Applied Statistics Comprehensive Exam

January 2020

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. Health policy researchers are interested in comparing the daily volume of hospital emergency room patients in three different locations (a Denver ER, a Colorado Springs ER, and a Fort Collins ER). The researchers also suspect daily volume will differ between weekdays (Monday through Thursday) and weekends (Friday through Sunday). For their study, these researchers randomly selected 20 weekdays and 20 weekend days from each hospital's records in the last year (giving 120 total observations) and recorded the daily patient volume for each.

   i. Present an appropriate *cell means model* to compare mean patient volume across both location and weekday / weekend.

   ii. Provide a contrast that could be used to compare the mean patient volume for Fort Collins to *both* Denver and Colorado Springs. Explain how the significance of this contrast can be tested.

   iii. Give a second contrast that is *orthogonal* to the contrast of part ii. What does this new contrast represent, in terms of the original problem?

   iv. Provide a contrast that can be used to evaluate whether the difference in mean patient volume between weekdays and weekends *differs* across the three locations. What is the common term for such a contrast in factorial ANOVA situations?

---

2. Auditing of caps for food jars gave the following results for caps with functional defects:

Shift	Total inspected	Number defective
1	74740	9
2	86880	5
3	83680	29

   Test that the defective ratio is the same for all three shifts, using $\alpha = .01$.

---

3. In an experiment 16 mongrel dogs were divided into four groups. Groups 1 and 2 received morphine sulphate intravenously, and groups 3 and 4 received intravenous trimethaphan. In addition, the dogs in groups 2 and 4 were treated so that their supplies of available histamine were depleted at the time of treatment, whereas groups 1 and 3 had intact histamine supplies. The blood histamine levels (μg/ml) were measured at minutes 1, 3, and 5 of treatment. Before the analyses, the responses were transformed by taking the base 10 logarithm to maintain consistency with the assumptions.

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

   Using the SAS output answer the following questions. For each question, state the null and alternative hypotheses, F-value, the p-value, and conclusion, and use $\alpha = .05$.

   ii. Is there a significant histamine supply effect?

   iii. Test whether or not histamine supply effect equally at different times.

   iv. Was there an interaction effect of treatment and histamine supply?

   v. Was the interaction effect of treatment and histamine supply the same at different times?

---

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.

   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, $\mathbf{c}^T\boldsymbol{\beta}$, and explain why estimability is important for Least Squares Estimation.

   iii. 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 General Linear Mixed Model (GLMM)

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

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

   Prove

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{V} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}' & \boldsymbol{R} \end{bmatrix} \right)$$

---

6. Let

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$. Define $Q = (X_1 - X_2)^2$ and $L = (X_1 + X_2)/2$.

   i. Show that $Q/2(1-\rho)$ is distributed as a χ^2 random variable, and give the parameters of the distribution.

   ii. Are Q and L distributed independently? Why?

---

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

---

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 $Y_1, \ldots, Y_N$ be independent and identically distributed random variables with pdf $f(y_i;\theta)$, where θ is the only parameter of the distribution.

   i. Describe in detail the process of Maximum Likelihood Estimation for the parameter θ, given the random sample $Y_1, \ldots, Y_N$. Introduce likelihood function notation as needed.

   ii. Assuming the *true value* of θ is indicated by θ_0, and assuming all necessary regularity conditions are met, discuss the *asymptotic distribution* of the MLE $\hat{\theta}$. Be sure to include discussion of the Fisher Information.

   iii. Now suppose the random variables $Y_1, \ldots, Y_N$ are independent and identically distributed as Poisson random variables, $Y_i \sim Poi(\lambda_0 = 10)$. Find the MLE $\hat{\lambda}$, and provide the asymptotic distribution of $\hat{\lambda}$.

---

10. Please respond to the following.

    i. Suppose a random vector $\mathbf{X}$ has joint pdf $f(\mathbf{x}, \boldsymbol{\theta})$, and let $\mathbf{S}$ indicate a statistic based on $\mathbf{X}$. Provide a definition of what it means for $\mathbf{S}$ to be jointly sufficient for $\boldsymbol{\theta}$.

    ii. Let $X_1, \ldots, X_n$ be a random sample such that $X_i \sim geo(p)$, independently and identically distributed. Find the MLE $\hat{p}$ for p, and show that $S = \sum_{i=1}^{n} X_i$ is sufficient for p.

    $$\left(\text{HINT: } f(x,p) = p(1-p)^{(x-1)} \ , \ x \in \mathcal{Z}^{+} \right)$$

    iii. State why sufficiency is of interest in situations in which a unique MLE can be identified.

SAS output for question 3

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
Treatment	2	Morphine Trimethaphan
Histamine Supply	2	Depleted Intact

Number of Observations Read	16
Number of Observations Used	16

The GLM Procedure
Repeated Measures Analysis of Variance

Sphericity Tests				
Variables	**DF**	**Mauchly's Criterion**	**Chi-Square**	**Pr > ChiSq**
Transformed Variates	2	0.4616166	8.5032276	0.0142
Orthogonal Components	2	0.5695519	6.1919585	0.0452

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	1.41530042	1.41530042	4.08	0.0663
Histamine Supply	1	4.83597029	4.83597029	13.94	0.0029
Treatment*Histamine	1	1.64763216	1.64763216	4.75	0.0499
Error	12	4.16219352	0.34684946		

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**
Time	2	0.26402726	0.13201363	19.34	<.0001	0.0001	<.0001
Time*Treatment	2	0.03229083	0.01614542	2.37	0.1154	0.1360	0.1315
Time*Histamine Supply	2	0.36785338	0.18392669	26.95	<.0001	<.0001	<.0001
Time*Treatment*Histamine	2	0.03875012	0.01937506	2.84	0.0782	0.1006	0.0954
Error(Time)	24	0.16380107	0.00682504				

Greenhouse-Geisser Epsilon	0.6991
Huynh-Feldt-Lecoutre Epsilon	0.7629

Applied Statistics Comprehensive Exam

August 2019

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 recent study was conducted to assess the effectiveness of a new weight-loss drug. After random selection, eight mice were given a single dosage, eight mice were given a double dosage, and eight mice were given a placebo. Drug 1 is a single dosage and Drug 2 is a double dosage. The gender of the mice was also recorded, with four males and four females in each group. The summary data are given below.

Treatment	Gender	n	mean	variance
Drug 1	Male	4	7.25	1.58
Drug 1	Female	4	7.75	2.92
Drug 2	Male	4	16.00	15.33
Drug 2	Female	4	13.50	6.33
Placebo	Male	4	6.50	3.66
Placebo	Female	4	7.50	1.67

   i. Write down a contrast that compares Drug 1 and the Drug 2.

   ii. Write down a contrast that is orthogonal to the contrast in (i).

   iii. Test the contrasts in (i).

