Applied Statistics Comprehensive Exam

January 2024

Ph.D Day I - Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:__________

1. A company wanted to replace machines used to make a certain component in one of its factories. Three different brands of machines were available, so the management designed an experiment to evaluate the productivity of the three machines when operated by the company's own personnel. Six employees (persons) were randomly selected from the population of employees that are trained to operate such machines. Each selected employee was required to operate each machine during three different shifts. The data recorded were overall productivity scores that took into account the number and quality of components produced. A two-factor ANOVA was performed and the following partially completed ANOVA table is available. Using the appropriate model for the two-factor ANOVA, answer the following questions:

Source	SS	df	MS	F
Machine	1755.263			
Person	1241.895			
Person $\times$ Machine	426.530			
Error				
Total	3456.975	53		

   i. Complete the ANOVA table.
   ii. Using $\alpha = 0.01$, what conclusion will you reach about the interaction between Machine and Person? State the appropriate null and alternative hypothesis.
   iii. Using $\alpha = 0.01$, what conclusion should be reached about the results from main effect of Person? State the appropriate null and alternative hypothesis.

---

2. A report summarizing the results of a study on the relationship between a verbal aptitude test x and a mathematics aptitude test y states that $r = 0.60$ with a sample size of $n = 35$.

   i. Test $\rho > .70$.
   ii. Calculate a 95% confidence interval for the correlation coefficient ρ.

---

3. In a study, a researcher analyzed the Adaptive Behavior Scale (ABS) values of mentally challenged institutionalized people. There are four groups of individuals based on their mental ages, and the ABS values for three periods were recorded every 6 months.

   Using the SAS (PAGE 5) output answer the following questions. In each case, state the null and alternative hypotheses, the p-value, and conclusion, and use $\alpha = .05$.

i. Test the hypothesis of no differences in ABS values from group to group.

ii. Test the hypothesis of no differences in ABS values from period to period.

iii. Does the effect of group changes from period to period?

---

4. Consider the Two-Factor ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \quad i = 1, 2, \ j = 1, 2, \ k = 1, 2.$$

(a) Write an appropriate general linear model equation in the form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ by giving each term $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ explicitly.

i. Find a basis of linear functions that describes *all* estimable functions for this problem.

ii. Determine whether $\mu + \alpha_1 - \beta_2$ is estimable. Then Find the Best Linear Unbiased Estimator (BLUE) for it if it is estimable.

---

5. Consider the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

i. Prove that $\mathbf{Y}^T\left(\frac{\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ and $\mathbf{Y}^T\left(\frac{\mathbf{I}-\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ are both distributed as χ^2 random variables. Give the degrees of freedom associated with each.

ii. Prove that $\mathbf{Y}^T\left(\frac{\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ is distributed independently of $\mathbf{Y}^T\left(\frac{\mathbf{I}-\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$.

iii. Find the distribution of

$$\frac{\mathbf{Y}^T(\mathbf{P}_x)\mathbf{Y}/rank(\mathbf{X})}{\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_x)\mathbf{Y}/(n - rank(\mathbf{X}))}.$$

---

6. Let $\boldsymbol{A}$ be an $m \times n$ matrix. Show $\mathcal{N}(\boldsymbol{A})$ (Null space of $\boldsymbol{A}$) is a subspace of $\mathcal{R}^n$.

---

7. Let $X_1, \ldots, X_n$ be a random sample from density function

$$f(x, \theta) = (2x/\theta^2)I_{(0,\theta)}(x), \ \theta > 0.$$

i. Find a MLE for θ.

ii. Find the asymptotic distribution of the MLE of θ.

iii. Find a 95% confidence interval for θ.

---

8. Let $X_1, \ldots, X_n$ be a random sample from the density

$$f(x;\theta) = e^{-(x-\theta)} \exp\left(-e^{-(x-\theta)}\right),$$

where $-\infty < \theta < \infty$.

   i. Find the Cramér-Rao lower bound for unbiased estimators of θ.

   ii. Is there a function of θ for which there exists an unbiased estimator, the variance of which coincides with the Cramér-Rao lower bound? If so, find it.

   iii. Find the UMVUE of θ.

---

9. Let $X_1, \ldots, X_m$ be a random sample from the density $\theta_1 x^{\theta_1 - 1} I_{(0,1)}(x)$, and let $Y_1, \ldots, Y_n$ be a random sample from the density $\theta_2 y^{\theta_2 - 1} I_{(0,1)}(y)$. Assume that the samples are independent. Set $U_t = -\log_e X_i, \quad i = 1, \ldots, m$, and $V_j = -\log_e Y_j, j = 1, \ldots, n$

   i. Find the generalized likelihood-ratio for testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, and show that the generalized likelihood-ratio test can be expressed in terms of the statistic

$$T = \frac{\sum U_i}{\sum U_i + \sum V_j}.$$

   ii. If H_0 is true, what is the distribution of T? (You do not have to derive it if you know the answer.) Does the distribution of T depend on $\theta = \theta_1 = \theta_2$ given that $H_\circ$ is true?

---

10. Let X be a single observation from the density

$$f(x;\theta) = (1+\theta) x^{\theta} I_{(0,1)}(x),$$

where $\theta > -1$.

   i. Find the most powerful size-α test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$.

   ii. Is there uniformly most powerful size-α test of $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$? If so, what is it?

SAS output for Question 3

Sphericity Tests				
Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	2	0.2172023	10.688481	0.0048
Orthogonal Components	2	0.4992139	4.8630437	0.0879

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	3	38419.22222	12806.40741	5.88	0.0202
Error	8	17422.66667	2177.83333		

The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H-F-L
Period	2	4629.05556	2314.52778	3.61	0.0508	0.0767	0.0692
Period*Group	6	1670.94444	278.49074	0.43	0.8453	0.7812	0.8001
Error(Period)	16	10260.00000	641.25000				

Greenhouse-Geisser Epsilon	0.6663
Huynh-Feldt-Lecoutre Epsilon	0.7495

Applied Statistics Comprehensive Exam

August 2023

Ph.D Day I - Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. For each of the following scenarios, write the type of study design being implemented and the statistical method to be used.

   i. An exercise physiologist structured three types of exercise programs (EPRO) and conducted an experiment to evaluate and compare the effectiveness of each program. The experiment consisted of subjecting an individual to a given exercise program for 8 weeks. At the end of the training program, each individual ran for 6 min after which their heart rate was measured. An exercise program is deemed to be more effective if individuals on that program have lower heart rates after the 6-min run than individuals on another exercise program. Since individuals entered the experiment at differing degrees of fitness, the resting heart rate before beginning training was recorded.

   ii. A hospital is implementing a program to improve service quality and productivity. As part of this program the hospital management is attempting to measure and evaluate patient satisfaction. On a random sample of 25 recently discharged patients they collected data on the satisfaction (a subjective measure on an increasing scale), patient age, severity (an index measuring the severity of the patient's illness), an indicator of whether the patient is a surgical or medical patient, and an index measuring the patient's anxiety level.

   iii. To test the efficiency of its new programmable calculator, a computer company selected at random six engineers who were proficient in the use of both this calculator and an earlier model and asked them to work out two problems on both calculators. One of the problems was statistical in nature, the other was an engineering problem. The order of the four calculations was randomized independently for each engineer. The length of time (in minutes) required to solve each problem was observed.

   iv. A production engineer studied the effects of machine model and operator on the output in a bottling plant. Three bottling machines were used, each a different model. Twelve operators were randomly selected. Four operators were assigned to a machine and worked six-hour shifts each. Data on the number of cases produced by each machine and operator were collected for a week.

---

2. The Elves Toy Co. makes toy trains. For quality control purposes, toy trains coming off the production line are regularly inspected for defects; defective trains are always thrown away. Once every week, five trains are randomly selected from the production line. The probability that a train has a defect and is thrown away is 0.25. The table below shows the number of weeks in which $0, 1, ..., 5$ trains were defective and thus thrown away in 2007.

No. of trains thrown away	No. of weeks
0	10
1	21
2	14
3	6
4	1
5	0

Propose an appropriate probability distribution for the number of defective trains, and test to see if it is consistent with the data observed.

---

3. Dental fillings made with gold can vary in hardness depending on how the metal is treated prior to its placement in the tooth. Two factors are thought to influence the hardness: the gold alloy and the condensation method. In addition, some dentists doing the work are better at some types of fillings than others. Five dentists were selected at random. Each dentist prepares 24 fillings (in random order), one for each of the combinations of method (three levels) and alloy (eight levels). The fillings were then measured for hardness using the Diamond Pyramid Hardness Number (big scores are better).

   i. What type of analysis should be used here? Write the model and specify all the components. Using the SAS (PAGE 5) output answer the following questions. In each case, state the null and alternative hypotheses, the $p-$value, and conclusion, and use $\alpha = .05$.

   ii. Test the effect of dentists on the response. cost on likelihood of purchase.

   iii. Test the effect of alloy.

   iv. Is the method effect the same for all dentists?

---

4. Consider the General Linear Mixed Model (GLMM)

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{R})$, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$, and $\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{u}) = \boldsymbol{0}$.

   i. Find the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{u}$.
   ii. Find the distribution of $\boldsymbol{Y}$.
   iii. Find the distribution of $\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{u} \end{bmatrix}$.
   iv. Find the best unbiased predictor for $\boldsymbol{u}$.

---

5. Consider the linear model for one-way ANOVA

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, k, j = 1, \cdots, n$$

Show that if a linear function of the treatment parameters $\sum_{i=1}^{k} c_i \alpha_i$ is estimable then $\sum_{i}^{k} c_i = 0$.

---

6. Consider the General Linear Model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ with $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.

   i. Under which condition(s) will the Least Squares Estimator $\hat{\boldsymbol{\beta}}$ be a unique vector?
   ii. Provide a definition of estimability for any linear combination of parameters, $\boldsymbol{c}^T\boldsymbol{\beta}$ and explain why estimability is important for Least Squares Estimation.
   iii. Prove

$$\frac{\boldsymbol{Y}^T \boldsymbol{P_X} \boldsymbol{Y} / rank(\boldsymbol{X})}{\boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{P_X}) \boldsymbol{Y} / (n - rank(\boldsymbol{X}))}$$

   has an F distribution.

---

7. i. Assume that X and Y are independent random variables and X has binomial distribution with parameters 3 ans 1/4 and Y has binomial distribution with parameters 2 and 1/3. Find $P(X = Y)$.

   ii. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. What is the distribution of

$$\frac{1}{\sigma^2}\left[(k-1)S_k^2 + (n-k-1)S_{n-k}^2\right],$$

   where

$$\bar{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i \quad , \quad \bar{X}_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} X_i,$$

$$S_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \bar{X}_k)^2, \text{ and } S_{n-k}^2 = \frac{1}{n-k-1}\sum_{i=k+1}^{n}(X_i - \bar{X}_{n-k})^2?$$

---

8. Let $X_1, \ldots, X_n$ be a random sample from the discrete density function

$$f(x, \theta) = (1/\theta) I_{\{1,2,\ldots,\theta\}}(x),$$

   where $\theta \in \Theta = \{1, 2, \ldots\}$, the set of positive integers.

   i. Find a UMVUE for θ. Justify your answer.

   ii. Find a 95% confidence interval for θ.

---

9. Let $X_1, \ldots, X_m$ be a random sample from $N(\mu_1, 4)$ and $Y_1, \ldots, Y_n$ be a random sample from $N(\mu_2, 2)$.

   i. Find the GLRT of size $\alpha = .01$ for $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$

   ii. Find the power function of the test as a function of $\theta = \mu_1 - \mu_2$.

---

10. Let $\ldots, X_n$ be a random sample from

$$f(x; \theta, \alpha) = \frac{\theta^n x^{\alpha-1}}{\Gamma(\alpha)} e^{-x\theta};\ x > 0, \alpha > 0, \theta > 0.$$

   Suppose θ is known.

   i. Find an MLE for parameter α.

   ii. Find a UMP test of size γ for $H_0 : \alpha \leq \alpha_0$ vs $H_a : \alpha > \alpha_0$.

Applied Statistics Comprehensive Exam

January 2021

Ph.D Day I - Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:__________

1. Briefly describe the experimental design you would choose for each of the following situations, and explain why.

   (a) An investigative group at a television station wishes to determine if doctors treat patients on public assistance differently from those with private insurance. They measure this by how long the doctor spends with the patient. There are four large clinics in the city, and the station chooses three pediatricians at random from each of the four clinics. Ninety-six families on public assistance are located and divided into four groups of 24 at random. All 96 families have a one-year-old child and a child just entering school. Half the families will request a one-year checkup, and the others will request a preschool checkup. Half the families will be given temporary private insurance for the study, and the others will use public assistance. The four groupings of families are the factorial combinations of checkup type and insurance type. Each group of 24 is now divided at random into twelve sets of two, with each set of two assigned to one of the twelve selected doctors. Thus each doctor will see eight patients from the investigation. Recap: 96 units (families); the response is how long the doctor spends with each family.

   (b) An experiment was conducted to study the effects of irrigation, crop variety, and aerially sprayed pesticide on grain yield. There were two replicates. Within each replicate, three fields were chosen and randomly assigned to be sprayed with one of the pesticides. Each field was then divided into two east- west strips; one of these strips was chosen at random to be irrigated, and the other was left unirrigated. Each east-west strip was split into north-south plots, and the two varieties were randomly assigned to plots.

   (c) A company has 50 machines that make cardboard cartons for canned goods, and they want to understand the variation in strength of the cartons. They choose ten machines at random from the 50 and make 40 cartons on each machine, assigning 400 lots of feedstock cardboard at random to the ten chosen machines. The resulting cartons are tested for strength.

---

2. When the 2000 GSS asked whether human beings developed from earlier species of animals (variable SCITEST4), 53.8% of 1095 respondents answered that this was probably or definitely not true.

   (a) Can you conclude that a majority of Americans felt this way? Use $\alpha = 0.05$. Find the P-value for this test.

   (b) Find a 99% confidence interval for the corresponding population proportion.

---

3. In a study, 34 male albino rats, each weighing approximately 200 grams, were randomly divided into 17 pairs. One animal in each pair was selected at random to receive an experimental diet containing ethionine, whereas the other was a pair-fed control (that is, the control animal received the same amount of food as was eaten by the corresponding treated animal). After seven days, the 34 rats were sacrificed, and the liver of each animal was extracted and divided into three parts. The 17 pairs were then randomized to one of two groups. In the eight pairs randomized to group 1, the liver thirds from each animal were randomly assigned to be treated with radioactive iron in a solution of low pH (2.0–3.0), medium pH (4.5–5.5), or high pH (7.0–7.7) at a temperature of 37○C. The same procedure was followed in the nine pairs randomized to group 2, with the exception that the liver portions were treated at 25○C. The response variable of interest is the amount of iron absorbed by the variously treated liver thirds, which can be assumed to be continuous and normally distributed. Using the SAS output (PAGE 6) answer the following questions. Use $\alpha = .05$ for all the tests.

   (a) What type of design is used here? Write the model for your design and specify all the components.

   (b) Is there a significant treatment effect? State the null and alternative hypotheses, the p-value, and conclusion.

   (c) Is there a significant Ph effect? State the null and alternative hypotheses, the p-value, and conclusion.

   (d) The effect of pH is the same in temperature 37° as in temperature 25°? State the null and alternative hypotheses, the p-value, and conclusion.

   (e) Test whether or not Ph effect on rats in two temperatures equally for two treatments? State the null and alternative hypotheses, the p-value, and conclusion.

---

4. Consider the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ with $\sigma^2 > 0$ a positive constant.

   (a) Under which condition(s) will the Least Squares Estimator $\hat{\boldsymbol{\beta}}$ be a *unique* vector?

   (b) Provide a definition of *estimability* for any linear combination of parameters, $\mathbf{c}^T\boldsymbol{\beta}$, and explain why estimability is important for Least Squares Estimation.

   (c) Assuming $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are *two different* Least Squares Estimators for $\boldsymbol{\beta}$, and assuming $\mathbf{c}^T\boldsymbol{\beta}$ is estimable, find an expression for:

$$Cov(\mathbf{c}^T\hat{\boldsymbol{\beta}}_1, \mathbf{c}^T\hat{\boldsymbol{\beta}}_2).$$

---

5. Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

Let $\boldsymbol{C}'\boldsymbol{\beta} = \boldsymbol{0}$ be testable and assume $\boldsymbol{C}$ is full rank. Show $H_0 : \ \boldsymbol{C}'\boldsymbol{\beta} = \boldsymbol{0}$ can be written as a form of full and reduced model then prov the test statistic reduces to

$$F = \frac{\hat{\boldsymbol{\beta}}'_{LS}\boldsymbol{C}\left[\boldsymbol{C}'\boldsymbol{G}\boldsymbol{C}\right]^{-1}\boldsymbol{C}'\hat{\boldsymbol{\beta}}_{LS}/rank(\boldsymbol{C})}{SSE/(n-r)},$$

where $\boldsymbol{G} = (\boldsymbol{X}'\boldsymbol{X})^{-}$, and $r = rank(\boldsymbol{X})$. Note that since $\boldsymbol{C}$ is full rank $\boldsymbol{C}'\boldsymbol{G}\boldsymbol{C}$ is nonsingular.

---

6. Consider the linear model

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

where $\boldsymbol{X}$ is an $n \times k$ matrix.

(a) Find the orthogonal projection matrix $\boldsymbol{P_X}$.

(b) Let $\boldsymbol{A}$ be an $n \times n$ orthogonal matrix. Show that the matrix $\boldsymbol{A}'\boldsymbol{P_X}\boldsymbol{A}$ is an orthogonal projection.

---

7. Consider a random sample from a Poisson distribution, $X_i \sim \text{POI}(\mu)$. Let $\hat{\mu}_n$ be the MLE for μ.

(a) Show that $Y_n = e^{\hat{\mu}_n}$ converges in probability to $P[X = 0] = e^{-\mu}$

(b) Find the asymptotic normal distribution of Y_n.

---

8. Assume $X_1, \cdots, X_n$ iid $\sim N(\theta, \theta), \theta > 0$. Give an example of a pivotal quantity, and use it to obtain a confidence-interval estimator of θ.

---

9. Let $X_1, \cdots, X_n$ be a random sample from the Bernoulli distribution, say

$$P[X = 1] = \theta = 1 - P[X = 0].$$

   (a) Find the Cramer-Rao lower bound for the variance of unbiased estimators of $\theta(1-\theta)$.

   (b) Find the UMVUE of $\theta(1-\theta)$ if such exists.

---

10. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ.

   (a) Assume all the parameters are unknow and derive the generalized likelihood-ratio test for testing $H_0 : \rho = 0$.

   (b) Find the asymptotic distribution of the test in (a).

SAS output for Question 3

The GLM Procedure
Repeated Measures Analysis of Variance

Sphericity Tests				
Variables	**DF**	**Mauchly's Criterion**	**Chi-Square**	**Pr > ChiSq**
Transformed Variates	2	0.7150262	9.7276459	0.0077
Orthogonal Components	2	0.9945524	0.1584136	0.9238

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treatment	1	26.5264762	26.5264762	7.19	0.0118
temperature	1	0.7615669	0.7615669	0.21	0.6529
treatment*temperatur	1	0.3787115	0.3787115	0.10	0.7509
Error	30	110.7114314	3.6903810		

The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	**DF**	**Type III SS**	**Mean Square**	**F Value**	**Pr > F**	**G - G**	**H-F-L**
Ph	2	120.9305348	60.4652674	22.46	<.0001	<.0001	<.0001
Ph*treatment	2	2.5610637	1.2805319	0.48	0.6238	0.6228	0.6238
Ph*temperature	2	38.0933270	19.0466635	7.07	0.0017	0.0018	0.0017
Ph*treatment*temperatur	2	6.7900049	3.3950025	1.26	0.2908	0.2907	0.2908
Error(Ph)	60	161.5516171	2.6925270				

Greenhouse-Geisser Epsilon	0.9946
Huynh-Feldt-Lecoutre Epsilon	1.0650

Applied Statistics Comprehensive Exam

August 2020

Ph.D Day I - Exam

This comprehensive exam consists of 10 questions pertaining to methodological and applied statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. In a bumper test, three types of autos were deliberately crashed into a barrier at 5 mph, and the resulting damage (in dollars) was estimated. Five test vehicles of each type were crashed, with the summary results shown below.

TYPE	N	mean	variance
Goliath	5	1282	246320
Varmint	5	1376	103580
Weasel	5	1638	195170

   We are interested in testing whether, at the significance level $\alpha = .01$, the population variances of amount of damage to the auto for the three types of autos are the same. Assume the normality of amount of damage to the auto for three types of autos.

   i. State the null and alternate hypotheses,
   ii. Report the value of the test statistic,
   iii. Find the critical value, and
   iv. Report your decision regarding the null hypothesis and your conclusion in the context of the problem.

---

2. The progeny of a certain mating were classified by a physical attribute into three groups, the numbers being $10, 53$, and 46. According to a genetic model the frequencies should be in the ratios $p^2/2p(1-p)/(1-p)^2$. Are the data consistent with the model at the .05 level?

---

3. Education researchers are concerned about differences in standardized test scores across grade levels in a specific school district. The following table shows a random selection of scores for third-, fourth-, fifth-, and sixth- grade students from *all three* schools within the district of interest.

	Third Grade	Fourth Grade	Fifth Grade	Sixth Grade
School A	94, 81, 87, 70, 85	86, 90, 82, 91, 90	76, 54, 75, 96, 83	86, 93, 81, 90, 80
School B	74, 70, 84, 87, 91	92, 71, 70, 65, 91	73, 82, 89, 60, 93	85, 72, 75, 89, 80
School C	90, 91, 80, 70, 75	74, 89, 80, 90, 94	84, 85, 72, 85, 90	94, 83, 80, 78, 97

i. Present an appropriate model that could be used to compare mean test scores across grades, while accounting for differences across schools.

ii. List the assumptions of your model from part i, and explain how each assumption could be assessed.

iii. Provide an ANOVA table for your model, showing sources of variation, the formulas for sums of squares (no need to calculate the sums of squares by hand), and degrees of freedom.

iv. Suppose researchers believe mean test scores for sixth-graders tend to be greater than those of the other grades, due to testing familiarity. Descrbie in detail how this suspicion can be tested. Provide appropriate hypotheses, test statistic, and decision rule.

4. Consider the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

   i. Prove that $\mathbf{Y}^T\left(\frac{\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ and $\mathbf{Y}^T\left(\frac{\mathbf{I}-\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ are both distributed as χ^2 random variables. Give the degrees of freedom associated with each.

   ii. Prove that $\mathbf{Y}^T\left(\frac{\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$ is distributed independently of $\mathbf{Y}^T\left(\frac{\mathbf{I}-\mathbf{P}_x}{\sigma^2}\right)\mathbf{Y}$.

   iii. Find the distribution of

$$\frac{\mathbf{Y}^T(\mathbf{P}_x)\mathbf{Y}/rank(\mathbf{X})}{\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_x)\mathbf{Y}/(n-rank(\mathbf{X}))}.$$

---

5. Consider a traditional Two-Factor ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

   where $i = 1, 2, 3$, $j = 1, 2$, and $k = 1, 2, 3$.

   i. Construct an appropriate design matrix $\mathbf{X}$ for this model. What is the rank of your design matrix?

   ii. Provide an ANOVA table for this model, including sources of variation, formulas for sums of squares, and degrees of freedom. Explain how the degrees of freedom from this table relate to the rank of the design matrix.

   iii. Suppose it is of interest to test the hypothesis $H_0 : \beta_j = 0, \forall i, j$. Express this as a General Linear Hypothesis, explicitly showing $\mathbf{C}^T$ and $\mathbf{d}$.

   iv. Explicitly explain how the hypothesis H_0 can be evaluated. Include an expression for the test statistic, the distribution of your test statistic, along with degrees of freedom.

---

6. Consider the balanced One-Factor random effect model

$$Y_{ij} = \mu + a_i + \epsilon_{ij}$$
$$i = 1, \ldots, A \qquad , \qquad j = 1, \ldots, n$$

   where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, and $a_i \sim \mathcal{N}(0, \sigma_a^2)$, independent of each other. Let $N = nA$, and show that

   i. $\text{E}[SSB] = (A-1)\sigma^2 + \left(N - \frac{\sum_i n_i^2}{N}\right)\sigma_a^2$.

   ii. $\text{E}[SSW] = \sigma^2(N - A)$.

---

7. Suppose $\{\hat{\theta}_n\}$ is a sequence of Maximum Likelihood Estimators for the parameter θ such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to \mathcal{N}(0,1),$$

where $\theta_0 = 2$ and convergence is in distribution.

   i. Describe how the asymptotic distributional properties of smooth transformations of statistics can be derived, given the asymptotic distributional properties of the original statistic.

   ii. Assume $g(x) = 2x^2$. Find the asymptotic distribution of

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta_0)\right).$$

   iii. Explain why the approach used for part ii would *not* be appropriate in this case for $g(x) = |x+5|$.

---

8. Let $X_1, \ldots, X_n$ be a random sample from $\theta x^{\theta-1} I_{(0,1)}(x)$, where $\theta > 0$

   i. Find the maximum-likelihood estimator of $\mu = \theta/(1+\theta)$.

   ii. Is there a function of θ for which there exists an unbiased estimator whose variance coincides with the Cramer-Rao lower bound?

   iii. Find a pivotal quantity, and use it to find a confidence-interval estimator of θ

---

9. Let $X \sim Unif(0,1)$, and $Y \sim Unif(2,3)$ independent of each other. Find the distribution of $X+Y$.

---

10. A sample of size n is drawn from each of k normal populations with the same variance. Derive the generalized likelihood-ratio test for testing the hypothesis that the means are all 0. Show that the test is a function of a ratio which has the F distribution.

Applied Statistics Comprehensive Exam

January 2020

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. Sampling is usually distinguished from the closely related field of experimental design.

   i. What is experimental design? Explain what are the distinctions between sampling and experimental design.

   ii. What is observational studies? How is sampling distinguished from observational studies?

   iii. What are sampling and nonsampling errors? Give an example.

   iv. Describe the distinctions between design-based approach to sampling and model-based approach to sampling.

---

2. The y-values in the population are denoted as $y_1, y_2, \cdots, y_N$ and the y-values in the sample s are denoted as $y_{s1}, y_{s2}, \cdots, y_{sn}$ distinguishing that the first unit in the sample is not necessarily the same unit as the first unit in the population. For each unit i in the population, define an indicator variable z_i such that $z_i = 1$ if unit i is included in the sample and $z_i = 0$ otherwise. Then the sample mean can be written in the alternative form

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} y_i z_i$$

   i. With simple random sampling of n units from a population of N units, what is the probability π_i that the ith unit of the population is included in the sample? What probability does each possible sample s of distinct n units have?

   ii. How do you select a simple random sample in practice?

   ii. Is $\bar{y}$ unbiased? Justify your answer.

   iii. Show that

$$\text{var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}$$

.

---

3. Data from Alaska Department of Fish and Game shrimp surveys in the vicinity of Kodiak Island, Alaska, were used to estimate the spatial covariance function, which in turn will be used to predict the amount of catch at a new location. The catch data, plotted by location on a chart of the study region, were originally recorded in pounds (lb), with distances measured in nautical miles (nmi.). A research vessel made tows of a trawl net approximately 1 nmi. apart in a grid pattern. Sample covariances were computed using pairs of data lumped into distance intervals. Then a curve of the form $a\exp(-bx)$ was fitted by nonlinear least squares to the covariance estimates. The fitted covariance function was

$$c(x) = 5.1e^{-0.49x}$$

Suppose that one tow has been made with a catch of $y_1 = 5.526$ (units are thousands of pounds) and a second tow 6 nmi. away produced $y_2 = 1.417$.

i. What would be the predicted catch y_0 at a location 1 nmi. from the first tow and 5.4 nmi. from the second using linear prediction (kriging) and the associated prediction mean square error?

ii. What would be the predicted catch y_0 at a location 1 nmi. from the first tow and 5.4 nmi. from the second using the semivariogram and the associated prediction mean square error?

---

4. Suppose that observations y_i have been observed at the ith site for $i = 1, \cdots, n$. Let c_{ij} be $cov(y_i, y_j)$, the covariance between the y-values at the ith and jth sites. From observed y-values at sites $t_1, \cdots, t_n$, it is desired to predict the value of the random variable y_0 at the site t_0. For simplicity, the means $E(y_i)$ are assumed equal. Show that the best unbiased linear predictor (called kriging predictor) is

$$\hat{y}_0 = \sum_{i=1}^{n} a_i y_i$$

where $\boldsymbol{f} = \boldsymbol{G}^{-1}\boldsymbol{h}$ with

$$\boldsymbol{f} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m \end{pmatrix}, \boldsymbol{h} = \begin{pmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{n0} \\ 1 \end{pmatrix} \text{ and } \boldsymbol{G} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n1} & c_{m2} & \cdots & c_{nn} & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

The constant m, which is obtained along with the coefficients a_i, is the Lagrange multiplier and is used in calculating the mean square prediction error.

---

5. With any design, with or without replacement, the ith unit of the population is included in the sample with probability π_i, for $i = 1, 2, \cdots, N$. Let the probability that both unit i and unit j are included in the sample be denoted by π_{ij}. Define the indicator variable z_i to be 1 if the ith unit of the population is included in the sample and 0 otherwise, for $i = 1, 2, \cdots, N$. Define z_{ij} to be 1 if both units i and j are included in the sample and 0 otherwise. Let y_s denote the sample y-value and π_s the selection probability for the unit in the sample.

i. What is the Horvitz-Thompson estimator $\hat{\tau}_\pi$ of the population total τ in terms of z_i?

ii. Is the Horvitz-Thompson estimator unbiased? Justify your answer.

iii. What is the variance of the Horvitz-Thompson estimator of the population total τ? What is an unbiased estimator of this variance? Justify your answer.

iv. What is the generalized unequal-probability estimator $\hat{\mu}_g$ of the population mean μ? Is the estimator $\hat{\mu}_g$ unbiased?

v. The population mean is defined implicitly by the population estimation equation

$$\sum_{i=1}^{N}(y_i - \mu) = 0$$

Notice that solving this equation for μ gives $\mu = (1/N)\sum_{i=1}^{N} y_i$, the usual definition of the population mean. Obtain the generalized unequal-probability estimator $\hat{\mu}_g$ using a (modified) "estimating equation" approach.

vi. Use Taylor series approximation to prove

$$\hat{\mu}_g - \mu \approx \frac{1}{N}\sum_{i\in s}\frac{y_i - \mu}{\pi_i}.$$

Then obtain the approximate variance and variance estimator formulas by the approximation from the usual formulas for a Horvitz-Thompson estimator.

---

6. (**Prediction versus Inference**) Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \cdots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \cdots, X_p)$, which can be written in the very general form

$$y = f(x) + \epsilon,$$

where f is some fixed but unknown function of $X_1, \cdots, X_p$, and ϵ is a random error term, which is independent of X and has mean zero. In essence, statistical learning refers to a set of approaches for estimating f.

i. There are two main reasons that we may wish to estimate f: prediction and (target) inference. Explain the distinctions between prediction and (target) inference with graph such as DAG.

ii. Which is the study, prediction or inference, refering to Table 1 on what you **perceive** is not what you **hear**? Explain.

