7738 1.7731 5.1607 5.1559
Total Sample Size = 530				

## **Problem IV.**

This a Simple Linear Regression problem with several parts.

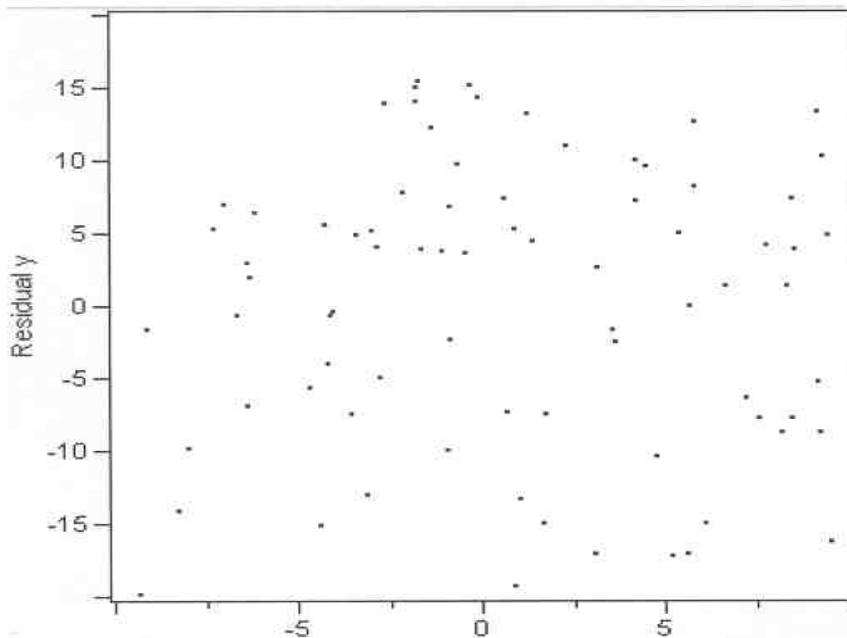
- 1.) Assume that the model is valid with respect to the mean function, that is,

$$E(Y) = \beta_0 + \beta_1 * X$$

- 2.) The errors may not be iid.
- 3.) There are three parts to this question; with each part standing on its own merits.
- 4.) Explain your answers. No long answers are needed.

### **Part 1.**

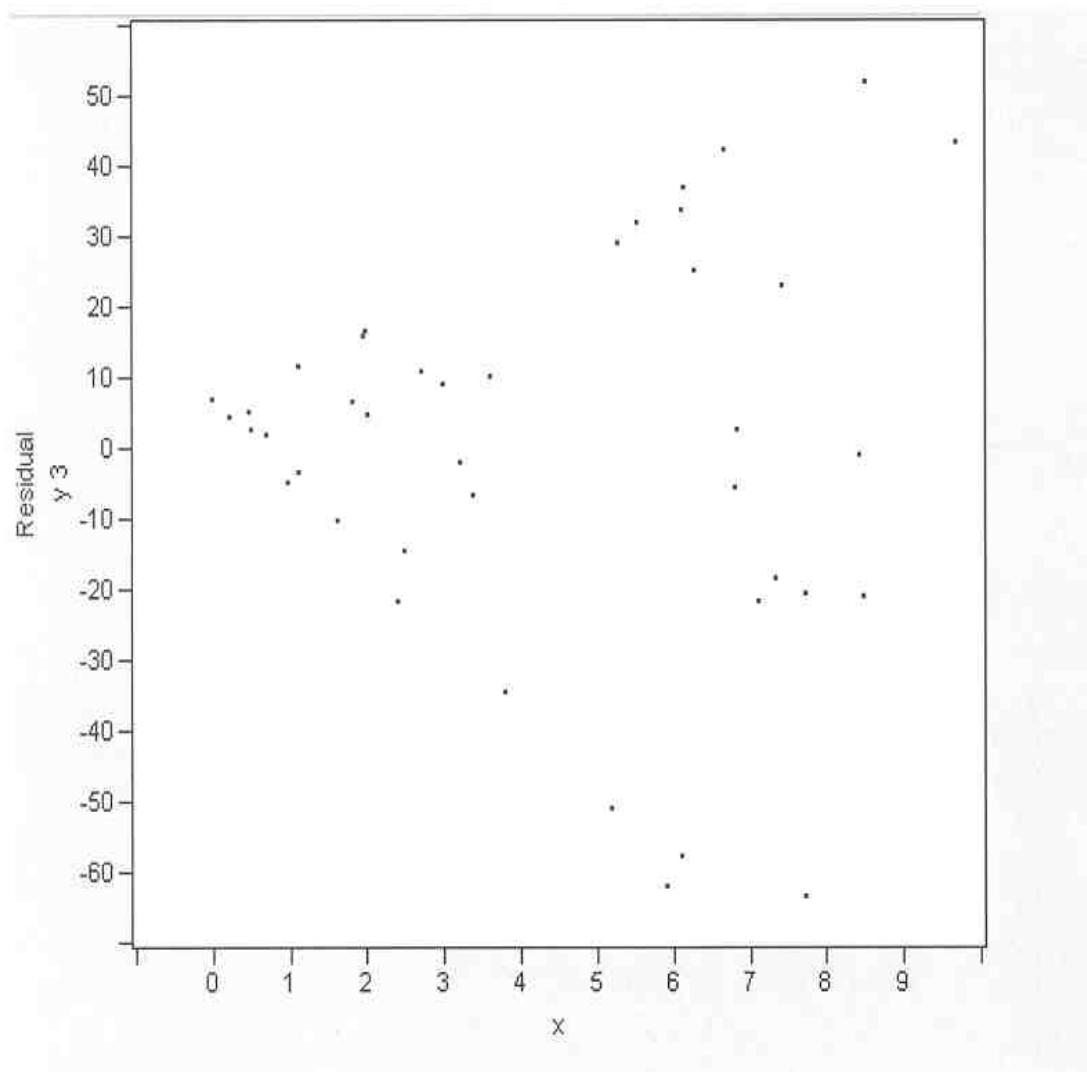
- a.) Given Graph 1 below, is there any reason to suggest that the errors are not iid?
- b.) Can you tell from Graph 1 if the residuals are normally distributed?
- c.) Are there any remedial measures you need to take based on Graph 1. If yes, please indicate what they are.



**GRAPH 1**

**Part 2.**

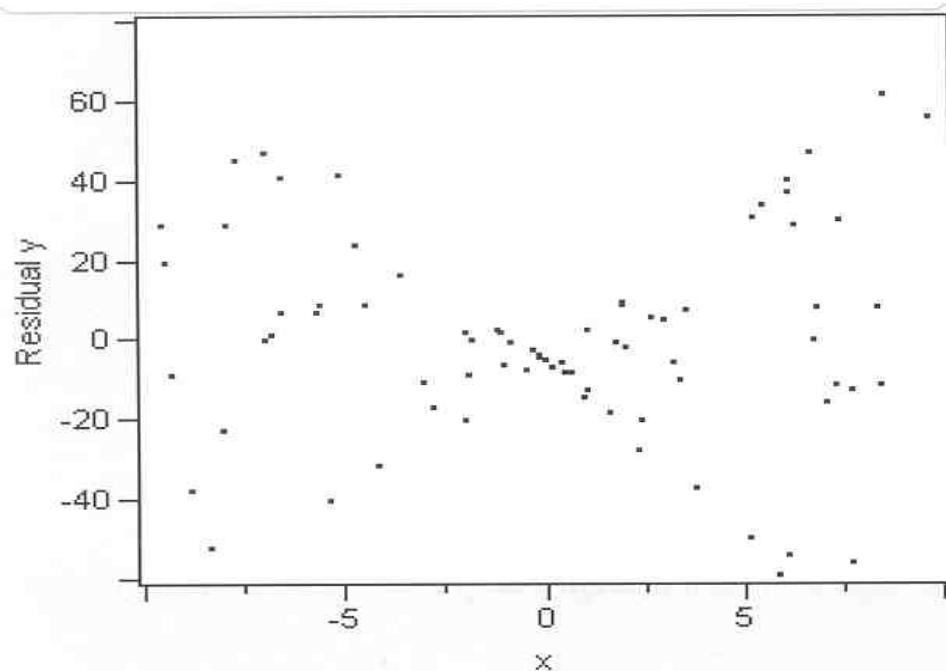
- a.) Given Graph 2 below, is there any reason to suggest that the errors are not iid?
- b.) Can you tell from Graph 2 if the residuals are normally distributed?
- c.) Are there any remedial measures you need to take based on Graph 2. If yes, please indicate what they are.



GRAPH 2

### Part 3.

- a.) Given Graph 3 below, is there any reason to suggest that the errors are not iid?
- b.) Can you tell from Graph 3 if the residuals are normally distributed?
- c.) Are there any remedial measures you need to take based on Graph 3. If yes, please indicate what they are.
- d.) Given the data in Dataset #3 below, what is the null hypothesis for testing homogeneity of variances? Give  $H_0 : \dots$



GRAPH 3

Dataset #3

Id	Y	X
1	-48.906	-9.488
2	-57.390	-9.407
3	-84.613	-9.245
4	-109.74	-8.815
5	-120.01	-8.324
6	-87.559	-7.987
7	-34.971	-7.879
8	-17.096	-7.686

9	-9.396	-6.986	52	32.524	2.712
10	-56.142	-6.984	53	34.096	3.005
11	-53.162	-6.783	54	25.316	3.221
12	-11.369	-6.518	55	22.761	3.396
13	-45.517	-6.515	56	41.879	3.610
14	-37.812	-5.660	57	-0.731	3.801
15	-35.617	-5.592	58	-1.605	5.198
16	-81.905	-5.324	59	79.410	5.276
17	1.668	-5.073	60	84.994	5.539
18	-12.897	-4.677	61	-4.279	5.929
19	-25.544	-4.435	62	1.950	6.119
20	-62.664	-4.102	63	93.363	6.122
21	-10.673	-3.578	64	96.671	6.133
22	-32.863	-3.008	65	86.779	6.301
23	-36.823	-2.750	66	108.040	6.680
24	-32.754	-1.936	67	62.021	6.823
25	-10.763	-1.933	68	70.248	6.840
26	-21.630	-1.886	69	49.236	7.135
27	-12.070	-1.822	70	55.136	7.355
28	-11.733	-1.776	71	97.180	7.429
29	-3.675	-1.167	72	57.155	7.733
30	-3.673	-1.105	73	14.233	7.739
31	-11.111	-1.021	74	84.284	8.431
32	-4.039	-0.832	75	65.244	8.499
33	-7.244	-0.408	76	138.130	8.519
34	-1.422	-0.318	77	142.630	9.690
35	-1.345	-0.148			
36	-1.781	-0.141			
37	-1.188	0.007			
38	-1.365	0.234			
39	1.921	0.471			
40	-0.024	0.498			
41	1.211	0.692			
42	-2.098	0.984			
43	0.740	1.114			
44	15.724	1.122			
45	-0.533	1.626			
46	18.468	1.807			
47	29.165	1.951			
48	30.305	1.988			
49	18.895	2.016			
50	-3.365	2.398			
51	4.689	2.486			

# METHODS QUALIFYING EXAM

August 2009

Student's Name \_\_\_\_\_

## INSTRUCTIONS FOR STUDENTS:

1. 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 solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer Questions I, II, and III but **select only ONE** of Questions IV and V to answer.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.
8. You are to answer Questions I, II, and III and then select **ONE** of Questions IV and V in this exam.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

## INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to **mspeed@stat.tamu.edu**. Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_ and the time at which the student completed the exam was \_\_\_\_\_.
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **msspeed@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

**Problem I.** For the following two experiments, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units.
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures
6. An ANOVA Table Including: Sources of Variation and Degrees of Freedom Freedom

**Experiment A:**

An experiment studies the effect of insect infestation and weeds on the yield of cotton plants. The experiment will include four levels of infestation, three weed treatments (no additional weeds, addition of weed species 1, addition of weed species 2), and two herbivore treatments (clipping or no clipping). There are eight fields; each field has three plots of land. Each of the plots receive 200 cotton plants to start growth. The eight fields are randomly assigned to the four levels of insect infestation, with two fields to each level. Within each field, the three plots are randomly assigned to the three weed treatments. Each plot is then split into two subplots; with one subplot randomly assigned to be clipped and the other subplot is not clipped. At the end of 18 weeks, the researcher determines the total cotton yield of each subplot of land. The yields are given here with the following notation: Field (F), Infestation (I), weed treatment (W), and clipping (C).

		W1		W2		W3				W1		W2		W3	
F	I	C1	C2	C1	C2	C1	C2	F	I	C1	C2	C1	C2	C1	C2
F1	I1	83.2	81.8	67.4	79.7	75.9	80.6	F5	I1	78.2	80.5	65.1	68.3	65.3	66.6
F2	I2	77.5	78.2	69.2	71.5	75.9	78.2	F6	I2	79.8	85.2	57.6	61.4	58.5	61.6
F3	I3	72.7	69.3	70.1	71.2	75.9	81.3	F7	I3	82.4	83.1	50.5	54.0	51.6	54.7
F4	I4	75.3	78.9	72.7	74.6	75.9	82.8	F8	I4	75.5	78.7	39.0	43.9	41.9	45.1

## Experiment B:

A human nutrition researcher conducted an experiment to determine the acceptability of cakes baked with sucrose substitutes as the sweetening agent. Specifically, there were 6 recipes formed by combinations of 3 sweeteners and 2 leavening agents:

$S_1$  : 100% sucrose     $S_2$  : 75% corn syrup, 25% sucrose     $S_3$  : 75% fructose, 25% sucrose

$L_1$  : Baking soda     $L_2$  : Baking soda plus “additional acid”

A panel of 6 taste testers were used to evaluate various characteristics of the cakes. On each of three days, cakes were baked from all six recipes. On each day, the six tasters evaluated six cake samples, one from each of the six recipes. The tasters then assigned a taste evaluation score to each of the recipes. The following table provides the tasting regimen for the three days:

Taster	Day 1						Day 2						Day 3					
	Order						Order						Order					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
T1	R1	R2	R3	R4	R5	R6	R6	R3	R4	R2	R5	R1	R1	R3	R5	R2	R4	R6
T2	R2	R3	R4	R5	R6	R1	R1	R4	R5	R3	R6	R2	R3	R5	R2	R4	R6	R1
T3	R3	R4	R5	R6	R1	R2	R2	R5	R6	R4	R1	R3	R5	R2	R4	R6	R1	R3
T4	R4	R5	R6	R1	R2	R3	R3	R6	R1	R5	R2	R4	R2	R4	R6	R1	R3	R5
T5	R5	R6	R1	R2	R3	R4	R4	R1	R2	R6	R3	R5	R4	R6	R1	R3	R5	R2
T6	R6	R1	R2	R3	R4	R5	R5	R2	R3	R1	R4	R6	R6	R1	R3	R5	R2	R4

## Problem II.

Two types of models for an experiment with a continuous response  $Y$  and four treatments are under consideration. The two models are given here:

“**Dummy variable**” model

$$(1) \quad \mu_i = E(Y_{ij}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3; \quad \text{for } i = 1, 2, 3, 4; \quad j = 1, \dots, n,$$

where  $x_i = 1$  if treatment  $i$  and  $x_i = 0$ , otherwise, for  $i = 1, 2, 3$ .

“**Design effects**” model

$$(2) \quad \mu_i = E(Y_{ij}) = \alpha^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3; \quad \text{for } i = 1, 2, 3, 4; \quad j = 1, \dots, n,$$

where, for  $i = 1, 2, 3$

$x_i = 1$  if treatment  $i$ ,  $x_i = -1$  if treatment 4, and  $x_i = 0$ , otherwise.

1. By considering the mean responses,  $\mu_1, \mu_2, \mu_3, \mu_4$ , of the four treatments, determine the relationship between  $\alpha, \beta_1, \beta_2, \beta_3$  and  $\alpha^*, \beta_1^*, \beta_2^*, \beta_3^*$ .
2. Obtain expressions for the difference in mean responses of treatments 2 and 3 for both models (1) and (2).
3. Suppose that there are factors  $A$  and  $B$ , each with two levels, that were used to define the treatments as follows: Treatment 1=  $A_1B_1$ , Treatment 2=  $A_1B_2$ , Treatment 3=  $A_2B_1$ , Treatment 4=  $A_2B_2$ . Express each of the following null hypotheses in term of the coefficients for the dummy variable model (1):
  - i. No interaction between  $A$  and  $B$
  - ii. No effect due to  $A$
  - iii. No effect due to  $B$
4. Consider now the “**dummy variable**” model for the logarithm of the mean:

$$(3) \quad \mu_i^* = \log(E(Y_{ij})) = \alpha^{**} + \beta_1^{**} x_1 + \beta_2^{**} x_2 + \beta_3^{**} x_3; \quad \text{for } i = 1, 2, 3, 4; \quad j = 1, \dots, n,$$

where  $x_i = 1$  if treatment  $i$  and  $x_i = 0$ , otherwise, for  $i = 1, 2, 3$ .

- i. Explain what the coefficients  $\alpha^{**}, \beta_1^{**}, \beta_2^{**}, \beta_3^{**}$  represent in terms of the mean responses  $\mu_i = E(Y_{ij})$   $i = 1, 2, 3, 4$ .
- ii. Explain what the difference in coefficients,  $\beta_1^{**} - \beta_2^{**}$  represents in terms of the mean responses,  $\mu_i$ 's, of the four treatments.

### Problem III.

This is the "Nambeware Polishing Times" data file. It concerns the efforts of a metal tableware manufacturer (Nambe Mills, Santa Fe, N. M.) to plan its production schedule. Each case represents a different item in the product line. The diameter, polishing time, price, and product type (there are 5 product types) are recorded for each item. Price is the dependent variable. The model fit to the data was

$$\text{Price} = \beta_{0i} + \beta_{1i} * \text{Type} * \text{Time} + \beta_{2i} * \text{Type} * \text{Diameter} + \text{error} \quad \text{for } i = 1, 2, 3, 4, 5$$

Using the output on this page and on Pages 5 and 6, answer the following questions: Do not make any additional assumptions.

1. Since Price is not normally distributed, what course of action should you take. Transform Price; transform both Price and the Predictors; do nothing. Explain your answer.
2. The R-Square is quite high. Does this mean that the model assumptions have been met. Explain.
3. In the following table, what null hypothesis is being tested by  $F = 151.95$ ? Spell out the null hypothesis.

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	15	583240.0866	38882.6724	151.95	< 0.0001

4. In the following table, what null hypothesis is being tested by  $F = 15.22$ ?

Spell out the null hypothesis.

Source	DF	Type III SS	Mean Square	F Value	Pr>F
Type	5	3891.92579	778.38516	3.04	0.0192
Time*Type	5	19470.94059	3894.18812	15.22	< 0.0001

## Distribution analysis of: price

The UNIVARIATE Procedure  
Variable: price (price)

Basic Statistical Measures			
Location		Variability	
<b>Mean</b>	86.38136	<b>Std Deviation</b>	51.57129
<b>Median</b>	75.00000	<b>Variance</b>	2660
<b>Mode</b>	99.00000	<b>Range</b>	238.50000
		<b>Interquartile Range</b>	64.00000

Tests for Location: Mu0=0				
Test	Statistic	p Value		
Student's t	t	12.86584	Pr >  t	<.0001
Sign	M	29.5	Pr >=  M	<.0001
Signed Rank	S	885	Pr >=  S	<.0001

Tests for Normality				
Test	Statistic	p Value		
Shapiro-Wilk	W	0.904181	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.132164	Pr > D	0.0112
Cramer-von Mises	W-Sq	0.219789	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.472978	Pr > A-Sq	<0.0050

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	583240.0866	38882.6724	151.95	<.0001
Error	44	11259.1634	255.8901		
Uncorrected Total	59	594499.2500			

R-Square	Coeff Var	Root MSE	price Mean
0.927010	18.51854	15.99656	86.38136

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	5	3891.92579	778.38516	3.04	0.0192
time*type	5	19470.94059	3894.18812	15.22	<.0001
diam*type	5	10339.64186	2067.92837	8.08	<.0001

Parameter	Estimate	Standard Error	t Value	Pr >  t
type Bowl	-25.43340388	11.96199181	-2.13	0.0391
type Casserole	-29.73215408	37.86173748	-0.79	0.4365
type Dish	-25.90176313	26.04032795	-0.99	0.3253
type Plate	-2.74650494	32.51765971	-0.08	0.9331
type Tray	-48.87508509	16.22377283	-3.01	0.0043
time*type Bowl	1.37132089	0.36671340	3.74	0.0005
time*type Casserole	2.18134783	0.32535774	6.70	<.0001
time*type Dish	1.78116744	0.71406240	2.49	0.0164
time*type Plate	-0.94400915	1.55388850	-0.61	0.5466
time*type Tray	1.81331318	0.55784223	3.25	0.0022
diam*type Bowl	5.89371075	1.42682164	4.13	0.0002
diam*type Casserole	3.94831670	2.76362585	1.43	0.1602
diam*type Dish	4.91642503	4.30069811	1.14	0.2592
diam*type Plate	7.54121327	1.86103147	4.05	0.0002
diam*type Tray	5.00293936	2.64552608	1.89	0.0652

**Problem IV.** Assume the two-part linear regression model:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

where  $\mathbf{Y}$  is an  $n \times p$  matrix of response variables,  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are, respectively,  $n_1 \times p_1$  and  $n_2 \times p_2$  matrices of fixed (i.e., nonrandom) explanatory variables,  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$  are, respectively,  $p_1 \times 1$  and  $p_2 \times 1$  vectors of unknown parameters, and  $p_1 + p_2 = p$ . Assume that the  $n \times p$  matrix  $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$  is of full column rank. The  $n \times 1$  vector  $\boldsymbol{\epsilon}$  is assumed to be comprised of independent, identically distributed  $(0, \sigma^2)$  random errors.

Hint: To answer parts a) through c) below you may wish to use the following result:

*For any real symmetric matrix  $\mathbf{Q}$ , the expectation of the quadratic form  $\mathbf{Y}^T \mathbf{Q} \mathbf{Y}$  is*

$$E(\mathbf{Y}^T \mathbf{Q} \mathbf{Y}) = \text{tr}[\mathbf{Q} \text{Var}(\mathbf{Y})] + [E(\mathbf{Y})]^T \mathbf{Q} [E(\mathbf{Y})].$$

1. Show that:

$$E(\mathbf{Y}^T \mathbf{P}_1 \mathbf{Y}) = p_1 \sigma^2 + (\boldsymbol{\beta}_1 + \mathbf{A} \boldsymbol{\beta}_2)^T \mathbf{X}_1^T \mathbf{X}_1 (\boldsymbol{\beta}_1 + \mathbf{A} \boldsymbol{\beta}_2),$$

where  $\mathbf{P}_1 = \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$  and  $\mathbf{A} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2$ .

Explain why it is not necessary that  $\boldsymbol{\beta}_1 = 0$  in order for  $E[(\mathbf{Y}^T \mathbf{P}_1 \mathbf{Y}) / (p_1 \sigma^2)] = 1$ .

Give simple sufficient conditions (that  $\mathbf{X}_1$  and  $\mathbf{X}_2$  satisfy)

in order for  $E[(\mathbf{Y}^T \mathbf{P}_1 \mathbf{Y}) / (p_1 \sigma^2)] = 1$  to imply that  $\boldsymbol{\beta}_1 = 0$ .

2. Show that:

$$E[\mathbf{Y}^T (\mathbf{P} - \mathbf{P}_1) \mathbf{Y}] = p_2 \sigma^2 + \boldsymbol{\beta}_2^T \mathbf{X}_2^T (I - \mathbf{P}_1) \mathbf{X}_2 \boldsymbol{\beta}_2,$$

where  $\mathbf{P} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ . Give sufficient conditions (that  $\mathbf{X}_1$  and  $\mathbf{X}_2$  satisfy) in order for  $E[\mathbf{Y}^T (\mathbf{P} - \mathbf{P}_1) \mathbf{Y} / (p_2 \sigma^2)] = 1$  to imply that  $\boldsymbol{\beta}_2 = 0$ .

3. Show that:

$$E[\mathbf{Y}^T (\mathbf{I} - \mathbf{P}) \mathbf{Y}] = (n - p) \sigma^2.$$

4. Suppose that we are interested in testing:

$$H_o : \boldsymbol{\beta}_1 = 0 \text{ versus } H_1 : \boldsymbol{\beta}_1 \neq 0.$$

Would the statistic  $F = (\mathbf{Y}^T \mathbf{P}_1 \mathbf{Y}) / (p_1 s^2)$ , where  $s^2 = [\mathbf{Y}^T (\mathbf{I} - \mathbf{P}) \mathbf{Y}] / (n - p)$  is the least squares estimator of  $\sigma^2$ , be a useful test statistic in general? Justify your answer heuristically using the results in parts (1.) and (3.).

5. Assume  $\mathbf{X}$  is of full rank and suppose that we are interested in testing:

$$H_o : \boldsymbol{\beta}_2 = 0 \text{ versus } H_1 : \boldsymbol{\beta}_2 \neq 0.$$

Would the statistic  $F = [\mathbf{Y}^T (\mathbf{P} - \mathbf{P}_1) \mathbf{Y}] / (p_2 s^2)$  ever be a reasonable test statistic? Again, justify your answer heuristically, now using the results in parts (2.) and (3.).

## Problem V.

### Part 1:

Suppose that  $Y$  has mean equal to  $\mu$  and variance equal to  $\mu^4$  show that the appropriate transformation of  $Y$  for stabilizing variance is the reciprocal transformation (i.e.,  $1/Y$ ).

### Part 2:

Chu (1996, Diamond ring pricing using linear regression *Journal of Statistics Education*, 4, <http://www.amstat.org/publications/jse/v4n3/datasets.chu.html>) discusses the development of a regression model to predict the price of diamond rings from the size of their diamond stones (in terms of their weight in carats). Data on both variables were obtained from a full page advertisement placed in the Straits Times newspaper by a Singaporebased retailer of diamond jewelry. Only rings made with 20 carat gold and mounted with a single diamond stone were included in the data set. There were 48 such rings of varying designs. (Information on the designs was available but not used in the modeling.) The weights of the diamond stones ranged from 0.12 to 0.35 carats (a one carat diamond stone weighs 0.2 gram) and were priced between \$223 and \$1086.

An analyst fit two models to the data. The first model fit to the data was

$$\text{Price} = \beta_0 + \beta_1 \text{Size} + e \quad (1)$$

On Page 9 is some output from fitting model (1) as well as some plots.

The second model fit to the data was

$$\text{Log(Price)} = \beta_0 + \beta_1 \text{Log(Size)} + e \quad (2)$$

Output from model (2) and plots appear on Page 10.

- (a) Based on the output for models (1) and (2) the analyst concluded the following:  
*Since model (1) has a higher  $R^2$  than model (2), model (1) is a more effective model for producing prediction intervals for Price.*  
Provide a detailed critique of this conclusion.
- (b) Carefully describe any shortcomings evident in model (1). For any shortcoming, describe the steps needed to overcome the shortcoming.
- (c) Is model (2) an improvement over model (1) in terms of predicting Price? If so, please describe all the ways in which it is an improvement.
- (d) Interpret the estimated coefficient of  $\text{log(Size)}$  in model (2).
- (e) List any weaknesses apparent in model (2).

## Output from R for model (1)

Call:

```
lm(formula = Price ~ Size)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-258.05	16.94	-15.23	<2e-16 ***
Size	3715.02	80.41	46.20	<2e-16 ***

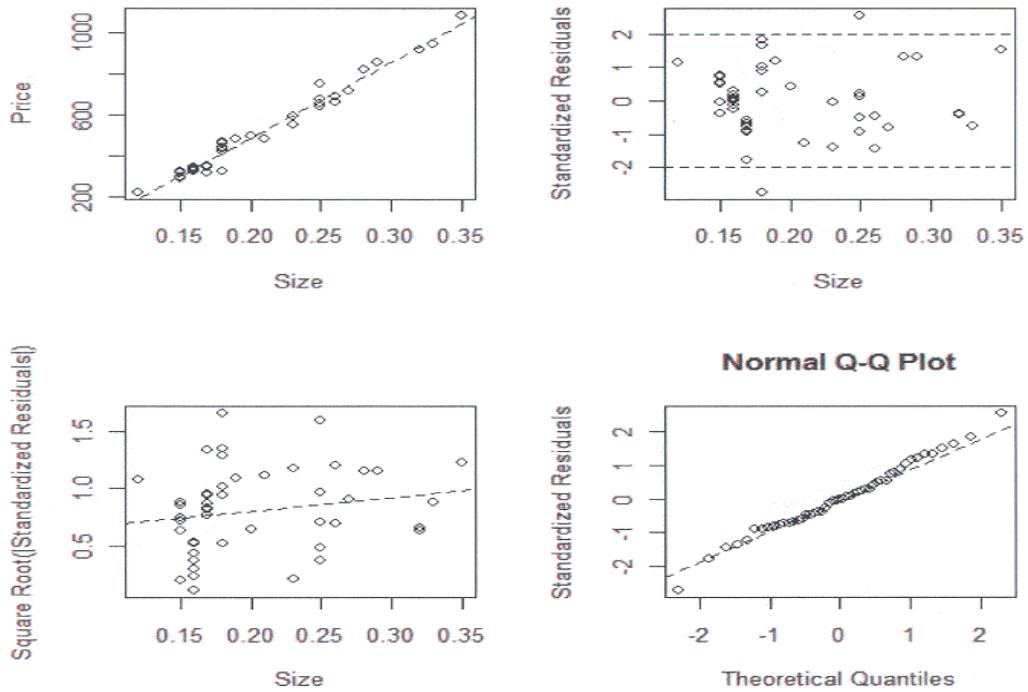
---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

Residual standard error: 31.6 on 47 degrees of freedom

Multiple R-squared: 0.9785, Adjusted R-squared: 0.978

F-statistic: 2135 on 1 and 47 DF, p-value: < 2.2e-16



## Output from R for model (2)

Call:

```
lm(formula = log(Price) ~ log(Size))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.56317	0.06221	137.65	<2e-16 ***
log(Size)	1.49566	0.03772	39.65	<2e-16 ***

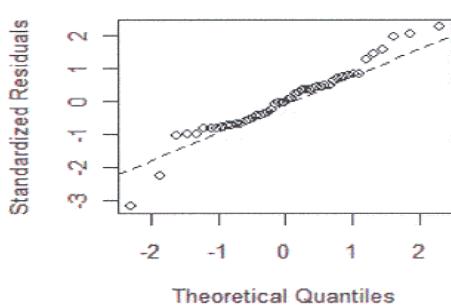
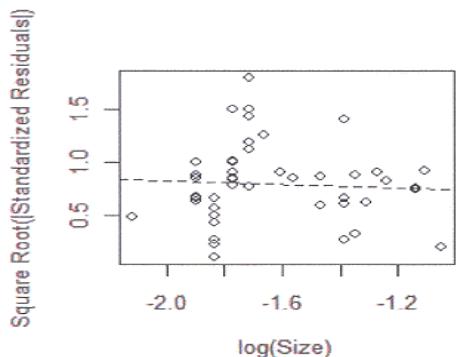
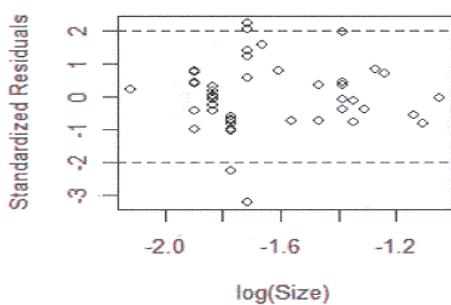
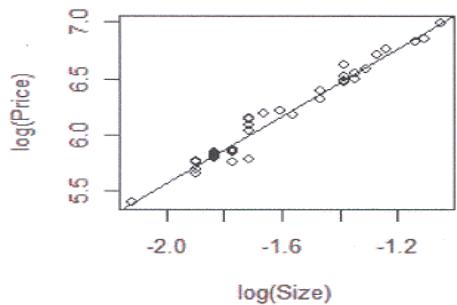
---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

Residual standard error: 0.06796 on 47 degrees of freedom

Multiple R-squared: 0.971, Adjusted R-squared: 0.9704

F-statistic: 1572 on 1 and 47 DF, p-value: < 2.2e-16



# METHODS QUALIFYING EXAM

January 2010

Student's Name \_\_\_\_\_

## INSTRUCTIONS FOR STUDENTS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer Questions I, II, and III but **select only ONE** of Questions IV and V to answer.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.
8. You are to answer Questions I, II, and III and then select **ONE** of Questions IV and V in this exam.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

## INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to **longneck@stat.tamu.edu** Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was  
\_\_\_\_\_.
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

**Problem I. Part A:**

For the following experiment, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units.
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures
6. An ANOVA Table Including: Sources of Variation and Degrees of Freedom

An industrial engineer is studying the hand insertion of electronic components on printed circuit boards in order to improve the speed of the assembly operation. She has designed three assembly fixtures ( $F_1, F_2, F_3$ ) and two workplace layouts ( $L_1, L_2$ ) that seem promising. Specialized operators are required to perform the assembly and it was initially decided to randomly select four operators from the many qualified operators at the plant. However, because the workplaces are in different locations within the plant, it is difficult to use the same operators for each layout. Therefore, the four operators randomly chosen for layout 1 are different individuals from the four operators randomly chosen for layout 2. Each of the operators assembles four circuit boards for each of the three fixture types with the 12 circuit boards assembled in random order. The 96 assembly times are measured in seconds. The engineer is interested in the effects of Assembly Fixtures (F), Workplace Layout (L), and Operator (O) on the average time required to assemble the circuit boards.

**Problem I. Part B:**

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations. **Place** your selection in the space to the **left** of each situation.

**SITUATION:**

- ..... 1. A CRD was conducted with Factor  $F_1$  having four fixed quantitative levels and Factor  $F_2$  with six randomly selected levels. The AOV table reveals that the interaction between  $F_1$  and  $F_2$  was significant. The researcher wanted to investigate the change in the mean of the response with increasing levels of factor  $F_1$ .
- ..... 2. An experiment was designed to compare four techniques, the levels of  $F_1$ , for removing mercury contamination from drinking water. The researcher wanted to also evaluate the variability in the many devices, the levels of  $F_2$ , of measuring mercury levels in water. Five devices for detecting mercury were randomly selected from the list of all such devices. A specified amount of mercury was placed in 200 water samples. Ten of the 200 water samples were randomly assigned to each of the twenty combinations of a level of  $F_1$  and a level of  $F_2$ . There was significant evidence of an interaction between factors  $F_1$  and  $F_2$ . The researcher wants to determine which of the four techniques removed the greatest amount of mercury.
- ..... 3. A three factor experiment is run with Factor  $F_1$  having five fixed levels, Factor  $F_2$  with six fixed selected levels and Factor  $F_3$  with four fixed levels. The effects  $F_1 * F_2$ ,  $F_1 * F_3$ , and  $F_1 * F_2 * F_3$  were all found to be nonsignificant. The statistician wants to evaluate the pairwise differences in the levels of factor  $F_1$ .
- ..... 4. An experiment is conducted using a factorial treatment structure with factor  $F_1$  having values  $40^{\circ}C$ ,  $50^{\circ}C$ ,  $60^{\circ}C$ ,  $70^{\circ}C$  crossed with factor  $F_2$  having levels A, B, C in a CRD with three reps per treatment. There is not significant evidence of an interaction between  $F_1$  and  $F_2$ . The researcher wants to determine the temperature that yields the maximum mean response.
- ..... 5. In an experiment having the levels of factor  $F_1$ -qualitative and the levels of factor  $F_2$ -quantitative, there was significant evidence of an interaction between  $F_1$  and  $F_2$ . The experimenter wants to compare the mean responses across the levels of factor  $F_1$ , averaged over the levels of factor  $F_2$ .
- ..... 6. In an experiment was designed to compare the performance of three new types of machine tools to the machine tool currently in use, factor  $F_1$ , with four levels. A random sample of five machinists, factor  $F_2$ , were randomly selected from the workforce. Each machinists produced ten units of product from each of the four types of machines. A quality rating was determined for each of the 200 units produced in the study. There was significant evidence of an interaction between factors  $F_1$  and  $F_2$ . The company wants to know if any of the new types of machines have a higher mean quality rating than the type of machine the company is currently using.

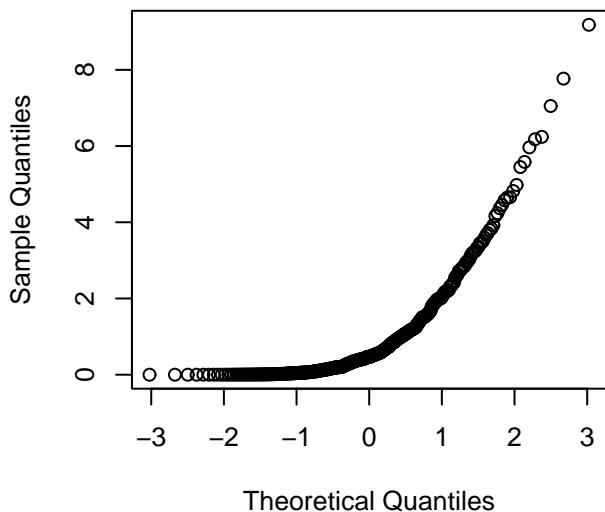
**TECHNIQUE:**

- A. Trend analysis using Scheffe contrasts
- B. Trend analysis using Bonferroni contrasts
- C. Trend analysis in the levels of  $F_1$  averaged over levels of the other factors
- D. Trend analysis in the levels of  $F_1$  separately at each level of the other factors
- E. Scheffe's test for contrast differences
- F. Dunnett's comparison technique
- G. Dunnett's comparison technique to all combinations of the factors
- H. Dunnett's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- I. Dunnett's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- J. Tukey's comparison technique
- K. Tukey's comparison technique to all combinations of the factors
- L. Tukey's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- M. Tukey's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- N. Hsu's comparison technique
- O. Hsu's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- P. Hsu's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- Q. Hsu's comparison technique applied to all combinations of the factors
- R. Nothing new is learned beyond the results of the F-tests from the AOV table.
- S. Comparison of marginal means is not appropriate.
- T. None of the above methods are appropriate.

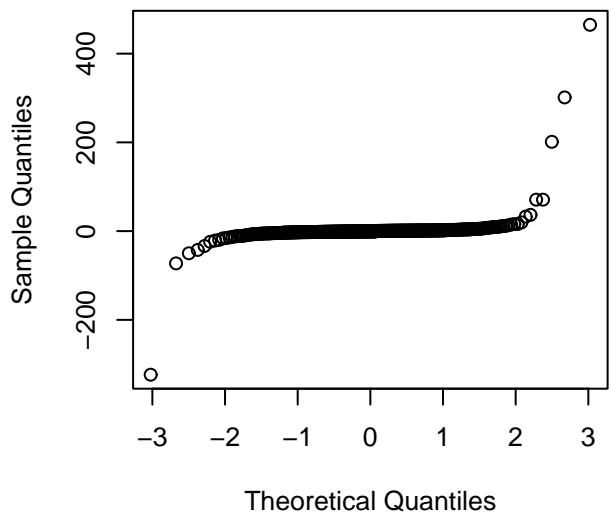
## Problem II.

Normal Reference Distribution (QQ) plots are often used to assess distributional assumptions. The following normal quantile plots were produced in R using the command `qqnorm`. For each of the following four plots, discuss the assumption of normality for the pictured data. If the data are nonnormal, describe the manner in which the data are nonnormal and a transformation (if possible) to make the data more normally distributed.

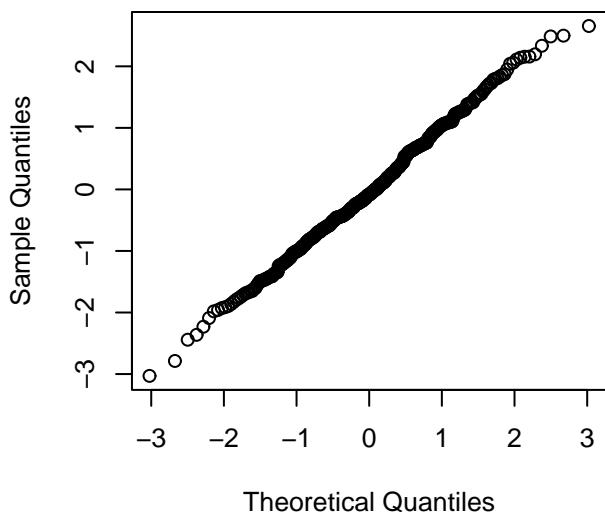
**Normal Q–Q Plot A.**



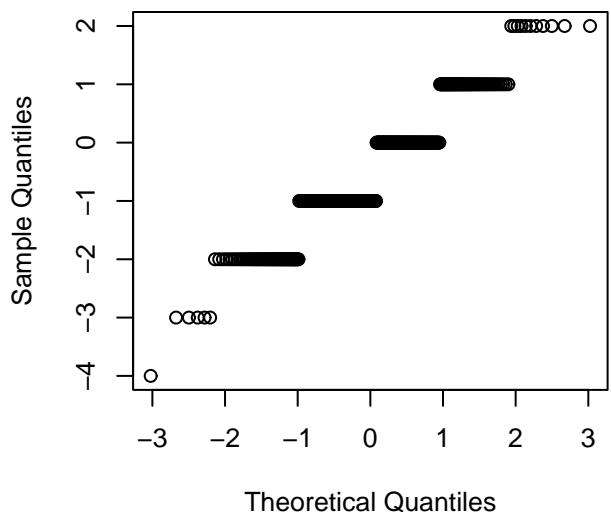
**Normal Q–Q Plot B.**



**Normal Q–Q Plot C.**



**Normal Q–Q Plot D.**

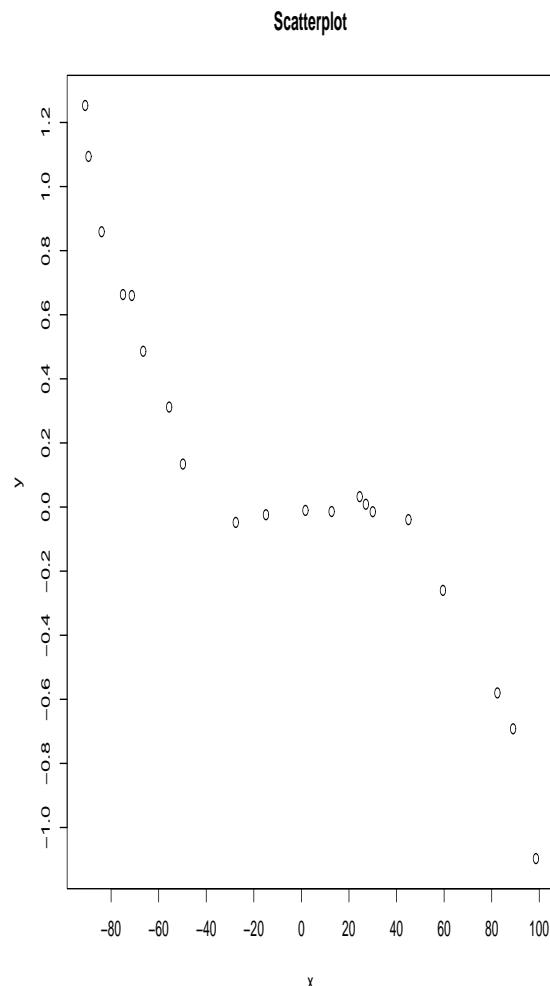


### Problem III.

An engineer wants to evaluate the relationship between the standardized amount of additive,  $x$ , used in a chemical reaction and the deviation from the standard yield of a chemical reaction,  $y$ . Given the data in Table given below and the scatterplot of the data, answer the following questions. Please explain your answers. Do not make any calculations.

1. Does a reasonable model for this data satisfy the requirement for multiple linear regression?
2. If the model  $y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \epsilon$  is fit to the data, how many distinct populations are involved in the modelling?
3. If you use the model in part 2. above and you test for equality of variance, what is the null hypothesis in terms of  $H_o : \dots$ ?
4. If you use Anderson-Darling statistic or Shapiro-Wilks statistic to test  $H_o : y$  has a normal distribution and reject that null hypothesis, would this conclusion violate the assumptions for multiple linear regression?

Data Table		
ID	x	y
1	-90.9	1.25300
2	-89.5	1.09400
3	-84.0	0.85900
4	-75.0	0.66300
5	-71.3	0.66000
6	-66.5	0.48600
7	-55.6	0.31200
8	-49.8	0.13400
9	-27.6	-0.04800
10	-14.9	-0.02400
11	1.7	-0.01050
12	12.7	-0.01400
13	24.5	0.03240
14	27.1	0.00871
15	30.0	-0.01460
16	45.0	-0.03930
17	59.5	-0.26000
18	82.4	-0.58000
19	89.0	-0.69200
20	98.6	-1.09700



#### Problem IV.

Consider the usual multiple linear regression model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{Y}$  is an  $n \times 1$  vector of response variables,  $\mathbf{X}$  is an  $n \times p$  matrix of predictor variables,  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of unknown parameters and  $\boldsymbol{\epsilon}$  is an  $n \times 1$  vector of unobservable independent and identically distributed random variables, each with mean zero and variance  $\sigma^2$ . In what follows, you may not assume that the matrix  $\mathbf{X}$  is of full column rank.

Four results are given below about least squares estimation of  $\boldsymbol{\beta}$  for this multiple linear regression model. You are to prove any **three of the four results** that you choose. Clearly indicate **which three** results are to be graded.

Result 1:

The normal equations,  $\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y}$ , are consistent, that is, they always have at least one solution.

Result 2:

Any solution to the normal equations satisfies the least squares criterion, that is, if  $\hat{\boldsymbol{\beta}}$  satisfies:

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y}$$

then the minimum value  $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$  can attain is  $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ .

Result 3:

The least squares predicted values,  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ , where  $\hat{\boldsymbol{\beta}}$  is any solution to the normal equations, are unique.

Result 4:

If the model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  contains an intercept parameter, the mean (that is, arithmetic average) of the least squares residuals that result from the fitting of this model is zero.

## Problem V.

A book<sup>1</sup> on robust biostatistical methods published in 2009 considers a data set taken from Everitt (1994), *The Handbook of Statistical Analysis Using Splus*, Chapman & Hall. The data consist of the IQ scores (IQ) and behavioral problem scores (BP.score) of children of age five, labeled according to whether or not their mothers had suffered an episode of postnatal depression (state.mother = 1 if yes and 0 if no). We seek to model IQ as a function of BP.score and to determine whether the effect of BP.score differs significantly across the two groups of mothers.

The two models under consideration are as follows:

$$(1) \quad IQ = \beta_0 + \beta_1 BP.score + \beta_2 state.mother + e$$

and

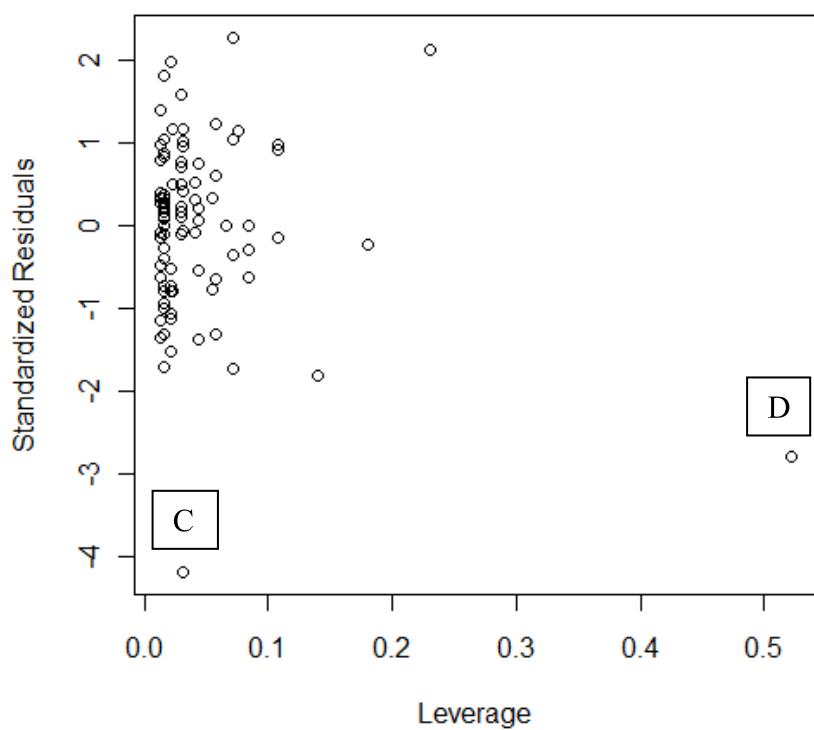
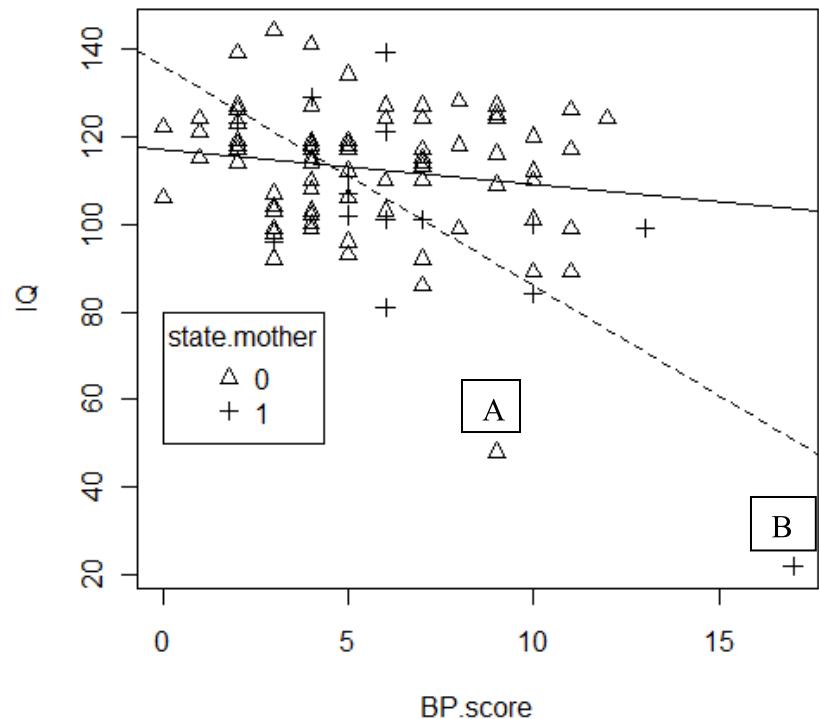
$$(2) \quad IQ = \beta_0 + \beta_1 BP.score + \beta_2 state.mother + \beta_3 BP.score \times state.mother + e$$

A plot of the data and the two regression lines from model (2) along with a plot of the standardized residuals and leverage values from model (2) can be found below. In addition, numerical output from R for models (1) and (2) appears below.

- a) Two points are marked as “A” and “B” in the plot of IQ and BP.score. Two points are marked as “C” and “D” in the plot of the standardized residuals and leverage values from model (2). Match “A” and “B” with “C” and “D”. Give reasons to support your choices.
- b) Using the output from R provided below, calculate the value of the F-statistic for testing the null hypothesis,  $H_0 : \beta_3 = 0$ .
- c) Briefly describe the steps you would follow in order to obtain a final model for the data on IQ and BP.score , labelled according to whether or not their mothers had suffered an episode of postnatal depression.

<sup>1</sup>Heritier, S., E. Cantoni, S. Copt, & M.-P. Victoria-Feser (2009) *Robust Methods in Biostatistics*. Wiley, New York

Plots associated with model (2)



## Output from R for model (1)

Call:  
lm(formula = IQ ~ BP.score + state.mother)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	122.5832	3.3674	36.403	< 2e-16 ***
BP.score	-1.8171	0.5281	-3.441	0.000878 ***
state.mother	-8.7970	4.5782	-1.922	0.057797 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.9731 on 91 degrees of freedom  
Multiple R-squared: 0.1699, Adjusted R-squared: 0.1516  
F-statistic: 9.312 on 2 and 91 DF, p-value: 0.0002093

## Edited Output from R for model (2)

Call:  
lm(formula = IQ ~ BP.score + state.mother + BP.score:state.mother)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	117.1334	3.5034	33.434	< 2e-16 ***
BP.score	-0.8064	0.5693	-1.417	0.160063
state.mother	18.9970	8.8000	2.159	0.033531 *
BP.score:state.mother	-4.2027	???????	???????	0.000486 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.0061 on 90 degrees of freedom  
Multiple R-squared: 0.2754, Adjusted R-squared: 0.2513  
F-statistic: 11.4 on 3 and 90 DF, p-value: 2.084e-06

## METHODS QUALIFYING EXAM

August 2010

Student's Name \_\_\_\_\_

### INSTRUCTIONS FOR STUDENTS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer Questions I, II, and III but **select only ONE** of Questions IV and V to answer.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.
8. You are to answer Questions I, II, and III and then select **ONE** of Questions IV and V in this exam.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

### INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to **longneck@stat.tamu.edu** Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or  
emailed to **longneck@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

**PROBLEM I. Part A:**

For the following experiment, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units;
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures;
6. A partial ANOVA Table containing just Sources of Variation (SV) and Degrees of Freedom (DF);

Commercial cheese is manufactured by bacterial fermentation of Pasteurized milk. Selected bacteria is added to the milk to implement the fermentation, referred to as starter cultures. However, some *Wild* bacteria may also be present in cheese, nonstarter bacteria, which may alter the desired quality of the cheese. Thus, cheese manufactured under seemingly identical conditions in two cheese making facilities may produce cheese of differing quality due to the present of different indigenous nonstarter bacteria. To test the impact of two nonstarter bacteria, R50 and R21, on cheese quality, the nonstarter bacteria was added to the cheese to see if it impacted the quality of the cheese.

The researchers decided to use four types of nonstarter bacteria: a control (no nonstarter bacteria added), addition of R50, addition of R21, and addition of a blend of R50 and R21. Twelve containers of cheeses were made, three of each of the four types of nonstarter bacteria, with the type of bacteria randomly assigned to the cheese containers. Each of the 12 containers of cheese was then divided into four portions. The four portions are then randomly assigned to one of four aging times: 1 day, 28 days, 56 days, and 84 days. At the end of the specified aging period, the cheese is measured for total free amino acids. The researcher was particularly interested in the bacterial effects and their interaction with aging times.

Bacteria	Container	Days			
		1	28	56	84
Control	1	0.637	1.250	1.697	2.892
	2	0.549	0.794	1.601	2.922
	3	0.604	0.871	1.830	3.198
R50	1	0.678	1.062	2.032	2.567
	2	0.736	0.817	2.017	3.000
	3	0.659	0.968	2.409	3.022
R21	1	0.607	1.228	2.211	3.705
	2	0.661	0.944	1.673	2.905
	3	0.755	0.924	1.973	2.478
R50+R21	1	0.643	1.100	2.091	3.757
	2	0.581	1.245	2.255	3.891
	3	0.754	0.968	2.987	3.322

**PROBLEM I. Part B:**

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations.

**SITUATION:**

- (1) A CRD with three factors:  $F_1$ -fixed levels,  $F_2$ -fixed levels,  $F_3$ -fixed levels, was conducted. The experimenter obtained the following results from the AOV  $F$ -tests:  $F_1 * F_2 * F_3$  is significant,  $F_1 * F_2$ -not significant,  $F_1 * F_3$ -significant,  $F_2 * F_3$ - not significant, and  $F_1, F_2, F_3$  are all not significant. She wants to determine which pairs of means are different across the levels of  $F_1$ .
- (2) A three factor experiment is run with Factor  $F_1$ -fixed, Factor  $F_2$ -fixed having levels nested within the levels of Factor  $F_1$ , Factor  $F_3$ -fixed crossed with Factor  $F_1$ . The interaction between factors  $F_1$  and  $F_3$  was not significant and the interaction between factors  $F_2(F_1)$  and  $F_3$  was not significant. The researcher was interested in determining which pairs of means are different across the levels of  $F_1$ .
- (3) A RCBD with three factors:  $F_1$ -fixed,  $F_2$ -fixed,  $F_3$ -random, was conducted. The experimenter obtained the following results from the AOV  $F$ -tests:  $F_1 * F_2 * F_3$  is not significant,  $F_1 * F_2$ -not significant,  $F_1 * F_3$ -significant,  $F_2 * F_3$ -significant, and  $F_1, F_2, F_3$  are all not significant. She wants to determine if there are pairwise differences in the levels of  $F_1$ .
- (4) In a quality control experiment, the production engineer was interested in evaluating factors which may have caused a high defective rate in a product. There are five Rates,  $F_1$ , at which a platinum coating is applied to the product, with levels, 1.0, 1.1, 1.2, 1.3, 1.4 mm/second. The second factor,  $F_2$ , is four Types of Machines used to apply the coating to the product, with levels, M1, M2, M3, M4. The third factor,  $F_3$ , was the Operators the coating machines. Twenty operators of the coating machines were randomly selected from the workforce. Each operator applied the coating to 80 units of the product, four units for each combination of a Rate and Type of Machine. There was significant evidence of a 3-factor interaction and all 2-factor interactions were found to be significant. The company wants to know if the mean defective rate,  $\mu_{ijk}$ , increased as the Rate,  $F_1$ , of applying the coating was increased.
- (5) An experiment is designed to investigate plant growth involving a factorial treatment structure with factor  $F_1$ , the temperature in a growth chamber,  $15^{\circ}C, 20^{\circ}C, 30^{\circ}C, 35^{\circ}C$  crossed with factor  $F_2$ , four brands of growth stimulants at three dose levels: 0 ml/mg, 10 ml/mg, 15 ml/mg, factor  $F_3$ . The experiment was conducted as a completely randomized design with 10 flowers randomly assigned to each of the 36 treatments. The experimenter determined from the AOV  $F$ -tests that only the following effects were significant:  $F_1 * F_3, F_1 * F_2, F_2, F_3$ . The researcher wants to determine the temperature that yields the maximum mean growth.