---

2. Integrated Pest Management (IPM) adopters apply significantly less insecticides and fungicides than nonadopters among grape producers. A research contained the following adoption rates for the six states that account for most of the U.S. production. A survey of 712 grape-producing growers asked whether or not the growers were using an IPM program on the farms.

	State						
	Cal.	Mich.	New York	Oregon	Penn.	Wash.	Total
IPM Adopted	39	55	19	22	24	30	189
IPM Not Adopted	92	69	114	88	83	77	523
Total	131	124	133	110	107	107	712

   Is there significant evidence that the proportion of grape farmers who have adopted IPM is different across the six states? Take $\alpha = .05$.

---

3. Consider a study intended to assess the drying times of oil-based paints with differing levels of turpentine mixed in. Researchers painted identical wooden boards using three different levels of turpentine (low, medium, high) and exposed to four different temperatures (low, medium, high, extreme), and recorded the number of minutes until the paint dried. The process was replicated five times for each combination of turpentine and temperature. The data follow.

		Temperature Level			
		Low	Medium	High	Extreme
Turpentine	Low	43, 52, 55, 41, 48	44, 51, 54, 55, 51	54, 60, 59, 58, 61	75, 72, 58, 68, 71
	Medium	38, 43, 50, 40, 38	33, 38, 46, 49, 52	50, 58, 63, 54, 65	70, 71, 68, 62, 67
	High	34, 39, 46, 40, 32	34, 30, 31, 53, 36	40, 50, 55, 61, 60	62, 78, 63, 64, 57

i. Propose *at least three* descriptive statistics that could be used to explore the data.

ii. Suppose the researchers would like to know whether there is evidence that the trends in typical time to dry across increasing temperatures *differ* according to turpentine levels. What type of model would you recommend? Show the model equation, list the assumptions, and present an appropriate hypothesis test to address this question.

iii. Show a *partial* ANOVA table for your model from part ii, including only sources of variation, degrees of freedom, and *formulas* for the appropriate sums of squares. What would be the distribution of the associated test statistic for your test from part ii?

iv. Explain how your response to part ii would change if the cells did *not* all include five observations each, but instead showed variation in the number of replicates. How would you perform your hypothesis test?

v. Explain how your response to parts ii and iii would change if the three levels of turpentine were selected randomly from a population of turpentine levels.

---

4. Please respond to the following.

   i. Suppose a random vector $\mathbf{X}$ has joint pdf $f(\mathbf{x}, \boldsymbol{\theta})$, and let $\mathbf{S}$ indicate a statistic based on $\mathbf{X}$. Provide a definition of what it means for $\mathbf{S}$ to be jointly sufficient for $\boldsymbol{\theta}$.

   ii. Let $X_1, \ldots, X_n$ be a random sample such that $X_i \sim geo(\theta)$, independently and identically distributed. Find the MLE $\hat{\theta}$ for θ.

   $$\left(\text{HINT: } f(x, \theta) = \theta(1-\theta)^{(x-1)} \ , \ x \in \mathcal{Z}^+ \right)$$

   iii. Show that $S = \sum_{i=1}^{n} X_i$ is sufficient for θ.

---

5. Assume $X_1, \cdots, X_n$ is a random sample from $(1/\theta)x^{(1-\theta)/\theta} I_{[0,1]}(x)$, where $\theta > 0$. Find a 100γ percent confidence interval for θ. Find its expected length.

---

6. Let $X_1, \cdots, X_n$ be a random sample from the Bernoulli distribution, say $P[X = 1] = \theta = 1 - P[X = 0]$.

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

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

---

7. Let $X_1, \cdots, X_m$ be a random sample from the density $\theta_1 x^{\theta_1 - 1} I_{[0,1]}(x)$, and let $Y_1, \cdots, Y_n$ be a random sample from the density $\theta_2 y^{\theta_2 - 1} I_{[0,1]}(y)$. Assume that the samples are independent. Set $U_i = -\ln X_i$, $i = l, ..., m$, and $V_j = -\ln Y_j$.

   i. Find the generalized likelihood-ratio for testing $H_0 : \theta_1 = \theta_2$ versus $H_a : \theta_1 \neq \theta_2$.

   ii. Show that the generalized likelihood-ratio test can be expressed in terms of the statistic

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

   iii. If H_0 is true, what is the distribution of T? (You do not have to derive it if you know the answer.)

---

8. In the model $Y_{ij} = \mu + \tau_j + \epsilon_{ij}, \quad j = 1, \cdots, n_i, i = 1, \cdots, 3$, $\epsilon_{ij} \sim N(0, \sigma^2)$, derive a test for $H_0 : (\mu + \tau_1)/5 = (\mu + \tau_2)/10 = (\mu + \tau_3)/15$.

---

9. Consider the General Linear Model with independent errors (with mean 0 and variance σ^2). If the *true* design matrix and parameter vector for a regression model are $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1|\boldsymbol{\beta}_2]^T$, respectively, but the model is *underfit* using design matrix $\mathbf{X}_1$ and parameter vector $\boldsymbol{\beta}_1$, find the bias in the least squares estimator $\hat{\boldsymbol{\beta}}_1$.

---

10. Consider a balanced Blocked ANOVA model:

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

where $i = 1, \ldots, 4$, $j = 1, \ldots, 3$, and b_j represents the blocking factor.

i. Write the model in the vector notation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, showing the design matrix explicitly. Is your design matrix full rank?

ii. Provide an expression for the Least Squares Estimator for $\boldsymbol{\beta}$.

iii. Suppose it is of interest to perform mean comparisons on the factor of interest, α_i. In scalar notation, it is of interest to test $H_0 : \alpha_i = \alpha_j \ , \forall i \neq j$. Write this as a *single* General Linear Hypothesis, explicitly showing all relevant matrices and vectors.

iv. Determine whether the test from part iii is testable, and explain your answer.

v. Show the test statistic (in terms of your notation from part iii) and associated null distribution for the test from part iii, explaining why the test statistic is distributed as described. Be sure to provide the name of the distribution and degrees of freedom.

Applied Statistics Comprehensive Exam

January 2019

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. Consider an experiment designed to assess typical walking activity relative to different shoe brands. Fitness researchers have randomly selected twelve participants (six males and six females), and have assigned half to wear Nike shoes, half to wear Adidas shoes. (Three males were given Nike shoes, three were given Adidas shoes, and similarly for the female participants.). Using activity watches, the participants have their daily step counts recorded for four consecutive days. The data follow.