Table 1: What you perceive is not what you hear:

Actual Sound	Perceived Words
1. The ?eel is on the shoe	The heel is on the shoe
2. The ?eel is on the car	The wheel is on the car
3. The ?eel is on the table	The meal is on the table
4. The ?eel is on the orange	The peel is on the orange

iii. Most statistical learning methods for estimating linear and non-linear f can be characterized as either parametric or non-parametric. Define and explain parametric method and non-parametric method.

iv. Describe the different performances in estimation and inference between a parametric statistical learning approach and a non-parametric statistical learning approach: What are the advantages and disadvantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)?

---

7. (**Curse of Dimensionality**) The curse of dimensionality appears in increasing applications of the new generation of nonparametric statistical methods branded as "machine learning" (ML) techniques, such as linear methods and k-NN methods, widely arising in physical, biological, and social sciences and engineering.

   i. Write the mathematics for k-NN methods and then explain it in words.

   ii. Table 2 below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbors. Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$. What is your prediction with k = 1? Show the detail of your calculation.

Table 2: Training Data Set

obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

   iii. Describe the curse of dimensionality in general?

   iv. Illustrate the curse of dimensionality using an example, e.g., k-NN.

   v. How does the curse of dimensionality affect the performance of statistical and machine learning methods?

   vi. How do you tackle the curse of dimensionality in the k-NN method.

---

8. (**Model Assessment and Selection**) A central problem in all statistical learning situations involves choosing the best learning method for a given application. The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the quality of the ultimately chosen model. The central problem of statistical learning theory is specifically the **complexity** of the model and the **Bias-Variance** dilemma.

i. Assume that $Y = f(X) + \epsilon$, where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma_\epsilon^2$. Derive the bias-variance decomposition for the k-nearest-neighbor regression fit. Then point out the bias term, variance term and irreducible error term.

ii. The cross-validation can be used to estimate the accuracy of a number of different methods in order to choose the best one. Explain how k-fold cross-validation is implemented.

iii. The bootstrap can be used to estimate the accuracy of a number of different methods in order to choose the best one. We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store=rep (NA , 10000)
> for (i in 1:10000) {
store[i]=sum(sample(1:100, rep =TRUE)==4) >0
}
> mean(store)
```

Comment on the results obtained.

---

9. (**High-Dimensional Data**) Much of the recent research in statistical learning has concentrated on non-linear methods. However, linear methods often have advantages over their non-linear competitors in terms of interpretability and sometimes also accuracy. The lasso is a relatively recent alternative to overcome the disadvantage in ridge regression.

i. Explain what is the phenomenon called overfitting and when it occurs, how it affects estimation and inference, and how to avoid it.

ii. The linear method lasso offers improvements over standard linear regression. Write the mathematical representation for the lasso and illustrate the variable selection property of the lasso using contours of the error and constraint functions for the lasso.

---

10. (**Kernel Trick**) The high dimensionality of the feature space raises both **sample complexity** and **computational complexity** challenges. Support vector machines (SVM) are a set of approaches for performing both linear and non-linear classification.

i. The SVM algorithmic paradigm tackles the **sample complexity** challenge by searching for "large margin" separators with regularization that can yield a small **sample complexity** even if the dimensionality of the feature space is high (and even infinite).

Table 3: Data Set

obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

a. Sketch the observations from Table 3 and the optimal separating hyperplane, and provide the equation for this hyperplane.

b. Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.

c. On your sketch, indicate the margin for the maximal margin hyperplane.

d. Indicate the support vectors for the maximal margin classifier.

e. Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

f. Sketch a hyperplane that is not the optimal separating hyperplane, and provide the equation for this hyperplane.

g. Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

h. Write the optimization problem for this problem in the form of maximize *objective* subject to *conditions*.

ii. Describe how SVM algorithmic paradigm tackles the computational complexity challenge using the method of kernel trick for the the labeled dataset shown in Figure 1 that is not linearly separable in 2D input space (left) so that we can use our linear algorithm from part i on a transformed version of the data (right) to get a non-linear algorithm with no effort.

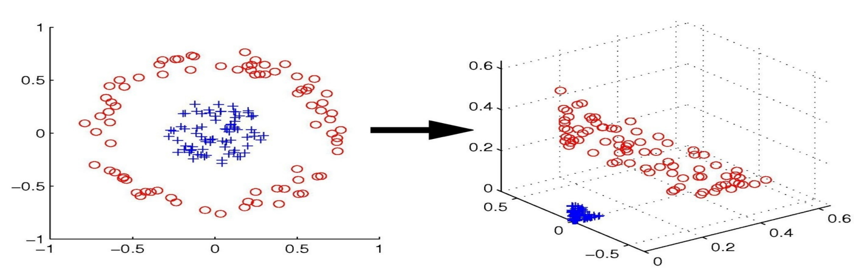


Figure 1: From input space R^2 to feature space R^3

Applied Statistics Comprehensive Exam

August 2019

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1. This Ph.D level exam will run from **8:30 AM to 3:30 PM**.
2. Please label each page with your identification number.

   DO NOT USE YOUR NAME OR BEAR NUMBER.

3. Please write only on one side of each page.
4. Please leave one inch margins on all sides of each page.
5. Please number all pages consecutively.
6. Please label the day number (Day 1 or Day 2) on each page.
7. Please begin each question on a new page, and number each question.
8. Please do not staple pages together.
9. No electronic devices, formula sheets, or other outside materials are permitted.
10. Statistical tables and paper will be provided.
11. Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. Prior to a nationally televised debate between the two presidential candidates, a random sample of 100 persons stated their choice of candidates as follows. Eighty-four persons favored the Democratic candidate, and the remaining 16 favored the Republican. After the debate the same 100 people expressed their preference again. Of the persons who formerly favored the Democrat, exactly one-fourth of them changed their minds, and also one-fourth of the people formerly favoring the Republican switched to the Democratic side.

   i. What nonparametric test would you use to know if there has been a change in the proportion of all voters who favor the Democrat after the debate.

   ii. State the assumptions, the hypotheses, the test statistic and the null distribution of the test statistic.

   ii. Perform the test and draw your conclusion.

---

2. Fifty two-digit numbers were drawn at random from a telephone book, and the chi-squared test for goodness-of-fit is used to see if they could have been observations on a normally distributed random variable. The numbers, after being arranged in order from the smallest to the largest in column, are in Table 1.

Table 1: Two-digit Numbers

23	36	54	61	73	23	37	54	61	73
24	40	56	62	74	27	42	57	63	75
29	43	57	64	77	31	43	58	65	81
32	44	58	66	87	33	45	58	68	89
33	48	58	68	93	35	48	59	70	97

   i. Formulate hypotheses for the χ^2 goodness-of-fit test problem.

   ii. State the assumptions for the χ^2 goodness-of-fit test.

   iii. Perform the chi-squared goodness-of-fit test with the data and draw your conclusion.

---

3. A simple experiment was designed to see if flint in area A tended to have the same degree of hardness as flint in area B. Four sample pieces of flint were collected in area A and five sample pieces of flint were collected in area B. To determine which of two pieces of flint was harder, the two pieces were rubbed against each other. The piece sustaining less damage was judged the harder of the two. In this manner all nine pieces of flint were ordered according to hardness. The rank 1 was assigned to the softest piece, rank 2 to the next softest, and so on. See the Table 2.

Table 2: Flint Data

Origin of Piece	Rank
A	1
A	2
A	3
B	4
A	5
B	6
B	7
B	8
B	9

   i. State the assumptions, the hypotheses, the test statistic for the Mann-Whitney test.
   ii. Perform the Mann-Whitney test and draw your conclusion.
   iii. Why may ranks be considered preferable to the actual data?

---

4. Twelve sets of identical twins were given psychological tests to measure in some sense the amount of aggressiveness in each person's personality. We are interested in comparing the twins with each other to see if the firstborn twin tends to be more aggressive than the other. The results are in Table 3, where the higher score indicates more aggressiveness.

Table 3: Twin Set

	1	2	3	4	5	6	7	8	9	10	11	12
Firstborn X_i	86	71	77	68	91	72	77	91	70	71	88	87
Second twin Y_i	88	77	76	64	96	72	65	90	65	80	81	72

   i. State the assumptions, the hypotheses, the test statistic for the Wilcoxon signed ranks test.
   ii. Perform the Wilcoxon signed ranks test and draw your conclusion.

---

5. Suppose that we want to see whether a random sample agrees with the hypothesized distribution function that is not completely specified, that is, there are unknown parameters that must be estimated from the sample.

   i. Which goodness-of-fit test can be applied to the problem?
      - the chi-squared goodness-of-fit test,
      - the Kolmogorov goodness-of-fit test,
      - the Lilliefors test,
      - the Shapiro-Wilk test.

   ii. State the assumptions, the hypotheses, the test statistic, the test procedure, the advantages and disadvantages for each of the above tests.

---

6. Define a generalized quadratic form $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)'$ and $\boldsymbol{A}$ is a constant $n \times n$ symmetric matrix. If $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$ are independent with $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ and $cov(\boldsymbol{y}_i) = \boldsymbol{\Sigma}$ of the $\boldsymbol{Y}$.

   i. Show that

$$E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = (tr\boldsymbol{A})\boldsymbol{\Sigma} + E(\boldsymbol{Y}')\boldsymbol{A}E(\boldsymbol{Y}).$$

   ii. Use Question i. to obtain a proof that $E(\boldsymbol{S}) = \boldsymbol{\Sigma}$, where $\boldsymbol{S}$ is the sample covariance matrix

---

7. Table 4 gives partial data from Kramer and Jensen (1969).

Table 4: Calcium in Soil and Turnip Greens

Location Number	y_1	y_2	y_3
1	35	3.5	2.80
2	35	4.9	2.70
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	35	4.6	2.88
8	35	10.9	2.90
9	35	8.0	3.28
10	30	1.6	3.2

Three variables were measured (in milliequivalents per 100g) at 10 different locations in the South. The variables are

- y_1 = available soil calcium
- y_2 = exchangeable soil calcium
- y_3 = turnip green calcium

i. Find the sample mean vector $\bar{\boldsymbol{y}}$ and sample covariance matrix $\boldsymbol{S}$.

ii. How many measures of overall variability? What are they?

iii. Calculate the overall variability of the calcium data in Table 1.

iv. Interpret the results.

---

8. In statistics, the Wishart distribution is a generalization to multiple dimensions of the gamma distribution. It is named in honor of John Wishart, who first formulated the distribution in 1928.

   i. What is the formal definition of a Wishart random variable? How is it related to $\sigma^2\chi^2$ distribution?

   ii. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)'$, where $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n$ are independent and each $\boldsymbol{z}_i$ is $N_p(0, \boldsymbol{\Sigma})$. Then $\boldsymbol{Z'AZ}$ is distributed as $W_p(r, \boldsymbol{\Sigma})$ if $\boldsymbol{A}$ is an $n \times n$ constant idempotent matrix of rank r.

---

9. We often measure several dependent variables on each experimental unit instead of just one variable. In the multivariate case, we assume that k independent random samples of size n are obtained from p-variate normal populations with equal covariance matrices for balanced one-way multivariate analysis of variance (MANOVA).

   i. Write the statistical model for MANOVA with the data in terms of (1) each observation vector and (2) the p variables. (define your notation)

   ii. Write the hypotheses in terms of (1) each observation vector and (2) the p variables if we wish to compare the mean vectors of the k samples for significant differences.

   iii. Write the "hypothesis" matrix $\boldsymbol{H}$ and the "error" matrix $\boldsymbol{E}$.

---

10. The four MANOVA test statistics can be summarized in terms of $\boldsymbol{E}$ and $\boldsymbol{H}$ associated with the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_s$ of $\boldsymbol{E}^{-1}\boldsymbol{H}$ as follows, where $s = min(\nu_H, p)$, p is the number of variables, ν_H is the hypothesis degree of freedom and ν_E is the error degrees of freedom.

    - Wilks' Lambda: $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$
    - Pillai's trace: $V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} = tr[(\boldsymbol{E}+\boldsymbol{H})^{-1}\boldsymbol{H}]$

- Lawley-Hotelling: $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\boldsymbol{E}^{-1}\boldsymbol{H})$
- Roy's largest root: $\theta = \frac{\lambda_1}{1+\lambda_1} = \max_{\boldsymbol{a}} \frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}/(k-1)}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}/(kn-k)} = \frac{\boldsymbol{a}_1'\boldsymbol{H}\boldsymbol{a}_1/(k-1)}{\boldsymbol{a}_1'\boldsymbol{E}\boldsymbol{a}_1/(kn-k)}$, where $\boldsymbol{a}_1$ is the eigenvector of λ_1.

i. State the assumptions for the four MANOVA test statistics.

ii. Are all four tests exact tests? Why?

iii. If there are questions about either the multivariate normality or the equality of covariance matrices, then which statistic may be more robust than the other three statistics suggested by simulation studies (related to R output)?

iv. Why do we use four different tests?

Applied Statistics Comprehensive Exam

January 2019

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:______________________________________ Date:__________

1. Briefly describe the Metropolis-Hastings algorithm in Bayesian analysis.

---

2. Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from $Beta(\theta, 1)$. Find

   i. the conjugate prior,

   ii. the posterior distribution.

   iii. prior predictive distribution, and

   iv. the posterior predictive distribution.

---

3. Suppose we have n independent observations from $U(0, \theta)$, $\theta > 0$.

   i. Find a conjugate prior distribution for θ.

   ii. Find the posterior mean and variance for θ.

---

4. For $j = 1, 2$, suppose that

$$\begin{aligned} (y_{j1}, \cdots y_{jn_j} | \mu_j, \sigma_j^2) &\sim \quad iid \;\; N(\mu_j, \sigma_j^2) \\ p(\mu_j, \sigma_j^2) &\propto \sigma_j^{-2}, \end{aligned}$$

and (μ_1, σ_1^2) are independent of (μ_2, σ_2^2) in the prior distribution. Show that the posterior distribution of $\frac{(S_1^2/S_2^2)}{(\sigma_1^2/\sigma_2^2)}$ is F with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

---

5. Observations $y_1, y_2, \cdots, y_n$ are independently distributed given parameters $\theta_1, \cdots, \theta_n$ according to the Poisson distribution

$$p(y_i|\theta) = \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!}.$$

The prior distribution for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ is constructed hierarchically. First, the θ_i s are assumed to be independently identically distributed given a hyperparameter ϕ according to the exponential distribution $p(\theta_i|\phi) = \phi exp(-\phi\theta_i)$ for $\theta_i > 0$ and then ϕ is given the improper uniform prior $p(\phi) \propto 1$ for $\phi > 0$. Provided that $\sum y_i > 1$, prove that the posterior distribution of $z = 1/(1 + \phi)$ is a Beta distribution.

---

6. Consider $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (2,2)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and the matrices $\boldsymbol{A} = (1,1)$, $\boldsymbol{B} = (1,-1)$. Show that $\boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent.

---

7. Due to the curse of dimensionality, the tests for multivariate normality may not be very powerful. However, some check on the distribution is often desirable. One of the procedures for assessing multivariate normality is a generalization of the univariate test based on the skewness and kurtosis measures. The kurtosis for multivariate populations is defined by Mardia as

$$\beta_{2,p} = E\left[(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right]^2 .$$

To estimate $\beta_{2,p}$ using a sample $\boldsymbol{y_1}, \cdots, \boldsymbol{y_n}$, we first define

$$g_{ij} = (\boldsymbol{y_i} - \bar{\boldsymbol{y}})' \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y_j} - \bar{\boldsymbol{y}}),$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$ is the maximum likelihood estimator. Then the estimate of $\beta_{2,p}$ is given by

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} g_{ii}^2 .$$

i. Show that $\beta_{2,p} = p(p+2)$ if $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

ii. Show that $b_{2,p}$ is invariant under the transformation $\boldsymbol{z_i} = \boldsymbol{A}\boldsymbol{y_i} + \boldsymbol{b}$, where $\boldsymbol{A}$ is nonsingular. (hint: $g_{ij}(\boldsymbol{z}) = g_{ij}(\boldsymbol{y})$)

---

8. Consider the hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ that is p-dimensional. In order to test H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus H_a: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, we assume that a random sample $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$ is available from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ unknown. We use the sample mean $\bar{\boldsymbol{y}}$ and the sample covariance $\boldsymbol{S}$ to construct the test statistic,

$$T^2 = n(\bar{\boldsymbol{y}} - \boldsymbol{\mu_0})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{y}} - \boldsymbol{\mu_0}).$$

i. What is the null distribution for the T^2-statistic? What are the parameters to index this null distribution?

ii. What is the key assumption in the null distribution for the T^2-statistic?

iii. Show that the T^2-statistic is the likelihood ratio test statistic.

---

9. The formal definition of a T^2 random variable is similar to the formal definition of the t random variable. Let $\boldsymbol{z}$ be distributed as the multivariate normal $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{W}$ be distributed as the Wishart $W_p(v, \boldsymbol{\Sigma})$ with $\boldsymbol{z}$ and $\boldsymbol{W}$ independent. Then the T^2 random variable with dimension p and degrees of freedom ν is defined as

$$T^2 = z' \left(\frac{\boldsymbol{W}}{\nu} \right)^{-1} z.$$

   i. Show that the distribution of Hotelling's T^2 in problem 3.) can be expressed as this formal definition.

   ii. The square of a univariate t has an F-distribution. In the multivariate case, a simple function of T^2 also has an F-distribution. Show that T^2-statistic can be converted to an F-statistic as follows:

$$\frac{\nu - p + 1}{\nu p} T^2_{p,v} = F_{p,\nu-p+1}.$$

---

10. Explain what is principle component analysis and what are differences between principal component analysis and factor analysis.

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. In paragraph forms answer the following questions regarding a Multiple Linear Regression Analysis.

    i. Describe piecewise linear regression models. Explain when/why they are used.

    ii. Describe the consequences of incorrect model specification.

    iii. Give two interpretations of VIF.

---

2. Consider the model given by

$$Y_i = \beta_0 + W_{i1}\gamma_1 + W_{i2}\gamma_2 + \epsilon_i,$$

where $\epsilon_i \sim NID(0, \sigma^2)$, and

$$\begin{aligned} \sum_i W_{i1} &= \sum_i W_{i2} = \sum_i W_{i1}W_{i2} = 0 \\ \sum_i W_{i1}^2 &= 1+\rho, \quad \sum_i W_{i2}^2 = 1-\rho. \end{aligned}$$

Consider the estimator

$$\widehat{\gamma}_{1(k_1)} = \frac{\sum_i W_{i1}Y_i}{\sum_i W_{i1}^2 + k_1}$$

i. For $k_1 > 0$, show that $\widehat{\gamma}_{1(k_1)}$ is a biased estimate of γ_1.

ii. Find the mean squared errors of $\widehat{\gamma}_{1(k_1)}$.

---

3. Consider the test for lack of fit for a multiple linear regression. Find $E(MS_{PE})$ and $E(MS_{LOF})$. Note that

$$MS_{PE} = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$MS_{LOF} = \frac{1}{m-p}\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

---

4. A data set was collected to model relationship between selling price to nine regressors. Using this data set, the attached SAS output has been compiled to test for the potential effects of the nine regressors on the selling price. Based on SAS output page 6, answer the following questions.

i. Test for significance of regression. What conclusions can you draw?

ii. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

iii. What is the contribution of lot size and living space to the model given that all of the other regressors are included?

iv. Is multicollinearity a potential problem in this model?

---

5. DUPLEX algorithm was used to split a data set on the gasoline mileage performance of 30 different automobiles into estimation and prediction sets. Based on SAS output page 7, answer the following questions.

i. Evaluate the statistical properties of these data sets.

[Hint: Use the relative volumes of the regions spanned by the two data sets.]

ii. Fit a model involving x_1 and x_6 to the estimation data. Do the coefficients values from this model seem reasonable?

iii. Use this model to predict the observations in the prediction data set. What is your evaluation of this model's predictive performance?

---

6. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Then

i. show that $\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}) = tr(\boldsymbol{\Sigma}^{-1}[\boldsymbol{W} + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'])$, where $\boldsymbol{W} = \sum_{i=1}^n (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$

ii. show that the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$;

iii. show that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}$.

---

7. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ unknown. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Let $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the likelihood function for the sample. Then for H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$,

i. show that $\sum_{i=1}^{n}(\boldsymbol{y_i}-\boldsymbol{\mu_0})'\widehat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{y_i}-\boldsymbol{\mu_0}) = np$, where $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_i}-\boldsymbol{\mu_0})(\boldsymbol{y_i}-\boldsymbol{\mu_0})'$ that maximizes $L(\boldsymbol{\mu}_0,\boldsymbol{\Sigma})$ under H_0.

ii. show that the likelihood ratio

$$LR = \frac{\max\limits_{H_0} L(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\max\limits_{H_1} L(\boldsymbol{\mu},\boldsymbol{\Sigma})}$$

leads to the test statistic $T^2 = (\bar{\boldsymbol{y}}-\boldsymbol{\mu_0})'(\frac{\boldsymbol{S}}{n})^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu_0})$, where $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y_i}-\bar{\boldsymbol{y}})(\boldsymbol{y_i}-\bar{\boldsymbol{y}})'$.

iii. what is the distribution of T^2 that was obtained by Hotelling (1931), assuming H_0 is true and sampling is from $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$? What are the parameters the distribution of T^2 is indexed by?

---

8. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. We define sample totals and means as follows:

$$\boldsymbol{y}_{i.} = \sum_{j=1}^{n}\boldsymbol{y}_{ij}, \qquad \boldsymbol{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n}\boldsymbol{y}_{ij},$$

$$\bar{\boldsymbol{y}}_{i.} = \frac{\boldsymbol{y}_{i.}}{n}, \qquad \bar{\boldsymbol{y}}_{..} = \frac{\boldsymbol{y}_{..}}{kn}.$$

To summarize variation in the data, we use "between" and "within" matrices $\boldsymbol{H}$ and $\boldsymbol{E}$, defined as

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.}-\bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.}-\bar{\boldsymbol{y}}_{..})'$$

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_{i.})'$$

The likelihood ratio test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ is given by

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}$$

where p=the number of variables (dimension), ν_H=the hypothesis degrees of freedom and ν_E =the error degrees of freedom, which is known as Wilks' Λ. Show that Wilks' Λ can be expressed in terms of the eigenvalues λ_1, λ_2, $\cdots$, λ_p of $\boldsymbol{E}^{-1}\boldsymbol{H}$

$$\Lambda = \prod_{i=1}^{s}\frac{1}{1+\lambda_i}$$

where the number of nonzero eigenvalues of $\boldsymbol{E}^{-1}\boldsymbol{H}$ is $s = min(p,\nu_H)$.

---

9. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. Consider H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, the Lawley-Hotelling statistic (Lawley 1938, Hotelling 1951) defined as $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\boldsymbol{E^{-1}H})$ can be expressed as a linear combination of Hotelling T^2-statistics so that Lawley-Hotelling statistic is also known as Hotelling's generalized T^2-statistic, where

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})',$$

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'.$$

   i. Show that $\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})' - kn(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})'$, where $\boldsymbol{\mu}$ is the common value of $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\cdots$, $\boldsymbol{\mu_k}$ under H_0.

   ii. Show that $U^{(s)} = \frac{n}{\nu_E}\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu}) - \frac{kn}{\nu_E}(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})$, where $\boldsymbol{S}_{pl} = \boldsymbol{E}/\nu_E$. Write the terms on the right side in terms of T^2-statistics.

---

10. If $\boldsymbol{y}_{ij}$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n$, are independently observed from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, the hypothesis matrix and error matrix for H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ are $\boldsymbol{H}$ and $\boldsymbol{E}$. Show that the maximum value of $\lambda = \frac{\boldsymbol{a'Ha}}{\boldsymbol{a'Ea}}$ and the vector $\boldsymbol{a}$ that produces the maximum are given by the largest eigenvalue λ_1 and the associated eigenvector of $\boldsymbol{E^{-1}H}$, respectively (hint: differentiate λ with respect to $\boldsymbol{a}$ and set the result equal to $\boldsymbol{0}$).

SAS output for Question 4

Model: model1

Dependent Variable: y sale price of the house (in thousands of dollars)

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	701.66438	87.70805	10.33	<.0001
Error	15	127.38187	8.49212		
Corrected Total	23	829.04625			

Root MSE	2.91412	**R-Square**	0.8464
Dependent Mean	34.61250	**Adj R-Sq**	0.7644
Coeff Var	8.41928		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	12.75192	5.19667	2.45	0.0268	0
x1	taxes (in thousands of dollars)	1	1.72633	0.98816	1.75	0.1011	6.61889
x2	number of baths	1	8.08784	4.03447	2.00	0.0634	2.55561
x3	lot size(in thousands of square feet)	1	0.28738	0.45392	0.63	0.5362	2.15393
x4	living space(in thousands of square feet)	1	2.28954	4.27510	0.54	0.6001	3.77798
x5	number of garage stalls	1	2.20354	1.33560	1.65	0.1198	1.76578
x6	number of rooms	1	0.50740	2.06293	0.25	0.8090	9.02035
x7	number of bedrooms	1	-2.87189	2.82979	-1.01	0.3263	6.91500
x8	age of the home (in years)	1	-0.01681	0.06102	-0.28	0.7867	1.98792

Model: model2

Dependent Variable: y sale price of the house (in thousands of dollars)

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	693.77015	115.62836	14.53	<.0001
Error	17	135.27610	7.95742		
Corrected Total	23	829.04625			

Root MSE	2.82089	**R-Square**	0.8368
Dependent Mean	34.61250	**Adj R-Sq**	0.7792
Coeff Var	8.14992		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.94158	5.02453	2.58	0.0196
x1	taxes (in thousands of dollars)	1	2.14748	0.85729	2.50	0.0227
x2	number of baths	1	8.83544	3.57210	2.47	0.0242
x5	number of garage stalls	1	1.98550	1.27411	1.56	0.1376
x6	number of rooms	1	0.66117	1.98627	0.33	0.7433
x7	number of bedrooms	1	-2.71535	2.62127	-1.04	0.3148
x8	age of the home (in years)	1	-0.01859	0.05903	-0.31	0.7566

SAS output for Question 5

vol_est	vol_pred	ratio
0.546669	0.6313279	0.9531395

The CORR Procedure

3 Variables:	y x1 x6

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
y	15	20.47867	6.98018	307.18000	11.20000	36.50000	y
x1	15	275.04000	124.86596	4126	85.30000	500.00000	x1
x6	15	2.53333	1.12546	38.00000	1.00000	4.00000	x6

Pearson Correlation Coefficients, N = 15 Prob > |r| under H0: Rho=0			
	y	x1	x6
y y	1.00000	-0.85105 <.0001	-0.42115 0.1180
x1 x1	-0.85105 <.0001	1.00000	0.67330 0.0059
x6 x6	-0.42115 0.1180	0.67330 0.0059	1.00000

The REG Procedure
Model: PREDY
Dependent Variable: y y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	522.83015	261.41508	19.69	0.0002
Error	12	159.29022	13.27418		
Corrected Total	14	682.12037			

Root MSE	3.64338	R-Square	0.7665
Dependent Mean	20.47867	Adj R-Sq	0.7276
Coeff Var	17.79108		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	32.07476	2.55105	12.57	<.0001
x1	x1	1	-0.05803	0.01055	-5.50	0.0001
x6	x6	1	1.72291	1.17016	1.47	0.1667

Obs	y	PREDY	residual
1	17.00	18.6556	-1.65561
2	18.25	15.1518	3.09824
3	21.47	20.3165	1.15350
4	30.40	29.8974	0.50261
5	16.50	18.6556	-2.15561
6	21.50	25.5973	-4.09731
7	19.70	18.8257	0.87428
8	14.89	13.4328	1.45716
9	16.41	17.0668	-0.65678
10	23.54	22.1155	1.42454
11	21.47	14.6295	6.84052
12	31.90	29.8974	2.00261
13	13.27	12.2722	0.99778
14	13.90	15.1518	-1.25176
15	13.77	18.0753	-4.30530

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1. Compare and contrast the Wilcoxon Signed Rank Test and the Wilcoxon Rank Sum Test. Give examples of both tests.

---

2. Describe the procedure behind the Binomial Test and the applications it can be used for.

---

3. What are the non-parametric alternative procedures for the following parametric tests?
   i. Repeated Measures ANOVA
   ii. Two Population Test for the Difference Between Two Means
   iii. Test for Slope in Simple Linear Regression

---

4. What are some alternative applications of the Sign Test in nonparametric statistics? (i.e. name several tests that employ the basic principle of the Sign Test to its procedure).

---

5. Name the parametric tests that the following nonparametric procedures replace.
   i. Mann-Whitney Test
   ii. RxC Median Test
   iii. Sign Test

---

6. The Education Commissioner of Colorado has hired you to estimate the average reading score of kindergartner students in the state of Colorado. Devise a sampling scheme (including how you would determine sample size) and explain your reasoning as to why you chose such a plan. Include costs in your plan.

---

7. Explain in detail how the Adaptive Cluster Sampling procedure works.

---

8. Compare and contrast Two-Stage Sampling and Double Sampling. Give examples of both.

---

9. Describe the conditions within the population when Stratified Sampling works best.

---

10. Explain how an auxiliary variable may be used to develop a Ratio Estimate of the population total.

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:__________

1. Compare and contrast the Wilcoxon Signed Rank Test and the Wilcoxon Rank Sum Test. Give examples of both tests.

2. Describe the procedure behind the Binomial Test and the applications it can be used for.

3. What are the non-parametric alternative procedures for the following parametric tests?

   i. Repeated Measures ANOVA

   ii. Two Population Test for the Difference Between Two Means

   iii. Test for Slope in Simple Linear Regression

4. What are some alternative applications of the Sign Test in nonparametric statistics? (i.e. name several tests that employ the basic principle of the Sign Test to its procedure).

5. Name the parametric tests that the following nonparametric procedures replace.

   i. Mann-Whitney Test

   ii. RxC Median Test

   iii. Sign Test

6. Explain how to set up the control limits of an x-bar and R chart in Phase I.

7. Compare and contrast the following charts. Give examples of each.

   i. p-chart

   ii. np-chart

   iii. u-chart

8. Discuss why multiple univariate x-bar charts are not used to follow p-variables simultaneously in a quality control process.

9. Compare and contrast the CUSUM and EWMA charts. Explain how each chart is set up and how the monitoring statistic is defined. Which chart has the better overall average run length performance, if so?

10. Discuss how to perform a gage R & R study.

Applied Statistics Comprehensive Exam

August 2018

Ph.D Day 2 - Exam

This comprehensive exam consists of 10 questions pertaining to two topics of your choice.

Before you start, Please make sure the topics are the one you have chosen.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No electronic devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:__________

1. Briefly describe the Gibbs sampler in Bayesian analysis.

---

2. if $\pi_m(\theta)$, for $m = 1, \cdots, , M$, are conjugate prior densities for the sampling model $y|\theta$, show that the class of finite mixture prior densities given by

$$\pi(\theta) = \sum_{m=1}^{M} \lambda_i \pi_m(\theta)$$

is also a conjugate class, where the λ_m's are nonnegative weights that sum to 1.

---

3. Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from $N(\mu, \sigma)$, assuming both μ and σ are unknown. Let $\theta = (\mu, \sigma)$.

   i. Find the Jeffreys' prior.

   ii. Find the posterior distribution of $\frac{\sqrt{n}(\mu-\bar{y})}{s}$, where s is the sample standard deviation.

   iii. Use part (b) to find a 95% HPD credible set for μ.

---

4. Suppose we have n independent observations from $Unif(0, \theta)$, $\theta > 0$.

   i. Find a conjugate prior distribution for θ.

   ii. Find the posterior mean and variance for θ.

---

5. A set of n counts $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$ are modeled as

$$P(Y_i = 0|\gamma_i = 0, \pi, \lambda) = 1,$$

$$Y_i|(\gamma_i = 1, \pi, \lambda) \sim \text{Poisson}(\lambda), \quad \text{(independent across } i)$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ is a set of n latent binary counts, $\pi \in [0, 1]$ and $\lambda > 0$ have the hierarchical prior:

$$\gamma_i|(\pi, \lambda) \sim iid \quad \text{Bernoulli}(\pi), i = 1, \cdots, n,$$

$$\pi|\lambda \sim \text{Beta}(c\lambda, 1), \quad \lambda \sim \text{Gamma}(a, b)$$

for some positive constants a, b and c

i. Show that the conditional prior of λ given π is $Gamma(a+1, b-c\log\pi)$.

ii. Write down the posterior conditional probability distributions of $\pi|(\gamma, \lambda, \boldsymbol{y})$, $\lambda|(\boldsymbol{\gamma}, \pi, \boldsymbol{y})$,and $\boldsymbol{\gamma}|(\pi, \lambda, \boldsymbol{y})$, where $\boldsymbol{y}$ is the given data on $\boldsymbol{Y}$. Answer in terms of conditional distributions with explicit formulas for their parameters and with appropriate use of conditional independence.

---

6. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Then

i. show that $\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu}) = tr(\boldsymbol{\Sigma}^{-1}[\boldsymbol{W} + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'])$, where $\boldsymbol{W} = \sum_{i=1}^n (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$

ii. show that the maximum likelihood estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$;

iii. show that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W}$.

---

7. If the observation vectors $\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_n}$ is a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ unknown. The density function for $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}.$$

Let $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the likelihood function for the sample. Then for H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$,

i. show that $\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu_0})'\widehat{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_0}) = np$, where $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y_i} - \boldsymbol{\mu_0})(\boldsymbol{y_i} - \boldsymbol{\mu_0})'$ that maximizes $L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ under H_0.

ii. show that the likelihood ratio

$$LR = \frac{\max_{H_0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\max_{H_1} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

leads to the test statistic $T^2 = (\bar{\boldsymbol{y}} - \boldsymbol{\mu_0})'(\frac{\boldsymbol{S}}{n})^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu_0})$, where $\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{y_i} - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})'$.

iii. what is the distribution of T^2 that was obtained by Hotelling (1931), assuming H_0 is true and sampling is from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$? What are the parameters the distribution of T^2 is indexed by?

---

8. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. We define sample totals and means as follows:

$$\boldsymbol{y}_{i.} = \sum_{j=1}^{n} \boldsymbol{y}_{ij}, \qquad \boldsymbol{y}_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n} \boldsymbol{y}_{ij},$$

$$\bar{\boldsymbol{y}}_{i.} = \frac{\boldsymbol{y}_{i.}}{n}, \qquad \bar{\boldsymbol{y}}_{..} = \frac{\boldsymbol{y}_{..}}{kn}.$$

To summarize variation in the data, we use "between" and "within" matrices $\boldsymbol{H}$ and $\boldsymbol{E}$, defined as

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})'$$

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'$$

The likelihood ratio test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ is given by

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|}$$

where p=the number of variables (dimension), ν_H=the hypothesis degrees of freedom and ν_E =the error degrees of freedom, which is known as Wilks' Λ. Show that Wilks' Λ can be expressed in terms of the eigenvalues λ_1, λ_2, $\cdots$, λ_p of $\boldsymbol{E}^{-1}\boldsymbol{H}$

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i}$$

where the number of nonzero eigenvalues of $\boldsymbol{E}^{-1}\boldsymbol{H}$ is $s = min(p, \nu_H)$.

---

9. In a one-way multivariate analysis of variance (MANOVA), we assume that a random sample of p-variate observations is available from each of k multivariate normal populations with equal covariance matrices $\boldsymbol{\Sigma}$. Consider H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, the Lawley-Hotelling statistic (Lawley 1938, Hotelling 1951) defined as $U^{(s)} = \sum_{i=1}^{s} \lambda_i = tr(\boldsymbol{E^{-1}H})$ can be expressed as a linear combination of Hotelling T^2-statistics so that Lawley-Hotelling statistic is also known as Hotelling's generalized T^2-statistic, where

$$\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{\boldsymbol{..}})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{\boldsymbol{..}})',$$

$$\boldsymbol{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{\boldsymbol{i.}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{\boldsymbol{i.}})'.$$

i. Show that $\boldsymbol{H} = n\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})' - kn(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})'$, where $\boldsymbol{\mu}$ is the common value of $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\cdots$, $\boldsymbol{\mu_k}$ under H_0.

ii. Show that $U^{(s)} = \frac{n}{\nu_E}\sum_{i=1}^{k}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{i.} - \boldsymbol{\mu}) - \frac{kn}{\nu_E}(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})'\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_{..} - \boldsymbol{\mu})$, where $\boldsymbol{S}_{pl} = \boldsymbol{E}/\nu_E$. Write the terms on the right side in terms of T^2-statistics.

---

10. If $\boldsymbol{y}_{ij}$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, n$, are independently observed from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, the hypothesis matrix and error matrix for H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$ are $\boldsymbol{H}$ and $\boldsymbol{E}$. Show that the maximum value of $\lambda = \frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}$ and the vector $\boldsymbol{a}$ that produces the maximum are given by the largest eigenvalue λ_1 and the associated eigenvector of $\boldsymbol{E}^{-1}\boldsymbol{H}$, respectively (hint: differentiate λ with respect to $\boldsymbol{a}$ and set the result equal to $\boldsymbol{0}$).

Applied Statistics Comprehensive Exam

January 2018

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) Over the past 5 years, an insurance company has had a mix of 40% whole life policies, 20% universal life policies, 25% annual renewable-term (ART) policies, and 15% other types of policies. A change in this mix over the long haul could require a change in the commission structure, reserves, and possibly investments. A sample of 1,000 policies issued over the last few months gave the results shown here.

Category	n
Whole life	320
Universal life	280
ART	240
Other	160

Use these data to assess whether there has been a shift from the historical percentages. Use the 5% significant level. Clearly identify the hypothesis you are testing, report the test statistic and state your conclusion.

2.) A researcher randomly assigned forty male subjects to one of two treatment groups to compare the treatments. Each patient had his BPRS factor measured before treatment (week 0) and at weekly intervals for eight weeks. Use SAS output to answer the following questions. Use $\alpha = .05$.

Repeated Measures Analysis of Variance

Sphericity Tests				
Variables	**DF**	**Mauchly's Criterion**	**Chi-Square**	**Pr > ChiSq**
Transformed Variates	35	0.0002916	285.9191	<.0001
Orthogonal Components	35	0.0048335	187.29315	<.0001

Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treatment	1	33.61111	33.61111	0.03	0.8539
Error	38	37177.44444	978.35380		

Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	**DF**	**Type III SS**	**Mean Square**	**F Value**	**Pr > F**	**G - G**	**H-F-L**
Time	8	13717.80556	1714.72569	34.62	<.0001	<.0001	<.0001
Time*treatment	8	830.33889	103.79236	2.10	0.0360	0.0961	0.0886
Error(Time)	304	15057.85556	49.53242				

Greenhouse-Geisser Epsilon	0.4240
Huynh-Feldt-Lecoutre Epsilon	0.4706

a) What type of design is used here?

b) Is there a difference between two treatments?

c) Is there a significant time effect?

d) Did the time affect equally across the two treatments?

3.) A drug company claims that a certain vitamin will increase a sprinter's speed in the 200 yard dash. Eight runners were initially timed (in sec.), given the supplement, and then timed again on the next day.

Athlete:	1	2	3	4	5	6	7	8
Day one:	21.0	23.0	18.2	20.5	26.2	25.3	21.9	21.6
Day two:	21.9	23.6	17.9	20.4	27.0	25.0	22.2	21.6

Test for a vitamin effect using $\alpha = .05$. State the null and alternate hypotheses, and report the value of the test statistic, and the critical value used to conduct the test. Report your decision regarding the null hypothesis and your conclusion in the context of the problem.

4.) Briefly describe the experimental design you would choose for each of the following situations, and explain why.

a) Modern zoos try to reproduce natural habitats in their exhibits as much as possible. They therefore use appropriate plants, but these plants can be infested with inappropriate insects. Zoos need to take great care with pesticides, because the variety of species in a zoo makes it more likely that a sensitive species is present. Cycads (plants that look vaguely like palms) can be infested with mealybug, and the zoo wishes to test three treatments: water (a control), horticultural oil (a standard no-mammalian-toxicity pesticide), and fungal spores in water (Beauveria bassiana, a fungus that grows exclusively on insects). Five infested cycads are removed to a testing area. Three branches are randomly chosen on each cycad, and two 3 cm by 3 cm patches are marked on each branch; the number of mealybugs in these patches is noted. The three branches on each cycad are randomly assigned to the three treatments. After three days, the patches are counted again, and the response is the change in the number of mealybugs (before - after).

b) An investigative group at a television station wishes to determine if doctors treat patients on public assistance differently from those with private insurance. They measure this by how long the doctor spends with the patient. There are four large clinics in the city, and the station chooses three pediatricians at random from each of the four clinics. Ninety-six families on public assistance are located and divided into four groups of 24 at random. All 96 families have a one-year-old child and a child just entering school. Half the families will request a one-year checkup, and the others will request a preschool checkup. Half the families will be given temporary private insurance for the study, and the others will use public assistance. The four groupings of families are the factorial combinations of checkup type and insurance type. Each group of 24 is now divided at random into twelve sets of two, with each set of two assigned to one of the twelve selected doctors. Thus each doctor will see eight patients from the investigation. Recap: 96 units (families); the response is how long the doctor spends with each family.

5.) The SAS output gives a regression analysis of the systolic blood pressure (SBP), body size (QUET) a measure of size defined by QUET=100 (weight/height^2), age (AGE), and smoking history (SMK=0 if nonsmoker,SMK=1 if a current or previous smoker) for a hypothetical sample of 32 white males over 40 years old from the town of Angina. Note that QUMK=QUET*SMK.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4184.10759	1394.70253	17.42	<.0001
Error	28	2241.86116	80.06647		
Corrected Total	31	6425.96875			

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.36649	2060.18325	25.91	<.0001
Error	29	2305.60226	79.50353		
Corrected Total	31	6425.96875			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	49.31176	19.97235	2.47	0.0199
QUET	1	26.30283	5.70349	4.61	<.0001
SMK	1	29.94357	24.16355	1.24	0.2256
QUMK	1	-6.18478	6.93171	-0.89	0.3799

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	63.87603	11.46811	5.57	<.0001
QUET	1	22.11560	3.22996	6.85	<.0001
SMK	1	8.57101	3.16670	2.71	0.0113

a) Determine a single multiple model that uses the data for both smokers and nonsmokers and that defines straight-line models for each group with possibly differing intercepts and slopes. Obtain the least-square line for smokers and nonsmokers by using the single multiple model.

b) Test H_0 : the two lines are parallel. State the appropriate null hypothesis in terms of the regression coefficients of the regression model.

c) Suppose we fail to reject the null hypothesis in part (b) above. State the appropriate ANACOVA regression model to use for comparing the mean blood pressure in the two smoking categories, controlling for QUET.

6.) In an exercise study, pulse gains (the difference in the post-exercise and pre-exercise pulse rates) were computed on forty persons. In this study, gender (female or male) and smoking status (smoker or nonsmoker) for individuals were also registered. The summary data is in the table below.

Group	n	mean	Variance
MALE SMOKER	10	59.0	101.556
FEMALE SMOKER	10	85.4	913.822
MALE NONSMOKER	10	56.6	174.267
FEMALE NONSMOKER	10	63.8	72.400

Consider the group (given in the table) as a four-level factor and answer the following questions.

a) Write down a contrast that compares

   i. smokers and nonsmokers
   ii. female and male
   iii. female smokers and female nonsmokers
   iv. male smokers and male nonsmokers

b) Estimate all the contrasts in part a).

c) Find the MSE for the one-way ANOVA with the group as the only factor.

d) Use the Bonferroni method to test all the comparisons in a). Consider $\alpha_\Sigma = .05$.

7.) Given the data, use the Sign Test to test $H_0: \tilde{\mu} = 8.41$ vs $H_1: \tilde{\mu} > 8.41$.

8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30

8.) Repeat previous question using the Wilcoxon Signed Rank Test.

9.) We wish to investigate the shelf life of a particular carbonated beverage. Ten cans are randomly selected and their respective shelf lives measured. The following results were obtained (in days):