**TECHNIQUE:**

- A. Trend analysis using Scheffe contrasts
- B. Trend analysis using Bonferroni contrasts
- C. Trend analysis in the levels of  $F_1$  averaged over levels of the other factors
- D. Trend analysis in the levels of  $F_1$  separately at each level of the other factors
- E. Trend analysis in the levels of  $F_1$  separately at each level of  $F_2$  but averaged over the other factors
- F. Scheffe's test for contrast differences
- G. Dunnett's comparison technique
- H. Dunnett's comparison technique to all combinations of the factors
- I. Dunnett's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- J. Dunnett's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- K. Tukey's comparison technique
- L. Tukey's comparison technique to all combinations of the factors
- M. Tukey's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- N. Tukey's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- O. Hsu's comparison technique
- P. Hsu's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- Q. Hsu's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- R. Hsu's comparison technique applied to all combinations of the factors
- S. Nothing new is learned beyond the results of the F-tests from the AOV table.
- T. Comparison of marginal means is not appropriate.
- U. None of the above methods are appropriate.

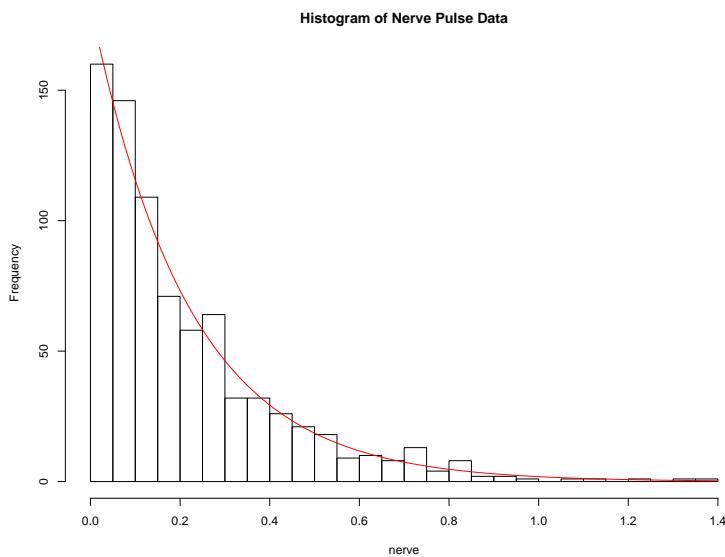
## PROBLEM II.

Consider a set of observations  $Y_1, Y_2, \dots, Y_{400}$ , which are assumed to be independent and identically distributed with a mean  $\mu$  and variance  $\sigma^2$ . NOTE that due to the large sample size involved here, you may use normal-distribution tables (distributed with this examination) for any probability calculations or decision rules required below.

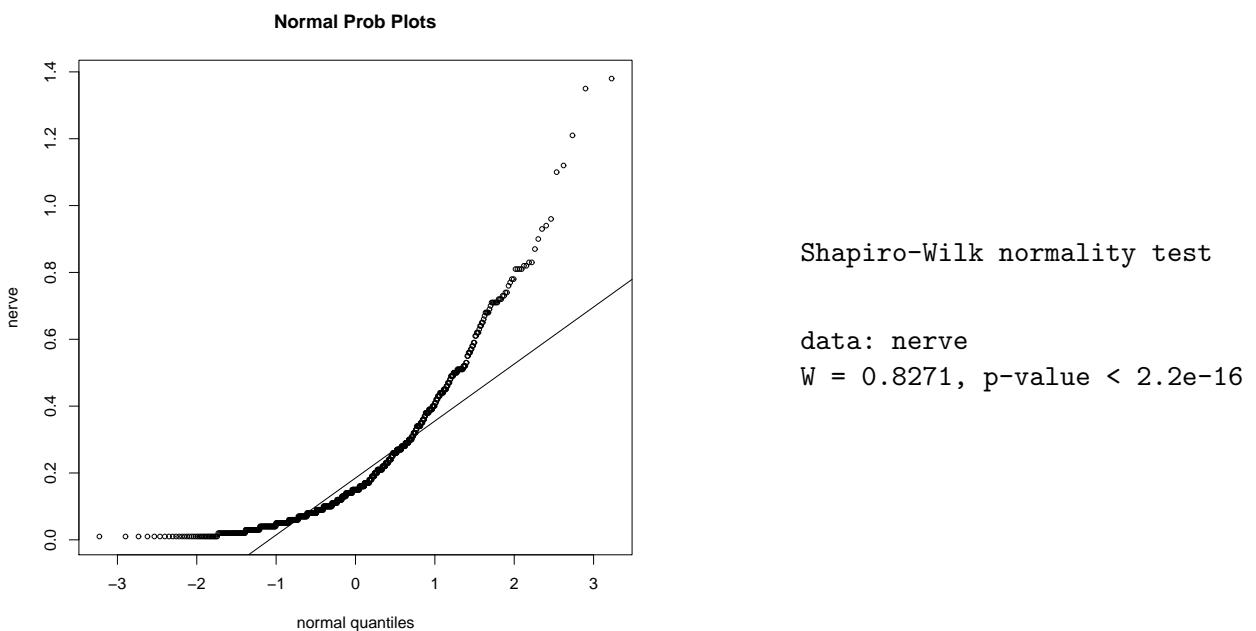
1. Consider the hypotheses  $H_0 : \mu = 12$  versus the two-sided alternative hypothesis  $H_1 : \mu \neq 12$ .
  - a. Write down a general formula for the  $t$  test statistic commonly used for this hypothesis test.
  - b. Write down the decision rule for this hypothesis test. Use  $\alpha = 0.05$ .
2. For this part of the question, you may assume that  $\sigma$  is known with  $\sigma = 1$ .
  - a. Calculate the power of your test for the following six values of the true parameter  $\mu$ : 11.9, 11.95, 11.975, 12.025, 12.05, 12.1.
  - b. Use your results from part 3.a to sketch a power curve for your test. Be sure to label your axes clearly.
3. An agronomist reviews your work from parts 1. through 2. and comments, “The power you have for  $\mu = 12.025$  is lower than I was hoping to get. How can I increase it?” Answer your agronomist’s question by providing at least two ways in which the power can be increased.
4. The 400 observations considered above represent the weight (in grams) of pecans (a type of nut). However, (unknown to you) the 400 pecans were actually collected from 10 trees, with 40 pecans picked from each tree. Also, your agronomist admits that within a given tree, pecan weights cannot be considered independent, and will have a strong *positive* correlation, due to common genetic and environmental factors. Given this additional information, answer the following questions without carrying out additional calculations.
  - a. How will this positive correlation within trees affect the expectation of the variance estimator of the sample mean that you used in part 1.a.?
  - b. Suppose you ignored the positive correlation within trees and proceeded to use the test you proposed in part 1. Will the actual values of the power of the test be larger or smaller than the values you calculated in part 2.? Explain.
5. In light of your answer to part 4., your agronomist states, “OK, I see that it’s wrong to use the test statistics from part 1. to evaluate the hypotheses. What should I do instead?” Answer your agronomist’s question by presenting a standard testing method that will account appropriately for the correlated data described in part 4. Be sure to give clear, explicit statements of both your test statistic formula and your decision rule.

### PROBLEM III.

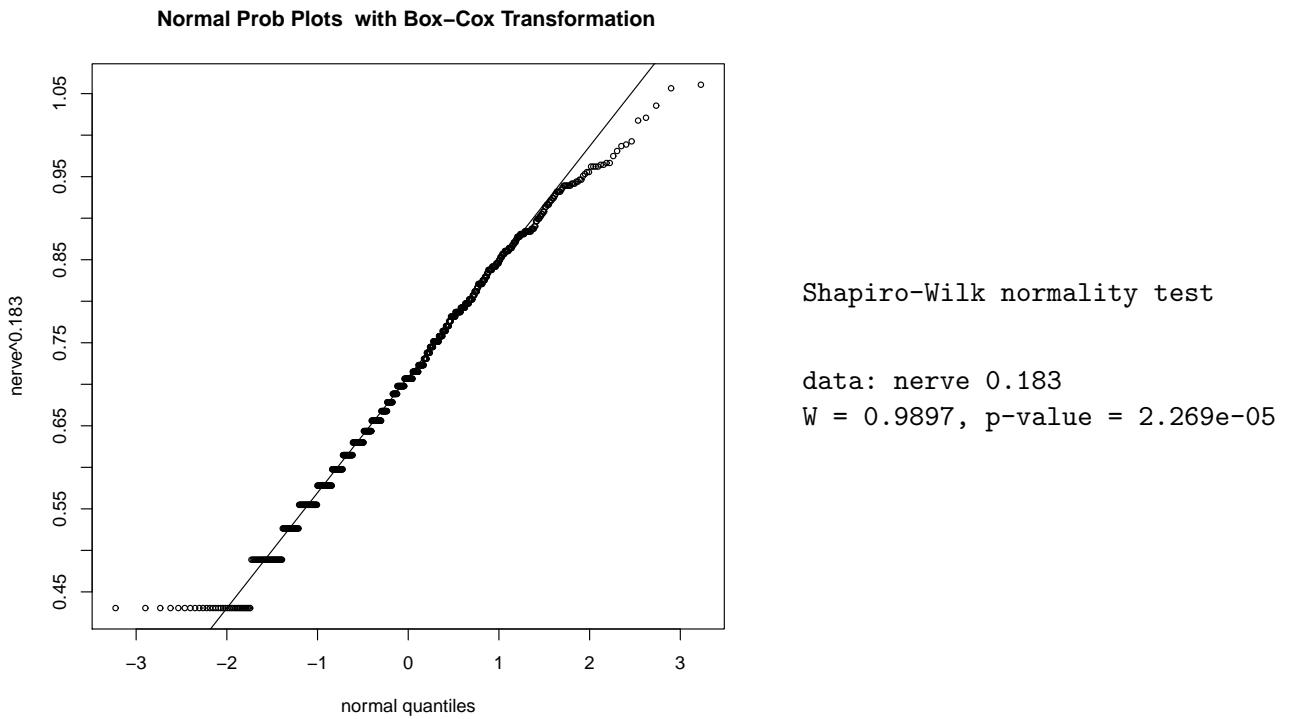
We consider a data set consisting of 799 waiting times between successive pulses along a nerve fiber. The data appear in the following histogram with a smooth curve superimposed on the histogram.



- a.) QQ plots and goodness-of-fit tests are often used to assess distributional assumptions. The following normal quantile plots were produced in R using the command `qqnorm`. Also, the command `shapiro.test` was used on the data set. Use following plot and the results of the R command `shapiro.test` to discuss the assumption of normality for the pictured data. If the data are nonnormal, describe the manner in which the data are nonnormal.



- b.) The Box-Cox transformation is often used to produce a data set that is better fit by the normal distribution. For this data set, the value of the Box-Cox parameter that produced the best fit was  $\lambda = 0.183$ . Describe carefully the transformation that corresponds to this value. Then use the following plot and results of the R command `shapiro.test` to discuss the assumption of normality for the transformed data.



- c.) Looking for a transformation to normality may not be appropriate for this data set. Discuss some reasons why the exponential distribution with cumulative distribution function

$$F(x) = \begin{cases} 1 - e^{-\beta x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

or the Weibull distribution with cumulative distribution function

$$F(x) = \begin{cases} 1 - e^{-(x/\alpha)^\gamma} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

might be more reasonable for the above data.

- d.) The normal qq-plot in parts a. and b. was used to assess the fit of the normal distribution to the data. Discuss similar plots that could be used to assess the fit of the exponential or the Weibull distributions to the data. Be sure to explain why these plots are appropriate.

#### PROBLEM IV.

The World Health Organization defines low birth weight as a baby weighing less than 2500 grams at birth. This data was taken from [Stat Labs: Mathematical Statistics Through Applications](#) by Deb Nolan and Terry Speed, University of California, Berkeley. This data is available at <http://www.statsci.org/datasets.html>.

We want to determine if gestation, age, weight, height and the mother smoking can be used to predict low birth weight.

Variable	Description
bwt	1 if less than 2500 grams ; 0 otherwise
Gestation	Length of pregnancy in days
age	mother's age in years
height	mother's height in inches
weight	Mother's prepregnancy weight in pounds
smoke	Smoking status of mother 0=not now, 1=yes now

1. Since the dependent variable (low birth weight – bwt) is a one (1) or a zero (0), multiple linear regression is not appropriate. Please give some reasons why multiple linear regression is not appropriate.

2. What is the logit equation for the model that predicts low birth weight from the predictors given above?
3. Given the graphics on pages 10, 11, 12, and 13; the researcher decided to use a log transformation on the predictor variables. Does this seem reasonable? Please explain your answer.
4. Suppose that your model for low birth weight contains the following predictor variables:

<b>smoke</b>
<b>LOG_GEST</b>
<b>LOG_AGE</b>
<b>LOG_HEIGHT</b>
<b>LOG_WEIGHT</b>

Given the table below the researcher wants to delete all three of the non-significant variables all at once. Is this the correct approach? Explain your answer.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	109.0	18.7220	33.8971	<.0001
smoke	0	1	-1.2378	0.3058	16.3821
LOG_GEST	1	-18.4168	2.1783	71.4834	<.0001
LOG_AGE	1	0.8558	0.6826	1.5719	0.2099
LOG_HEIGHT	1	-2.1324	3.9408	0.2928	0.5884
LOG_WEIGHT	1	-0.4369	1.0929	0.1598	0.6893

5. Suppose that your model for low birth weight has just smoke and **LOG\_GEST** as predictors.

Page 14 has the marginal model plots. Do these indicate a valid model? What transformation would you suggest?

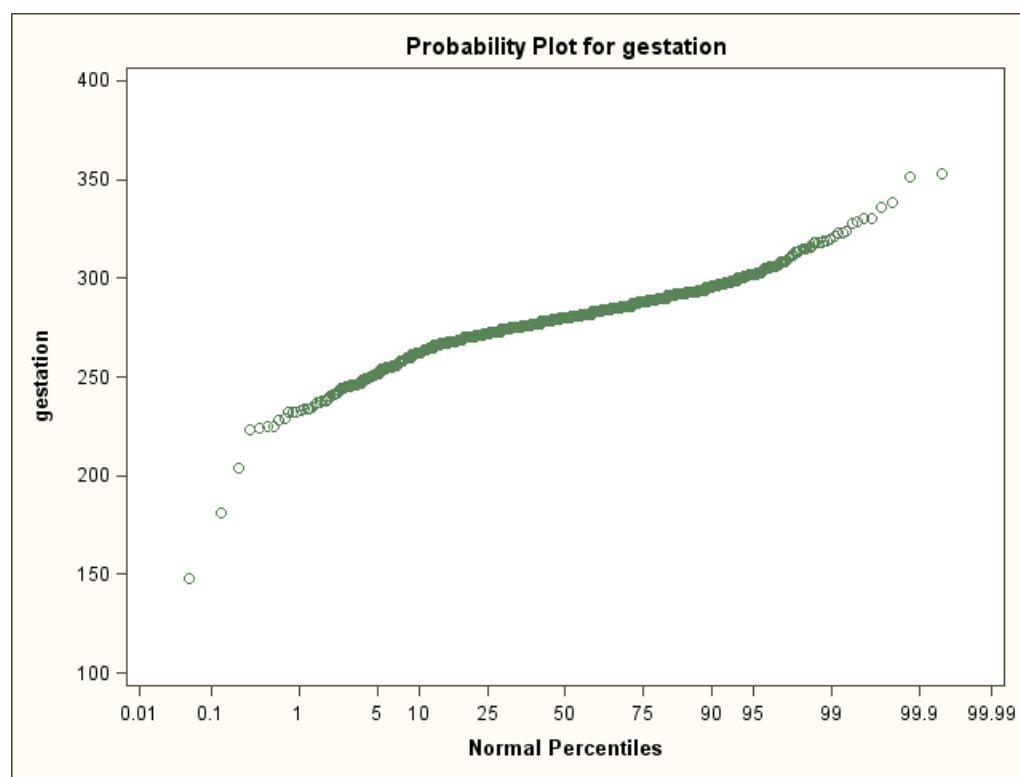
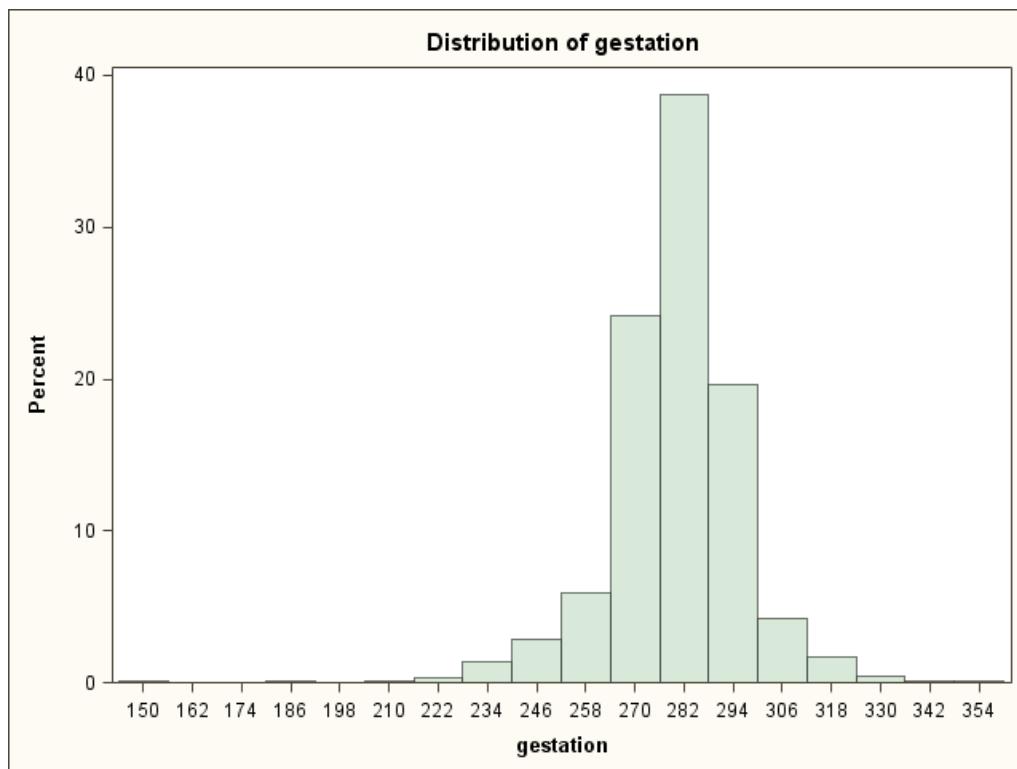
6. After an appropriate transformation, the marginal model plots on page 15 were obtained. Do these indicate a valid model? Explain your answer.

7. If the probability of a low birth weight baby is .437 for a log gestation of 5.5 for a woman who smokes and the probability of a low birth weight baby is .204 for a log gestation of 5.5 for a woman who does not smoke, what is the odds ratio of getting a low birth weight baby?

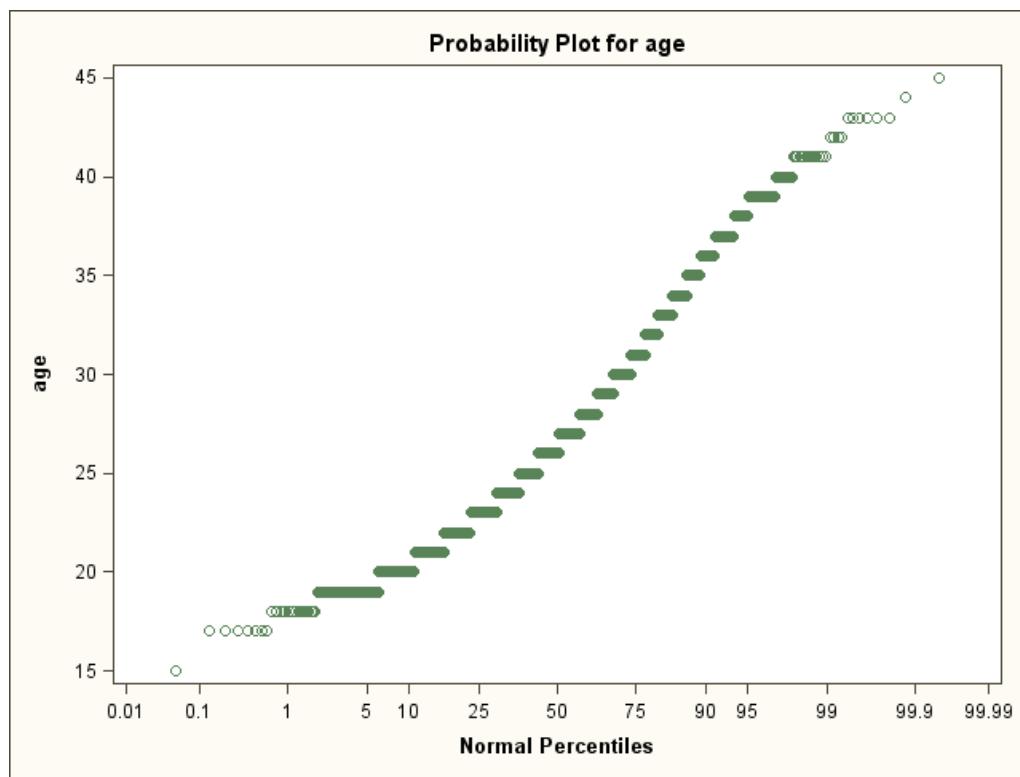
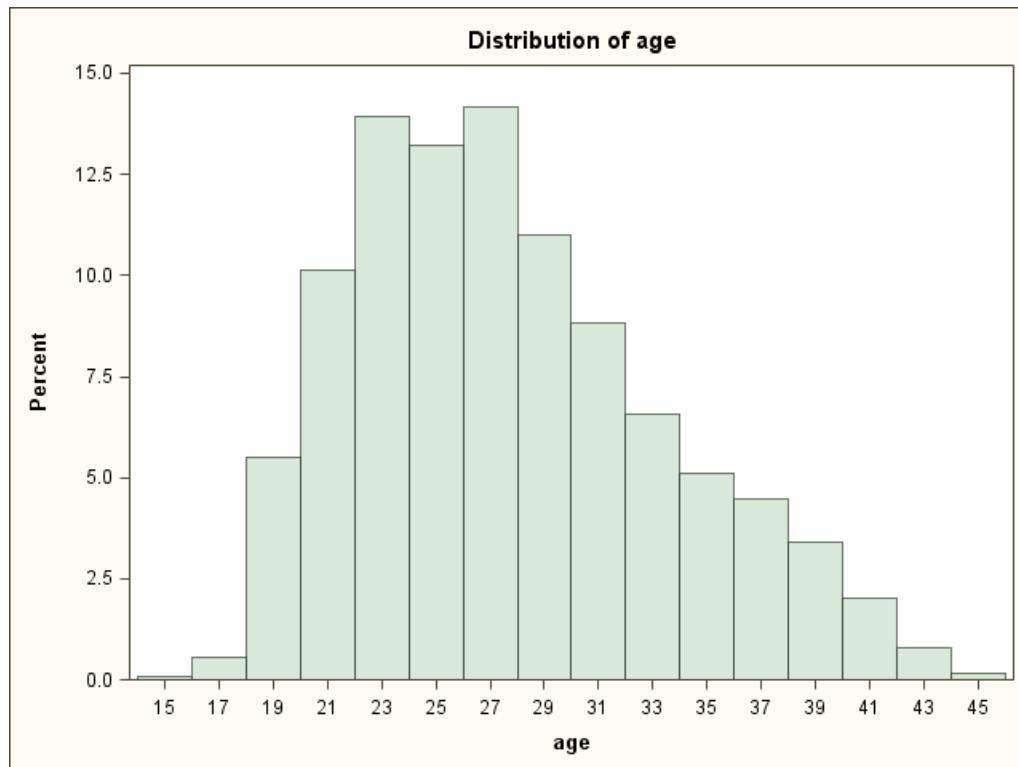
8. If the probability of a low birth weight baby is .559 for a log gestation of 5.48 for a woman who smokes and the probability of a low birth weight baby is .295 for a log gestation of 5.48 for a woman who does not smoke, what is the odds ratio of getting a low birth weight baby?

9. A researcher found that the odds ratio in parts 7 and part 8 are identical. How can you explain this phenomenon?

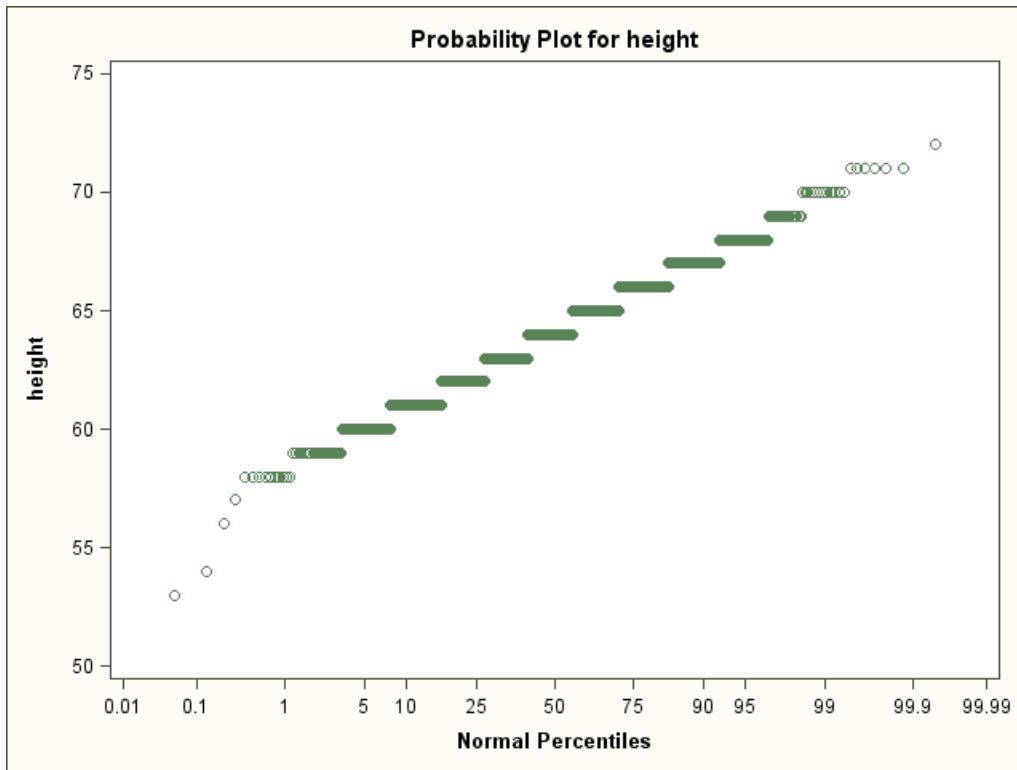
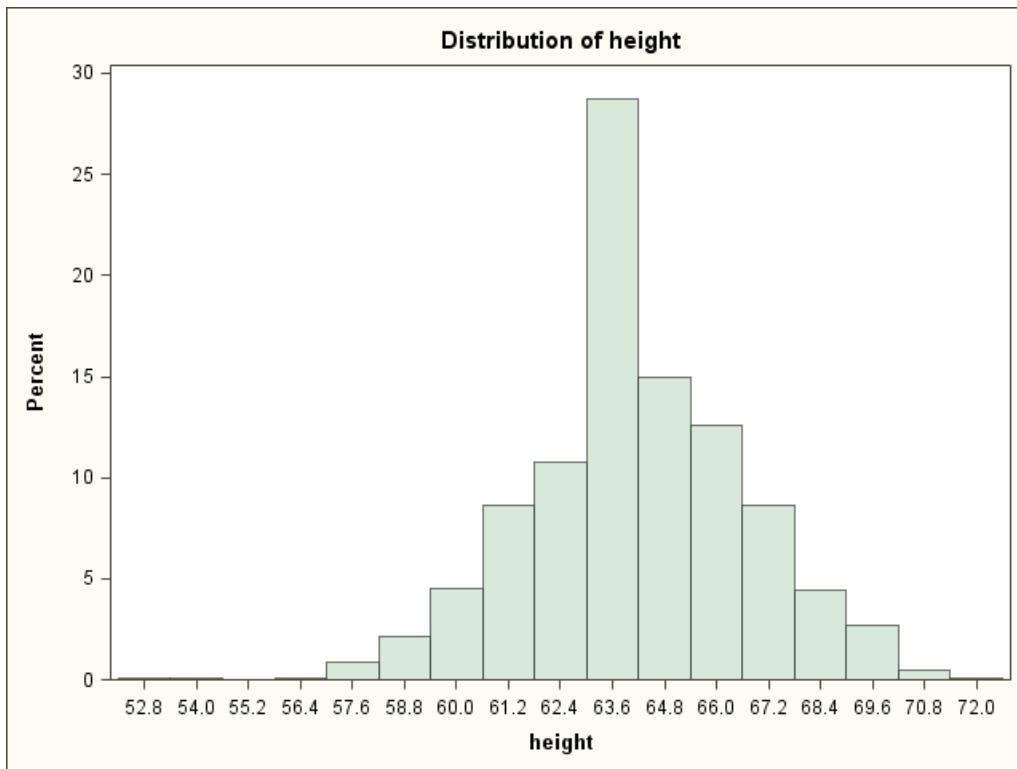
GESTATION:



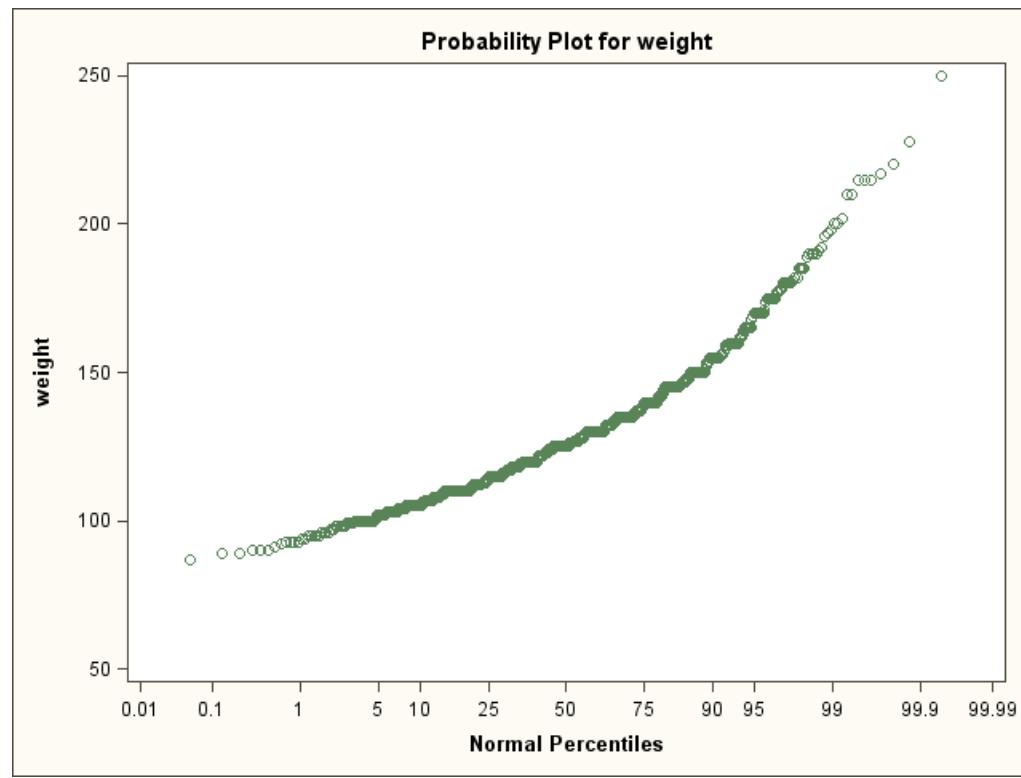
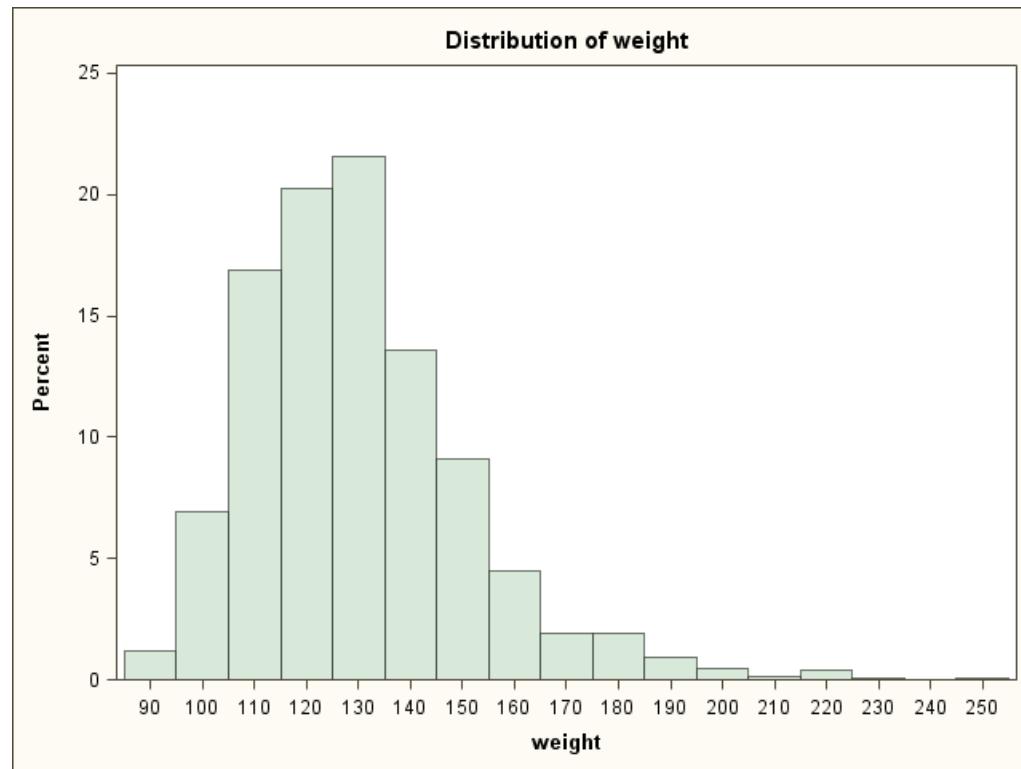
AGE:

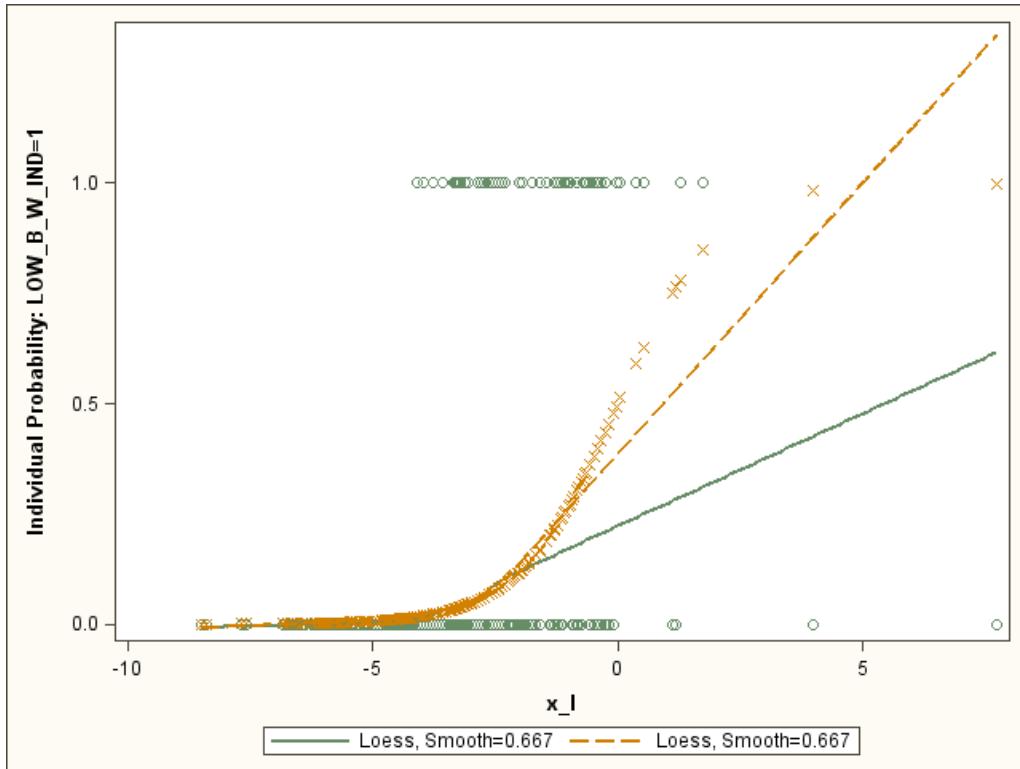
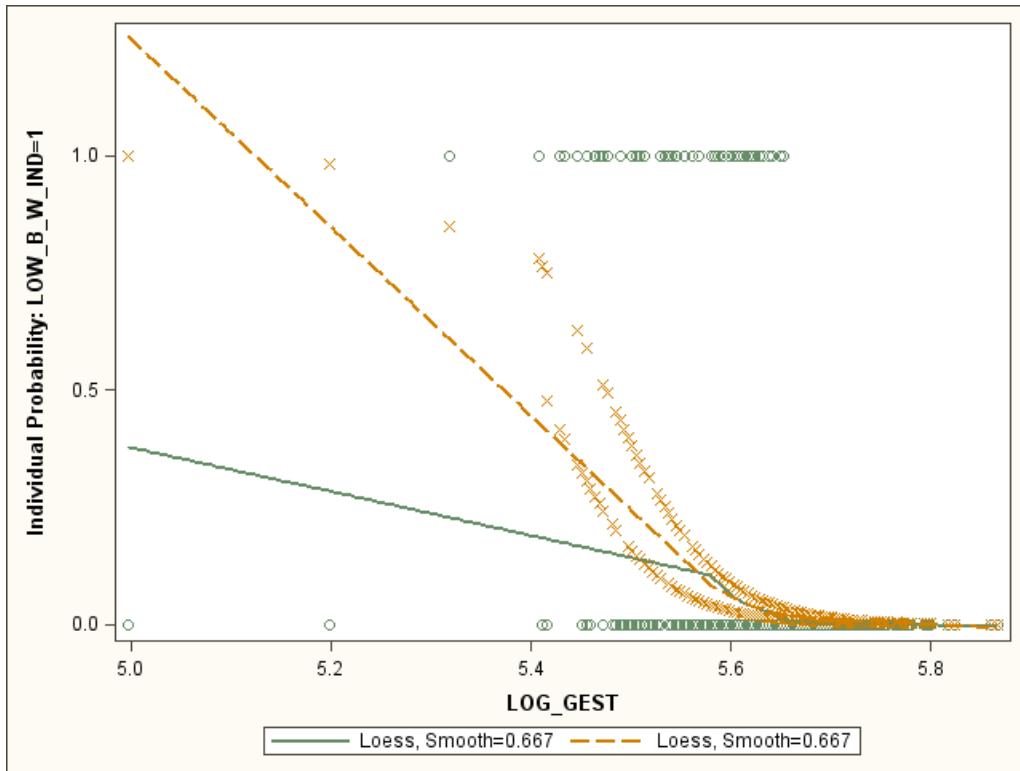


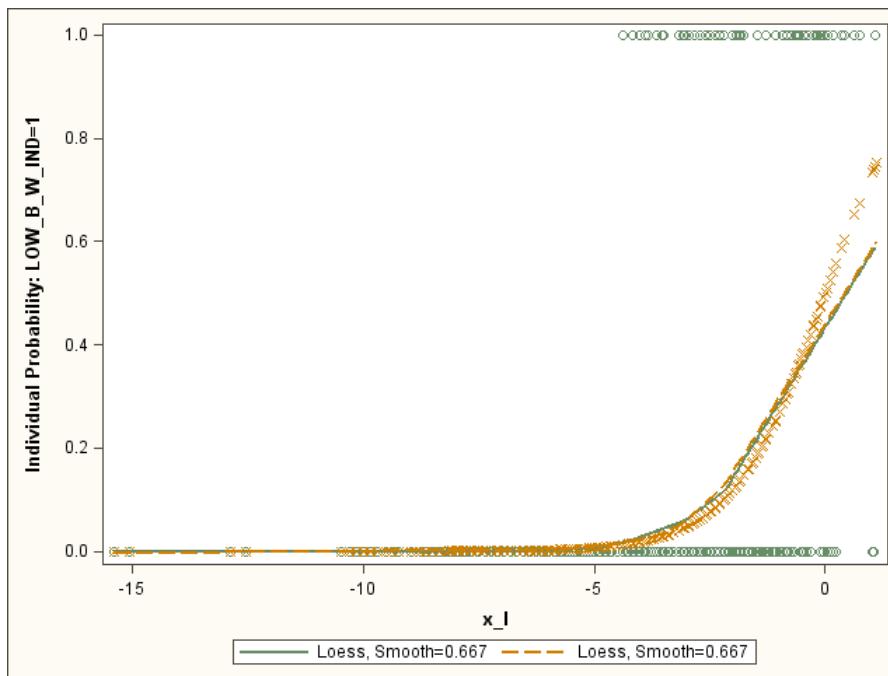
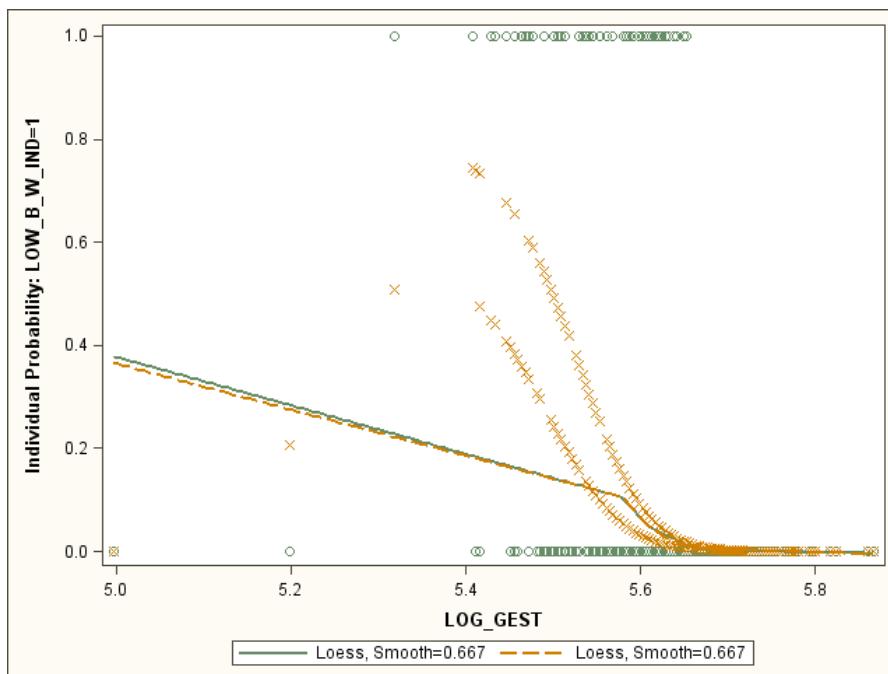
HEIGHT:



WEIGHT:







## PROBLEM V.

Let

$$Y_1 = \alpha_1 + \varepsilon_1$$

$$Y_2 = 2\alpha_1 - \alpha_2 + \varepsilon_2$$

$$Y_3 = \alpha_1 + 2\alpha_2 + \varepsilon_3 ,$$

where  $\varepsilon_1, \varepsilon_2$  and  $\varepsilon_3$  are iid  $N(0, \sigma^2)$  random variables. Derive the F statistic for testing

$$H_0 : \alpha_1 = \alpha_2 .$$

Note: “Derive the F statistic” means to produce an expression which is a function only of real numbers and the random variables  $Y_1, Y_2$ , and  $Y_3$ . You may (and are encouraged to) introduce simplifying notation in your derivation. However, be sure to carefully define, in terms of real numbers and  $Y_1, Y_2$ , and  $Y_3$ , all notation that you introduce.

Hint: Consider the usual multiple linear regression model,  $Y = X\beta + \varepsilon$ .  $Y$  is an  $n \times 1$  vector of response variables,  $X$  is an  $n \times p$  matrix (of rank  $p$ ) of predictor variables,  $\beta$  is a  $p \times 1$  vector of unknown parameters and  $\varepsilon$  is an  $n \times 1$  vector of unobservable independent and identically normally distributed random variables, each with mean zero and variance  $\sigma^2$ . Either of two (equivalent) forms of the F statistic for testing the null hypothesis  $H_0 : A\beta = c$  versus  $H_1 : A\beta \neq c$ , where  $A$  is a known  $q \times p$  matrix of rank  $q$ , are

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{\left( A\hat{\beta} - c \right)^T \left[ A \left( X^T X \right)^{-1} A^T \right]^{-1} \left( A\hat{\beta} - c \right) / q}{RSS/(n-p)},$$

where  $RSS$  and  $RSS_H$  are the residual sum of squares for the least squares fit of the unconstrained model and the model constrained by  $H_0 : A\beta = c$ , respectively.

**Tables of the Normal Distribution**

z	Probability Content from 0.00 to z										
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	

Far Right Tail Probabilities											
z	P{Z to $\infty$ }	z	P{Z to $\infty$ }	z	P{Z to $\infty$ }	z	P{Z to $\infty$ }				
2.0	0.02275	3.0	0.001350	4.0	0.00003167	5.0	2.867 E-7				
2.1	0.01786	3.1	0.0009676	4.1	0.00002066	5.5	1.899 E-8				
2.2	0.01390	3.2	0.0006871	4.2	0.00001335	6.0	9.866 E-10				
2.3	0.01072	3.3	0.0004834	4.3	0.00000854	6.5	4.016 E-11				
2.4	0.00820	3.4	0.0003369	4.4	0.000005413	7.0	1.280 E-12				
2.5	0.00621	3.5	0.0002326	4.5	0.000003398	7.5	3.191 E-14				
2.6	0.004661	3.6	0.0001591	4.6	0.000002112	8.0	6.221 E-16				
2.7	0.003467	3.7	0.0001078	4.7	0.000001300	8.5	9.480 E-18				
2.8	0.002555	3.8	0.00007235	4.8	7.933 E-7	9.0	1.129 E-19				
2.9	0.001866	3.9	0.00004810	4.9	4.792 E-7	9.5	1.049 E-21				

# **MASTER'S DIAGNOSTIC EXAMINATION**

**January 2011**

Student's Name \_\_\_\_\_

## **INSTRUCTIONS FOR STUDENTS:**

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer all four questions: Questions I, II, III and IV.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature\_\_\_\_\_

## **INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to **longneck@stat.tamu.edu** Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu**.

Proctor's Signature\_\_\_\_\_

## PROBLEM I.

A computer programmer claims that  $U_1, \dots, U_{30}$  is a random sample of size 30 from a uniform (0,1) distribution.

1. Describe a graphical method to evaluate the programmer's claim. Be sure to label your axes.
2. Describe a test of hypotheses to evaluate the programmer's claim.
3. If  $U_1, \dots, U_{30}$  are determined to be in fact a random sample from a uniform on (0,1) distribution, show how  $U_1, \dots, U_{30}$  could be used to generate a random sample of 30 observations,  $Y_1, \dots, Y_{30}$  from a distribution having cdf given by

$$F(y) = 1 - \exp(-\alpha(y - \beta)) \text{ if } y \geq \beta,$$

where  $\alpha$  and  $\beta$  are known constants.

4. Suppose we have  $n$  independent observations,  $Y_1, \dots, Y_n$  on a random variable having cdf,  $F(y)$ . Describe a graphical procedure to evaluate whether  $F(y)$  has the form:

$$F(y) = 1 - \exp(-\alpha(y - \beta)) \text{ if } y \geq \beta,$$

where  $\alpha$  and  $\beta$  are **unknown** constants.

5. How would you modify your graphical procedure in Part (4.) if you were told that the data was Type I censored data with  $m < n$  of the observations censored?

## PROBLEM II:

There are two experimental situations described below. For each of the experiments provide the following information:

1. Type of Randomization, for example, CRD, RCB, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units;
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures;
6. A partial ANOVA Table containing just Sources of Variation (SV) and Degrees of Freedom (DF);

**Experiment A:** A research specialist for a large seafood company investigated bacterial growth on oysters and mussels subjected to three different storage temperatures. Nine cold storage units were available. Three storage units were randomly assigned to be used for each of the storage temperatures: 0, 5, and 10 degrees C. Oysters and mussels were stored for two weeks in each of the cold storage units. A bacterial count was made from a sample of oysters and a sample of mussels from each storage unit at the end of two weeks so that for each storage unit there is a bacterial value for oysters and a bacterial value for mussels, yielding a total of 18 observations.

**Experiment B:** A study was designed to compare the effect of a vitamin E supplement on the growth of guinea pigs. There are 15 guinea pigs available for the study. The guinea pigs are randomly assigned to one of the three dose levels of vitamin E with 5 animals per level. For each animal the body weight was recorded at the end of weeks 1, 3, 4, 5, 6, and 7. All 15 animals were given a growth-inhibiting substance during week 1 and given identical diets during the first four weeks of the study. At the beginning of week 5, the vitamin E treatments were implemented. The three treatment levels (doses of vitamin E) were 0, L (low), and H (high). The data include the response variable WEIGHT for each of the 15 animals for each of the 6 weekly weighings (total of 90 measurements). The other information available for each observation are the levels of DOSE (0, L, H) and the WEEK (1, 3, 4, 5, 6, 7). The animals are numbered 1 through 15. In addition, a variable called BEFAFT is created which has the following values:

BEFAFT = B for weeks 1, 3, and 4, that is, before the start of the vitamin E doses

BEFAFT = A for weeks 5, 6, and 7, that is, after starting the vitamin E doses

### **PROBLEM III:**

1. A researcher, who has run a simple linear regression through the origin, states "My residuals do not sum to 0. Thus my model assumptions have been violated. I must transform Y or X." Is this a correct statement? Must the residuals sum to 0? Please explain your answer.
  2. A researcher believes that the log of the odds is a linear combination of  $x$ ,  $x^*x$  and  $\log x$ . Write out the logit function.
  3. Carefully study the results for the three models given below. What model would you recommend to your client? Explain why. Discuss the issues with validity of the models. Is there anything you can tell the client about his predictor variables?