Gender	Shoe Brand	Day 1	Day 2	Day 3	Day 4
Female	Nike	7200	1100	11000	8300
		5600	1200	8500	9300
		13056	7048	16212	10012
	Adidas	8593	9726	10152	14000
		12000	6093	9072	8984
		9000	10000	10000	11000
Male	Nike	8123	12345	13333	10451
		12000	5048	11000	9012
		7000	10345	8000	11451
	Adidas	8000	8500	9000	12984
		10000	12012	12000	13073
		13200	15012	14089	10073

   i. Assuming the researchers' primary interest is in changes in typical step counts over time, propose *at least three* descriptive statistics (or visuals) to explore the data. Using the output on pages 3-6 provide descriptive interpretations for the data.

   ii. Propose an appropriate model that could be used to assess differences in trends over time. Be sure to explain the meaning of all terms in your model.

   iii. Provide the assumptions of your model. How can each assumptions be evaluated? Using the output on pages 3-6, assess any assumptions that you can.

   iv. Give a *partial* ANOVA table, in which sources of variation, sums of squares formulas (no need to calculate the values!), and degrees of freedom are shown.

   v. Using the output on pages 3-6, perform a test to assess whether the change in step counts appears to differ between shoe brands. Be sure to include the distribution of the test statistic (under H_0), degrees of freedom, and an explanation of the meaning of the result.

2. Quality control experts are interested in comparing the weights of transistors produced using three different processes (process 1, process 2, and process 3) and also two different types of materials (material 1 and material 2). For their study, these experts randomly selected 20 transistors produced from each of the six combinations of process and material (giving 120 total observations) and recorded the weight of each.

   i. Present an appropriate *cell means model* to compare mean weight across both process and type of material.

   ii. Give a contrast that could be used to compare the mean weight for process 3 to *both* process 1 and process 2. Explain how the significance of this contrast can be tested.

   iii. Give a second contrast that is *orthogonal* to the contrast of part ii. What does this new contrast represent, in terms of the original problem?

   iv. Provide a contrast that can be used to evaluate whether the difference in mean weight between the two types of materials *differs* across the three processes. What is the common term for such a contrast in factorial ANOVA situations?

---

3. Nylon bars were tested for brittleness. Each of 280 bars was molded under similar conditions and was tested in five places. Assuming that each bar has uniform composition, the number of breaks on a given bar should be binomially distributed with five trials and an unknown probability π of failure. If the bars are all of the same uniform strength, π should be the same for all of them; if they are of different strengths, π should vary from bar to bar. The following table summarizes the outcome of the experiment:

Breaks/Bar	0	1	2	3	4	5
Frequency	157	69	35	17	1	1

   Test the agreement of the observed frequency distribution with the binomial distribution. Take $\alpha = .05$.

---

4. If Z_1 and Z_2 iid $\sim N(0,1)$, does Z_1^2 have the same distribution as $|Z_1 Z_2|$? Justify.

---

5. Assume $X_1, \cdots, X_n$ iid $\sim f(x;\alpha,\beta) = \frac{1}{2\beta} e^{-\frac{|x-\alpha|}{\beta}}$, where $\beta > 0, \theta = (\alpha,\beta)$.
   i. Find MME of (α,β).
   ii. Find MLE of (α,β).

---

6. Assume $X_1, \cdots, X_n$ iid $\sim N(\theta,\theta), \theta > 0$. Do we have an exponential class? If so, find a complete sufficient statistic.

---

7. Assume $X_1, \cdots, X_n$ iid $\sim N(\theta,\theta), \theta > 0$. Show that $\bar{X}$ is not UMVUE of θ. Argue that an UMVUE of θ exists. (you don't have to find it!) Find the $CRLB_\theta(\theta)$.

---

8. For the normal general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with design matrix of full rank k, identify and prove the distribution of
$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2},$$
where $\hat{\boldsymbol{\beta}}$ indicates Least Squares Estimators.

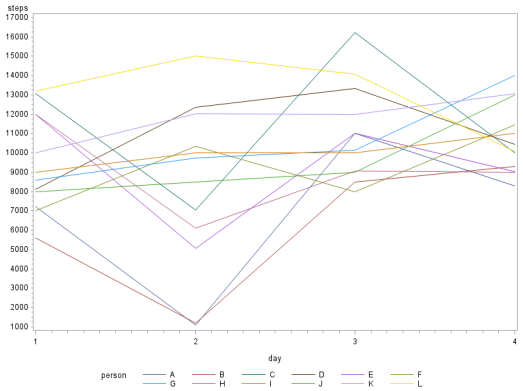
---

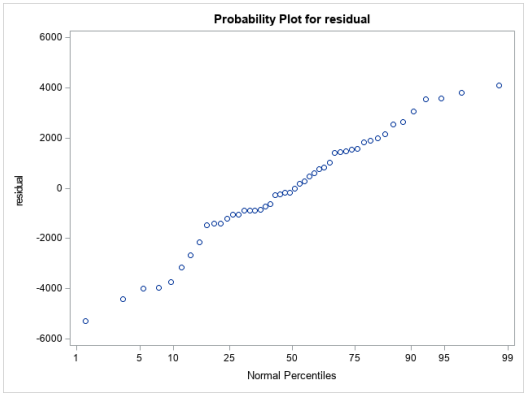
9. For a real-valued matrix $\mathbf{X}$, assume $\mathbf{P}_X$ projects onto the column space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$. Prove that $\mathbf{I} - \mathbf{P}_X$ projects onto the perp space, $\mathcal{C}(\mathbf{X})^\perp$.

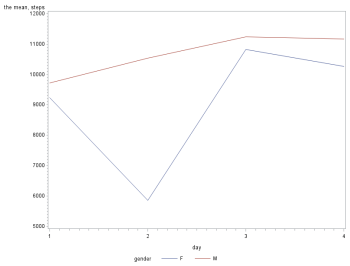
---

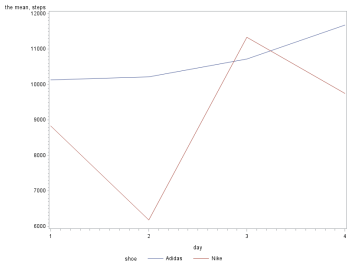
10. Consider the two-factor balanced additive random-effects model without interaction
$$Y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$
$$i = 1,2, \quad j = 1,2, \quad k = 1,2.$$
Suppose ϵ_{ijk} are iid $N(0,\sigma^2)$ variables, a_i are iid $N(0,\sigma_a^2)$ variables, b_i are iid $N(0,\sigma_b^2)$ variables, $Cov(\epsilon_{ijk}, a_i) = 0$, $Cov(\epsilon_{ijk}, b_j) = 0$, and $Cov(a_i, b_j) = 0$.
   i. Write this model in the General Linear Mixed Model form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$.
   ii. Find an expression for $\boldsymbol{V} = \text{Cov}(\boldsymbol{Y})$.