```
108 138 124 163 124 159 106 134 115 139
```

Assuming the normal distribution for these data and the corresponding conjugate prior for $\theta = (\mu, \sigma^2)$ with $E(\mu|\sigma^2) = 120, Var(\mu|\sigma^2) = \sigma^2/2$, and $\sigma^2 \sim Inv - \chi^2(3, 2)$.

a) Find a Bayesian estimate for the population mean.

b) Find a 95% credible interval for the mean of population μ and interpret this interval.

10.) Let $y_1 = 2$ and $y_2 = 7$ be two independent observations from a Poisson distribution

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad \lambda = 2, 6, 8.$$

Assume a uniform prior for λ and find the posterior mean.

Applied Statistics Comprehensive Exam

August 2017

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) 502

2.) 502

3.) The effect of three different lubricating oils on fuel economy in diesel truck engines is being studied. (Fuel economy is measured using brake-specific fuel consumption after the engine has been running for 15 minutes.) Five different truck engines are available for the study. Experimenters have collected the following data.

	Truck				
Oil	1	2	3	4	5
1	0.500	0.634	0.487	0.329	0.512
2	0.535	0.675	0.520	0.435	0.540
3	0.513	0.595	0.488	0.400	0.510

i. Describe as completely as you can the type of design applied for this experiment. What are the advantages of using such a design?

ii. Present an appropriate model equation to describe how the two separate factors will be analyzed.

iii. Provide an ANOVA table for this experiment. Include columns for source of variation, sums of squares (*you can show only the SS formulas in place of calculating the numbers*), degrees of freedom, mean squares (*in terms of each SS*), and F-statistic.

iv. Assuming $SSA = 0.0013$ (Oil), $SSB = 0.0307$ (Truck), and $SSE = 0.0710$, perform a test of the significance of the Oil effect, and explain the result.

4.) 614

5.) 610

6.) Consider the Randomized Complete Block design, and the Latin Square design.

i. Describe each design in detail, providing notation for different factors and replicates within each design.

ii. Provide an outline of an ANOVA table for a specific case of each type of design, including only sources of variation and degrees of freedom.

iii. Discuss the relative advantages and disadvantages of each type of design with respect to the other.

7.) Given the data, use the Wilcoxon Signed Ranks Test to test:

$$H_0 : \tilde{\mu} = 107 \text{ versus } H_1 : \tilde{\mu} \neq 107$$

$$99, 100, 90, 94, 135, 108, 107, 111, 119, 104, 127, 109, 117, 105, 125$$

8.) Compare and contrast stratified sampling to simple random sampling. What do these designs have in common? How are they different? Give examples/applications of each design. Under what conditions is stratified sampling preferred over simple random sampling?

9.) 606

10.) Higher education researchers are interested in predicting the chances of first-generation students completing a four-year college degree within six years. Using historical records, they have retrieved the six-year completion status (graduated / not graduated) of approximately 12,000 first-generation students from 520 undergraduate US institutions. They have also recorded each student's gender (female / male), high school composite score (continuous), and parents' income level at admission (in thousands of dollars).

i. Considering the interest of modeling six-year completion status, propose an appropriate model.

ii. Describe *at least two* descriptive statistics you would consult before applying your model from part i. What are these descriptives used for?

iii. Describe *at least three* measures of fit that could be used to assess the adequacy of your model from part i.

iv. Specifically, discuss the Hosmer-Lemeshow fit statistic. How is it calculated? How would your perspective of this statistic change if the researchers had sampled 120,000 students, instead of 12,000?

v. Specifically, discuss the Receiver Operating Characteristic (ROC) curve. How is it produced, and how do you use the area under the curve to assess fit?

vi. Specifically, discuss the (studentized) deviance residuals. How are they used to assess fit?

Applied Statistics Comprehensive Exam

January 2017

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) Briefly describe the experimental design you would choose for each of the following situations, and explain why.

i. An investigative group at a television station wishes to determine if doctors treat patients on public assistance differently from those with private insurance. They measure this by how long the doctor spends with the patient. There are four large clinics in the city, and the station chooses three pediatricians at random from each of the four clinics. Ninety-six families on public assistance are located and divided into four groups of 24 at random. All 96 families have a one-year-old child and a child just entering school. Half the families will request a one-year checkup, and the others will request a preschool checkup. Half the families will be given temporary private insurance for the study, and the others will use public assistance. The four groupings of families are the factorial combinations of checkup type and insurance type. Each group of 24 is now divided at random into twelve sets of two, with each set of two assigned to one of the twelve selected doctors. Thus each doctor will see eight patients from the investigation. Recap: 96 units (families); the response is how long the doctor spends with each family.

ii. Food scientists wish to study how urban and rural consumers rate cheddar cheeses for bitterness. Four 50-pound blocks of cheddar cheese of different types are obtained. Each block of cheese represents one of the segments of the market (for example, a sharp New York style cheese). The raters are students from a large introductory food science class. Ten students from rural backgrounds and ten students from urban backgrounds are selected at random from the pool of possible raters. Each rater will taste eight bites of cheese presented in random order. The eight bites are two each from the four different cheeses, but the raters don't know that. Each rater rates each bite for bitterness.

iii. A small travel agency is interested to better understand the effect of age of customer (x) on the amount of money spent in a tour (y), in the last twelve months. Agency has recognized two important customer segments. The first segment, which we will denote by A, consists of those customers who have purchased an adventure tour in the last twelve months. The second segment, which we will denote by C, consists of those customers who have purchased a cultural tour in the last twelve months. Note that the two segments are completely separate in the sense that there are no customers who are in both segments.

2.) In an experiment to investigate the effect of color of paper (blue,green, orange) on response rates for questionnaires distributed by the "windshield method" in supermarket parking lots, 15 representative supermarket parking lots were chosen in a metropolitan area and each color was assigned at random to five of the lots. It has been suggested to the investigator that size of parking lot might be a useful concomitant variable. For this question use the **SAS output on Page 6**.

i. Test for color effects after adjusting for the size of parking lot; use $\alpha = .10$.

ii. Make all pairwise comparisons between the color effects after adjusting for the size of parking lot; use the Bonferroni procedure with a .05 percent family rate. Also use the adjusted MSE for the denominator of your t-tests.

3.) Kuehl (2000) reports the results of an experiment conducted at a large seafood company to investigate the effect of storage temperature and type of seafood upon bacterial growth on oysters and mussels. Three storage temperatures were studied (0°C, 5°C, and 10°C). Three cold storage units were randomly assigned to be operated at each temperature. Within each storage unit, oysters and mussels were randomly assigned to be stored on one of the two shelves. The seafood was stored for 2 weeks at the assigned temperature, and at the end of the time the bacterial count was obtained from a sample on each shelf. The resulting data (log bacterial count) is shown below.

Storage		Seafood Type	
Unit	Temp.	Oysters	Mussels
1	0	3.6882	0.3565
2	0	1.8275	1.7023
3	0	5.2327	4.5780
4	5	7.1950	5.0169
5	5	9.3224	7.9519
6	5	7.4195	6.3861
7	10	9.7842	10.1352
8	10	6.4703	5.0482
9	10	9.4442	11.0329

i. What is the experimental unit for temperature?

ii. Why was it necessary to include nine storage units instead of three?

iii. What is the experimental unit for seafood type?

iv. What design was used to conduct the experiment? Justify your answer. In particular, write the factors, their levels, and the outcome variable.

v. Write the statistical model for the data and identify the model components.

vi. List the assumptions that need to be checked for the model that you considered above.

4.) Consider the model,

$$\begin{aligned} y_1 &= \theta_1 + \theta_2 + \epsilon_1, \\ y_2 &= 2\theta_1 + \epsilon_2, \\ y_3 &= \theta_1 - \theta_2 + \epsilon_3. \end{aligned}$$

where ϵ_i, $i = 1, 2, 3$ are i.i.d. $N(0, \sigma^2)$. Assume $\boldsymbol{y}^T = \begin{bmatrix} 2 & 1 & 4 \end{bmatrix}$ is observed.

i. Find a 95% confidence intervals for $\theta_1 - \theta_2$.

ii. Find the value of the GLRT test statistic for testing $H_0 : \theta_1 = \theta_2$ versus $H_a : \theta_1 \neq \theta_2$.

iii. What is your conclusion for the hypothesis in (ii)?

5.) The following data are a random sample from a three variates normal distribution.

Subject Number	y_1	y_2	y_3
1	51	36	50
2	27	20	26
3	37	22	41
4	42	36	32
5	27	18	33
6	43	32	43
7	41	22	36
8	38	21	31
9	36	23	27
10	26	31	31
11	29	20	25

Here are the sample mean vector and sample covariance matrix of the above data:

$$\bar{\mathbf{y}} = \begin{bmatrix} 36.09 \\ 25.55 \\ 34.09 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 65.09 & 33.65 & 47.59 \\ 33.65 & 46.07 & 28.95 \\ 47.59 & 28.95 & 60.69 \end{bmatrix}$$

Using these data, at the level of significance 0.05, perform the following testing of hypothesis and report your conclusion.

$$H_0 : \mu = \begin{bmatrix} 30 \\ 20 \\ 25 \end{bmatrix} \text{ against } H_a : \mu \neq \begin{bmatrix} 30 \\ 20 \\ 25 \end{bmatrix}.$$

6.) Explain in your words the purpose of response surface methods in the context of design and analysis of an experiment. In particular attempt to explain the following items:

i. Provide a brief outline of response surface methodology.

ii. What are the key differences and similarities of response surface design with a 2^k factorial experiment? When would you perform a response surface analysis?

iii. Illustrate with diagrams or figures the three types of stationary points that come along with a response surface analysis.

7.) Given the data, use the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ versus $H_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) A researcher wishes to estimate the average income of employees in a large firm. Records have the employees listed by seniority, and, generally speaking, salary increases with seniority. Discuss the relative merits of simple random sampling and stratified random sampling in this case.

9.) In the **SAS output on Pages 7-8** you see analysis of a data set. Based on this output, answer the following questions.

i. Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

ii. Do you see any evidence of multicollinearity? Give your reasons?

iii. What criteria is chosen for model selection? What is the best selected model?

iv. Do you think the best selected model has taken care of the multicollinearity?

10.) Housing finance researchers are interested in predicting current college students' chances of living in various levels of housing in five years, using parents' income (in thousands of dollars per year), gender (female, male), and race (African-American, Caucasian, Hispanic / Latino) as predictors. The researchers conducted a retrospective study in which 20 former college students (approximately five years removed from college) were randomly selected from each of the six combinations of gender and race (giving 120 total subjects). Each subject was asked to estimate his/her parents' annual income and to indicate whether he/she rents an apartment, rents a condo or townhouse, owns a condo or townhouse (without land), or owns a house (with land).

i. Propose an appropriate Generalized Linear Model (GLM) to predict the chances of reaching each level of housing. Clearly explain the meaning of each component and each parameter included in your model.

ii. Compare your proposed model from part i with at least one other possible model using a *different link function.*

iii. Assuming all 120 subjects provide different values for "parents' income," calculate the "error" degrees of freedom for your model.

iv. Explain the meaning of the "proportional odds assumption." Does your model include such an assumption?

v. Suppose a single parameter for "parents' income" is estimated to be $\hat{\beta} \approx 0.55$. Provide an interpretation of this parameter estimate.

SAS output for Question 2

The GLM Procedure

Dependent Variable: response rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	122.6838120	40.8946040	341.78	<.0001
Error	11	1.3161880	0.1196535		
Corrected Total	14	124.0000000			

R-Square	Coeff Var	Root MSE	response rate Mean
0.989386	1.192792	0.345910	29.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Color	2	7.6000000	3.8000000	31.76	<.0001
size	1	115.0838120	115.0838120	961.81	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Color	2	23.3918639	11.6959320	97.75	<.0001
size	1	115.0838120	115.0838120	961.81	<.0001

Least Squares Means

Color	response rate LSMEAN
Blue	29.1436089
Green	30.4884248
Orange	27.3679662

SAS output for Question 9

The SAS System

Model: MODEL1
Dependent Variable: y y

Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	7	5729.27961	818.46852	7.26	0.0003
Error	19	2140.83249	112.67539		
Corrected Total	26	7870.11210			

Root MSE	10.61487	**R-Square**	0.7280
Dependent Mean	24.73037	**Adj R-Sq**	0.6278
Coeff Var	42.92239		

Parameter Estimates									
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**	**Type II SS**	**Variance Inflation**
Intercept	Intercept	1	53.93702	57.42895	0.94	0.3594	16513	99.38965	0
x1	x1	1	-0.12765	0.28150	-0.45	0.6553	3909.46676	23.17091	3.67457
x2	x2	1	-0.22918	0.23264	-0.99	0.3370	0.06700	109.34435	7.72641
x3	x3	1	0.82485	0.76527	1.08	0.2946	271.28437	130.90367	19.20339
x4	x4	1	-0.43822	0.35855	-1.22	0.2366	49.20818	168.31209	7.46364
x5	x5	1	-0.00194	0.00965	-0.20	0.8431	417.17861	4.53643	4.69700
x6	x6	1	0.01989	0.00809	2.46	0.0237	704.98814	681.08914	7.73152
x7	x7	1	1.99349	1.08970	1.83	0.0831	377.08655	377.08655	1.11945

The SAS System

Model: MODEL1
Dependent Variable: y y

Collinearity Diagnostics										
			Proportion of Variation							
Number	**Eigenvalue**	**Condition Index**	**Intercept**	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**
1	6.32655	1.00000	0.00003102	0.00144	0.00001004	0.00002864	0.00060092	0.00133	0.00076761	0.00649
2	1.06079	2.44213	0.00000145	0.02325	3.02933E-7	0.00001692	0.00447	0.01945	0.01134	0.00918
3	0.43116	3.83055	0.00003233	0.00084028	0.00000755	0.00001076	0.00042097	0.00878	0.00448	0.88354
4	0.08564	8.59496	0.00027294	0.42757	0.00001098	0.00023206	0.19501	0.00024423	0.00205	0.00078779
5	0.06036	10.23816	0.00271	0.19364	0.00086074	0.00134	0.03862	0.55978	0.02657	0.03168
6	0.03157	14.15552	0.00190	0.16202	0.00085247	0.00251	0.06273	0.39946	0.76762	0.02151
7	0.00367	41.54161	0.16062	0.13203	0.00015639	0.15720	0.25780	0.00495	0.09246	0.04614
8	0.00026063	155.80205	0.83442	0.05921	0.99810	0.83866	0.44035	0.00600	0.09472	0.00067480

SAS output for Question 9-cont'd

Model: MODEL2
Dependent Variable: y

C(p) Selection Method

Number of Observations Read	27
Number of Observations Used	27

Number in Model	C(p)	R-Square	Variables in Model
2	-0.0208	0.6996	x6 x7

Model: MODEL2
Dependent Variable: y y

Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5506.27694	2753.13847	27.95	<.0001
Error	24	2363.83516	98.49313		
Corrected Total	26	7870.11210			

Root MSE	9.92437	**R-Square**	0.6996
Dependent Mean	24.73037	**Adj R-Sq**	0.6746
Coeff Var	40.13030		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	Intercept	1	2.52646	3.61005	0.70	0.4908	16513	48.23955	0
x6	x6	1	0.01852	0.00275	6.74	<.0001	5008.93619	4476.98336	1.02039
x7	x7	1	2.18575	0.97270	2.25	0.0341	497.34075	497.34075	1.02039

Model: MODEL2
Dependent Variable: y y

Collinearity Diagnostics

			Proportion of Variation		
Number	**Eigenvalue**	**Condition Index**	**Intercept**	**x6**	**x7**
1	2.44190	1.00000	0.04111	0.04703	0.05975
2	0.37728	2.54410	0.03123	0.28130	0.81885
3	0.18082	3.67487	0.92766	0.67167	0.12141

Applied Statistics Comprehensive Exam

August 2016

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:______________________________________ Date:____________

1.) A university health center tracks the number of flu-related visits during each month of the fall semester. The center director wonders whether students come down with the flu more often around midterm (mid-October) and final (mid December) exams. Can these data shed any light on this issue?

Flu-Related Visits to the University Health Center (by months)

September	October	November	December
20	48	27	56

Is there any significant difference among the flu-related visits during the fall semester? Use an α level of .05 to test the appropriate hypothesis.

2.) In an exercise study, pulse gains (the difference in the post-exercise and pre-exercise pulse rates) were computed on forty persons. In this study, gender (female or male) and smoking status (smoker or nonsmoker) for individuals were also registered. The summary data is in the table below.

Group	n	mean	Variance
MALE SMOKER	10	59.0	101.556
FEMALE SMOKER	10	85.4	913.822
MALE NONSMOKER	10	56.6	174.267
FEMALE NONSMOKER	10	63.8	72.400

Consider the group (given in the table) as a four-level factor and answer the following questions.

i. Write down a contrast that compares

 a) smokers and nonsmokers

 b) female and male

 c) female smokers and female nonsmokers

 d) male smokers and male nonsmokers

ii. Estimate all the contrasts in part i.

iii. Find the MSE for the one-way ANOVA with the group as the only factor.

iv. Use the Bonferroni method to test all the comparisons in i. Consider $\alpha_\Sigma = .05$.

3.) An experiment concerned the evaluation of eight drugs (factor A at $a = 8$ levels) for the treatment of arthritis. A second factor was the dose of the drug (factor B at $b = 2$ levels), and the third factor was the length of time (factor C at $c = 2$ levels) that a measurement was taken after injection by a substance known to cause an inflammatory reaction.

The experimental unit used in the study were $n = 64$ rats. The response was the amount of fluid (in milliliter) measured in the pleural cavity of an animal after having been administered a particular treatment combination.

In pharmacological studies, time of the day has an effect on the response due to changing laboratory conditions, etc. Consequently, the experiment was divided into blocks. It was possible to make the blocks sizes to be 32, each set of 32 observations being measured on a single day. Each treatment combination was measured once per day.

For the researcher, the effect of the drug (A) was of primary importance, and the effects of B and C were of interest only in the form of an interaction with A.

Propose a design for this experiment and justify your opinion. Please provide sufficient details on the following items:

i. The justification of your choice of the design

ii. Clearly indicate how would the factors in this experiment be used (for example, what factor to be confounded, what factor or factors to be interacted with others, and so on.)

iii. Create a dummy data table putting factors in rows or columns as appropriate. The response variable takes positive values between 5 and 15.

iv. Write the statistical model appropriate for your design and identify the model components.

v. Create the ANOVA table clearly showing the mean-squares expressions for each of them (i.e., the SS divided by appropriate denominator). You may use notations only, you do not have to write the actual mathematical expressions. You must show the columns such as the source of variation, df, SS, MS, and F.

4.) Consider the linear model defined in scalar notation by the following:

$$Y_{ij} = \mu_i + \beta(x_{ij} - \bar{x}_{..}) + \epsilon_{ij},$$

where $i = 1, 2, 3$, $j = 1, 2, 3, 4, 5$, and $\mathbf{x}^T = [4, 2, -1, 0, 3, 5, 5, 8, 6, 8, -3, -4, -1, 0, -1]$. ($\bar{x}_{..} = 2.07$)

i. Write the model in vector notation. Explicitly show $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$.

ii. Consider the hypothesis that all group means are equal, that is, $H_0 : \mu_1 = \mu_2, \mu_1 = \mu_3,$ *and* $\mu_2 = \mu_3$ (versus the alternative that at least one equality does not hold). Write this hypothesis as a General Linear Hypothesis, explicitly showing $\mathbf{C}^T$ and $\mathbf{d}$.

iii. Determine whether the hypothesis from part ii is testable.

iv. Assuming testability, explain the process that could be followed to test H_0.

5.) Respond to both parts of the question.

I. Briefly explain the concept of MANOVA understandable by someone with basic knowledge of ANOVA. When you answer this question, please address the following items.

i. Explanation of MANOVA and how does it relate to or differ from ANOVA?

ii. Name the test statistics to perform tests of significance in MANOVA.

iii. List the assumptions necessary to perform a MANOVA.

iv. Explain the concept of homogeneity of covariance matrices in the context of MANOVA. Why it is important to check for this assumption? What test statistic would you use for testing this assumption? If the p value for this test is 0.3001, what would be your conclusion abut this assumption?

II. Consider the following data. Provide some research questions that you would be able to answer using ANOVA and MANOVA for this data set. State your null and alternative hypotheses both in writing and using statistical terms. Sketch the MANOVA table related to your research question showing the sources of variation, the structure of the matrices of sum of squares within and sum of squares between, and degrees of freedom for each items in the MANOVA table.

Gender	Achievement Scores		
	Math	Science	Social Studies
Male	81	84	78
Male	88	91	86
Male	90	95	91
Female	83	82	94
Female	90	93	91
Female	85	87	88

6.) An experiment was conducted to evaluate in which of five sound models the experimenter best played a certain video game. The first three sound modes corresponded to three different types of background music, as well as game sounds expected to enhance play. The fourth mode had game sounds but no background music. The fifth mode had no music or game sounds. Denote these sound modes by the treatment factor levels 1-5, respectively.

The experimenter observed that the game required no warm up, that boredom and fatigue would be a factor after 4 to 6 games, and that his performance varied considerably on a day-to-day basis. Hence, he used a Latin square design, with the two blocking factors being "day" and "time order of the game".

The response measured was the game score, with higher scores being better. The design and resulting data are shown in the table below. The treatment factors are labeled 1-5 and the response is within the parenthesis.

		Day				
		1	2	3	4	5
	1	1 (94)	3 (100)	4 (98)	2 (101)	5 (112)
Time	2	3 (103)	2 (11)	1 (51)	5 (110)	4 (90)
Order	3	4 (114)	1 (75)	5 (94)	3 (85)	2 (100)
	4	5 (100)	4 (74)	2 (70)	1 (93)	3 (106)
	5	2 (106)	5 (95)	3 (81)	4 (90)	1 (73)

Analyze the data and draw conclusions. In particular:

i. State the model and identify the model components.

ii. State the null and alternative hypothesis both in terms of statistical notation and in writing.

iii. Analyze the data, create an ANOVA table.

iv. Draw conclusions.

7.) Discuss the differences between the Sign Test and Wilcoxons Signed Ranks Test. Include in your discussion the advantages / disadvantages of the two techniques.

8.) The Colorado Commission of Higher Education has recently hired you to estimate the average amount of scholarship money (in dollars) that each student receives per semester. Only state-supported universities/colleges in the state of Colorado are to be used. Private schools are not to be included in the population for this study. What type of sampling design would you use? Why? Explain, in detail, how you would obtain the data for your sample. What are the advantages/disadvantages of your design? What costs might be involved in collecting your data?