**Model 1:**

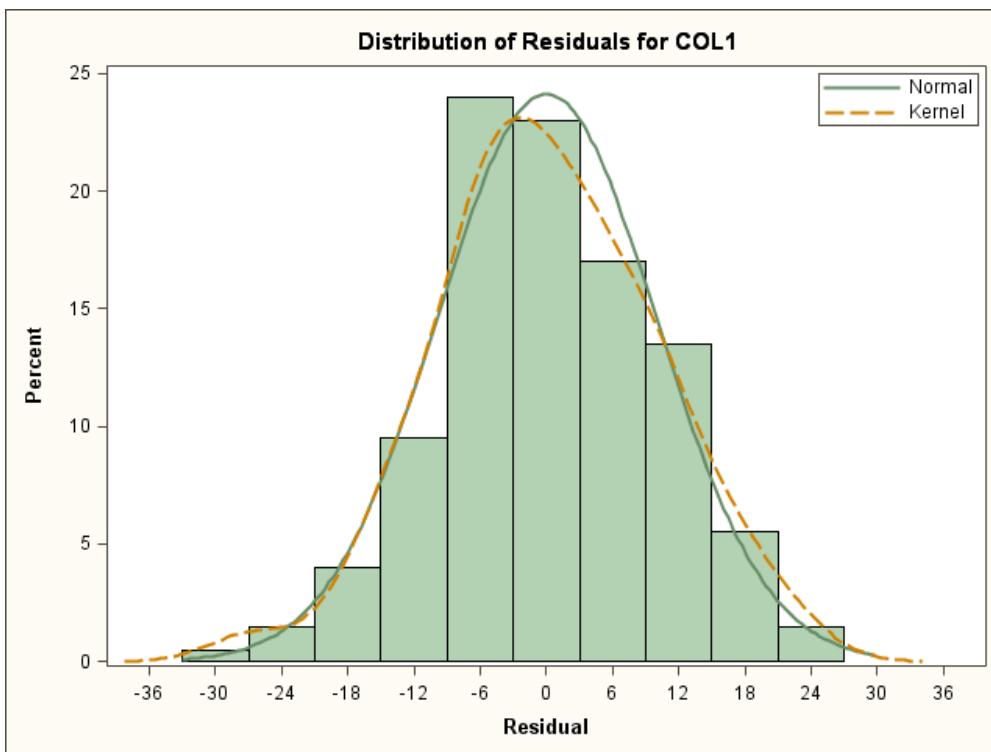
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	154584	17176	166.62	<.0001
Error	190	19586	103.08628		
Corrected Total	199	174171			

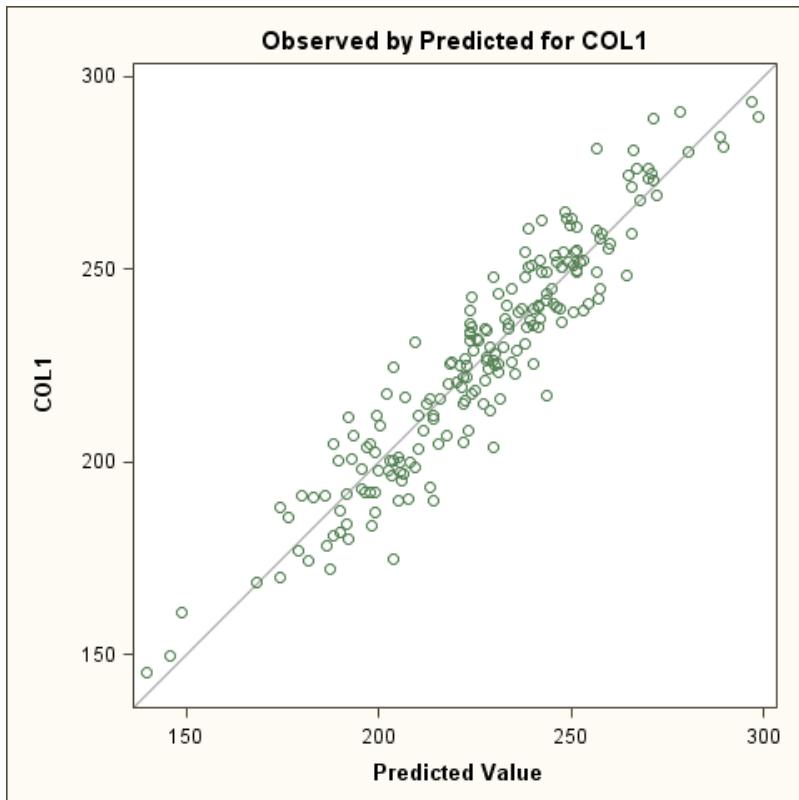
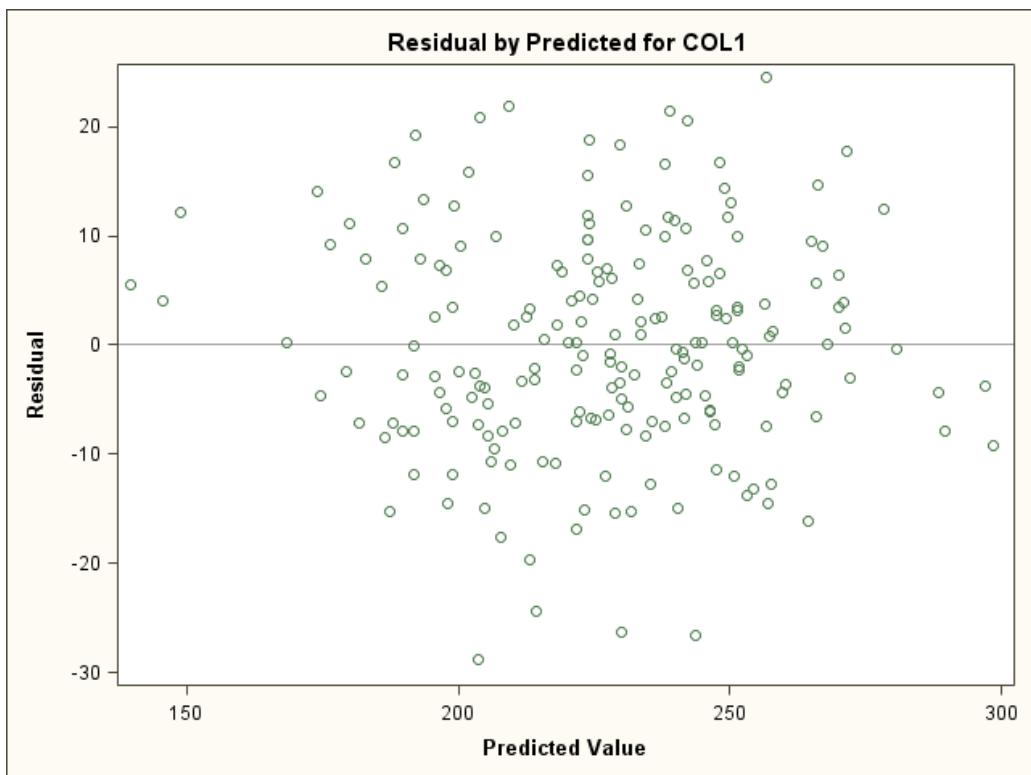
  

Root MSE	10.15314	R-Square	0.8875
Dependent Mean	227.04725	Adj R-Sq	0.8822
Coeff Var	4.47182		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	3.81085	6.13929	0.62	0.5355	0
COL2	1	0.93090	1.53281	0.61	0.5444	12.23132
COL3	1	2.46945	0.90784	2.72	0.0071	4.27408
COL4	1	3.41763	1.17315	2.91	0.0040	8.77643
COL5	1	3.68754	0.42996	8.58	<.0001	1.05793
COL6	1	4.71660	0.41765	11.29	<.0001	1.06188
COL7	1	6.02569	0.42912	14.04	<.0001	1.03296
COL8	1	7.19138	0.45166	15.92	<.0001	1.06395
COL9	1	7.95603	0.44949	17.70	<.0001	1.05266
COL10	1	4.39536	1.49354	2.94	0.0037	19.06114





## Model 2

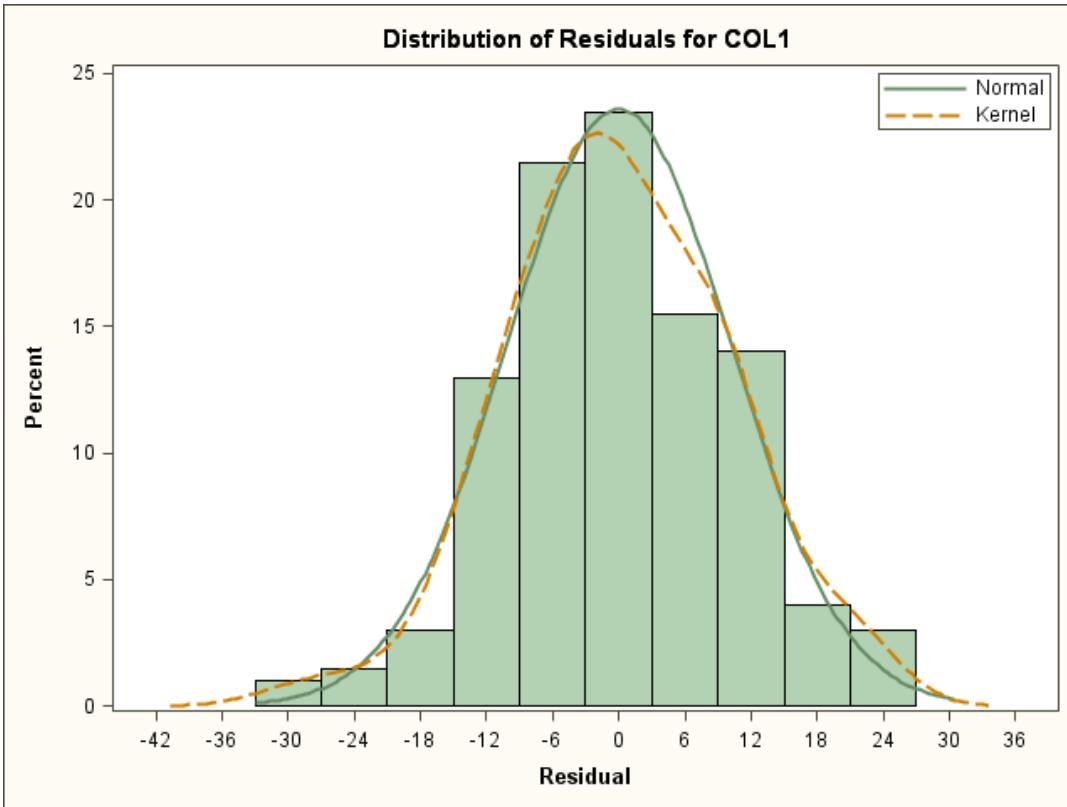
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	153692	19211	179.18	<.0001
Error	191	20479	107.22094		
Corrected Total	199	174171			

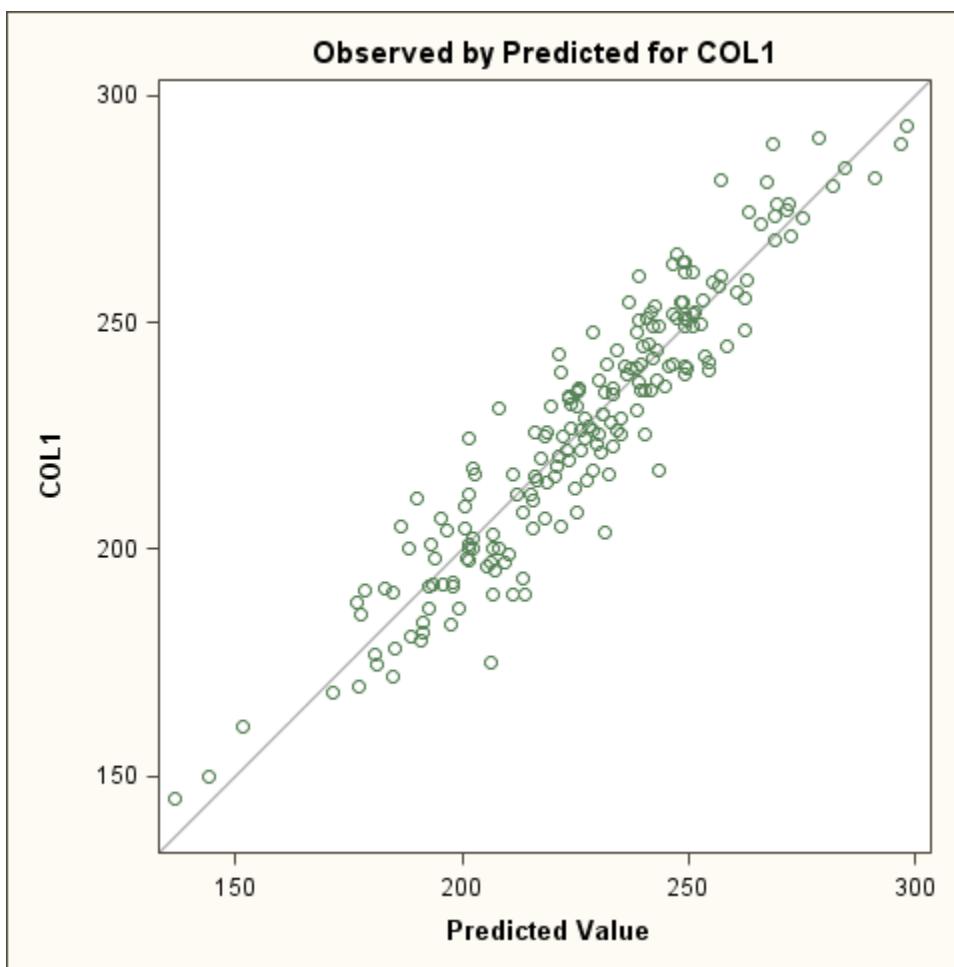
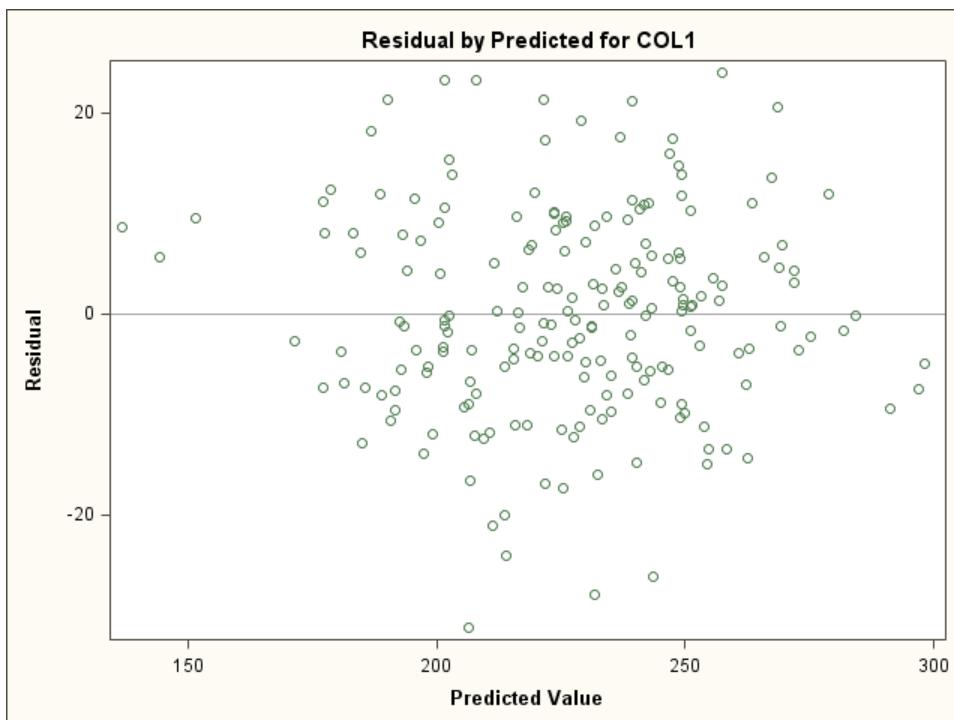
  

Root MSE	10.35475	R-Square	0.8824
Dependent Mean	227.04725	Adj R-Sq	0.8775
Coeff Var	4.56062		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	4.08374	6.26048	0.65	0.5150	0
COL2	1	5.24085	0.46140	11.36	<.0001	1.06556
COL3	1	4.79871	0.45351	10.58	<.0001	1.02544
COL4	1	6.65044	0.41997	15.84	<.0001	1.08134
COL5	1	3.60344	0.43753	8.24	<.0001	1.05326
COL6	1	4.57544	0.42313	10.81	<.0001	1.04787
COL7	1	6.08528	0.43715	13.92	<.0001	1.03066
COL8	1	7.25810	0.46005	15.78	<.0001	1.06127
COL9	1	7.96753	0.45840	17.38	<.0001	1.05259





### Model 3 – Stepwise

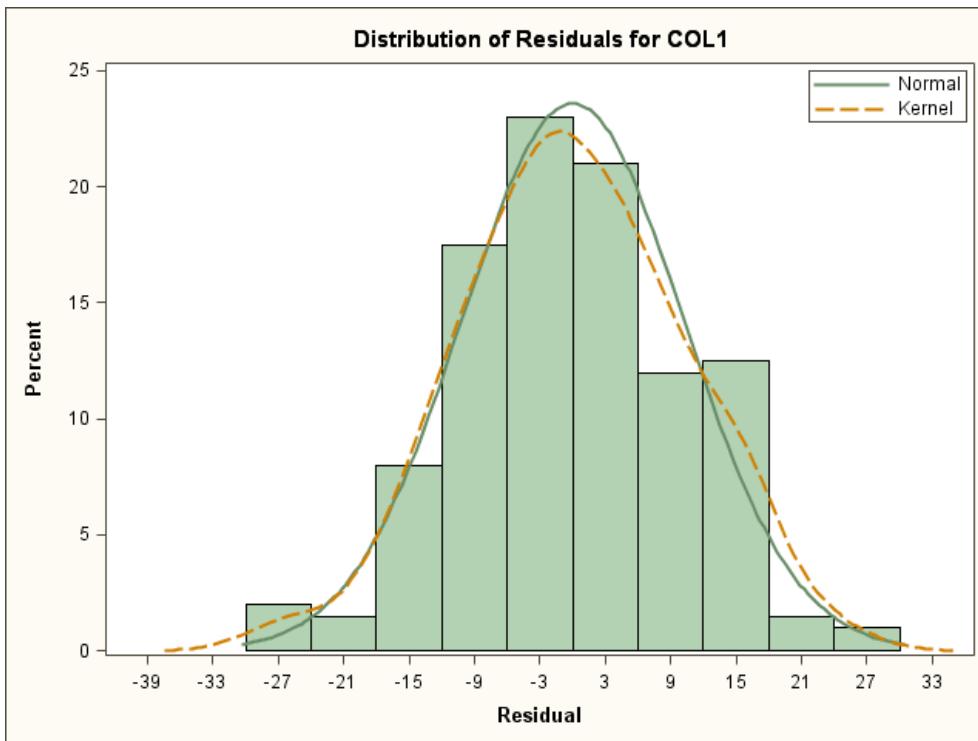
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	153676	21954	205.66	<.0001
Error	192	20495	106.74650		
Corrected Total	199	174171			

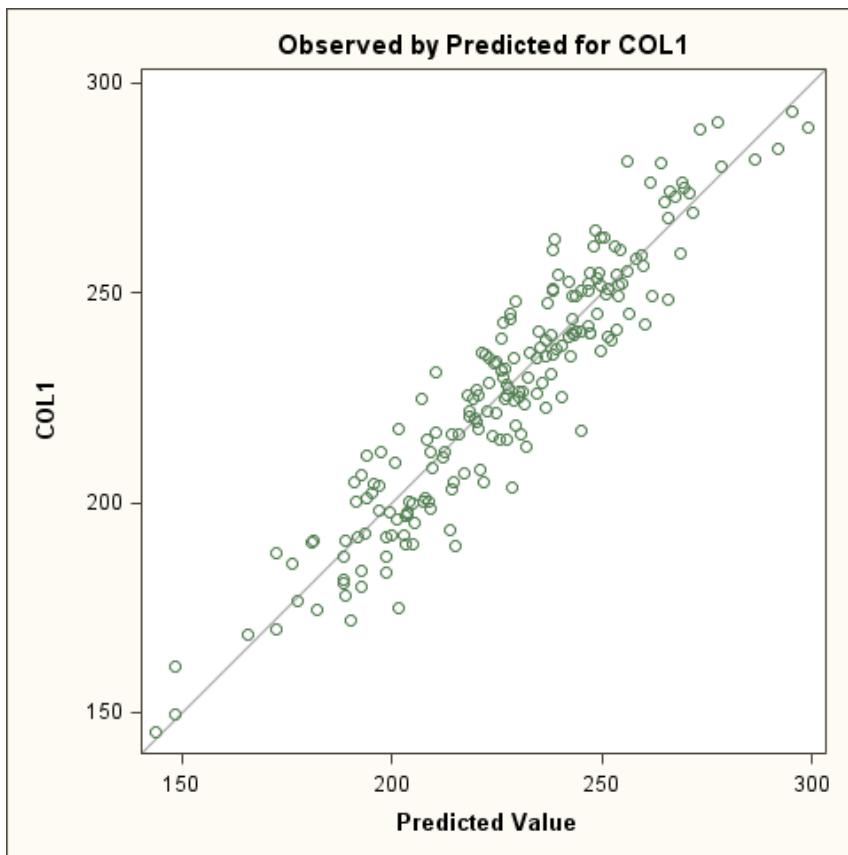
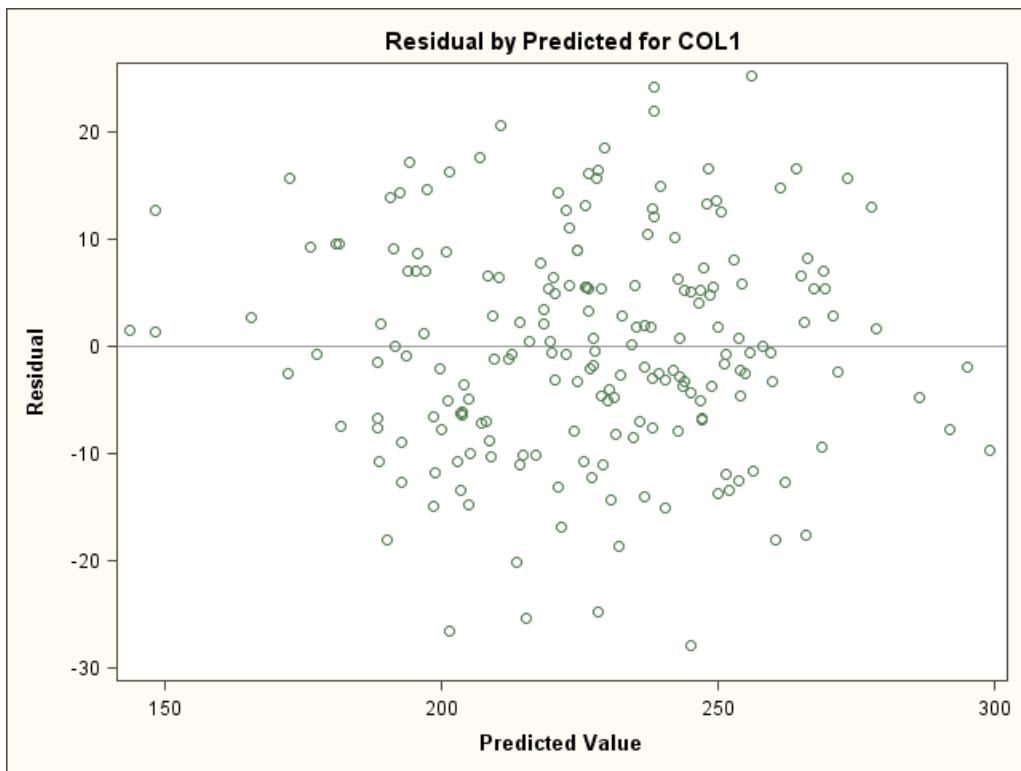
  

Root MSE	10.33182	R-Square	0.8823
Dependent Mean	227.04725	Adj R-Sq	0.8780
Coeff Var	4.55052		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	5.19190	6.22902	0.83	0.4056	0
COL2	1	-3.29178	0.58148	-5.66	<.0001	1.69988
COL5	1	3.73159	0.43698	8.54	<.0001	1.05532
COL6	1	4.88690	0.41914	11.66	<.0001	1.03276
COL7	1	5.97025	0.43561	13.71	<.0001	1.02796
COL8	1	7.05863	0.45679	15.45	<.0001	1.05090
COL9	1	7.98222	0.45577	17.51	<.0001	1.04517
COL10	1	8.62440	0.45761	18.85	<.0001	1.72803





**PROBLEM IV:** Provide complete solutions to the following two problems:

**Part 1.** A study was carried out to examine the effect of injecting Botox on eye pain. Fifteen Patients received a high-dose injection in one eye (experimental treatment E) and a low-dose injection in the other eye (control treatment C). Patients were asked to rate the level of pain in each eye on a 1 to 10 scale, with higher values indicating more pain. Which eye received which treatment was randomized. The pain scores for the two eyes recorded on the last of several visits are given in the table:

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
E eye ( $x_1$ )	1.3	7.3	0	0	3	0	3.5	0	0	2.0	0	3.0	5.0	0.3	0
C eye ( $x_2$ )	8.8	1.3	0.8	9.5	7.8	9.0	5.0	2.3	2.5	8.0	4.5	4.5	9.0	7.5	0.5

Some possibly useful statistics follow:

$$\bar{x}_1 = 1.693, s_1 = 2.2569, \bar{x}_2 = 5.400, s_2 = 3.3058, d = x_2 - x_1, \bar{d} = 3.707, s_d = 3.9773$$

1. What is the design of the experiment? Discuss its appropriateness for this experiment.
2. Obtain a 95% confidence interval for the mean difference pain scores between the two eyes. Interpret your results.
3. Another way to look at these data is to consider the percentage of subjects who have less pain with the E eye versus the C eye. Carry out a test of whether the proportion of subjects who have less pain with the E eye differs from 0.5. Interpret your results.

**Part 2.** We wish to compare two different chemotherapy regimens for breast cancer after mastectomy. Patients are placed into pairs based on age and clinical condition. Then a random member of each pair receives treatment A and the other member treatment B. The patients are followed for five years and the number surviving is recorded. The data can be recorded into either of two tables:

Table A		Outcome		Table B	
		Survive for 5 years	Die within 5 years		
Treatment	A	526	95	Outcome for A patient	Survive for 5 years
	B	515	106	Die within 5 years	16

1. Discuss which table is more appropriate for presenting these data. Explain why.
2. Carry out a test for association between treatment and survival for 5 years. You may use the SAS output on the next pages in your solution.

## The FREQ Procedure

## Table of Survive by Treatment

Survive	Treatment		
Frequency	A	B	Total
Percent	526	515	1041
yes	42.35	41.47	83.82
no	95	106	201
	7.65	8.53	16.18
Total	621	621	1242
	50.00	50.00	100.00

## Statistics for Table of Survive by Treatment

Statistic	DF	Value	Prob
Chi-Square	1	0.7182	0.3967
Likelihood Ratio Chi-Square	1	0.7185	0.3966
Continuity Adj. Chi-Square	1	0.5936	0.4410
Mantel-Haenszel Chi-Square	1	0.7176	0.3969
Phi Coefficient		0.0240	
Contingency Coefficient		0.0240	
Cramer's V		0.0240	

## Fisher's Exact Test

Cell (1,1) Frequency (F)	526
Left-sided Pr <= F	0.8224
Right-sided Pr >= F	0.2205
Table Probability (P)	0.0429
Two-sided Pr <= P	0.4411

## Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	1.1396	0.8422 1.5420
Cohort (Col1 Risk)	1.0691	0.9129 1.2520
Cohort (Col2 Risk)	0.9381	0.8118 1.0840

Sample Size = 1242

## Analysis for Table B

The FREQ Procedure

## Table of SurviveA by SurviveB

SurviveA      SurviveB

		Frequency		Total
Percent	yes	no		
yes	510 82.13	16 2.58		526 84.70
no	5 0.81	90 14.49		95 15.30
Total	515 82.93	106 17.07		621 100.00

## Statistics for Table of SurviveA by SurviveB

## McNemar's Test

---

Statistic (S)      5.7619  
 DF                  1  
 Pr > S            0.0164

## Simple Kappa Coefficient

---

Kappa	0.8754
ASE	0.0266
95% Lower Conf Limit	0.8233
95% Upper Conf Limit	0.9275

Sample Size = 621

Table A.3 Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$

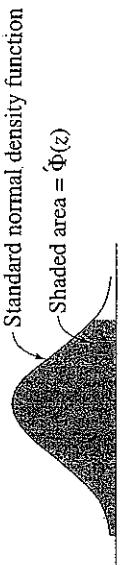
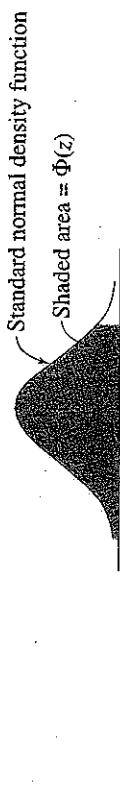


Table A.3 Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)



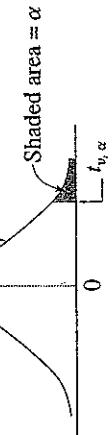
$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0024	0.0023	0.0023	0.0022	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0394	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table A.3 Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)

**Table A.1** Cumulative Binomial Probabilities (cont.)

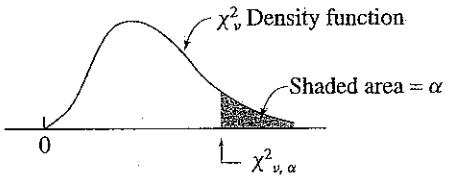
		p														
		.01	.05	.10	.20	.25	.30	.40	.50	.60	.70	.75	.80	.90	.95	.99
	0	0.860	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000
x	1	.990	.829	.549	.167	.080	.035	.015	.000	.000	.000	.000	.000	.000	.000	.000
	2	1.000	.964	.816	.398	.236	.127	.027	.004	.000	.000	.000	.000	.000	.000	.000
	3	1.000	.995	.944	.648	.461	.297	.091	.018	.002	.000	.000	.000	.000	.000	.000
	4	1.000	.999	.987	.836	.686	.515	.217	.059	.009	.001	.000	.000	.000	.000	.000
	5	1.000	1.000	.998	.939	.852	.722	.403	.151	.034	.004	.001	.000	.000	.000	.000
	6	1.000	1.000	1.000	.982	.943	.869	.610	.304	.095	.015	.004	.001	.000	.000	.000
x	7	1.000	1.000	1.000	.996	.983	.950	.787	.500	.213	.050	.017	.004	.000	.000	.000
	8	1.000	1.000	1.000	.999	.996	.985	.905	.696	.390	.131	.057	.018	.000	.000	.000
	9	1.000	1.000	1.000	1.000	.999	.996	.966	.849	.597	.278	.148	.061	.002	.000	.000
	10	1.000	1.000	1.000	1.000	1.000	.999	.991	.941	.783	.485	.314	.164	.013	.001	.000
	11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.982	.909	.703	.539	.352	.056	.005
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.973	.873	.764	.602	.184	.036
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.920	.833	.451	.171
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.987	.965	.794	.537

**Table A.4** Critical Values  $t_{v,\alpha}$  for the t-Distribution



v	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

**Table A.5** Critical Values  $\chi^2_{v,\alpha}$  for the Chi-square Distribution



v	$\alpha$									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.420	65.473
40*	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

**Table A.10** Upper-Tail Probabilities of the Null Distribution of the Wilcoxon Signed Rank Statistic

$n$	$w$	$P(W \geq w   H_0)$	$n$	$w$	$P(W \geq w   H_0)$
3	6	0.125		78	0.011
4	9	0.125		79	0.009
5	10	0.062		81	0.005
6	13	0.094	14	73	0.108
7	14	0.062		74	0.097
8	15	0.031		79	0.052
9	17	0.109		84	0.025
10	19	0.047		89	0.010
11	20	0.031		92	0.005
12	21	0.016	15	83	0.104
13	22	0.109		84	0.094
14	24	0.055		89	0.053
15	26	0.023		90	0.047
16	28	0.008		95	0.024
17	30	0.098	100	101	0.011
18	32	0.027	104	105	0.009
19	34	0.012	112	113	0.016
20	35	0.008	94	95	0.096
21	36	0.004	100	101	0.052
22	39	0.102	106	107	0.025
23	37	0.049	112	113	0.011
24	42	0.010		116	0.005
25	44	0.004		104	0.103
26	41	0.097	105	106	0.095
27	44	0.053	112	113	0.049
28	47	0.024	118	119	0.005
29	50	0.010	125	126	0.010
30	52	0.005	129	130	0.005
31	59	0.009	138	139	0.010
32	61	0.005	143	144	0.005
33	64	0.026	116	117	0.103
34	68	0.051	124	125	0.049
35	69	0.055	136	137	0.024
36	71	0.005	157	158	0.027
37	61	0.046	140	141	0.101
38	65	0.095	150	151	0.049
39	69	0.055	158	159	0.024
40	70	0.047	167	168	0.010
41	74	0.024	172	173	0.005

**Table A.11** Upper-Tail Probabilities of the Null Distribution of the Wilcoxon-Mann-Whitney Statistic

$n_1$	$n_2$	$w_1$	$u_1$	$P(U \geq u_1)$	$n_1$	$n_2$	$w_1$	$u_1$	$P(U \geq u_1)$	$n_1$	$n_2$	$w_1$	$u_1$	$P(U \geq u_1)$
$P(W \geq w_1) =$					$P(W \geq w_1) =$					$P(W \geq w_1) =$				
3	3	15	9	0.050	4	9	44	34	0.006	6	10	66	45	0.059
4	4	17	11	0.057	4	10	42	32	0.053	10	10	69	48	0.028
5	5	18	12	0.029	10	44	34	0.027	10	10	72	51	0.011	
6	6	22	16	0.048	10	46	36	0.012	10	10	74	53	0.005	
7	7	24	18	0.018	7	47	37	0.007	7	7	66	38	0.049	
8	8	28	22	0.024	5	5	36	21	0.048	8	7	68	40	0.027
9	9	30	24	0.006	6	41	26	0.026	8	7	71	43	0.009	
10	10	31	25	0.050	6	43	28	0.009	8	8	78	50	0.005	
11	10	33	27	0.024	6	44	29	0.004	9	9	75	47	0.057	
12	11	35	29	0.012	7	43	28	0.053	9	9	78	50	0.027	
13	12	36	30	0.006	7	45	30	0.024	9	9	81	53	0.011	
14	13	40	33	0.056	8	51	36	0.009	10	10	83	55	0.006	
15	14	44	37	0.024	8	52	37	0.005	10	10	80	52	0.054	
16	15	48	33	0.018	7	48	33	0.005	8	8	84	48	0.052	
17	16	50	35	0.009	8	47	32	0.047	10	10	83	55	0.028	
18	17	53	29	0.003	9	53	38	0.021	8	8	90	54	0.010	
19	18	55	33	0.005	9	55	40	0.009	8	9	92	56	0.005	
20	19	56	41	0.029	9	56	41	0.006	9	9	89	53	0.057	
21	20	57	41	0.014	10	54	39	0.056	8	8	87	51	0.025	
22	21	57	41	0.056	10	56	41	0.027	9	9	96	60	0.010	
23	22	58	48	0.032	10	59	44	0.010	9	9	98	62	0.006	
24	23	59	49	0.016	10	60	45	0.006	10	10	95	59	0.051	
25	24	60	50	0.008	6	6	29	0.047	10	10	98	62	0.027	
26	25	60	50	0.007	9	54	39	0.050	9	9	93	57	0.023	
27	26	61	51	0.005	10	55	38	0.021	8	8	90	54	0.010	
28	27	61	51	0.057	9	55	40	0.009	8	9	92	56	0.005	
29	28	62	52	0.029	9	56	41	0.006	9	9	89	53	0.057	
30	29	62	52	0.008	6	6	29	0.047	10	10	102	66	0.010	
31	30	63	53	0.007	6	52	31	0.021	10	10	104	68	0.006	
32	31	63	53	0.019	6	54	33	0.008	9	9	114	69	0.005	
33	32	64	54	0.012	7	60	39	0.004	10	10	110	65	0.056	
34	33	64	54	0.006	8	58	37	0.054	10	10	114	69	0.027	
35	34	64	54	0.005	7	54	33	0.051	9	9	111	66	0.012	
36	35	65	55	0.098	8	55	35	0.026	9	9	118	73	0.011	
37	36	65	55	0.049	6	56	35	0.021	10	10	121	76	0.005	
38	37	66	56	0.048	8	56	37	0.011	9	9	127	72	0.053	
39	38	66	56	0.048	8	57	38	0.010	10	10	131	76	0.026	
40	39	67	57	0.008	8	57	38	0.004	10	10	135	80	0.012	
41	40	67	57	0.008	9	58	39	0.025	10	10	138	83	0.006	

## MASTER'S DIAGNOSTIC EXAMINATION

August 2011

Student's Name \_\_\_\_\_

### **INSTRUCTIONS FOR STUDENTS:**

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer all four questions: Questions I, II, III and IV.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature\_\_\_\_\_

### **INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, the student's solutions should be

**faxed to 979-845-6060    or    emailed to longneck@stat.tamu.edu**

Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or  
emailed to **longneck@stat.tamu.edu**.

Proctor's Signature\_\_\_\_\_

**PROBLEM I.** In a genetics experiment, investigators examined 300 chromosomes of a particular type and counted the number of sister-chromatid exchanges on each chromosome. The following table displays the number of exchanges for each of the 300 chromosomes and information about model fit.

Number of Exchanges	0	1	2	3	4	5	6	7	8	9	Total
Number of Chromosomes	5	21	40	59	62	44	42	17	7	3	300
$\hat{p}_i$	.018	.074	.146	.195	.196	.156	.104	.060	.030	.021	1.00
$\hat{E}_i$	5.4	22.2	43.8	58.5	58.8	46.8	31.2	18.0	9.0	6.3	300
$O_i$	5	21	40	59	62	44	42	17	7	3	300
$\frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$	.030	.065	.330	.004	.174	.168	3.738	.056	.444	1.729	6.74

Summary Statistics for the number of exchanges for the 300 chromosomes were

$$N = 300$$

$$\bar{X} = 4.0$$

$$S^2 = 3.48$$

1. Using the information given above, does a Poisson model appear to be an appropriate model for the data?
  2. Estimate, using an appropriate 95% confidence interval, the average number of exchanges per chromosome.
  3. Assume that the Poisson model is the correct model. Using an  $\alpha=0.05$  test, test the hypothesis that the number of exchanges per chromosome is less than 4. What is the p-value of your test?
  4. What is the power of the test developed in Question c) if the true mean number of exchanges is 3.8 per chromosome?
  5. A new study is being designed. Determine the minimum sample size which would yield an  $\alpha = .05$  test of the research hypothesis that the mean number of exchanges is less than 4 with the test having power of at least 0.85 whenever the true value of the mean number of exchanges is 3.5 or smaller.
- Hint: You may assume that the sample size is large enough for the central limit theorem to be applicable.

**PROBLEM II.** Greens for golf courses present a unique problem for use of fertilizers in maintaining turf growth. The soil is usually very sandy and as a consequence has little capacity to retain nitrogen in the root zone after irrigation. Large initial doses of nitrogen are harmful to the grass, but slow release forms may leach out of the soil and be ineffective. A second factor that may affect nitrogen retention is the build-up of thatch, or dead grass.

A soil scientist wanted to investigate the effects of nitrogen supplied in different chemical forms and evaluate those effects combined with the effects of thatch accumulation on the quality of established turf. The first factor in his experiment is N, the form of nitrogen. He will use four forms of nitrogen, two fast release fertilizers: U = urea and AS = Ammonium Sulphate; and two slow release fertilizers: ID = Isobutylidene Diurea and SC = Sulphur Coated urea. Each fertilizer will be applied to the turf at the rate of one pound nitrogen per 1000 square feet of turf. The fertilizer will be applied each year of the experiment. The second factor is T = thatch. The thatch will be allowed to accumulate on an experimental plot for 2, 4, or 6 years. One variable of interest will be the chlorophyll content of grass clippings from the experimental plots.

The soil scientist has two choices for the experimental design:

**Design A.** This design would use a Randomized Complete Block Design with 12 treatment combinations (N at 4 levels and T at 3 levels). There would be two complete blocks (total of 24 observations).

**Design B.** This design would use a Split Plot Design with the 4 N levels as the whole plot treatments and the 3 T levels as the subplot treatments. There would be two complete blocks for the whole plot design (total of 24 observations).

1. Discuss the conditions under which each of the designs would be favored.
2. Describe the randomization procedure to be followed if the soil scientist chooses to use Design B.
3. The soil scientist uses Design B. A partial AOV table and Means table are given on the next page with notation Blocks (B), Nitrogen (N), Thatch (T). Provide the values for degree of freedom.

**AOV Table**

Source	DF	Mean Square
B		0.51
N		12.44
N*B		0.42
T		1.91
N*T		0.69
Error		0.21

**Table of Treatment Means**

	Nitrogen	Thatch			Nitrogen Mean
		2	4	6	
Nitrogen	U	3.85	5.35	5.10	4.77
	AS	5.60	5.85	5.80	5.75
	ID	6.50	6.00	7.80	6.77
	SC	7.35	8.60	8.45	8.13
Thatch Mean		5.82	6.45	6.79	6.35

- (a) Using  $\alpha=0.05$  for any tests that you may want to conduct, what is the effect of Nitrogen and Thatch on the average chlorophyll content of the grass clippings.
- (b) Are there any additional analyses you would suggest? If so, give a brief outline of these analyses. (You **do not** need to actually perform the analyses.)

### PROBLEM III.

Two treatments, A and B, showed promise for treating a potentially fatal disease. A randomized experiment was conducted to determine whether there is a significant difference in the survival rate between patients who receive treatment A and those who receive treatment B. Of 154 patients who received treatment A, 38 survived for at least 15 years, whereas 16 of the 164 patients who received treatment B survived at least 15 years.

1. A double-blind experiment is one in which neither the patient nor the doctor can identify the treatment given to the patient. Treatment A can be administered only as a pill, and treatment B can be administered only as an injection. Can this randomized experiment be performed as a double-blind experiment? Why or why not?
2. Construct and interpret a 95 percent confidence interval for the difference between the proportion of the population who would survive at least 15 years if given treatment A and the proportion of the population who would survive at least 15 years if given treatment B.

In many of these types of studies, physicians are interested in the ratio of survival probabilities,  $\frac{p_A}{p_B}$ , where  $p_A$  represents the true 15-year survival rate for all patients who receive treatment A and  $p_B$  represents the true 15-year survival rate for all patients who receive treatment B. This ratio is usually referred to as the *relative risk* of the two treatments. For example, a relative risk of 1 indicates the survival rates for patients receiving the two treatments are equal, whereas a relative risk of 1.5 indicates that the survival rate for patients receiving treatment A is 50 percent higher than the survival rate for patients receiving treatment B. An estimator of the relative risk is the ratio of estimated probabilities,  $\frac{\hat{p}_A}{\hat{p}_B}$ .

3. Using the data from the randomized experiment described above, compute the estimate of the relative risk.

The sampling distribution of  $\frac{\hat{p}_A}{\hat{p}_B}$  is skewed. However, when both sample sizes,  $n_A$  and  $n_B$ , are relatively large, the distribution of  $\log_e \left( \frac{\hat{p}_A}{\hat{p}_B} \right)$ , the natural logarithm of relative risk, is approximately normal with a mean of  $\log_e \left( \frac{p_A}{p_B} \right)$  and a standard deviation of  $\sqrt{\frac{1 - p_A}{n_A p_A} + \frac{1 - p_B}{n_B p_B}}$  where  $p_A$  and  $p_B$  can be estimated by using  $\hat{p}_A$  and  $\hat{p}_B$ . When a 95 percent confidence interval for  $\log_e \left( \frac{p_A}{p_B} \right)$  is known, an approximate 95 percent confidence interval for  $\frac{p_A}{p_B}$ , the relative risk of the two treatments, can be constructed by exponentiating the endpoints of the confidence interval for  $\log_e \left( \frac{p_A}{p_B} \right)$ .

4. Construct and interpret a 95 percent confidence interval for the relative risk,  $\frac{p_A}{p_B}$ , of the two treatments.

The *odds ratio*,  $OR = \frac{p_A/(1-p_A)}{p_B/(1-p_B)} = \frac{p_A(1-p_B)}{p_B(1-p_A)}$  is another quantity that can be used to compare proportions. An estimator of the odds ratio is the sample odds ratio,  $\widehat{OR} = \frac{\hat{p}_A(1-\hat{p}_B)}{\hat{p}_B(1-\hat{p}_A)}$ .

5. Using the data from the randomized experiment above, compute the estimate of the odds ratio.

We will now look at a setting where we can further explore the three measures used to compare proportions. In a long-term study of British male physicians, the sample proportion who died from heart disease was 0.00140 for smokers and 0.00669 for nonsmokers.

6. Compute sample estimates of the difference in proportions, relative risk, and odds ratio of having heart disease for smokers versus nonsmokers.
7. Explain why the estimated relative risk and the estimated odds ratio have similar values.
8. Which measure(s) best indicate the relationship between the probabilities for having heart disease for smokers and nonsmokers? Explain why.

## Problem IV:

### Part A.

In a study of the percentage of raw material that responds in a reaction, researchers identified the following five factors:

- the feed rate of the chemicals (*FeedRate*), ranging from 10 to 15 liters per minute
- the percentage of the catalyst (*Catalyst*), ranging from 1% to 2%
- the agitation rate of the reactor (*AgitRate*), ranging from 100 to 120 revolutions per minute
- the temperature (*Temperature*), ranging from 140 to 180 degrees Celsius
- the concentration (*Concentration*), ranging from 3% to 6%

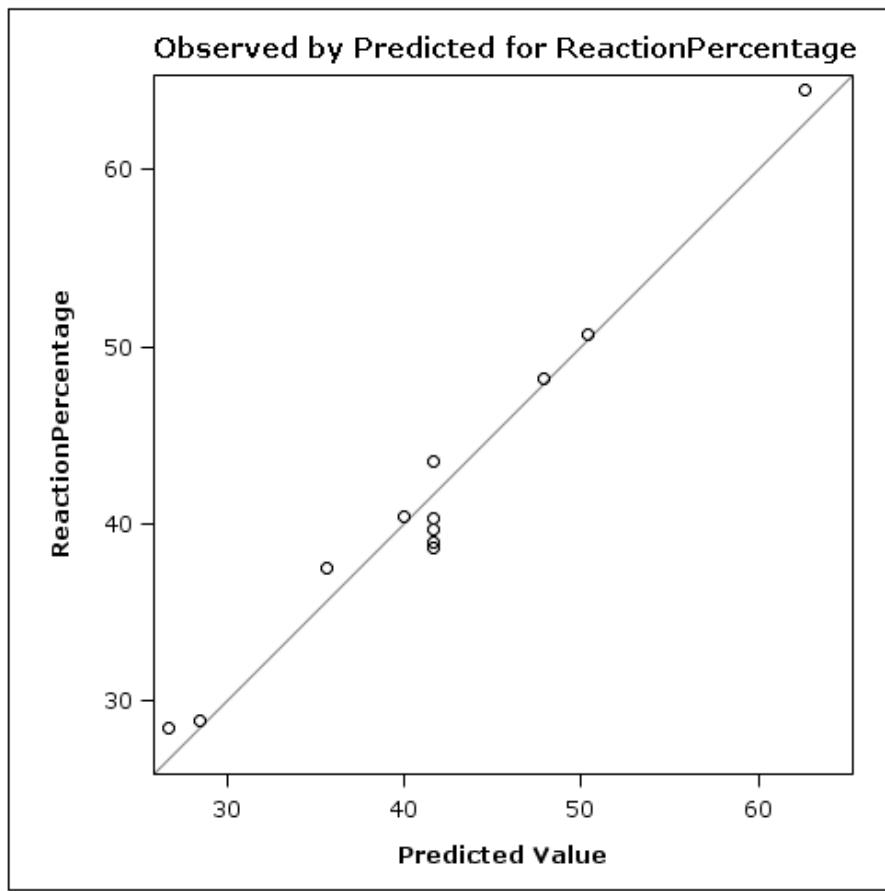
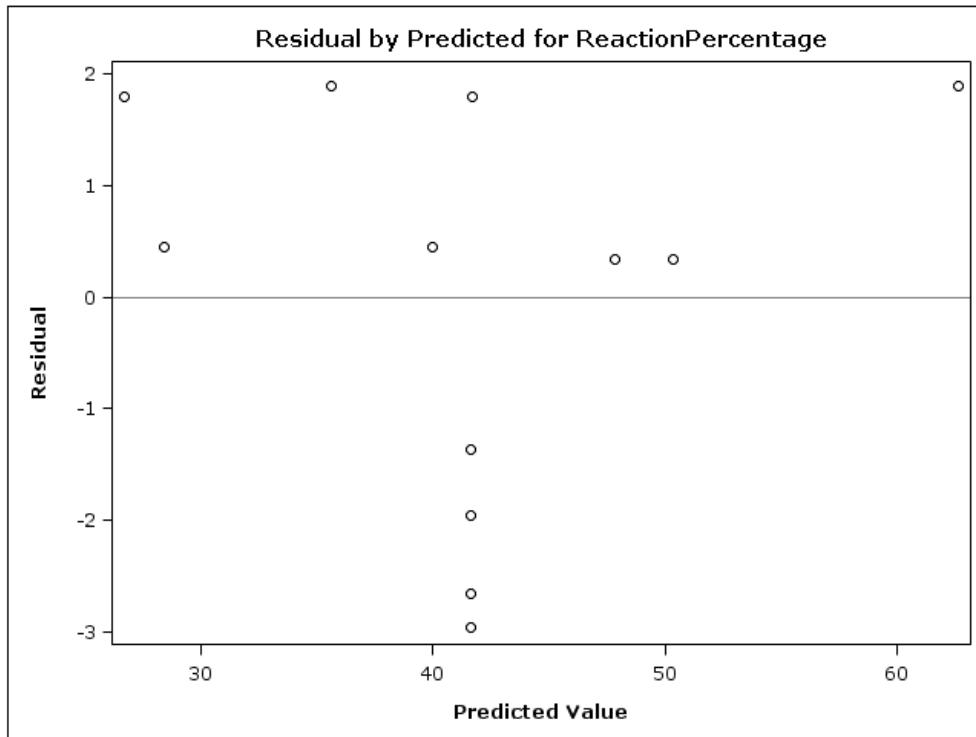
The following data set contains the results of an experiment designed to estimate main effects for all factors:

	FeedRate	Catalyst	AgitRate	Temperature	Concentration	ReactionPercentage
1	10	1	100	140	6	37.5
2	10	1	120	180	3	28.5
3	10	2	100	180	3	40.4
4	10	2	120	140	6	48.2
5	15	1	100	180	6	50.7
6	15	1	120	140	3	28.9
7	15	2	100	140	3	43.5
8	15	2	120	180	6	64.5
9	12.5	1.5	110	160	4.5	39
10	12.5	1.5	110	160	4.5	40.3
11	12.5	1.5	110	160	4.5	38.7
12	12.5	1.5	110	160	4.5	39.7

The model is :

$$\text{ReactionPercentage} = \beta_0 + \beta_1 * \text{FeedRate} + \beta_2 * \text{Catalyst} + \beta_3 * \text{AgitRate} + \beta_4 * \text{Temperature} + \beta_5 * \text{Concentration} + \varepsilon$$

Some of the results of the regression analyses are given below:



Source	Analysis of Variance				
	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	990.27000	198.05400	33.29	0.0003
Error	6	35.69917	5.94986		
Corrected Total	11	1025.96917			

Root MSE	2.43923	R-Square	0.9652
Dependent Mean	41.65833	Adj R-Sq	0.9362
Coeff Var	5.85533		

Variable	Parameter Estimates					
	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	-43.69167	13.04097	-3.35	0.0154	0
FeedRate	1	1.65000	0.34496	4.78	0.0031	1.00000
Catalyst	1	12.75000	1.72480	7.39	0.0003	1.00000
AgitRate	1	-0.02500	0.08624	-0.29	0.7817	1.00000
Temperature	1	0.16250	0.04312	3.77	0.0093	1.00000
Concentration	1	4.96667	0.57493	8.64	0.0001	1.00000

What would you recommend to your client?

Answer the above question in terms of the following questions:

1) Is this a valid model? Why or why not?

2) If you need to add interactions and squared terms, can you just add them as a group to the model and run the analyses? Why or why not?

3) What is being tested by the F value of 33.29 in the ANOVA table? Answer in terms of the  $\beta$ 's ?

## **Part B.**

The usual multiple linear regression model can be written as:

$$Y = X\beta + \varepsilon$$

where  $V(\varepsilon) = \sigma^2 I$  and  $I$  is the  $(nxn)$  identity matrix so that  $V(Y|X) = \sigma^2 I$ .

However, if  $V(\varepsilon) = R_{nn}$ , then (Show work)

a) What is the  $V(Y|X)$  ?

b) If  $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$  then what is  $V(\hat{Y}|X)$ ?

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)

Standard normal density function  
Shaded area =  $\Phi(z)$

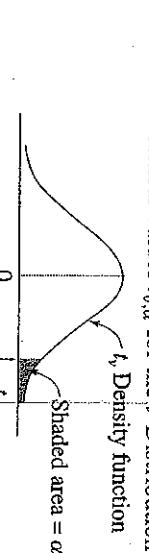
$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0.9982
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)

Standard normal density function  
Shaded area =  $\Phi(z)$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0197	0.0192	0.0188	0.0183	0.0180
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0394	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3839
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

**Table A.4** Critical Values  $t_{v,\alpha}$  for the  $t$ -Distribution



$v$	.10	.05	.025	.01	.005	.001	.0005
	$\alpha$						
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

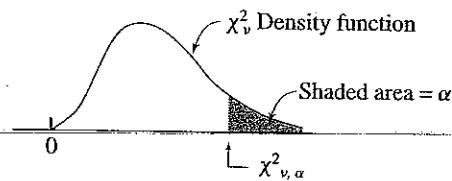
**SOURCE:** This table is produced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.), *The Biometrika Tables for Statisticians*, vol. 1, 3rd ed. (1966), Biometrika.

**Table A.2** Cumulative Poisson Probabilities (cont.)

	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	15.0	20.0
1	.406	.135	.050	.018	.007	.002	.001	.000	.000	.000	.000
2	.677	.423	.238	.125	.062	.030	.014	.006	.003	.000	.000
3	.857	.647	.433	.265	.151	.082	.042	.021	.010	.000	.000
4	.947	.815	.629	.440	.285	.173	.100	.055	.029	.001	.000
5	.983	.916	.785	.616	.446	.301	.191	.116	.067	.003	.000
6	.995	.966	.889	.762	.606	.450	.313	.207	.130	.008	.000
7	.999	.988	.949	.867	.744	.599	.453	.324	.220	.018	.001
8	1.000	.996	.979	.932	.847	.729	.593	.456	.333	.037	.002
9	1.000	.999	.992	.968	.916	.830	.717	.587	.458	.070	.005
10	1.000	.997	.986	.957	.901	.816	.706	.583	.466	.066	.006
11	1.000	.999	.995	.980	.947	.888	.803	.697	.568	.105	.011
12	1.000	.998	.991	.973	.936	.876	.792	.668	.531	.157	.021
13	1.000	.999	.996	.987	.966	.926	.864	.739	.604	.221	.063
14	1.000	.999	.994	.983	.959	.917	.846	.718	.586	.297	.059
15	1.000	.999	.998	.992	.978	.951	.893	.764	.634	.466	.105
16	1.000	.999	.996	.989	.973	.944	.889	.759	.624	.434	.096
17	1.000	.999	.998	.995	.986	.955	.905	.875	.745	.559	.131
18	1.000	.999	.998	.998	.993	.963	.913	.883	.753	.568	.131
19	1.000	.999	.999	.998	.998	.993	.933	.893	.863	.673	.131
20	1.000	.999	.999	.999	.998	.998	.993	.953	.923	.733	.131
21	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
22	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
23	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
24	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
25	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
26	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
27	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
28	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
29	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
30	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131
36	1.000	.999	.999	.999	.999	.999	.999	.999	.999	.999	.131

**SOURCE:** L. L. Chao (1974), *Statistics: Methods and Analysis*, 2nd ed. New York: McGraw-Hill.

**Table A.5 Critical Values  $\chi^2_{v,\alpha}$  for the Chi-square Distribution**

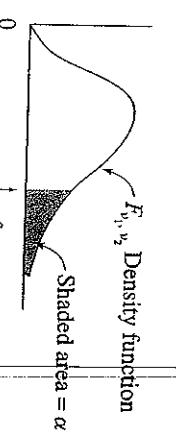


v	$\alpha$									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.420	65.473
40*	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

\* For  $v > 40$ ,  $\chi^2_{v,\alpha} \approx \sqrt{1 - \frac{2}{9v}} + z_\alpha \sqrt{\frac{2}{9v}}$ .

SOURCE: This table is produced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.) *The Biometrika Tables for Statisticians*, vol. 1, 3rd ed. (1966) Biometrika.

Table A.6 Critical Values  $f_{v_1, v_2, \alpha}$  for the  $F$ -Distribution ( $\alpha = .05$ ) (cont.)



	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
	Degrees of freedom for the numerator ( $v_1$ )																		
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.55	8.53	
4	7.71	6.94	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.59	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	
8	5.32	4.46	4.07	3.84	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.01	2.97	2.93		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.75	2.71	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.53	2.49	2.45	2.40	2.36	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.78	2.70	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.26	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.49	2.40	2.33	2.29	2.25	2.20	2.16	2.11	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	
17	4.45	3.59	3.20	2.96	2.81	2.69	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.10	2.06	2.01	1.96	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.06	2.02	1.97	1.92	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.78	
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.81	1.80	1.75	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	
28	4.20	3.34	2.95	2.71	2.56	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.76	1.70	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	
31	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.62	
32	4.00	3.15	2.76	2.53	2.37	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39	
33	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.81	1.75	1.66	1.61	1.55	1.55	1.43	1.35	
34	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	

# **MASTER'S DIAGNOSTIC EXAMINATION**

**January 2012**

Student's Name \_\_\_\_\_

## **INSTRUCTIONS FOR STUDENTS:**

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer all four questions: Questions I, II, III and IV.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature\_\_\_\_\_

## **INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to **longneck@stat.tamu.edu** Do not send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu**.

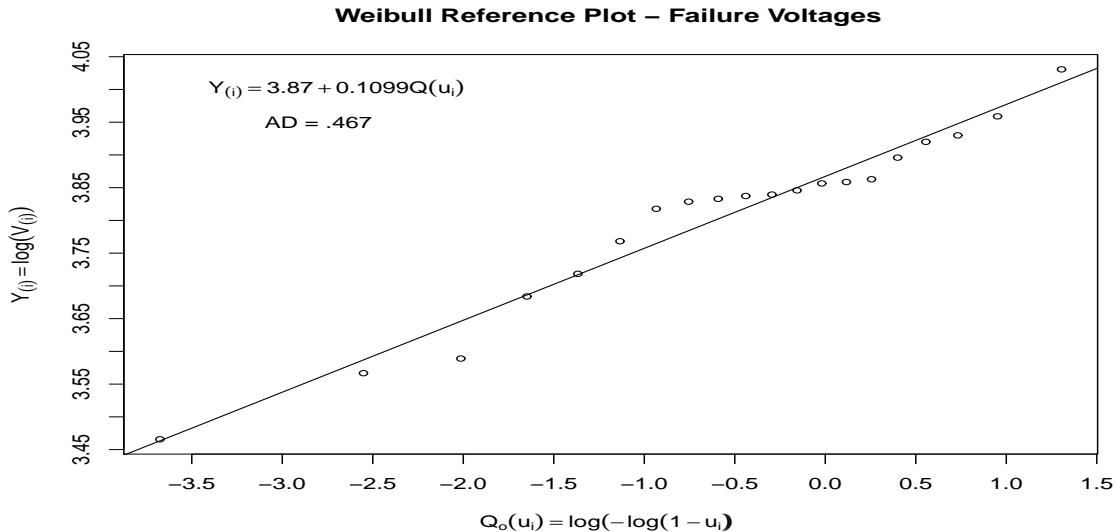
Proctor's Signature\_\_\_\_\_

## QUESTION I.

An electronics firm is testing the insulation used on its electrical cable. A study was conducted to determine the voltage level at which the insulation failed. The study consisted of 20 specimens and the data given in the following table is failure voltages (in kilovolts per millimeter).

32.0	35.4	36.2	39.8	41.2	43.3	45.5	46.0	46.2	46.4
46.5	46.8	47.3	47.4	47.6	49.2	50.4	50.9	52.4	56.3

A plot of  $Y_i = \log(V_i)$  versus  $Q(u_i) = \log(-\log(1 - u_i))$  is given here, where  $V_i$  is the failure voltage of the  $i$ th unit and  $u_i = (i - .5)/20$  for  $i = 1, \dots, 20$ .



- Explain how the Weibull Reference Plot is able to assess the fit of the Weibull model to the data without specifying the parameters,  $\gamma$  and  $\beta$ .
- The Anderson-Darling statistic yielded a value of  $AD = .467$  as a measure of how well a Weibull model would fit the data. Based on AD and the above plot, assess the fit of a Weibull distribution to the data. Make sure to provide a p-value from the AD statistic (tables attached).
- Use the following SAS output to obtain the MLE's for the parameters,  $\gamma$  and  $\beta$  from a Weibull distribution:

$$F(x) = 1 - e^{-x^{\gamma}/\beta}$$

Parameter	DF	Estimate	Standard	95% Confidence		Chi-Square	Pr > ChiSq
			Error	Limits			
Intercept	1	3.8667	0.0251	3.8176	3.9159	23792.0	<.0001
Scale	1	0.1065	0.0184	0.0759	0.1495		
Weibull Scale	1	47.7866	1.1979	45.4954	50.1931		
Weibull Shape	1	9.3863	1.6224	6.6891	13.1712		

Hint: SAS uses the Weibull cdf in the form  $F(x) = 1 - e^{-(x/\alpha)^{\gamma}}$

4. The above table contains 95% Confidence Limits for  $\alpha$  and  $\gamma$ . Using the information in the above table, provide approximate 95% Lower Limits for  $\alpha$  and  $\gamma$ .

Hint: The following upper percentiles are from the standard normal distribution:

$$Z_{.005} = 2.576; \quad Z_{.01} = 2.326; \quad Z_{.025} = 1.96; \quad Z_{.05} = 1.645; \quad Z_{.10} = 1.282$$

5. The researcher was uncertain about using a Weibull distribution in the analysis and decides to use a distribution-free method of estimating the survival function. The Kaplan-Meier Product-Limit estimator of the survival function is displayed in the SAS output given below.

- a. Obtain the estimate of the 78th percentile provided by the Product-Limit estimator.
- b. Obtain the MLE estimate of the 78th percentile based on fitting the Weibull distribution. Hint: Use your results from part 3.
- c. Which of the two estimators would you recommend?