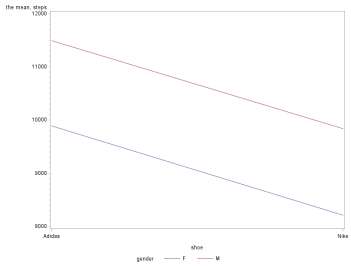
Output for Question 1











Output for Question 1

The UNIVARIATE Procedure
Variable:
residual

Moments			
N	48	**Sum Weights**	48
Mean	0	**Sum Observations**	0
Std Deviation	2248.31388	**Variance**	5054915.29
Skewness	-0.314118	**Kurtosis**	-0.2511913
Uncorrected SS	237581019	**Corrected SS**	237581019
Coeff Variation	.	**Std Error Mean**	324.516156

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.97709	**Pr < W**	0.4640
Kolmogorov-Smirnov	**D**	0.090644	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.041196	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.30959	**Pr > A-Sq**	>0.2500

Output for Question 1

The GLM Procedure
Repeated Measures Analysis of Variance

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|				
DF = 8	**day1**	**day2**	**day3**	**day4**
day1	1.000000	0.237488 0.5384	0.733543 0.0245	-0.464736 0.2075
day2	0.237488 0.5384	1.000000	0.693547 0.0383	0.256219 0.5058
day3	0.733543 0.0245	0.693547 0.0383	1.000000	-0.094414 0.8091
day4	-0.464736 0.2075	0.256219 0.5058	-0.094414 0.8091	1.000000

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|		
DF = 8	**day_1**	**day_2**
day_1	1.000000	0.465521 0.2067
day_2	0.465521 0.2067	1.000000
day_3	0.856089 0.0032	0.740882 0.0224

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|	
DF = 8	**day_3**
day_1	0.856089 0.0032
day_2	0.740882 0.0224
day_3	1.000000

Sphericity Tests				
Variables	**DF**	**Mauchly's Criterion**	**Chi-Square**	**Pr > ChiSq**
Transformed Variates	5	0.0877076	16.36019	0.0059
Orthogonal Components	5	0.316263	7.7384944	0.1712

Output for Question 1

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
gender	1	31530071.0	31530071.0	2.34	0.1648
shoe	1	33211777.7	33211777.7	2.46	0.1553
gender*shoe	1	1376.0	1376.0	0.00	0.9922
Error	8	107918084.3	13489760.5		

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

Source	DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H-F-L
day	3	59991865.7	19997288.6	3.94	0.0203	0.0378	0.0222
day*gender	3	37924079.1	12641359.7	2.49	0.0845	0.1108	0.0878
day*shoe	3	33260740.1	11086913.4	2.18	0.1161	0.1417	0.1194
day*gender*shoe	3	13526348.7	4508782.9	0.89	0.4613	0.4346	0.4581
Error(day)	24	121818927.7	5075788.7				

Greenhouse-Geisser Epsilon	0.7014
Huynh-Feldt-Lecoutre Epsilon	0.9575

Applied Statistics Comprehensive Exam

August 2018

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. Researchers have taken a stratified random sample including 1000 male undergraduate students and 1000 female undergraduate students. As part of their study, they have asked students about their majors, and have totaled the numbers of students falling into each category: 450 males in STEM (Science, Technology, Engineering, and Math), 250 males in Social Sciences, 200 males in Liberal Arts, and 100 males with undeclared majors. For females the totals are 275 in STEM, 350 in Social Sciences, 275 in Liberal Arts, and 100 undeclared.

   i. Present the data in a table to summarize the totals.

   ii. What is the proportion of males who reported liberal arts majors? What is the proportion of liberal arts majors who were male? Of these two statistics, which is more meaningful, given the type of sampling?

   iii. Researchers are interested in whether the distribution of major differs between male and female students. What type of test would you apply, and why would you choose that test?

---

2. Suppose education researchers are interested in evaluating high school students' standardized math exam performances over three years, with one exam given at the end of each year. The researchers have developed an experimental classroom activity that is intended to improve students' critical thinking ability. Two groups of students are randomly selected: one group is given the experimental instruction while the second is not. Both groups are followed for three years to track standardized math exam scores. At the end of the study, researchers would like to determine whether (i) mean exam scores differ between the two groups, and (ii) trends in mean exam scores (over time) differ between the two groups.

   i. Propose *at least three* descriptive statistics you would use to explore the data. (Here descriptives can be calculated values or visuals / plots.).

   ii. Propose an appropriate statistical model to address the researchers' questions, and identify the model parameters associated with each of the researchers' two questions.

   iii. List the assumptions of your model from part ii.

   Now suppose the experiment has been expanded to include *two groups each* from *ten different schools*, followed for three years as before.

   iv. How would you change your model from part ii to account for the "school" effect?

   v. Describe how you would address the attrition of students; i.e. how would you handle students with missing exam scores?

---

3. A recent study was conducted to assess the effectiveness of a new weight-loss drug. After random selection, eight mice were given a single dosage, eight mice were given a double dosage, and eight mice were given a placebo. Drug 1 is a single dosage and Drug 2 is a double dosage. The gender of the mice was also recorded, with four males and four females in each group. The summary data are given below.

Treatment	Gender	n	mean	variance
Drug 1	Male	4	7.25	1.58
Drug 1	Female	4	7.75	2.92
Drug 2	Male	4	16.00	15.33
Drug 2	Female	4	13.50	6.33
Placebo	Male	4	6.50	3.66
Placebo	Female	4	7.50	1.67

   i. Write down a contrast that compares Drug 1 and the placebo.

   ii. Write down a contrast that is orthogonal to the contrast in (a).

   iii. Test the contrasts in (a) and (b).

---

4. If $X_1, X_2, \ldots, X_n$ is a random sample from $X \sim Exp(\theta)$, find the likelihood ratio test for $H_0: \ \theta \geq \theta_0$ vs $H_1: \ \theta < \theta_0$. Do not find the distribution of the test statistic.

---

5. If a random sample of size 5 yields: $\{15, 8, 19, 10, 11\}$, use this information to find a 90% two-sided confidence interval for θ based on $Y_n = X_{n:n}$ if:

$$f(x;\theta) = \begin{cases} \frac{2x}{\theta^2} & 0 < x < \theta, \quad \theta > 0 \\ 0 & \text{elsewhere} \end{cases}$$

---

6. If $X \sim N(\mu = 7, \sigma^2 = 25)$, $Y \sim \chi^2(6)$, and $Z \sim \chi^2(10)$ are all mutually independent, identify the distribution of:

   i. $Y + Z$

   ii. $\frac{X-7}{5\sqrt{Z/10}}$

   iii. $\frac{25Y}{6(X-7)^2}$

---

7. If $\{5.1, 4.3, 2.8, 3.3, 2.8\}$ is a random sample from $X \sim Gam(\theta, \kappa = 1)$, find a UMP test of size $\alpha = 0.05$ for $H_0: \ \theta \leq 2$ vs $H_1: \ \theta > 2$ and then run the test with the data given.

---

8. State and prove the Gauss-Markov Theorem.

---

9. Consider a traditional 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 $k = 1, 2, 3$.

i. Construct an appropriate design matrix $\mathbf{X}$ for this model.

ii. Suppose it is of interest to test the hypothesis $H_0 : (\alpha\beta)_{ij} = 0, \forall i, j$. 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.