9.) In a study of faculty salaries in a small college in the Midwest, a linear regression model was fit, giving the fitted mean function,

$$\widehat{E(\text{Salary}|\text{Sex})} = 24697 - 3340 \times \text{Sex},$$

where Sex equals one if the faculty member was female and zero if male. The response Salary is measured in dollars.

i. Give a sentence that describes the meaning of the two estimated coefficients.

ii. An alternative mean function fit to these data with an additional term, Years, the number of years employed at this college, gives the estimated mean function,

$$\widehat{E(\text{Salary}|\text{Sex}, \text{Years})} = 18065 + 201 \times \text{Sex} + 759 \times \text{Years}.$$

Now give a sentence that describes the meaning of the estimated coefficient of Sex.

iii. The important difference between these two mean functions is that the coefficient for Sex has changed signs. Explain how this could happen.

10.) Health researchers are interested in modeling the likelihood of myocardial infarction (MI, "heart attack") using professional attributes. For their study they randomly selected 75 individuals between 55 and 65 years of age and recorded each individual's annual income (in thousands of dollars), whether the individual has a college degree, and whether the individual has experienced at least one MI within the last 10 years.

i. Describe some descriptive statistics that should be used to investigate the data (tables, plots, etc.). What does each descriptive statistic tell you about the response or predictors?

ii. Describe an appropriate statistical model that could be used to answer the research interests.

iii. Using the output on pages 7-8, assess the fit of the statistical model provided.

iv. Using the output on pages 7-8, provide interpretations of the effects of both income and college degree.

v. Assume researchers had instead gathered data on the *number* of heart attacks within the last ten years. Explain how your model from part ii would change to accommodate this new response.

SAS Output for Question 10

'

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MIDATA
Response Variable	MI
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Probability modeled is MI='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	101.106	86.535
SC	103.423	93.487
-2 Log L	99.106	80.535

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	18.5712	2	<.0001
Score	16.2263	2	0.0003
Wald	12.7642	2	0.0017

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-15.0145	5.0909	8.6985	0.0032
Income	1	0.1504	0.0504	8.9175	0.0028
CollegeDegree	1	1.8937	1.1298	2.8092	0.0937

SAS Output for Question 10

The LOGISTIC Procedure

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Income	1.162	1.053	1.283
CollegeDegree	6.644	0.726	60.830

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.3	Somers' D	0.568
Percent Discordant	21.6	Gamma	0.568
Percent Tied	0.1	Tau-a	0.269
Pairs	1316	c	0.784

Partition for the Hosmer and Lemeshow Test					
		MI = 1		MI = 0	
Group	Total	Observed	Expected	Observed	Expected
1	8	1	0.36	7	7.64
2	8	0	0.84	8	7.16
3	8	2	1.57	6	6.43
4	8	4	2.21	4	5.79
5	8	1	2.81	7	5.19
6	8	1	3.35	7	4.65
7	8	4	3.84	4	4.16
8	8	5	4.60	3	3.40
9	11	10	8.41	1	2.59

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.2674	7	0.1739

Applied Statistics Comprehensive Exam

January 2016

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) Over the past 5 years, an insurance company has had a mix of 40% whole life policies, 20% universal life policies, 25% annual renewable-term (ART) policies, and 15% other types of policies. A change in this mix over the long haul could require a change in the commission structure, reserves, and possibly investments. A sample of 1,000 policies issued over the last few months gave the results shown here.

Category	n
Whole life	320
Universal life	280
ART	240
Other	160

Use these data to assess whether there has been a shift from the historical percentages. Use the 5% significant level. Clearly identify the hypothesis you are testing, report the test statistic and state your conclusion.

2.) A small travel agency is interested to better understand the effect of age of customer (x) on the amount of money spent in a tour (y), in the last twelve months. Agency has recognized two important customer segments. The first segment, which we will denote by A, consists of those customers who have purchased an adventure tour in the last twelve months. The second segment, which we will denote by C, consists of those customers who have purchased a cultural tour in the last twelve months. Note that the two segments are completely separate in the sense that there are no customers who are in both segments. Assume the relation between y and x is linear.

i. Write down a single multiple linear model with possibly differing intercepts and slopes for two different segments of the customers. Interpret each parameter in the model.

ii. Write down the hypothesis that the two lines are coincident in terms of the regression coefficients.

3.) The effect of five different ingredients A, B, C, D, E on the reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximatley 1.5 hours, so only five runs can be made in one day. The experimenter wants to control the two sources of variation and choses an appropriate design for this.

The data are given in the table below. Answer the following questions.

Table 1: Data for the reaction time of a chemical process

Batch	Day				
	1	2	3	4	5
1	A=8	B=7	C=1	D=7	E=3
2	B=11	C=2	D=7	E=3	A=8
3	C=4	D=9	E=10	A=1	B=5
4	D=6	E=8	A=6	B=6	C=10
5	E=4	A=2	B=3	C=8	D=8

i. What is the name of this design?

ii. There are two nuisance factors to be "averaged out" in the design. What are those?

iii. Calculate the total corrected sum of squares for this experiment as well as all other required sum of squares.

iv. State the hypothesis, and construct the analysis of variance table including the calculated F-statistic. Comment on the results.

4.) Consider a Two-Factor ANOVA model (with interaction),

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2$.

i. Write this model in vector form, showing all components explicitly. Determine the rank of your design matrix, $\mathbf{X}$.

ii. Provide an expression for the Least Squares Estimators (LSE) for the parameters of your model.

iii. Explain the meaning of the term "estimable." Why are we concerned with estimability?

iv. Determine whether each of the following functions is estimable:

$$\mu + \alpha_1,\ \beta_2 - \beta_1,\ \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$$

v. For each of the estimable functions from part iv, provide an expression for the Best Linear Unbiased Estimator, and explain in what sense these are "best".

5.) Amitriptyline is prescribed by some physicians as an antidepressant. However, there are also conjectured side effects that seem to be related to the use of the drug: irregular heartbeat, abnormal blood pressures, and irregular waves on the electrocardiogram, among other things. Data gathered on 17 patients who were admitted to the hospital after an amitriptyline overdose are given in the Output 1.1 on Page 7.

The two response (dependent) variables are

Y_1 = Total TACD plasma level (TOT)
Y_2 = Amount of amitriptyline present in TCAD plasma level (AMI)

The five predictor (independent) variables are

X_1 = Gender: 1 if female, 0 if male (GEN)
X_2 = Amount of antidepressants taken at time of overdose (AMT)
X_3 = PR wave measurement (PR)
X_4 = Diastolic blood pressure (DIAP)
X_5 = QRS wave measurement (QRS)

i. Perform a regression analysis using only the first response (Y_1). Use Output 1.2 on Page 8 to write the regression model that predicts TOT (Y_1) from AMT, PR, DIAP, and QRS. Comment on the overall significance of the model.

ii. Use the fitted model (i.e., the estimated model) above to predict total TCAD plasma level (TOT) for a patient whose AMT=750, PR=200, DIAP=80, and QRS=90.

iii. For the model fitted in a), suppose, we are interested in testing for the significance of AMT in predicting TOT. Write the appropriate null and alternative hypothesis that you would be testing for this. Comment on the significance of AMT.

iv. Construct a 95% confidence interval for the parameter associated with AMT. Comment on the interval. Is the CI consistent with the test result in part iii?

v. Suppose you would like to test for significance of AMT and DIAP together. For this we would perform a partial F-test by fitting two models–one with all the predictors, and the other without AMT and DIAP. See the Output 1.3 on Page 9.

   Write appropriate null and alternative hypotheses and carry out the test. Calculate the value of the appropriate test statistic. Comment on your findings.

vi. List the assumptions concerning the model fitted in part i. Analyze the residuals. Are there any sign of violation of any of the regression assumptions?

6.) Suppose a design of Resolution III is desired for $N - 1 = 7$ factors in $N = 8$ runs. Each factor is considered to have two levels– low and high. The eight runs would be a 1/16th fraction of the $2^7 = 128$ runs. This design is a 2_{III}^{7-4} fractional factorial. Suppose the design generators are ABD, ACE, BCF, and $ABCG$.

Answer the following questions.

i. Construct the standard order table for the above design.

ii. Show/derive the complete defining relation for the above design.

iii. Show the complete alias structure for the above design.

iv. When do you think such a design would be applicable. Give a real life example with details of factors and their levels.

7.) Given the data, use the Wilcoxon Signed Ranks Test to test:

$$H_0 : \tilde{\mu} = 107 \text{ vs } H_1 : \tilde{\mu} \neq 107.$$

$$99, 100, 90, 94, 135, 108, 107, 111, 119, 104, 127, 109, 117, 105, 125$$

8.) A researcher wishes to estimate the average income of employees in a large firm. Records have the employees listed by seniority, and, generally speaking, salary increases with seniority. Discuss the relative merits of simple random sampling and stratified random sampling in this case.

9.) The carbonation level of a soft drink beverage is affected by the temperature of the product and the filler operating pressure. Twelve observations were obtained and analyzed using SAS software. Use the SAS output on Page 10 to answer the following questions.

i. Fit a second-order polynomial.

ii. Test for significance of regression.

iii. What is the lack of fit test for? Why can we do this test for this data set? Test for lack of fit and draw conclusions.

iv. Does the interaction term contribute significantly to the model?

v. Do the second-order terms contribute significantly to the model?

vi. Is there any sign of multicollinearity? If yes, how would you deal with it?

10.) A financial analyst is interested in factors that impact fraudulent activity on credit cards. Using historical data, she has collected monthly counts of fraudulent activities identified on various credit cards supported by regional stores. She has also recorded whether each card is available for international use, whether each card was made available to individuals under the age of 18, whether each card was awarded with a promotional gift, and a continuous measure of each card's monthly usage volume. The analyst would like to determine whether any of the recorded factors has an effect on the volume of monthly fraudulent activities.

i. Propose some descriptive statistics that could be used to help answer the analyst's questions.

ii. Using the SAS output on Pages 11-13, describe the response of interest, monthly fraudulent activities.

iii. Propose an appropriate model that could be used to determine whether any of the associated factors affect monthly fraudulent activity.

iv. Using the SAS output on Pages 11-13, perform an analysis of the data. Be sure to present conclusions using the language of the original question.

v. Suppose the experiment is expanded to include *all* US credit cards. When the analyst collects a similar but larger data set, she notices that approximately 40% of the observations are zeros. How would you change your analysis approach to account for this phenomenon?

SAS Output for Question 5

1

Output 1.1: Drug Overdose Data Set

Obs	tot	ami	gen	amt	pr	diap	qrs
1	3389	3149	1	7500	220	0	140
2	1101	653	1	1975	200	0	100
3	1131	810	0	3600	205	60	111
4	596	448	1	675	160	60	120
5	896	844	1	750	185	70	83
6	1767	1450	1	2500	180	60	80
7	807	493	1	350	154	80	98
8	1111	941	0	1500	200	70	93
9	645	547	1	375	137	60	105
10	628	392	1	1050	167	60	74
11	1360	1283	1	3000	180	60	80
12	652	458	1	450	160	64	60
13	860	722	1	1750	135	90	79
14	500	384	0	2000	160	60	80
15	781	501	0	4500	180	0	100
16	1070	405	0	1500	170	90	120
17	1754	1520	1	3000	180	0	129

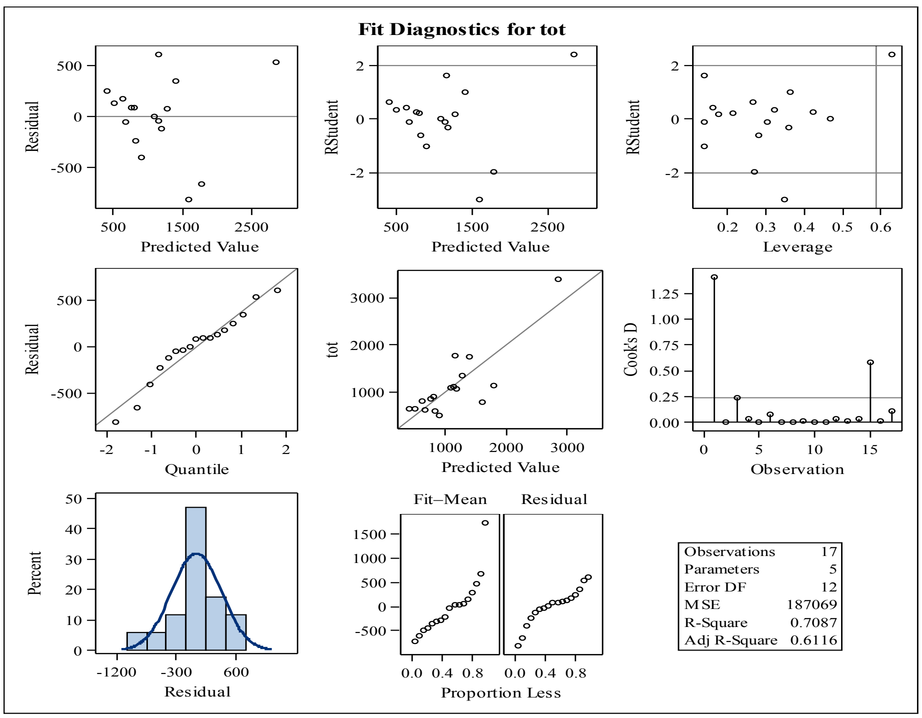
The REG Procedure
Model: MODEL1
Dependent Variable: tot

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5461108	1365277	7.30	0.0032
Error	12	2244832	187069		
Corrected Total	16	7705940			

SAS Output for Question 5

Output 1.2
Regression of AMT PR DIAP QRS on TOT

Parameter Estimates								
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	1	-1308.68851	1245.59727	-1.05	0.3141	0	-4022.61182	1405.23481
amt	1	0.24839	0.09271	2.68	0.0201	2.45782	0.04641	0.45038
pr	1	6.41246	6.38695	1.00	0.3352	1.88054	-7.50352	20.32844
diap	1	3.06008	4.71297	0.65	0.5284	1.86750	-7.20860	13.32877
qrs	1	6.33545	5.90093	1.07	0.3041	1.39799	-6.52157	19.19247



SAS Output for Question 5

3

Output 1.3
Partial Significance of AMT and DIAP

The REG Procedure
Model: FullModel
Dependent Variable: tot

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	5461108	1365277	7.30	0.0032
Error	12	2244832	187069		
Corrected Total	16	7705940			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-1308.68851	1245.59727	-1.05	0.3141
amt	1	0.24839	0.09271	2.68	0.0201
pr	1	6.41246	6.38695	1.00	0.3352
diap	1	3.06008	4.71297	0.65	0.5284
qrs	1	6.33545	5.90093	1.07	0.3041

The REG Procedure
Model: Submodel
Dependent Variable: tot

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	4081088	2040544	7.88	0.0051
Error	14	3624853	258918		
Corrected Total	16	7705940			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-2675.83752	989.89628	-2.70	0.0172
pr	1	15.57400	5.94574	2.62	0.0202
qrs	1	11.03858	6.37119	1.73	0.1051

SAS Output for Question 9

SAS output for question 9

The REG Procedure
Model: MODEL1
Dependent Variable: y Carbonation

Number of Observations Read	12
Number of Observations Used	12

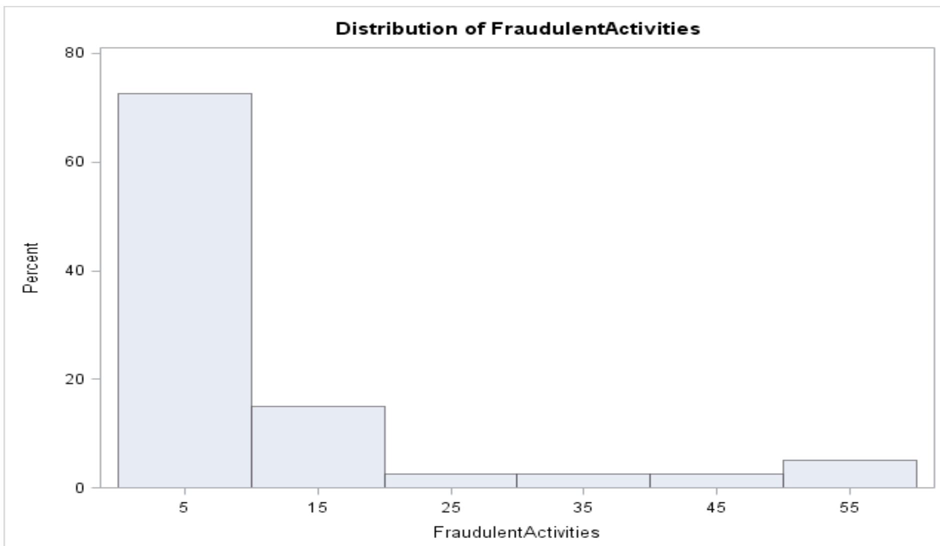
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	339.88774	67.97755	177.17	<.0001
Error	6	2.30216	0.38369		
Lack of Fit	3	0.72590	0.24197	0.46	0.7297
Pure Error	3	1.57627	0.52542		
Corrected Total	11	342.18990			

Root MSE	0.61943	**R-Square**	0.9933
Dependent Mean	7.94500	**Adj R-Sq**	0.9877
Coeff Var	7.79648		

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	1	3025.31858	2045.74638	1.48	0.1897	0
x1	Temperature	1	-194.27289	132.06428	-1.47	0.1917	90911
x2	Pressure	1	-6.05067	20.60625	-0.29	0.7789	15401
x1_2	$x1^2$	1	3.62587	2.20978	1.64	0.1519	97845
x2_2	$x2^2$	1	1.15425	0.32373	3.57	0.0118	7759.76159
x1x2	x1*x2	1	-1.33171	0.89619	-1.49	0.1878	37872

SAS Output for Question 10

Moments			
N	40	**Sum Weights**	40
Mean	8.9	**Sum Observations**	356
Std Deviation	14.9611463	**Variance**	223.835897
Skewness	2.18258178	**Kurtosis**	4.05213894
Uncorrected SS	11898	**Corrected SS**	8729.6
Coeff Variation	168.102767	**Std Error Mean**	2.36556493



SAS Output for Question 10

Model Information	
Data Set	WORK.CREDITDATA
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	FraudulentActivities

Number of Observations Read	40
Number of Observations Used	34
Missing Values	6

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	27	57.1062	2.1150
Scaled Deviance	27	57.1062	2.1150
Pearson Chi-Square	27	53.0935	1.9664
Scaled Pearson X2	27	53.0935	1.9664
Log Likelihood		758.8075	
Full Log Likelihood		-77.4864	
AIC (smaller is better)		168.9727	
AICC (smaller is better)		173.2804	
BIC (smaller is better)		179.6573	

Algorithm converged.

SAS Output for Question 10

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.7512	1.0850	-4.8778	-0.6246	6.43	0.0112
International	1	-2.7066	1.4921	-5.6311	0.2178	3.29	0.0697
Under18	1	-1.5216	1.1519	-3.7792	0.7360	1.75	0.1865
GiftPromotion	1	0.1837	0.1301	-0.0713	0.4387	1.99	0.1580
UsageVolume	1	0.6157	0.1327	0.3556	0.8757	21.53	<.0001
Under18*UsageVolume	1	0.2073	0.1400	-0.0671	0.4817	2.19	0.1387
Internati*UsageVolum	1	0.2341	0.1653	-0.0899	0.5582	2.01	0.1568
Dispersion	0	0.0000	0.0000	.	.		

Note: The negative binomial dispersion parameter was held fixed.

Lagrange Multiplier Statistics			
Parameter	**Chi-Square**	**Pr > ChiSq**	
Dispersion	0.8224	0.1822	*
*** One-sided p-value**			

Applied Statistics Comprehensive Exam

August 2015

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:________________________________ Date:____________

1.) [**This problem should be answered based on a 1-page SAS output that follows all questions, beginning on Page 7.**]

A team of researchers is interested in determining whether two methods of hypnotic induction, I and II, differ with respect to their effectiveness. They begin by randomly sorting 20 volunteer subjects into two independent groups of 10 subjects each, with the aim of administering Method I to one group and Method II to the other. Before either of the induction methods is administered, each subject is pre-measured on a standard index of "primary suggestibility," which is a variable known to be correlated with receptivity to hypnotic induction group. Use the SAS output to answer the following questions.

i) Identify the dependent and independent variables.

ii) Write an appropriate model for this scenario.

iii) Specify appropriate null and alternate hypotheses for testing the linear relationship between the dependent variable and standard index of primary suggestibility. Report values for the F-statistic, degrees of freedom, and p-value and state the conclusion of this test.

iv) Specify appropriate null and alternate hypotheses for testing the group effect. Report the values for the F-statistic, degrees of freedom, and p-value and state the conclusion of this test.

2.) For each of the following scenarios, write the null and alternative hypotheses, and the statistical method and test statistic to be used.

i) The director of food services in a school district is considering the addition of new items to the cafeteria menu. One of the new items is a green salad topped with strips of grilled chicken breast. After tasting the salad, students in the district's elementary, middle, and high schools are asked to indicate their preference by circling one of the following options: (a) Add it to the menu, (b) Do not add it to the menu, and (c) No opinion. The director of food services analyzes the data to determine if there are differences in the numbers of students in the district's elementary, middle, and high schools that chose each of the three response options.

ii) A statistics instructor at a liberal arts college has noticed that psychology and sociology students seem to have more positive attitudes toward statistics compared with history and English students. The professor administers the Statistics Attitudes Inventory (SAI) scale to all students on the first day of the fall semester. The inventory contains twenty Likert-scale items with responses ranging from Strongly Disagree to Strongly Agree. The responses of psychology, sociology, English, and history students are then compared to determine if there are significant differences in attitudes among the four groups of students.

Question 2, Continued

iii) Many school districts administer readiness tests to students upon kindergarten entry. A publisher of a new kindergarten readiness test wants to convince potential users of the test that it can accurately predict students' academic performance in first grade. The test publisher offers to administer the readiness test, free of charge, to all kindergarten students in the district. After the test is administered, the readiness test score for each student is recorded. At the end of first grade, all students are administered a standardized achievement test. The readiness test scores obtained a year earlier and the scores on the standardized achievement test administered at the end of the first grade are studied to determine whether, as the readiness test publisher predicted, the readiness test can serve as a good predictor of end-of-year achievement of first-grade students.

3.) Respond to the following.

i) Explain in non-technical terms the concept of analysis of variance (ANOVA), and write down the fundamental ANOVA identity. (This question is not about the use of ANOVA for testing several population means.)

ii) ANOVA is used to test for the equality of several treatments. In performing ANOVA, we compare mean squares for treatments (MS_{Treat}) and mean squares for errors (MS_{Err}). Technically, the expected value for MS_{Treat} is the true population variance, σ^2.

Explain in your words: How does the comparison of these two mean squares help us testing for the treatment effects?

4.) Consider a Blocked One-Factor ANOVA model,

$$Y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij},$$

where $i = 1, \ldots, 3$ indicates the three groups of interest, $j = 1, \ldots, 4$ indicates the four blocks, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, independent.

i. Present the response vector $\mathbf{Y}$, parameter vector $\boldsymbol{\beta}$ and a *full-rank* design matrix $\mathbf{X}$.

ii. For what values of i and j is $\mu + \alpha_i + b_j$ estimable? Justify your answer.

iii. Find an expression for the BLUE of $\mu + \alpha_1 + b_1$, and explain in what sense it is "best."

iv. Find an expression for the variance of the BLUE of $\mu + \alpha_1 + b_1$, and explain how this variance compares to the variance of other estimators of $\mu + \alpha_1 + b_1$.

5.) [**This problem should be answered based on a 7-page SAS output that follows all questions, beginning on Page 8.**]

The admissions officer of a graduate school has used an "index" of undergraduate GPA and graduate management aptitude test (GMAT) scores to help decide which applicants should be admitted to the graduate programs. The scatter plot of GPA vs GMAT (shown in the attached SAS output) shows recent applicants who have been classified as "Admit (A)", "Borderline (B)", and "Reject (R)".

A discriminant analysis and classification have been performed on the data and the results are shown in the attached SAS output. Answer the following questions. **Note:** when you answer, make sure to include the associated statistics. For example, if you decide to reject a null hypothesis, you should mention the value of the appropriate test statistic and the corresponding p-value.

i) Is there significant association between admission status (admitted, rejected, borderline) and the scores on GPA and GMAT?

ii) If there is significant association, we would like to perform a discriminant analysis. How many discriminant functions (DF) are possible for the given problem?

iii) Comment on the significance of the discriminant function(s).

iv) What is the overall effect size for the discriminant analysis? Comment on the effect size of each of the discriminant functions.

v) Write the classification functions corresponding to each discriminant function. Use the classification function(s) for classifying an applicant as "Admit" or "Reject" or "Borderline" who has GPA = 3.7 and GMAT score = 650

vi) In the SAS output, both resubstitution summary and crossvalidation summary for classification are provided. Comment on the error of misclassification based on these output.

vii) Is there a reason to believe that the classification function produces noticeably higher error rate than what we would have obtained by chance alone? Would you use the discriminant functions obtained from this analysis to classify an applicant to either admit, reject, or borderline? Justify.

6.) [**This problem should be answered based on a 14-page SAS output that follows all questions, beginning on Page 15.**]

An experiment on the yield of three varieties of oats (factor A) and four different levels of manure (factor B) was described by F. Yates in his 1935 paper *Complex Experiments.* The experimental area was divided into 6 blocks. Each of these was then subdivided into 3 whole plots.

The varieties of oat were sown on the whole plots according to a randomized complete block design (so that every variety appeared in every block exactly once). Each whole plot was then divided into 4 split plots, and the levels of manure were applied to the split plots according to a randomized complete block design (so that every level of B appeared in every whole plot exactly once).