Product-Limit Survival Estimates						
V	Survival	Failure	Survival	Number Failed	Number Left	
			Standard Error			
0.0000	1.0000	0	0	0	20	
32.0000	0.9500	0.0500	0.0487	1	19	
35.4000	0.9000	0.1000	0.0671	2	18	
36.2000	0.8500	0.1500	0.0798	3	17	
39.8000	0.8000	0.2000	0.0894	4	16	
41.2000	0.7500	0.2500	0.0968	5	15	
43.3000	0.7000	0.3000	0.1025	6	14	
45.5000	0.6500	0.3500	0.1067	7	13	
46.0000	0.6000	0.4000	0.1095	8	12	
46.2000	0.5500	0.4500	0.1112	9	11	
46.4000	0.5000	0.5000	0.1118	10	10	
46.5000	0.4500	0.5500	0.1112	11	9	
46.8000	0.4000	0.6000	0.1095	12	8	
47.3000	0.3500	0.6500	0.1067	13	7	
47.4000	0.3000	0.7000	0.1025	14	6	
47.6000	0.2500	0.7500	0.0968	15	5	
49.2000	0.2000	0.8000	0.0894	16	4	
50.4000	0.1500	0.8500	0.0798	17	3	
50.9000	0.1000	0.9000	0.0671	18	2	
52.4000	0.0500	0.9500	0.0487	19	1	
56.3000	0	1.0000	.	20	0	

**Table 1: Percentiles for GOF Measures (Completely Specified Distributions)**

Statistic	Modified Statistic	Upper Percentiles							
		.25	.15	.10	.05	.025	.01	.005	.001
$D_n$	$D_n(\sqrt{n} + .12 + .11/\sqrt{n})$	1.019	1.138	1.224	1.358	1.480	1.628	1.731	1.950
$W_n^2$	$(W_n^2 - \frac{4}{n} + \frac{6}{n^2})(1 + \frac{1}{n})$	0.209	0.284	0.347	0.461	0.581	0.743	0.869	1.167
$A_n^2$	For all $n \geq 5$	1.248	1.610	1.933	2.492	3.070	3.857	4.500	6.000

**Table 2: CDF for Anderson-Darling (Completely Specified Distributions)**

z	G(z)										
0.05	0.0000	0.75	0.4815	1.45	0.8111	2.15	0.9239	2.85	0.9674	3.80	0.9891
0.10	0.0000	0.80	0.5190	1.50	0.8235	2.20	0.9285	2.90	0.9692	3.90	0.9902
0.15	0.0000	0.85	0.5537	1.55	0.8350	2.25	0.9328	2.95	0.9710	4.00	0.9913
0.20	0.0096	0.90	0.5858	1.60	0.8457	2.30	0.9368	3.00	0.9726	4.25	0.9934
0.25	0.0296	0.95	0.6154	1.65	0.8556	2.35	0.9405	3.25	0.9795	4.50	0.9950
0.30	0.0618	1.00	0.6427	1.70	0.8648	2.40	0.9441	3.30	0.9807	4.60	0.9955
0.35	0.1036	1.05	0.6680	1.75	0.8734	2.45	0.9474	3.35	0.9818	4.70	0.9960
0.40	0.1513	1.10	0.6912	1.80	0.8814	2.50	0.9504	3.40	0.9828	4.80	0.9964
0.45	0.2019	1.15	0.7127	1.85	0.8888	2.55	0.9534	3.45	0.9837	4.90	0.9968
0.50	0.2532	1.20	0.7324	1.90	0.8957	2.60	0.9561	3.50	0.9846	5.00	0.9971
0.55	0.3036	1.25	0.7503	1.95	0.9021	2.65	0.9586	3.55	0.9855	5.50	0.9983
0.60	0.3520	1.30	0.7677	2.00	0.9082	2.70	0.9610	3.60	0.9863	6.00	0.9990
0.65	0.3930	1.35	0.7833	2.05	0.9138	2.75	0.9633	3.65	0.9870	7.00	0.9997
0.70	0.4412	1.40	0.7973	2.10	0.9190	2.80	0.9654	3.70	0.9878	8.00	0.9999

**Table 3: Modifications and Percentiles for GOF Measures for Normal Distributions with  $\mu$  and  $\sigma$  Unknown**

Statistic	Modified Statistic	Upper Percentiles							
		.50	.25	.15	.10	.05	.025	.01	.005
$D_n$	$D_n(\sqrt{n} - .01 + .85/\sqrt{n})$	-	-	0.775	0.819	0.895	0.995	1.035	-
$W_n^2$	$W_n^2(1 + \frac{5}{n})$	0.051	0.074	0.091	0.104	0.126	0.148	0.179	0.201
$A_n^2$	$A_n^2(1 + \frac{.75}{n} + \frac{2.25}{n^2})$	0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

**Table 4: Modifications and Percentiles for GOF Measures for Exponential Distribution with  $\beta$  Unknown**

Statistic	Modified Statistic	Upper Percentiles							
		.25	.20	.15	.10	.05	.025	.01	.005
$D_n$	$(D_n - \frac{0.2}{n})(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}})$	-	-	0.926	0.995	1.094	1.184	-	-
$W_n^2$	$W_n^2(1.0 + \frac{0.16}{n})$	0.116	0.130	0.148	0.175	0.222	0.271	0.338	0.390
$A_n^2$	$A_n^2(1.0 + \frac{0.6}{n})$	0.736	0.816	0.916	1.062	1.321	1.591	1.959	2.244

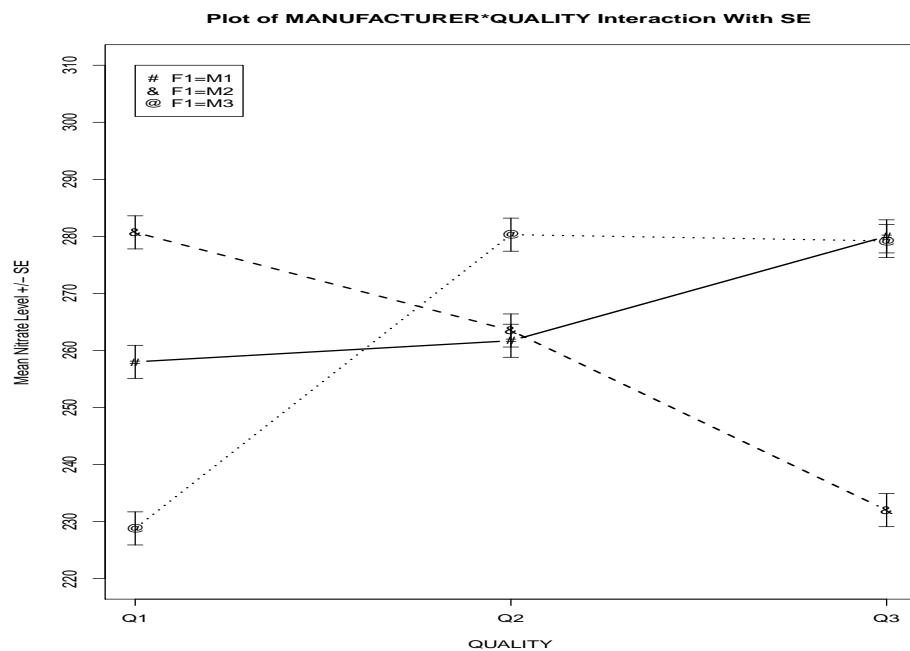
**Table 5: Modifications and Percentiles for A-D Measure for Extreme Value Distribution with Unspecified Parameters**

Statistic	Modified Statistic	Upper Percentiles				
		.25	.10	.05	.025	.01
$A_n^2$	$A_n^2(1.0 + \frac{0.2}{\sqrt{n}})$	0.474	0.637	0.757	0.877	1.038

## QUESTION II.

The USDA is evaluating the level of nitrate ( $NO_3$ ) from sausages obtained from three largest manufacturers, M1, M2, M3 in the US. Each manufacturer produces three grades of quality, either Q1, Q2, or Q3 of their sausage. The processing of different grades of sausage from a common production run may involve different sources of raw materials and processing environments, and these factors sometimes are problematic. Each manufacturer submits two sausages of each grade from each of three production runs. The amount of  $NO_3$  is determined by an USDA lab and is reported in the following table. The three manufacturers are the only manufacturers under evaluation, the production runs were randomly selected, and are representative of general production runs of each manufacturer.

Grade	Manufacture								
	M1			M2			M3		
	Run	Run	Run	Run	Run	Run	Run	Run	Run
Grade	R1	R2	R3	R4	R5	R6	R7	R8	R9
Q1	253	265	253	230	234	231	225	228	232
	256	270	251	226	239	232	229	227	232
Q2	262	263	255	257	268	265	277	276	289
	260	266	264	267	258	266	276	277	287
Q3	279	285	277	275	286	284	280	278	282
	279	288	272	272	283	284	276	277	282



Use the above plots, data, and the attached SAS output to answer the following questions.

- Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.

$C_1$  Normality:

$C_2$  Equal Variance:

$C_3$  Independence:

2. At the  $\alpha = .05$  level, which main effects and interactions are significant? Justify your answer by including the relevant p-values along with their pair of degrees of freedom ( $df_{NUM.}, df_{DEN.}$ ).
3. What is the expected value of  $MS_{Manufacturer}$ ?
4. Separate the three Grade Levels of Quality into groups of levels such that all levels in a group are not significantly different from any other member of the group with respect to their mean  $NO_3$  level. Use an experimentwise error rate of  $\alpha = .05$ .
5. Provide a 95% confidence on the mean  $NO_3$  level of a sausage having Quality Grade Q1 produced by Manufacturer M1.
6. Identify each of the following sums of squares formulas with its correct source of variation in the AOV table for the experiment. For example, if  $y_{ijkl}$  is the  $NO_3$  level of sausage  $l$  of Quality Grade  $i$  from Manufacturer  $j$  on production run  $k$ , then

$$\sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \sum_{l=1}^2 (y_{ijkl} - \bar{y}_{....})^2 \text{ is } SS_{TOTAL}$$

a.  $18 \sum_{i=1}^3 (\bar{y}_{i...} - \bar{y}_{....})^2$  is  $SS_{\dots}$

b.  $6 \sum_{i=1}^3 \sum_{j=1}^3 (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{j..} + \bar{y}_{....})^2$  is  $SS_{ij..}$

c.  $6 \sum_{j=1}^3 \sum_{k=1}^3 (\bar{y}_{.jk.} - \bar{y}_{.j..})^2$  is  $SS_{.jk.}$

```

OPTIONS LS=90 PS=55 nocenter nodate;
TITLE 'SAS OUTPUT FOR QUESTION II';
DATA MANU;
INPUT M $ Q $ R $ Y @@;
TRT=COMPRESS (M) || COMPRESS (Q);
LABEL M="MANUFACTURER" Q="QUALITY";
CARDS;
M1 Q1 R1 253 M1 Q1 R2 265 M1 Q1 R3 253 M1 Q1 R1 256 M1 Q1 R2 270 M1 Q1 R3 251
M1 Q2 R1 262 M1 Q2 R2 263 M1 Q2 R3 255 M1 Q2 R1 260 M1 Q2 R2 266 M1 Q2 R3 264
M1 Q3 R1 279 M1 Q3 R2 285 M1 Q3 R3 277 M1 Q3 R1 279 M1 Q3 R2 288 M1 Q3 R3 272
M2 Q3 R4 230 M2 Q3 R5 234 M2 Q3 R6 231 M2 Q3 R4 226 M2 Q3 R5 239 M2 Q3 R6 232
M2 Q2 R4 257 M2 Q2 R5 268 M2 Q2 R6 265 M2 Q2 R4 267 M2 Q2 R5 258 M2 Q2 R6 266
M2 Q1 R4 275 M2 Q1 R5 286 M2 Q1 R6 284 M2 Q1 R4 272 M2 Q1 R5 283 M2 Q1 R6 284
M3 Q1 R7 225 M3 Q1 R8 228 M3 Q1 R9 232 M3 Q1 R7 229 M3 Q1 R8 227 M3 Q1 R9 232
M3 Q2 R7 277 M3 Q2 R8 276 M3 Q2 R9 289 M3 Q2 R7 276 M3 Q2 R8 277 M3 Q2 R9 287
M3 Q3 R7 280 M3 Q3 R8 278 M3 Q3 R9 282 M3 Q3 R7 276 M3 Q3 R8 277 M3 Q3 R9 282
PROC GLM;
CLASS M Q R;
MODEL Y = M Q M*Q R(M) Q*R(M);
RANDOM R(M) Q*R(M)/TEST;
LSMEANS M Q M*Q/STDEERR PDIFF ADJUST=TUKEY;
RUN;
PROC MIXED CL ALPHA=.05 COVTEST;
CLASS M Q R;
MODEL Y = M Q M*Q ;
RANDOM R(M) Q*R(M);
LSMEANS M Q M*Q/ ADJUST=TUKEY;
RUN;
PROC GLM;
CLASS TRT;
MODEL Y = TRT;
MEANS TRT/HOVTEST=BF;
OUTPUT OUT=ASSUMP R=RESID P=MEANS;
PROC GPLOT; PLOT MEANS*TRT ;PLOT RESID*TRT/VREF=0;
PROC UNIVARIATE DEF=5 PLOT NORMAL;
VAR RESID;
RUN;

```

SAS OUTPUT FOR QUESTION II

Class	Levels	Values
M	3	M1 M2 M3
Q	3	Q1 Q2 Q3
R	9	R1 R2 R3 R4 R5 R6 R7 R8 R9
Number of Observations Read		54

Dependent Variable: Y

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	26	20826.14815	801.00570	92.62	<.0001
Error	27	233.50000	8.64815		
Corrected Total	53	21059.64815			

Source	DF	Sum of		F Value	Pr > F
		Type III SS	Mean Square		
M	2	552.48148	276.24074	31.94	<.0001
Q	2	1473.03704	736.51852	85.16	<.0001
M*Q	4	17878.96296	4469.74074	516.84	<.0001
R(M)	6	729.00000	121.50000	14.05	<.0001
Q*R(M)	12	192.66667	16.05556	1.86	0.0887

Source	Type III Expected Mean Square				
M	Var(Error) + 2 Var(Q*R(M)) + 6 Var(R(M)) + Q(M,M*Q)				
Q	Var(Error) + 2 Var(Q*R(M)) + Q(Q,M*Q)				
M*Q	Var(Error) + 2 Var(Q*R(M)) + Q(M*Q)				
R(M)	Var(Error) + 2 Var(Q*R(M)) + 6 Var(R(M))				
Q*R(M)	Var(Error) + 2 Var(Q*R(M))				

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

*      Source	DF	Sum of		F Value	Pr > F
		Type III SS	Mean Square		
*      M	2	552.481481	276.240741	2.27	0.1841
Error: MS(R(M))	6	729.000000	121.500000		
*      Source	DF	Type III SS	Mean Square	F Value	Pr > F
*      Q	2	1473.037037	736.518519	45.87	<.0001
M*Q	4	17879	4469.740741	278.39	<.0001
R(M)	6	729.000000	121.500000	7.57	0.0016
Error: MS(Q*R(M))	12	192.666667	16.055556		
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Q*R(M)	12	192.666667	16.055556	1.86	0.0887
Error: MS(Error)	27	233.500000	8.648148		

SAS OUTPUT FOR QUESTION II

The GLM Procedure

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

M	Y LSMEAN	Standard		LSMEAN Number
		Error	Pr >  t	
M1	266.555556	0.693147	<.0001	1
M2	258.722222	0.693147	<.0001	2
M3	262.777778	0.693147	<.0001	3

Least Squares Means for effect M  
 $\text{Pr} > |t| \text{ for } H_0: \text{LSMean}(i) = \text{LSMean}(j)$

i/j	Dependent Variable: Y		
	1	2	3
1		<.0001	0.0018
2	<.0001		0.0009
3	0.0018	0.0009	

Least Squares Means  
 Adjustment for Multiple Comparisons: Tukey

Q	Y LSMEAN	Standard		LSMEAN Number
		Error	Pr >  t	
Q1	255.833333	0.693147	<.0001	1
Q2	268.500000	0.693147	<.0001	2
Q3	263.722222	0.693147	<.0001	3

Least Squares Means for effect Q  
 $\text{Pr} > |t| \text{ for } H_0: \text{LSMean}(i) = \text{LSMean}(j)$

i/j	Dependent Variable: Y		
	1	2	3
1		<.0001	<.0001
2	<.0001		0.0001
3	<.0001	0.0001	

Least Squares Means  
 Adjustment for Multiple Comparisons: Tukey

M	Q	Y LSMEAN	Standard		LSMEAN	
			Error	Pr >  t	Number	
M1	Q1	258.000000	1.200566	<.0001	1	
M1	Q2	261.666667	1.200566	<.0001	2	
M1	Q3	280.000000	1.200566	<.0001	3	
M2	Q1	280.666667	1.200566	<.0001	4	
M2	Q2	263.500000	1.200566	<.0001	5	
M2	Q3	232.000000	1.200566	<.0001	6	
M3	Q1	228.833333	1.200566	<.0001	7	
M3	Q2	280.333333	1.200566	<.0001	8	
M3	Q3	279.166667	1.200566	<.0001	9	

Least Squares Means for effect M\*Q  
 $\text{Pr} > |t| \text{ for } H_0: \text{LSMean}(i) = \text{LSMean}(j)$

i/j	Dependent Variable: Y								
	1	2	3	4	5	6	7	8	9
1	0.4574	<.0001	<.0001	0.0659	<.0001	<.0001	<.0001	<.0001	<.0001
2	0.4574		<.0001	<.0001	0.9724	<.0001	<.0001	<.0001	<.0001
3	<.0001	<.0001		1.0000	<.0001	<.0001	<.0001	1.0000	0.9999
4	<.0001	<.0001	1.0000		<.0001	<.0001	<.0001	1.0000	0.9921
5	0.0659	0.9724	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
6	<.0001	<.0001	<.0001	<.0001	<.0001		0.6414	<.0001	<.0001
7	<.0001	<.0001	<.0001	<.0001	<.0001	0.6414		<.0001	<.0001
8	<.0001	<.0001	1.0000	1.0000	<.0001	<.0001	<.0001		0.9986
9	<.0001	<.0001	0.9999	0.9921	<.0001	<.0001	<.0001	0.9986	

SAS OUTPUT FOR QUESTION II

The Mixed Procedure

Model Information

Data Set	WORK.MANU
Dependent Variable	Y
Covariance Structure	Variance Components
Estimation Method	REML

Class Level Information

Class	Levels	Values
M	3	M1 M2 M3
Q	3	Q1 Q2 Q3
R	9	R1 R2 R3 R4 R5 R6 R7 R8 R9

Number of Observations  
Number of Observations Read 54  
Number of Observations Used 54

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
R(M)	17.5741	11.7423	1.50	0.0672	0.05	6.5803	122.55
Q*R(M)	3.7037	3.4822	1.06	0.1438	0.05	1.0582	101.54
Residual	8.6481	2.3537	3.67	0.0001	0.05	5.4058	16.0224

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
M	2	6	2.27	0.1841
Q	2	12	45.87	<.0001
M*Q	4	12	278.39	<.0001

SAS OUTPUT FOR QUESTION II

The Mixed Procedure

Least Squares Means

Effect	MANUFACTURER	QUALITY	Standard		DF	t Value	Pr >  t
			Estimate	Error			
M	M1		266.56	2.5981	6	102.60	<.0001
M	M2		258.72	2.5981	6	99.58	<.0001
M	M3		262.78	2.5981	6	101.14	<.0001
Q		Q1	255.83	1.6866	12	151.69	<.0001
Q		Q2	268.50	1.6866	12	159.20	<.0001
Q		Q3	263.72	1.6866	12	156.36	<.0001
M*Q	M1	Q1	258.00	2.9213	12	88.32	<.0001
M*Q	M1	Q2	261.67	2.9213	12	89.57	<.0001
M*Q	M1	Q3	280.00	2.9213	12	95.85	<.0001
M*Q	M2	Q1	280.67	2.9213	12	96.08	<.0001
M*Q	M2	Q2	263.50	2.9213	12	90.20	<.0001
M*Q	M2	Q3	232.00	2.9213	12	79.42	<.0001
M*Q	M3	Q1	228.83	2.9213	12	78.33	<.0001
M*Q	M3	Q2	280.33	2.9213	12	95.96	<.0001
M*Q	M3	Q3	279.17	2.9213	12	95.56	<.0001

SAS OUTPUT FOR QUESTION II

The Mixed Procedure

Differences of Least Squares Means

Effect	MANUFACTURER	QUALITY	MANUFACTURER	QUALITY	Pr >  t	Adjustment	Adj P
M	M1		M2		0.0770	Tukey	0.1632
M	M1		M3		0.3435	Tukey	0.5878
M	M2		M3		0.3120	Tukey	0.5463
Q		Q1		Q2	<.0001	Tukey-Kramer	<.0001
Q		Q1		Q3	<.0001	Tukey-Kramer	0.0002
Q		Q2		Q3	0.0038	Tukey-Kramer	0.0098
M*Q	M1	Q1	M1	Q2	0.1390	Tukey-Kramer	0.7963
M*Q	M1	Q1	M1	Q3	<.0001	Tukey-Kramer	<.0001
M*Q	M1	Q2	M1	Q3	<.0001	Tukey-Kramer	<.0001
M*Q	M2	Q1	M2	Q2	<.0001	Tukey-Kramer	0.0002
M*Q	M2	Q1	M2	Q3	<.0001	Tukey-Kramer	<.0001
M*Q	M2	Q2	M2	Q3	<.0001	Tukey-Kramer	<.0001
M*Q	M3	Q1	M3	Q2	<.0001	Tukey-Kramer	<.0001
M*Q	M3	Q1	M3	Q3	<.0001	Tukey-Kramer	<.0001
M*Q	M3	Q2	M3	Q3	0.6232	Tukey-Kramer	0.9998

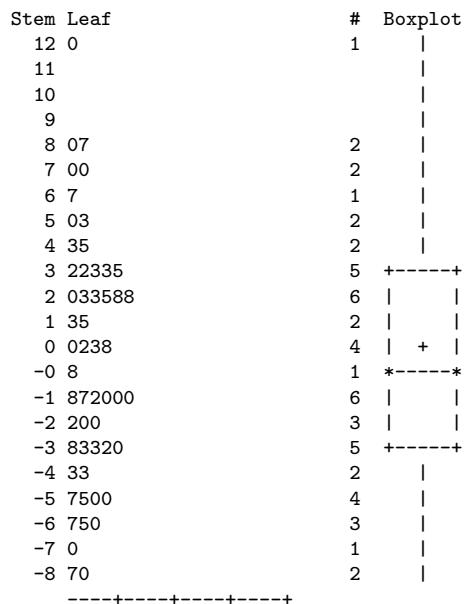
SAS OUTPUT FOR QUESTION II

Brown and Forsythe's Test for Homogeneity of Y Variance  
 ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
TRT	8	59.5926	7.4491	0.50	0.8510
Error	45	672.8	14.9519		

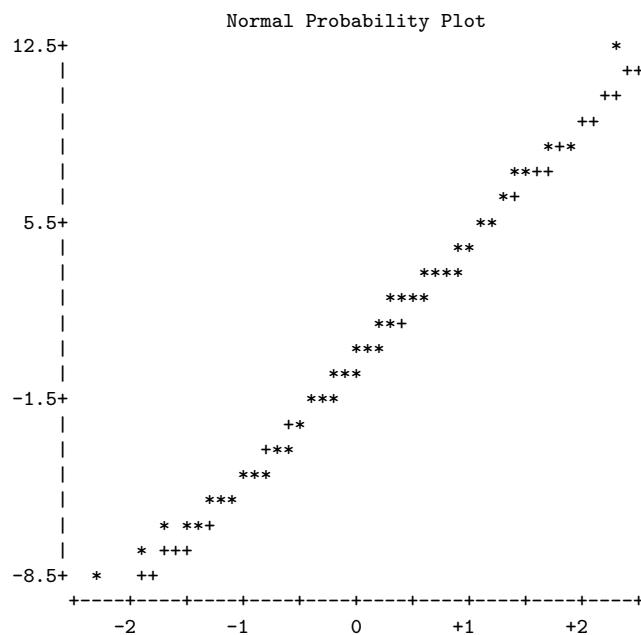
Tests for Normality

Test	--Statistic--	-----p Value-----
Shapiro-Wilk	W	0.984389 Pr < W 0.7025
Kolmogorov-Smirnov	D	0.070834 Pr > D >0.1500
Cramer-von Mises	W-Sq	0.037375 Pr > W-Sq >0.2500
Anderson-Darling	A-Sq	0.237817 Pr > A-Sq >0.2500

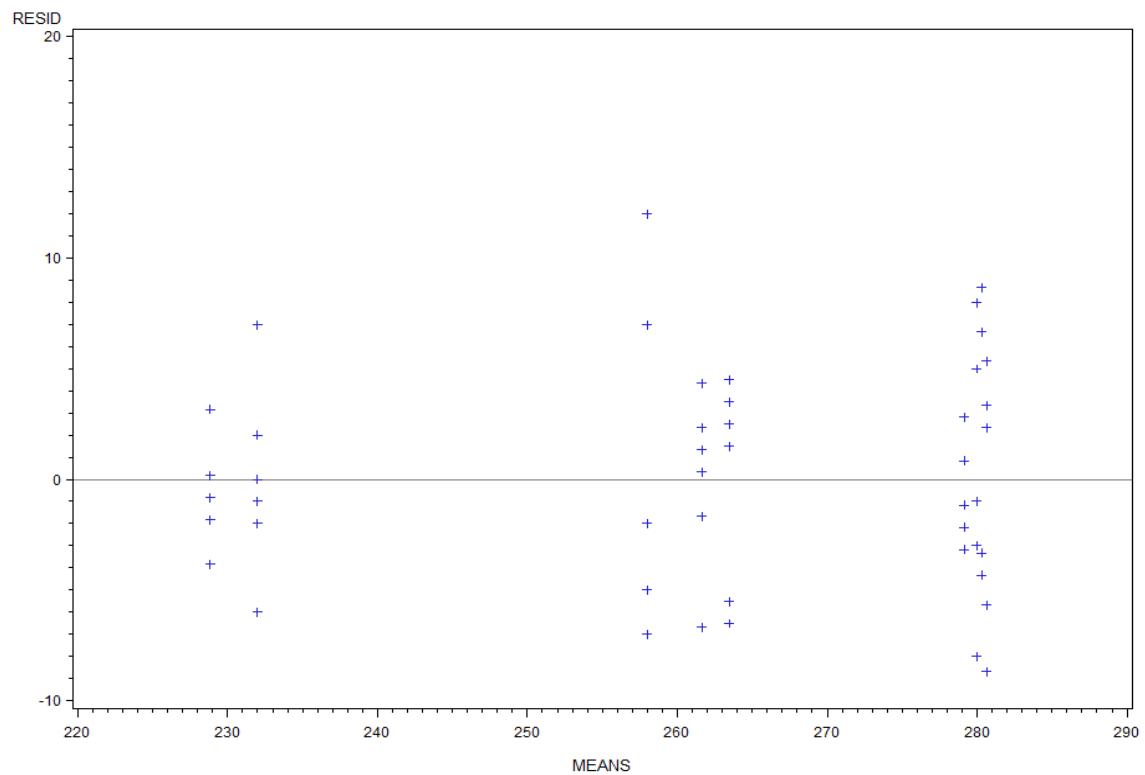


SAS OUTPUT FOR PROBLEM II

Variable: RESID



**SAS OUTPUT FOR PROBLEM II**



### QUESTION III.

1. Let  $X_1, \dots, X_7$  be independent standard normal random variables. Identify the distribution of each of the following random variables. Be sure to justify your answers.

(a)  $U = X_1^2 + X_3^2 + X_5^2 + X_6^2 + X_7^2$ .

(b)  $W = X_7 / \sqrt{[X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 + X_6^2]/6}$ .

(c)  $Y = W^2 = 6X_2^2/[X_1^2 + X_3^2 + X_4^2 + X_5^2 + X_6^2 + X_7^2 + X_8^2]$ .

(d)  $T = X_1/X_4$ .

(e)  $S = 3(X_2^2 + X_4^2)/[2(X_1^2 + X_3^2 + X_5^2)]$ .

2. Let  $X \sim N(2, 8)$  and  $Y \sim N(-3, 5)$  be independent normal random variables. (Note: The notation  $N(a, b)$  indicates a normal distribution with mean  $a$  and variance  $b$ .)

(a) Let  $U = 2X + 3Y - 5$  and  $V = X - CY$ , where  $C$  is a constant. Identify the distributions of  $U$  and  $V$ .

(b) For  $U$  and  $V$  defined in part (a), what is the value of  $C$  that makes  $U$  and  $V$  independent?

(c) Let

$$W = C_1(X + C_2)^2 + C_3(Y + C_4)^2$$

Find values of  $C_1, C_2, C_3, C_4$ , and  $C_5$  (with  $C_1 \neq 0$  and  $C_3 \neq 0$ ) so that  $W$  has a  $\chi^2$  distribution with  $C_5$  degrees of freedom.

(d) Let

$$T = \frac{C_1(X + C_2)^{C_3}}{(Y + C_4)^{C_5}}.$$

Find values of  $C_1, C_2, C_3, C_4, C_5$  and  $C_6$  so that  $T$  has a  $t$  distribution with  $C_6$  degrees of freedom.

(e) Let

$$T = \frac{C_1(X + C_2)^{C_3}}{(Y + C_4)^{C_5}}.$$

Find values of  $C_1, C_2, C_3, C_4, C_5, C_6$  and  $C_7$  so that  $T$  has a  $F$  distribution with  $(C_6, C_7)$  degrees of freedom.

**QUESTION IV:**

1. The multiple regression matrix formulation is given by:

$$Y_{nx1} = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$$

Where:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$E(\varepsilon) = 0_{nx1} \quad V(\varepsilon) = \sigma^2 I_{nxn}$$

- a. What distribution is usually assumed for  $\varepsilon$ ?
- b. Is it correct to test that Y has a normal distribution using a univariate test such as Sharipo-Wilks? Explain your answer.
- c. What is  $I_{nxn}$ ?
- d. Assuming the conditions for MLR are met, how many unknown parameters are there? Display the unknown parameters.
- e. If  $\hat{\beta} = (X^t X)^{(-1)} X^t Y$  derive the expectation and variance of  $\hat{\beta}$ .
- f. Is  $\hat{\beta}$  an unbiased estimator of  $\beta$ ? Explain your answer.

2. A researcher states that he has two models:

Model 1:  $Y_{nx1} = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$  with R-squared = .7

Model 2:  $\log(Y_{nx1}) = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$  with R-squared = .8

Note: n/p = 50.

He also states that since the R-squared for Model 2 is greater than the R-squared for Model 1, then Model 2 is better than Model 1.

Do you agree? Explain your answer.

# MASTER'S DIAGNOSTIC EXAMINATION

August 2012

Student's Name \_\_\_\_\_

## INSTRUCTIONS FOR STUDENTS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer all four questions: Questions I, II, III and IV.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature\_\_\_\_\_

## INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax the student's solutions to 979-845-6060 or email to longneck@stat.tamu.edu.**

Do not send the exam booklet or SAS output, just send the student's solutions.

1. I certify that the time at which the student started the exam was \_\_\_\_\_ and the time at which the student completed the exam was \_\_\_\_\_
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu.**

Proctor's Signature\_\_\_\_\_

## QUESTION I.

A randomized trial was conducted to investigate the relationship between a continuous response  $y$  and four treatments A, B, C, and D. The sample size was  $n = 200$ , with 50 observations in each of the four treatment groups. Let  $\mathbf{y}$  be the  $200 \times 1$  vector of response values, ordered so that the first 50 entries are for treatment group A, the next 50 for B, then C, and finally D. The regression model  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  was fit, where  $\mathbf{X}$  is the  $200 \times 4$  design matrix given by

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

and where each entry is a column vector of length 50. The estimated regression coefficients were

$$\hat{\beta}' = (37.5, -11.5, 1.0, -27.7), \quad \text{with standard errors } 2.75, 3.89, 3.89, 3.89$$

and residual standard deviation  $\hat{\sigma} = 19.45$ .

- (1.) Interpret each of the **four** regression parameters. As in, “the intercept,  $\beta_0$ , is the mean response when ...”.
- (2.) What assumptions are required for the regression model?
- (3.) What is an approximate 95% confidence interval for the mean response in treatment group B?

**Hint:** If  $\mathbf{v}$  is a  $4 \times 1$  column vector, then the variance of

$$\mathbf{v}'\hat{\beta}$$
 is equal to  $\hat{\sigma}^2\mathbf{v}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}$ .

In our example,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.02 & -0.02 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0.02 & 0.02 \\ -0.02 & 0.02 & 0.04 & 0.02 \\ -0.02 & 0.02 & 0.02 & 0.04 \end{pmatrix}$$

- (4.) What is an approximate 95% confidence interval for the mean difference in response between treatment groups B and A (so, the difference  $\mu_B - \mu_A$ )?
- (5.) Suppose the observations in treatment group A were positively correlated with those in treatment group B, and you correctly fit a correlated-data regression model. How would a 95% confidence interval for the mean difference in response between treatment groups B and A compare to the one reported in (4.) above, and why?

## QUESTION II.

(1.) Define, using formulas in addition to words, the least squares criterion.

(2.) Show that the least squares criterion applied to the “intercept-only” model, i.e.,

$$y_i = \beta_0 + \epsilon_i, \quad i = 1, 2, \dots, n$$

results in the least squares estimator of  $\beta_0$ :  $\hat{\beta}_0 = \bar{y}$ .

(3.) Consider the analysis of variance table for the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

What sum of squares in the analysis of variance table for the simple linear regression model would correspond to the sum of squares error,  $SSE$ , from the “intercept-only” model in part (2.)?

(4.) With reference to your answers in the above questions, briefly explain in words why statisticians almost always perform a “corrected” analysis of variance, a partition of  $\sum_{i=1}^n (y_i - \bar{y})^2$ , rather than an “uncorrected” analysis of variance, a partition of  $\sum_{i=1}^n y_i^2$

### QUESTION III

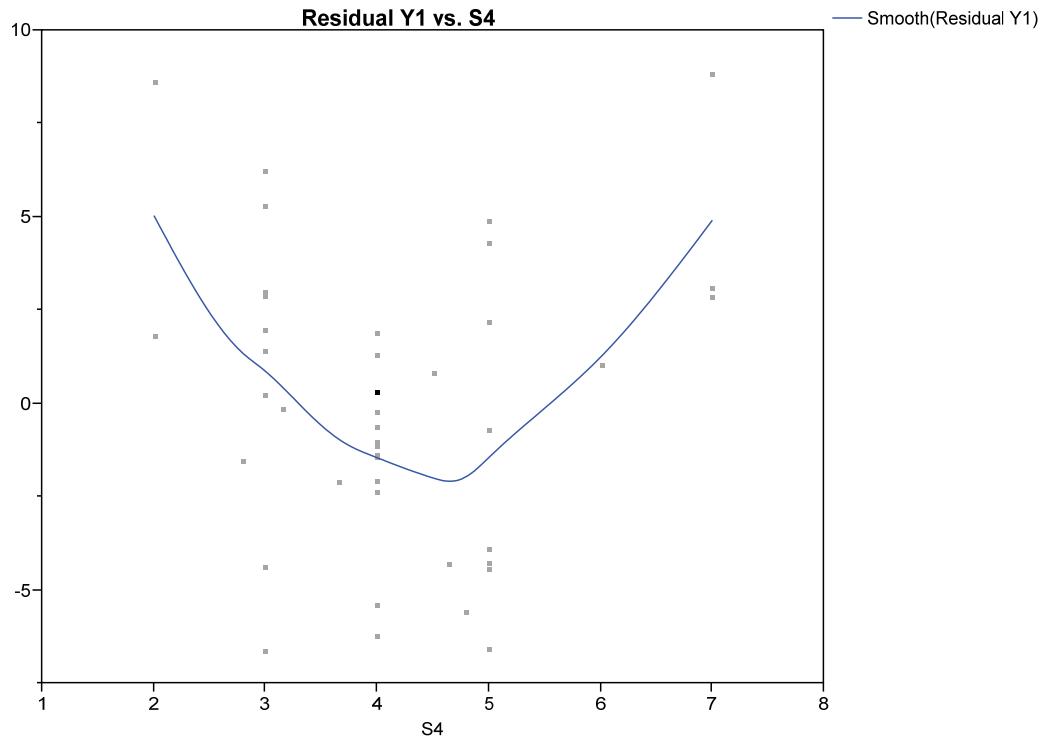
#### Part 1:

We have a dataset with  $y_1$  (a measure of diabetes & dependent variable) and 3 predictors BMI (Body mass index); BP (Blood pressure) and  $s_4$  a measure of glucose. We begin by considering the classical linear model relating  $y_1$  to the three explanatory variables.

Given the graph below (Residual is the “raw” residual), answer the following questions:

1. Based on the graph can you determine if the residuals are normally distributed? Explain.
2. Based on the graph below can you conclude that this is a valid model? Explain.
3. Does the graph suggest a change in the model? (That is: should you add a term; transform  $y_1$  etc.). Explain.

#### Graph Builder



## **Part 2:**

Suppose that we want to predict whether a person will get the Tourette Syndrome based on Gender, HDL, Glucose & Chol. Note: Let Y = 1 if the Syndrome is detected and Y = 0 if not.

A logistic regression was run and the following output was obtained:

Gender	M	F
HDL	40	40
Glucose	125	125
Chol	190	190
Prob Y = 1	0.001	0.0001

1. What are the odds that a male with those characteristics will get the Syndrome? **Justify your answer.**
  2. What are the odds that a female with those characteristics will get the Syndrome? **Justify your Answer.**
  3. What is the odds ratio of a male getting the Syndrome as opposed to a female? **Justify your answer.**
  4. The local newspaper front page story is: "Males are 10.01 times more likely to get Tourette Syndrome than Females. More research money is needed to help the males..."
- Is there something other than the odds ratio that needs to be considered here? Please explain your answer.

## **Part 3.**

This part is similar to Part 1. Here we have y (a measure of diabetes & dependent variable) and 10 predictors: Age, Sex, BMI,BP, s1,s2,s3,s4,s5,s6.

A stepwise regression was run with the following results:

### **Parameter Estimates**

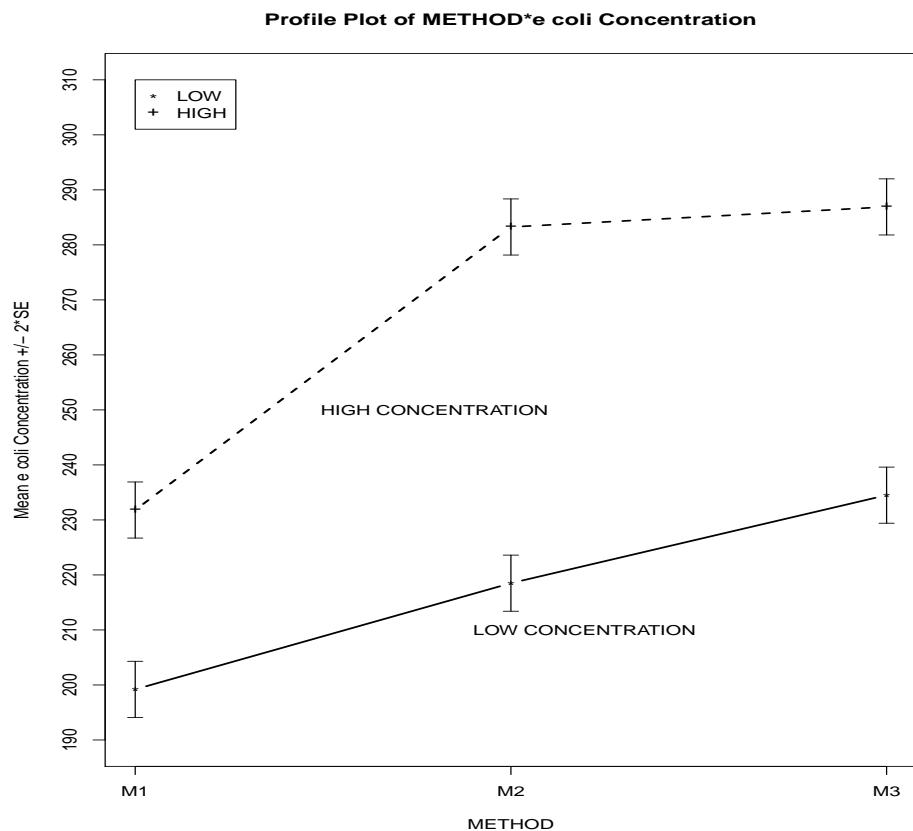
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-483.9509	71.29641	-6.79	<.0001*
SEX	-31.32679	16.06386	-1.95	0.0586
BP	1.7477267	0.568201	3.08	0.0039*
S2	0.6081179	0.320221	1.90	0.0652
S5	52.425839	16.93224	3.10	0.0037*
S6	2.11306	0.72281	2.92	0.0058*

Since SEX and S2 are not significant at alpha= .05, can we remove them both from the model at the same time? Explain.

#### QUESTION IV.

The FDA is investigating methods to control *e coli* in beef products. There are three methods under consideration: M1, M2, and M3. The scientist at FDA decide to evaluate the three methods under two levels of *e coli* contamination, Low ( $< 300$  cfu/g) and High ( $\geq 300$  cfu/g). The researchers randomly selected six herds of cattle having a Low level of *e coli* contamination and six herds of cattle having a High level of *e coli* contamination. From each of the herds, six cattle were randomly selected for slaughter and then a meat sample from two cattle were randomly assigned to each of the three methods for *e coli* treatment. The amount of *e coli* was determined in a FDA lab and is reported in the following table. The three methods are the only methods under evaluation. The FDA is interested in determining which method provides the lowest *e coli* concentration after treatment.

METHOD	LEVEL OF CONTAMINATION											
	LOW						HIGH					
	HERD						HERD					
H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	
M1	203	215	203	190	194	191	225	228	232	235	233	237
	206	210	201	186	199	192	229	227	232	231	238	235
M2	222	223	215	217	218	215	277	276	289	284	285	286
	230	216	224	217	209	216	276	277	287	286	287	289
M3	239	245	237	235	236	234	280	278	282	292	296	295
	229	238	222	232	233	234	276	277	282	294	294	297



Use the above plot, data, and the attached SAS output to answer the following questions.

- (1.) Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.

$C_1$  Normality:

$C_2$  Equal Variance:

$C_3$  Independence:

- (2.) Write a model for  $y_{ijkl}$ , the level of *e coli* in the  $\ell$ th meat sample from Herd  $k$  of Concentration  $i$  treated by Method  $j$ .

- (3.) At the  $\alpha = .05$  level, which main effects and interactions are significant? Justify your answer by including the relevant p-values along with their pair of degrees of freedom ( $df_{NUM.}, df_{DEN.}$ ).

- (4.) Compute that the estimated standard error in the estimated mean difference of *e coli* concentrations between Methods M1 and M2 from a Low concentration.

- (5.) Separate the three Methods of treating *e coli* into groups of levels such that all levels in a group are not significantly different from any other member of the group with respect to their mean concentration of *e coli*. Use an experimentwise error rate of  $\alpha = .05$ .

- (6.) Provide a 95% confidence on the mean *e coli* level of a High concentration meat sample treated by Method M3.

- (7.) The researchers were criticized for using as their response variable the level of *e coli* concentration in the meat after the meat was treated by the three methods. Suggest a more appropriate modeling of the data still using the level of *e coli* concentration in the meat after the meat was treated as the response.

## Methods Exam August 2012 SAS PROGRAM FOR PROBLEM II

```
ods html;ods graphics on;
OPTIONS LS=90 PS=55 nocenter nodate;
TITLE 'SAS OUTPUT FOR PROBLEM II';
DATA MANU;
INPUT C $ M $ H $ Y @@;
TRT=COMPRESS (C) || COMPRESS (M);
LABEL C="LEVEL OF CONTAMINATION" M="METHOD OF CONTROL";
CARDS;

LOW M1 H1 203    LOW M1 H2 215    LOW M1 H3 203
LOW M1 H1 206    LOW M1 H2 210    LOW M1 H3 201
LOW M2 H1 222    LOW M2 H2 223    LOW M2 H3 215
LOW M2 H1 230    LOW M2 H2 216    LOW M2 H3 224
LOW M3 H1 239    LOW M3 H2 245    LOW M3 H3 237
LOW M3 H1 229    LOW M3 H2 238    LOW M3 H3 222
LOW M1 H4 190    LOW M1 H5 194    LOW M1 H6 191
LOW M1 H4 186    LOW M1 H5 199    LOW M1 H6 192
LOW M2 H4 217    LOW M2 H5 218    LOW M2 H6 215
LOW M2 H4 217    LOW M2 H5 209    LOW M2 H6 216
LOW M3 H4 235    LOW M3 H5 236    LOW M3 H6 234
LOW M3 H4 232    LOW M3 H5 233    LOW M3 H6 234
HIGH M1 H7 225   HIGH M1 H8 228   HIGH M1 H9 232
HIGH M1 H7 229   HIGH M1 H8 227   HIGH M1 H9 232
HIGH M2 H7 277   HIGH M2 H8 276   HIGH M2 H9 289
HIGH M2 H7 276   HIGH M2 H8 277   HIGH M2 H9 287
HIGH M3 H7 280   HIGH M3 H8 278   HIGH M3 H9 282
HIGH M3 H7 276   HIGH M3 H8 277   HIGH M3 H9 282
HIGH M1 H10 235  HIGH M1 H11 233  HIGH M1 H12 237
HIGH M1 H10 231  HIGH M1 H11 238  HIGH M1 H12 235
HIGH M2 H10 284  HIGH M2 H11 285  HIGH M2 H12 286
HIGH M2 H10 286  HIGH M2 H11 287  HIGH M2 H12 289
HIGH M3 H10 292  HIGH M3 H11 296  HIGH M3 H12 295
HIGH M3 H10 294  HIGH M3 H11 294  HIGH M3 H12 297
RUN;

PROC GLM;
CLASS C M H;
MODEL Y = C M C*M H(C) M*H(C);
RANDOM H(C) M*H(C)/TEST;
LSMEANS C M C*M/STDEERR PDIFF ADJUST=TUKEY;
RUN;

PROC MIXED CL ALPHA=.05 COVTEST;
CLASS C M H;
MODEL Y = C M C*M /ddfmsatterth;
RANDOM H(C) M*H(C);
LSMEANS C M C*M/ ADJUST=TUKEY;
RUN;

PROC GLM;
CLASS TRT;
MODEL Y = TRT;
MEANS TRT/HOVTEST=BF;
OUTPUT OUT=ASSUMP R=RESID P=MEANS;
PROC GPLOT; PLOT RESID*MEANS/VREF=0;
PROC UNIVARIATE DEF=5 PLOT NORMAL;
VAR RESID;
RUN;
ods graphics off;ods html close;
```

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

Class Level Information		
Class	Levels	Values
C	2	HIGH LOW
M	3	M1 M2 M3
H	12	H1 H10 H11 H12 H2 H3 H4 H5 H6 H7 H8 H9

Number of Observations Read 72

Number of Observations Used 72

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	77515.61111	2214.73175	184.99	<.0001
Error	36	431.00000	11.97222		
Corrected Total	71	77946.61111			

R-Square	Coeff Var	Root MSE	Y Mean
0.994471	1.427659	3.460090	242.3611

Source	DF	Type I SS	Mean Square	F Value	Pr > F
C	1	44900.05556	44900.05556	3750.35	<.0001
M	2	27135.02778	13567.51389	1133.25	<.0001
C*M	2	3143.02778	1571.51389	131.26	<.0001
H(C)	10	1611.22222	161.12222	13.46	<.0001
M*H(C)	20	726.27778	36.31389	3.03	0.0018

Source	DF	Type III SS	Mean Square	F Value	Pr > F
C	1	44900.05556	44900.05556	3750.35	<.0001
M	2	27135.02778	13567.51389	1133.25	<.0001
C*M	2	3143.02778	1571.51389	131.26	<.0001
H(C)	10	1611.22222	161.12222	13.46	<.0001
M*H(C)	20	726.27778	36.31389	3.03	0.0018

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

Source	Type III Expected Mean Square
C	Var(Error) + 2 Var(M*H(C)) + 6 Var(H(C)) + Q(C,C*M)
M	Var(Error) + 2 Var(M*H(C)) + Q(M,C*M)
C*M	Var(Error) + 2 Var(M*H(C)) + Q(C*M)
H(C)	Var(Error) + 2 Var(M*H(C)) + 6 Var(H(C))
M*H(C)	Var(Error) + 2 Var(M*H(C))

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* C	1	44900	44900	278.67	<.0001
Error: MS(H(C))	10	1611.22222	161.12222		

\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* M	2	27135	13568	373.62	<.0001
C*M	2	3143.027778	1571.513889	43.28	<.0001
H(C)	10	1611.22222	161.12222	4.44	0.0022
Error: MS(M*H(C))	20	726.277778	36.313889		

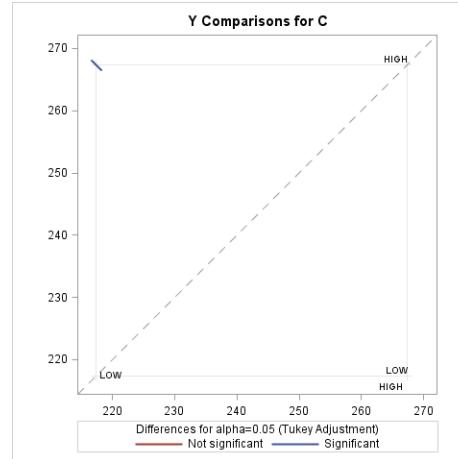
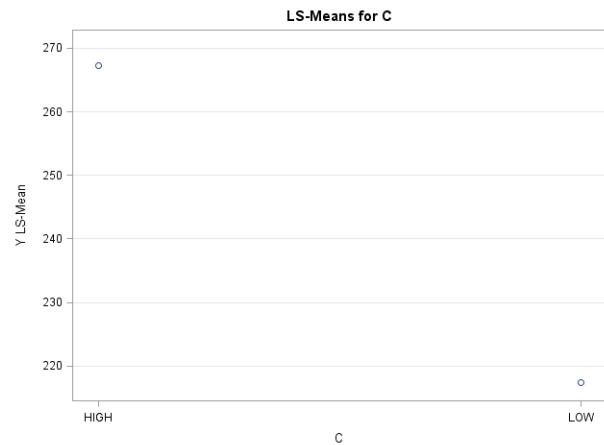
\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
M*H(C)	20	726.277778	36.313889	3.03	0.0018
Error: MS(Error)	36	431.000000	11.97222		

**SAS OUTPUT FOR PROBLEM II**

Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

C	Y LSMEAN	Standard Error	H0:LSMEAN=0	H0:LSMean1=LSMean2
			Pr >  t	Pr >  t
HIGH	267.333333	0.576682	<.0001	<.0001
LOW	217.388889	0.576682	<.0001	



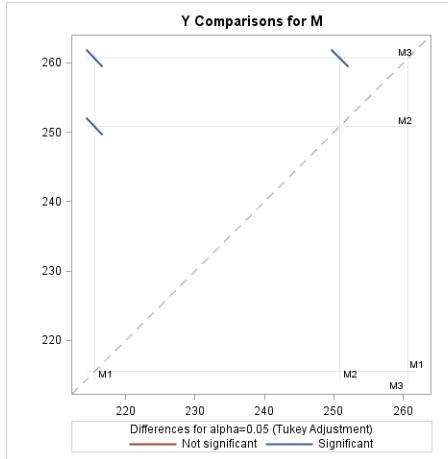
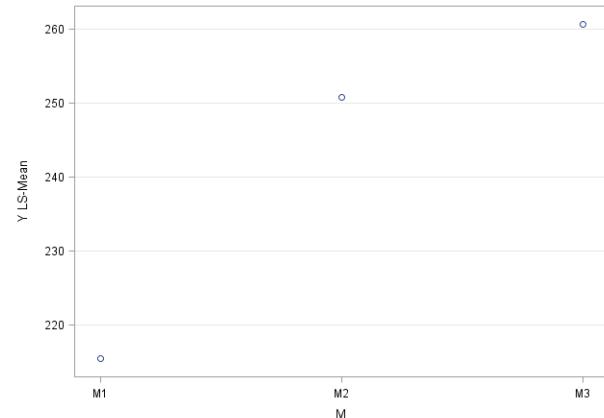
**SAS OUTPUT FOR PROBLEM II**

Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

M	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
M1	215.500000	0.706288	<.0001	1
M2	250.875000	0.706288	<.0001	2
M3	260.708333	0.706288	<.0001	3

Least Squares Means for effect M  
Pr > |t| for H0: LSMean(i)=LSMean(j)  
Dependent Variable: Y

i\j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

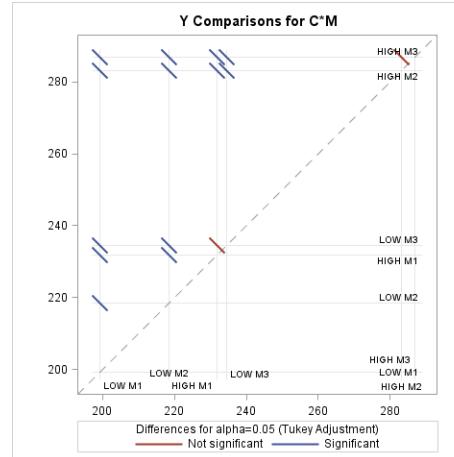
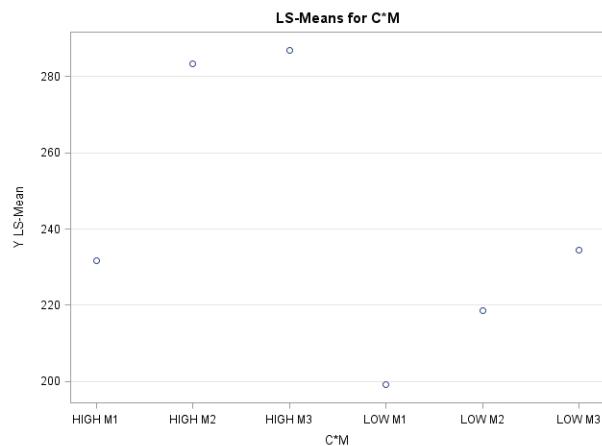
**LS-Means for M**

**SAS OUTPUT FOR PROBLEM II**

Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

C	M	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
HIGH	M1	231.833333	0.998842	<.0001	1
HIGH	M2	283.250000	0.998842	<.0001	2
HIGH	M3	286.916667	0.998842	<.0001	3
LOW	M1	199.166667	0.998842	<.0001	4
LOW	M2	218.500000	0.998842	<.0001	5
LOW	M3	234.500000	0.998842	<.0001	6

Least Squares Means for effect C*M						
Pr >  t  for H0: LSMean(i)=LSMean(j)						
Dependent Variable: Y						
i\j	1	2	3	4	5	6
1		<.0001	<.0001	<.0001	<.0001	0.4257
2	<.0001		0.1245	<.0001	<.0001	<.0001
3	<.0001	0.1245		<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001		<.0001	<.0001
5	<.0001	<.0001	<.0001	<.0001		<.0001
6	0.4257	<.0001	<.0001	<.0001	<.0001	



**SAS OUTPUT FOR PROBLEM II**

The Mixed Procedure

Model Information	
Data Set	WORK.MANU
Dependent Variable	Y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Class Level Information		
Class	Levels	Values
C	2	HIGH LOW
M	3	M1 M2 M3
H	12	H1 H10 H11 H12 H2 H3 H4 H5 H6 H7 H8 H9

Dimensions	
Covariance Parameters	3
Columns in X	12
Columns in Z	48
Subjects	1
Max Obs Per Subject	72

Number of Observations	
Number of Observations Read	72
Number of Observations Used	72
Number of Observations Not Used	0

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	448.81213510	
1	1	414.24817360	0.00000000

Convergence criteria met.