---

10. Assume that first ($\mu_{\boldsymbol{Y}_0}$) and second ($\boldsymbol{\Sigma}_{\boldsymbol{Y}_0}$) moments for $\boldsymbol{Y}_0$ exist. we also assume the existence of $E(\boldsymbol{Y}) = \boldsymbol{\mu}_{\boldsymbol{Y}}$, $Cov(\boldsymbol{Y}) = \Sigma_{\boldsymbol{Y}}$, and $Cov(\boldsymbol{Y}, \boldsymbol{Y}_0) = \Sigma_{\boldsymbol{Y}\boldsymbol{Y}_0}$. We know, a best linear predictor (BLUP)of $\boldsymbol{Y_0}$ based on $\boldsymbol{Y}$ has the form of

$$\hat{Y}_0 = \mu_{\boldsymbol{Y}_0} + \boldsymbol{\beta}^{*\prime}(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Y}})$$

where $\boldsymbol{\beta}^*$ is a solution to

$$\Sigma_{\boldsymbol{Y}}\boldsymbol{\beta} = \Sigma_{\boldsymbol{Y}\boldsymbol{Y}_0}.$$

Now,consider the General Linear Mixed Model (GLMM)

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

where $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}$, $\mathrm{E}[\boldsymbol{u}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{G}$, $\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{u}) = \mathbf{0}$.

Show for the GLMM if $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{V}$ is known, the BLUP of $\boldsymbol{v} = \boldsymbol{m}'\boldsymbol{u}$ is $\hat{\boldsymbol{v}} = \boldsymbol{m}'\hat{\boldsymbol{u}}$ where

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y} \\ \hat{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

Applied Statistics Comprehensive Exam

January 2018

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, X_2, \cdots, X_{10}$ be a random sample from $X \sim Gam(\theta = 8, \kappa = 6)$. Find the pdf of $Y = 4X_1 + 4X_2 + \cdots + 4X_{10}$.

2.) Let $X_1, X_2, \cdots, X_n$ be a random sample from $X \sim Unif(0, \theta)$, $\theta > 0$.

(a) Find the MME of θ. Is it unbiased?

(b) Find the MLE of θ. Is it unbiased?

3.) Let $X_1, X_2, \cdots, X_5$ be a random sample from $X \sim Poi(\theta)$. Find the UMP size $\alpha = 0.05$ test for $H_0: \ \theta = 1$ vs $H_1: \ \theta > 1$. Identify the test statistic and the corresponding critical value for this test. Also, find $\pi_\phi(\theta = 3)$.

4.) If X and Y have the joint pdf given by:

$$f(x,y) = \begin{cases} x + y, & 0 < x < 1, \ 0 < y < 1 \\ 0, & elsewhere \end{cases}$$

Find:

a. $f(x|y)$

b. $E[3X|y]$

c. $E[XY]$.

5.) Compare and contrast the Mann-Whitney (aka Wilcoxon Rank Sum) Test with the Median Test. Discuss their similarities/differences.

6.) For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, recall $SSR = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} = \mathbf{Y}^T\mathbf{P}_x\mathbf{Y}$ and $SSE = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} = \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_x)\mathbf{Y}$. Assume $\mathbf{Y}$ has n components.

a. Find $\mathrm{E}[SSR]$ in terms of σ^2, $\boldsymbol{\beta}$, $\mathbf{X}$, n, and $rank(\mathbf{X})$.

b. Find $\mathrm{E}[SSE]$ in terms of σ^2, $\boldsymbol{\beta}$, $\mathbf{X}$, n, and $rank(\mathbf{X})$.

c. Show that SSR and SSE are distributed independently of each other.

7.) For the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, prove that $\hat{\boldsymbol{\beta}}$ is a minimizer of the quadratic form $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ *if and only if* $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_X\mathbf{Y}$, where $\mathbf{P}_X$ is the projection onto the column space $\mathcal{C}(\mathbf{X})$.

8.) Consider the two-factor balanced additive random-effects model without interaction

$$Y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$
$$i = 1, 2, \quad j = 1, 2, \quad k = 1, 2.$$

Suppose ϵ_{ijk} are iid $N(0, \sigma^2)$ variables, a_i are iid $N(0, \sigma_a^2)$ variables, b_i are iid $N(0, \sigma_b^2)$ variables, $Cov(\epsilon_{ijk}, a_i) = 0$, $Cov(\epsilon_{ijk}, b_j) = 0$, and $Cov(a_i, b_j) = 0$.

a. Write this model in the General Linear Mixed Model form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$.

b. Find an expression for $\boldsymbol{V} = \text{Cov}(\boldsymbol{Y})$.

9.) Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from $Y_i \sim Gamma(v, \theta),\ i = 1, \cdots, n$ assuming that v is known

a. Find the Jeffreys' prior.

b. Prove that the Gamma distribution is a conjugate prior for θ.

c. Find the posterior mean and variance for θ.

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

a. Find a conjugate prior distribution for θ.

b. Find the posterior mean and variance for θ.

Applied Statistics Comprehensive Exam

August 2017

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, X_2, \ldots, X_{10}$ be a random sample from $X \sim GAM(\theta = 8, \kappa = 6)$. Find the pdf of $Y = 4X_1 + 4X_2 + \cdots + 4X_{10}$.

2.) Let $X_1, \ldots, X_n$ be a random sample from $X_i \sim Pois(\theta)$ and we want to estimate $\tau(\theta) = \theta$.

i. Find an unbiased estimator T of $\tau(\theta)$.

ii. Find the variance of T.

iii. Find the CRLB of $\tau(\theta) = \theta$.

iv. Is T the UMVUE of $\tau(\theta) = \theta$? Why or why not?

3.) Consider the family of distributions,

$$f(x|p) = p^x(1-p)^{1-x}, \;\; x = 0, 1,$$

where $p \in \{1/2, 2/3\}$. Is this family of distributions complete? Justify your answer.

4.) In general terms, describe the procedure used to perform the Sign Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$, where μ_D = mean of the difference "Pre-Post"). If needed, construct a sample table of data to explain the procedure.

5.) Consider and experiment in which an industrial engineer would like to assess six different two-level factors (think of these as factors A, B, C, D, E, and F, each with "+" and "−" levels). The experimental units to be sacrificed for testing are very expensive laptop circuits.

i. Explain why this industrial engineer may *not* want to use a full factorial 2^6 design. If she is not financed sufficiently to perform $2^6 = 64$ runs, but can afford 16 runs, determine a more appropriate type of design that can be applied.

ii. Assume the industrial engineer will apply a 2^{6-2} design, using $I = ABCE$ and $I = BCDF$ as the design generators. Produce a table showing the treatment combinations applied for each of the 16 runs.

iii. Explain the meaning of the *alias structure* for this design, and produce the alias structure for all main effects and two-factor interactions.

iv. Explain the meaning of the term *resolution* in this context. What is the resolution of this design?