The design, after randomization, is shown in the Table below. The yield is measured in quarter pounds.

Table 1: Split-plot design and yields (in quarter lb) for the oat experiment.

Block	Level of A	Level of B (yield)		Block	Level of A	Level of B (yield)	
1	2	3(156)	2(118)	2	2	2(109)	3(99)
		1(140)	0(105)			0(63)	1(70)
	0	0(111)	1(130)		1	0(80)	2(94)
		3(174)	2(157)			3(126)	1(82)
	1	0(117)	1(114)		0	1(90)	2(100)
		2(161)	3(141)			3(116)	0(62)
3	2	2(104)	0(70)	4	1	3(96)	0(60)
		1(89)	3(117)			2(89)	1(102)
	0	3(122)	0(74)		0	2(112)	3(86)
		1(89)	2(81)			0(68)	1(64)
	1	1(103)	0(64)		2	2(132)	3(124)
		2(132)	3(133)			1(129)	0(89)
5	1	1(108)	2(126)	6	0	2(118)	0(53)
		3(149)	0(70)			3(113)	1(74)
	2	3(144)	1(124)		1	3(104)	2(86)
		2(121)	0(96)			0(89)	1(82)
	0	0(61)	3(100)		2	0(97)	1(99)
		1(91)	2(97)			2(119)	3(121)

i) Explain briefly why it is a split-plot design?

ii) List the whole plot and split-plot factors along with their levels and types. Is there any or more factors that you would consider as random? Justify.

iii) Write down the statistical model that you consider appropriate for predicting yield and define all the terms in the model.

iv) Now, consider someone else has conducted an analysis of the data. Only the SAS output is available. Use the results produced by SAS to create a table to show the expected mean squares expressions.

v) Use the provided SAS output to draw conclusions.

7.) Given the data, use the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ versus $H_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) Compare and contrast stratified sampling to simple random sampling. What do these designs have in common? How are they different? Give examples/applications of each design. Under what conditions is stratified sampling preferred over simple random sampling?

9.) [**This problem should be answered based on a 1-page SAS output that follows all questions, beginning on Page 29.**]

An investigator is interested in understanding the relationship, if any, between the house selling price (price) and size of home in square feet (size), number of bedrooms (beds), number of bathrooms (baths), anual taxes (Taxes), and house condition (New), which takes the values of new or not new. Data were collected for 100 houses sold in a large city. Use the SAS output provided to answer the following questions. In this output, IntNT represents the interaction between Taxes and New.

i) Specify appropriate null and alternate hypotheses for testing the adequacy of the overall model and state the conclusion of this test.

ii) Do a group test for the effects of Taxes, New, and their interaction.

iii) Describe the differences between two models; the model with the interaction of New and Taxes, and the model without this interaction.

iv) Test the partial effect of New in two models; the model with the interaction of New and Taxes, and the model without this interaction.

v) What would you tell to the investigator for the effect of New?

10.) Consider an experiment intended to evaluate the relationships between the volume of patients who utilize local Urgent Care (UC) services and the properties of the different UC locations. To collect data, various research assistants are recruited to sit in different UC waiting rooms and count the number of patients who enter; however, research assistants cannot be trusted to sit in waiting rooms for the same amounts of time. In addition, the number of staff and the age of each location is also recorded.

i) Clearly present a standard count regression model that could be used to model the mean number of patients using staff and age as predictors. What are the assumptions of your model?

ii) Assume the parameter estimate for "staff" is $\hat{\beta}_s = 1.45$. How would you interpret this estimate?

iii) What is "overdispersion"? What are the consequences of ignoring overdispersion in a count regression model? Explain why overdispersion can be expected for this research situation.

iv) Clearly describe two options for accounting for overdispersion in these data (assume you have recorded a variable for "research assistant" that can be used to group the observations). Compare your two options to each other.

SAS output for Question 1

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
method	2	I II

Number of Observations Read	20
Number of Observations Used	20

The GLM Procedure

Dependent Variable: effectiveness

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	763.2920226	381.6460113	44.54	<.0001
Error	17	145.6579774	8.5681163		
Corrected Total	19	908.9500000			

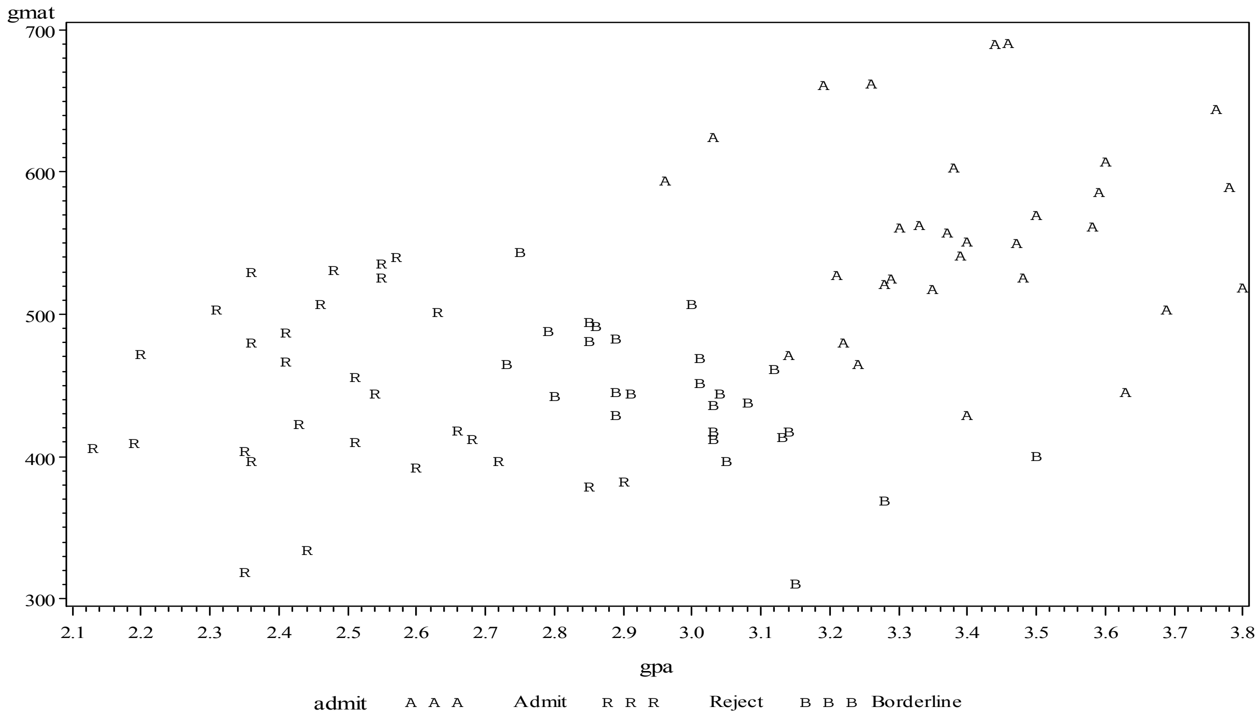
R-Square	Coeff Var	Root MSE	effectiveness Mean
0.839751	18.82402	2.927134	15.55000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
primary_suggestibili	1	585.8772306	585.8772306	68.38	<.0001
method	1	177.4147921	177.4147921	20.71	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
primary_suggestibili	1	643.2420226	643.2420226	75.07	<.0001
method	1	177.4147921	177.4147921	20.71	0.0003

SAS output for question 5

1



The SAS System

The DISCRIM Procedure

Total Sample Size	85	**DF Total**	84
Variables	2	**DF Within Classes**	82
Classes	3	**DF Between Classes**	2

Number of Observations Read	85
Number of Observations Used	85

Class Level Information					
admit	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
Admit	Admit	31	31.0000	0.364706	0.333333
Borderline	Borderline	26	26.0000	0.305882	0.333333
Reject	Reject	28	28.0000	0.329412	0.333333

Pooled Covariance Matrix Information	
Covariance Matrix Rank	**Natural Log of the Determinant of the Covariance Matrix**
2	4.85035

3

The SAS System

The DISCRIM Procedure
Canonical Discriminant Analysis

Generalized Squared Distance to admit			
From admit	Admit	Borderline	Reject
Admit	0	10.06344	31.28880
Borderline	10.06344	0	7.43364
Reject	31.28880	7.43364	0

Multivariate Statistics and F Approximations					
S=2 M=-0.5 N=39.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.12637661	73.43	4	162	<.0001
Pillai's Trace	1.00963002	41.80	4	164	<.0001
Hotelling-Lawley Trace	5.83665601	117.72	4	96.17	<.0001
Roy's Greatest Root	5.64604452	231.49	2	82	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.921702	0.920516	0.016417	0.849535	5.6460	5.4554	0.9673	0.9673
2	0.400119	.	0.091641	0.160095	0.1906		0.0327	1.0000

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.12637661	73.43	4	162	<.0001
2	0.83990454	15.63	1	82	0.0002

4

The SAS System

The DISCRIM Procedure
Canonical Discriminant Analysis

Total Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.969922	-0.243416
gmat	0.662832	0.748768

Between Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.994118	-0.108305
gmat	0.897852	0.440298

Pooled Within Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.860161	-0.510023
gmat	0.350860	0.936428

5

The SAS System

The DISCRIM Procedure

Total-Sample Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	2.148737595	-0.805087984
gmat	0.698531804	1.178084322

Pooled Within-Class Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	0.9512430832	-.3564113077
gmat	0.5180918168	0.8737695880

Raw Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	5.008766354	-1.876682204
gmat	0.008568593	0.014451060

Class Means on Canonical Variables		
admit	**Can1**	**Can2**
Admit	2.773788370	0.246102784
Borderline	-0.271055133	-0.644045724
Reject	-2.819285930	0.325571519

Linear Discriminant Function for admit			
Variable	**Admit**	**Borderline**	**Reject**
Constant	-240.37168	-177.31575	-133.89892
gpa	106.24991	92.66953	78.08637
gmat	0.21218	0.17323	0.16541

6

The SAS System

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.GPA
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into admit				
From admit	Admit	Borderline	Reject	Total
Admit	27 87.10	4 12.90	0 0.00	31 100.00
Borderline	1 3.85	25 96.15	0 0.00	26 100.00
Reject	0 0.00	2 7.14	26 92.86	28 100.00
Total	28 32.94	31 36.47	26 30.59	85 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for admit				
	Admit	Borderline	Reject	Total
Rate	0.1290	0.0385	0.0714	0.0796
Priors	0.3333	0.3333	0.3333	

7

The SAS System

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.GPA
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into admit				
From admit	**Admit**	**Borderline**	**Reject**	**Total**
Admit	26 83.87	5 16.13	0 0.00	31 100.00
Borderline	1 3.85	24 92.31	1 3.85	26 100.00
Reject	0 0.00	2 7.14	26 92.86	28 100.00
Total	27 31.76	31 36.47	27 31.76	85 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for admit				
	Admit	**Borderline**	**Reject**	**Total**
Rate	0.1613	0.0769	0.0714	0.1032
Priors	0.3333	0.3333	0.3333	

SAS output for question 6

1

Oats Experiment Data

Obs	block	WP	A	B	yield
1	1	1	2	3	156
2	1	1	2	1	140
3	1	1	2	2	118
4	1	1	2	0	105
5	1	2	0	0	111
6	1	2	0	3	174
7	1	2	0	1	130
8	1	2	0	2	157
9	1	3	1	0	117
10	1	3	1	2	161
11	1	3	1	1	114
12	1	3	1	3	141
13	2	1	2	2	109
14	2	1	2	0	63
15	2	1	2	3	99
16	2	1	2	1	70
17	2	2	1	0	80
18	2	2	1	3	126
19	2	2	1	2	94
20	2	2	1	1	82
21	2	3	0	1	90
22	2	3	0	3	116
23	2	3	0	2	100
24	2	3	0	0	62
25	3	1	2	2	104
26	3	1	2	1	89
27	3	1	2	0	70
28	3	1	2	3	117
29	3	2	0	3	122
30	3	2	0	1	89
31	3	2	0	0	74
32	3	2	0	2	81
33	3	3	1	1	103
34	3	3	1	2	132
35	3	3	1	0	64

2

Oats Experiment Data

Obs	block	WP	A	B	yield
36	3	3	1	3	133
37	4	1	1	3	96
38	4	1	1	2	89
39	4	1	1	0	60
40	4	1	1	1	102
41	4	2	0	2	112
42	4	2	0	0	68
43	4	2	0	3	86
44	4	2	0	1	64
45	4	3	2	2	132
46	4	3	2	1	129
47	4	3	2	3	124
48	4	3	2	0	89
49	5	1	1	1	108
50	5	1	1	3	149
51	5	1	1	2	126
52	5	1	1	0	70
53	5	2	2	3	144
54	5	2	2	2	121
55	5	2	2	1	124
56	5	2	2	0	96
57	5	3	0	0	61
58	5	3	0	1	91
59	5	3	0	3	100
60	5	3	0	2	97
61	6	1	0	2	118
62	6	1	0	3	113
63	6	1	0	0	53
64	6	1	0	1	74
65	6	2	1	3	104
66	6	2	1	0	89
67	6	2	1	2	86
68	6	2	1	1	82
69	6	3	2	0	97
70	6	3	2	2	119

3

Oats Experiment Data

Obs	block	WP	A	B	yield
71	6	3	2	1	99
72	6	3	2	3	121

4

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
A	3	0 1 2
B	4	0 1 2 3
block	6	1 2 3 4 5 6

Number of Observations Read	72
Number of Observations Used	72

The GLM Procedure

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	44017.19444	1692.96902	9.56	<.0001
Error	45	7968.75000	177.08333		
Corrected Total	71	51985.94444			

R-Square	Coeff Var	Root MSE	yield Mean
0.846713	12.79887	13.30727	103.9722

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	5	15875.27778	3175.05556	17.93	<.0001
A	2	1786.36111	893.18056	5.04	0.0106
A*block	10	6013.30556	601.33056	3.40	0.0023
B	3	20020.50000	6673.50000	37.69	<.0001
A*B	6	321.75000	53.62500	0.30	0.9322

Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	5	15875.27778	3175.05556	17.93	<.0001
A	2	1786.36111	893.18056	5.04	0.0106
A*block	10	6013.30556	601.33056	3.40	0.0023
B	3	20020.50000	6673.50000	37.69	<.0001
A*B	6	321.75000	53.62500	0.30	0.9322

The GLM Procedure

5

Source	Type III Expected Mean Square
block	Var(Error) + 4 Var(A*block) + 12 Var(block)
A	Var(Error) + 4 Var(A*block) + Q(A,A*B)
A*block	Var(Error) + 4 Var(A*block)
B	Var(Error) + Q(B,A*B)
A*B	Var(Error) + Q(A*B)

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: yield

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	block	5	15875	3175.055556	5.28	0.0124
*	A	2	1786.361111	893.180556	1.49	0.2724
	Error: MS(A*block)	10	6013.305556	601.330556		
* This test assumes one or more other fixed effects are zero.						

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	A*block	10	6013.305556	601.330556	3.40	0.0023
*	B	3	20021	6673.500000	37.69	<.0001
	A*B	6	321.750000	53.625000	0.30	0.9322
	Error: MS(Error)	45	7968.750000	177.083333		
* This test assumes one or more other fixed effects are zero.						

The GLM Procedure

Class Level Information		
Class	Levels	Values
A	3	0 1 2
B	4	0 1 2 3
block	6	1 2 3 4 5 6

Number of Observations Read	72
Number of Observations Used	72

The GLM Procedure

6

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	44017.19444	1692.96902	9.56	<.0001
Error	45	7968.75000	177.08333		
Corrected Total	71	51985.94444			

R-Square	Coeff Var	Root MSE	yield Mean
0.846713	12.79887	13.30727	103.9722

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	5	15875.27778	3175.05556	17.93	<.0001
A	2	1786.36111	893.18056	5.04	0.0106
A*block	10	6013.30556	601.33056	3.40	0.0023
B	3	20020.50000	6673.50000	37.69	<.0001
A*B	6	321.75000	53.62500	0.30	0.9322

Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	5	15875.27778	3175.05556	17.93	<.0001
A	2	1786.36111	893.18056	5.04	0.0106
A*block	10	6013.30556	601.33056	3.40	0.0023
B	3	20020.50000	6673.50000	37.69	<.0001
A*B	6	321.75000	53.62500	0.30	0.9322

The GLM Procedure

Source	Type III Expected Mean Square
block	Var(Error) + 4 Var(A*block) + 12 Var(block)
A	Var(Error) + 4 Var(A*block) + Q(A,A*B)
A*block	Var(Error) + 4 Var(A*block)
B	Var(Error) + Q(B,A*B)
A*B	Var(Error) + Q(A*B)

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: yield

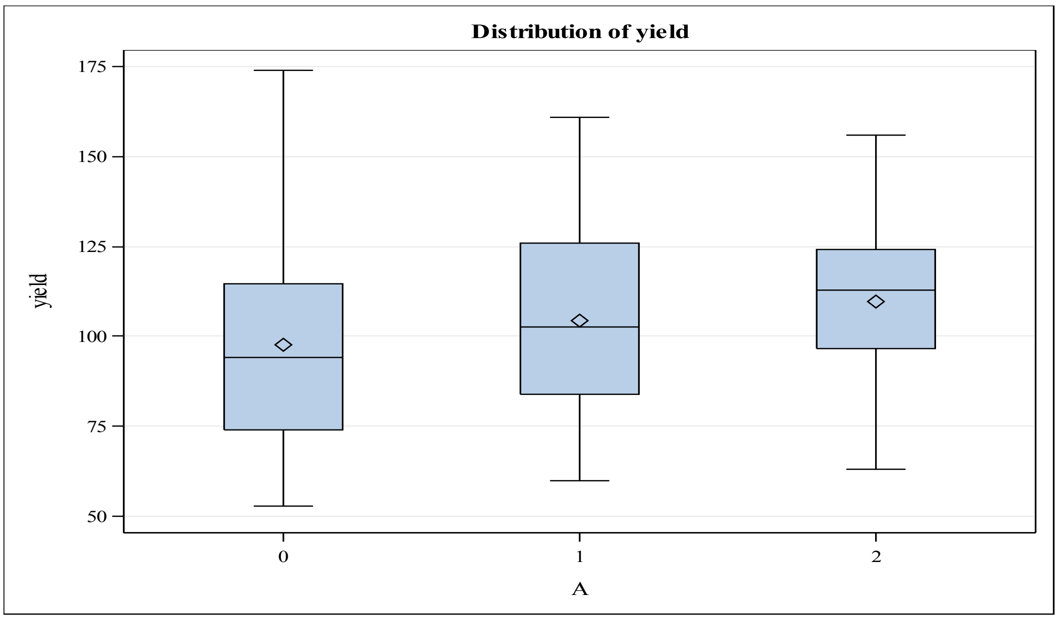
7

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	block	5	15875	3175.055556	5.28	0.0124
*	**A**	2	1786.361111	893.180556	1.49	0.2724
	Error: MS(A*block)	10	6013.305556	601.330556		

*** This test assumes one or more other fixed effects are zero.**

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	A*block	10	6013.305556	601.330556	3.40	0.0023
*	**B**	3	20021	6673.500000	37.69	<.0001
	A*B	6	321.750000	53.625000	0.30	0.9322
	Error: MS(Error)	45	7968.750000	177.083333		

*** This test assumes one or more other fixed effects are zero.**



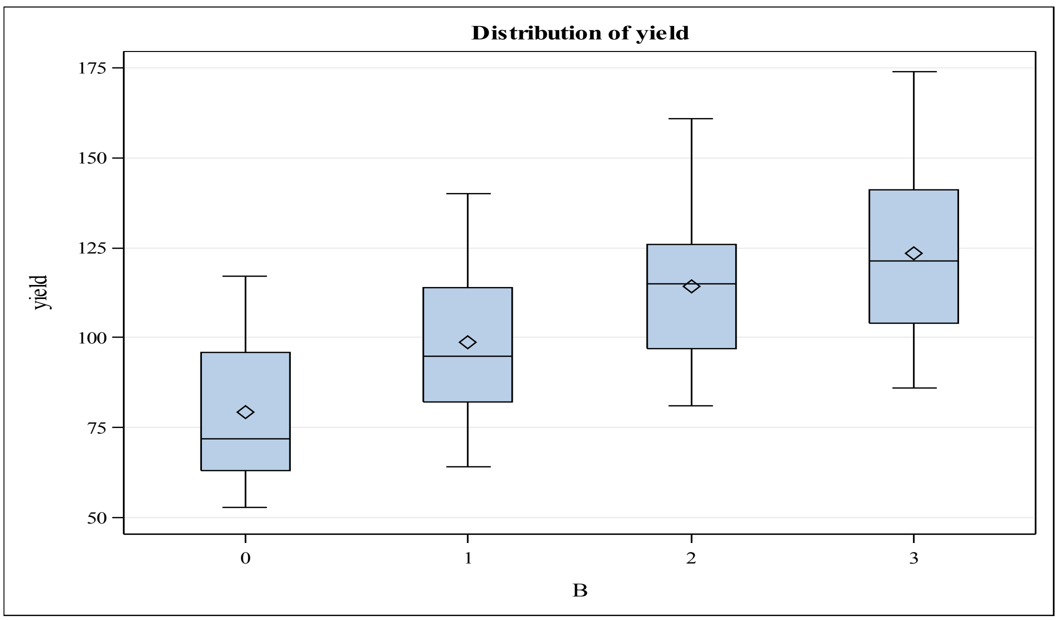
Dunnett's t Tests for yield

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

8

Alpha	0.05
Error Degrees of Freedom	45
Error Mean Square	177.0833
Critical Value of Dunnett's t	2.28350
Minimum Significant Difference	8.772

Comparisons significant at the 0.05 level are indicated by ***.				
A Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
2 - 0	12.167	3.395	20.939	***
1 - 0	6.875	-1.897	15.647	



Dunnett's t Tests for yield

9

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	45
Error Mean Square	177.0833
Critical Value of Dunnett's t	2.43088
Minimum Significant Difference	10.783

Comparisons significant at the 0.05 level are indicated by *.**				
B Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
3 - 0	44.000	33.217	54.783	***
2 - 0	34.833	24.051	45.616	***
1 - 0	19.500	8.717	30.283	***

10

The Mixed Procedure

Model Information	
Data Set	WORK.OATS
Dependent Variable	yield
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**
A	3	0 1 2
B	4	0 1 2 3
block	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	3
Columns in X	20
Columns in Z	24
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	564.36420957	
1	1	529.02850701	0.00000000

Convergence criteria met.

11

Covariance Parameter Estimates	
Cov Parm	**Estimate**
block	214.48
A*block	106.06
Residual	177.08

Fit Statistics	
-2 Res Log Likelihood	529.0
AIC (Smaller is Better)	535.0
AICC (Smaller is Better)	535.5
BIC (Smaller is Better)	534.4

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
A	2	10	1.49	0.2724
B	3	45	37.69	<.0001
A*B	6	45	0.30	0.9322

Variance Components Estimation Procedure

Class Level Information		
Class	**Levels**	**Values**
A	3	0 1 2
B	4	0 1 2 3
block	6	1 2 3 4 5 6

Number of Observations Read	72
Number of Observations Used	72

MIVQUE(0) SSQ Matrix				
Source	**block**	**A*block**	**Error**	**yield**
block	720.00000	240.00000	60.00000	190503.3
A*block	240.00000	240.00000	60.00000	87554.3
Error	60.00000	60.00000	60.00000	29857.3

MIVQUE(0) Estimates	
Variance Component	**yield**
Var(block)	214.47708
Var(A*block)	106.06181
Var(Error)	177.08333

The Mixed Procedure

Model Information	
Data Set	WORK.OATS
Dependent Variable	yield
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**
A	3	0 1 2
B	4	0 1 2 3
block	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	3
Columns in X	20
Columns in Z	24
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

13

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	564.36420957	
1	1	529.02850701	0.00000000

Convergence criteria met.

Covariance Parameter Estimates	
Cov Parm	**Estimate**
block	214.48
A*block	106.06
Residual	177.08

Fit Statistics	
-2 Res Log Likelihood	529.0
AIC (Smaller is Better)	535.0
AICC (Smaller is Better)	535.5
BIC (Smaller is Better)	534.4

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
A	2	10	1.49	0.2724
B	3	45	37.69	<.0001
A*B	6	45	0.30	0.9322

Least Squares Means							
Effect	**A**	**B**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
A	0		97.6250	7.7975	10	12.52	<.0001
A	1		104.50	7.7975	10	13.40	<.0001
A	2		109.79	7.7975	10	14.08	<.0001
B		0	79.3889	7.1747	45	11.07	<.0001
B		1	98.8889	7.1747	45	13.78	<.0001
B		2	114.22	7.1747	45	15.92	<.0001
B		3	123.39	7.1747	45	17.20	<.0001

Differences of Least Squares Means											
Effect	**A**	**B**	**_A**	**_B**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Adjustment**	**Adj P**
A	0		1		-6.8750	7.0789	10	-0.97	0.3544	Tukey-Kramer	0.6104
A	0		2		-12.1667	7.0789	10	-1.72	0.1164	Tukey-Kramer	0.2458
A	1		2		-5.2917	7.0789	10	-0.75	0.4720	Tukey-Kramer	0.7419
B		0		1	-19.5000	4.4358	45	-4.40	<.0001	Tukey-Kramer	0.0004
B		0		2	-34.8333	4.4358	45	-7.85	<.0001	Tukey-Kramer	<.0001
B		0		3	-44.0000	4.4358	45	-9.92	<.0001	Tukey-Kramer	<.0001
B		1		2	-15.3333	4.4358	45	-3.46	0.0012	Tukey-Kramer	0.0064
B		1		3	-24.5000	4.4358	45	-5.52	<.0001	Tukey-Kramer	<.0001
B		2		3	-9.1667	4.4358	45	-2.07	0.0446	Tukey-Kramer	0.1797

SAS output for Question 9

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	8.068673E11	1.344779E11	60.05	<.0001
Error	93	2.082822E11	2239593539		
Corrected Total	99	1.01515E12			

Root MSE	47324	**R-Square**	0.7948
Dependent Mean	155331	**Adj R-Sq**	0.7816
Coeff Var	30.46677		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	6868.82264	24689	0.28	0.7815
Size	1	66.42947	14.16148	4.69	<.0001
Beds	1	-10509	9178.44892	-1.14	0.2552
Baths	1	-2391.56172	11491	-0.21	0.8356
Taxes	1	37.50482	6.87171	5.46	<.0001
New	1	10796	41717	0.26	0.7964
IntNT	1	10.32976	12.74113	0.81	0.4196

The REG Procedure
Model: MODEL2
Dependent Variable: Price

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	7.117919E11	2.37264E11	75.08	<.0001
Error	96	3.033576E11	3159974854		
Corrected Total	99	1.01515E12			

Root MSE	56214	**R-Square**	0.7012
Dependent Mean	155331	**Adj R-Sq**	0.6918
Coeff Var	36.18959		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-27290	28241	-0.97	0.3363
Size	1	130.43397	11.95115	10.91	<.0001
Beds	1	-14466	10583	-1.37	0.1749
Baths	1	6890.26655	13540	0.51	0.6120

The REG Procedure
Model: MODEL3
Dependent Variable: Price

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	8.053952E11	1.61079E11	72.19	<.0001
Error	94	2.097543E11	2231428577		
Corrected Total	99	1.01515E12			

Root MSE	47238	**R-Square**	0.7934
Dependent Mean	155331	**Adj R-Sq**	0.7824
Coeff Var	30.41119		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4525.75265	24474	0.18	0.8537
Size	1	68.35009	13.93646	4.90	<.0001
Beds	1	-11259	9115.00315	-1.24	0.2198
Baths	1	-2114.37153	11465	-0.18	0.8541
Taxes	1	38.13524	6.81512	5.60	<.0001
New	1	41711	16887	2.47	0.0153

Applied Statistics Comprehensive Exam

January 2015

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:______________________________________ Date:____________

1.) An experiment on sugar beets compared times and methods of applying mixed artificail fertilizers (NPK). The mean yields of sugar (cwt per acre) were as follows: no artificial, $\bar{y}_{1.} = 38.7$, artificial applied in January (plowed), $\bar{y}_2 = 48.7$, artificial applied in January (broadcast), $\bar{y}_3 = 48.8$, artificial applied in April (broadcast), $\bar{y}_4 = 45.0$.

i. Write down a contrast that compares January and April applications. Call this ψ_1.

ii. Construct a contrast ψ_2 that compares the means for artifact and no artifact. show that this contrast is orthogonal to ψ_1.

iii. If the experiment has three replications ($n = 3$ per treatment), and $SSE = 62.51$, perform a test of the significance of ψ_1 at $\alpha = .05$ level and explain the result. We know: ${}_{0.05}t(14) = 1.761$, ${}_{0.05}t(15) = 1.753$, ${}_{0.05}t(16) = 1.745$.

2.) Food scientists wish to study how urban and rural consumers rate cheddar cheeses for bitterness. Four 50-pound blocks of cheddar cheese of different types are obtained. Each block of cheese represents one of the segments of the market (for example, a sharp New York style cheese). The raters are students from a large introductory food science class. Ten students from rural backgrounds and ten students from urban backgrounds are selected at random from the pool of possible raters. Each rater will taste eight bites of cheese presented in random order. The eight bites are two each from the four different cheeses, but the raters dont know that. Each rater rates each bite for bitterness.

i. Describe the experimental design you would use.

ii. Specify the type of factors / independent variables, i.e within-subjects, between-subjects, nested, split-plot, whole plot, etc.

iii. Write down the model and clearly identify each term in your model and assumptions of the model.

3.) An industrial engineer employed by a beverage bottler is interested in the effects of two different types of 32-ounce bottles on the time to deliver 12-bottle cases of the product. The two bottle types are glass and plastic. Two workers are used to perform a task consisting of moving 40 cases of the product 40 feet on a standard type of hand truck and stacking the cases in a display.

Four replicates of a 2^2 factorial design are performed, and the times observed are listed in the following table.

Suppose we denote the two factors as B for bottle type, and W for workers. Also the levels of the factors may be arbitrarily called "low" and "high". For instance, plastic may be called "low" and glass may be called "high". Similarly, worker 1 is considered "low" and worker 2 is considered "high". Consider the following data where a yield was recorded when the above mentioned factorial experiment was run in a completely randomized design with four replicates. *(CONTINUED ON NEXT PAGE)*

Factor			Replicate			
B-Bottle type	W- Worker	Treatment Combination	I	II	III	IV
$-$	$-$	B low, W low	4.95	4.43	4.27	4.25
$+$	$-$	B high, W low	5.12	4.89	4.98	5.00
$-$	$+$	B low, W high	5.28	4.91	4.75	4.71
$+$	$+$	B high, W high	6.65	6.24	5.49	5.55

You may want to construct a standard order table (also known as Yates' order) in order to answer the following questions.

i. Obtain the estimates of main effects of B, W, and the BW interaction.

ii. Obtain all sums of squares including the interaction effect and complete the ANOVA table. Show the ANOVA table.

iii.) Write the appropriate hypotheses with regard to analyzing the above data. Comment on the significance of the main effects and the interaction effect.

4.) Consider the following general linear model, which represents a regression line that changes at the point $X = \tilde{X}$,

$$Y_i = \beta_0 + \beta_1(X_i - \bar{X}_.) + \beta_2 I_i + \beta_3(I_i \times (X_i - \bar{X}_.)) + \epsilon_i,$$

where I_i is an indicator that takes the value 1 if $X_i > \tilde{X}$ and is 0 otherwise. Assume $\tilde{X} = 30$ and the X-values observed are $\mathbf{X} = [12\ \ 15\ \ 26\ \ 31\ \ 39\ \ 42\ \ 48]^T$ (so $\bar{X}_. = 30.43$).

i. Write the model in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, giving $\mathbf{X}$ and $\boldsymbol{\beta}$ explicitly.

ii. Give interpretations of all model parameters.

iii. A reduced model with equal slopes on either side of $\tilde{X}$ is proposed. Explicitly give the design matrix $\mathbf{X}_0$ and parameter vector $\boldsymbol{\beta}_0$ for this reduced model.

d. Explain how to test the significance of the difference in regression slopes on either side of $\tilde{X}$. Include detaiivls such as the test statistic, distribution, and degrees of freedom.

5.) [**This problem should be answered based on a 20-page SAS output beginning on page 7.**]

Peanuts are an important crop in parts of the southern United States. In an effort to develop improved plants, crop scientists routinely compare varieties with respect to several variables. The data for one two-factor experiment is available in the SAS output provided with this exam.

Three varieties (5, 6, and 8) were grown at two geographical locations (location 1 and 2) and, in this case, the three variables representing yield and the two important grade-grain characteristics were measured. The three variables are

Yield =	Plot weight
SdMtKer =	Sound mature kernels (weight in grams–max of 250 grams)
SeedSize =	Seed size (weight, in grams, of 100 seeds)

The experiment was replicated twice.

i. Perform a two-factor MANOVA. Test for a location effect, a variety effect, and a location-variety interaction effect. Use $\alpha = .05$. Comment on the findings for each of the factors.

ii. List the MANOVA assumptions. In particular, comment on why the assumption of homogeneity of covariance matrices is important.