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
H(C)	20.8014	12.1609	1.71	0.0436	0.05	8.5641	103.69
M*H(C)	12.1708	5.9125	2.06	0.0198	0.05	5.6533	42.5624
Residual	11.9722	2.8219	4.24	<.0001	0.05	7.9174	20.2007

Fit Statistics	
-2 Res Log Likelihood	414.2
AIC (smaller is better)	420.2
AICC (smaller is better)	420.6
BIC (smaller is better)	421.7

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
C	1	10	278.67	<.0001
M	2	20	373.62	<.0001
C*M	2	20	43.28	<.0001

Least Squares Means							
Effect	LEVEL OF CONTAMINATION	METHOD OF CONTROL	Estimate	Standard Error	DF	t Value	Pr >  t
C	HIGH		267.33	2.1156	10	126.36	<.0001
C	LOW		217.39	2.1156	10	102.76	<.0001
M		M1	215.50	1.8018	19.1	119.60	<.0001
M		M2	250.87	1.8018	19.1	139.23	<.0001
M		M3	260.71	1.8018	19.1	144.69	<.0001
C*M	HIGH	M1	231.83	2.5481	19.1	90.98	<.0001
C*M	HIGH	M2	283.25	2.5481	19.1	111.16	<.0001
C*M	HIGH	M3	286.92	2.5481	19.1	112.60	<.0001
C*M	LOW	M1	199.17	2.5481	19.1	78.16	<.0001
C*M	LOW	M2	218.50	2.5481	19.1	85.75	<.0001
C*M	LOW	M3	234.50	2.5481	19.1	92.03	<.0001

Differences of Least Squares Means											
Effect	LEVEL OF CONTAMINATION	METHOD OF CONTROL	LEVEL OF CONTAMINATION	METHOD OF CONTROL	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P
C	HIGH		LOW		49.9444	2.9919	10	16.69	<.0001	Tukey-Kramer	<.0001
M		M1		M2	-35.3750	1.7396	20	-20.34	<.0001	Tukey-Kramer	<.0001
M		M1		M3	-45.2083	1.7396	20	-25.99	<.0001	Tukey-Kramer	<.0001
M		M2		M3	-9.8333	1.7396	20	-5.65	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M1	HIGH	M2	-51.4167	2.4601	20	-20.90	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M1	HIGH	M3	-55.0833	2.4601	20	-22.39	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M1	LOW	M1	32.6667	3.6036	19.1	9.06	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M1	LOW	M2	13.3333	3.6036	19.1	3.70	0.0015	Tukey-Kramer	0.0154
C*M	HIGH	M1	LOW	M3	-2.6667	3.6036	19.1	-0.74	0.4683	Tukey-Kramer	0.9743
C*M	HIGH	M2	HIGH	M3	-3.6667	2.4601	20	-1.49	0.1517	Tukey-Kramer	0.6739
C*M	HIGH	M2	LOW	M1	84.0833	3.6036	19.1	23.33	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M2	LOW	M2	64.7500	3.6036	19.1	17.97	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M2	LOW	M3	48.7500	3.6036	19.1	13.53	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M3	LOW	M1	87.7500	3.6036	19.1	24.35	<.0001	Tukey-Kramer	<.0001
C*M	HIGH	M3	LOW	M2	68.4167	3.6036	19.1	18.99	<.0001	Tukey-Kramer	<.0001

C*M	HIGH	M3	LOW	M3	52.4167	3.6036	19.1	14.55	<.0001	Tukey-Kramer	<.0001
C*M	LOW	M1	LOW	M2	-19.3333	2.4601	20	-7.86	<.0001	Tukey-Kramer	<.0001
C*M	LOW	M1	LOW	M3	-35.3333	2.4601	20	-14.36	<.0001	Tukey-Kramer	<.0001
C*M	LOW	M2	LOW	M3	-16.0000	2.4601	20	-6.50	<.0001	Tukey-Kramer	<.0001

---

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

Class Level Information		
Class	Levels	Values
TRT	6	HIGHM1 HIGHM2 HIGHM3 LOWM1 LOWM2 LOWM3

Number of Observations Read 72

Number of Observations Used 72

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

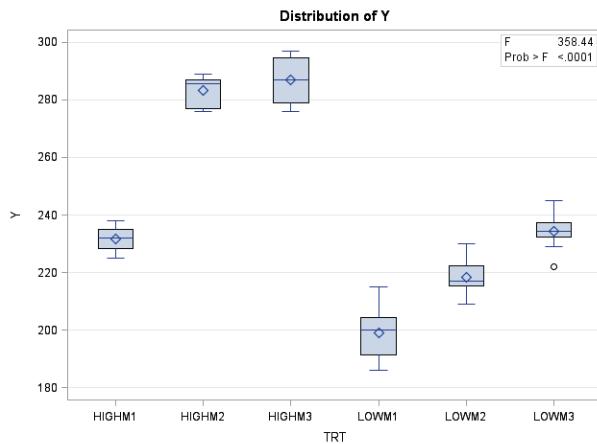
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	75178.11111	15035.62222	358.44	<.0001
Error	66	2768.50000	41.94697		
Corrected Total	71	77946.61111			

R-Square	Coeff Var	Root MSE	Y Mean
0.964482	2.672313	6.476648	242.3611

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	5	75178.11111	15035.62222	358.44	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	5	75178.11111	15035.62222	358.44	<.0001

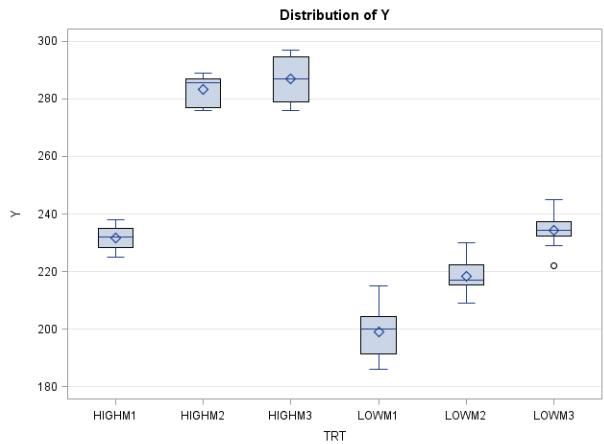
**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure

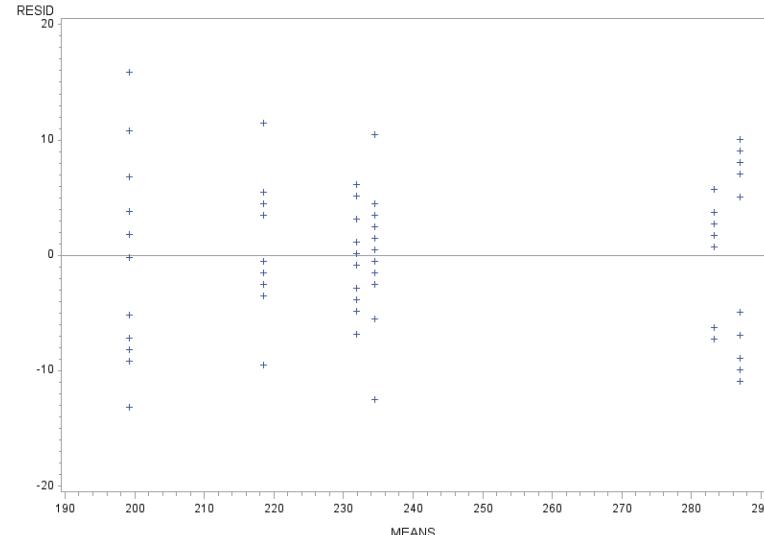
Brown and Forsythe's Test for Homogeneity of Y Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
TRT	5	230.1	46.0222	3.55	0.0066
Error	66	854.8	12.9520		

**SAS OUTPUT FOR PROBLEM II**

The GLM Procedure



Level of TRT	N	Y	
		Mean	Std Dev
HIGHM1	12	231.833333	4.04145188
HIGHM2	12	283.250000	5.18958748
HIGHM3	12	286.916667	8.36071260
LOWM1	12	199.166667	8.78876694
LOWM2	12	218.500000	5.45227225
LOWM3	12	234.500000	5.61653403

**SAS OUTPUT FOR PROBLEM II**

**SAS OUTPUT FOR PROBLEM II**

The UNIVARIATE Procedure  
Variable: RESID

Moments		
N	72	Sum Weights
Mean	0	Sum Observations
Std Deviation	6.24443414	Variance
Skewness	0.02301717	Kurtosis
Uncorrected SS	2768.5	Corrected SS
Coeff Variation	.	Std Error Mean

Basic Statistical Measures		
Location	Variability	
Mean	0.00000	Std Deviation
Median	0.16667	Variance
Mode	-7.25000	Range
		Interquartile Range

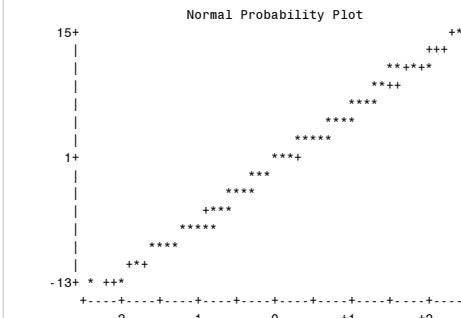
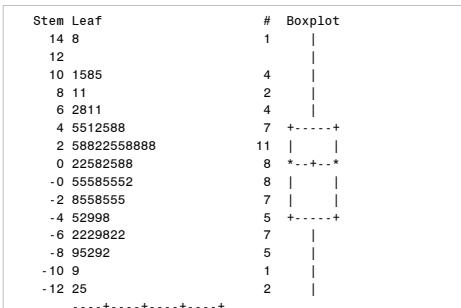
Note: The mode displayed is the smallest of 14 modes with a count of 2.

Tests for Location: Mu0=0		
Test	Statistic	p Value
Student's t	t	0 Pr >  t  1.0000
Sign	M	1 Pr >=  M  0.9063
Signed Rank	S	-2 Pr >=  S  0.9911

Tests for Normality		
Test	Statistic	p Value
Shapiro-Wilk	W	0.990551 Pr < W 0.8709
Kolmogorov-Smirnov	D	0.059062 Pr > D >0.1500
Cramer-von Mises	W-Sq	0.03672 Pr > W-Sq >0.2500
Anderson-Darling	A-Sq	0.222575 Pr > A-Sq >0.2500

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	15.833333
99%	15.833333
95%	10.500000
90%	7.083333
75% Q3	4.166667
50% Median	0.166667
25% Q1	-4.916667
10%	-8.166667
5%	-9.916667
1%	-13.166667
0% Min	-13.166667

Extreme Observations			
Lowest	Obs	Highest	Obs
-13.16667	22	10.0833	72
-12.50000	18	10.5000	14
-10.91667	52	10.8333	5
-9.91667	53	11.5000	10
-9.50000	29	15.8333	2



## **MASTER'S DIAGNOSTIC EXAMINATION - January 2013**

Student's Name \_\_\_\_\_

### **INSTRUCTIONS FOR STUDENTS:**

1. The exam has an Instruction page, 8 pages of questions, and 21 pages of SAS code and output.
2. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the  
UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
3. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
4. Use only one side of each sheet of paper.
5. You must answer all four questions: Questions I, II, III and IV.
6. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
7. 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.
8. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for your solutions to the questions on this examination
  - No other materials are allowed  
  - I attest that I spent no more than 4 hours to complete the exam,
  - I used only the materials described above, and
  - I did not receive assistance from anyone during the taking of this exam.

Student's Signature\_\_\_\_\_

### **INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or **scan** the solutions into a pdf file and **email** to **longneck@stat.tamu.edu**.

**DO NOT send the exam booklet or SAS output, just send the student's solutions.**

1. I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the student's solutions were faxed to **979-845-6060** or  
emailed to **longneck@stat.tamu.edu**.

Proctor's Signature\_\_\_\_\_

## QUESTION I.

You are the consulting statistician on a study of the genetic components of exercise. The PI plans to raise 30 mice in which a particular gene suspected to be involved in regulating exercise has been knocked out (you can think of it like turning off this gene), as well as 30 wild-type mice (mice in which the gene has not been knocked out). Each mouse will then be observed over a two week period, and the average number of minutes mouse  $i$  spends running on the exercise wheel per day,  $y_i$ , will be recorded. The variables in the study are

- the 60 response values  $y_1, y_2, \dots, y_{60}$
- a variable indicating comparison group ( $x_i = 1$  if knockout,  $x_i = 0$  if wild-type,  $i = 1, 2, \dots, 60$ )
- gender ( $g_i = 1$  if female,  $g_i = 0$  if male,  $i = 1, 2, \dots, 60$ )
- average daily food consumption ( $f_i$ , a continuous number, standardized so that a value of 0 indicates average consumption, negative values indicate less than average, positive values indicate greater than average,  $i = 1, 2, \dots, 60$ ).

Based on your exploratory analysis of preliminary data, you and the PI agree that a sensible model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 f_i + \beta_4(x_i \times g_i) + \epsilon_i,$$

where the  $\epsilon_i$  are i.i.d. with mean 0 and s.d.  $\sigma$ .

- 1.) Interpret each of the coefficients in the above model.
- 2.) Using the following R output construct an approximate 95% confidence interval for  $\beta_1 - \beta_4$ ?

Based on the interval you report, comment on your conclusions regarding whether there is a genotype effect.

**Hint:** If  $\mathbf{v}$  is a column vector of length 5, then the standard error of  $\mathbf{v}'\hat{\boldsymbol{\beta}}$  is equal to the square root of  $\hat{\sigma}^2 \mathbf{v}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}$ , where  $\hat{\sigma}$  is the estimated residual s.d., and  $\mathbf{X}$  is the model matrix. In our example,  $\hat{\sigma} = 1.716$  and

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.07 & -0.07 & -0.07 & -0.01 & 0.07 \\ -0.07 & 0.14 & 0.07 & 0.01 & -0.13 \\ -0.07 & 0.07 & 0.13 & 0.00 & -0.13 \\ -0.01 & 0.01 & 0.00 & 0.05 & 0.00 \\ 0.07 & -0.13 & -0.13 & 0.00 & 0.27 \end{pmatrix}$$

Here's the R output from fitting the above model:

```
Call:
lm(formula = y ~ 1 + x + g + f + x * g)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3862 -1.0618 -0.0837  1.1362  4.2091 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.6375    0.4463  23.835 < 2e-16 ***
x           0.8728    0.6325   1.380  0.17319    
g          -0.9642    0.6265  -1.539  0.12955    
f           1.1632    0.3880   2.998  0.00407 **  
x:g        0.0586    0.8867   0.066  0.94755    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.716 on 55 degrees of freedom
Multiple R-squared:  0.221, Adjusted R-squared:  0.1643 
F-statistic:  3.9 on 4 and 55 DF,  p-value: 0.007369
```

- 3.) An F-test of the null hypothesis that  $\beta_1 = \beta_4 = 0$  returned a p-value of 0.15. Meanwhile, stepwise model selection tells you that the best-fitting model is the one without the treatment-by-gender interaction, with abbreviated R output:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	10.6230	0.3853	27.573	< 2e-16 ***
x	0.9018	0.4404	2.048	0.04535 *
g	-0.9349	0.4392	-2.128	0.03771 *
f	1.1622	0.3843	3.025	0.00375 **

The PI is primarily interested in whether there is sufficient evidence of an effect of treatment, the knockout variable, on time spent exercising in mice. Based on the F-test and stepwise regression results, how would you advise the PI? Provide a detail explanation of your advice.

## QUESTION II.

The Storm Prediction Center (an agency of NOAA) tracks the number and characteristics of tornadoes. In this problem we will consider the variable **Killer\_tornadoes**, which is the number of tornadoes in a given year which resulted in one or more deaths. The summary statistics for the variable **Killer\_tornadoes** over 48 years of data were found using the SAS function `proc means`:

```

The MEANS Procedure

Analysis Variable : Killer_tornadoes

      N        Mean       Std Dev      Minimum      Maximum
      48     2.0416667    2.9532408         0    13.0000000
  
```

Use the above information, the SAS output, and the tables on page 5 to answer the following questions.

- 1.) Explain why  $\bar{X} \pm Z_{\alpha/2}\sqrt{\bar{X}/n}$  is a reasonable  $1 - \alpha$  confidence set for  $\lambda$ , the mean of a Poisson distribution. Then assuming that the variable **Killer\_tornadoes** has a Poisson distribution, obtain a 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 chi-squared test statistic was computed to be  $\chi^2 = 32.35$ . 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.

Using the 48 years of data on the pair (**Tornadoes**, **Killer\_tornadoes**), a simple linear regression model was fit with **Killer\_tornadoes** as the response and **Tornadoes** as the predictor. Some SAS output is given below. A scatter plot of the data and a plot of residuals versus the predictor are on the next page. Discuss the appropriateness of using this prediction model for the number of killer tornadoes.

```

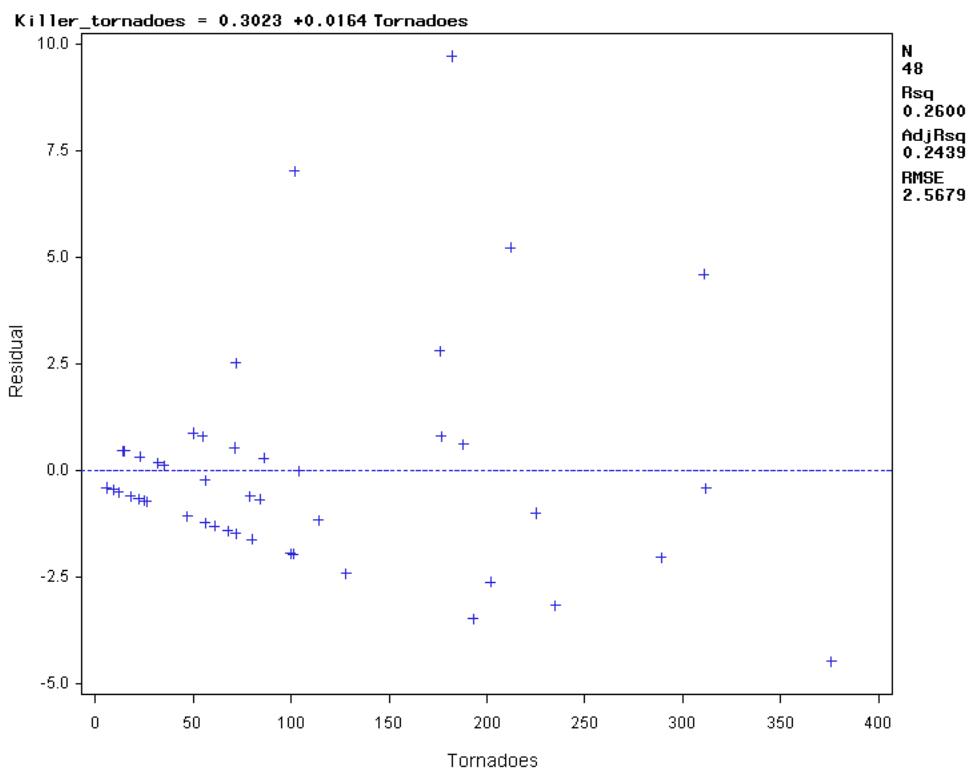
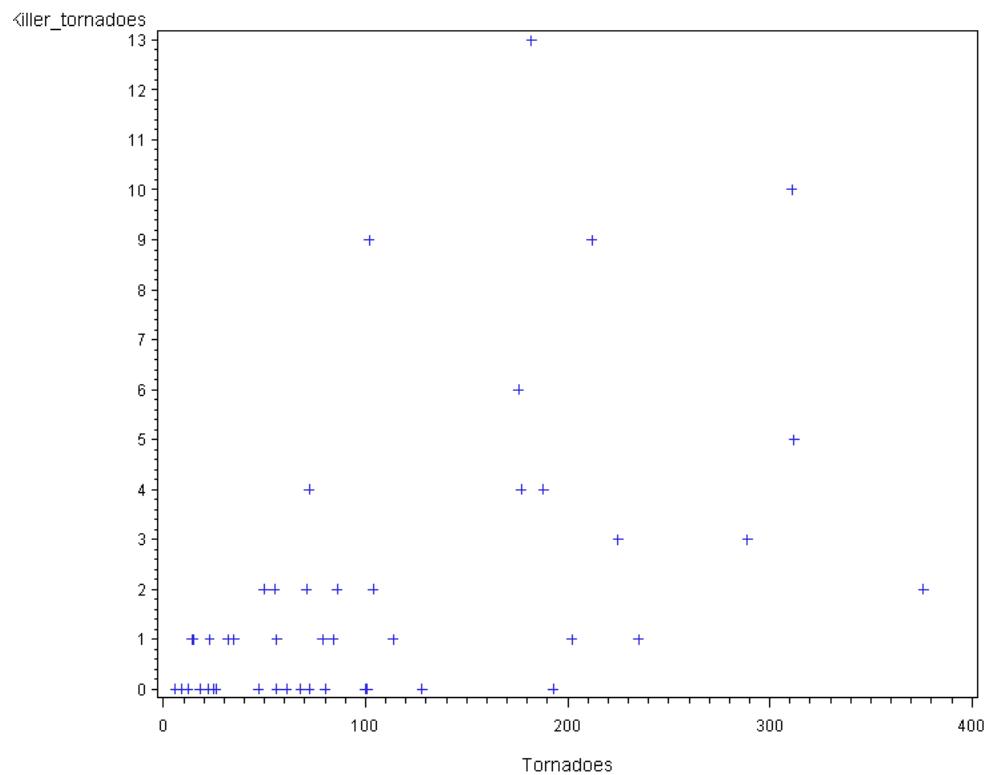
Analysis of Variance

      Source        DF      Sum of Squares      Mean 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

      Variable        DF      Parameter Estimate      Standard 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
  
```



### Some Chi-Squared Percentiles

df	Right-Tail Probability			
	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

### Some t-Distribution Percentiles

df	Right-Tail Probability			
	0.100	0.050	0.025	0.010
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
6	1.440	1.943	2.447	3.143
8	1.397	1.860	2.306	2.896
12	1.356	1.782	2.179	2.681
27	1.314	1.703	2.052	2.473
53	1.298	1.674	2.006	2.399
54	1.297	1.674	2.005	2.397
$\infty$	1.282	1.645	1.960	2.326

### QUESTION III.

Consider a regression model for an experiment with a continuous response  $Y$  and two factors,  $A$  and  $B$ , each with two levels, *High* and *Low*. Consider the two models:

- **Dummy Variable Model :**  $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

where

$$x_1 = \begin{cases} 1 & \text{if } A = \text{High} \\ 0 & \text{if } A = \text{Low} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if } B = \text{High} \\ 0 & \text{if } B = \text{Low} \end{cases}$$

- **Design Effects Model :**  $E(Y) = \alpha^* + \beta_1^* x_1^* + \beta_2^* x_2^* + \beta_3^* x_1^* x_2^*$ ,

where

$$x_1^* = \begin{cases} 1 & \text{if } A = \text{High} \\ -1 & \text{if } A = \text{Low} \end{cases} \quad x_2^* = \begin{cases} 1 & \text{if } B = \text{High} \\ -1 & \text{if } B = \text{Low} \end{cases}$$

- 1.) Express the following null hypotheses in term of the coefficients for the Dummy Variable Model :
  - No interaction between  $A$  and  $B$
  - No effect due to  $A$
  - No effect due to  $B$
- 2.) By considering the mean response for the four treatments (combinations of  $A$  and  $B$ ), determine the relationship between  $\alpha, \beta_1, \beta_2, \beta_3$  and  $\alpha^*, \beta_1^*, \beta_2^*, \beta_3^*$ .
- 3.) Using the Dummy Variable Model, obtain expressions for the difference in mean response between
  - $A_{\text{high}}B_{\text{high}}$  and  $A_{\text{low}}B_{\text{low}}$
  - $A_{\text{high}}B_{\text{high}}$  and  $A_{\text{high}}B_{\text{low}}$
  - $A_{\text{low}}B_{\text{high}}$  and  $A_{\text{low}}B_{\text{low}}$ .
- 4.) Using the Design Effects Model, obtain expressions for the difference in mean response between
  - $A_{\text{high}}B_{\text{high}}$  and  $A_{\text{low}}B_{\text{low}}$
  - $A_{\text{high}}B_{\text{high}}$  and  $A_{\text{high}}B_{\text{low}}$
  - $A_{\text{low}}B_{\text{high}}$  and  $A_{\text{low}}B_{\text{low}}$ .
- 5.) Suppose that there is also a third factor  $C$  with two levels, *High* and *Low*. Describe how to form a Dummy Variable Model for the mean response in each of the following cases:
  - Main effects only for  $A$ ,  $B$ , and  $C$ .
  - Main effects and all two-way interactions for  $A$ ,  $B$ , and  $C$ .
  - Main effects and all possible interactions for  $A$ ,  $B$ , and  $C$ .
- 6.) Consider now the Dummy Variable Model for the logarithm of the mean response:

- **Dummy Variable Model :**  $\log(E(Y)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ ,

where

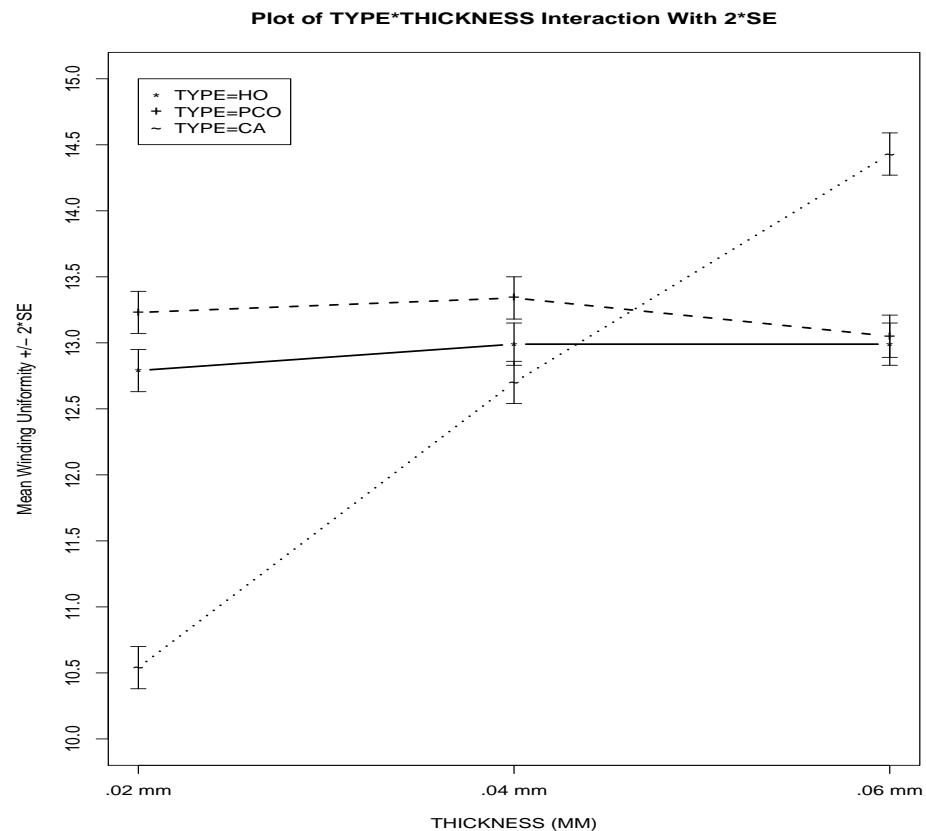
$$x_1 = \begin{cases} 1 & \text{if } A = \text{High} \\ 0 & \text{if } A = \text{Low} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if } B = \text{High} \\ 0 & \text{if } B = \text{Low} \end{cases}$$

Explain what the coefficients  $\alpha, \beta_1, \beta_2, \beta_3$  represent in terms of the mean response  $E(Y)$ .

#### QUESTION IV.

A company is interested in the ability of a machine to consistently place electrical wire on a coil. There are three types of machines available: hand operated(HO), partially computer operated(PCO), and completely automated(CA). Three machines of each type are randomly selected from their suppliers for use in the study. The wire placed on the coils comes in one of three thicknesses: .02mm, .04mm, or .06mm. Each of the machines assembles two coils of each of the three wire thicknesses. Each wound coil is then measured for the uniformity of windings at a middle position on the coil. These measurements are given in the following table.

		TYPE OF MACHINE								
		HO			PCO			CA		
MACHINE ID		1	2	3	1	2	3	1	2	3
THICKNESS	.02mm	12.30	13.46	12.35	13.01	13.46	13.15	10.47	10.75	10.24
	.02mm	12.59	14.00	12.06	12.63	13.92	13.20	10.96	19.68	10.15
	.04mm	13.15	13.29	12.50	12.74	13.84	13.46	12.73	12.60	12.92
	.04mm	13.00	13.62	12.39	12.68	13.75	13.57	12.64	12.65	12.64
	.06mm	12.87	13.46	12.73	12.47	13.62	13.36	14.01	14.80	14.19
	.06mm	12.92	13.82	12.15	12.15	13.28	13.42	14.62	14.71	14.23



Use the above plot, data, the attached SAS output, and the tables on page 5 to answer the following questions.

- 1.) Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.

$C_1$  - Normality:

$C_2$  - Equal Variance:

$C_3$  - Independence:

- 2.) Write a model for  $y_{ijk\ell}$ , the uniformity of windings on the  $\ell$ th Coil wound by the  $k$ th Machine of Type  $i$  using Wire Thickness  $j$ .

- Make sure to include all necessary conditions on parameters and random variables in your model.

- 3.) At the  $\alpha = .05$  level, which main effects and interactions are significant? Justify your answer by including the relevant p-values along with their pair of degrees of freedom ( $df_{NUM.}, df_{DEN.}$ ).

- 4.) The researchers were interested in estimating the difference in the mean uniformity in the windings between wires of thickness .02 mm and .06 mm wound using a Type CA machine:  $\mu_{31} - \mu_{33}$

Determine the variance of the estimated mean difference:  $\text{Var}(\hat{\mu}_{31} - \hat{\mu}_{33})$ , in terms of the variance components from your model.

Using the estimated variance components from the SAS output, compute the estimated standard error of the estimated mean difference:  $\widehat{SE}(\hat{\mu}_{31} - \hat{\mu}_{33})$ .

- Provide supporting details for your expressions.

- 5.) Place the three types of machines into groups such that the machines within a group are not significantly different from any machine in the same group with respect to their mean uniformity of windings. Use an experimentwise error rate of  $\alpha = .05$ .

- 6.) Provide a 95% confidence on the mean uniformity of windings of a wire of thickness .04 mm wound with a CA machine.

- 7.) The wire regulatory agency reviewed the study and were concerned that the rigidity of the different wire samples was not included in the study. Suggest a method by which the rigidity measurements of the 54 wire samples could be used in the analysis.

The following SAS program was used to analyze the data. The output is contained in the following pages.

```
ods html;ods graphics on;
OPTIONS LS=90 PS=55 nocenter nodate;
TITLE 'SAS OUTPUT FOR PROBLEM IV';
DATA MANU;
INPUT TYPE $ MACH $ THICK $ Y @@;
TRT=COMPRESS (TYPE) || COMPRESS (THICK);
CARDS;

HO M1 .02 12.30 HO M2 .02 13.46 HO M3 .02 12.35
HO M1 .02 12.59 HO M2 .02 14.00 HO M3 .02 12.06
HO M1 .04 13.15 HO M2 .04 13.29 HO M3 .04 12.50
HO M1 .04 13.00 HO M2 .04 13.62 HO M3 .04 12.39
HO M1 .06 12.87 HO M2 .06 13.46 HO M3 .06 12.73
HO M1 .06 12.92 HO M2 .06 13.82 HO M3 .06 12.15

PCO M1 .02 13.01 PCO M2 .02 13.46 PCO M3 .02 13.15
PCO M1 .02 12.63 PCO M2 .02 13.92 PCO M3 .02 13.20
PCO M1 .04 12.74 PCO M2 .04 13.84 PCO M3 .04 13.46
PCO M1 .04 12.68 PCO M2 .04 13.75 PCO M3 .04 13.57
PCO M1 .06 12.47 PCO M2 .06 13.62 PCO M3 .06 13.36
PCO M1 .06 12.15 PCO M2 .06 13.28 PCO M3 .06 13.42

CA M1 .02 10.47 CA M2 .02 10.75 CA M3 .02 10.24
CA M1 .02 10.96 CA M2 .02 10.68 CA M3 .02 10.15
CA M1 .04 12.73 CA M2 .04 12.60 CA M3 .04 12.92
CA M1 .04 12.64 CA M2 .04 12.65 CA M3 .04 12.64
CA M1 .06 14.01 CA M2 .06 14.80 CA M3 .06 14.19
CA M1 .06 14.62 CA M2 .06 14.71 CA M3 .06 14.23

RUN;

PROC GLM ORDER=DATA;
CLASS TYPE MACH THICK;
MODEL Y = TYPE THICK TYPE*THICK MACH(TYPE) THICK*MACH(TYPE);
RANDOM MACH(TYPE) THICK*MACH(TYPE)/TEST;
LSMEANS TYPE THICK THICK*TYPE/STDERR PDIFF ADJUST=TUKEY;
RUN;

PROC MIXED CL ALPHA=.05 COVTEST;
CLASS TYPE MACH THICK;
MODEL Y = TYPE THICK TYPE*THICK;
RANDOM MACH(TYPE) THICK*MACH(TYPE);
LSMEANS TYPE THICK TYPE*THICK/ ADJUST=TUKEY;
RUN;

PROC GLM;
CLASS TRT;
MODEL Y = TRT;
MEANS TRT/HOVTEST=BF;
OUTPUT OUT=ASSUMP R=RESID P=MEANS;
PROC GPLOT; PLOT RESID*MEANS/VREF=0;
PROC UNIVARIATE DEF=5 PLOT NORMAL;
VAR RESID;
RUN;
ods graphics off;ods html close;
```

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

Class Level Information		
Class	Levels	Values
TYPE	3	HO PCO CA
MACH	3	M1 M2 M3
THICK	3	.02 .04 .06

Number of Observations Read	54
Number of Observations Used	54

---

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure****Dependent Variable: Y**

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	59.31313333	2.28127436	51.34	<.0001
Error	27	1.19980000	0.04443704		
Corrected Total	53	60.51293333			

R-Square	Coeff Var	Root MSE	Y Mean
0.980173	1.634679	0.210801	12.89556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TYPE	2	3.83981111	1.91990556	43.21	<.0001
THICK	2	15.59923333	7.79961667	175.52	<.0001
TYPE*THICK	4	30.27515556	7.56878889	170.33	<.0001
MACH(TYPE)	6	8.46555556	1.41092593	31.75	<.0001
MACH*THICK(TYPE)	12	1.13337778	0.09444815	2.13	0.0507

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TYPE	2	3.83981111	1.91990556	43.21	<.0001
THICK	2	15.59923333	7.79961667	175.52	<.0001
TYPE*THICK	4	30.27515556	7.56878889	170.33	<.0001
MACH(TYPE)	6	8.46555556	1.41092593	31.75	<.0001
MACH*THICK(TYPE)	12	1.13337778	0.09444815	2.13	0.0507

---

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

Source	Type III Expected Mean Square
TYPE	Var(Error) + 2 Var(MACH*THICK(TYPE)) + 6 Var(MACH(TYPE)) + Q(TYPE,TYPE*THICK)
THICK	Var(Error) + 2 Var(MACH*THICK(TYPE)) + Q(THICK,TYPE*THICK)
TYPE*THICK	Var(Error) + 2 Var(MACH*THICK(TYPE)) + Q(TYPE*THICK)
MACH(TYPE)	Var(Error) + 2 Var(MACH*THICK(TYPE)) + 6 Var(MACH(TYPE))
MACH*THICK(TYPE)	Var(Error) + 2 Var(MACH*THICK(TYPE))

---

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure****Tests of Hypotheses for Mixed Model Analysis of Variance****Dependent Variable: Y**

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
*	<b>TYPE</b>	2	3.839811	1.919906	1.36	0.3256
	<b>Error</b>	6	8.465556	1.410926		

**Error: MS(MACH(TYPE))**

\* This test assumes one or more other fixed effects are zero.

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
*	<b>THICK</b>	2	15.599233	7.799617	82.58	<.0001
	<b>TYPE*THICK</b>	4	30.275156	7.568789	80.14	<.0001
	<b>MACH(TYPE)</b>	6	8.465556	1.410926	14.94	<.0001
	<b>Error</b>	12	1.133378	0.094448		

**Error: MS(MACH\*THICK(TYPE))**

\* This test assumes one or more other fixed effects are zero.

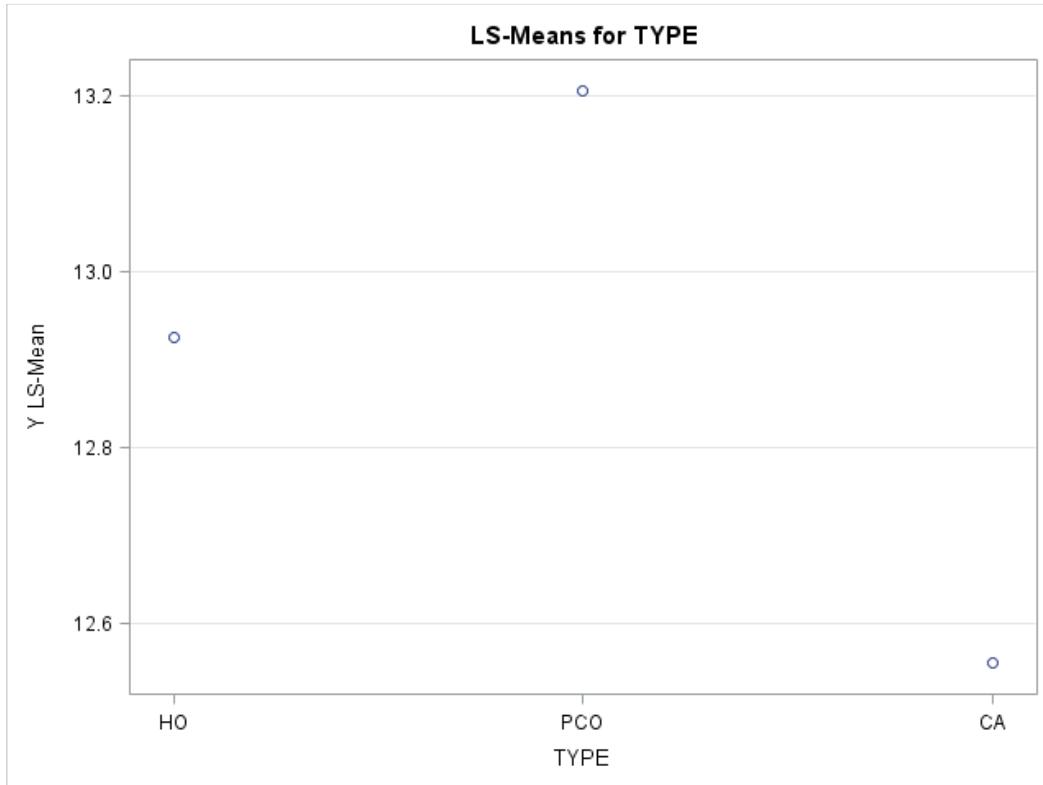
Source	DF	Type III SS	Mean Square	F Value	Pr > F
<b>MACH*THICK(TYPE)</b>	12	1.133378	0.094448	2.13	0.0507
<b>Error: MS(Error)</b>	27	1.199800	0.044437		

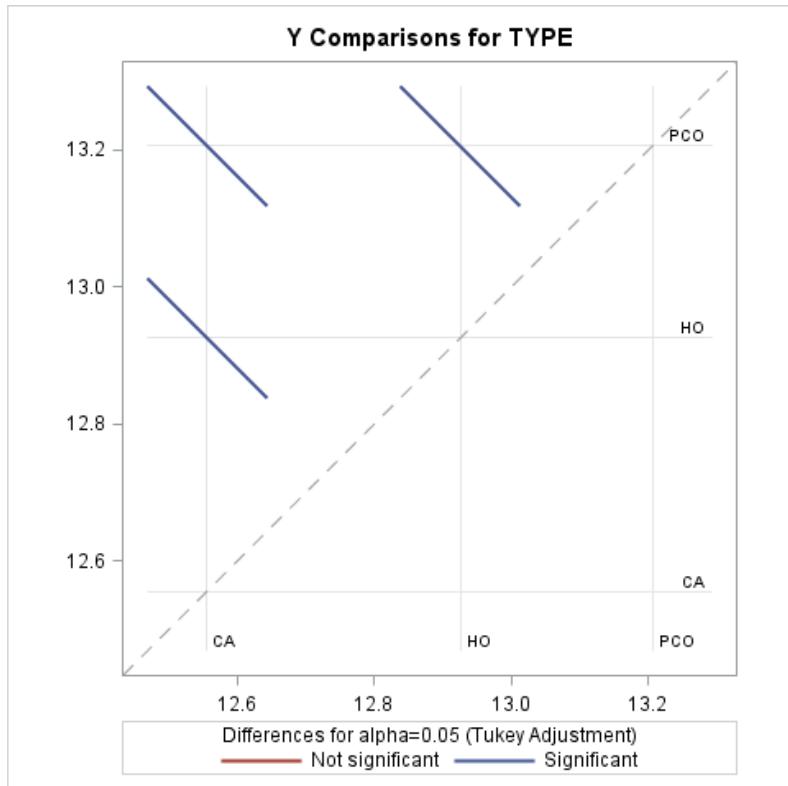
**SAS OUTPUT FOR PROBLEM IV**

Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

TYPE	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
HO	12.9255556	0.0496863	<.0001	1
PCO	13.2061111	0.0496863	<.0001	2
CA	12.5550000	0.0496863	<.0001	3

Least Squares Means for effect TYPE				
Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: Y				
i/j	1	2	3	
1		0.0013	<.0001	
2	0.0013		<.0001	
3	<.0001	<.0001		



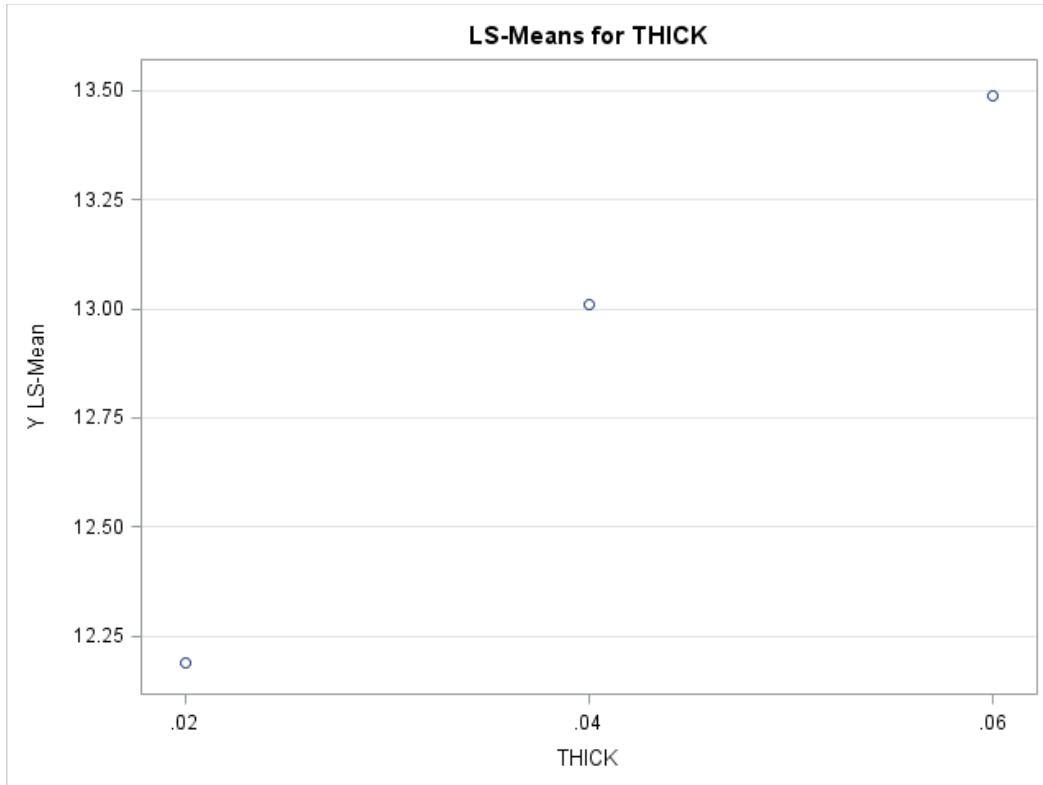


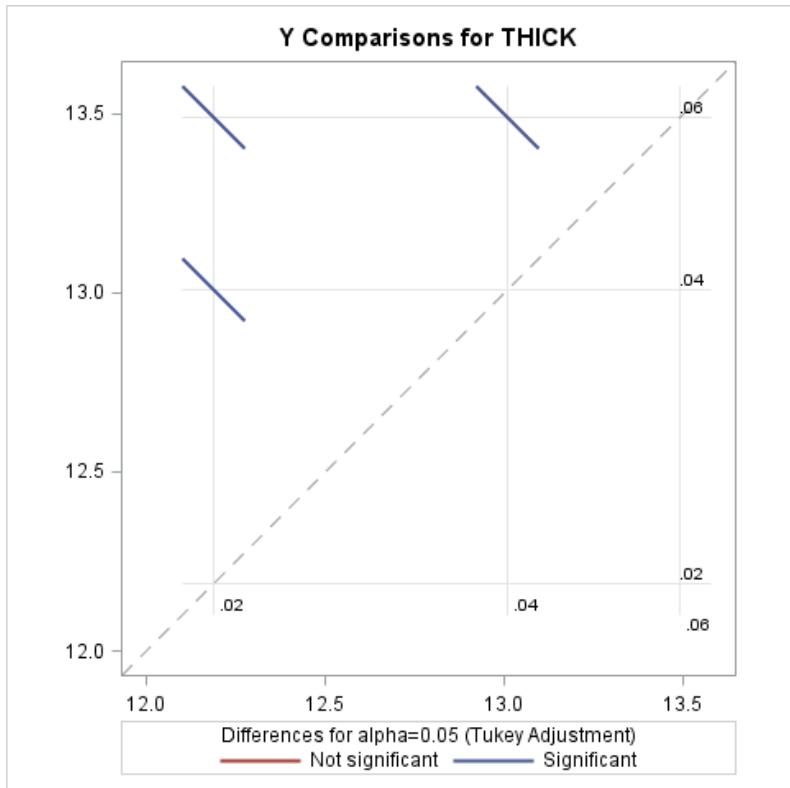
**SAS OUTPUT FOR PROBLEM IV**

Least Squares Means  
Adjustment for Multiple Comparisons: Tukey

THICK	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
.02	12.1877778	0.0496863	<.0001	1
.04	13.0094444	0.0496863	<.0001	2
.06	13.4894444	0.0496863	<.0001	3

Least Squares Means for effect THICK				
Pr >  t  for H0: LSMean(i)=LSMean(j)				
Dependent Variable: Y				
i/j	1	2	3	
1		<.0001	<.0001	
2	<.0001		<.0001	
3	<.0001	<.0001		



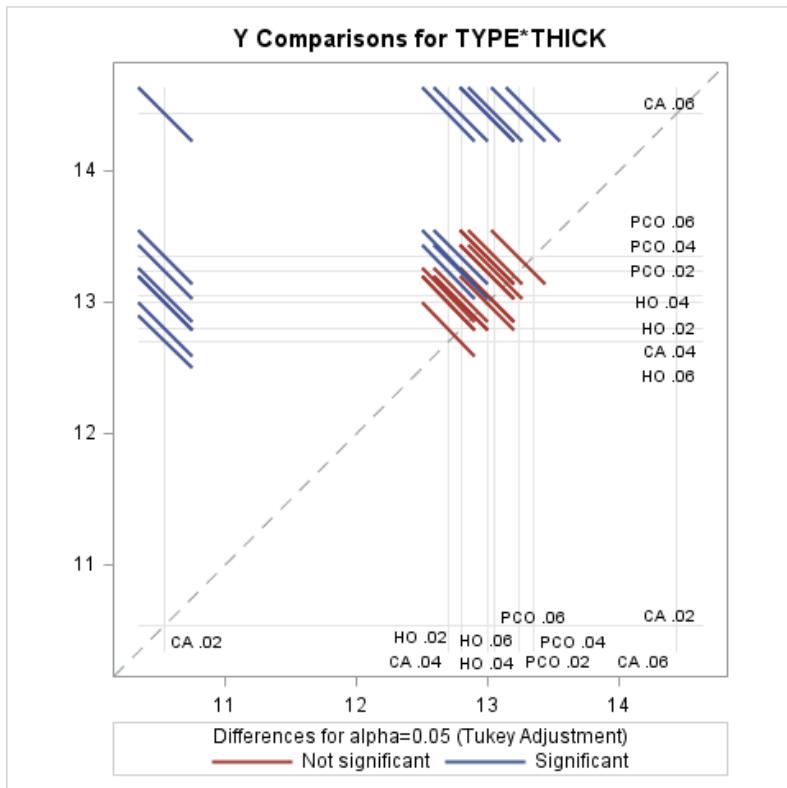
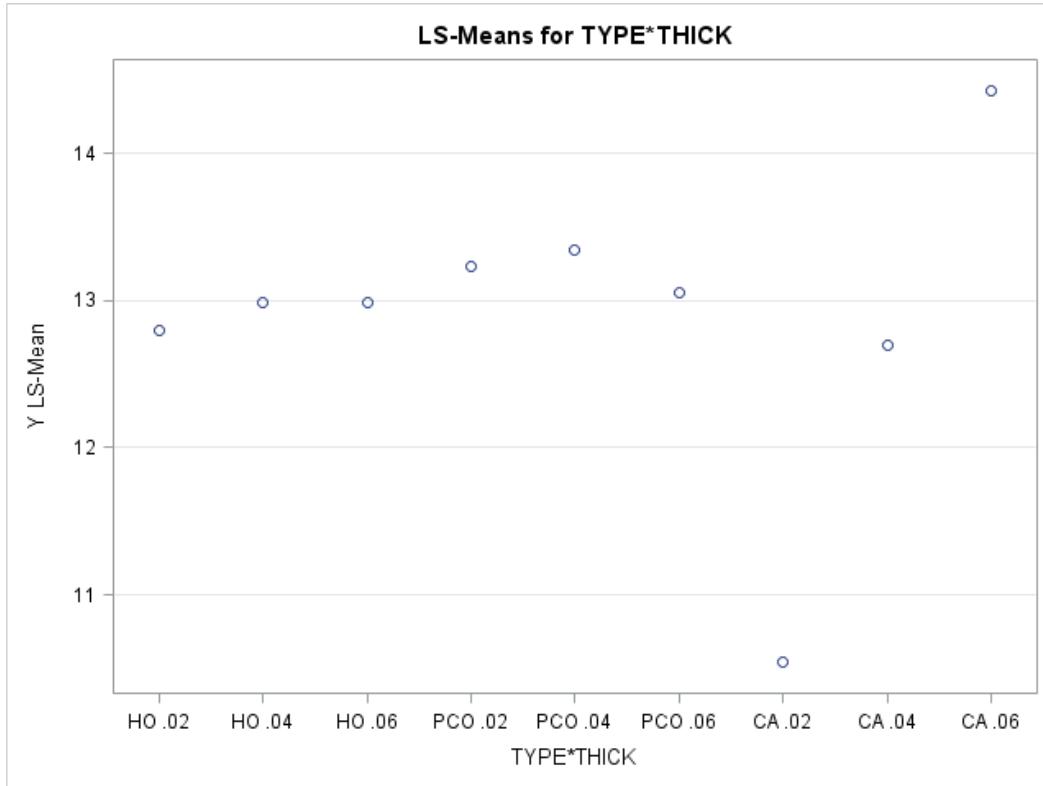


**SAS OUTPUT FOR PROBLEM IV**

**Least Squares Means**  
**Adjustment for Multiple Comparisons: Tukey**

TYPE	THICK	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
HO	.02	12.7933333	0.0860591	<.0001	1
HO	.04	12.9916667	0.0860591	<.0001	2
HO	.06	12.9916667	0.0860591	<.0001	3
PCO	.02	13.2283333	0.0860591	<.0001	4
PCO	.04	13.3400000	0.0860591	<.0001	5
PCO	.06	13.0500000	0.0860591	<.0001	6
CA	.02	10.5416667	0.0860591	<.0001	7
CA	.04	12.6966667	0.0860591	<.0001	8
CA	.06	14.4266667	0.0860591	<.0001	9

Least Squares Means for effect TYPE*THICK Pr >  t  for H0: LSMean(i)=LSMean(j) Dependent Variable: Y									
i\j	1	2	3	4	5	6	7	8	9
1		0.7808	0.7808	0.0310	0.0033	0.4882	<.0001	0.9961	<.0001
2	0.7808		1.0000	0.5913	0.1434	0.9999	<.0001	0.3117	<.0001
3	0.7808	1.0000		0.5913	0.1434	0.9999	<.0001	0.3117	<.0001
4	0.0310	0.5913	0.5913		0.9899	0.8614	<.0001	0.0045	<.0001
5	0.0033	0.1434	0.1434	0.9899		0.3323	<.0001	0.0004	<.0001
6	0.4882	0.9999	0.9999	0.8614	0.3323		<.0001	0.1324	<.0001
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
8	0.9961	0.3117	0.3117	0.0045	0.0004	0.1324	<.0001		<.0001
9	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	



**SAS OUTPUT FOR PROBLEM IV****The Mixed Procedure**

Model Information	
Data Set	WORK.MANU
Dependent Variable	Y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	Levels	Values
TYPE	3	CA HO PCO
MACH	3	M1 M2 M3
THICK	3	.02 .04 .06

Dimensions	
Covariance Parameters	3
Columns in X	16
Columns in Z	36
Subjects	1
Max Obs Per Subject	54

Number of Observations	
Number of Observations Read	54
Number of Observations Used	54
Number of Observations Not Used	0

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	79.60478912	
1	1	33.50991503	0.00000000

Convergence criteria met.