6.) Consider the *balanced* one-factor random effect model,

$$Y_{ij} = \mu + a_i + \epsilon_{ij},$$

where $i = 1, \ldots, I$, $j = 1, \ldots, n$, $a_i \sim \mathcal{N}(0, \sigma_a^2)$, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ with all random effects independent.

i. Derive an expression for the response variance-covariance structure, $\mathbf{V}$.

ii. Show that the GLS estimator for μ is equivalent to the OLS estimator, $\bar{Y}_{\cdot\cdot}$.

(HINT: Use $(\mathbf{I} + \gamma\mathbf{J})^{-1} = (\mathbf{I} - \frac{\gamma}{1+n\gamma}\mathbf{J})$.)

7.) Consider the General Linear Mixed Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $\text{Cov}(\boldsymbol{u}) = \mathbf{G} = \sigma_u^2\mathbf{I}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The BLUP for $\boldsymbol{u}$ is given by

$$\hat{\boldsymbol{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbf{V} = \text{Cov}(\mathbf{Y})$ and $\hat{\boldsymbol{\beta}}$ indicates GLS estimators for $\boldsymbol{\beta}$. Prove the "U" in BLUP. That is, prove that

$$\text{E}[\hat{\boldsymbol{u}} - \boldsymbol{u}] = \mathbf{0}.$$

8.) The zero-truncated Poisson distribution can be derived from the Poisson distribution by conditioning on positive counts.

i. Show that the zero-truncated Poisson pmf can be written,

$$f_{>0}(y; \lambda) = \frac{\lambda^y}{y!(e^\lambda - 1)},$$

where λ is the rate parameter from the original Poisson distribution.

(HINTS: $f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$, $P(Y = y|\lambda) = P(Y = y|\lambda, y > 0) \times P(Y > 0|\lambda)$)

ii. Show that the mean of the zero-truncated Poisson can be written,

$$\text{E}[Y] = \frac{\lambda}{1 - e^{-\lambda}}.$$

(HINT: For $y \in (1, \infty)$, define $z = y - 1$.)

9.) Consider the $p \times 1$ vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ representing a random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are mutually independent and each with distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

i. Write down the joint density function of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

ii. Explain in plain language suitable for a layperson the concept of likelihood function with reference to the above problem. Are there any differences between a joint density function and likelihood function? Justify your answer.

iii. Derive the maximum likelihood estimates of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ for the problem stated above. You must show your work to receive credit.

10.) Assume the model is

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

where $\boldsymbol{X}$ is full rank, $E\boldsymbol{\epsilon} = \boldsymbol{0}, \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, $\boldsymbol{V}$ is a known positive definite matrix. Find the generalized least-squares estimator of $\boldsymbol{\beta}$.

Applied Statistics Comprehensive Exam

January 2017

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let X_1 and X_2 be a random sample of size 2 from a Poison distribution with mean λ.

i. Find distribution of $X_1 + X_2$.

ii. Find conditional distribution of $X_1 + X_2$ given that $X_1 > 0$.

iii. Compute $P(X_1 \leq 3 | X_2 \leq 5)$.

iv. Show that $Var(X_1 + X_2) \geq Var[E(X_1 + X_2|X_1)]$

2.) Let $X_1, \ldots, X_n$ be a random sample from $X_i \sim UNIF(0, \theta)$, $\theta > 0$.

i. Find the MME of θ. Is it unbiased?

ii. Find the MLE of θ. Is it unbiased?

3.) Let $X_1, X_2, \ldots, X_5$ be a random sample from $X_i \sim Poi(\theta)$, Find the UMP size $\alpha = 0.05$ test for $H_0 : \theta = 1$ versus $H_1 : \theta > 1$. Identify the test statistic and the corresponding critical value for this test. Also, find $\pi(\theta = 3)$.

4.) Often times in practice, we are focused on the nature of a population distribution function. For example, the validity of a parametric test depends on the shape of the population from which the sample was drawn. When we do not know the functional form of the population, we first want to test whether the population of interest is likely to be distributed according to the assumptions underlying the proposed parametric procedure. Identify three nonparametric procedures that test if a sample has been drawn from a population that is normally distributed. Discuss the advantages and disadvantages of these three tests.

5.) Suppose you are interested in designing an experiment in which you would like to control two nuisance sources of variation. Propose a design and justify your selection. Your answer must address the following points:

i. Provide a list of factors and the outcome variable appropriate for this design. You do not have to give a real example. You may use notations such as A, B etc. to denote factors and Y to denote the outcome variable.

ii. What particular design would you choose and why?

iii. Write the statistical model appropriate for the proposed design and identify its components.

iv. Provide the list of assumptions necessary for the model you consider in this study.

v. Sketch the ANOVA table showing the columns: sources of variation, df, SS, MS, and F.

6.) Consider the two-factor balanced additive random-effects model without interaction,

$$Y_{ijk} = \mu + a_i + b_j + \epsilon_{ijk}$$
$$i = 1, 2, \quad j = 1, 2, \quad k = 1, 2.$$

Suppose ϵ_{ijk} are iid $N(0, \sigma^2)$ variables, a_i are iid $N(0, \sigma_a^2)$ variables, b_i are iid $N(0, \sigma_b^2)$ variables, $Cov(\epsilon_{ijk}, a_i) = 0$, $Cov(\epsilon_{ijk}, b_j) = 0$, and $Cov(a_i, b_j) = 0$.

i. Write this model in the General Linear Mixed Model form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$.

ii. Find an expression for $\boldsymbol{V} = \text{Cov}(\boldsymbol{Y})$.

7.) Let $\boldsymbol{x} \sim \mathcal{N}_n(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n)$.

i. Find the distributions of $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $Q = \sum_{i}^{n}(x_i - \bar{x})^2$.

ii. Show Q and $\bar{X}$ are independent.

8.) Consider an $I \times J$ contingency table under the assumption of Multinomial sampling; that is, all cells are jointly distributed according to a multinomial distribution.

i. Using notation to represent cell, row, and column probabilities, state the null and alternative hypotheses for the Test of Independence.

ii. Derive the formula for the likelihood ratio test statistic, G^2. (HINT: $l(\boldsymbol{\pi}, \mathbf{n}) = \frac{n_{++}!}{n_{11}!\ldots n_{IJ}!}\Pi_{i,j}\pi_{ij}^{n_{ij}}$.)

9.) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a distribution with mean vector μ and covariance matrix $\boldsymbol{\Sigma}$.

(*a*) Show that $\bar{\mathbf{X}}$ is an unbiased estimator for μ

(*b*) Let $\mathbf{S_n}$ denote the sample covariance matrix. Show that $\frac{n\mathbf{S_n}}{n-1}$ is an unbiased estimator for $\boldsymbol{\Sigma}$.

(*c*) Is $E[tr(\mathbf{S_n})] = tr(\boldsymbol{\Sigma})$? Justify your answer.

(*d*) Is it true that $\boldsymbol{\Sigma}$ is always positive definite matrix? Justify your answer.