iii. Analyze the residuals from Part i. Do the usual MANOVA assumptions appear to be satisfied? Discuss.

iv. Using the results in part i, can we conclude that the location and/or variety effects are additive? Additivity of effects means that a model without the interaction effect is good. If not, does the interaction effect show up for some (dependent) variables, but not for others? Check by using the three separate univariate two-factor ANOVA results taking each DV at a time. Comment on your findings.

v. In the data, larger numbers correspond to better yield and grade-grain characteristics. Can we conclude that one variety is better than the other two for each characteristic (the DVs)? Discuss your answer using 95% Bonferroni simultaneous intervals for pairs of varieties.

6.) Consider a one-fourth fraction of a 2^5 factorial design with factors A, B, C, D, E. Answer the following questions:

i. Suppose that the design generators for this design are $I = ACE, I = BCDE$. Write the complete defining relation for this design.

ii. Show the standard order table for this 2^{5-2} design with the design generators given above.

iii. What is the resolution of this design? Justify.

iv. For a 2^{5-2} design with design generators considered above, assuming all three-factor and higher-order interactions as negligible, write the alias structure of the main effects A, B, C, D, E.

v. Demonstrate how you would estimate the "pure" or de-aliased effect of A from such a design. You should show the procedure in detail including necessary "new" alias structures and a table showing how the de-aliasing of the effect of A can be obtained.

vi. Is it possible to have a better 2^{5-2} design for this situation? Briefly explain.

7.) Given the data, use the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ vs $h_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) The Colorado Commission of Higher Education has recently hired you to estimate the average amount of scholarship money (in dollars) that each student receives per semester. Only state-supported universities/colleges in the state of Colorado are to be used. Private schools are not to be included in the population for this study. What type of sampling design would you use? Why? Explain, in detail, how you would obtain the data for your sample. What are the advantages/disadvantages of your design? What costs might be involved in collecting your data?

9.) Explain what multicollinearity is. What are the sources and effects of multicollinearity?

10.) Higher education researchers are interested in predicting current high school students' chances of completing various levels of education, using parents' income (in thousands of dollars per year), gender (female, male), and race (African-American, Caucasian, Hispanic / Latino) as predictors. The researchers conducted a retrospective study in which 20 high school graduates over the age of 35 were randomly selected from each of the six combinations of gender and race (giving 120 total subjects). Each subject was asked to estimate his/her parents' annual income and to indicate whether he/she has completed high school, has completed an associate's degree, has completed a bachelor's degree, or has completed a graduate degree.

i. Propose an appropriate Generalized Linear Model (GLM) to predict the chances of reaching each level of education. Clearly explain the meaning of each component and each parameter included in your model.

ii. Compare your proposed model from part i with at least one other possible model using a *different link function*.

iii. Assuming all 120 subjects provide different values for "parents' income," calculate the "error" degrees of freedom for your model.

iv. Explain the meaning of the "proportional odds assumption." Does your model include such an assumption?

v. Suppose a single parameter for "parents' income" is estimated to be $\hat{\beta} \approx 0.35$. Provide an interpretation of this parameter estimate.

SAS output for question 5

Peanuts Data 1

Obs	location	variety	yield	SdMtKer	SeedSize
1	1	5	195.3	153.1	51.4
2	1	5	194.3	167.7	53.7
3	2	5	189.7	139.5	55.5
4	2	5	180.4	121.1	44.4
5	1	6	203.0	156.8	49.8
6	1	6	195.9	166.0	45.8
7	2	6	202.7	166.1	60.4
8	2	6	197.6	161.8	54.1
9	1	8	193.5	164.5	57.8
10	1	8	187.0	165.1	58.6
11	2	8	201.5	166.8	65.0
12	2	8	200.0	173.8	67.2

Peanuts Data

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
location	2	1 2
variety	3	5 6 8

Number of Observations Read	12
Number of Observations Used	12

The GLM Procedure

Dependent Variable: yield

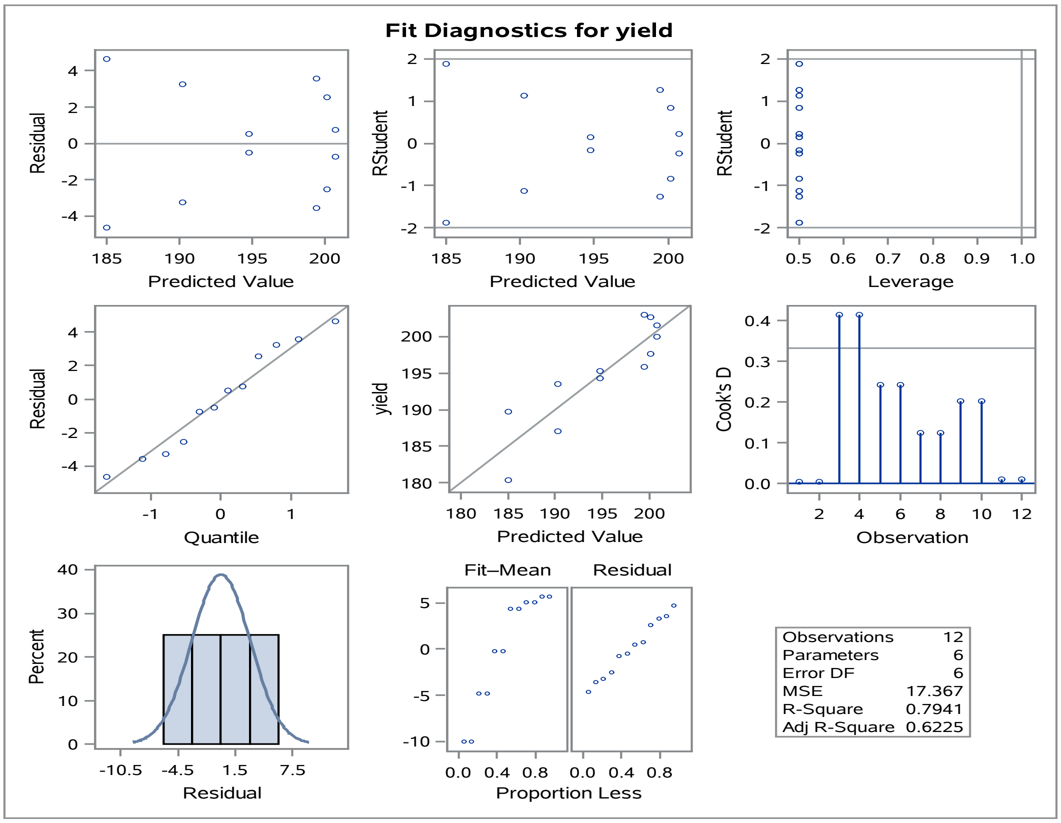
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	401.9175000	80.3835000	4.63	0.0446
Error	6	104.2050000	17.3675000		
Corrected Total	11	506.1225000			

R-Square	Coeff Var	Root MSE	yield Mean
0.794111	2.136324	4.167433	195.0750

Source	DF	Type III SS	Mean Square	F Value	Pr > F
location	1	0.7008333	0.7008333	0.04	0.8474
variety	2	196.1150000	98.0575000	5.65	0.0418
location*variety	2	205.1016667	102.5508333	5.90	0.0382

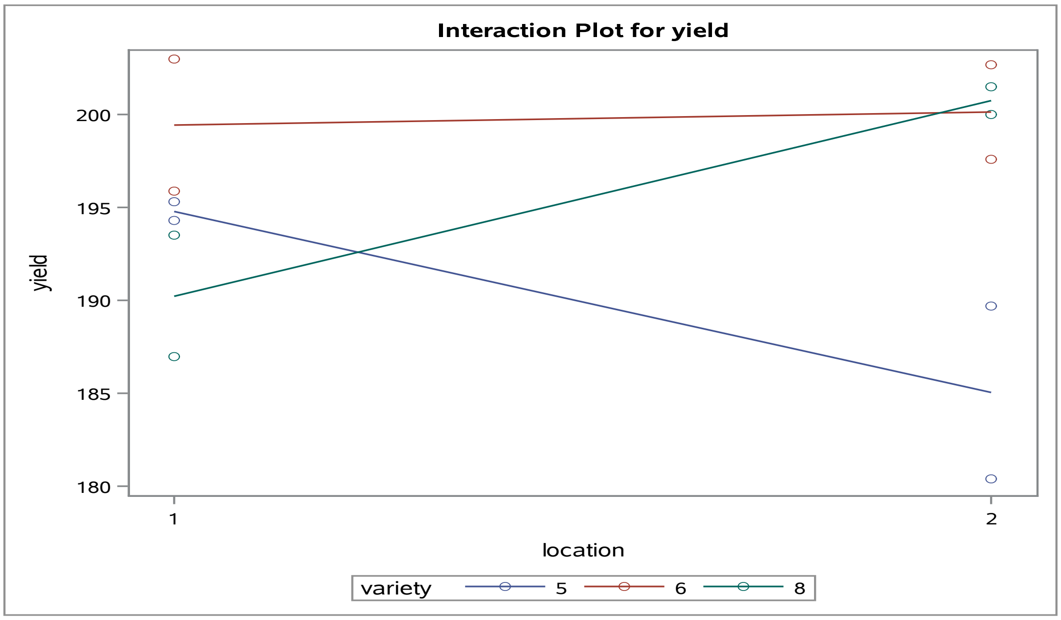
The GLM Procedure

Dependent Variable: yield



The GLM Procedure

Dependent Variable: yield



The GLM Procedure

Dependent Variable: SdMtKer

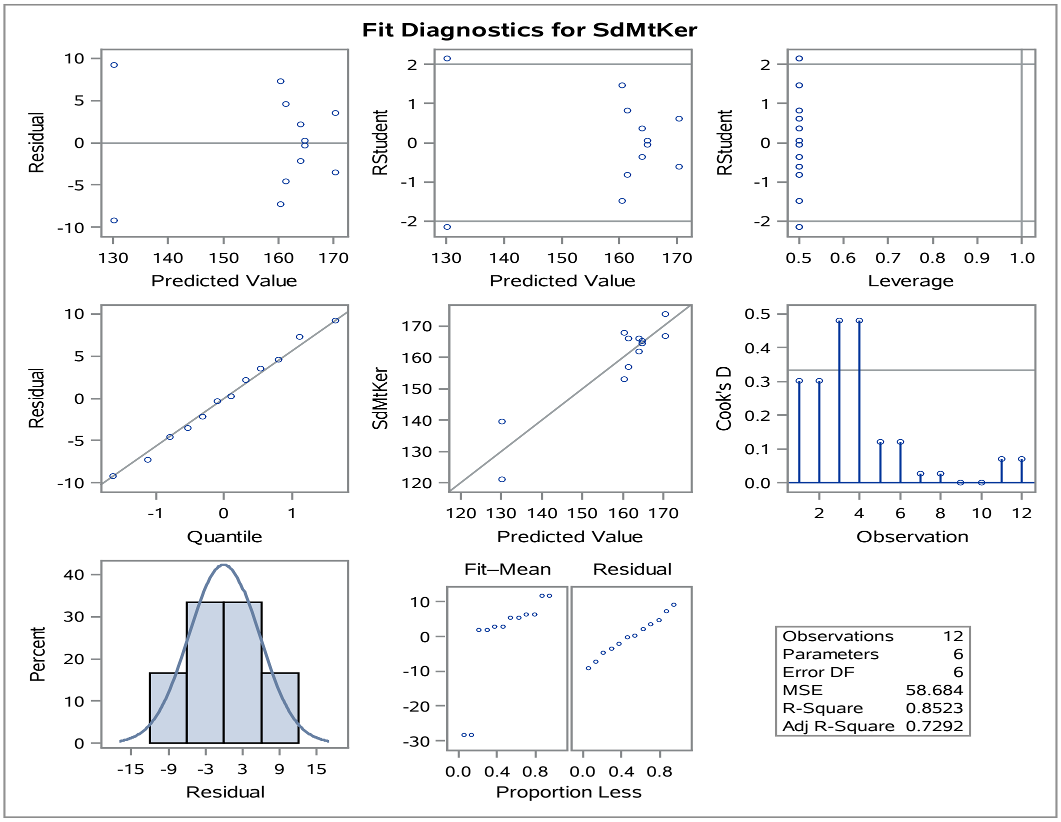
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2031.777500	406.355500	6.92	0.0177
Error	6	352.105000	58.684167		
Corrected Total	11	2383.882500			

R-Square	Coeff Var	Root MSE	SdMtKer Mean
0.852298	4.832398	7.660559	158.5250

Source	DF	Type III SS	Mean Square	F Value	Pr > F
location	1	162.067500	162.067500	2.76	0.1476
variety	2	1089.015000	544.507500	9.28	0.0146
location*variety	2	780.695000	390.347500	6.65	0.0300

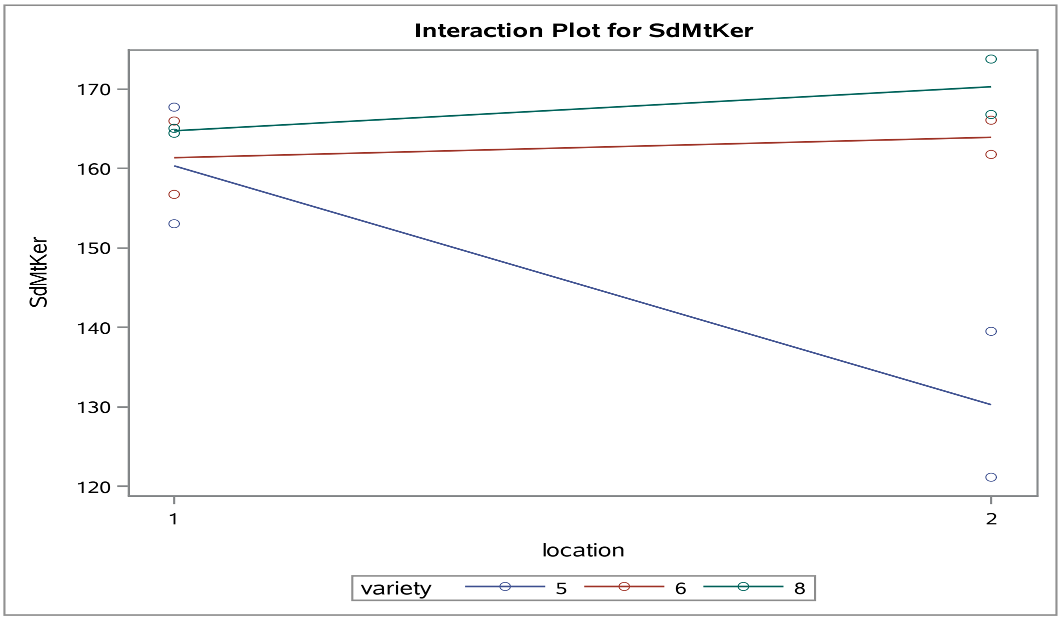
The GLM Procedure

Dependent Variable: SdMtKer



The GLM Procedure

Dependent Variable: SdMtKer



The GLM Procedure

Dependent Variable: SeedSize

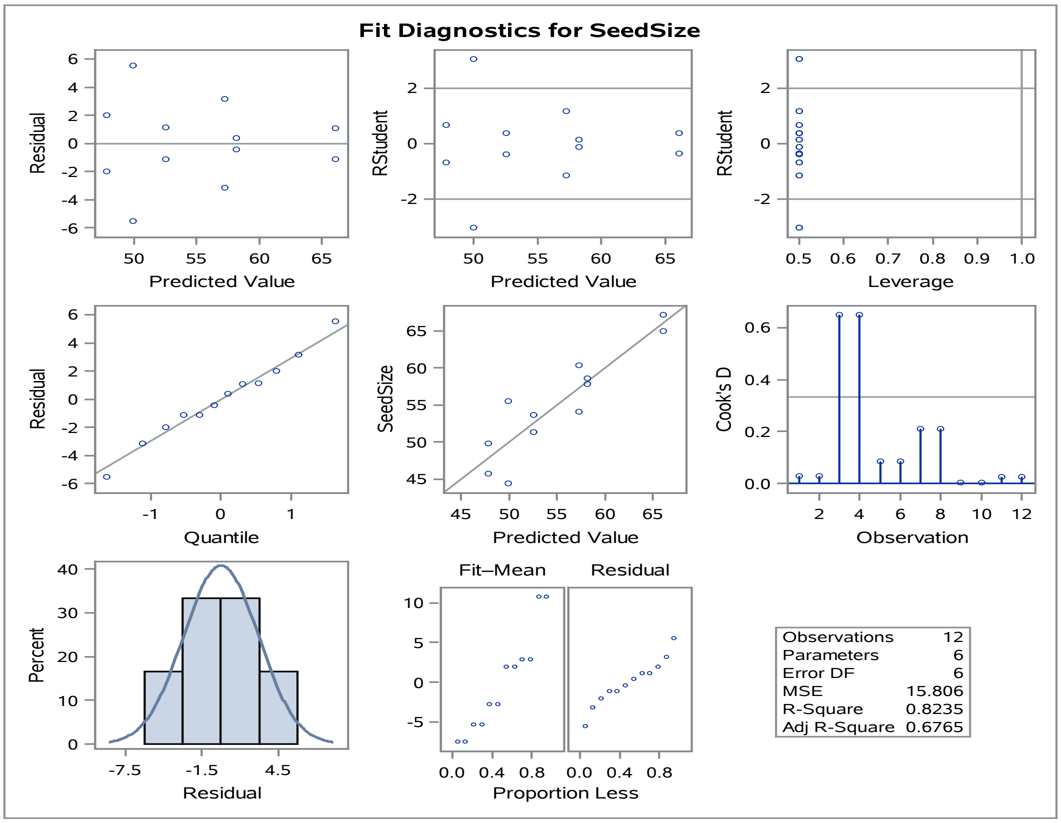
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	442.5741667	88.5148333	5.60	0.0292
Error	6	94.8350000	15.8058333		
Corrected Total	11	537.4091667			

R-Square	Coeff Var	Root MSE	SeedSize Mean
0.823533	7.188166	3.975655	55.30833

Source	DF	Type III SS	Mean Square	F Value	Pr > F
location	1	72.5208333	72.5208333	4.59	0.0759
variety	2	284.1016667	142.0508333	8.99	0.0157
location*variety	2	85.9516667	42.9758333	2.72	0.1443

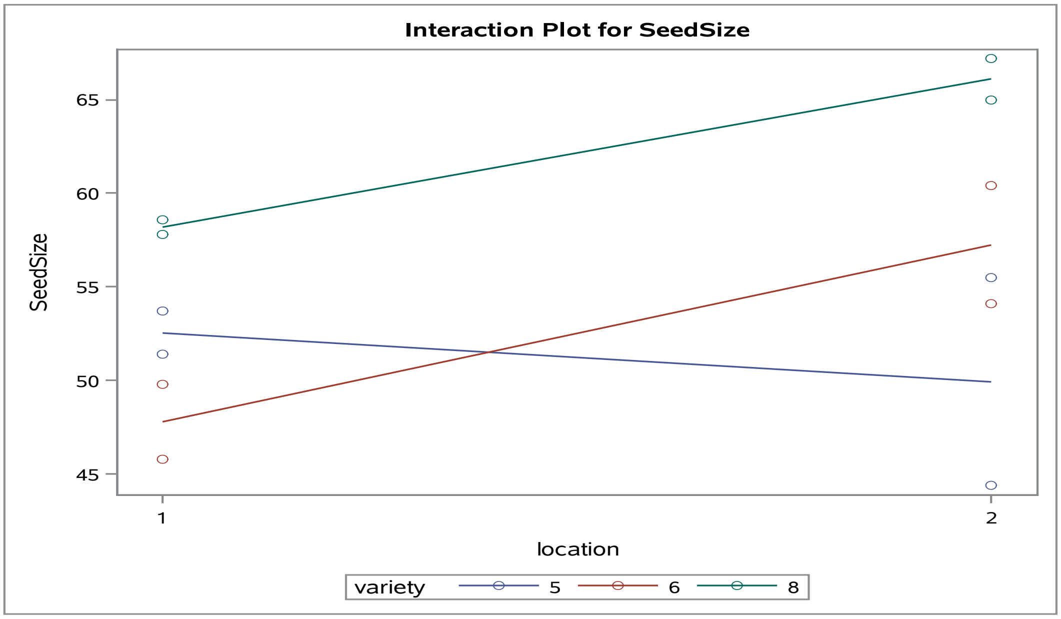
The GLM Procedure

Dependent Variable: SeedSize



The GLM Procedure

Dependent Variable: SeedSize



The GLM Procedure
Multivariate Analysis of Variance

E = Error SSCP Matrix			
	yield	**SdMtKer**	**SeedSize**
yield	104.205	49.365	76.48
SdMtKer	49.365	352.105	121.995
SeedSize	76.48	121.995	94.835

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|			
DF = 6	**yield**	**SdMtKer**	**SeedSize**
yield	1.000000	0.257714 0.5769	0.769342 0.0432
SdMtKer	0.257714 0.5769	1.000000	0.667608 0.1013
SeedSize	0.769342 0.0432	0.667608 0.1013	1.000000

The GLM Procedure
Multivariate Analysis of Variance

H = Type III SSCP Matrix for location			
	yield	SdMtKer	SeedSize
yield	0.7008333333	-10.6575	7.1291666667
SdMtKer	-10.6575	162.0675	-108.4125
SeedSize	7.1291666667	-108.4125	72.520833333

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for location E = Error SSCP Matrix				
		Characteristic Vector V'EV=1		
Characteristic Root	Percent	yield	SdMtKer	SeedSize
8.38824348	100.00	-0.13688388	-0.07628041	0.23952166
0.00000000	0.00	0.10187838	0.00216080	-0.00678495
0.00000000	0.00	-0.06307410	0.03725453	0.06189287

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall location Effect H = Type III SSCP Matrix for location E = Error SSCP Matrix S=1 M=0.5 N=1					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.10651620	11.18	3	4	0.0205
Pillai's Trace	0.89348380	11.18	3	4	0.0205
Hotelling-Lawley Trace	8.38824348	11.18	3	4	0.0205
Roy's Greatest Root	8.38824348	11.18	3	4	0.0205

H = Type III SSCP Matrix for variety			
	yield	SdMtKer	SeedSize
yield	196.115	365.1825	42.6275
SdMtKer	365.1825	1089.015	414.655
SeedSize	42.6275	414.655	284.10166667

The GLM Procedure
Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for variety E = Error SSCP Matrix				
		Characteristic Vector V'EV=1		
Characteristic Root	**Percent**	**yield**	**SdMtKer**	**SeedSize**
18.1876113	85.09	-0.16986539	-0.06425268	0.23943636
3.1880638	14.91	0.00137509	0.03769309	0.03800092
0.0000000	0.00	0.06510456	-0.04076880	0.04973481

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall variety Effect H = Type III SSCP Matrix for variety E = Error SSCP Matrix S=2 M=0 N=1					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.01244417	10.62	6	8	0.0019
Pillai's Trace	1.70910921	9.79	6	10	0.0011
Hotelling-Lawley Trace	21.37567504	14.25	6	4	0.0113
Roy's Greatest Root	18.18761127	30.31	3	5	0.0012
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

H = Type III SSCP Matrix for location*variety			
	yield	**SdMtKer**	**SeedSize**
yield	205.10166667	363.6675	107.78583333
SdMtKer	363.6675	780.695	254.22
SeedSize	107.78583333	254.22	85.951666667

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for location*variety E = Error SSCP Matrix				
		Characteristic Vector V'EV=1		
Characteristic Root	**Percent**	**yield**	**SdMtKer**	**SeedSize**
6.82409388	90.45	0.15723347	0.06948572	-0.18762316
0.72019649	9.55	-0.08644203	0.01400396	0.12612011
0.00000000	0.00	0.03000259	-0.04676424	0.10069089

Peanuts Data

The GLM Procedure
Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall location*variety Effect H = Type III SSCP Matrix for location*variety E = Error SSCP Matrix S=2 M=0 N=1					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.07429984	3.56	6	8	0.0508
Pillai's Trace	1.29086073	3.03	6	10	0.0587
Hotelling-Lawley Trace	7.54429038	5.03	6	4	0.0699
Roy's Greatest Root	6.82409388	11.37	3	5	0.0113
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

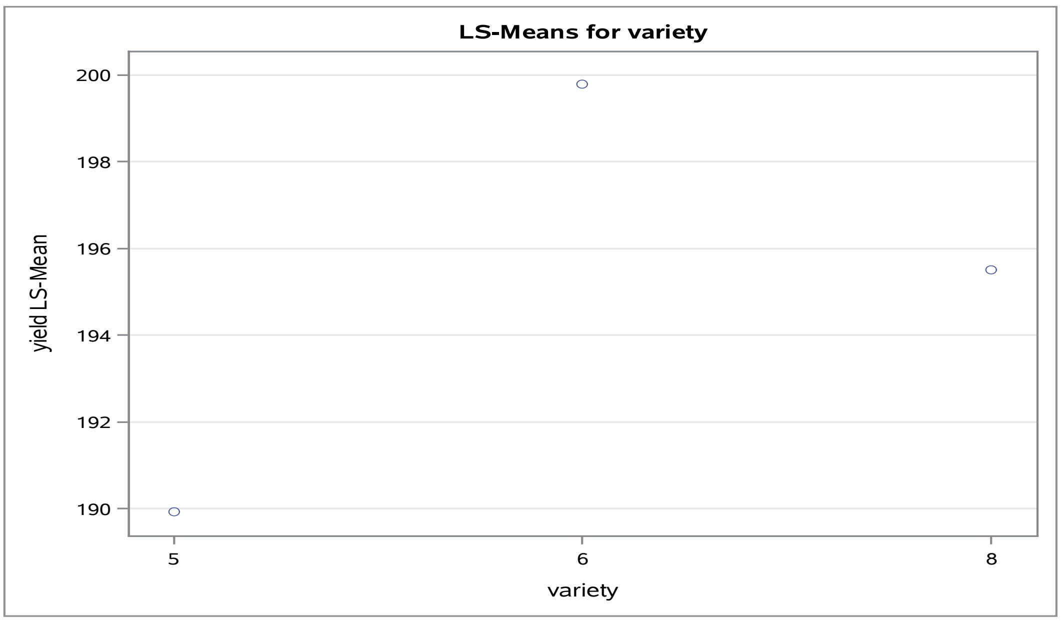
variety	yield LSMEAN	LSMEAN Number
5	189.925000	1
6	199.800000	2
8	195.500000	3

Least Squares Means for effect variety Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: yield			
i/j	**1**	**2**	**3**
1		0.0462	0.3221
2	0.0462		0.5844
3	0.3221	0.5844	

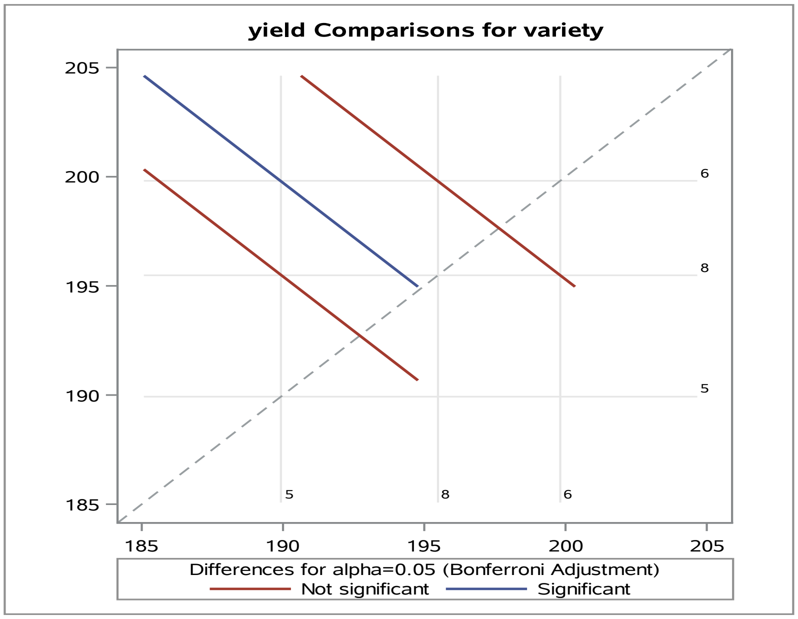
variety	yield LSMEAN	95% Confidence Limits	
5	189.925000	184.826329	195.023671
6	199.800000	194.701329	204.898671
8	195.500000	190.401329	200.598671

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

Least Squares Means for Effect variety				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-9.875000	-19.562540	-0.187460
1	3	-5.575000	-15.262540	4.112540
2	3	4.300000	-5.387540	13.987540



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni



variety	SdMtKer LSMEAN	LSMEAN Number
5	145.350000	1
6	162.675000	2
8	167.550000	3

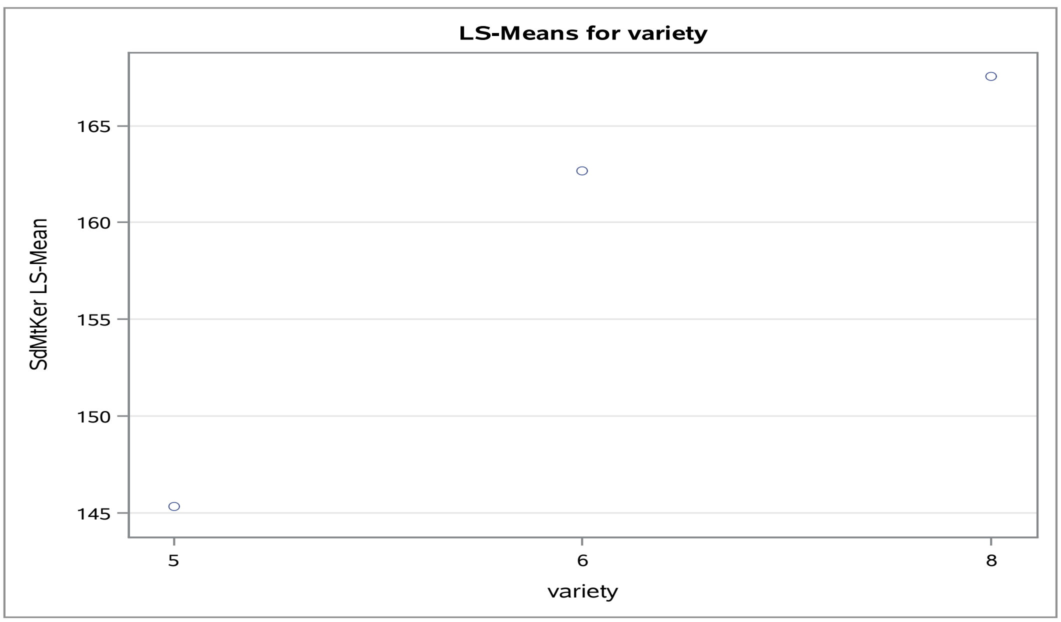
Least Squares Means for effect variety Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: SdMtKer			
i/j	**1**	**2**	**3**
1		0.0559	0.0191
2	0.0559		1.0000
3	0.0191	1.0000	

The GLM Procedure
Least Squares Means

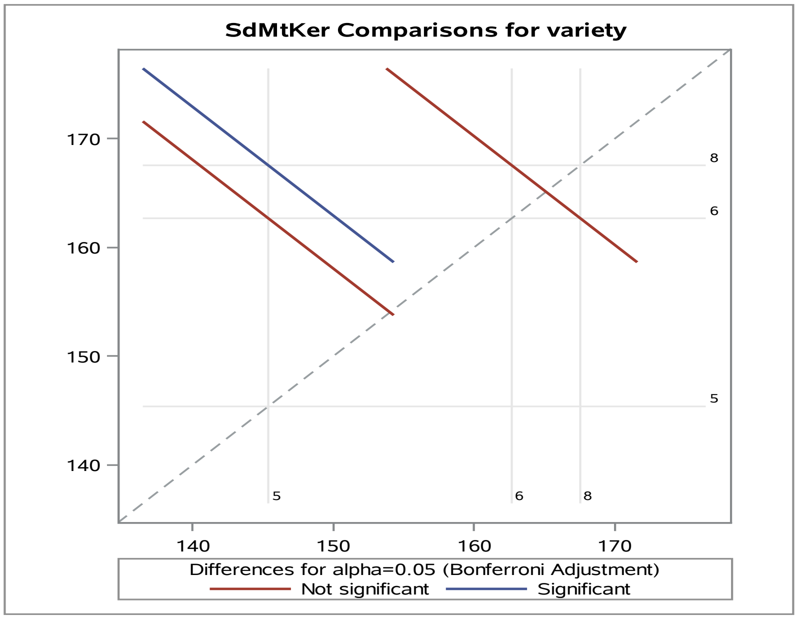
variety	SdMtKer LSMEAN	95% Confidence Limits	
5	145.350000	135.977644	154.722356
6	162.675000	153.302644	172.047356
8	167.550000	158.177644	176.922356

Least Squares Means for Effect variety

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-17.325000	-35.132597	0.482597
1	3	-22.200000	-40.007597	-4.392403
2	3	-4.875000	-22.682597	12.932597



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni



variety	SeedSize LSMEAN	LSMEAN Number
5	51.2500000	1
6	52.5250000	2
8	62.1500000	3

Least Squares Means for effect variety
Pr > \|t\| for H0: LSMean(i)=LSMean(j)

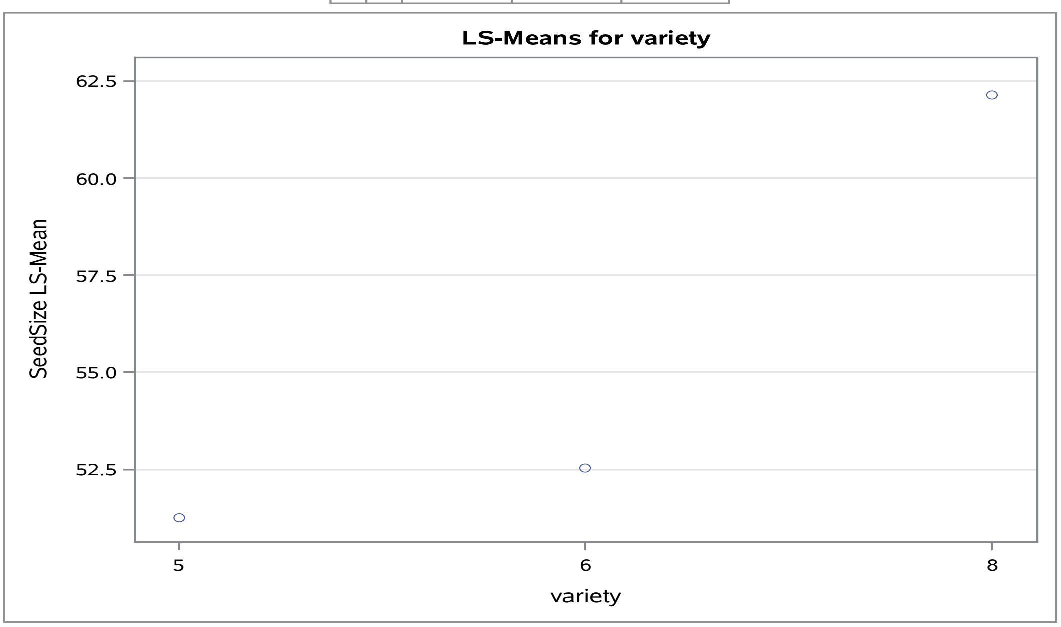
Dependent Variable: SeedSize

i/j	1	2	3
1		1.0000	0.0246
2	1.0000		0.0422
3	0.0246	0.0422	

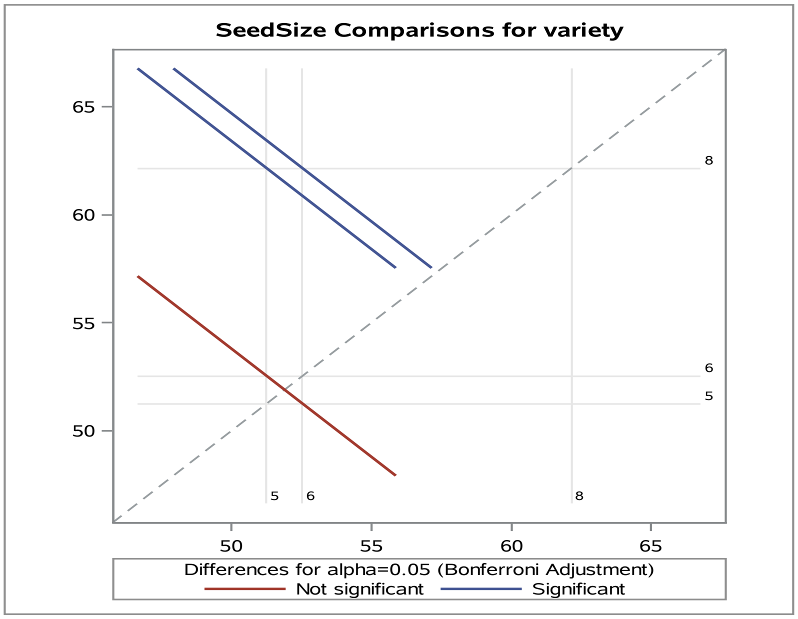
The GLM Procedure
Least Squares Means

variety	SeedSize LSMEAN	95% Confidence Limits	
5	51.250000	46.385961	56.114039
6	52.525000	47.660961	57.389039
8	62.150000	57.285961	67.014039

Least Squares Means for Effect variety				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-1.275000	-10.516736	7.966736
1	3	-10.900000	-20.141736	-1.658264
2	3	-9.625000	-18.866736	-0.383264



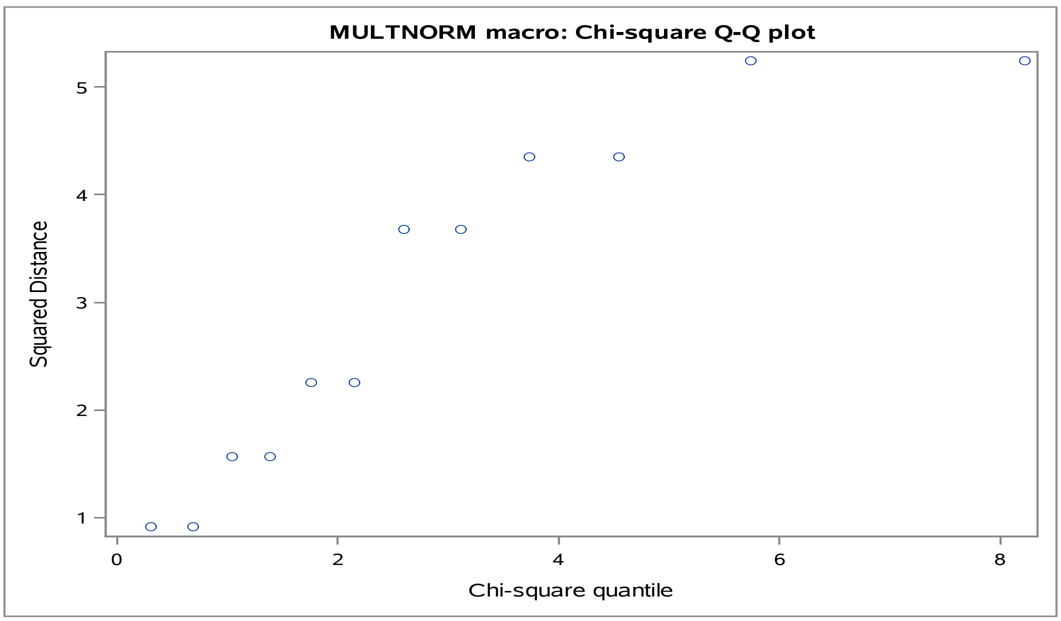
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni



Obs	location	variety	yield	SdMtKer	SeedSize	r1	r2	r3
1	1	5	195.3	153.1	51.4	0.50	-7.30	-1.15
2	1	5	194.3	167.7	53.7	-0.50	7.30	1.15
3	2	5	189.7	139.5	55.5	4.65	9.20	5.55
4	2	5	180.4	121.1	44.4	-4.65	-9.20	-5.55
5	1	6	203.0	156.8	49.8	3.55	-4.60	2.00
6	1	6	195.9	166.0	45.8	-3.55	4.60	-2.00
7	2	6	202.7	166.1	60.4	2.55	2.15	3.15
8	2	6	197.6	161.8	54.1	-2.55	-2.15	-3.15
9	1	8	193.5	164.5	57.8	3.25	-0.30	-0.40
10	1	8	187.0	165.1	58.6	-3.25	0.30	0.40
11	2	8	201.5	166.8	65.0	0.75	-3.50	-1.10
12	2	8	200.0	173.8	67.2	-0.75	3.50	1.10

The MODEL Procedure

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
r1	Shapiro-Wilk W	0.95	0.6552
r2	Shapiro-Wilk W	0.98	0.9949
r3	Shapiro-Wilk W	0.99	1.0000
System	Mardia Skewness	0.00	1.0000
	Mardia Kurtosis	-1.14	0.2527
	Henze-Zirkler T	0.37	0.8770



Discriminant Analysis Results

The DISCRIM Procedure

Total Sample Size	12	**DF Total**	11
Variables	3	**DF Within Classes**	10
Classes	2	**DF Between Classes**	1

Number of Observations Read	12
Number of Observations Used	12

Class Level Information					
location	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
1	_1	6	6.0000	0.500000	0.500000
2	_2	6	6.0000	0.500000	0.500000

Within Covariance Matrix Information		
location	**Covariance Matrix Rank**	**Natural Log of the Determinant of the Covariance Matrix**
1	3	9.06175
2	3	10.04024
Pooled	3	11.39958

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
12.477965	6	0.0521

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

The DISCRIM Procedure

Generalized Squared Distance to location		
From location	**1**	**2**
1	9.06175	17.74659
2	12.67195	10.04024

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.PEANUTS
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into location

From location	1	2	Total
1	6 100.00	0 0.00	6 100.00
2	1 16.67	5 83.33	6 100.00
Total	7 58.33	5 41.67	12 100.00
Priors	0.5	0.5	

Error Count Estimates for location

	1	2	Total
Rate	0.0000	0.1667	0.0833
Priors	0.5000	0.5000	

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.PEANUTS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into location

From location	1	2	Total
1	4 66.67	2 33.33	6 100.00
2	1 16.67	5 83.33	6 100.00
Total	5 41.67	7 58.33	12 100.00
Priors	0.5	0.5	

Error Count Estimates for location

	1	2	Total
Rate	0.3333	0.1667	0.2500
Priors	0.5000	0.5000	

Applied Statistics Comprehensive Exam

August 2014

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1. This Ph.D level exam will run from **8:30 AM to 3:30 PM**.
2. Please label each page with your identification number.

   DO NOT USE YOUR NAME OR BEAR NUMBER.
3. Please write only on one side of each page.
4. Please leave one inch margins on all sides of each page.
5. Please number all pages consecutively.
6. Please label the day number (Day 1 or Day 2) on each page.
7. Please begin each question on a new page, and number each question.
8. Please do not staple pages together.
9. No wireless devices, formula sheets, or other outside materials are permitted.
10. Statistical tables and paper will be provided.
11. Relax and good luck!

I have read and understand the rules of this exam.

Signature:______________________________ Date:__________

1.) Cancer rehabilitation researchers are interested in evaluating patients' post-treatment cardiopulmonary function using a continuous measure of "VO2 peak." They would like to compare this measure across four cancer stages (I, II, III, IV) while controlling for gender (female / male) and also patient age (measured in years). Researchers are not interested in testing hypotheses across gender and different ages, as it is accepted that cardiopulmonary function differs for males and females and at different ages.

i. Describe an appropriate model that could be used to assess differences in cardiopulmonary function across cancer stages while accounting for gender and age.

ii. State the assumptions of your model.

iii. Your model must include an assumption about the relationship between age and lung capacity. Describe how your model could be adjusted to change this assumption.

iv. Provide an interpretation of the intercept / constant term in your model. Is this meaningful for making conclusions?

v. Provide an interpretation of the term(s) associated with cancer stage in your model. Give a detailed expression that could be used to test for the significance of this term, including the steps in the testing process, any calculations or formulas involved, and the distribution of any test statistic(s) used.

2.) For each of the following three descriptions,

1. Determine the dependent and independent variables.

2. Determine which independent variable is nested within the other.

3. Sketch a representation of the data structure that makes the nesting clear. Do this in any way that makes sense to you!

i. Education researchers recorded the individual students' standardized reading scores for five classrooms selected from four different schools in a district. They are interested in the effects of schools and of classrooms on these scores.

ii. A hospital is investigating basic supply expenditures. Three nurses are selected from each of four different floors and surveyed about supply usage once a month for a year. Nurses work only on a single floor. It is of interest to know the effect of the individual and also of the floor on average supply usage.

iii. Public health researchers are interested in rating residents on a "health index" scored on a scale from 1 to 100. Three states participated in their study, with three cities selected from each state. Researchers recorded the health index value for five households in each city, and are interested in the effect of the state and the city on these values.

3.) Consider a 2^2 factorial design with two factors A and B. The levels of the factors may be arbitrarily called "low" and "high". Consider the following data where an yield was recorded when the above mentioned factorial experiment was run in a completely randomized design with three replicates.

Factor			Replicate			Total
A	B	Treatment Combination	I	II	III	
$-$	$-$	A low, B low	28	25	27	80
$+$	$-$	A high, B low	36	32	32	100
$-$	$+$	A low, B high	18	19	23	60
$+$	$+$	A high, B high	31	30	29	90

You may want to construct a standard order table (also known as Yates' order) in order to answer the following questions.

i. Obtain the estimates of main effects of A, B, and AB interaction.

ii. Obtain the sum of squares estimates SS_A, SS_B, and SS_{AB} for A, B, and AB, respectively.

iii. The total corrected sum of squares for this experiment, SS_T is 323. Calculate the sum of squares for the error, SS_E by subtracting SS_A, SS_B, and SS_{AB} from the SS_T.

iv. Construct the analysis of variance table including the calculated F-statistic. Comment on the significance of the main effects and the interaction.

4.) Consider a Two-Factor ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2$.

i. Write the model in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, giving $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ explicitly.

ii. Provide (but do not simplify!) an expression for the Least Squares Estimator $\hat{\boldsymbol{\beta}}$.

iii. Determine whether the linear expression $\beta_k - \beta_l$ is estimable, for any combination k and l. What is the importance of identifying "estimable" functions?

iv. Describe how to find the Best Linear Unbiased Estimator for

$$\begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \\ \beta_3 - \beta_1 \end{bmatrix}.$$

Explain in what sense the estimator is "best."

5.) [**This problem should be answered based on a 7-page SAS output on pages 7 through 13.**]

The admission officer of a graduate school has used an "index" of undergraduate GPA and graduate management aptitude test (GMAT) scores to help decide which applicants should be admitted to the graduate programs. The scatter plot of GPA vs GMAT (shown in the attached SAS output) shows recent applicants who have been classified as "Admit (A)", "Borderline (B)", and "Reject (R)".

A discriminant analysis and classification have been performed on the data and the results are shown in the attached SAS output. Answer the following questions. **Note:** when you answer, make sure to include the associated statistics. For example, if you decide to reject a null hypothesis, you should mention the value of the appropriate test statistic and the corresponding p-value.

i. Is there significant association between admission status (admitted, rejected, borderline) and the scores on GPA and GMAT?

ii. If there is significant association, we would like to perform a discriminant analysis. How many discriminant functions (DF) are possible for the given problem?

iii. Comment on the significance of the discriminant function(s).

iv. What is the overall effect size for the discriminant analysis? Comment on the effect size of each of the discriminant functions.

v. Write the classification functions corresponding to each discriminant function. Use the classification function(s) for classifying an applicant as "Admit" or "Reject" or "Borderline" who has GPA = 3.7 and GMAT score = 650

vi. In the SAS output, both resubstitution summary and crossvalidation summary for classification are provided. Comment on the error of misclassification based on these output.

vii. Is there a reason to believe that the classification function produces noticeably higher error rate than what we would have obtained by chance alone? Would you use the discriminant functions obtained from this analysis to classify an applicant to either admit, reject, or borderline? Justify.

6.) Consider a one-fourth fraction of a 2^5 factorial design with factors A, B, C, D, E. Answer the following questions:

i. Suppose that the design generators for this design are $I = ACE, I = BCDE$. Write the complete defining relation for this design.

ii. Show the standard order table for this 2^{5-2} design with the design generators given above.

iii. What is the resolution of this design? Justify.

iv. For a 2^{5-2} design with design generators considered above, assuming all three-factor and higher-order interactions as negligible, write the alias structure of the main effects A, B, C, D, E.

v. Demonstrate how would you estimate the "pure" or de-aliased effect of A from such a design. You should show the procedure in detail including necessary "new" alias structures and a table showing how the de-aliasing of the effect of A can be obtained.

vi. Is it possible to have a better 2^{5-2} design for this situation? Briefly explain.

7.) Given the data, use the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ versus $H_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) Compare and contrast adaptive cluster sampling to simple random sampling. What do these designs have in common? How are they different? Give examples/applications of each design.

9.) The Berkeley Guidance Study was a longitudinal monitoring of girls born in Berkeley, California between January 1928 and June 1929, and followed for at least 18 years. The variables are described as follows.

Variable	Description
HT18	Age 18 height (cm)
HT2	Age 2 height (cm)
LG9	Age 9 leg circumference (cm)
ST9	Age 9 strength (kg)
WT9	Age 9 weight (kg)
WT18	Age 18 weight (kg)

Use the SAS output on page 14 to answer the following questions.

i. Test for significance of regression for the relationship between $HT18$ and $HT2, LG9, ST9, WT9$ and $WT18$.

ii. Assess multicollinearity in the model, and describe the results.

iii. Use the model with 5 regressors to find the prediction for $\mathbf{x}_0$ with the following values

$HT2$	$LG9$	$ST9$	$WT9$	$WT18$
91.4	26.61	62	30.1	76.3

iv. Using the partial F test, determine the contribution of $WT9$ and $WT18$ to the model. Note $F_{.05}(2, 130) = 3.065$.

10.) Health researchers are interested in explaining the likelihood of myocardial infarction (MI, "heart attack") using professional attributes. For their study they randomly selected 75 individuals between 55 and 65 years of age and recorded each individual's annual income (in thousands of dollars), whether the individual has a college degree, and whether the individual has experienced at least one MI within the last 10 years.

i. Describe an appropriate Generalized Linear Model for this research situation. Clearly explain each term in your systematic component.

ii. Using the output on pages 15 to 16, assess the fit of this model.

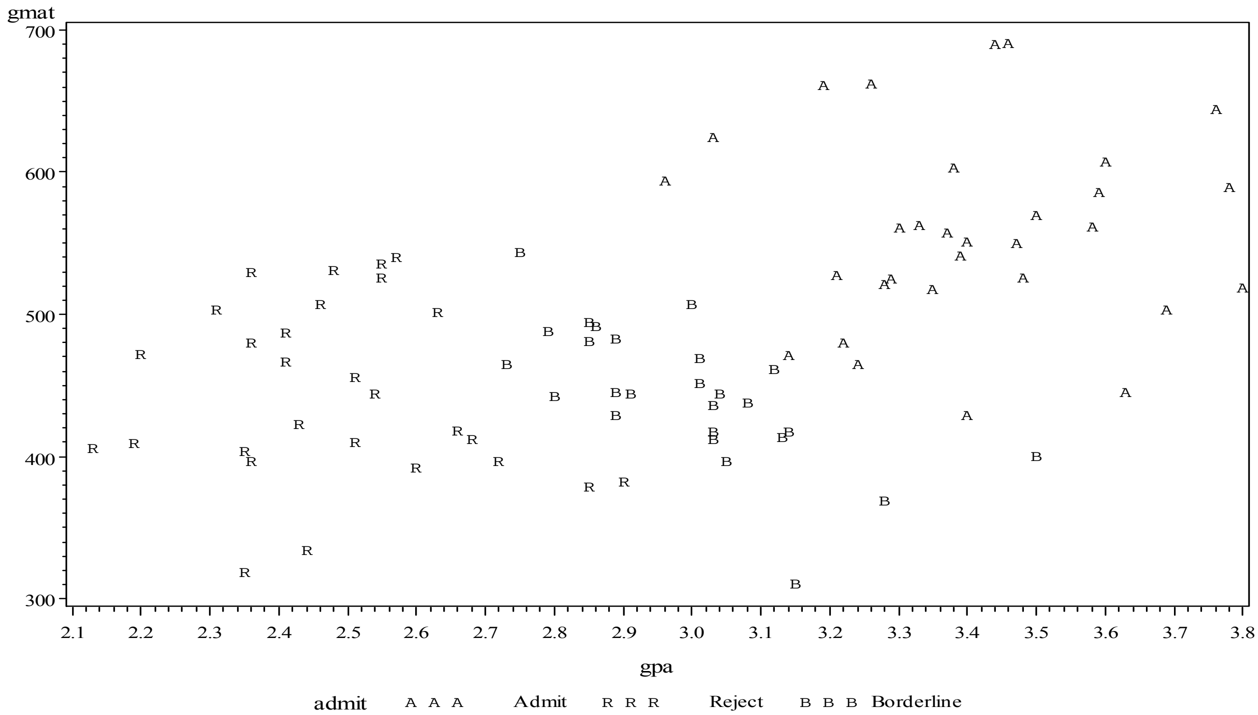
iii. Using the output on pages 15 to 16, provide an interpretation of the coefficient for "college degree." Also provide an interpretation of the coefficient for "annual income."