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
MACH(TYPE)	0.2194	0.1359	1.61	0.0532	0.05	0.08676	1.2528
MACH*THICK(TYPE)	0.02501	0.02021	1.24	0.1079	0.05	0.008092	0.3330
Residual	0.04444	0.01209	3.67	0.0001	0.05	0.02778	0.08233

Fit Statistics	
-2 Res Log Likelihood	33.5

AIC (smaller is better)	39.5
AICC (smaller is better)	40.1
BIC (smaller is better)	40.1

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
TYPE	2	6	1.36	0.3256
THICK	2	12	82.58	<.0001
TYPE*THICK	4	12	80.14	<.0001

Least Squares Means							
Effect	TYPE	THICK	Estimate	Standard Error	DF	t Value	Pr >  t
TYPE	CA		12.5550	0.2800	6	44.84	<.0001
TYPE	HO		12.9256	0.2800	6	46.17	<.0001
TYPE	PCO		13.2061	0.2800	6	47.17	<.0001
THICK		.02	12.1878	0.1721	12	70.81	<.0001
THICK		.04	13.0094	0.1721	12	75.58	<.0001
THICK		.06	13.4894	0.1721	12	78.37	<.0001
TYPE*THICK	CA	.02	10.5417	0.2981	12	35.36	<.0001
TYPE*THICK	CA	.04	12.6967	0.2981	12	42.59	<.0001
TYPE*THICK	CA	.06	14.4267	0.2981	12	48.39	<.0001
TYPE*THICK	HO	.02	12.7933	0.2981	12	42.91	<.0001
TYPE*THICK	HO	.04	12.9917	0.2981	12	43.58	<.0001
TYPE*THICK	HO	.06	12.9917	0.2981	12	43.58	<.0001
TYPE*THICK	PCO	.02	13.2283	0.2981	12	44.37	<.0001
TYPE*THICK	PCO	.04	13.3400	0.2981	12	44.75	<.0001
TYPE*THICK	PCO	.06	13.0500	0.2981	12	43.77	<.0001

Differences of Least Squares Means												
Effect	TYPE	THICK	_TYPE	_THICK	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P	
TYPE	CA		HO		-0.3706	0.3959	6	-0.94	0.3855	Tukey	0.6396	
TYPE	CA		PCO		-0.6511	0.3959	6	-1.64	0.1512	Tukey	0.2996	
TYPE	HO		PCO		-0.2806	0.3959	6	-0.71	0.5052	Tukey	0.7676	
THICK		.02		.04	-0.8217	0.1024	12	-8.02	<.0001	Tukey-Kramer	<.0001	
THICK		.02		.06	-1.3017	0.1024	12	-12.71	<.0001	Tukey-Kramer	<.0001	
THICK		.04		.06	-0.4800	0.1024	12	-4.69	0.0005	Tukey-Kramer	0.0014	
TYPE*THICK	CA	.02	CA	.04	-2.1550	0.1774	12	-12.15	<.0001	Tukey-Kramer	<.0001	
TYPE*THICK	CA	.02	CA	.06	-3.8850	0.1774	12	-21.90	<.0001	Tukey-Kramer	<.0001	
TYPE*THICK	CA	.02	HO	.02	-2.2517	0.4216	12	-5.34	0.0002	Tukey-Kramer	0.0037	
TYPE*THICK	CA	.02	HO	.04	-2.4500	0.4216	12	-5.81	<.0001	Tukey-Kramer	0.0018	
TYPE*THICK	CA	.02	HO	.06	-2.4500	0.4216	12	-5.81	<.0001	Tukey-Kramer	0.0018	
TYPE*THICK	CA	.02	PCO	.02	-2.6867	0.4216	12	-6.37	<.0001	Tukey-Kramer	0.0008	
TYPE*THICK	CA	.02	PCO	.04	-2.7983	0.4216	12	-6.64	<.0001	Tukey-Kramer	0.0005	
TYPE*THICK	CA	.02	PCO	.06	-2.5083	0.4216	12	-5.95	<.0001	Tukey-Kramer	0.0015	

<b>TYPE*THICK</b>	CA	.04	CA	.06	-1.7300	0.1774	12	-9.75	<.0001	Tukey-Kramer	<.0001
<b>TYPE*THICK</b>	CA	.04	HO	.02	-0.09667	0.4216	12	-0.23	0.8225	Tukey-Kramer	1.0000
<b>TYPE*THICK</b>	CA	.04	HO	.04	-0.2950	0.4216	12	-0.70	0.4975	Tukey-Kramer	0.9978
<b>TYPE*THICK</b>	CA	.04	HO	.06	-0.2950	0.4216	12	-0.70	0.4975	Tukey-Kramer	0.9978
<b>TYPE*THICK</b>	CA	.04	PCO	.02	-0.5317	0.4216	12	-1.26	0.2313	Tukey-Kramer	0.9252
<b>TYPE*THICK</b>	CA	.04	PCO	.04	-0.6433	0.4216	12	-1.53	0.1530	Tukey-Kramer	0.8245
<b>TYPE*THICK</b>	CA	.04	PCO	.06	-0.3533	0.4216	12	-0.84	0.4184	Tukey-Kramer	0.9927
<b>TYPE*THICK</b>	CA	.06	HO	.02	1.6333	0.4216	12	3.87	0.0022	Tukey-Kramer	0.0391
<b>TYPE*THICK</b>	CA	.06	HO	.04	1.4350	0.4216	12	3.40	0.0052	Tukey-Kramer	0.0835
<b>TYPE*THICK</b>	CA	.06	HO	.06	1.4350	0.4216	12	3.40	0.0052	Tukey-Kramer	0.0835
<b>TYPE*THICK</b>	CA	.06	PCO	.02	1.1983	0.4216	12	2.84	0.0148	Tukey-Kramer	0.1974
<b>TYPE*THICK</b>	CA	.06	PCO	.04	1.0867	0.4216	12	2.58	0.0242	Tukey-Kramer	0.2867
<b>TYPE*THICK</b>	CA	.06	PCO	.06	1.3767	0.4216	12	3.27	0.0068	Tukey-Kramer	0.1039
<b>TYPE*THICK</b>	HO	.02	HO	.04	-0.1983	0.1774	12	-1.12	0.2855	Tukey-Kramer	0.9599
<b>TYPE*THICK</b>	HO	.02	HO	.06	-0.1983	0.1774	12	-1.12	0.2855	Tukey-Kramer	0.9599
<b>TYPE*THICK</b>	HO	.02	PCO	.02	-0.4350	0.4216	12	-1.03	0.3225	Tukey-Kramer	0.9743
<b>TYPE*THICK</b>	HO	.02	PCO	.04	-0.5467	0.4216	12	-1.30	0.2191	Tukey-Kramer	0.9144
<b>TYPE*THICK</b>	HO	.02	PCO	.06	-0.2567	0.4216	12	-0.61	0.5540	Tukey-Kramer	0.9992
<b>TYPE*THICK</b>	HO	.04	HO	.06	-866E-17	0.1774	12	-0.00	1.0000	Tukey-Kramer	1.0000
<b>TYPE*THICK</b>	HO	.04	PCO	.02	-0.2367	0.4216	12	-0.56	0.5849	Tukey-Kramer	0.9995
<b>TYPE*THICK</b>	HO	.04	PCO	.04	-0.3483	0.4216	12	-0.83	0.4248	Tukey-Kramer	0.9934
<b>TYPE*THICK</b>	HO	.04	PCO	.06	-0.05833	0.4216	12	-0.14	0.8923	Tukey-Kramer	1.0000
<b>TYPE*THICK</b>	HO	.06	PCO	.02	-0.2367	0.4216	12	-0.56	0.5849	Tukey-Kramer	0.9995
<b>TYPE*THICK</b>	HO	.06	PCO	.04	-0.3483	0.4216	12	-0.83	0.4248	Tukey-Kramer	0.9934
<b>TYPE*THICK</b>	HO	.06	PCO	.06	-0.05833	0.4216	12	-0.14	0.8923	Tukey-Kramer	1.0000
<b>TYPE*THICK</b>	PCO	.02	PCO	.04	-0.1117	0.1774	12	-0.63	0.5409	Tukey-Kramer	0.9989
<b>TYPE*THICK</b>	PCO	.02	PCO	.06	0.1783	0.1774	12	1.01	0.3347	Tukey-Kramer	0.9779
<b>TYPE*THICK</b>	PCO	.04	PCO	.06	0.2900	0.1774	12	1.63	0.1281	Tukey-Kramer	0.7715

---

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

Class Level Information		
Class	Levels	Values
TRT	9	CA.02 CA.04 CA.06 HO.02 HO.04 HO.06 PCO.02 PCO.04 PCO.06
Number of Observations Read		54
Number of Observations Used		54

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

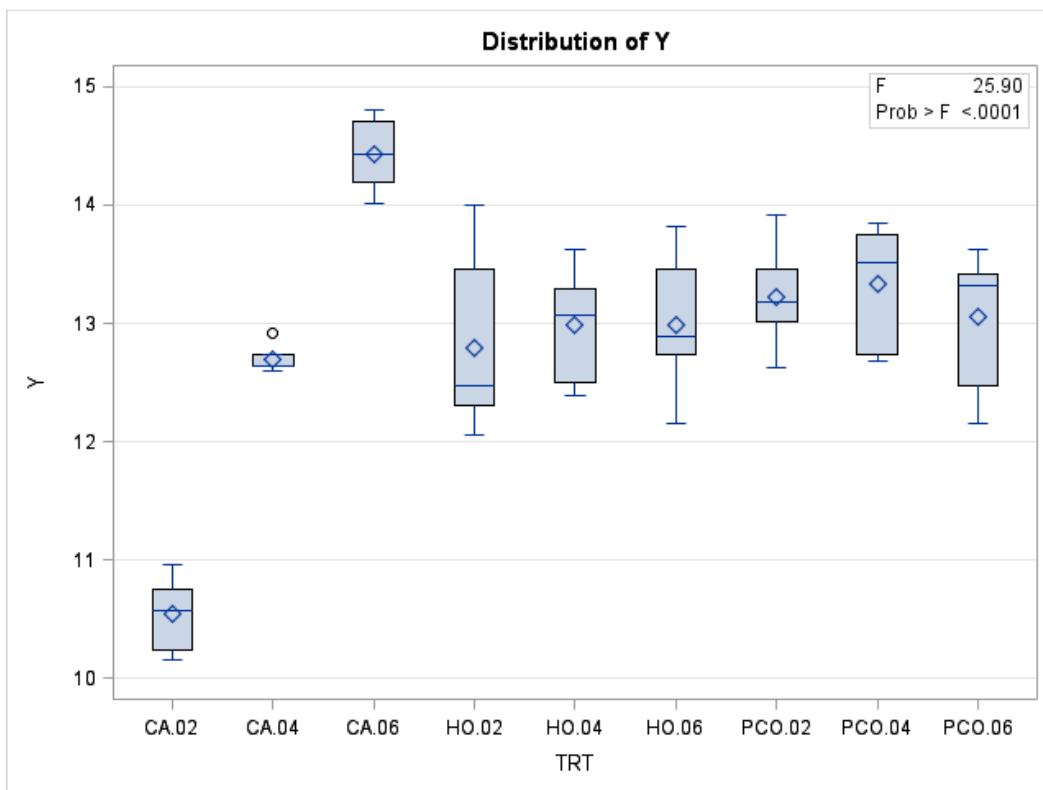
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	49.71420000	6.21427500	25.90	<.0001
Error	45	10.79873333	0.23997185		
Corrected Total	53	60.51293333			

R-Square	Coeff Var	Root MSE	Y Mean
0.821547	3.798745	0.489869	12.89556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRT	8	49.71420000	6.21427500	25.90	<.0001

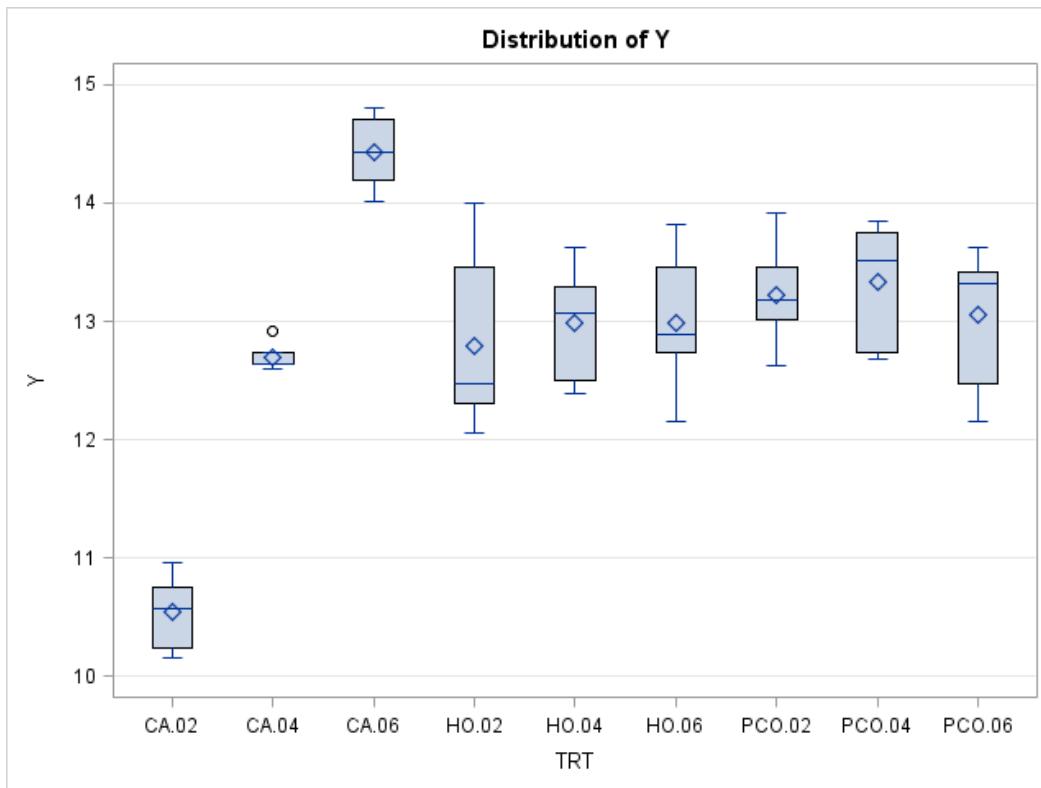
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRT	8	49.71420000	6.21427500	25.90	<.0001



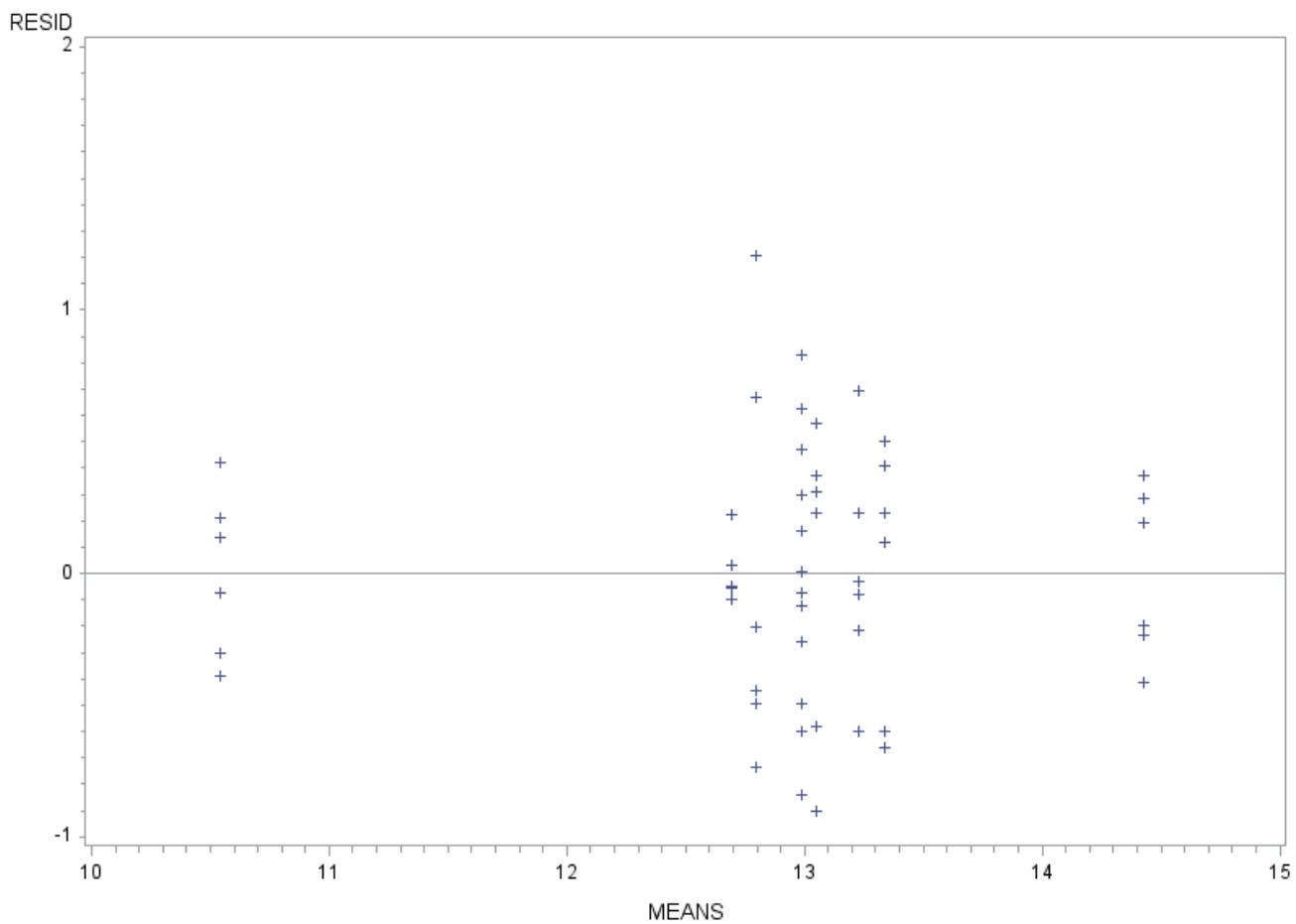
---

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

<b>Brown and Forsythe's Test for Homogeneity of Y Variance ANOVA of Absolute Deviations from Group Medians</b>					
<b>Source</b>	<b>DF</b>	<b>Sum of Squares</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
TRT	8	0.8672	0.1084	0.94	0.4948
Error	45	5.1969	0.1155		

**SAS OUTPUT FOR PROBLEM IV****The GLM Procedure**

Level of TRT	N	Y	
		Mean	Std Dev
CA.02	6	10.5416667	0.31211643
CA.04	6	12.6966667	0.11741664
CA.06	6	14.4266667	0.32413989
HO.02	6	12.7933333	0.76413786
HO.04	6	12.9916667	0.47173792
HO.06	6	12.9916667	0.58348665
PCO.02	6	13.2283333	0.43466846
PCO.04	6	13.3400000	0.50616203
PCO.06	6	13.0500000	0.59282375

**SAS OUTPUT FOR PROBLEM IV**

**SAS OUTPUT FOR PROBLEM IV**

The UNIVARIATE Procedure  
Variable: RESID

Moments			
N	54	Sum Weights	54
Mean	0	Sum Observations	0
Std Deviation	0.4513864	Variance	0.20374969
Skewness	0.13033828	Kurtosis	-0.1764522
Uncorrected SS	10.7987333	Corrected SS	10.7987333
Coeff Variation	.	Std Error Mean	0.06142591

Basic Statistical Measures				
Location		Variability		
Mean	0.00000	Std Deviation	0.45139	
Median	-0.03750	Variance	0.20375	
Mode	-0.05667	Range	2.10667	
		Interquartile Range	0.60000	

Note: The mode displayed is the smallest of 2 modes with a count of 2.

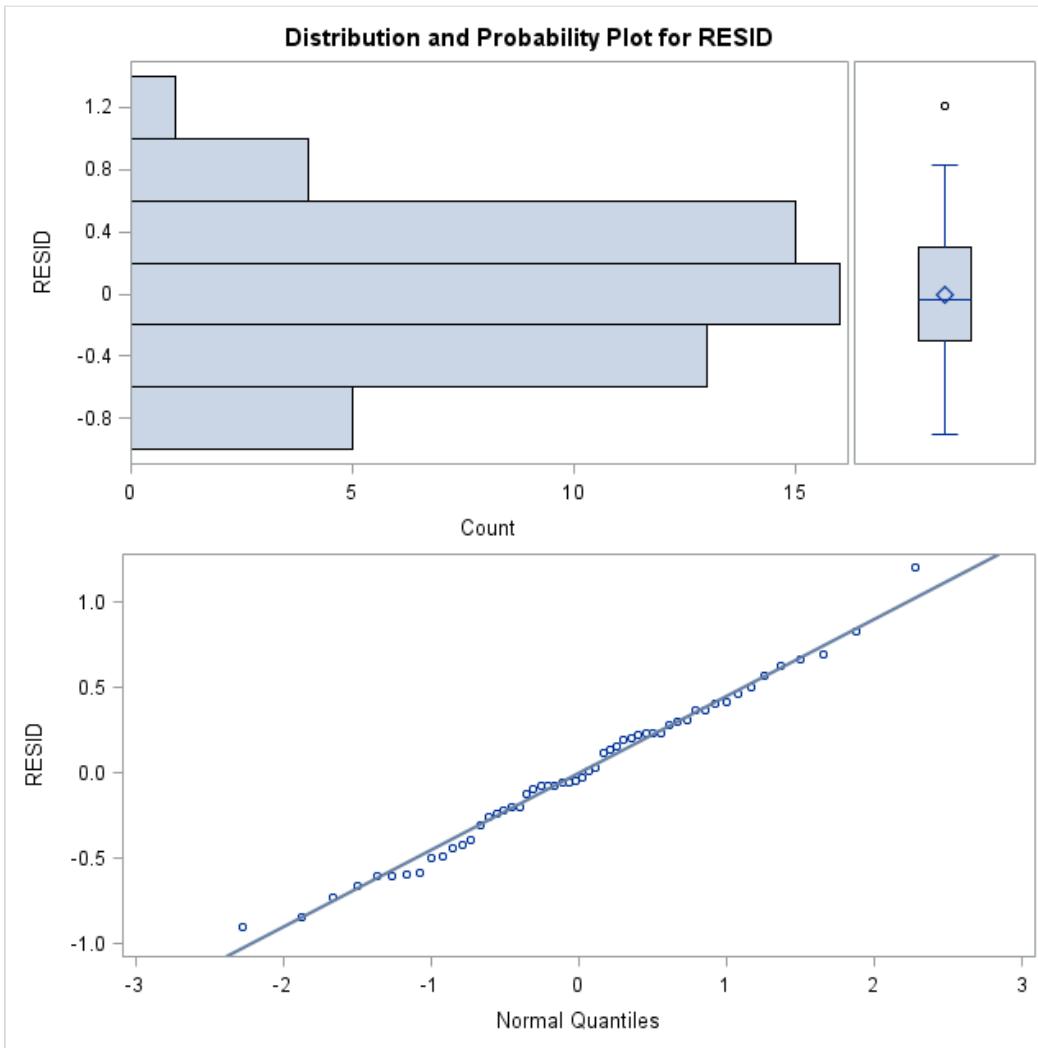
Tests for Location: Mu0=0				
Test	Statistic	p Value		
Student's t	t	0	Pr >  t	1.0000
Sign	M	-1	Pr >=  M	0.8919
Signed Rank	S	-4.5	Pr >=  S	0.9695

Tests for Normality				
Test	Statistic	p Value		
Shapiro-Wilk	W	0.98865	Pr < W	0.8877
Kolmogorov-Smirnov	D	0.054676	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.028584	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.190357	Pr > A-Sq	>0.2500

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1.206667
99%	1.206667
95%	0.691667
90%	0.570000
75% Q3	0.298333
50% Median	-0.037500
25% Q1	-0.301667
10%	-0.600000

5%	-0.733333
1%	-0.900000
0% Min	-0.900000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-0.900000	34	0.628333	11
-0.841667	18	0.666667	2
-0.733333	6	0.691667	23
-0.660000	28	0.828333	17
-0.601667	12	1.206667	5



## MASTER'S DIAGNOSTIC EXAMINATION - August 15, 2013

Student's Name \_\_\_\_\_

### **INSTRUCTIONS FOR STUDENTS:**

1. The exam is to be started at 1 pm (CDT) and completed by 5 pm (CDT) on August 15 2013.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the  
UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
8. 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.
9. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed
- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

### **INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or  
**Scan** the solutions into a **single** pdf file and **email to longneck@stat.tamu.edu**

**Do not** send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or  
emailed to **longneck@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

## **QUESTION I.**

Sleep-disordered breathing is common among adults. To estimate the prevalence of this disorder, a questionnaire concerning sleep habits was mailed to 3514 individuals from thirty to sixty years of age who worked for three large state agencies in Wisconsin. Subjects were classified as habitual snorers if they reported snoring, snorting, or breathing pauses every night or almost every night.

Use the attached SAS output in answering the following questions.

1. Are either of the two potential explanatory variables, gender and agegroup associated with being a snorer? Explain your reasoning.
2. Carry out a test that the odds ratio for being a snorer between women and men differs for the three age groups.
3. Carry out a test for conditional dependence of snoring and gender, controlling for age group.
4. Carry out a test for dependence of snoring and gender, ignoring the effect of age group.
5. Discuss which test is more appropriate, the one in part (3.) or the one in part (4.).
6. Construct and interpret a confidence interval for an assumed common odds ratio. Is it appropriate to use for these data.
7. What is Simpson's paradox? Does Simpson's paradox occur for this data set? Explain.

## SAS OUTPUT FOR QUESTION I

The FREQ Procedure

Table of agegroup by gender

agegroup	gender		
Frequency	female	male	Total
30-39	799	536	1335
40-49	709	696	1405
50-59	336	438	774
Total	1844	1670	3514

## Statistics for Table of agegroup by gender

Statistic	DF	Value	Prob
Chi-Square	2	56.8978	<.0001
Likelihood Ratio Chi-Square	2	57.1344	<.0001
Sample Size = 3514			

---

The FREQ Procedure

Table of agegroup by snorer

agegroup	snorer		
Frequency	yes	no	Total
30-39	384	951	1335
40-49	536	869	1405
50-59	335	439	774
Total	1255	2259	3514

## Statistics for Table of agegroup by snorer

Statistic	DF	Value	Prob
Chi-Square	2	51.0224	<.0001
Likelihood Ratio Chi-Square	2	51.4389	<.0001
Sample Size = 3514			

The FREQ Procedure  
Table of gender by snorer

gender	snorer		
Frequency	yes	no	Total
female	522	1322	1844
male	733	937	1670
Total	1255	2259	3514

Statistics for Table of gender by snorer

Statistic	DF	Value	Prob
Chi-Square	1	92.7017	<.0001
Likelihood Ratio Chi-Square	1	92.9613	<.0001

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	0.5047	0.4388 0.5806
Cohort (Col1 Risk)	0.6449	0.5891 0.7061
Cohort (Col2 Risk)	1.2778	1.2140 1.3449

---

Table 1 of gender by snorer  
Controlling for agegroup=30-39

gender	snorer		
Frequency	yes	no	Total
female	196	603	799
male	188	348	536
Total	384	951	1335

Statistics for Table 1 of gender by snorer  
Controlling for agegroup=30-39

Statistic	DF	Value	Prob
Chi-Square	1	17.4056	<.0001
Likelihood Ratio Chi-Square	1	17.2335	<.0001
Continuity Adj. Chi-Square	1	16.8948	<.0001

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
Case-Control (Odds Ratio)	0.6017	0.4734 0.7646
Cohort (Col1 Risk)	0.6994	0.5915 0.8269
Cohort (Col2 Risk)	1.1624	1.0798 1.2513
Sample Size = 1335		

Table 2 of gender by snorer  
Controlling for agegroup=40-49

gender	snorer		
Frequency	yes	no	Total
female	223	486	709
male	313	383	696
Total	536	869	1405

Statistics for Table 2 of gender by snorer  
Controlling for agegroup=40-49

Statistic	DF	Value	Prob
Chi-Square	1	27.2023	<.0001
Likelihood Ratio Chi-Square	1	27.3005	<.0001

#### Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.5615	0.4516	0.6981
Cohort (Col1 Risk)	0.6994	0.6103	0.8015
Cohort (Col2 Risk)	1.2457	1.1457	1.3543
Sample Size = 1405			

Table 3 of gender by snorer  
Controlling for agegroup=50-59

gender	snorer		
Frequency	yes	no	Total
female	103	233	336
male	232	206	438
Total	335	439	774

Statistics for Table 3 of gender by snorer  
Controlling for agegroup=50-59

Statistic	DF	Value	Prob
Chi-Square	1	38.5631	<.0001
Likelihood Ratio Chi-Square	1	39.1621	<.0001

#### Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.3925	0.2913	0.529
Cohort (Col1 Risk)	0.5787	0.4817	0.6953
Cohort (Col2 Risk)	1.4744	1.3048	1.6661
Sample Size = 774			

Summary Statistics for gender by snorer  
 Controlling for agegroup

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	78.3284	<.0001
2	Row Mean Scores Differ	1	78.3284	<.0001
3	General Association	1	78.3284	<.0001

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	0.5296	0.4598	0.6100
	Logit	0.5305	0.4604	0.6114
Cohort (Col1 Risk)	Mantel-Haenszel	0.6667	0.6083	0.7306
	Logit	0.6672	0.6088	0.7311
Cohort (Col2 Risk)	Mantel-Haenszel	1.2520	1.1901	1.3171
	Logit	1.2411	1.1801	1.3053

Breslow-Day Test for  
 Homogeneity of the Odds Ratios

Chi-Square	5.2478
DF	2
Pr > ChiSq	0.0725

Total Sample Size = 3514

## QUESTION II.

1. Suppose we have  $n$  independent observations of the pair  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , where  $y_i$  is binary, taking only the values 0 or 1, and  $x_i$  is continuous. Provide two reasons why linear regression of  $y$  on  $x$ , with the usual linear regression assumptions, would not be appropriate as an analysis method.
2. In a study of a particular disease in humans, researchers observed the following data.

	Female	Male	Total
Diseased	20	19	39
Healthy	20	41	61
Total	40	60	100

Let  $y_i$  be disease status (1 if diseased, 0 if healthy) and  $x_i$  the gender (1 if female, 0 if male) of the  $i$ th individual,  $i = 1, 2, \dots, 100$ . Let  $\pi(x)$  be the conditional probability of an individual being diseased, given its gender. Consider the logistic regression model:

$$\log \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \beta_0 + \beta_1 x_i.$$

- Write down the likelihood function  $L(\boldsymbol{\beta})$ , where  $\boldsymbol{\beta}' = (\beta_0, \beta_1)$ .
- Report the maximum likelihood estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .
- One way to test  $H_0 : \beta_1 = 0$  is with a likelihood ratio test. The test statistic equals  $-2 \log \left[ \frac{L_0(\hat{\beta}_0^0)}{L(\hat{\boldsymbol{\beta}})} \right]$ , where  $L_0$  is the likelihood function for the null model, the model with only the intercept,  $\beta_0^0$ , included:

$$\log \left( \frac{\pi_0(x_i)}{1 - \pi_0(x_i)} \right) = \beta_0^0,$$

and where  $\hat{\beta}_0^0$  and  $\hat{\boldsymbol{\beta}}$  are the MLEs under the null and unrestricted models, respectively. Under  $H_0$ , the statistic is approximately chi-square distributed with 1 degree of freedom. Test  $H_0 : \beta_1 = 0$ . Some chi-square cumulative probabilities are provided.

3. A similar study to the one described above was conducted, in which the independent variable was race:

	White	Black	Hispanic	Other	Total
Diseased	5	20	15	10	50
Healthy	20	10	10	10	50
Total	25	30	25	20	100

Consider the logistic regression model

$$\log \left( \frac{\pi(r_i)}{1 - \pi(r_i)} \right) = \beta_0 + \beta_1 r_{1i} + \beta_2 r_{2i} + \beta_3 r_{3i},$$

where the  $r_1, r_2, r_3$  independent variables code for race as follows:

	$r_1$	$r_2$	$r_3$
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1

Here is partial output from the model fit using R code:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.3863	0.5000	-2.773	0.00556 **
r1	2.0794	0.6325	3.288	0.00101 **
r2	1.7918	0.6455	2.776	0.00551 **
r3	1.3863	0.6708	2.067	0.03878 *

- (a.) According to the model, what is the estimated probability of a white individual being diseased? Show your calculations, in terms of the model coefficients.
- (b.) Report a 95% confidence interval for the odds ratio (not *log* odds ratio) comparing Hispanic to White.

### QUESTION III.

An analyst for a grocery store chain was interested in the effect of product placement on shelves (Knee, Waist, and Eye levels) and facings (or the amount of shelf space required by the products: Half or Full) on sales, the number of products sold. Three products were placed at each shelf / facing combination, for a total of 18 products. The first model fit to the data was:

$$Sales = \alpha_i + \beta_j + e_{ij}; \quad i = 1, 2, 3; \quad j = 1, 2 \quad (\text{Model 1})$$

1. Interpret the parameter estimate  $\hat{\beta}_1$  from Model 1 in context.
2. If the variable Sales had been an indicator variable for whether or not a product had sold rather than the number of each product sold, would the model above have been a valid model? Why or why not?

It was also suggested that the sugar content of the products may play a part in the number sold. The second fitted model (this time with an intercept) was:

$$Sales = \beta_0 + \beta_1 Sugars + \beta_2 i + \beta_3 j; \quad i = 1, 2, 3; \quad j = 1, 2 \quad (\text{Model 2})$$

Output from this second model is found below and on the following page.

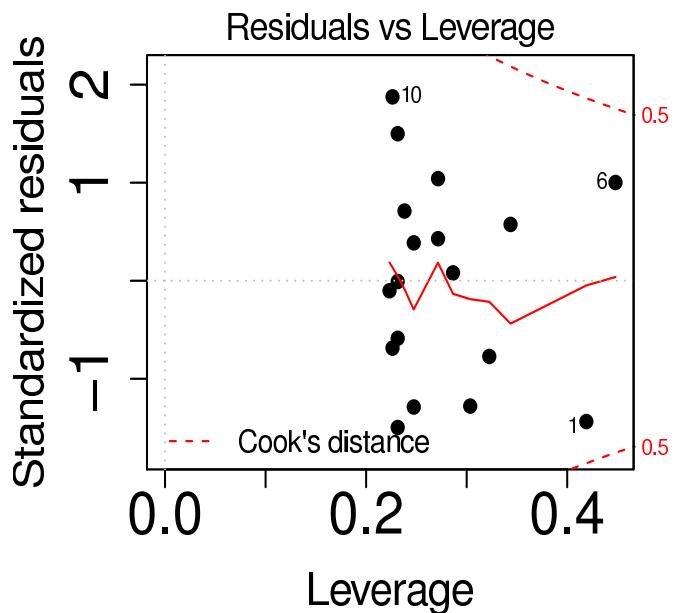
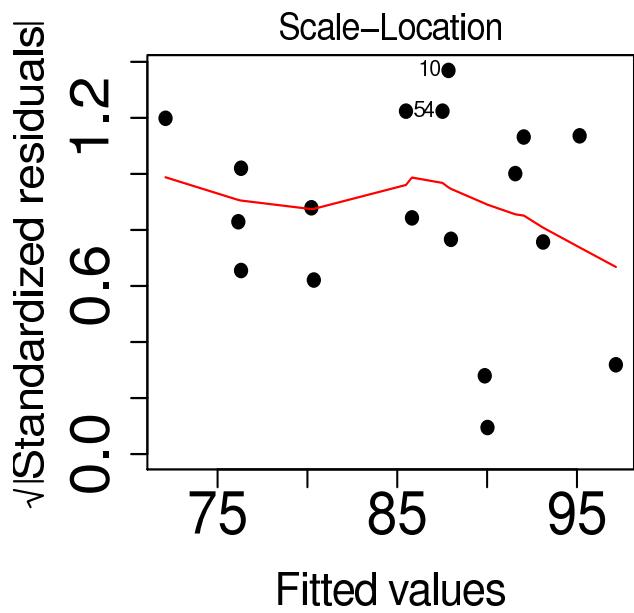
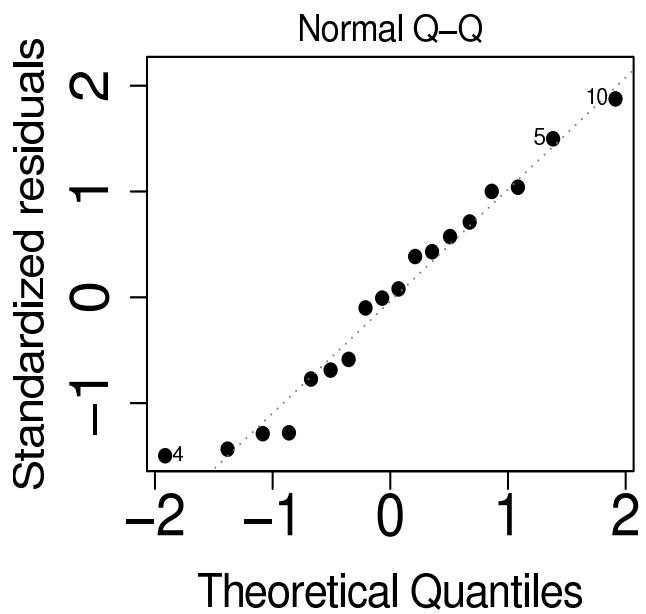
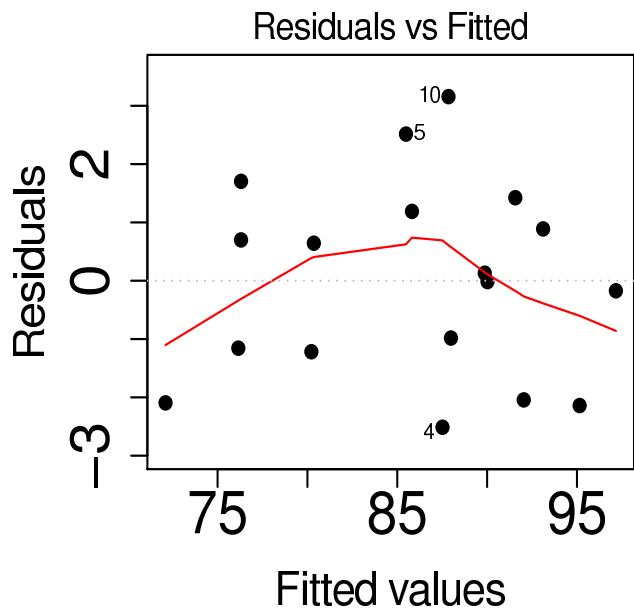
#### Model 2:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	73.774	2.875	25.661	1.60e-12 ***
FacingHalf	-7.627	1.025	-7.443	4.88e-06 ***
HeightKnee	-2.172	1.107	-1.962	0.0715
HeightWaist	3.096	1.394	2.222	0.0447 *
Sugars	2.030	0.364	5.578	8.95e-05 ***

3. Is Model 2 a valid model? Discuss why or why not.
4. Regardless of your answer above, use Model 2 to test whether Sugars have an effect on Sales, after controlling for shelf placement and facing. Be sure to include the hypotheses used and your conclusion in the context of the problem.
5. Suppose that a third model is fit, and an interaction between shelf placement and facing is found to be significant. How would you explain the interaction to others in the company without using statistical jargon?

## Model 2: Graphical Displays

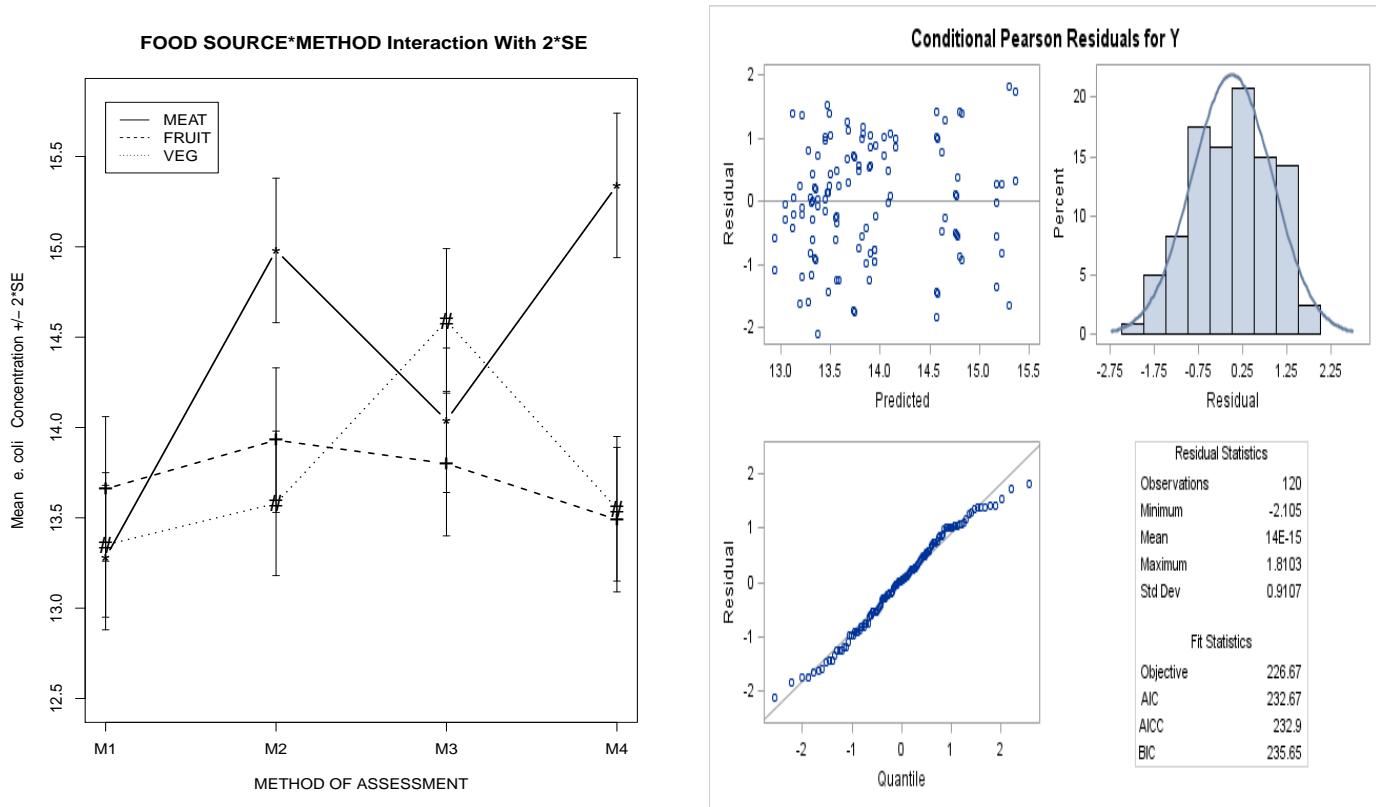


#### QUESTION IV.

The Center for Disease Control (CDC) conducted an experiment to evaluate the reliability of assessing the level of contamination of *e. coli* in three food sources, Meat, Fruit, and Vegetables. There are four unique methods for assessing *e. coli* : M1, M2, M3, M4 and hundreds of laboratories which use one or more of these methods in the USA. For each of the methods of assessment, five laboratories are randomly selected to participate in the study. Forty containers are prepared for each food source by spiking the container with a known level of contamination of *e. coli* and then placing the container in a controlled climate for three weeks to allow the *e. coli* level to stabilize. Six containers, two of each of the three food sources, are then sent to each of the 20 laboratories selected for the study. The *e. coli* level (cfu/g),  $Y_{ijkl}$ , determined by the  $k$ th Lab using Assessment Method  $j$  for the  $\ell$ th Container of Food Source  $i$  is recorded for the 120 containers. The CDC wants to compare the mean *e. coli* level of the four assessment methods and their differences across the food sources. Also, CDC wants to determine if there are major differences in the mean *e. coli* determinations across the many laboratories in the USA.

Source	Assessment Methods																			
	M1					M2					M3					M4				
	Lab		Lab		Lab		Lab		Lab		Lab		Lab		Lab		Lab		Lab	
Source	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	L16	L17	L18	L19	L20
Meat	12.3 12.6	13.2 13.0	12.9 13.0	13.2 14.0	12.9 13.9	14.5 15.0	14.3 15.6	14.5 14.8	14.3 15.6	14.5 14.8	14.4 14.1	13.5 13.4	14.7 14.6	13.5 13.4	14.7 14.2	14.8 15.3	16.4 14.3	15.6 16.4	14.8 15.4	15.2 14.4
Fruit	13.2 13.4	14.4 14.5	12.9 13.7	14.1 14.1	12.8 13.4	13.2 14.2	14.2 13.4	14.4 14.6	13.3 13.6	14.5 13.8	13.4 14.1	14.5 14.4	12.7 14.2	13.5 14.4	12.7 14.2	12.2 13.3	14.4 13.6	12.4 13.8	13.4 13.8	13.2 13.3
Veg	13.1 12.5	13.4 14.0	13.6 13.0	13.8 14.1	12.8 13.3	13.5 14.0	13.3 12.6	13.5 12.8	14.3 13.6	13.5 12.8	14.3 15.1	13.5 15.4	13.7 15.2	14.5 15.4	13.7 15.2	12.2 13.3	14.3 13.9	13.6 12.7	13.4 13.9	14.4 14.1

Use the following plots, the attached tables, and SAS output to answer the following questions.



1. Write a model which displays an appropriate relationship between level of contamination,  $Y_{ijkl}$ , and its possible sources of variation. Include any restrictions on the parameters in your model and any distributional properties on the random variables in your model.
2. Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.

$C_1$  Normality:

$C_2$  Equal Variance:

$C_3$  Independence:

3. Construct a partial ANOVA table for this experiment.

Include only the following: Source of Variation, df, and Expected Mean Squares.

4. At the  $\alpha = .05$  level, which Main effects and Interaction effects are significant? Justify your answer by including the relevant p-values along with their pair of degrees of freedom ( $df_{NUM.}, df_{DEN.}$ ). Test for Main effects even if an interaction is significant. Also, provide a test for all random effects.
5. Separate the four Assessment Methods into groups such that all Assessment Methods in a group are not significantly different from any other member of the group with respect to their mean *e. coli* level. Use an experimentwise error rate of  $\alpha = .05$ .
6. Provide justification for the following standard errors listed in the PROC MIXED output using the following estimates of the variance components:

$$\sigma_{LAB(METH)}^2 = .008708 \quad \sigma_{FOOD*LAB(METH)}^2 = .03236 \quad \sigma_{RESIDUAL}^2 = .3357$$

- a. The estimated standard error of the estimated mean *e. coli* level of a container of Meat which was measured by Method M1 is .2044
  - b. The estimated standard error of the estimated difference between the mean *e. coli* level of a container of Meat and a container of Fruit is .1415
7. Provide a 95% confidence on the mean *e. coli* level of a container of Meat which was measured by Method M1.

SAS Code:

```
ods html;ods graphics on;

OPTIONS LS=90 PS=55 nocenter nodate;
TITLE 'SAS OUTPUT FOR PROBLEM IV';
DATA ECOLI;
INPUT FOOD $ METH $ LAB $ Y @@;
TRT=COMPRESS (METH) || COMPRESS (FOOD);
CARDS;

MEAT M1 L1 12.30 MEAT M2 L6 14.46 MEAT M3 L11 14.35 MEAT M4 L16 14.85
MEAT M1 L1 12.59 MEAT M2 L6 15.00 MEAT M3 L11 14.06 MEAT M4 L16 15.32
MEAT M1 L2 13.15 MEAT M2 L7 14.29 MEAT M3 L12 13.50 MEAT M4 L17 16.35
MEAT M1 L2 13.00 MEAT M2 L7 15.62 MEAT M3 L12 13.39 MEAT M4 L17 14.34
MEAT M1 L3 12.87 MEAT M2 L8 14.46 MEAT M3 L13 14.73 MEAT M4 L18 15.55
MEAT M1 L3 13.02 MEAT M2 L8 14.82 MEAT M3 L13 14.65 MEAT M4 L18 16.36
MEAT M1 L4 13.15 MEAT M2 L9 14.29 MEAT M3 L14 13.50 MEAT M4 L19 14.75
MEAT M1 L4 14.00 MEAT M2 L9 15.62 MEAT M3 L14 13.39 MEAT M4 L19 15.38
MEAT M1 L5 12.87 MEAT M2 L10 14.46 MEAT M3 L15 14.73 MEAT M4 L20 15.15
MEAT M1 L5 13.92 MEAT M2 L10 14.82 MEAT M3 L15 14.15 MEAT M4 L20 14.39
FRUIT M1 L1 13.20 FRUIT M2 L6 13.16 FRUIT M3 L11 13.35 FRUIT M4 L16 12.25
FRUIT M1 L1 13.39 FRUIT M2 L6 14.20 FRUIT M3 L11 14.06 FRUIT M4 L16 13.33
FRUIT M1 L2 14.45 FRUIT M2 L7 14.23 FRUIT M3 L12 14.50 FRUIT M4 L17 14.35
FRUIT M1 L2 14.50 FRUIT M2 L7 13.42 FRUIT M3 L12 14.39 FRUIT M4 L17 13.55
FRUIT M1 L3 12.86 FRUIT M2 L8 14.45 FRUIT M3 L13 12.73 FRUIT M4 L18 12.36
FRUIT M1 L3 13.72 FRUIT M2 L8 14.62 FRUIT M3 L13 14.15 FRUIT M4 L18 13.75
FRUIT M1 L4 14.11 FRUIT M2 L9 13.29 FRUIT M3 L14 13.50 FRUIT M4 L19 13.38
FRUIT M1 L4 14.10 FRUIT M2 L9 13.62 FRUIT M3 L14 14.39 FRUIT M4 L19 13.79
FRUIT M1 L5 12.83 FRUIT M2 L10 14.46 FRUIT M3 L15 12.73 FRUIT M4 L20 13.15
FRUIT M1 L5 13.42 FRUIT M2 L10 13.82 FRUIT M3 L15 14.15 FRUIT M4 L20 13.32
VEG M1 L1 13.10 VEG M2 L6 13.46 VEG M3 L11 14.35 VEG M4 L16 12.15
VEG M1 L1 12.52 VEG M2 L6 14.00 VEG M3 L11 15.06 VEG M4 L16 13.32
VEG M1 L2 13.35 VEG M2 L7 13.29 VEG M3 L12 13.50 VEG M4 L17 14.32
VEG M1 L2 14.04 VEG M2 L7 12.62 VEG M3 L12 15.39 VEG M4 L17 13.85
VEG M1 L3 13.57 VEG M2 L8 13.46 VEG M3 L13 13.73 VEG M4 L18 13.55
VEG M1 L3 12.96 VEG M2 L8 12.82 VEG M3 L13 15.15 VEG M4 L18 12.65
VEG M1 L4 13.75 VEG M2 L9 14.29 VEG M3 L14 14.50 VEG M4 L19 13.37
VEG M1 L4 14.10 VEG M2 L9 13.62 VEG M3 L14 15.39 VEG M4 L19 13.85
VEG M1 L5 12.82 VEG M2 L10 13.46 VEG M3 L15 13.73 VEG M4 L20 14.39
VEG M1 L5 13.32 VEG M2 L10 12.82 VEG M3 L15 15.15 VEG M4 L20 14.05

RUN;

PROC GLM ORDER=DATA;
CLASS FOOD METH LAB;
MODEL Y = FOOD METH FOOD*METH LAB(METH) FOOD*LAB(METH);
RANDOM LAB(METH) FOOD*LAB(METH)/TEST;
LSMEANS FOOD METH FOOD*METH/STDERR PDIFF ADJUST=TUKEY;

PROC MIXED ORDER=DATA;
CLASS FOOD METH LAB;
MODEL Y = FOOD METH FOOD*METH/RESIDUAL ;
RANDOM LAB(METH) FOOD*LAB(METH);
LSMEANS FOOD METH FOOD*METH/ADJUST=TUKEY;

ods graphics off;ods html close;
```

## SAS OUTPUT FOR PROBLEM IV

The GLM Procedure

### Class Level Information

Class Levels Values

**FOOD** 3 MEAT FRUIT VEG

**METH** 4 M1 M2 M3 M4

**LAB** 20 L1 L6 L11 L16 L2 L7 L12 L17 L3 L8 L13 L18 L4 L9 L14 L19 L5 L10 L15 L20

**Number of Observations Read** 120

**Number of Observations Used** 120

---

**SAS OUTPUT FOR PROBLEM IV**

---

The GLM Procedure

Dependent Variable: Y

<b>Source</b>	<b>DF</b>	<b>Sum of Squares</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>Model</b>	59	68.49452917	1.16092422	3.46	<.0001
<b>Error</b>	60	20.14305000	0.33571750		
<b>Corrected Total</b>	119	88.63757917			

<b>R-Square</b>	<b>Coeff Var</b>	<b>Root MSE</b>	<b>Y Mean</b>
0.772748	4.169802	0.579411	13.89542

<b>Source</b>	<b>DF</b>	<b>Type I SS</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>FOOD</b>	2	9.38181167	4.69090583	13.97	<.0001
<b>METH</b>	3	11.45262250	3.81754083	11.37	<.0001
<b>FOOD*METH</b>	6	27.60327500	4.60054583	13.70	<.0001
<b>LAB(METH)</b>	16	7.24294000	0.45268375	1.35	0.1995
<b>FOOD*LAB(METH)</b>	32	12.81388000	0.40043375	1.19	0.2734

<b>Source</b>	<b>DF</b>	<b>Type III SS</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>FOOD</b>	2	9.38181167	4.69090583	13.97	<.0001
<b>METH</b>	3	11.45262250	3.81754083	11.37	<.0001
<b>FOOD*METH</b>	6	27.60327500	4.60054583	13.70	<.0001
<b>LAB(METH)</b>	16	7.24294000	0.45268375	1.35	0.1995
<b>FOOD*LAB(METH)</b>	32	12.81388000	0.40043375	1.19	0.2734

The GLM Procedure

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* FOOD	2	9.381812	4.690906	11.71	0.0002
FOOD*METH	6	27.603275	4.600546	11.49	<.0001
LAB(METH)	16	7.242940	0.452684	1.13	0.3705
Error	32	12.813880	0.400434		

Error: MS(FOOD\*LAB(METH))

\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* METH	3	11.452623	3.817541	8.43	0.0014
Error: MS(LAB(METH))	16	7.242940	0.452684		

\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FOOD*LAB(METH)	32	12.813880	0.400434	1.19	0.2734
Error: MS(Error)	60	20.143050	0.335717		

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

FOOD	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
MEAT	14.2900000	0.0916130	<.0001	1
FRUIT	13.6757500	0.0916130	<.0001	2
VEG	13.7205000	0.0916130	<.0001	3

**Least Squares Means for effect FOOD**

**Pr > |t| for H0: LSMean(i)=LSMean(j)**

**Dependent Variable: Y**

i/j	1	2	3
1		<.0001	0.0001
2	<.0001		0.9364
3	0.0001	0.9364	

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

METH	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
M1	13.3660000	0.1057856	<.0001	1
M2	14.0316667	0.1057856	<.0001	2
M3	14.1450000	0.1057856	<.0001	3
M4	14.0390000	0.1057856	<.0001	4

**Least Squares Means for effect METH**

**Pr > |t| for H0: LSMean(i)=LSMean(j)**

**Dependent Variable: Y**

i/j	1	2	3	4
1		0.0002	<.0001	0.0002
2	0.0002		0.8731	1.0000
3	<.0001	0.8731		0.8933
4	0.0002	1.0000	0.8933	

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

FOOD	METH	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
MEAT	M1	13.0870000	0.1832260	<.0001	1
MEAT	M2	14.7840000	0.1832260	<.0001	2
MEAT	M3	14.0450000	0.1832260	<.0001	3
MEAT	M4	15.2440000	0.1832260	<.0001	4
FRUIT	M1	13.6580000	0.1832260	<.0001	5
FRUIT	M2	13.9270000	0.1832260	<.0001	6
FRUIT	M3	13.7950000	0.1832260	<.0001	7
FRUIT	M4	13.3230000	0.1832260	<.0001	8
VEG	M1	13.3530000	0.1832260	<.0001	9
VEG	M2	13.3840000	0.1832260	<.0001	10
VEG	M3	14.5950000	0.1832260	<.0001	11
VEG	M4	13.5500000	0.1832260	<.0001	12

Least Squares Means for effect FOOD*METH												
Pr >  t  for H0: LSMean(i)=LSMean(j)												
Dependent Variable: Y												
i/j	1	2	3	4	5	6	7	8	9	10	11	12
1		<.0001	0.021	<.0001	0.553	0.075	0.236	0.998	0.996	0.991	<.000	0.818
			9		6	3	3	8	4	0	1	5
2	<.0001		0.185	0.824	0.002	0.063	0.015	<.000	<.000	<.000	0.999	0.000
			4	5	9	7	4	1	1	1	8	7
3	0.021	0.185		0.001	0.936	1.000	0.997	0.212	0.266	0.330	0.609	0.748
	9	4		1	7	0	9	2	1	2	3	2
4	<.0001	0.824	0.001		<.000	0.000	<.000	<.000	<.000	<.000	0.357	<.000
		5	1		1	2	1	1	1	1	2	1
5	0.553	0.002	0.936	<.0001		0.996	1.000	0.977	0.988	0.995	0.027	1.000
	6	9	7			1	0	0	8	4	6	0
6	0.075	0.063	1.000	0.000	0.996		1.000	0.467	0.545	0.627	0.315	0.946
	3	7	0	2	1		0	2	7	7	0	7
7	0.236	0.015	0.997	<.0001	1.000	1.000		0.799	0.858	0.907	0.109	0.998
	3	4	9		0	0		8	5	4	6	3
8	0.998	<.0001	0.212	<.0001	0.977	0.467	0.799		1.000	1.000	0.000	0.999
	8		2		0	2	8		0	0	4	1
9	0.996	<.0001	0.266	<.0001	0.988	0.545	0.858	1.000		1.000	0.000	0.999
	4		1		8	7	5	0		0	6	8
10	0.991	<.0001	0.330	<.0001	0.995	0.627	0.907	1.000	1.000		0.001	1.000
	0		2		4	7	4	0	0		0	0
11	<.0001	0.999	0.609	0.357	0.027	0.315	0.109	0.000	0.000	0.001		0.008
		8	3	2	6	0	6	4	6	0		0
12	0.818	0.000	0.748	<.0001	1.000	0.946	0.998	0.999	0.999	1.000	0.008	
	5	7	2		0	7	3	1	8	0	0	0

---

**SAS OUTPUT FOR PROBLEM IV**

---

## The Mixed Procedure

**Model Information**

<b>Data Set</b>	WORK.ECOLI
<b>Dependent Variable</b>	Y
<b>Covariance Structure</b>	Variance Components
<b>Estimation Method</b>	REML
<b>Residual Variance Method</b>	Profile
<b>Fixed Effects SE Method</b>	Model-Based
<b>Degrees of Freedom Method</b>	Satterthwaite

**Class Level Information**

Class	Levels	Values
FOOD	3	MEAT FRUIT VEG
METH	4	M1 M2 M3 M4
LAB	20	L1 L6 L11 L16 L2 L7 L12 L17 L3 L8 L13 L18 L4 L9 L14 L19 L5 L10 L15 L20

**Number of Observations**

<b>Number of Observations Read</b>	120
<b>Number of Observations Used</b>	120
<b>Number of Observations Not Used</b>	0

**Covariance Parameter Estimates**

Cov Parm	Estimate
LAB(METH)	0.008708
FOOD*LAB(METH)	0.03236
Residual	0.3357

**Type 3 Tests of Fixed Effects**

Effect	Num DF	Den DF	F Value	Pr > F
FOOD	2	32	11.71	0.0002
METH	3	16	8.43	0.0014
FOOD*METH	6	32	11.49	<.0001

**Least Squares Means**

Effect	FOOD	METH	Estimate	Standard Error	DF	t Value	Pr >  t
FOOD	MEAT		14.2900	0.1022	47.8	139.81	<.0001
FOOD	FRUIT		13.6758	0.1022	47.8	133.80	<.0001
FOOD	VEG		13.7205	0.1022	47.8	134.24	<.0001
METH		M1	13.3660	0.1228	16	108.81	<.0001
METH		M2	14.0317	0.1228	16	114.23	<.0001
METH		M3	14.1450	0.1228	16	115.15	<.0001
METH		M4	14.0390	0.1228	16	114.29	<.0001
FOOD*METH	MEAT	M1	13.0870	0.2044	47.8	64.02	<.0001
FOOD*METH	MEAT	M2	14.7840	0.2044	47.8	72.32	<.0001
FOOD*METH	MEAT	M3	14.0450	0.2044	47.8	68.71	<.0001
FOOD*METH	MEAT	M4	15.2440	0.2044	47.8	74.57	<.0001
FOOD*METH	FRUIT	M1	13.6580	0.2044	47.8	66.82	<.0001
FOOD*METH	FRUIT	M2	13.9270	0.2044	47.8	68.13	<.0001
FOOD*METH	FRUIT	M3	13.7950	0.2044	47.8	67.49	<.0001
FOOD*METH	FRUIT	M4	13.3230	0.2044	47.8	65.18	<.0001
FOOD*METH	VEG	M1	13.3530	0.2044	47.8	65.32	<.0001
FOOD*METH	VEG	M2	13.3840	0.2044	47.8	65.48	<.0001
FOOD*METH	VEG	M3	14.5950	0.2044	47.8	71.40	<.0001
FOOD*METH	VEG	M4	13.5500	0.2044	47.8	66.29	<.0001

Differences of Least Squares Means												
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P	
FOOD	MEAT		FRUIT		0.6142	0.1415	32	4.34	0.0001	Tukey-Kramer	0.0004	
FOOD	MEAT		VEG		0.5695	0.1415	32	4.02	0.0003	Tukey-Kramer	0.0009	
FOOD	FRUIT		VEG		-0.04475	0.1415	32	-0.32	0.7539	Tukey-Kramer	0.9465	
METH		M1		M2	-0.6657	0.1737	16	-3.83	0.0015	Tukey	0.0072	
METH		M1		M3	-0.7790	0.1737	16	-4.48	0.0004	Tukey	0.0019	
METH		M1		M4	-0.6730	0.1737	16	-3.87	0.0013	Tukey	0.0066	
METH		M2		M3	-0.1133	0.1737	16	-0.65	0.5234	Tukey	0.9132	
METH		M2		M4	-0.00733	0.1737	16	-0.04	0.9669	Tukey	1.0000	
METH		M3		M4	0.1060	0.1737	16	0.61	0.5503	Tukey	0.9274	
FOOD*METH	MEAT	M1	MEAT	M2	-1.6970	0.2891	47.8	-5.87	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M1	MEAT	M3	-0.9580	0.2891	47.8	-3.31	0.0018	Tukey-Kramer	0.0799	
FOOD*METH	MEAT	M1	MEAT	M4	-2.1570	0.2891	47.8	-7.46	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M1	FRUIT	M1	-0.5710	0.2830	32	-2.02	0.0521	Tukey-Kramer	0.6790	
FOOD*METH	MEAT	M1	FRUIT	M2	-0.8400	0.2891	47.8	-2.91	0.0055	Tukey-Kramer	0.1864	
FOOD*METH	MEAT	M1	FRUIT	M3	-0.7080	0.2891	47.8	-2.45	0.0180	Tukey-Kramer	0.4047	
FOOD*METH	MEAT	M1	FRUIT	M4	-0.2360	0.2891	47.8	-0.82	0.4183	Tukey-Kramer	0.9994	
FOOD*METH	MEAT	M1	VEG	M1	-0.2660	0.2830	32	-0.94	0.3543	Tukey-Kramer	0.9980	
FOOD*METH	MEAT	M1	VEG	M2	-0.2970	0.2891	47.8	-1.03	0.3094	Tukey-Kramer	0.9957	
FOOD*METH	MEAT	M1	VEG	M3	-1.5080	0.2891	47.8	-5.22	<.0001	Tukey-Kramer	0.0006	
FOOD*METH	MEAT	M1	VEG	M4	-0.4630	0.2891	47.8	-1.60	0.1158	Tukey-Kramer	0.8960	
FOOD*METH	MEAT	M2	MEAT	M3	0.7390	0.2891	47.8	2.56	0.0138	Tukey-Kramer	0.3439	
FOOD*METH	MEAT	M2	MEAT	M4	-0.4600	0.2891	47.8	-1.59	0.1181	Tukey-Kramer	0.8999	
FOOD*METH	MEAT	M2	FRUIT	M1	1.1260	0.2891	47.8	3.90	0.0003	Tukey-Kramer	0.0200	
FOOD*METH	MEAT	M2	FRUIT	M2	0.8570	0.2830	32	3.03	0.0048	Tukey-Kramer	0.1465	
FOOD*METH	MEAT	M2	FRUIT	M3	0.9890	0.2891	47.8	3.42	0.0013	Tukey-Kramer	0.0627	
FOOD*METH	MEAT	M2	FRUIT	M4	1.4610	0.2891	47.8	5.05	<.0001	Tukey-Kramer	0.0009	
FOOD*METH	MEAT	M2	VEG	M1	1.4310	0.2891	47.8	4.95	<.0001	Tukey-Kramer	0.0012	
FOOD*METH	MEAT	M2	VEG	M2	1.4000	0.2830	32	4.95	<.0001	Tukey-Kramer	0.0012	
FOOD*METH	MEAT	M2	VEG	M3	0.1890	0.2891	47.8	0.65	0.5164	Tukey-Kramer	0.9999	
FOOD*METH	MEAT	M2	VEG	M4	1.2340	0.2891	47.8	4.27	<.0001	Tukey-Kramer	0.0076	

Differences of Least Squares Means												
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P	
FOOD*METH	MEAT	M3	MEAT	M4	-1.1990	0.2891	47.8	-4.15	0.0001	Tukey-Kramer	0.0104	
FOOD*METH	MEAT	M3	FRUIT	M1	0.3870	0.2891	47.8	1.34	0.1870	Tukey-Kramer	0.9670	
FOOD*METH	MEAT	M3	FRUIT	M2	0.1180	0.2891	47.8	0.41	0.6850	Tukey-Kramer	1.0000	
FOOD*METH	MEAT	M3	FRUIT	M3	0.2500	0.2830	32	0.88	0.3836	Tukey-Kramer	0.9989	
FOOD*METH	MEAT	M3	FRUIT	M4	0.7220	0.2891	47.8	2.50	0.0160	Tukey-Kramer	0.3766	
FOOD*METH	MEAT	M3	VEG	M1	0.6920	0.2891	47.8	2.39	0.0206	Tukey-Kramer	0.4379	
FOOD*METH	MEAT	M3	VEG	M2	0.6610	0.2891	47.8	2.29	0.0267	Tukey-Kramer	0.5052	
FOOD*METH	MEAT	M3	VEG	M3	-0.5500	0.2830	32	-1.94	0.0608	Tukey-Kramer	0.7249	
FOOD*METH	MEAT	M3	VEG	M4	0.4950	0.2891	47.8	1.71	0.0933	Tukey-Kramer	0.8496	
FOOD*METH	MEAT	M4	FRUIT	M1	1.5860	0.2891	47.8	5.49	<.0001	Tukey-Kramer	0.0003	
FOOD*METH	MEAT	M4	FRUIT	M2	1.3170	0.2891	47.8	4.56	<.0001	Tukey-Kramer	0.0035	
FOOD*METH	MEAT	M4	FRUIT	M3	1.4490	0.2891	47.8	5.01	<.0001	Tukey-Kramer	0.0010	
FOOD*METH	MEAT	M4	FRUIT	M4	1.9210	0.2830	32	6.79	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M1	1.8910	0.2891	47.8	6.54	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M2	1.8600	0.2891	47.8	6.43	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M3	0.6490	0.2891	47.8	2.25	0.0294	Tukey-Kramer	0.5319	
FOOD*METH	MEAT	M4	VEG	M4	1.6940	0.2830	32	5.99	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	FRUIT	M1	FRUIT	M2	-0.2690	0.2891	47.8	-0.93	0.3568	Tukey-Kramer	0.9982	
FOOD*METH	FRUIT	M1	FRUIT	M3	-0.1370	0.2891	47.8	-0.47	0.6377	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M1	FRUIT	M4	0.3350	0.2891	47.8	1.16	0.2523	Tukey-Kramer	0.9887	
FOOD*METH	FRUIT	M1	VEG	M1	0.3050	0.2830	32	1.08	0.2892	Tukey-Kramer	0.9937	
FOOD*METH	FRUIT	M1	VEG	M2	0.2740	0.2891	47.8	0.95	0.3480	Tukey-Kramer	0.9979	
FOOD*METH	FRUIT	M1	VEG	M3	-0.9370	0.2891	47.8	-3.24	0.0022	Tukey-Kramer	0.0937	
FOOD*METH	FRUIT	M1	VEG	M4	0.1080	0.2891	47.8	0.37	0.7104	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M2	FRUIT	M3	0.1320	0.2891	47.8	0.46	0.6500	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M2	FRUIT	M4	0.6040	0.2891	47.8	2.09	0.0420	Tukey-Kramer	0.6332	
FOOD*METH	FRUIT	M2	VEG	M1	0.5740	0.2891	47.8	1.99	0.0528	Tukey-Kramer	0.6991	
FOOD*METH	FRUIT	M2	VEG	M2	0.5430	0.2830	32	1.92	0.0640	Tukey-Kramer	0.7397	
FOOD*METH	FRUIT	M2	VEG	M3	-0.6680	0.2891	47.8	-2.31	0.0252	Tukey-Kramer	0.4897	
FOOD*METH	FRUIT	M2	VEG	M4	0.3770	0.2891	47.8	1.30	0.1984	Tukey-Kramer	0.9725	

Differences of Least Squares Means													
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P		
FOOD*METH	FRUIT	M3	FRUIT	M4	0.4720	0.2891	47.8	1.63	0.1091	Tukey-Kramer	0.8840		
FOOD*METH	FRUIT	M3	VEG	M1	0.4420	0.2891	47.8	1.53	0.1329	Tukey-Kramer	0.9211		
FOOD*METH	FRUIT	M3	VEG	M2	0.4110	0.2891	47.8	1.42	0.1616	Tukey-Kramer	0.9503		
FOOD*METH	FRUIT	M3	VEG	M3	-0.8000	0.2830	32	-2.83	0.0080	Tukey-Kramer	0.2162		
FOOD*METH	FRUIT	M3	VEG	M4	0.2450	0.2891	47.8	0.85	0.4009	Tukey-Kramer	0.9992		
FOOD*METH	FRUIT	M4	VEG	M1	-0.03000	0.2891	47.8	-0.10	0.9178	Tukey-Kramer	1.0000		
FOOD*METH	FRUIT	M4	VEG	M2	-0.06100	0.2891	47.8	-0.21	0.8338	Tukey-Kramer	1.0000		
FOOD*METH	FRUIT	M4	VEG	M3	-1.2720	0.2891	47.8	-4.40	<.0001	Tukey-Kramer	0.0054		
FOOD*METH	FRUIT	M4	VEG	M4	-0.2270	0.2830	32	-0.80	0.4284	Tukey-Kramer	0.9995		
FOOD*METH	VEG	M1	VEG	M2	-0.03100	0.2891	47.8	-0.11	0.9151	Tukey-Kramer	1.0000		
FOOD*METH	VEG	M1	VEG	M3	-1.2420	0.2891	47.8	-4.30	<.0001	Tukey-Kramer	0.0071		
FOOD*METH	VEG	M1	VEG	M4	-0.1970	0.2891	47.8	-0.68	0.4989	Tukey-Kramer	0.9999		
FOOD*METH	VEG	M2	VEG	M3	-1.2110	0.2891	47.8	-4.19	0.0001	Tukey-Kramer	0.0094		
FOOD*METH	VEG	M2	VEG	M4	-0.1660	0.2891	47.8	-0.57	0.5685	Tukey-Kramer	1.0000		
FOOD*METH	VEG	M3	VEG	M4	1.0450	0.2891	47.8	3.61	0.0007	Tukey-Kramer	0.0398		

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)

Standard normal density function  
Shaded area =  $\Phi(z)$

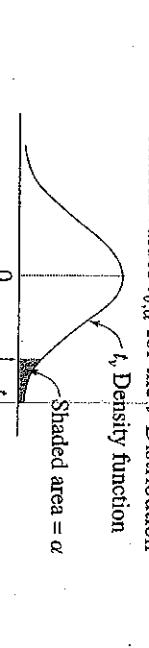
$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	0.9982
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (cont.)