10.) Consider the following two multiple linear regression models, with $E(\boldsymbol{\epsilon}) = 0$ and $var(\boldsymbol{\epsilon}) = \sigma^2 I$,

Model A: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$,

Model B: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$.

Show that $R_A^2 \leq R_B^2$.

Applied Statistics Comprehensive Exam

August 2016

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let X be a random variable such that $X \sim Bin(n, p)$. Show that $\mathrm{E}[X] = np$.

$$\text{Hint: } \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j} = (a+b)^n$$

2.) Let X be a random variable such that $X \sim Geo(p)$. Show that the Geometric distribution is memoryless (i.e. show $P(X > j + k | X > j) = P(X > k)$).

3.) Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim Exp(\theta, \eta)$. Find the MLE of θ and η.

4.) Discuss how the sign test may be used to test H_0 : median $= c$ where c is some hypothesized median.

5.) Respond to the following.

i. Construct a 2^{4-1} fractional factorial design with $I = ABCD$ as a defining relation. Show the detailed steps including alias structure and the construction of the design. Are there any better designs than this?

ii. What do you know about fold-over designs? Could you think of a situation where a fold-over design would be appropriate and justify why? Consider a one-sixteenth fraction of a 2^7 factorial experiment. What are the advantages of this particular design as opposed to a completely randomized full factorial design involving the same factors and their levels?

iii. Think of the 1/16th fraction of a full 2^7 factorial design. Construct the fractional design with $I = ABD$, $I = ACE$, $I = BCF$, and $I = ABCG$ as the design generators. Show the complete defining relation and tell us how many words are in there. Show complete alias structure of the main effects ignoring all the three- and higher-order interactions.

iv. Now show us the steps necessary to obtain de-aliased estimates of the main effects for this 1/16th fractional factorial design.

6.) Consider the linear model

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

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

i. Find the orthogonal projection matrix $\boldsymbol{P_X}$.

ii. Let $\boldsymbol{A}$ be an $n \times n$ orthogonal matrix. Show that the matrix $\boldsymbol{A'P_XA}$ is an orthogonal projection.

7.) Suppose a *true* regression model is given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, but in practice it is *underfit* with the model,

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

where the subscripts "U" indicate underfit. Let $\mathbf{X}_O$ and $\boldsymbol{\beta}_O$ indicate the columns and parameters that are omitted by the underfit model.

i. Assuming $\hat{\boldsymbol{\beta}}_U$ is obtained using ordinary least squares, find an expression for $\mathrm{E}[\hat{\boldsymbol{\beta}}_U]$.

ii. Find an expression for the bias of $\hat{\boldsymbol{\beta}}_U$. Under what circumstances would this bias be equal to zero?

8.) Consider the proportional odds cumulative logic multinomial logistic regression model, in which the log-odds is modeled as follows:

$$\ln\left(\frac{\pi_{\leq j}}{1-\pi_{\leq j}}\right) = \beta_{0j} + \sum_{k=1}^{J} \beta_k x_{kj},$$

where $\pi_{\leq j} = \sum_{k=1}^{j} \pi_k$, and π_k indicates the probability associated with ordered category k.

i. Find an expression for each π_k. (Hint: use something like η_j as a placeholder for the right-hand-side of the model equation.)

ii. For a unit increase of predictor x_k, show that the odds ratio is given by $exp(\beta_k)$.

9.) Respond to the following.

i. Explain in non-technical terms the concept of maximum likelihood estimation. Mention one key difference between a likelihood function and a joint density function.

ii. Consider the observation vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ are from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Write down the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and obtain the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. You may use any known results from matrix algebra to simplify the steps.

10.) Assume the model is

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

where $\boldsymbol{X}$ is full rank, $E\boldsymbol{\epsilon} = \boldsymbol{0}, \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, $\boldsymbol{V}$ is a known positive definite matrix. Find the generalized least-squares estimator of $\boldsymbol{\beta}$.

Applied Statistics Comprehensive Exam

January 2016

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, X_2, \ldots, X_{10}$ be a random sample from $X \sim Gam(\theta = 8, \kappa = 6)$. Find the pdf of $Y = 4X_1 + 4X_2 + \ldots 4X_{10}$.

2.) Let $X_1, \ldots, x_n$ be a random sample from $X \sim Pois(\theta)$ and we want to estimate $\tau(\theta) = \theta$.

i. Find an unbiased estimator T of $\tau(\theta)$.

ii. Find the variance of T.

iii. Find the CRLB $\tau(\theta) = \theta$.

iv. Is T the UMVUE of $\tau(\theta) = \theta$? Why or why not?

3.) If X and Y have the joint pdf given by:

$$f(x,y) = \begin{cases} x + y, & 0 < x < 1, 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find

i. $f(x|y)$

ii. E$[3X|y]$

iii. E$[XY]$

4.) Often times in practice, we are focused on the nature of a population distribution function. For example, the validity of a parametric test depends on the shape of the population from which the sample was drawn. When we do not know the functional form of the population, we first want to test whether the population of interest is likely to be distributed according to the assumptions underlying the proposed parametric procedure. Identify three nonparametric procedures that test if a sample has been drawn from a population that is normally distributed. Discuss the advantages and disadvantages of these three tests.

5.) An experimenter is studying the effects of five different formulations of a rocket propellant used in aircrew escape systes on the observed burning rate. Each formulation is mixed from a batch of raw material that is only large enough for five formulations to be tested. Furthermore, the formulations are prepared by several operators, and there may be substantial differences in the skills and experience of the operators. Thus, it would seem that there are two nuisance factors to be "averaged out" in the design.

The data are given in the table below.

Table 1: Data for the rocket propellant problem

Batches of	Operators				
Raw materials	1	2	3	4	5
1	A=24	B=20	C=19	D=24	E=24
2	B=17	C=24	D=30	E=27	A=36
3	C=18	D=38	E=26	A=27	B=21
4	D=26	E=31	A=26	B=23	C=22
5	E=22	A=30	B=20	C=29	D=31

Answer the following questions.

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. The total corrected sum of squares for this experiment, SS_T is 676. Calculate the necessary sum of squares.

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

6.) Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, with $\mathbf{X}$ full rank and $\mathbf{V}$ positive definite.

i. Explain why Ordinary Least Squares (OLS) is not appropriate for this model.

ii. Show that the Generalized Least Squares (GLS) estimator is given by

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}.$$

(HINT: $\mathbf{Z} = \mathbf{Q}^{-1}\mathbf{Y}$.)