iv. The significance of each independent variable can be assessed using Wald Statistics. Briefly explain the process of a Wald hypothesis test.

v. Suppose the researchers also want to model the *variation* in MI (using a variance multiplier). Thinking of the properties of variance, describe an appropriate Generalized Linear Model for modeling variance (assume the same independent variables as part i).

SAS Output for Question 5

1



SAS Output for Question 5

The SAS System

The DISCRIM Procedure

Total Sample Size	85	DF Total	84
Variables	2	DF Within Classes	82
Classes	3	DF Between Classes	2

Number of Observations Read	85
Number of Observations Used	85

Class Level Information					
admit	Variable Name	Frequency	Weight	Proportion	Prior Probability
Admit	Admit	31	31.0000	0.364706	0.333333
Borderline	Borderline	26	26.0000	0.305882	0.333333
Reject	Reject	28	28.0000	0.329412	0.333333

Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
2	4.85035

SAS Output for Question 5

The SAS System

The DISCRIM Procedure
Canonical Discriminant Analysis

Generalized Squared Distance to admit			
From admit	**Admit**	**Borderline**	**Reject**
Admit	0	10.06344	31.28880
Borderline	10.06344	0	7.43364
Reject	31.28880	7.43364	0

Multivariate Statistics and F Approximations					
S=2 M=-0.5 N=39.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.12637661	73.43	4	162	<.0001
Pillai's Trace	1.00963002	41.80	4	164	<.0001
Hotelling-Lawley Trace	5.83665601	117.72	4	96.17	<.0001
Roy's Greatest Root	5.64604452	231.49	2	82	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	0.921702	0.920516	0.016417	0.849535	5.6460	5.4554	0.9673	0.9673
2	0.400119	.	0.091641	0.160095	0.1906		0.0327	1.0000

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	**Approximate F Value**	**Num DF**	**Den DF**	**Pr > F**
1	0.12637661	73.43	4	162	<.0001
2	0.83990454	15.63	1	82	0.0002

SAS Output for Question 5

The SAS System

The DISCRIM Procedure
Canonical Discriminant Analysis

Total Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.969922	-0.243416
gmat	0.662832	0.748768

Between Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.994118	-0.108305
gmat	0.897852	0.440298

Pooled Within Canonical Structure		
Variable	**Can1**	**Can2**
gpa	0.860161	-0.510023
gmat	0.350860	0.936428

SAS Output for Question 5

5

The SAS System

The DISCRIM Procedure

Total-Sample Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	2.148737595	-0.805087984
gmat	0.698531804	1.178084322

Pooled Within-Class Standardized Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	0.9512430832	-.3564113077
gmat	0.5180918168	0.8737695880

Raw Canonical Coefficients		
Variable	**Can1**	**Can2**
gpa	5.008766354	-1.876682204
gmat	0.008568593	0.014451060

Class Means on Canonical Variables		
admit	**Can1**	**Can2**
Admit	2.773788370	0.246102784
Borderline	-0.271055133	-0.644045724
Reject	-2.819285930	0.325571519

Linear Discriminant Function for admit			
Variable	**Admit**	**Borderline**	**Reject**
Constant	-240.37168	-177.31575	-133.89892
gpa	106.24991	92.66953	78.08637
gmat	0.21218	0.17323	0.16541

SAS Output for Question 5

6

The SAS System

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.GPA
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into admit				
From admit	**Admit**	**Borderline**	**Reject**	**Total**
Admit	27 87.10	4 12.90	0 0.00	31 100.00
Borderline	1 3.85	25 96.15	0 0.00	26 100.00
Reject	0 0.00	2 7.14	26 92.86	28 100.00
Total	28 32.94	31 36.47	26 30.59	85 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for admit				
	Admit	**Borderline**	**Reject**	**Total**
Rate	0.1290	0.0385	0.0714	0.0796
Priors	0.3333	0.3333	0.3333	

SAS Output for Question 5

The SAS System

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.GPA
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into admit				
From admit	**Admit**	**Borderline**	**Reject**	**Total**
Admit	26 83.87	5 16.13	0 0.00	31 100.00
Borderline	1 3.85	24 92.31	1 3.85	26 100.00
Reject	0 0.00	2 7.14	26 92.86	28 100.00
Total	27 31.76	31 36.47	27 31.76	85 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for admit				
	Admit	**Borderline**	**Reject**	**Total**
Rate	0.1613	0.0769	0.0714	0.1032
Priors	0.3333	0.3333	0.3333	

SAS Output for Question 9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6619.20745	1323.84149	43.67	<.0001
Error	130	3940.64072	30.31262		
Corrected Total	135	10560			

Root MSE	5.50569	R-Square	0.6268
Dependent Mean	172.57868	Adj R-Sq	0.6125
Coeff Var	3.19025		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	95.21271	16.09348	5.92	<.0001	0
HT2	HT2	1	0.92626	0.17012	5.44	<.0001	1.45551
LG9	LG9	1	-1.89219	0.49505	-3.82	0.0002	6.61156
ST9	ST9	1	0.15934	0.03663	4.35	<.0001	1.42637
WT9	WT9	1	0.21869	0.21919	1.00	0.3203	7.62382
WT18	WT18	1	0.48105	0.05527	8.70	<.0001	1.54992

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4054.68423	1351.56141	27.43	<.0001
Error	132	6505.16393	49.28154		
Corrected Total	135	10560			

Root MSE	7.02008	R-Square	0.3840
Dependent Mean	172.57868	Adj R-Sq	0.3700
Coeff Var	4.06776		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	67.42450	16.62263	4.06	<.0001
HT2	HT2	1	1.23598	0.20826	5.93	<.0001
LG9	LG9	1	-0.57232	0.28929	-1.98	0.0500
ST9	ST9	1	0.19329	0.04629	4.18	<.0001

SAS Output for Question 10

'

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MIDATA
Response Variable	MI
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Probability modeled is MI='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	101.106	86.535
SC	103.423	93.487
-2 Log L	99.106	80.535

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	18.5712	2	<.0001
Score	16.2263	2	0.0003
Wald	12.7642	2	0.0017

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-15.0145	5.0909	8.6985	0.0032
Income	1	0.1504	0.0504	8.9175	0.0028
CollegeDegree	1	1.8937	1.1298	2.8092	0.0937

SAS Output for Question 10

'

The LOGISTIC Procedure

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Income	1.162	1.053	1.283
CollegeDegree	6.644	0.726	60.830

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.3	**Somers' D**	0.568
Percent Discordant	21.6	**Gamma**	0.568
Percent Tied	0.1	**Tau-a**	0.269
Pairs	1316	**c**	0.784

Partition for the Hosmer and Lemeshow Test					
		MI = 1		MI = 0	
Group	Total	Observed	Expected	Observed	Expected
1	8	1	0.36	7	7.64
2	8	0	0.84	8	7.16
3	8	2	1.57	6	6.43
4	8	4	2.21	4	5.79
5	8	1	2.81	7	5.19
6	8	1	3.35	7	4.65
7	8	4	3.84	4	4.16
8	8	5	4.60	3	3.40
9	11	10	8.41	1	2.59

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.2674	7	0.1739

Applied Statistics Comprehensive Exam

January 2014

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) Briefly respond to the following.

i. Explain the difference(s) between "random" factors and "fixed" factors in an ANOVA model.

ii. Explain the difference of a between-subjects factor and a within-subjects factor.

iii. Explain the difference between crossed and nested factors.

2.) A pharmaceutical manufacturer wants to investigate the bioactivity of a new drug. He considers three dosage levels of 20g, 30g and 40g and randomly assigns 4 subjects to each dosage level. A response variable y was measured for each of 12 subjects.

i. Give an appropriate model for this research problem.

ii. Explain the meaning of the contrast $\psi_1 = \frac{\mu_1+\mu_3}{2} - \mu_2$, where $\mu_i,\ i = 1, 2, 3$ refers to the mean of the measurement for the i^{th} dosage level.

iii. Construct a contrast ψ_2 that compares the means for the 20g dosage level and 40g dosage level **and** show that this contrast is orthogonal to ψ_1.

iv. Given $\bar{y}_{1.} = 29.75$, $\bar{y}_{2.} = 36.75$, $\bar{y}_{3.} = 44.75$, and $SSE = 288.25$, perform a test of the significance of ψ_1 at $\alpha = .05$ level and explain the result. We know, ${}_{0.05}t(11) = 1.80$, ${}_{0.05}t(10) = 1.$, ${}_{0.05}t(9) = 1.83$

3.) Respond to the following.

i. Explain in non-technical terms the concept of analysis of variance (ANOVA), and write down the fundamental ANOVA identity. (This question is not about the use of ANOVA for testing several population means.)

ii. ANOVA is used to test for the equality of several treatments. In performing ANOVA, we compare mean squares for treatments (MS_{Treat}) and mean squares for errors (MS_{Err}). Technically, expected value for MS_{Treat} is the true population variance, σ^2.

Explain in your words how does the comparison of these two mean squares help us in testing for the treatment effects.

4.) Consider an Additive Two-Factor ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

where $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2, 3$.

i. Construct an appropriate design matrix $\mathbf{X}$ for this model.

ii. Suppose it is of interest to test the hypothesis $H_0 : \alpha_1 = \alpha_2$. Express this as a General Linear Hypothesis.

iii. Construct a *reduced model* associated with H_0 from part ii. Give the design matrix $\mathbf{X}_0$ for this reduced model.

iv. Using both design matrices, explicitly show how the hypothesis H_0 can be evaluated by comparing full and reduced models. Include an expression for the test statistic, the distribution of your test statistic, along with degrees of freedom.

5.) The salmon fishery is a valuable resource for both the United States and Canada. Because it is a limited resource, it must be managed efficiently. Moreover, since more than one country is involved, problems must be solved equitably. To help regulate catches, samples of fish taken during the harvest must be identified as coming from Alaskan or Canadian waters. The fish carry some information about their birth place in the growth rings on their scales. Typically, the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon.

The data set contains three variables measured on 100 specimens. The variables are country of origin (American or Canadian), freshwater growth ring size, and marine growth ring size.

We perform a discriminant analysis using country of origin as the population into which we wish to classify the fish, and the variables `freshwater` and `marine` as the discriminators. Selected SAS output is provided on page 6.

i. Comment about overall significance of the discriminators in predicting country of origin. Clearly mention what test statistic/statistics did you consider. How many discriminant functions would you obtain?

ii. Calculate overall effect size, effect size for the discriminant function and the effect sizes for each of the discriminator (predictor) variables. In each case, explain what do these effect sizes mean in the context of this problem.

iii. Considering a linear discriminant function analysis, if you have a new measurement on a salmon fish with `freshwater ring size = 90` and `marine ring size = 420`, would you classify it as American or Canadian? Justify your answer.

6.) The factors that influence longevity of AA-sized batteries are being studied. Three brands of batteries (Premium, Average, Low) each with two types (regular, alkaline) will be compared.

i. Design an experiment to answer the research question. Give a scenario and identify your experimental units. What is the outcome variable and how would you measure it in your context? In particular, comment about the allocation of treatments to the experimental units. Justify your selection of the design.

ii. Create a dummy data set that would be collected from such an experiment.

iii. Construct a partial ANOVA and show the sources of variation, appropriate degrees of freedoms, sum of squares, mean squares, and the F-statistic. Do not use the dummy data in ii to prepare the ANOVA table. How would you make conclusions from the experiment?

7.) Given the data, use the Wilcoxon Signed Rank Test to test:

$$H_0 : \tilde{\mu} = 8.41 \qquad vs \qquad H_1 : \tilde{\mu} > 8.41.$$

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) Compare and contrast simple random sampling to adaptive cluster sampling. What do these designs have in common? How are they different? Give examples/applications of each design.

9.) The Berkeley Guidance Study was a longitudinal monitoring of girls born in Berkeley, California between January 1928 and June 1929, and followed for at least 18 years. The variables are described as follows.

Variable	Description
HT18	Age 18 height (cm)
HT2	Age 2 height (cm)
LG9	Age 9 leg circumference (cm)
ST9	Age 9 strength (kg)
WT9	Age 9 weight (kg)
WT18	Age 18 weight (kg)

Use the SAS output on page 7 to answer the following questions.

i. Test for significance of regression for the relationship between $HT18$ and $HT2, LG9, ST9, WT9$ and $WT18$.

ii. Assess multicollinearity in the model, and describe the results.

iii. Use the model with 5 regressors to find the prediction for $\mathbf{x}_0$ with the following values

$HT2$	$LG9$	$ST9$	$WT9$	$WT18$
91.4	26.61	62	30.1	76.3

iv. Using the partial F test, determine the contribution of $WT9$ and $WT18$ to the model. Note $F_{.05}(2, 130) = 3.065$.

10.) The National Medical Expenditure Survey (NMES) includes records that allow researchers to model "number of annual physician office visits" using "age" (in years), "gender" (an indicator for females) and "married" (indicator).

i. Using the output on pages 8-9, clearly describe an appropriate model for this research situation.

ii. Using the output on pages 8-9, evaluate the fit of the selected model.

iii. Provide interpretations of the parameter estimates for "age" and "married" in terms of the original problem.

iv. Suppose researchers have observed an excess of zeros in this type of data in the past (significantly more than expected). Describe in detail how your model could be adjusted to account for this expectation.

SAS output for question 5

The SAS System

The DISCRIM Procedure

Total Sample Size	100	**DF Total**	99
Variables	2	**DF Within Classes**	98
Classes	2	**DF Between Classes**	1

Number of Observations Read	100
Number of Observations Used	100

Class Level Information					
region	Variable Name	Frequency	Weight	Proportion	Prior Probability
A	A	50	50.0000	0.500000	0.500000
C	C	50	50.0000	0.500000	0.500000

Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
2	12.72333

SAS output for question 9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6619.20745	1323.84149	43.67	<.0001
Error	130	3940.64072	30.31262		
Corrected Total	135	10560			

Root MSE	5.50569	R-Square	0.6268
Dependent Mean	172.57868	Adj R-Sq	0.6125
Coeff Var	3.19025		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	95.21271	16.09348	5.92	<.0001	0
HT2	HT2	1	0.92626	0.17012	5.44	<.0001	1.45551
LG9	LG9	1	-1.89219	0.49505	-3.82	0.0002	6.61156
ST9	ST9	1	0.15934	0.03663	4.35	<.0001	1.42637
WT9	WT9	1	0.21869	0.21919	1.00	0.3203	7.62382
WT18	WT18	1	0.48105	0.05527	8.70	<.0001	1.54992

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4054.68423	1351.56141	27.43	<.0001
Error	132	6505.16393	49.28154		
Corrected Total	135	10560			

Root MSE	7.02008	R-Square	0.3840
Dependent Mean	172.57868	Adj R-Sq	0.3700
Coeff Var	4.06776		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	67.42450	16.62263	4.06	<.0001
HT2	HT2	1	1.23598	0.20826	5.93	<.0001
LG9	LG9	1	-0.57232	0.28929	-1.98	0.0500
ST9	ST9	1	0.19329	0.04629	4.18	<.0001

SAS output for question 10

'

The UNIVARIATE Procedure
Variable:
Visits

Moments			
N	1200	**Sum Weights**	1200
Mean	21.125	**Sum Observations**	25350
Std Deviation	28.3835762	**Variance**	805.627398
Skewness	2.98654565	**Kurtosis**	14.2713188
Uncorrected SS	1501466	**Corrected SS**	965947.25
Coeff Variation	134.360124	**Std Error Mean**	0.81936327

Basic Statistical Measures			
Location		**Variability**	
Mean	21.12500	**Std Deviation**	28.38358
Median	10.00000	**Variance**	805.62740
Mode	1.00000	**Range**	269.00000
		Interquartile Range	26.00000

SAS output for question 10

The GENMOD Procedure

Model Information	
Data Set	WORK.VISITSDATA
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	Visits

Number of Observations Read	1200
Number of Observations Used	1200

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	1196	1276.6287	1.0674
Scaled Deviance	1196	1276.6287	1.0674
Pearson Chi-Square	1196	1246.1493	1.0419
Scaled Pearson X2	1196	1246.1493	1.0419
Log Likelihood		67083.4112	
Full Log Likelihood		-3911.5938	
AIC (smaller is better)		7833.1876	
AICC (smaller is better)		7833.2378	
BIC (smaller is better)		7858.6380	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.2157	0.1556	-1.5208	-0.9106	61.01	<.0001
Age	1	0.0676	0.0026	0.0625	0.0727	672.64	<.0001
Female	1	-0.0204	0.0709	-0.1594	0.1186	0.08	0.7734
Married	1	0.0254	0.0650	-0.1021	0.1528	0.15	0.6967
Dispersion	1	0.2657	0.0147	0.2384	0.2962		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Applied Statistics Comprehensive Exam

August 2013

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____________________________________ Date:____________

1.) Cancer rehabilitation researchers are interested in evaluating patients' post-treatment cardiopulmonary function using a continuous measure of "VO2 peak." They would like to compare this measure across four cancer stages (I, II, III, IV) while controlling for gender (female / male) and also patient age (measured in years). Researchers are not interested in testing hypotheses across gender and different ages, as it is accepted that cardiopulmonary function differs for males and females and at different ages.

i. Describe an appropriate model that could be used to assess differences in cardiopulmonary function across cancer stages while accounting for gender and age.

ii. State the assumptions of your model.

iii. Your model must include an assumption about the relationship between age and lung capacity. Describe how your model could be adjusted to change this assumption.

iv. Provide an interpretation of the intercept / constant term in your model. Is this meaningful for making conclusions?

v. Provide an interpretation of the term(s) associated with cancer stage in your model. Give a detailed expression that could be used to test for the significance of this term, including the steps in the testing process, any calculations or formulas involved, and the distribution of any test statistic(s) used.

2.) Higher education researchers are interested in trends of GPA for first-generation college students during their first four semesters in college, and the corresponding effects of motivation and substance abuse. For their study they randomly selected 50 first-generation college students and initially classified their "motivation" level into one of three groups (low, medium, high). At the end of each of the first four semesters of school for each student, semester GPA is recorded as well as a self-reported continuous measure of "substance abuse."

i. As a factor in a longitudinal panel study, how would you classify "motivation"?

ii. Clearly describe a model for GPA, accounting for all of the factors described.

iii. State the assumptions of your model. Specifically, what have you assumed about the effect of "time"?

iv. Describe a process that could be used to test the effect of "motivation" on GPA. Include all steps in the testing process, any calculations or formulas involved, degrees of freedom and the distribution of any test statistic(s) used.

v. Assuming researchers are interested in assessing a "time trend" across the four semesters, describe in detail the types of trends that could be considered as well as how these trends could be assessed using your model.

3.) The following sample statistics were computed for a study of mercury contents in the wing muscles of Australian waterfowl. Calculate the 90% confidence interval for the contrast below assuming equal population variances.

$$C = \mu_3 - \frac{1}{2}(\mu_1 + \mu_2)$$

Species	N	Mean	SD
Shelduck	6	9	4
Shoveler	3	10	5
Blue-Billed	18	15	5

4.) Consider a Blocked One-Factor ANOVA model,

$$Y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij},$$

where $i = 1, \ldots, 4$ indicates the four groups of interest, $j = 1, \ldots, 3$ indicates the three blocks, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, independent.

i. Present the response vector $\mathbf{Y}$ and a *full-rank* design matrix $\mathbf{X}$.

ii. Is $\mu + \alpha_1 + b_2$ estimable? Justify your answer.

iii. Find an expression for the BLUE of $\mu + \alpha_1 + b_2$, and explain in what sense it is "best."

iv. Find an expression for the variance of the BLUE of $\mu + \alpha_1 + b_2$, and explain how this variance compares to the variance of other estimators of $\mu + \alpha_1 + b_2$.

5.) An experiment was conducted in which 30 patients at an inpatient alcohol rehabilitation center were randomly assigned to receive one of three therapies. After three months of treatment, two outcomes were measured using self-report questionnaires.

i. Write the one-way MANOVA model for this study in matrix form. Indicate the dimensions of each matrix in your model.

ii. Describe the steps you would take to check the assumptions of this statistical analysis.

6.) Suppose we wish to study the effects of three factors on corn yields: amount of nitrogen added, planting depth, and planting date. The nitrogen and depth factors each have two levels, and the date factor has three levels. There are 24 plots available for this experiment: twelve are on a farm near Greeley, CO, and twelve are on a different farm near Brighton, CO.

i. Describe the experimental design you would use. Specifically describe the process for assigning treatments to EUs (plots). Briefly explain why you selected that design.

ii. Construct a partial ANOVA table that includes sources of variation, degrees of freedom, expected mean squares, and appropriate F-ratios.

7.) Given the data, use the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ versus $H_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) Compare and contrast stratified sampling to simple random sampling. What do these designs have in common? How are they different? Give examples/applications of each design. Under what conditions is stratified sampling preferred over simple random sampling.

9.) The observations for delivery time, number of cases, and distance walked by the router drive were collected in four cities. A model was developed that relates delivery time y to cases x_1, distance x_2, and the city in which the delivery was made. Based on SAS output on pages 5-7, answer the following questions.

i. Is there an indication that delivery site is an important variable?

ii. What conclusions can you draw regarding model adequacy?

10.) Based on a random sample of 3 Colorado high school classrooms, researchers have recorded a measure of proficiency (proficient / not proficient) for a total of 93 students (31 in the first classroom, 27 in the second, and 35 in the third). As proficiency is determined by standardized exams, researchers would like to know if high school GPA is a reasonable predictor of proficiency. They would also like to control for gender (male = 1 / female = 0) and block by class.

i. Based on researcher interest, construct an appropriate model for "proficiency."

ii. Using the output on pages 8-10, assess the fit of this model.

iii. Using the output on pages 8-10, provide an interpretation for the coefficient for GPA, for gender, and also for the coefficient for the "class 2" indicator.

iv. Describe how your model would change if classes were treated as *random* blocks.

The REG Procedure
Model: MODEL1
Dependent Variable: y y

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	5615.09147	1123.01829	125.92	<.0001
Error	19	169.45113	8.91848		
Corrected Total	24	5784.54260			

Root MSE	2.98638	**R-Square**	0.9707
Dependent Mean	22.38400	**Adj R-Sq**	0.9630
Coeff Var	13.34159		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	0.41625	2.25783	0.18	0.8557
x1	x1	1	1.77028	0.18679	9.48	<.0001
x2	x2	1	0.01083	0.00379	2.86	0.0100
x3		1	2.28510	2.41624	0.95	0.3562
x4		1	3.73764	2.35702	1.59	0.1293
x5		1	-0.45264	2.68742	-0.17	0.8680

The SAS System

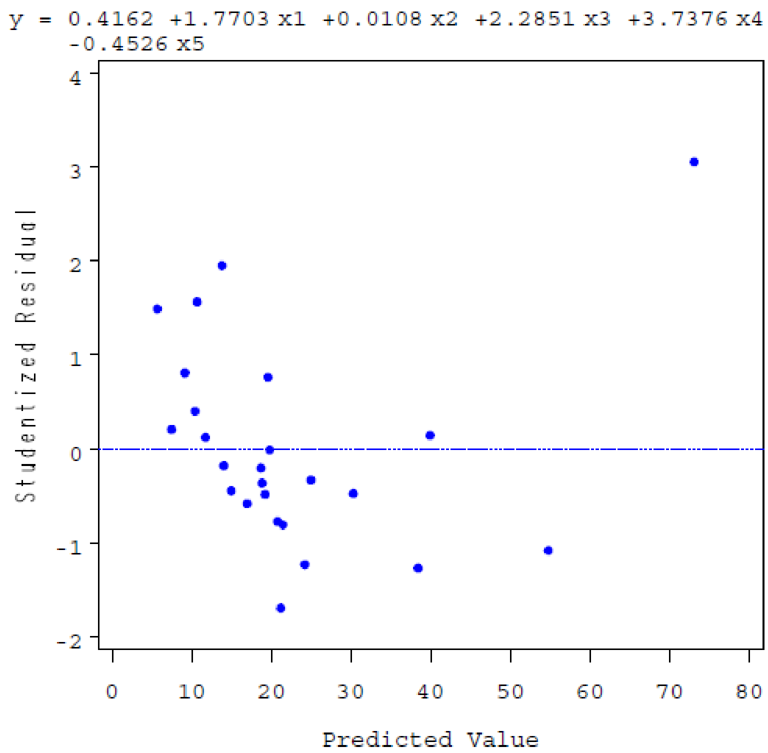
The REG Procedure
Model: MODEL2
Dependent Variable: y y

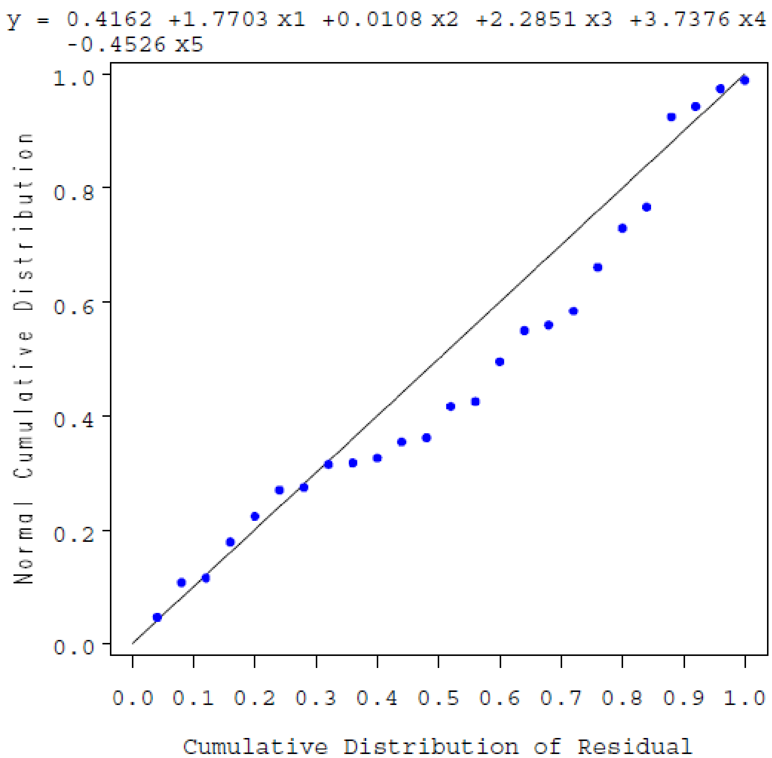
Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	5550.81092	2775.40546	261.24	<.0001
Error	22	233.73168	10.62417		
Corrected Total	24	5784.54260			

Root MSE	3.25947	**R-Square**	0.9596
Dependent Mean	22.38400	**Adj R-Sq**	0.9559
Coeff Var	14.56162		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	2.34123	1.09673	2.13	0.0442
x1	x1	1	1.61591	0.17073	9.46	<.0001
x2	x2	1	0.01438	0.00361	3.98	0.0006


y = 0.4162 +1.7703 x1 +0.0108 x2 +2.2851 x3 +3.7376 x4
-0.4526 x5
Studentized Residual
Predicted Value
4
3
2
1
0
-1
-2
0
10
20
30
40
50
60
70
80


y = 0.4162 +1.7703 x1 +0.0108 x2 +2.2851 x3 +3.7376 x4
-0.4526 x5
Normal Cumulative Distribution
Cumulative Distribution of Residual
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