Standard normal density function  
Shaded area =  $\Phi(z)$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0197	0.0192	0.0188	0.0183	0.0180
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0394	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3839
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

**Table A.4** Critical Values  $t_{v,\alpha}$  for the  $t$ -Distribution



$v$	.10	.05	.025	.01	.005	.001	.0005
	$\alpha$						
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

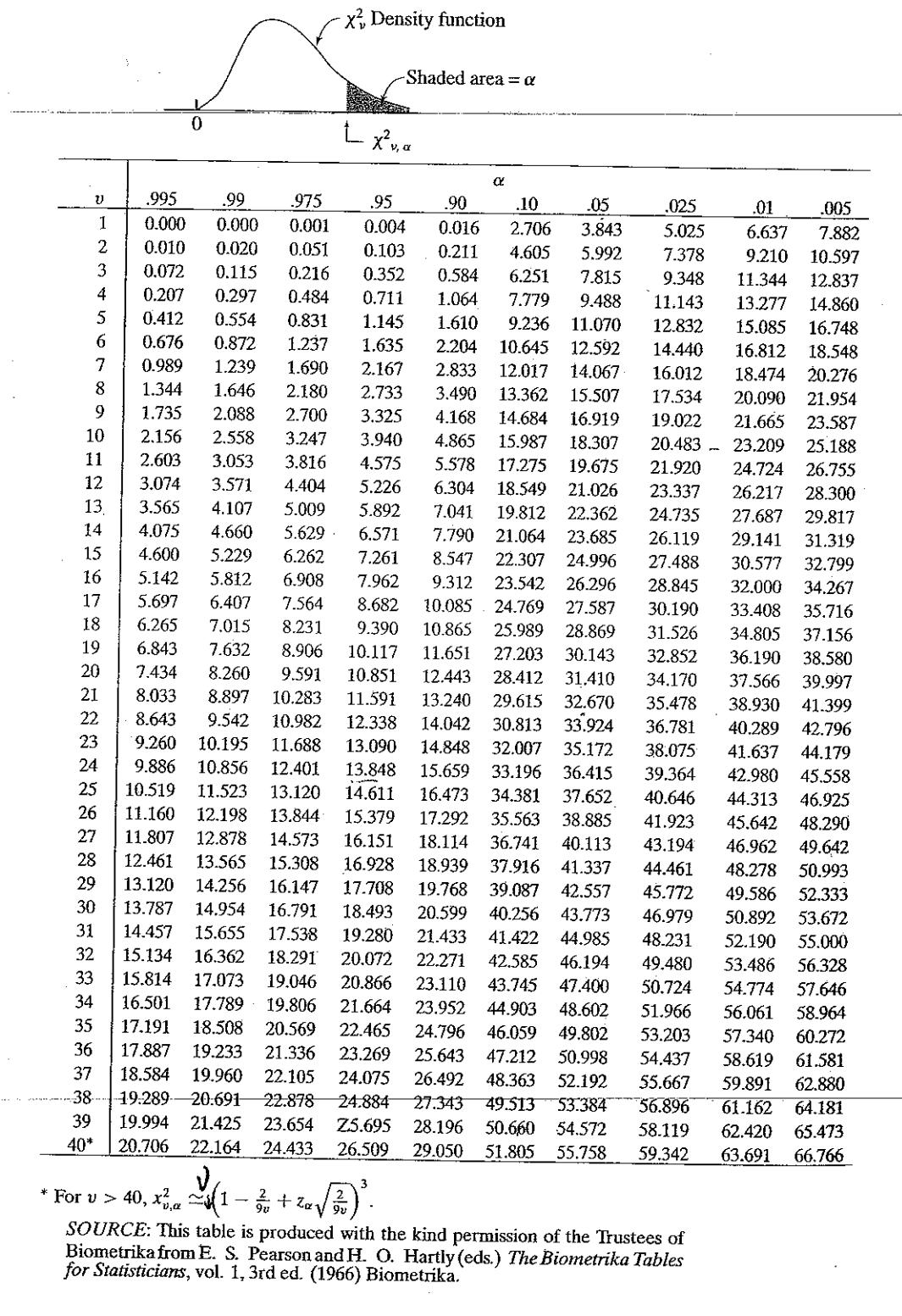
**SOURCE:** This table is produced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.), *The Biometrika Tables for Statisticians*, vol. 1, 3rd ed. (1966), Biometrika.

**Table A.2** Cumulative Poisson Probabilities (cont.)

	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	15.0	20.0
1	.406	.135	.050	.018	.007	.002	.001	.000	.000	.000	.000
2	.677	.423	.238	.125	.062	.030	.014	.006	.003	.000	.000
3	.857	.647	.433	.265	.151	.082	.042	.021	.010	.000	.000
4	.947	.815	.629	.440	.285	.173	.100	.055	.029	.001	.000
5	.983	.916	.785	.616	.446	.301	.191	.116	.067	.003	.000
6	.995	.966	.889	.762	.606	.450	.313	.207	.130	.008	.000
7	.999	.988	.949	.867	.744	.599	.453	.324	.220	.018	.001
8	1.000	.996	.979	.932	.847	.729	.593	.456	.333	.037	.002
9	1.000	.999	.992	.968	.916	.830	.717	.587	.458	.070	.005
10	1.000	.997	.986	.957	.901	.816	.706	.583	.466	.066	.006
11	1.000	.999	.995	.980	.947	.888	.803	.697	.568	.105	.011
12	1.000	.998	.991	.973	.936	.876	.792	.668	.531	.381	.021
13	1.000	.999	.996	.987	.966	.926	.864	.739	.604	.466	.157
14	1.000	.999	.994	.983	.959	.917	.886	.753	.621	.494	.221
15	1.000	.999	.998	.992	.978	.951	.917	.884	.750	.621	.494
16	1.000	.999	.996	.989	.973	.944	.917	.884	.851	.718	.588
17	1.000	.999	.999	.998	.995	.983	.970	.957	.934	.891	.851
18	1.000	.999	.999	.998	.995	.986	.974	.961	.948	.917	.884
19	1.000	.999	.999	.998	.995	.986	.974	.961	.948	.917	.884
20	1.000	.999	.999	.998	.995	.986	.974	.961	.948	.917	.884
21	1.000	.998	.997	.996	.995	.984	.973	.961	.948	.917	.884
22	1.000	.999	.999	.998	.997	.995	.993	.990	.987	.974	.951
23	1.000	.999	.999	.998	.997	.996	.995	.994	.992	.989	.976
24	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.990	.987
25	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.991	.988
26	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989
27	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989
28	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989
29	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989
30	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989
36	1.000	.999	.999	.998	.997	.996	.995	.994	.993	.992	.989

**SOURCE:** L. L. Chao (1974), *Statistics: Methods and Analysis*, 2nd ed. New York: McGraw-Hill.

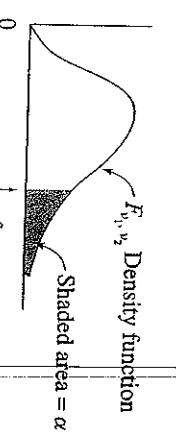
**Table A.5 Critical Values  $\chi^2_{v,\alpha}$  for the Chi-square Distribution**



$$* \text{ For } v > 40, \chi^2_{v,\alpha} \approx \sqrt{1 - \frac{2}{9v} + z_\alpha \sqrt{\frac{2}{9v}}}.$$

SOURCE: This table is produced with the kind permission of the Trustees of Biometrika from E. S. Pearson and H. O. Hartley (eds.) *The Biometrika Tables for Statisticians*, vol. 1, 3rd ed. (1966) Biometrika.

Table A.6 Critical Values  $f_{v_1, v_2, \alpha}$  for the  $F$ -Distribution ( $\alpha = .05$ ) (cont.)



	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
	Degrees of freedom for the numerator ( $v_1$ )																		
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.55	8.53	
4	7.71	6.94	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.59	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	
8	5.32	4.46	4.07	3.84	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.01	2.97	2.93		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.75	2.71	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.53	2.49	2.45	2.40	2.36	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.78	2.70	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.26	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.49	2.40	2.33	2.29	2.25	2.20	2.16	2.11	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	
17	4.45	3.59	3.20	2.96	2.81	2.69	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.10	2.06	2.01	1.96	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.06	2.02	1.97	1.92	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.78	
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.81	1.80	1.75	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	
28	4.20	3.34	2.95	2.71	2.56	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.76	1.70	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	
31	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.62	
32	4.00	3.15	2.76	2.53	2.37	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39	
33	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.81	1.75	1.66	1.61	1.55	1.55	1.43	1.35	
34	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	

## MASTER'S DIAGNOSTIC EXAMINATION - JANUARY 7, 2014

Student's Name \_\_\_\_\_

### INSTRUCTIONS FOR STUDENTS:

1. The exam is to be started at 1 pm (CDT) and completed by 5 pm (CDT) on January 7, 2014.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the  
UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
8. Be sure to hand in/send all of your pages to the solutions for the exam questions. No additional material will be accepted once the exam has ended and you have left the exam room or sent your solutions.
9. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed
- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

### INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** cover page and solutions to **979-845-6060** or **Scan** cover page and solutions into a **single** pdf file and **email** to **longneck@stat.tamu.edu**

**Do not** send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

**QUESTION I.** There are two parts to this Question.

**Question I - Part A.**

For the experiment described below, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units.
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures
6. An ANOVA Table with just the following information: Sources of Variation and Degrees of Freedom Freedom

**Description of the Experiment:**

An evaluation of the effectiveness of three weed treatments on the yield of wheat raised in the midwest of the U.S. was conducted. There were eight fields used in the study with each field containing three widely separated tracts of land. The eight fields are randomly assigned to the two varieties of wheat, V1, or V2, with four fields randomly assigned to each variety. Within each field, the three tracts are randomly assigned to the three weed treatments, W1, W2, or W3, with one weed treatment per tract. Finally, each tract is divided in half, with one half randomly assigned the L amount of weed treatment and the other half the H amount. At the end of the growing season, the total yield of wheat for each half of the twenty four tracts are recorded. The yields are given in the following table.

FIELD	VARIETY	WEED TREATMENT					
		W1		W2		W3	
		L	H	L	H	L	H
F1	V1	83.2	81.8	67.4	79.7	75.9	80.6
F2	V2	77.5	78.2	69.2	71.5	75.9	78.2
F3	V1	72.7	69.3	70.1	71.2	75.9	81.3
F4	V2	75.3	78.9	72.7	74.6	75.9	82.8
F5	V1	78.2	80.5	65.1	68.3	65.3	66.6
F6	V2	79.8	85.2	57.6	61.4	58.5	61.6
F7	V1	82.4	83.1	50.5	54.0	51.6	54.7
F8	V2	75.5	78.7	39.0	43.9	41.9	45.1

## **Question I - Part B.**

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations. Provide justification for your selection.

### **SITUATION:**

1. A CRD was conducted with Factor  $F_1$  having four fixed quantitative levels and Factor  $F_2$  with six randomly selected levels. The AOV table reveals that the interaction between  $F_1$  and  $F_2$  was significant. The researcher wanted to investigate the change in the mean of the response with increasing levels of factor  $F_1$ .
2. An experiment was designed to compare four techniques, the levels of  $F_1$ , for removing mercury contamination from drinking water. The researcher wanted to also evaluate the variability in the many devices, the levels of  $F_2$ , of measuring mercury levels in water. Five devices for detecting mercury were randomly selected from the list of all such devices. A specified amount of mercury was placed in 200 water samples. Ten of the 200 water samples were randomly assigned to each of the twenty combinations of a level of  $F_1$  and a level of  $F_2$ . There was significant evidence of an interaction between factors  $F_1$  and  $F_2$ . The researcher wants to determine which of the four techniques removed the greatest amount of mercury.
3. A three factor experiment is run with Factor  $F_1$  having five fixed levels, Factor  $F_2$  with six fixed selected levels and Factor  $F_3$  with four fixed levels. The results from the AOV were

$F_2$ ,  $F_1 * F_2$ ,  $F_1 * F_3$ ,  $F_1 * F_2 * F_3$  were nonsignificant.

$F_1$ ,  $F_3$ , and  $F_2 * F_3$ , were significant.

The statistician wants to evaluate the pairwise differences in the levels of factor  $F_1$ .

4. An experiment is conducted using a factorial treatment structure with factor  $F_1$  having values  $40^{\circ}C$ ,  $50^{\circ}C$ ,  $60^{\circ}C$ ,  $70^{\circ}C$  crossed with factor  $F_2$  having levels A, B, C in a CRD with three reps per treatment. There is not significant evidence of an interaction between  $F_1$  and  $F_2$ . The researcher wants to determine the temperature that yields the maximum mean response.
5. In an experiment having the levels of factor  $F_1$ -qualitative and the levels of factor  $F_2$ -quantitative, there was significant evidence of an interaction between  $F_1$  and  $F_2$ . The experimenter wants to compare the mean responses across the levels of factor  $F_1$ , averaged over the levels of factor  $F_2$ .
6. An experiment was designed to compare the performance of three new types of machine tools to the machine tool currently in use, factor  $F_1$ , with four levels. A random sample of five machinists, factor  $F_2$ , were randomly selected from the workforce. Each machinist produced ten units of product from each of the four types of machines. A quality rating was determined for each of the 200 units produced in the study. There was significant evidence of an interaction between factors  $F_1$  and  $F_2$ . The company wants to know if any of the new types of machines have a higher mean quality rating than the type of machine the company is currently using.

**TECHNIQUE:**

- A. Trend analysis using Scheffe contrasts
- B. Trend analysis using Bonferroni contrasts
- C. Trend analysis in the levels of  $F_1$  averaged over levels of the other factors
- D. Trend analysis in the levels of  $F_1$  separately at each level of the other factors
- E. Scheffe's test for contrast differences
- F. Dunnett's comparison technique
- G. Dunnett's comparison technique to all combinations of the factors
- H. Dunnett's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- I. Dunnett's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- J. Tukey's comparison technique
- K. Tukey's comparison technique to all combinations of the factors
- L. Tukey's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- M. Tukey's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- N. Hsu's comparison technique
- O. Hsu's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- P. Hsu's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- Q. Hsu's comparison technique applied to all combinations of the factors
- R. Nothing new is learned beyond the results of the F-tests from the AOV table.
- S. Comparison of marginal means is not appropriate.
- T. None of the above methods are appropriate.

## QUESTION II.

Consider the following linear model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3(x_i \times z_i) + \epsilon_i$$

where  $x$  is continuous and  $z$  is binary. The output from fitting this model to a sample of size  $n = 250$  is shown below:

Call:

```
lm(formula = y ~ x + z + x * z)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.28335	-0.34073	-0.04031	0.33622	1.36754

Coefficients:

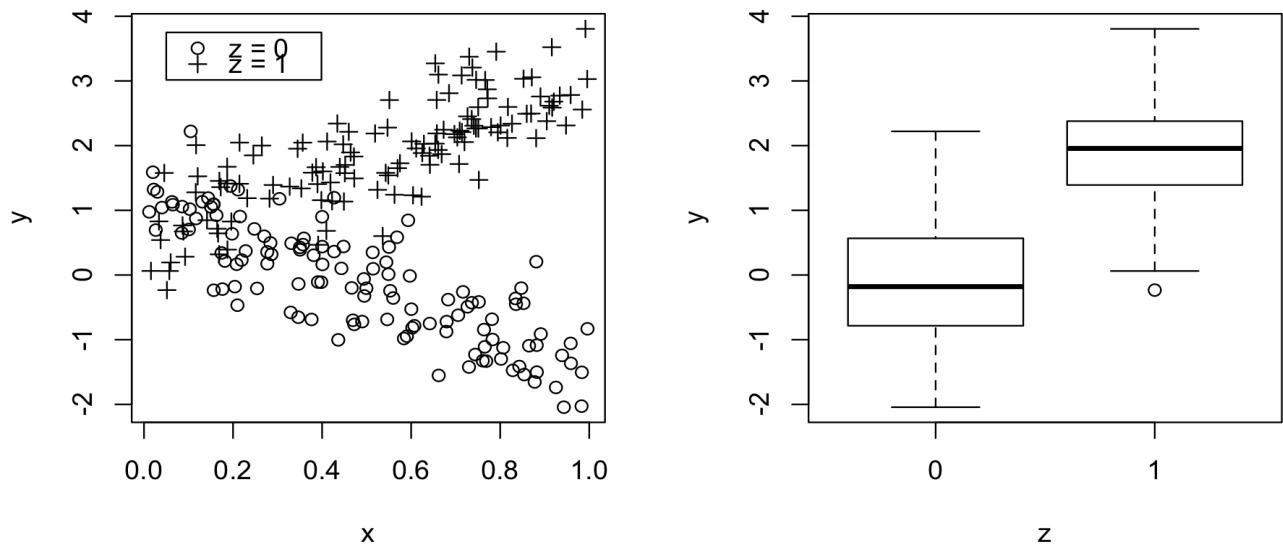
	Estimate	Std. Error	t value	Pr(> t )		
(Intercept)	1.15235	0.08933	12.900	< 2e-16 ***		
x	-2.62637	0.16003	-16.412	< 2e-16 ***		
z	-0.55213	0.13378	-4.127	5.03e-05 ***		
x:z	5.02318	0.23094	21.751	< 2e-16 ***		
---						
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Residual standard error: 0.5035 on 246 degrees of freedom

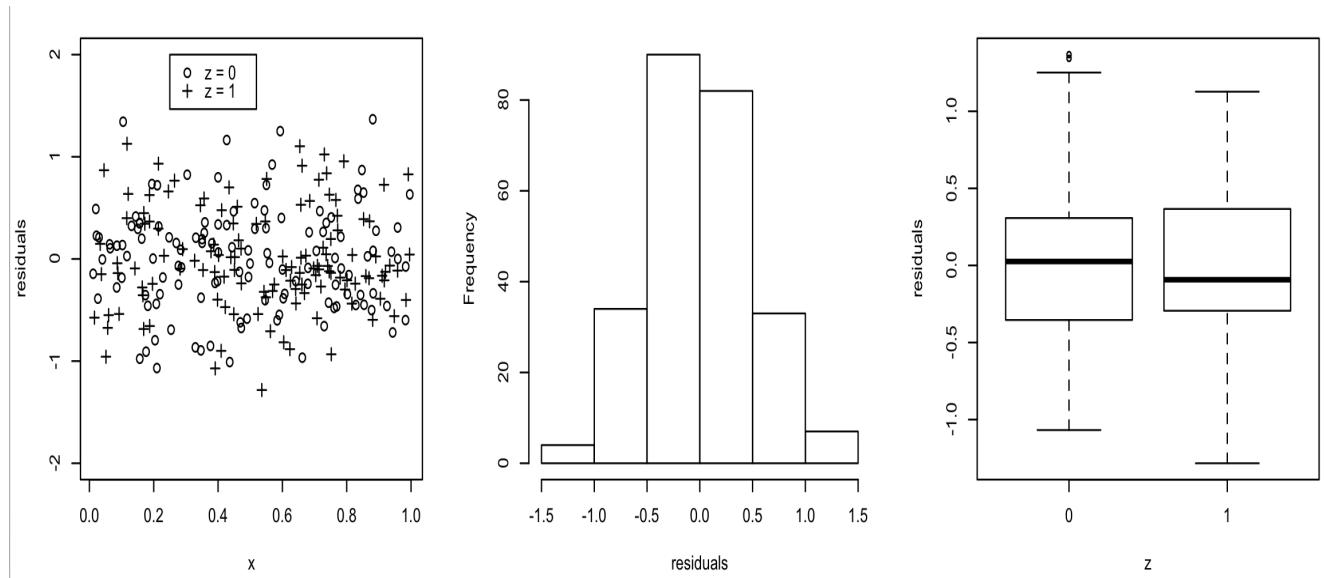
Multiple R-squared: 0.8554, Adjusted R-squared: 0.8536

F-statistic: 485.1 on 3 and 246 DF, p-value: < 2.2e-16

1. Interpret each of the model coefficients ( $\beta_0, \beta_1, \beta_2, \beta_3$ ) in terms of expected values.
2. Report a 95% confidence interval for the mean change in  $y$  associated with a one-unit increase in  $x$ , when  $z = 0$ .
3. What is the slope parameter (mean change in  $y$  associated with a one-unit increase in  $x$ ) for  $x$  when  $z = 1$ ?
4. What does the adjusted R squared measure?
5. What null and alternative hypotheses are tested using the  $F$  statistic at the bottom of the R output?
6. The figures below show diagnostic plots for the above model. Which of your model assumptions, if any, appear to not be met? And why?
7. One of the assumptions of our model is that the  $\epsilon_i$  are *i.i.d.* realizations from the Normal distribution with mean 0 and constant variance  $\sigma^2$ . What does the model report as an estimate of  $\sigma$ ?



Scatterplot of  $y$  vs.  $x$ , and side-by-side boxplots comparing  $y$  to  $z$ .



Residual plots: scatterplot of residuals vs.  $x$ ; histogram of residuals; and side-by-side boxplots of residuals vs.  $z$

### QUESTION III.

#### Question III - Part A.

Let  $X_1, \dots, X_{20}$  be a random sample from the normal distribution with mean 20 and standard deviation 5. and  $Y_1, \dots, Y_{25}$  be a random sample from the normal distribution with mean 24 and standard deviation 4. Assume that  $X_1, \dots, X_{20}, Y_1, \dots, Y_{25}$  are mutually independent.

1. Identify completely the distribution of  $\bar{X} = \sum_{i=1}^{20} X_i/20$  and the distribution of  $\bar{Y} = \sum_{j=1}^{25} Y_j/25$ .
2. Identify the distribution of  $W = \bar{X} - \bar{Y}$  and obtain an expression for  $P(\bar{X} < \bar{Y})$  in terms of the standard normal cumulative distribution function,  $\Phi$ .
3. Let  $U = (X_1 - \bar{X})^2 + \dots + (X_{20} - \bar{X})^2$ .  
Derive an expression for  $P(U > 50)$  in terms of the cumulative distribution function of a chi-squared distribution.
4. For what values of  $K$  and  $m$  is it true that the quantity

$$T = \frac{K(\bar{X} - 20)}{\sqrt{U}}$$

has a  $t$  distribution with  $m$  degrees of freedom.

#### Question III - Part B.

Let  $X \sim N(-2, 6)$ ,  $Y \sim N(3, 4)$ , and  $Z \sim N(0, 1)$  be independent normal random variables.  
(Note: The notation  $N(a, b)$  indicates a normal distribution with mean  $a$  and variance  $b$ .)

1. Let  $U = 2X + 3Y + Z - 5$  and  $V = X - 2Y - Z$ .  
Identify the distributions of  $U$  and  $V$ .
2. Let  $W = C_1(X + C_2)^2 + C_3(Y + C_4)^2$ .  
Find values of  $C_1, C_2, C_3, C_4$ , and  $C_5$  (with  $C_1 \neq 0$  and  $C_3 \neq 0$ ) so that  $W$  has a chi-squared distribution with  $C_5$  degrees of freedom.

3. Let

$$V = \frac{C_1(X + C_2)^{C_3}}{(Y + C_4)^{C_5}}.$$

Find values of  $C_1, C_2, C_3, C_4, C_5, C_6$  and  $C_7$  so that  $V$  has an  $F$  distribution with  $(C_6, C_7)$  degrees of freedom.

## QUESTION IV.

Consider the usual linear regression model, written either in non-matrix notation as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, 2, \dots, n, \quad (\text{A})$$

where  $e_1, e_2, \dots, e_n$  are independently and identically distributed as  $N(0, \sigma^2)$  random variables or, in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (\text{B})$$

where  $\mathbf{y}$  is an  $n \times 1$  vector of response variables,  $\mathbf{X}$  is an  $n \times p$  matrix of predictor variables (with  $p = k + 1$ ),  $\boldsymbol{\beta}$  is a  $p \times 1$  vector of unknown parameters and  $\mathbf{e}$  is an  $n \times 1$  vector of unobservable independent and identically distributed random  $N(0, \sigma^2)$  variables. In what follows, you may assume that the matrix  $\mathbf{X}$  is of full column rank. Using whichever notation above (A or B) that makes you more comfortable, answer the following parts to this problem. Please be concise with your answers - highly irrelevant statements may be counted against you!

1. The above model is called a linear regression model even though, for example, it encompasses polynomial (in the predictor variables) regression models. Explain what is linear about the above model.
2. Define the least squares criterion. That is, state what property must be satisfied for estimates of the unknown parameters of the above model to be called least squares estimates. Use formulas as part of your definition.
3. Specify which of the above assumptions made about the  $e_i$ 's,  $i = 1, 2, \dots, n$ , need not be true for the least squares estimators of the  $\beta_j$ 's,  $j = 0, 1, \dots, k$ , to be unbiased estimators. If all the above assumptions about the  $e_i$ 's need to be true, then state so.
4. Specify which of the above assumptions made about the  $e_i$ 's,  $i = 1, 2, \dots, n$ , need not be true for the least squares estimator of  $\sigma^2$  to be an unbiased estimator. If all the above assumptions about the  $e_i$ 's need to be true, then state so.
5. Specify which of the above assumptions made about the  $e_i$ 's,  $i = 1, 2, \dots, n$ , need not be true for the usual least squares t tests and F tests of hypotheses about the  $\beta_j$ 's,  $j = 0, 1, \dots, k$ , to be statistically valid. If all the above assumptions about the  $e_i$ 's need to be true, then state so.

**MASTER'S DIAGNOSTIC EXAMINATION - August 20, 2014**

**Student's Name** \_\_\_\_\_

**INSTRUCTIONS FOR STUDENTS:**

1. The exam is to be started at Noon (CDT) and completed by 4 pm (CDT) on August 20, 2014.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the  
UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
8. 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.
9. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed
- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

**Student's Signature** \_\_\_\_\_

**INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or **Scan** the solutions into a **single** pdf file and **email to longneck@stat.tamu.edu**

**Do not** send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **longneck@stat.tamu.edu**.

**Proctor's Signature** \_\_\_\_\_

**QUESTION I. Part A:**

For the following experiment, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units;
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures;
6. A partial ANOVA Table containing Sources of Variation (SV), Degrees of Freedom, and Expected Mean Squares;
7. Write a statistical model for this experiment and include all necessary conditions on the model parameters and variables.

A leading brand of ice cream designs an experiment to evaluate the impact of several artificial sweeteners on the texture of their product. It is well known that replacing natural sweeteners with artificial sweeteners in ice cream can result in a product which has an unappealing texture. A proposed method to overcome this problem is to increase the blending time in the production process. The researchers decided to use four types of sweeteners: a natural sweetener (Control), Aspartame, Saccharin, and Sucralose. Twelve containers of ice cream were made, three of each of the four types of sweeteners, with the type of sweetener randomly assigned to the containers. Each of the 12 containers of ice cream was then split into four portions. The four portions are then randomly assigned to one of four blending times: 1 minute, 2 minutes, 5 minutes, and 8 minutes. At the end of the specified blending period, the ice cream is assigned a texture score. The researcher was particularly interested in the impact of the four sweeteners and the blending times on the average texture scores.

Sweetener	Container	Blending Time(min.)			
		1	2	5	8
Control	1	7	10	17	22
	2	4	4	11	23
	3	4	11	10	31
Aspartame	1	8	12	22	27
	2	6	7	27	30
	3	9	8	29	32
Saccharin	1	7	8	21	35
	2	1	4	13	25
	3	5	4	13	28
Sucralose	1	3	11	21	37
	2	1	12	25	31
	3	4	9	27	32

## **PROBLEM I. Part B:**

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations.

### **SITUATION:**

- (1) A CRD with three factors:  $F_1$ -fixed levels,  $F_2$ -fixed levels,  $F_3$ -fixed levels, was conducted. The experimenter obtained the following results from the AOV  $F$ -tests:  $F_1 * F_2 * F_3$  is significant,  $F_1 * F_2$ -not significant,  $F_1 * F_3$ -significant,  $F_2 * F_3$ - not significant, and  $F_1, F_2, F_3$  are all not significant. She wants to determine which pairs of means are different across the levels of  $F_1$ .
- (2) A three factor experiment is run with Factor  $F_1$ -fixed, Factor  $F_2$ -fixed having levels nested within the levels of Factor  $F_1$ , Factor  $F_3$ -fixed crossed with Factor  $F_1$ . The interaction between factors  $F_1$  and  $F_3$  was not significant and the interaction between factors  $F_2(F_1)$  and  $F_3$  was not significant. The researcher was interested in determining which pairs of means are different across the levels of  $F_1$ .
- (3) A RCBD with three factors:  $F_1$ -fixed,  $F_2$ -fixed,  $F_3$ -random, was conducted. The experimenter obtained the following results from the AOV  $F$ -tests:  $F_1 * F_2 * F_3$  is not significant,  $F_1 * F_2$ -not significant,  $F_1 * F_3$ -significant,  $F_2 * F_3$ -significant, and  $F_1, F_2, F_3$  are all not significant. She wants to determine if there are pairwise differences in the levels of  $F_1$ .
- (4) In a quality control experiment, the production engineer was interested in evaluating factors which may have caused a high defective rate in a product. There are five Rates,  $F_1$ , at which a platinum coating is applied to the product, with levels, 1.0, 1.1, 1.2, 1.3, 1.4 mm/second. The second factor,  $F_2$ , is four Types of Machines used to apply the coating to the product, with levels, M1, M2, M3, M4. The third factor,  $F_3$ , was the Operators the coating machines. Twenty operators of the coating machines were randomly selected from the workforce. Each operator applied the coating to 80 units of the product, four units for each combination of a Rate and Type of Machine. There was significant evidence of a 3-factor interaction and all 2-factor interactions were found to be significant. The company wants to know if the mean defective rate,  $\mu_{ijk}$ , increased as the Rate,  $F_1$ , of applying the coating was increased.
- (5) An experiment is designed to investigate plant growth involving a factorial treatment structure with factor  $F_1$ , the temperature in a growth chamber,  $15^\circ C$ ,  $20^\circ C$ ,  $30^\circ C$ ,  $35^\circ C$  crossed with factor  $F_2$ , four brands of growth stimulants at three dose levels: 0 ml/mg, 10 ml/mg, 15 ml/mg, factor  $F_3$ . The experiment was conducted as a completely randomized design with 10 flowers randomly assigned to each of the 36 treatments. The experimenter determined from the AOV  $F$ -tests that only the following effects were significant:  $F_1 * F_3$ ,  $F_1 * F_2$ ,  $F_2$ ,  $F_3$ . The researcher wants to determine the temperature that yields the maximum mean growth.

**TECHNIQUE:**

- A. Trend analysis using Scheffe contrasts
- B. Trend analysis using Bonferroni contrasts
- C. Trend analysis in the levels of  $F_1$  averaged over levels of the other factors
- D. Trend analysis in the levels of  $F_1$  separately at each level of the other factors
- E. Trend analysis in the levels of  $F_1$  separately at each level of  $F_2$  but averaged over the other factors
- F. Scheffe's test for contrast differences
- G. Dunnett's comparison technique
- H. Dunnett's comparison technique to all combinations of the factors
- I. Dunnett's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- J. Dunnett's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- K. Tukey's comparison technique
- L. Tukey's comparison technique to all combinations of the factors
- M. Tukey's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- N. Tukey's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- O. Hsu's comparison technique
- P. Hsu's comparison technique applied to the levels of factor  $F_1$  separately at each level of the other factors
- Q. Hsu's comparison technique applied to the levels of factor  $F_1$  averaged over the levels of the other factors
- R. Hsu's comparison technique applied to all combinations of the factors
- S. Nothing new is learned beyond the results of the F-tests from the AOV table.
- T. Comparison of marginal means is not appropriate.
- U. None of the above methods are appropriate.

## QUESTION II.

Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \dots, n$$

Suppose we observe the following values for the  $y_i$  and  $x_i$ :

$y$	$x$
0.34	1
1.28	2
1.16	3
2.11	4
2.66	5

Recall that  $(\hat{\beta}_0, \hat{\beta}_1)' = \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , where  $\mathbf{X}$  is the model matrix and that  $\text{Var}(\mathbf{v}'\mathbf{Y}) = \mathbf{v}'\text{Var}(\mathbf{Y})\mathbf{v}$ , where  $\mathbf{v}$  is a vector of constants and  $\mathbf{Y}$  is a random vector.

1. Compute  $\hat{\beta}$ .
2. Compute  $\hat{\sigma}$ .
3. Compute the standard errors of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .
4. Compute a 95% confidence interval for the mean response when  $x = 3$ .
5. Compute a 95% prediction interval for a new observation when  $x = 3$ .

### QUESTION III.

**Part A.** The safety of people who live or work near nuclear-power plants has been under debate in recent years. One possible health hazard is an excess number of deaths due to cancer among those exposed. A problem with studying this is that the number of deaths from cancer is small, making it difficult to make statistical conclusions. An alternative approach used by epidemiologists is the *proportional mortality study*, in which the proportion of deaths in the study group is compared with the corresponding proportion in a large population. Suppose that 15 deaths have occurred among 55- to 64-year-old male workers in a nuclear power plant during a given time period and that in 6 of them the cause of death was cancer. In the general population of 55- to 64-year-old males it is known that approximately 20% of all deaths can be attributed to cancer.

1. Formulate the hypotheses, compute a  $p$ -value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer exceeds 20%. You may find it useful to use the attached tables of the binomial distribution.
2. Statisticians have found that using one-sided tests based upon a test statistic with a discrete distribution can be overly conservative (i.e., losing power by failing to reject too often for a test of the desired level of significance). To help adjust for this phenomenon, one can use the mid  $P$ -value for a one-sided test. The mid  $P$ -value equals one-half the probability of the observed result plus the probability of the more extreme results. Compute a the mid  $P$ -value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer exceeds 20%.
3. Taking into account the sample size, construct a 95% confidence interval for the true proportion of 55- to 64-year-old male workers in a nuclear power plant having deaths attributable to cancer.

**Part B.** The epidemiologist also considered data for all 55- to 64-year-old male workers in the local area. Suppose that 254 deaths occurred among 55- to 64-year-old male workers during the given time period and that in 59 of them the cause of death was cancer.

1. Formulate the hypotheses, compute a  $p$ -value, and state a conclusion at the 5% level for a test of whether the true proportion of 55- to 64-year-old male workers in the local area having deaths attributable to cancer exceeds 20%.
2. Taking into account the sample size, construct a 95% confidence interval for the true proportion of 55- to 64-year-old male workers in the local area having deaths attributable to cancer.
3. Give a brief explanation why you were able to answer the two previous parts of this problem without using binomial tables.

## QUESTION IV.

Chapter 2 of Miller (2013) Modelling Techniques in Predictive Analytics, Pearson, New Jersey makes extensive use of multiple regression to analyze attendance figures for the 81 Dodgers home games in the 2012 Major League Baseball season. Data are available on the following variables

- Attendance = home game attendance (i.e., the number of tickets sold to each game)
- Month = month in which each game was played
- Day\_of\_week = day of the week each game was played on
- OpponentsFromLargeMetroAreas = a dummy variable which is 1 if the opponent is the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and the Washington DC Nationals
- Temp = temperature at the stadium during the game
- Day\_night = day (for day games) and night (for night games)
- BobbleheadPromotion = a dummy variable which is 1 if the game involved a bobblehead promotion

According to Miller (2013):

“Dodger Stadium, with a capacity of 56,000, is the largest ballpark in the world. From the data, we can see that Dodger Stadium was filled to capacity only twice in 2012. . . . The eleven bobblehead promotions occurred on night games, six of those being Tuesday nights. . . . Opponents from the large metropolitan areas (the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and Washington D.C. Nationals) are consistently associated with higher attendance. . . . Explanatory graphics help us find models that might work for predicting attendance and for evaluating the effect of promotions on attendance.

Figure 2.1 shows distributions of attendance across days of the week, and

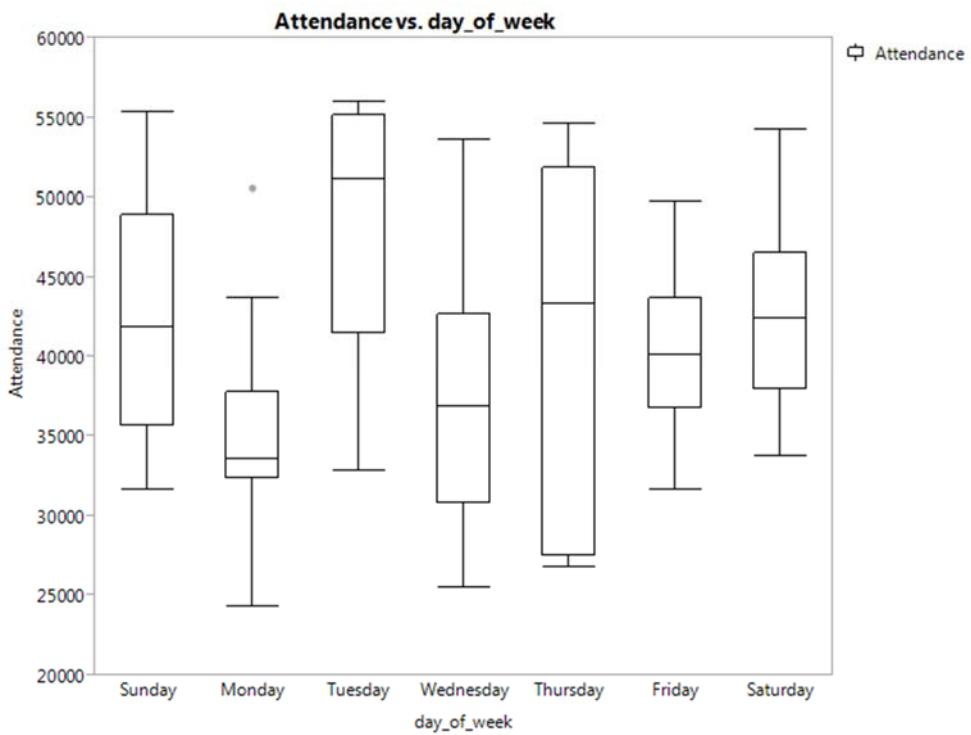
Figure 2.2 shows attendance by month.

To advise management regarding (bobblehead) promotions, we would like to know if promotions have a positive effect upon attendance, and if they do have a positive effect, how much that effect might be. To provide this advice we build a linear model for predicting attendance using month, day of the week and an indicator variable for the bobblehead promotion . . . ”

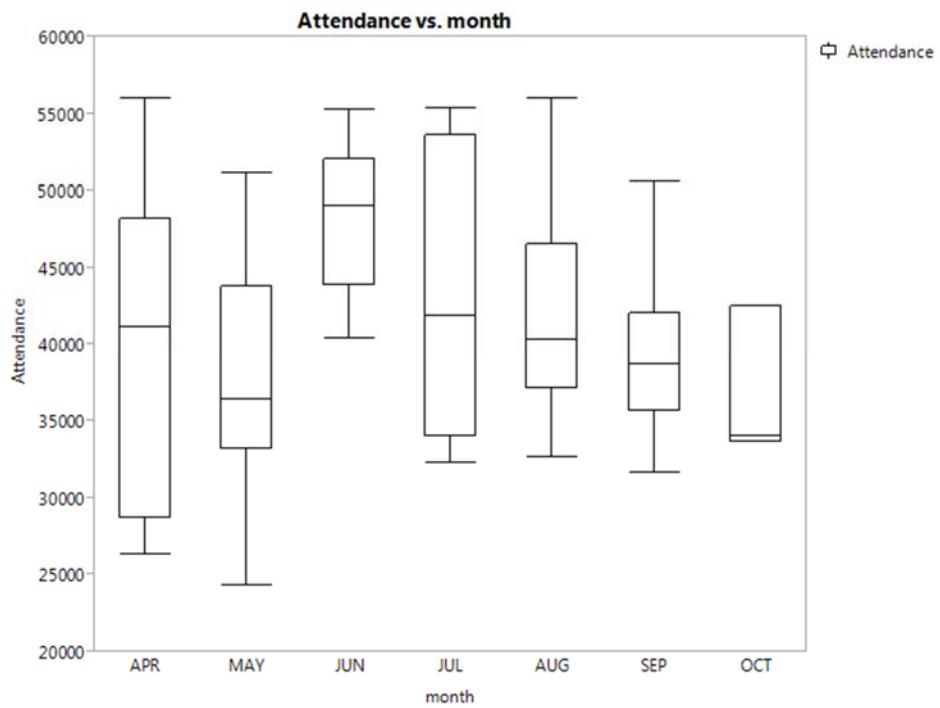
Given on the next few pages are Figures 2.1 and 2.2, JMP output from the least squares model fit by Miller (2013) and some additional plots. A statistics professor originally from Australia has taken a careful look at the data and found among other things that there is no evidence of significant autocorrelation in the attendance results. In other words, there is no evidence that attendance at home games on day  $t$  is statistically significantly related to attendance on days  $t - 1$ ,  $t - 2$ , . . . .

1. Describe in detail one major concern that potentially threatens the validity of the model fit by Miller (2013).
2. Explain the specific steps you would take to overcome the problem described in part (1).
3. On the basis of the plots presented what predictors would you recommend being included in the model you describe in part (2).

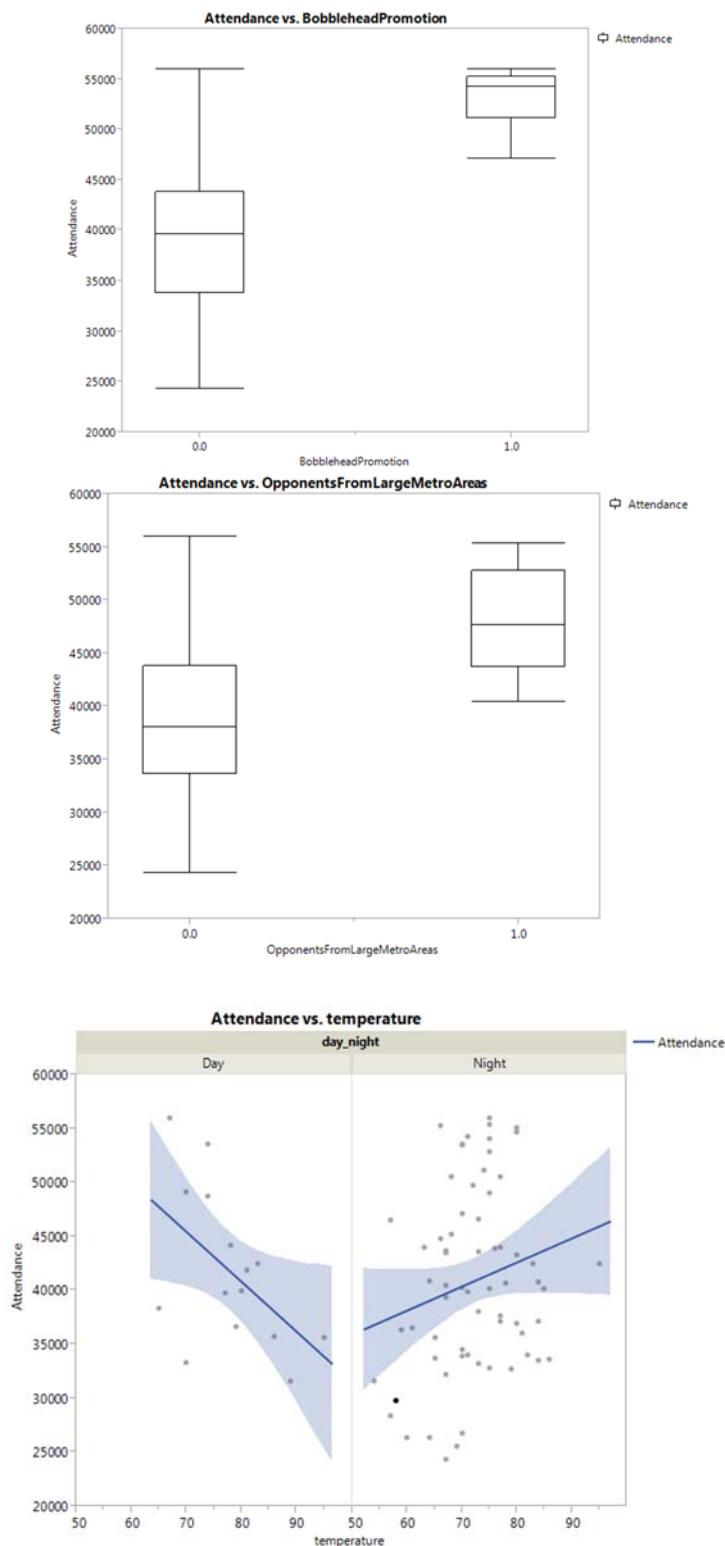
**Figure 2.1**



**Figure 2.2**



## Other plots not in Miller (2013)



## Model fit by Miller (2013)

**dodgersIndicatorVariables - Fit Least Squares - JMP Pro**

**Response Attendance**

**Summary of Fit**

RSquare	0.544371
RSquare Adj	0.455965
Root Mean Square Error	6120.158
Mean of Response	41040.07
Observations (or Sum Wgts)	81

**Analysis of Variance**

**Lack Of Fit**

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	39408.945	899.6271	43.81	<.0001*	.
month[APR]	-1338.818	1704.12	-0.79	0.4349	1.0854335
month[MAY]	-3724.442	1458.501	-2.55	0.0129*	1.0348791
month[JUN]	5824.4159	1936.695	3.01	0.0037*	1.1571459
month[JUL]	1511.0105	1743.695	0.87	0.3893	1.1364332
month[AUG]	1039.1067	1576.329	0.66	0.5120	1.076165
month[SEP]	-1309.788	1758.932	-0.74	0.4591	1.1563802
day_of_week[Sunday]	2563.0373	1627.647	1.57	0.1200	1.8389446
day_of_week[Monday]	-4160.965	1741.61	-2.39	0.0197*	2.0234943
day_of_week[Tuesday]	3750.5283	1782.685	2.10	0.0391*	2.2059578
day_of_week[Wednesday]	-1700.942	1725.411	-0.99	0.3278	1.9860279
day_of_week[Thursday]	-3385.602	2525.475	-1.34	0.1846	2.9304781
day_of_week[Friday]	722.85297	1662.321	0.43	0.6651	1.9181293
BobbleheadPromotion	10714.903	2419.52	4.43	<.0001*	1.4857266

**Effect Tests**

Source	Nparm	DF	Sum of Squares		Prob > F
			F Ratio	Prob > F	
month	6	6	620147363	2.7594	0.0186*
day_of_week	6	6	575839199	2.5623	0.0270*
BobbleheadPromotion	1	1	734587177	19.6118	<.0001*

**Table A.1** Cumulative Binomial Probabilities

c.  $n = 15$

## MASTER'S DIAGNOSTIC EXAMINATION - JANUARY 8, 2015

Student's Name \_\_\_\_\_

### INSTRUCTIONS FOR STUDENTS:

1. The exam is to be started at Noon (CST) and completed by 4 pm (CST) on January 8, 2015.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the  
UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
8. 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.
9. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed
- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

Student's Signature \_\_\_\_\_

### INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or  
**Scan** the solutions into a **single** pdf file and **email to longneck@stat.tamu.edu**

**Do not** send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or  
emailed to **longneck@stat.tamu.edu**.

Proctor's Signature \_\_\_\_\_

## QUESTION I.

A company is interested in the ability of a machine to consistently place electrical wire on a coil. There are three types of machines available: hand operated(HO), partially computer operated(PCO), and completely automated(CA). Three machines of each type are randomly selected from their suppliers for use in the study. The wire placed on the coils comes in one of three thicknesses: .02mm, .04mm, or .06mm. Each of the machines assembles two coils of each of the three wire thicknesses. Each wound coil is then measured for the uniformity of windings at a middle position on the coil. These measurements are given in the following table.

		TYPE OF MACHINE									
		HO			PCO			CA			
MACHINE ID		1	2	3	4	5	6	7	8	9	THICK MEANS
THICKNESS	.02mm	12.30	13.46	12.35	13.01	13.46	13.15	5.47	5.75	6.24	10.63
		12.59	14.00	12.06	12.63	13.92	13.20	5.96	5.68	6.15	
		(12.79)			(13.23)			(5.87)			
	.04mm	13.16	13.29	12.50	12.74	13.84	13.46	5.73	5.60	5.92	10.68
		13.00	13.62	12.39	12.68	13.75	13.57	5.64	5.65	5.64	
		(12.99)			(13.34)			(5.70)			
	.06mm	12.87	13.46	12.73	12.47	13.62	13.36	5.01	5.80	6.19	10.60
		12.92	13.82	12.15	12.15	13.28	13.42	5.62	5.71	6.23	
		(12.99)			(13.05)			(5.76)			
TYPE MEANS		12.93			13.21			5.78			10.64

A partial ANOVA table for the experiment is given below.

SOURCE	DF	MS	EMS
THICKNESS		0.0263	
TYPE		319.1202	
THICKNESS*TYPE		0.1152	
MACHINE(TYPE)		1.4935	
THICKNESS*MACHINE(TYPE)		0.0878	
ERROR		0.0445	

1. Complete the ANOVA table by determining the values of Degrees of Freedom and Expected Mean Squares for each of the sources of variation.

Use the information in the ANOVA table to answer the questions on the next page. Use  $\alpha = .05$  in reaching your answers to Questions 3 and 4.

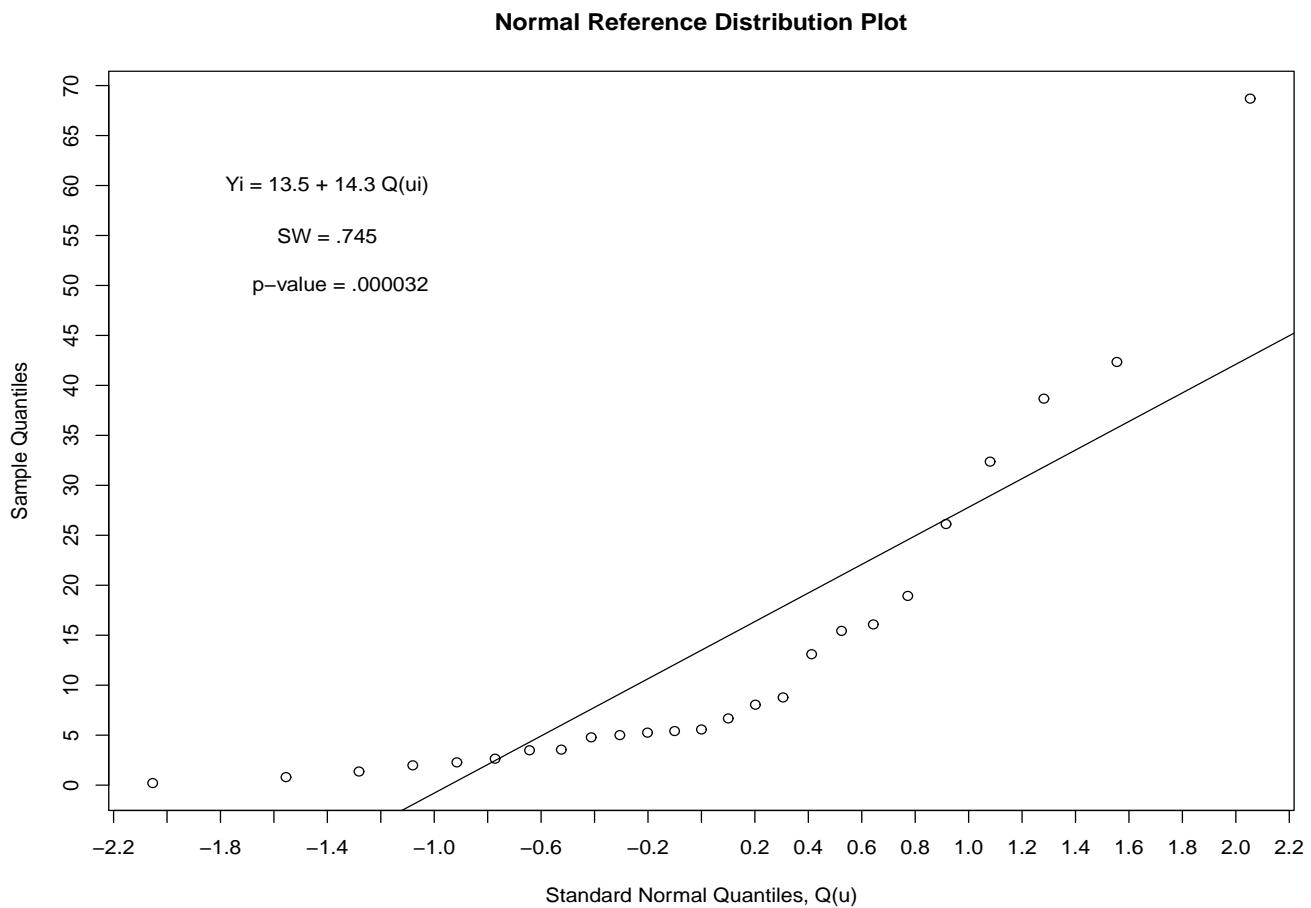
2. Write a linear model for the above experiment. Make sure to identify the terms in your model with respect to distributional properties or restrictions on population parameters.
3. Is the differences in the mean uniformity of the windings produced by the three types of Machines consistent across the three thicknesses?
4. Is there significant difference in the mean uniformity of the windings produced by the three types of Machines?
5. Using the numeric values of the MS's given above and your EMS's, provide the following information:
  - a. Proportionally allocate the variance of the uniformity in windings of a randomly selected coil to the various variance components.
  - b. An estimate of the standard error of the the estimated mean uniformity of windings from a CA winding machine.
  - c. An estimate of the standard error of the estimated difference in the mean uniformity of windings between type HO and CA winding machines.
  - d. An estimate of the standard error of the estimated difference in the mean uniformity of windings of coils using wire of thickness .02mm and .06mm assembled using a CA winding machine.

## QUESTION II.

Twenty five mice are exposed to a high level of radiation and then provided with treatment. The researchers are interested in the survival in days of these mice. The survival times in days for the random sample of 25 mice are given below along with summary statistics and a normal reference plot. The value of the Shapiro-Wilk test was .745 with a p-value of .0000032.

0.2	0.8	1.4	2.0	2.3	2.6	3.5	3.6	4.8	5.0	5.3	5.4	5.6
6.7	8.0	8.8	13.1	15.4	16.1	18.9	26.1	32.4	38.7	42.3	68.7	

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.	MAD
.2	3.49	5.57	13.50	16.80	69	16.55	5.32



Use the above information to answer the questions on the next page:

- Provide a 95% confidence interval for the mean survival of all mice subjected to the treatment using the assumption that the pivot

$$t = \frac{\sqrt{n} (\bar{X} - \mu)}{S}$$

has a t-distribution. A t-table is provided with the exam.

- The researchers feel the interval in part 1. is too imprecise, approximately how many mice would they need in a new experiment to obtain a 95% confidence interval for the mean having a width of 4 days?
- Explain why the confidence interval derived in part 1. may not be an appropriate confidence interval for the mean.
- Suppose the distribution of survival times is gamma, that is, has pdf

$$f(x) = \frac{1}{\Gamma(\alpha)} \beta^{-\alpha} x^{\alpha-1} \exp(-x/\beta) \text{ for some pair } (\alpha, \beta)$$

Explain how to use a simulation to determine the sampling distribution of the pivot

$t = \sqrt{n} (\bar{X} - \mu) / S$ . In particular, explain precisely how to determine  $t_\alpha$ , the  $\alpha$ th percentile of the distribution of  $t$  for any  $\alpha \in (0, 1)$ ?

- Suppose the simulation yielded  $t_{.025} = -3.75$ ,  $t_{.05} = -2.05$ ,  $t_{.95} = 1.45$  and  $t_{.975} = 1.75$ . Based on these percentiles, display a 95% confidence interval for mean survival?
- The treatment used previously for the applied level of radiation had a median survival time of 3 days. Is there significant ( $\alpha = .05$ ) evidence that the new treatment has increased the survival time in comparison to the previous treatment? What is the p-value of your test statistic?

### QUESTION III.

Consider a regression model for an experiment with a continuous response  $Y$ , a treatment factor,  $A$ , with three levels,  $A_1$ ,  $A_2$ , and  $A_3$ , and a continuous explanatory variable,  $X$ . Define the dummy variables,

$$D_1 = 1 \text{ if } A = A_1 \text{ and } D_1 = 0, \text{ otherwise,}$$

$$D_2 = 1 \text{ if } A = A_2 \text{ and } D_2 = 0, \text{ otherwise.}$$

Consider the regression model

$$E(Y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X + \beta_4 D_1 X + \beta_5 D_2 X$$

1. Obtain expressions for the mean response for each of the three treatments:
  - a.  $A_1$
  - b.  $A_2$
  - c.  $A_3$
2. Write out interpretations in terms of the mean response and its relationship to the treatments and explanatory variable for each of the parameters:
  - a.  $\beta_0$
  - b.  $\beta_1$
  - c.  $\beta_2$
  - d.  $\beta_3$
  - e.  $\beta_4$
  - f.  $\beta_5$
3. Formulate the hypotheses for a test of equal slopes for the three treatments. Explain how to carry out the test if you were provided statistical software that could fit a regression model and provide the usual analysis of variance table for the fitted regression model.
4. Does testing the hypothesis  $H_0 : \beta_1 = \beta_2 = 0$  provide a test of equal effects for the three treatments in the above model? Justify your answer.

**MASTER'S DIAGNOSTIC EXAMINATION - August 20, 2015**

**Student's Name** \_\_\_\_\_

**INSTRUCTIONS FOR STUDENTS:**

1. The exam will start at Noon (CDT) on August 20, 2015 and must be completed by 4 pm (CDT) on August 20, 2015.
2. Put your name above but **DO NOT put your NAME** on the **SOLUTIONS** to the exam.
3. Place the **NUMBER** assigned to you on the  
**UPPER RIGHT HAND CORNER** of EACH PAGE of your **SOLUTIONS**.
4. Please start your answer to each of the four questions on a **SEPARATE** sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four problems: Problems I, II, III and IV.
7. Be sure to attempt all parts of the four problems. It may be possible to answer a later part of a problem without having solved the earlier parts.
8. 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 or submitted your solutions.
9. You may use the following:
  - Calculator which does not have capability to phone, text, or access the Web
  - Pencil or pen
  - Blank paper for the solutions for this examination
  - No other materials are allowed
- I attest that I spent no more than 4 hours to complete the exam.
- I used only the materials described above.
- I did not receive assistance from anyone during the taking of this exam.

**Student's Signature** \_\_\_\_\_

**INSTRUCTIONS FOR PROCTOR:**

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or **Scan** the solutions into a **single pdf file** and **email to contact@stat.tamu.edu**

**Do not** send the questions, just send the student's solutions.

- (1) I certify that the time at which the student started the exam was \_\_\_\_\_  
and the time at which the student completed the exam was \_\_\_\_\_
- (2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
- (3) I certify that the student's solutions were faxed to **979-845-6060** or emailed to **contact@stat.tamu.edu**.

**Proctor's Signature** \_\_\_\_\_

## Problem I.

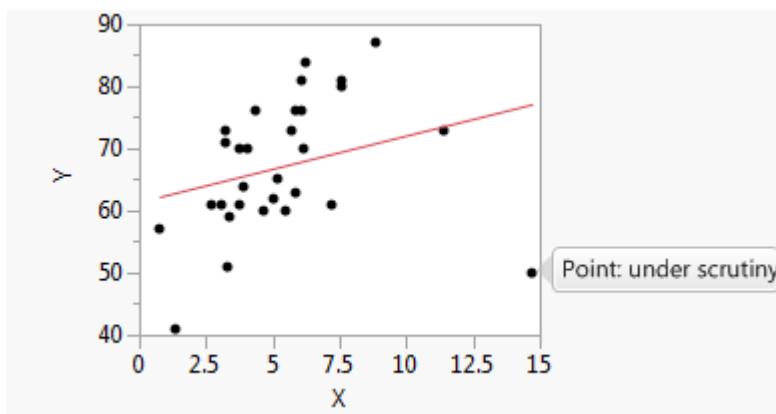
- 1) Given below is scatter plot with the least squares regression line marked on it. We shall consider the point in the bottom right hand corner of the plot which is marked as "Point under scrutiny". This point has a Cook's distance equal to 4.1 and a standardized residual equal to -3.4. An analyst added a dummy variable corresponding to "Point under scrutiny" to the model and refitted the model. Thus the fitted model became

$$Y = b_0 + b_1x_1 + b_2\text{PointUnderScrutiny} + e.$$

Comparing this second model to the first one fit

$$Y = b_0 + b_1x_1 + e.$$

- a. The slope of the regression line (i.e.  $b_1$ ) will be
  - i. The same in both models
  - ii. Higher in the second model
  - iii. Lower in the second model
- b.  $R^2$  will
  - i. The same in both models
  - ii. Higher in the second model
  - iii. Lower in the second model
- c. The regression coefficient of the dummy variable will be
  - i. Zero
  - ii. Positive
  - iii. Negative



2. Explain how to check whether the following modeling assumption is a reasonable for a given multiple regression model

$$Y = g(b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k) + e$$

where  $g$  is an unspecified function. In particular, what plot should you look at and what should you look for?

3. Consider the situation in which the following regression model has been fit to a set of data

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

Suppose that the hypothesis test based on Analysis of Variance leads to a strong rejection of the joint hypothesis that regression coefficients are all zero and that each individual t-tests of the form  $H_{01}: b_1 = 0$ ,  $H_{02}: b_2 = 0$  and  $H_{03}: b_3 = 0$  are strongly rejected. Suppose that the estimated values of  $b_1$  and  $b_2$  are positive as expected. On the other hand, suppose that the estimated value of  $b_3$  is negative, when it was expected to be positive. Describe why this is likely to have occurred and what plots and which statistics you would look at to diagnose the problem.

## Problem II.

An experiment was conducted to assess the effect of a new drug on weight loss. A random sample of 160 overweight people was obtained, consisting of 80 females and 80 males. Within each gender group, half of the people were randomized to receive the drug, with the remaining half receiving a placebo. Let  $\mathbf{y}' = [y_1, y_2, \dots, y_{160}]$  be the vector containing weight loss measurements in pounds for each of the participants, ordered so that the first 40 entries are for females in the placebo group, followed by females in the treatment group, then males in the placebo group, then males in the treatment group. Define the model matrix

$$\mathbf{X}_{160 \times 4} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}$$

where  $\mathbf{1}$  and  $\mathbf{0}$  are column vectors containing 40 ones and zeros, respectively. We have

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.025 & -0.025 & -0.025 & 0.025 \\ -0.025 & 0.050 & 0.025 & -0.050 \\ -0.025 & 0.025 & 0.050 & -0.050 \\ 0.025 & -0.050 & -0.050 & 0.100 \end{bmatrix}$$

The regression model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  was fit, where  $\boldsymbol{\epsilon}$  is a vector of random errors. The coefficient estimates are

$\hat{\boldsymbol{\beta}}' = [4.74, 3.13, 7.47, 3.77]$ , and the residual standard deviation is  $\hat{\sigma} = 4.92$ .

(1) What are the assumptions of the model?

(2) Interpret all model coefficients.

(3) Testing at significance level  $\alpha = 0.05$ , does weight loss in the placebo group differ between males and females? Support your answer. Also (while you do not have to compute it here), carefully interpret what the p-value means.

(4) In terms of the model coefficients, what is the appropriate null hypothesis for testing whether there is an overall treatment effect, and how would you test this?

(5) What is an approximate 95% confidence interval for mean weight loss among males in the treatment group? Hint: If  $\mathbf{v}$  is a vector of constants and  $\mathbf{Z}$  is a random vector of the same dimension as  $\mathbf{v}$ , then  $\text{Var}(\mathbf{v}'\mathbf{Z}) = \mathbf{v}'\text{Var}(\mathbf{Z})\mathbf{v}$ . Carefully interpret what your confidence interval means.

(6) Let  $[\text{lo}_{\text{ind}}, \text{hi}_{\text{ind}}]$  be the 95% confidence interval for the mean difference in weight loss comparing treated females to placebo females, based on the above model. Now consider an alternative design for the experiment, in which there were only 40 females and 40 males, each assessed both with and without the drug. Suppose we fit the same model as above to the resulting 160 weight loss observations from the alternative design, this time incorporating correlations when fitting the model. Let  $[\text{lo}_{\text{cor}}, \text{hi}_{\text{cor}}]$  be the 95% confidence interval for mean difference in weight loss comparing treated females to placebo females, according to the correlation-based model. How would  $[\text{lo}_{\text{cor}}, \text{hi}_{\text{cor}}]$  compare to  $[\text{lo}_{\text{ind}}, \text{hi}_{\text{ind}}]$ , and why?

### Problem III.

1. An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age and health was randomly divided into two groups of 12 lambs. One group of 12 lambs were injected with the drug, and the remaining 12 were left untreated. After a six-month period, the lambs were slaughtered and the worm counts are given in the following table:

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Drug-treated sheep ( $x_1$ )	18	43	28	50	16	32	13	35	38	33	6	7
Untreated sheep ( $x_2$ )	40	54	26	63	21	37	39	23	48	58	28	39

Some possibly useful statistics follow:

$$\bar{x}_1 = 26.583, s_1 = 14.362, \bar{x}_2 = 39.667, s_2 = 13.859, d = x_2 - x_1, \bar{d} = -13.083, s_d = 12.887$$

- (a) What is the design of the experiment? Discuss its appropriateness for this experiment.
- (b) Obtain a 95% confidence interval for the difference in mean tapeworm counts for the two treatments. Interpret your results.
- (c) The researcher states that she is certain that the data does not follow a normal distribution. Describe a procedure to determine whether the treatment was effective, that is, the worm counts for the treated sheep tend to be less than worm counts for the untreated sheep. Compute the value of the test statistic. It is not necessary to compute the p-value nor to compare the computed value of the test statistic to a critical value.
2. We wish to assess the efficacy of a psychiatric treatment for depression. Each patient is assessed as to whether they are depressed at the beginning of the study (Year 0) and then assessed as to whether they are depressed after a year of treatment.
- (a) The data can be recorded into either of two tables:

		Table A				Table B					
		Outcome		Total		Diagnosis at Year 0		Diagnosis at Year 1		Total	
Year	Depress		Depress		Total	Depress	Yes	Depress		Total	Total
	Yes	No						Depress	Yes		
0	317	164			481	Depress	Yes	276	41	317	
1	285	196			481	Depress	No	9	155	164	
						Total		285	196	481	

Discuss which table is more appropriate for presenting these data. Explain why.

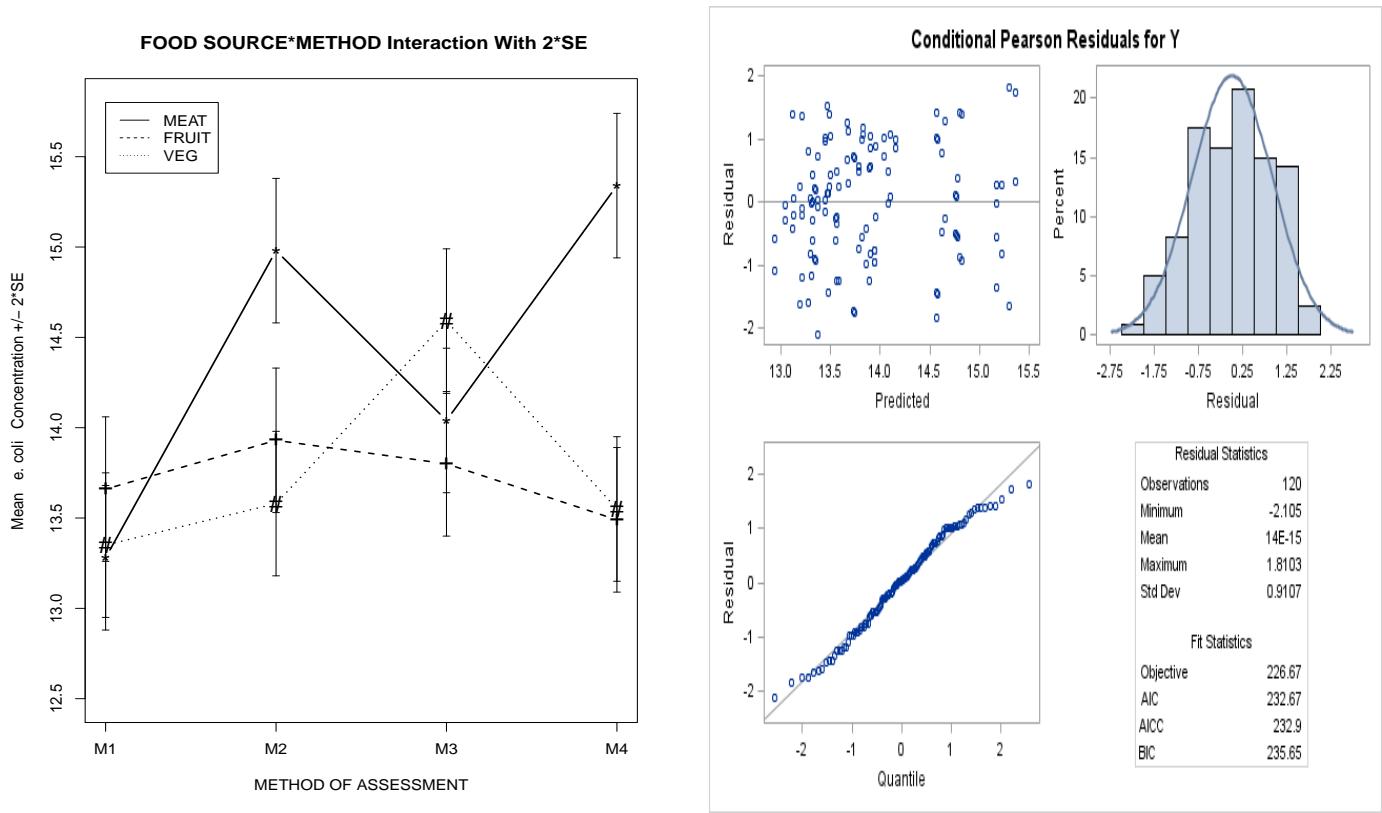
- (b) Describe (name) a procedure for testing for a difference in the proportion of depressed patients between Year 0 and Year 1. At the  $\alpha = .05$  level, conduct this test to determine if there is significant evidence of a difference in the proportions?

#### Problem IV.

The research department at a major food processor designed a study to assess the amount of salt (NaCl) in three processed foods: Meat, Fruit, and Vegetables. The accuracy of the determination of the amount of salt in foods varies considerably. The researchers wanted to evaluate the four major assessment methods Meth1, Meth2, Meth3, Meth4 and also the variation in the hundreds of laboratories which use one or more of these four methods in their determination of salt content in processed foods. Five laboratories were randomly selected from a long list of labs which used one of the four salt content methods to participate in the study. The researchers then randomly selected 120 containers of food product from the warehouse of the food processor, with 40 containers of each the 3 food sources. Six containers, two from each of the three food sources, are then sent to the 20 laboratories selected for the study. The salt content,  $Y_{ijkl}$ , is then determined by the  $k$ th Lab using Assessment Method  $j$  for the  $\ell$ th Container of Food Source  $i$  is recorded for the 120 containers. The researchers are interested in comparing the mean salt content of the four assessment methods and their differences across the food sources. Also, the researchers want to determine if there are major differences in determination of the mean salt content across the hundreds of laboratories in the USA.

AssessMethod	Lab	Food Source		
		Meat	Fruit	Vegetable
Meth1	L1	12.3, 12.6	13.2, 13.4	13.1, 12.5
	L2	13.2, 13.0	14.4, 14.5	13.4, 14.0
	L3	12.9, 13.0	12.9, 13.7	13.6, 13.0
	L4	13.2, 14.0	14.1, 14.1	13.8, 14.1
	L5	12.9, 13.9	12.8, 13.4	12.8, 13.3
Meth2	L6	14.5, 15.0	13.2, 14.2	13.5, 14.0
	L7	14.3, 15.6	14.2, 13.4	13.3, 12.6
	L8	14.5, 14.8	14.4, 14.6	13.5, 12.8
	L9	14.3, 15.6	13.3, 13.6	14.3, 13.6
	L10	14.5, 14.8	14.5, 13.8	13.5, 12.8
Meth3	L11	14.4, 14.1	13.4, 14.1	14.3, 15.1
	L12	13.5, 13.4	14.5, 14.4	13.5, 15.4
	L13	14.7, 14.6	12.7, 14.2	13.7, 15.2
	L14	13.5, 13.4	13.5, 14.4	14.5, 15.4
	L15	14.7, 14.2	12.7, 14.2	13.7, 15.2
Meth4	L16	14.8, 15.3	12.2, 13.3	12.2, 13.3
	L17	16.4, 14.3	14.4, 13.6	14.3, 13.9
	L18	15.6, 16.4	12.4, 13.8	13.6, 12.7
	L19	14.8, 15.4	13.4, 13.8	13.4, 13.9
	L20	15.2, 14.4	13.2, 13.3	14.4, 14.1

Use the following plots, the attached tables, and SAS output to answer the following questions. Pages 1-7 in the SAS output are from PROC GLM and pages 8-12 are from PROC MIXED.



2. Do the necessary conditions for testing hypotheses and constructing confidence intervals appear to be satisfied? Justify your answers.

$C_1$  Normality:

$C_2$  Equal Variance:

$C_3$  Independence:

3. Construct a partial ANOVA table for this experiment.

Include only the following: Source of Variation, df, and Expected Mean Squares.

4. At the  $\alpha = .05$  level, which Main effects and Interaction effects are significant? Justify your answer by including the relevant p-values along with their pair of degrees of freedom ( $df_{NUM}, df_{DEN}$ ). Test for Main effects even if an interaction is significant. Also, provide a test for all random effects.
5. Separate the four Assessment Methods into groups such that all four Assessment Methods in a group are not significantly different from any other member of the group with respect to their mean salt content. Use an experimentwise error rate of  $\alpha = .05$ .

6. Use the following estimates of the variance components from PROC MIXED to answer the following two questions:

$$\sigma_{LAB(METH)}^2 = .008708 \quad \sigma_{FOOD*LAB(METH)}^2 = .03236 \quad \sigma_{RESIDUAL}^2 = .3357$$

- a. Justify that the estimated standard error of the estimated mean salt content of a container of Meat which was measured using Method M1 is .2044. Hint, compute the variance of the estimated mean and use the estimated variance components to obtain the requested quantity.
  - b. Justify that the estimated standard error of the estimated difference between the mean salt content of a container of Meat and a container of Fruit is .1415. Hint, compute the variance of the estimated mean and use the estimated variance components to obtain the requested quantity.
7. The researchers state that from a health point of view two measurements of salt content have a practical difference only if the two means differ by 1.5 units. Using 95% confidence interval, is there a practical difference in the level of salt between a container of Meat and a container of Fruit using measurement Method M1?

SAS Code:

```
ods html;ods graphics on;

OPTIONS LS=90 PS=55 nocenter nodate;
TITLE 'SAS OUTPUT FOR PROBLEM IV';
DATA SALT;
INPUT FOOD $ METH $ LAB $ Y @@;
TRT=COMPRESS (METH) || COMPRESS (FOOD);
CARDS;

MEAT M1 L1 12.30 MEAT M2 L6 14.46 MEAT M3 L11 14.35 MEAT M4 L16 14.85
MEAT M1 L1 12.59 MEAT M2 L6 15.00 MEAT M3 L11 14.06 MEAT M4 L16 15.32
MEAT M1 L2 13.15 MEAT M2 L7 14.29 MEAT M3 L12 13.50 MEAT M4 L17 16.35
MEAT M1 L2 13.00 MEAT M2 L7 15.62 MEAT M3 L12 13.39 MEAT M4 L17 14.34
MEAT M1 L3 12.87 MEAT M2 L8 14.46 MEAT M3 L13 14.73 MEAT M4 L18 15.55
MEAT M1 L3 13.02 MEAT M2 L8 14.82 MEAT M3 L13 14.65 MEAT M4 L18 16.36
MEAT M1 L4 13.15 MEAT M2 L9 14.29 MEAT M3 L14 13.50 MEAT M4 L19 14.75
MEAT M1 L4 14.00 MEAT M2 L9 15.62 MEAT M3 L14 13.39 MEAT M4 L19 15.38
MEAT M1 L5 12.87 MEAT M2 L10 14.46 MEAT M3 L15 14.73 MEAT M4 L20 15.15
MEAT M1 L5 13.92 MEAT M2 L10 14.82 MEAT M3 L15 14.15 MEAT M4 L20 14.39
FRUIT M1 L1 13.20 FRUIT M2 L6 13.16 FRUIT M3 L11 13.35 FRUIT M4 L16 12.25
FRUIT M1 L1 13.39 FRUIT M2 L6 14.20 FRUIT M3 L11 14.06 FRUIT M4 L16 13.33
FRUIT M1 L2 14.45 FRUIT M2 L7 14.23 FRUIT M3 L12 14.50 FRUIT M4 L17 14.35
FRUIT M1 L2 14.50 FRUIT M2 L7 13.42 FRUIT M3 L12 14.39 FRUIT M4 L17 13.55
FRUIT M1 L3 12.86 FRUIT M2 L8 14.45 FRUIT M3 L13 12.73 FRUIT M4 L18 12.36
FRUIT M1 L3 13.72 FRUIT M2 L8 14.62 FRUIT M3 L13 14.15 FRUIT M4 L18 13.75
FRUIT M1 L4 14.11 FRUIT M2 L9 13.29 FRUIT M3 L14 13.50 FRUIT M4 L19 13.38
FRUIT M1 L4 14.10 FRUIT M2 L9 13.62 FRUIT M3 L14 14.39 FRUIT M4 L19 13.79
FRUIT M1 L5 12.83 FRUIT M2 L10 14.46 FRUIT M3 L15 12.73 FRUIT M4 L20 13.15
FRUIT M1 L5 13.42 FRUIT M2 L10 13.82 FRUIT M3 L15 14.15 FRUIT M4 L20 13.32
VEG M1 L1 13.10 VEG M2 L6 13.46 VEG M3 L11 14.35 VEG M4 L16 12.15
VEG M1 L1 12.52 VEG M2 L6 14.00 VEG M3 L11 15.06 VEG M4 L16 13.32
VEG M1 L2 13.35 VEG M2 L7 13.29 VEG M3 L12 13.50 VEG M4 L17 14.32
VEG M1 L2 14.04 VEG M2 L7 12.62 VEG M3 L12 15.39 VEG M4 L17 13.85
VEG M1 L3 13.57 VEG M2 L8 13.46 VEG M3 L13 13.73 VEG M4 L18 13.55
VEG M1 L3 12.96 VEG M2 L8 12.82 VEG M3 L13 15.15 VEG M4 L18 12.65
VEG M1 L4 13.75 VEG M2 L9 14.29 VEG M3 L14 14.50 VEG M4 L19 13.37
VEG M1 L4 14.10 VEG M2 L9 13.62 VEG M3 L14 15.39 VEG M4 L19 13.85
VEG M1 L5 12.82 VEG M2 L10 13.46 VEG M3 L15 13.73 VEG M4 L20 14.39
VEG M1 L5 13.32 VEG M2 L10 12.82 VEG M3 L15 15.15 VEG M4 L20 14.05

RUN;

PROC GLM ORDER=DATA;
CLASS FOOD METH LAB;
MODEL Y = FOOD METH FOOD*METH LAB(METH) FOOD*LAB(METH);
RANDOM LAB(METH) FOOD*LAB(METH)/TEST;
LSMEANS FOOD METH FOOD*METH/STDERR PDIFF ADJUST=TUKEY;

PROC MIXED ORDER=DATA;
CLASS FOOD METH LAB;
MODEL Y = FOOD METH FOOD*METH/RESIDUAL ;
RANDOM LAB(METH) FOOD*LAB(METH);
LSMEANS FOOD METH FOOD*METH/ADJUST=TUKEY;

ods graphics off;ods html close;
```

## SAS OUTPUT FOR PROBLEM IV

The GLM Procedure

### Class Level Information

Class Levels Values

**FOOD** 3 MEAT FRUIT VEG

**METH** 4 M1 M2 M3 M4

**LAB** 20 L1 L6 L11 L16 L2 L7 L12 L17 L3 L8 L13 L18 L4 L9 L14 L19 L5 L10 L15 L20

**Number of Observations Read** 120

**Number of Observations Used** 120

---

**SAS OUTPUT FOR PROBLEM IV**

---

The GLM Procedure

Dependent Variable: Y

<b>Source</b>	<b>DF</b>	<b>Sum of Squares</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>Model</b>	59	68.49452917	1.16092422	3.46	<.0001
<b>Error</b>	60	20.14305000	0.33571750		
<b>Corrected Total</b>	119	88.63757917			

<b>R-Square</b>	<b>Coeff Var</b>	<b>Root MSE</b>	<b>Y Mean</b>
0.772748	4.169802	0.579411	13.89542

<b>Source</b>	<b>DF</b>	<b>Type I SS</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>FOOD</b>	2	9.38181167	4.69090583	13.97	<.0001
<b>METH</b>	3	11.45262250	3.81754083	11.37	<.0001
<b>FOOD*METH</b>	6	27.60327500	4.60054583	13.70	<.0001
<b>LAB(METH)</b>	16	7.24294000	0.45268375	1.35	0.1995
<b>FOOD*LAB(METH)</b>	32	12.81388000	0.40043375	1.19	0.2734

<b>Source</b>	<b>DF</b>	<b>Type III SS</b>	<b>Mean Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
<b>FOOD</b>	2	9.38181167	4.69090583	13.97	<.0001
<b>METH</b>	3	11.45262250	3.81754083	11.37	<.0001
<b>FOOD*METH</b>	6	27.60327500	4.60054583	13.70	<.0001
<b>LAB(METH)</b>	16	7.24294000	0.45268375	1.35	0.1995
<b>FOOD*LAB(METH)</b>	32	12.81388000	0.40043375	1.19	0.2734

The GLM Procedure

Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: Y

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* FOOD	2	9.381812	4.690906	11.71	0.0002
FOOD*METH	6	27.603275	4.600546	11.49	<.0001
LAB(METH)	16	7.242940	0.452684	1.13	0.3705
Error	32	12.813880	0.400434		

Error: MS(FOOD\*LAB(METH))

\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
* METH	3	11.452623	3.817541	8.43	0.0014
Error: MS(LAB(METH))	16	7.242940	0.452684		

\* This test assumes one or more other fixed effects are zero.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FOOD*LAB(METH)	32	12.813880	0.400434	1.19	0.2734
Error: MS(Error)	60	20.143050	0.335717		

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

FOOD	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
MEAT	14.2900000	0.0916130	<.0001	1
FRUIT	13.6757500	0.0916130	<.0001	2
VEG	13.7205000	0.0916130	<.0001	3

**Least Squares Means for effect FOOD**

**Pr > |t| for H0: LSMean(i)=LSMean(j)**

**Dependent Variable: Y**

i/j	1	2	3
1		<.0001	0.0001
2	<.0001		0.9364
3	0.0001	0.9364	

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

METH	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
M1	13.3660000	0.1057856	<.0001	1
M2	14.0316667	0.1057856	<.0001	2
M3	14.1450000	0.1057856	<.0001	3
M4	14.0390000	0.1057856	<.0001	4

**Least Squares Means for effect METH**

**Pr > |t| for H0: LSMean(i)=LSMean(j)**

**Dependent Variable: Y**

i/j	1	2	3	4
1		0.0002	<.0001	0.0002
2	0.0002		0.8731	1.0000
3	<.0001	0.8731		0.8933
4	0.0002	1.0000	0.8933	

---

**SAS OUTPUT FOR PROBLEM IV**

---

Least Squares Means

Adjustment for Multiple Comparisons: Tukey

FOOD	METH	Y LSMEAN	Standard Error	Pr >  t	LSMEAN Number
MEAT	M1	13.0870000	0.1832260	<.0001	1
MEAT	M2	14.7840000	0.1832260	<.0001	2
MEAT	M3	14.0450000	0.1832260	<.0001	3
MEAT	M4	15.2440000	0.1832260	<.0001	4
FRUIT	M1	13.6580000	0.1832260	<.0001	5
FRUIT	M2	13.9270000	0.1832260	<.0001	6
FRUIT	M3	13.7950000	0.1832260	<.0001	7
FRUIT	M4	13.3230000	0.1832260	<.0001	8
VEG	M1	13.3530000	0.1832260	<.0001	9
VEG	M2	13.3840000	0.1832260	<.0001	10
VEG	M3	14.5950000	0.1832260	<.0001	11
VEG	M4	13.5500000	0.1832260	<.0001	12

Least Squares Means for effect FOOD*METH												
Pr >  t  for H0: LSMean(i)=LSMean(j)												
Dependent Variable: Y												
i/j	1	2	3	4	5	6	7	8	9	10	11	12
1	<.0001	0.021	<.0001	0.553	0.075	0.236	0.998	0.996	0.991	<.000	0.818	
		9		6	3	3	8	4	0	1	5	
2	<.0001		0.185	0.824	0.002	0.063	0.015	<.000	<.000	<.000	0.999	0.000
		4	5	9	7	4	1	1	1	8	7	
3	0.021	0.185		0.001	0.936	1.000	0.997	0.212	0.266	0.330	0.609	0.748
	9	4		1	7	0	9	2	1	2	3	2
4	<.0001	0.824	0.001		<.000	0.000	<.000	<.000	<.000	<.000	0.357	<.000
		5	1		1	2	1	1	1	1	2	1
5	0.553	0.002	0.936	<.0001		0.996	1.000	0.977	0.988	0.995	0.027	1.000
	6	9	7			1	0	0	8	4	6	0
6	0.075	0.063	1.000	0.000	0.996		1.000	0.467	0.545	0.627	0.315	0.946
	3	7	0	2	1		0	2	7	7	0	7
7	0.236	0.015	0.997	<.0001	1.000	1.000		0.799	0.858	0.907	0.109	0.998
	3	4	9		0	0		8	5	4	6	3
8	0.998	<.0001	0.212	<.0001	0.977	0.467	0.799		1.000	1.000	0.000	0.999
	8		2		0	2	8		0	0	4	1
9	0.996	<.0001	0.266	<.0001	0.988	0.545	0.858	1.000		1.000	0.000	0.999
	4		1		8	7	5	0		0	6	8
10	0.991	<.0001	0.330	<.0001	0.995	0.627	0.907	1.000	1.000		0.001	1.000
	0		2		4	7	4	0	0		0	0
11	<.0001	0.999	0.609	0.357	0.027	0.315	0.109	0.000	0.000	0.001		0.008
		8	3	2	6	0	6	4	6	0		0
12	0.818	0.000	0.748	<.0001	1.000	0.946	0.998	0.999	0.999	1.000	0.008	
	5	7	2		0	7	3	1	8	0	0	

---

**SAS OUTPUT FOR PROBLEM IV**

---

## The Mixed Procedure

**Model Information**

<b>Data Set</b>	WORK.SALT
<b>Dependent Variable</b>	Y
<b>Covariance Structure</b>	Variance Components
<b>Estimation Method</b>	REML
<b>Residual Variance Method</b>	Profile
<b>Fixed Effects SE Method</b>	Model-Based
<b>Degrees of Freedom Method</b>	Satterthwaite

**Class Level Information**

Class	Levels	Values
FOOD	3	MEAT FRUIT VEG
METH	4	M1 M2 M3 M4
LAB	20	L1 L6 L11 L16 L2 L7 L12 L17 L3 L8 L13 L18 L4 L9 L14 L19 L5 L10 L15 L20

**Number of Observations**

<b>Number of Observations Read</b>	120
<b>Number of Observations Used</b>	120
<b>Number of Observations Not Used</b>	0

**Covariance Parameter Estimates**

Cov Parm	Estimate
LAB(METH)	0.008708
FOOD*LAB(METH)	0.03236
Residual	0.3357

### Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
FOOD	2	32	11.71	0.0002
METH	3	16	8.43	0.0014
FOOD*METH	6	32	11.49	<.0001

### Least Squares Means

Effect	FOOD	METH	Estimate	Standard Error	DF	t Value	Pr >  t
FOOD	MEAT		14.2900	0.1022	47.8	139.81	<.0001
FOOD	FRUIT		13.6758	0.1022	47.8	133.80	<.0001
FOOD	VEG		13.7205	0.1022	47.8	134.24	<.0001
METH		M1	13.3660	0.1228	16	108.81	<.0001
METH		M2	14.0317	0.1228	16	114.23	<.0001
METH		M3	14.1450	0.1228	16	115.15	<.0001
METH		M4	14.0390	0.1228	16	114.29	<.0001
FOOD*METH	MEAT	M1	13.0870	0.2044	47.8	64.02	<.0001
FOOD*METH	MEAT	M2	14.7840	0.2044	47.8	72.32	<.0001
FOOD*METH	MEAT	M3	14.0450	0.2044	47.8	68.71	<.0001
FOOD*METH	MEAT	M4	15.2440	0.2044	47.8	74.57	<.0001
FOOD*METH	FRUIT	M1	13.6580	0.2044	47.8	66.82	<.0001
FOOD*METH	FRUIT	M2	13.9270	0.2044	47.8	68.13	<.0001
FOOD*METH	FRUIT	M3	13.7950	0.2044	47.8	67.49	<.0001
FOOD*METH	FRUIT	M4	13.3230	0.2044	47.8	65.18	<.0001
FOOD*METH	VEG	M1	13.3530	0.2044	47.8	65.32	<.0001
FOOD*METH	VEG	M2	13.3840	0.2044	47.8	65.48	<.0001
FOOD*METH	VEG	M3	14.5950	0.2044	47.8	71.40	<.0001
FOOD*METH	VEG	M4	13.5500	0.2044	47.8	66.29	<.0001

Differences of Least Squares Means												
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P	
FOOD	MEAT		FRUIT		0.6142	0.1415	32	4.34	0.0001	Tukey-Kramer	0.0004	
FOOD	MEAT		VEG		0.5695	0.1415	32	4.02	0.0003	Tukey-Kramer	0.0009	
FOOD	FRUIT		VEG		-0.04475	0.1415	32	-0.32	0.7539	Tukey-Kramer	0.9465	
METH		M1		M2	-0.6657	0.1737	16	-3.83	0.0015	Tukey	0.0072	
METH		M1		M3	-0.7790	0.1737	16	-4.48	0.0004	Tukey	0.0019	
METH		M1		M4	-0.6730	0.1737	16	-3.87	0.0013	Tukey	0.0066	
METH		M2		M3	-0.1133	0.1737	16	-0.65	0.5234	Tukey	0.9132	
METH		M2		M4	-0.00733	0.1737	16	-0.04	0.9669	Tukey	1.0000	
METH		M3		M4	0.1060	0.1737	16	0.61	0.5503	Tukey	0.9274	
FOOD*METH	MEAT	M1	MEAT	M2	-1.6970	0.2891	47.8	-5.87	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M1	MEAT	M3	-0.9580	0.2891	47.8	-3.31	0.0018	Tukey-Kramer	0.0799	
FOOD*METH	MEAT	M1	MEAT	M4	-2.1570	0.2891	47.8	-7.46	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M1	FRUIT	M1	-0.5710	0.2830	32	-2.02	0.0521	Tukey-Kramer	0.6790	
FOOD*METH	MEAT	M1	FRUIT	M2	-0.8400	0.2891	47.8	-2.91	0.0055	Tukey-Kramer	0.1864	
FOOD*METH	MEAT	M1	FRUIT	M3	-0.7080	0.2891	47.8	-2.45	0.0180	Tukey-Kramer	0.4047	
FOOD*METH	MEAT	M1	FRUIT	M4	-0.2360	0.2891	47.8	-0.82	0.4183	Tukey-Kramer	0.9994	
FOOD*METH	MEAT	M1	VEG	M1	-0.2660	0.2830	32	-0.94	0.3543	Tukey-Kramer	0.9980	
FOOD*METH	MEAT	M1	VEG	M2	-0.2970	0.2891	47.8	-1.03	0.3094	Tukey-Kramer	0.9957	
FOOD*METH	MEAT	M1	VEG	M3	-1.5080	0.2891	47.8	-5.22	<.0001	Tukey-Kramer	0.0006	
FOOD*METH	MEAT	M1	VEG	M4	-0.4630	0.2891	47.8	-1.60	0.1158	Tukey-Kramer	0.8960	
FOOD*METH	MEAT	M2	MEAT	M3	0.7390	0.2891	47.8	2.56	0.0138	Tukey-Kramer	0.3439	
FOOD*METH	MEAT	M2	MEAT	M4	-0.4600	0.2891	47.8	-1.59	0.1181	Tukey-Kramer	0.8999	
FOOD*METH	MEAT	M2	FRUIT	M1	1.1260	0.2891	47.8	3.90	0.0003	Tukey-Kramer	0.0200	
FOOD*METH	MEAT	M2	FRUIT	M2	0.8570	0.2830	32	3.03	0.0048	Tukey-Kramer	0.1465	
FOOD*METH	MEAT	M2	FRUIT	M3	0.9890	0.2891	47.8	3.42	0.0013	Tukey-Kramer	0.0627	
FOOD*METH	MEAT	M2	FRUIT	M4	1.4610	0.2891	47.8	5.05	<.0001	Tukey-Kramer	0.0009	
FOOD*METH	MEAT	M2	VEG	M1	1.4310	0.2891	47.8	4.95	<.0001	Tukey-Kramer	0.0012	
FOOD*METH	MEAT	M2	VEG	M2	1.4000	0.2830	32	4.95	<.0001	Tukey-Kramer	0.0012	
FOOD*METH	MEAT	M2	VEG	M3	0.1890	0.2891	47.8	0.65	0.5164	Tukey-Kramer	0.9999	
FOOD*METH	MEAT	M2	VEG	M4	1.2340	0.2891	47.8	4.27	<.0001	Tukey-Kramer	0.0076	

Differences of Least Squares Means												
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P	
FOOD*METH	MEAT	M3	MEAT	M4	-1.1990	0.2891	47.8	-4.15	0.0001	Tukey-Kramer	0.0104	
FOOD*METH	MEAT	M3	FRUIT	M1	0.3870	0.2891	47.8	1.34	0.1870	Tukey-Kramer	0.9670	
FOOD*METH	MEAT	M3	FRUIT	M2	0.1180	0.2891	47.8	0.41	0.6850	Tukey-Kramer	1.0000	
FOOD*METH	MEAT	M3	FRUIT	M3	0.2500	0.2830	32	0.88	0.3836	Tukey-Kramer	0.9989	
FOOD*METH	MEAT	M3	FRUIT	M4	0.7220	0.2891	47.8	2.50	0.0160	Tukey-Kramer	0.3766	
FOOD*METH	MEAT	M3	VEG	M1	0.6920	0.2891	47.8	2.39	0.0206	Tukey-Kramer	0.4379	
FOOD*METH	MEAT	M3	VEG	M2	0.6610	0.2891	47.8	2.29	0.0267	Tukey-Kramer	0.5052	
FOOD*METH	MEAT	M3	VEG	M3	-0.5500	0.2830	32	-1.94	0.0608	Tukey-Kramer	0.7249	
FOOD*METH	MEAT	M3	VEG	M4	0.4950	0.2891	47.8	1.71	0.0933	Tukey-Kramer	0.8496	
FOOD*METH	MEAT	M4	FRUIT	M1	1.5860	0.2891	47.8	5.49	<.0001	Tukey-Kramer	0.0003	
FOOD*METH	MEAT	M4	FRUIT	M2	1.3170	0.2891	47.8	4.56	<.0001	Tukey-Kramer	0.0035	
FOOD*METH	MEAT	M4	FRUIT	M3	1.4490	0.2891	47.8	5.01	<.0001	Tukey-Kramer	0.0010	
FOOD*METH	MEAT	M4	FRUIT	M4	1.9210	0.2830	32	6.79	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M1	1.8910	0.2891	47.8	6.54	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M2	1.8600	0.2891	47.8	6.43	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	MEAT	M4	VEG	M3	0.6490	0.2891	47.8	2.25	0.0294	Tukey-Kramer	0.5319	
FOOD*METH	MEAT	M4	VEG	M4	1.6940	0.2830	32	5.99	<.0001	Tukey-Kramer	<.0001	
FOOD*METH	FRUIT	M1	FRUIT	M2	-0.2690	0.2891	47.8	-0.93	0.3568	Tukey-Kramer	0.9982	
FOOD*METH	FRUIT	M1	FRUIT	M3	-0.1370	0.2891	47.8	-0.47	0.6377	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M1	FRUIT	M4	0.3350	0.2891	47.8	1.16	0.2523	Tukey-Kramer	0.9887	
FOOD*METH	FRUIT	M1	VEG	M1	0.3050	0.2830	32	1.08	0.2892	Tukey-Kramer	0.9937	
FOOD*METH	FRUIT	M1	VEG	M2	0.2740	0.2891	47.8	0.95	0.3480	Tukey-Kramer	0.9979	
FOOD*METH	FRUIT	M1	VEG	M3	-0.9370	0.2891	47.8	-3.24	0.0022	Tukey-Kramer	0.0937	
FOOD*METH	FRUIT	M1	VEG	M4	0.1080	0.2891	47.8	0.37	0.7104	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M2	FRUIT	M3	0.1320	0.2891	47.8	0.46	0.6500	Tukey-Kramer	1.0000	
FOOD*METH	FRUIT	M2	FRUIT	M4	0.6040	0.2891	47.8	2.09	0.0420	Tukey-Kramer	0.6332	
FOOD*METH	FRUIT	M2	VEG	M1	0.5740	0.2891	47.8	1.99	0.0528	Tukey-Kramer	0.6991	
FOOD*METH	FRUIT	M2	VEG	M2	0.5430	0.2830	32	1.92	0.0640	Tukey-Kramer	0.7397	
FOOD*METH	FRUIT	M2	VEG	M3	-0.6680	0.2891	47.8	-2.31	0.0252	Tukey-Kramer	0.4897	
FOOD*METH	FRUIT	M2	VEG	M4	0.3770	0.2891	47.8	1.30	0.1984	Tukey-Kramer	0.9725	

Differences of Least Squares Means													
Effect	FOOD	METH	_FOOD	_METH	Estimate	Standard Error	DF	t Value	Pr >  t	Adjustment	Adj P		
FOOD*METH	FRUIT	M3	FRUIT	M4	0.4720	0.2891	47.8	1.63	0.1091	Tukey-Kramer	0.8840		
FOOD*METH	FRUIT	M3	VEG	M1	0.4420	0.2891	47.8	1.53	0.1329	Tukey-Kramer	0.9211		
FOOD*METH	FRUIT	M3	VEG	M2	0.4110	0.2891	47.8	1.42	0.1616	Tukey-Kramer	0.9503		
FOOD*METH	FRUIT	M3	VEG	M3	-0.8000	0.2830	32	-2.83	0.0080	Tukey-Kramer	0.2162		
FOOD*METH	FRUIT	M3	VEG	M4	0.2450	0.2891	47.8	0.85	0.4009	Tukey-Kramer	0.9992		
FOOD*METH	FRUIT	M4	VEG	M1	-0.03000	0.2891	47.8	-0.10	0.9178	Tukey-Kramer	1.0000		
FOOD*METH	FRUIT	M4	VEG	M2	-0.06100	0.2891	47.8	-0.21	0.8338	Tukey-Kramer	1.0000		
FOOD*METH	FRUIT	M4	VEG	M3	-1.2720	0.2891	47.8	-4.40	<.0001	Tukey-Kramer	0.0054		
FOOD*METH	FRUIT	M4	VEG	M4	-0.2270	0.2830	32	-0.80	0.4284	Tukey-Kramer	0.9995		
FOOD*METH	VEG	M1	VEG	M2	-0.03100	0.2891	47.8	-0.11	0.9151	Tukey-Kramer	1.0000		
FOOD*METH	VEG	M1	VEG	M3	-1.2420	0.2891	47.8	-4.30	<.0001	Tukey-Kramer	0.0071		
FOOD*METH	VEG	M1	VEG	M4	-0.1970	0.2891	47.8	-0.68	0.4989	Tukey-Kramer	0.9999		
FOOD*METH	VEG	M2	VEG	M3	-1.2110	0.2891	47.8	-4.19	0.0001	Tukey-Kramer	0.0094		
FOOD*METH	VEG	M2	VEG	M4	-0.1660	0.2891	47.8	-0.57	0.5685	Tukey-Kramer	1.0000		
FOOD*METH	VEG	M3	VEG	M4	1.0450	0.2891	47.8	3.61	0.0007	Tukey-Kramer	0.0398		

**Table A.1** Cumulative Binomial Probabilities (*cont.*)

e.  $n = 25$

	<i>p</i>														
	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	.778	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.974	.642	.271	.027	.007	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.998	.873	.537	.098	.032	.009	.000	.000	.000	.000	.000	.000	.000	.000	.000
3	1.000	.966	.764	.234	.096	.033	.000	.000	.000	.000	.000	.000	.000	.000	.000
4	1.000	.993	.902	.421	.214	.090	.009	.000	.000	.000	.000	.000	.000	.000	.000
5	1.000	.999	.967	.617	.378	.193	.029	.002	.000	.000	.000	.000	.000	.000	.000
6	1.000	1.000	.991	.780	.561	.341	.074	.007	.000	.000	.000	.000	.000	.000	.000
7	1.000	1.000	.998	.891	.727	.512	.154	.022	.001	.000	.000	.000	.000	.000	.000
8	1.000	1.000	1.000	.953	.851	.677	.274	.054	.004	.000	.000	.000	.000	.000	.000
9	1.000	1.000	1.000	.983	.929	.811	.425	.115	.013	.000	.000	.000	.000	.000	.000
10	1.000	1.000	1.000	.994	.970	.902	.586	.212	.034	.002	.000	.000	.000	.000	.000
11	1.000	1.000	1.000	.998	.980	.956	.732	.345	.078	.006	.001	.000	.000	.000	.000
x 12	1.000	1.000	1.000	1.000	.997	.983	.846	.500	.154	.017	.003	.000	.000	.000	.000
13	1.000	1.000	1.000	1.000	.999	.994	.922	.655	.268	.044	.020	.002	.000	.000	.000
14	1.000	1.000	1.000	1.000	.998	.966	.788	.414	.098	.030	.006	.000	.000	.000	.000
15	1.000	1.000	1.000	1.000	1.000	.987	.885	.575	.189	.071	.017	.000	.000	.000	.000
16	1.000	1.000	1.000	1.000	1.000	.996	.946	.726	.323	.149	.047	.000	.000	.000	.000
17	1.000	1.000	1.000	1.000	1.000	.999	.978	.846	.488	.273	.109	.002	.000	.000	.000
18	1.000	1.000	1.000	1.000	1.000	1.000	.993	.926	.659	.439	.220	.009	.000	.000	.000
19	1.000	1.000	1.000	1.000	1.000	1.000	.998	.971	.807	.622	.383	.033	.001	.000	.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.910	.786	.579	.098	.007	.000	.000
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.967	.904	.766	.236	.034	.000	.000
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.968	.902	.463	.127	.002	.000
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.993	.973	.729	.358	.026	.000
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.996	.928	.723	.222	.000

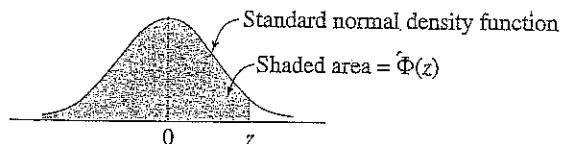
SOURCE: Adapted from L. L. Chao (1980), *Statistics for Management*, Wadsworth, Inc.

**Table A.2** Cumulative Poisson Probabilities (*cont.*)

	$\lambda$											
	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	15.0	20.0	
0	.135	.050	.018	.007	.002	.001	.000	.000	.000	.000	.000	
1	.406	.199	.092	.040	.017	.007	.003	.001	.000	.000	.000	
2	.677	.423	.238	.125	.062	.030	.014	.006	.003	.000	.000	
3	.857	.647	.433	.265	.151	.082	.042	.021	.010	.000	.000	
4	.947	.815	.629	.440	.285	.173	.100	.055	.029	.001	.000	
5	.983	.916	.785	.616	.446	.301	.191	.116	.067	.003	.000	
6	.995	.966	.889	.762	.606	.450	.313	.207	.130	.008	.000	
7	.999	.988	.949	.867	.744	.599	.453	.324	.220	.018	.001	
8	1.000	.996	.979	.932	.847	.729	.593	.456	.333	.037	.002	
9		.999	.992	.968	.916	.830	.717	.587	.458	.370	.005	
10	1.000	.997	.986	.957	.901	.816	.706	.583	.418	.211		
11		.999	.995	.980	.947	.888	.803	.697	.585	.421		
12		1.000	.998	.991	.973	.936	.876	.792	.668	.539		
13			.999	.996	.987	.966	.926	.864	.763	.666		
14				1.000	.999	.994	.983	.959	.917	.866	.705	
15					.999	.998	.992	.978	.951	.868	.757	
16					1.000	.999	.996	.989	.973	.864	.721	
17						.999	.998	.995	.986	.949	.797	
18						1.000	.999	.998	.993	.819	.381	
19							1.000	.998	.997	.875	.470	
20								1.000	.998	.917	.559	
21									.999	.947	.644	
22									1.000	.967	.721	
23										.981	.787	
24										.989	.843	
25										.994	.888	
26										.997	.922	
27										.998	.948	
28										.999	.966	
29										1.000	.978	
30											.987	
31											.992	
32											.995	
33											.997	
34											.999	
35											.999	
36											1.000	

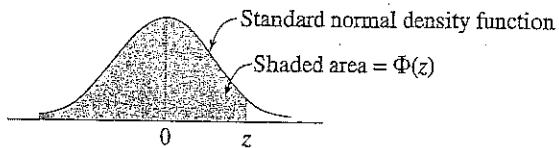
SOURCE: L. L. Chao (1974), *Statistics: Methods and Analysis*, 2nd ed. New York: McGraw-Hill.

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$

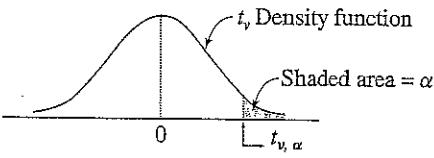


$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0394	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

**Table A.3** Standard Normal Curve Areas  $\Phi(z) = P(Z \leq z)$  (*cont.*)

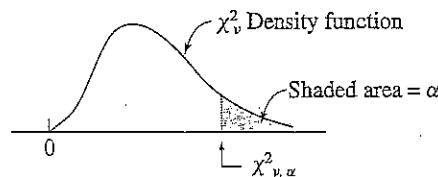


**Table A.4** Critical Values  $t_{v,\alpha}$  for the  $t$ -Distribution



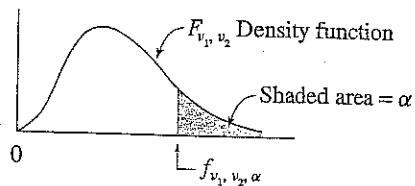
$v$	$\alpha$							
	.10	.05	.025	.01	.005	.001	.0005	
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62	
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598	
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924	
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869	
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140	
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073	
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015	
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965	
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922	
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883	
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850	
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819	
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792	
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767	
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745	
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725	
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707	
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690	
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674	
29	1.311	1.699	2.045	1.462	2.756	3.396	3.659	
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646	
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551	
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460	
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373	
$\infty$	1.282	1.645	1.960	2.326	2.576	3.090	3.291	

**Table A.5** Critical Values  $\chi_{v,\alpha}^2$  for the Chi-square Distribution



v	$\alpha$									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	Z5.695	28.196	50.660	54.572	58.119	62.420	65.473
40*	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

**Table A.6** Critical Values  $f_{v_1, v_2, \alpha}$  for the  $F$ -Distribution ( $\alpha = .05$ ) (cont.)



Degrees of freedom for the denominator ( $v_2$ )	Degrees of freedom for the numerator ( $v_1$ )																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.49	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.69	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.11	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.01	1.96	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.81	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.91	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.59	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.66	1.61	1.55	1.43	1.35	1.25
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.81	1.75	1.66	1.61	1.55	1.43	1.35	1.29	1.20
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00