```
                    The LOGISTIC Procedure

                      Model Information

        Data Set                      WORK.PROFICIENCY
        Response Variable             Proficiency
        Number of Response Levels     2
        Model                         binary logit
        Optimization Technique        Fisher's scoring


           Number of Observations Read          93
           Number of Observations Used          93


                       Response Profile

             Ordered                          Total
               Value     Proficiency      Frequency

                   1               1             59
                   2               0             34

           Probability modeled is Proficiency=1.


                   Model Convergence Status

        Convergence criterion (GCONV=1E-8) satisfied.
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	124.122	85.754
SC	126.654	98.417
-2 Log L	122.122	75.754

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	46.3675	4	<.0001
Score	40.2811	4	<.0001
Wald	26.2467	4	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.4772	1.8217	21.6537	<.0001
GPA	1	3.0771	0.6015	26.1689	<.0001
Gender	1	-0.0810	0.5912	0.0188	0.8910
Class2	1	-0.2019	0.7494	0.0726	0.7876
Class3	1	0.1339	0.6808	0.0387	0.8441

Odds Ratio Estimates

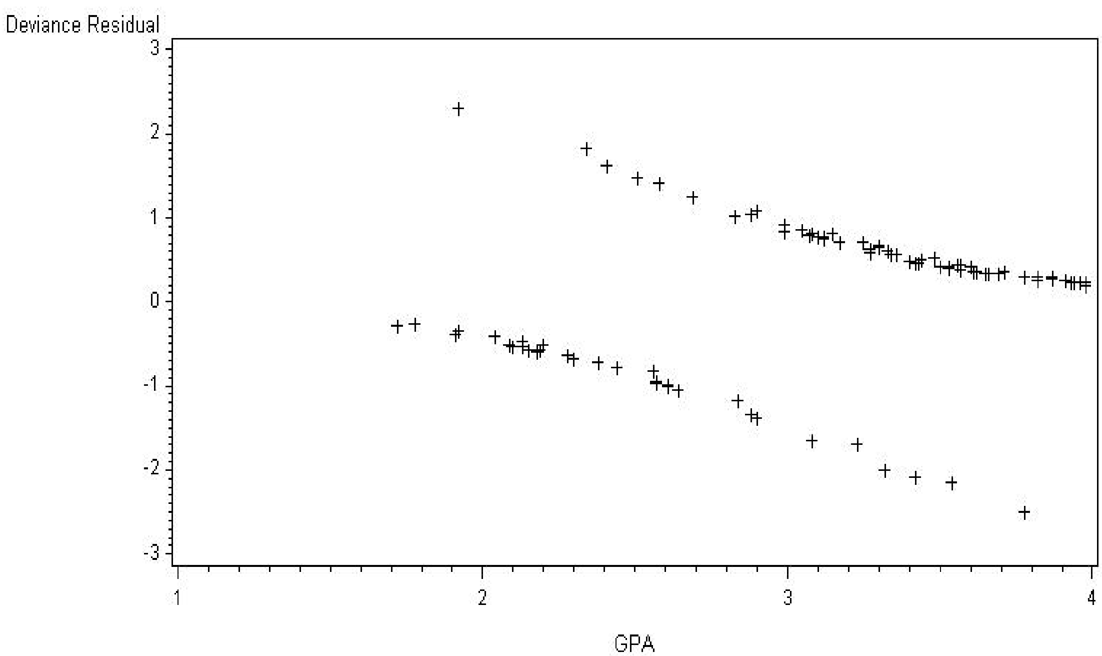
Effect	Point Estimate	95% Wald Confidence Limits	
GPA	21.696	6.674	70.533
Gender	0.922	0.289	2.938
Class2	0.817	0.188	3.550
Class3	1.143	0.301	4.341

Partition for the Hosmer and Lemeshow Test

		Proficiency = 1		Proficiency = 0	
Group	Total	Observed	Expected	Observed	Expected
1	9	1	0.69	8	8.31
2	9	1	1.43	8	7.57
3	9	3	2.85	6	6.15
4	9	4	4.50	5	4.50
5	9	7	6.31	2	2.69
6	9	8	7.08	1	1.92
7	9	7	7.84	2	1.16
8	9	8	8.21	1	0.79
9	10	9	9.42	1	0.58
10	11	11	10.66	0	0.34

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.6864	8	0.9525



Applied Statistics Comprehensive Exam

January 2013

Ph.D Methods Exam

This comprehensive exam consists of 10 questions pertaining to methodological statistical topics.

1 This Ph.D level exam will run from **8:30 AM to 3:30 PM**.

2 Please label each page with your identification number.

DO NOT USE YOUR NAME OR BEAR NUMBER.

3 Please write only on one side of each page.

4 Please leave one inch margins on all sides of each page.

5 Please number all pages consecutively.

6 Please label the day number (Day 1 or Day 2) on each page.

7 Please begin each question on a new page, and number each question.

8 Please do not staple pages together.

9 No wireless devices, formula sheets, or other outside materials are permitted.

10 Statistical tables and paper will be provided.

11 Relax and good luck!

I have read and understand the rules of this exam.

Signature:____ Date:____________

1.) Teachers are often evaluated using measures of "growth" based on students' standardized exam scores. Administrators are interested in whether different grades show different levels of typical growth. The following data present standardized measures of student growth for 5 randomly selected teachers from each of grades 3, 4, 5, and 6.

Growth Scores

	3^{rd} Grade	4^{th} Grade	5^{th} Grade	6^{th} Grade
1	12.3	8.6	10.2	7.5
2	11.0	9.5	10.0	8.6
3	8.4	7.1	6.5	7.9
4	5.7	4.3	5.3	3.2
5	7.6	6.7	8.2	5.8
Averages:	9.0	7.2	8.0	6.6

i. Construct an appropriate model that could be used to compare mean growth scores across grades. Include all assumptions for your model.

ii. Select one of your assumptions from i. Explain in detail how this assumption can be tested, and describe one change you could make to your model if this assumption is not met.

iii. Given $SSE = 97.25$, perform a test comparing growth across grades. Explain the meaning of the result using the language of student growth and grades.

Now suppose that teachers were randomly selected from 5 specific school districts (numbered 1 through 5 in the table), such that one teacher is selected from each grade within each district.

iv. Explain how your model from i. would change to account for the possible clustering within districts. Include all assumptions for your new model.

v. Explain in detail how the "no interaction" assumption can be assessed. Describe how you would change your model from iv if this assumption is not met.

2.) Suppose the mean income (in thousands of dollars) for an individual is to be predicted using years of education (from age 6), parents' mean income, age of the individual (in years), residential tax rate (in percent), and high school GPA (on a 0 - 100 scale).

i. Construct an appropriate model for this research interest. Include all assumptions of your model.

ii. Explain how each assumption can be assessed. Describe how you would change your model if the independence assumption fails due to cluster sampling.

iii. Describe what "multicollinearity" represents, and why it is a concern. Select variables from this data situation that you think may cause problems with multicollinearity, and explain why.

iv. Explain in detail at least two methods for detecting multicollinearity. Assuming you detect multicollinearity in this data set, describe how you would proceed with your model.

v. The following regression function is estimated.

$$\widehat{\text{Income}} = 9.63 + 1.2(\text{Years of Ed}) - 0.3(\text{Parents' Income}) + 0.6(\text{Age}) + 0.2(\text{Tax Rate}) - 0.4(\text{GPA})$$

Based on these parameter estimates, a colleague claims that "Years of Ed" is twice as important as "Age" and that higher "GPA" from high school tends to reduce future income. How would you respond to these statements?

3.) Consider a study in which high school basketball players are randomly assigned to one of three off-season training conditions: plyometric exercise for two months, weight-training for two months, or no extra training (control). Every players vertical leap ability is measured weekly for two months. The primary goal is to test whether the treatments have different effects on change in players jumping ability.

i. Identify the name of the design used.

ii. Describe how you would analyse these data.

iii. Discuss how this analysis would differ from a basic analysis of variance and why such modifications are necessary.

iv. Identify all of the effects that would be testable. For each of those effects describe what it means in practical language as if you were speaking to someone with minimal statistical training.

4.) Consider a One-Factor ANOVA Cell Means Model,

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where $i = 1, 2, 3$, $j = 1, \ldots, 5$, and

$$\mathbf{Y} = [1 \;\; 2 \;\; 4 \;\; 3 \;\; 0 \;\; 6 \;\; 4 \;\; 5 \;\; 8 \;\; 2 \;\; 5 \;\; 0 \;\; -2 \;\; 8 \;\; 4]^T.$$

i. Give the associated design matrix and parameter vector for this model.

ii. Consider performing pairwise comparisons between all combinations of group means. Present this as a General Linear Hypothesis of the form $\mathbf{C}^T\boldsymbol{\beta} = \mathbf{d}$.

iii. Show that your General Linear Hypothesis is testable.

iv. Perform a test of your General Linear Hypothesis. (HINTS: (1) Here $\hat{\boldsymbol{\beta}}$ will be group averages; (2) $\mathbf{Y}^T(\mathbf{I} - \mathbf{P})\mathbf{Y} = 94.0$; (3)

$$\begin{bmatrix} 2/5 & -1/5 & -1/5 \\ -1/5 & 2/5 & -1/5 \\ -1/5 & -1/5 & 2/5 \end{bmatrix}^{-} \approx \begin{bmatrix} 1.1 & -0.6 & -0.6 \\ -0.6 & 1.1 & -0.6 \\ -0.6 & -0.6 & 1.1 \end{bmatrix}.)$$

5.) Answer the following questions related to multivariate statistical methods.

i. Compare and contrast the use and purpose of predictive discriminant analysis vs. descriptive discriminant analysis.

ii. When discussing exploratory factor analysis, we often use the phrase "interpretable factor solution." Explain what this phrase means.

iii. Pretend you are working in the Research Consulting Lab. Use your imagination (do not use an example from SRM 610) to describe a research problem that a client might bring for which you would advocate the use of MANOVA. Then describe how you would convince them that MANOVA is the correct method.

6.) Experimental Design

i. Explain how a randomized complete block design with replication is different from a two-factor completely randomized design.

ii. What is a Latin square design? What are the advantages of Latin square designs? What are the disadvantages?

iii. Why would an experimenter use a fractional factorial design? What are some drawbacks of using that design? Briefly describe how you would construct a quarter-fraction design for a study involving 6 factors each with 2 levels.

7.) Given the data, us the Sign Test to test $H_0 : \tilde{\mu} = 8.41$ vs $H_1 : \tilde{\mu} > 8.41$.

$$8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80, 10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30$$

8.) Compare and contrast stratified random sampling to simple random sampling. What do these designs have in common? How are they different? Give examples/applications of each design.

9.) The SAS output gives a regression analysis of the systolic blood pressure (SBP), body size (QUET) a measure of size defined by QUET=100 (weight/height^2), age (AGE), and smoking history (SMK=0 if nonsmoker,SMK=1 if a current or previous smoker) for a hypothetical sample of 32 white males over 40 years old from the town of Angina. Note that QUMK=QUET*SMK.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4184.10759	1394.70253	17.42	<.0001
Error	28	2241.86116	80.06647		
Corrected Total	31	6425.96875			

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4120.36649	2060.18325	25.91	<.0001
Error	29	2305.60226	79.50353		
Corrected Total	31	6425.96875			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	49.31176	19.97235	2.47	0.0199
QUET	1	26.30283	5.70349	4.61	<.0001
SMK	1	29.94357	24.16355	1.24	0.2256
QUMK	1	-6.18478	6.93171	-0.89	0.3799

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	63.87603	11.46811	5.57	<.0001
QUET	1	22.11560	3.22996	6.85	<.0001
SMK	1	8.57101	3.16670	2.71	0.0113

i. Determine a single multiple model that uses the data for both smokers and nonsmokers and that defines straight-line models for each group with possibly differing intercepts and slopes. Obtain the least-square line for smokers and nonsmokers by using the single multiple model.

ii. Test H_0 : the two lines are parallel. State the appropriate null hypothesis in terms of the regression coefficients of the regression model.

iii. Suppose we fail to reject the null hypothesis in part (b) above. State the appropriate ANACOVA regression model to use for comparing the mean blood pressure in the two smoking categories, controlling for QUET.

10.) Consider the following data analysis, used to assess relationships between employment status (employed, self-employed, unemployed) and race (White, Black, Latino) for both males and females.

i. Explain the differences between a Homogeneous Association model and a Conditional Independence model (employment status and race are assumed independent, conditional on gender).

ii. Using the SAS output, evaluate the Conditional Independence assumption.

iii. Using the Conditional Independence model, give the estimated odds ratio for employed versus self-employed.

Conditional Independence

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8	13.7980	1.7248
Scaled Deviance	8	13.7980	1.7248
Pearson Chi-Square	8	13.2196	1.6525
Scaled Pearson X2	8	13.2196	1.6525
Log Likelihood		2124.9958	
Full Log Likelihood		-55.2658	
AIC (smaller is better)		130.5316	
AICC (smaller is better)		161.9602	
BIC (smaller is better)		139.4353	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates									
Parameter			**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept			1	3.7568	0.1160	3.5295	3.9842	1048.64	<.0001
Employment	emp		1	0.0397	0.1065	-0.1690	0.2483	0.14	0.7095
Employment	self		1	-1.3921	0.1704	-1.7261	-1.0581	66.74	<.0001
Race	black		1	0.3986	0.1306	0.1427	0.6546	9.32	0.0023
Race	latino		1	0.4389	0.1295	0.1850	0.6928	11.48	0.0007
Gender	f		1	0.1013	0.1619	-0.2160	0.4186	0.39	0.5315
Employment*Gender	emp	f	1	0.1523	0.1570	-0.1554	0.4601	0.94	0.3320
Employment*Gender	self	f	1	0.2563	0.2430	-0.2200	0.7326	1.11	0.2916
Race*Gender	black	f	1	-0.4325	0.1844	-0.7940	-0.0711	5.50	0.0190
Race*Gender	latino	f	1	-0.5169	0.1847	-0.8789	-0.1548	7.83	0.0051
Scale			0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Homogeneous Association

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	4	5.4046	1.3512
Scaled Deviance	4	5.4046	1.3512
Pearson Chi-Square	4	5.4024	1.3506
Scaled Pearson X2	4	5.4024	1.3506
Log Likelihood		2129.1925	
Full Log Likelihood		-51.0691	
AIC (smaller is better)		130.1382	
AICC (smaller is better)		270.1382	
BIC (smaller is better)		142.6034	

Algorithm converged.