7.) Consider the General Linear Mixed Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $\mathrm{Cov}(\boldsymbol{u}) = \mathbf{G} = \sigma_u^2\mathbf{I}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The BLUP for $\boldsymbol{u}$ is given by

$$\hat{\boldsymbol{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbf{V} = \mathrm{Cov}(\mathbf{Y})$ and $\hat{\boldsymbol{\beta}}$ indicates GLS estimators for $\boldsymbol{\beta}$. Prove the "U" in BLUP. That is, prove that

$$\mathrm{E}[\hat{\boldsymbol{u}} - \boldsymbol{u}] = \mathbf{0}.$$

8.) Consider the following presentation of the Negative Binomial pdf, where Y is the number of events before r non-events, and π is the probability of an event for each trial,

$$f(Y; r, \pi) = \begin{pmatrix} Y + r - 1 \\ Y \end{pmatrix} \pi^Y (1 - \pi)^r.$$

i. Write this pdf in the general exponential family form, identifying the canonical parameter θ and the three scalar functions $b(\theta)$, $a(\phi)$, and $c(Y, \phi)$.

$$f(Y; \theta, \phi) = e^{\left(\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right)}$$

ii. Use these functions to derive the expectation $\mathrm{E}[Y]$ and the variance $\mathrm{Var}(Y)$ of a Negative Binomial random variable. (HINT: Use $b(\theta)$.)

9.) Suppose $\boldsymbol{y}$ is $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{pmatrix} -2 \\ 3 \\ -1 \\ 5 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 11 & -8 & 3 & 9 \\ -8 & 9 & -3 & -6 \\ 3 & -3 & 2 & 2 \\ 9 & -6 & 3 & 9 \end{pmatrix}$$

Answer the following questions.

i. Find the distribution of $z = 4y_1 - 2y_2 + y_3 - 3y_4$.

ii. Find the joint distribution of $z_1 = 3y_1 + y_2 - 4y_3 - y_4$, $z_2 = -y_1 - 3y_2 + y_3 - 2y_4$, $z_3 = 2y_1 + 2y_2 + 4y_3 - 5y_4$.

iii. What is the distribution of $(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$?

iv. Is y_1 and y_2 independent? Justify.

10.) For the simple linear regression model, show that the elements of the hat matrix are

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \quad \text{and} \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

Discuss the behavior of these quantities as x_i moves farther from $\bar{x}$.

Applied Statistics Comprehensive Exam

August 2015

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) If $Z \sim \mathcal{N}(0,1)$, then show $Y = Z^2 \sim \chi^2(1)$.

2.) Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim \mathcal{N}(\theta, \sigma^2 = 3)$. Find the UMVUE of $\tau(\theta) = c$ where $P(X < c) = 0.90$. (i.e. find the UMVUE of the 90^{th} percentile.) Note c can be written as a function of θ.

3.) If $X_1, X_2, \ldots, X_n$ is a random sample from $f(x)$ with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$ then show: $\mathrm{E}[\bar{X}] = \mu$ and $\mathrm{E}[s^2] = \sigma^2$.

4.) In general terms, describe the procedure used to perform the Sign Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$ where μ_D = mean of the difference "Pre-Post"). If needed, construct a sample table of data to explain the procedure.

5.) An experiment was conducted comparing different formulations and methods of applying a pesticide to the leaves of cotton plants. The goal was to increase the amount of active pesticide remaining on cotton plant leaves one week after application.

The pesticide being studied degrades in sunlight and a certain additive to the formulation retards this process. Different application techniques may differ in the amount of pesticide delivered to the plant leaves. The treatment factors in this experiment were two different formulations of the pesticide and two different application methods, resulting in a 2^2 factorial experiment.

The experimental unit was a $20'$ row of cotton plants called a plot, because this was a convenient area within which the application of pesticide could be controlled. Eight plots were selected and two were randomly assigned to each of the four treatment combinations, resulting in two replicates per treatment combination.

One week after application, the experimenters were ready to determine the pesticide residue remaining on the plant leaves. However, there was too much plant material in an entire plot to send to the lab for analysis. Therefore, two samples of leaves in an amount convenient for laboratory analysis of the pesticide resides were selected from each plot. Each sample was sent to the lab resulting in the data shown in the Table on the following page.

i) Suggest an analysis plan for this study. In particular, mention and justify what type of analysis would be appropriate for this data (e.g., nested design analysis, split-plot design analysis, ordinary factorial experiment analysis etc.) If you decide to carry out n nested design, you must clearly mention which factor/factors is/are nested under what. If you decide to choose a split-plot design, you must clearly mention why is it a split-plot design and what are your whole-plot and split-plot factors. If you decide to chose ordinary factorial experiment, you must justify the reason for choosing this design.

ii) List all potential factors for determining the pesticide residue and mention whether each of the factors is fixed or random.

Table 1: Pesticide Residue on Cotton Plants

	Application		Sample	
Formulation	Technique	Plot	1	2
A	1	1	0.237	0.252
A	1	2	0.281	0.247
B	1	1	0.247	0.294
B	1	2	0.321	0.267
A	2	1	0.392	0.378
A	2	2	0.381	0.346
B	2	1	0.351	0.362
B	2	2	0.334	0.348

Question 5, Continued

iii) Write down an appropriate statistical model and define all the terms in the model. Your model must conform to the analysis plan that you've proposed in the above.

iv) Crete a dummy ANOVA table to summarize your findings about the factors that affect pesticide residue on cotton plants. Briefly discuss how would you make recommendation to the experimenter.

6.) Respond to each of the following.

i) State and prove the Gauss-Markov Theorem.

ii) Using the theorem from part i, respond to the question, "Why do you tend to use Least Squares Estimators for linear modeling?"

7.) Assume a Normal General Linear Model, such that $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. For the following questions, **do not** assume mean-corrected sums of squares.

i) Show that $SSE/\sigma^2 \sim \chi^2\left(n - rank(\mathbf{X})\right)$. (HINT: Think about $\mathbf{Y}/\sigma$ as a random variable.)

ii) What is the distribution of SSR/σ^2? Derive your answer.

iii) Show that SSR and SSE are distributed independently of each other.

iv) Using parts i, ii, and iii, what can you conclude about the distribution of $\frac{SSR/rank(\mathbf{X})}{SSE/(n-rank(\mathbf{X}))}$?

8.) Consider a contingency table of dimension $I \times J$ under the assumption of Independent Multinomial sampling; that is, each row is distributed as a multinomial, independent of all other rows.

i) Derive a formula for the likelihood ratio test statistic G^2 for the Test of Homogeneity.

ii) Show that the associated degrees of freedom for the Test of Homogeneity is $(I-1)(J-1)$.

9.) The spectral decomposition of a $k \times k$ symmetric matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2' + \cdots + \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k'$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ are the associated normalized eigenvectors.

i) Let $\boldsymbol{X}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| > 0$. Then, by spectral decomposition with $\boldsymbol{A} = \boldsymbol{\Sigma}$, we get

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i',$$

where $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_i = (1/\lambda_i)\boldsymbol{e}_i$. Show that $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ is distributed as χ^2_p, where χ^2_p denotes the chi-square distribution with p degrees of freedom.

ii) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent observations from any population with mean vector $\boldsymbol{\mu}$ and finite covariance matrix $\boldsymbol{\Sigma}$. Then by central limit theorem, for large sample sizes and $n >> p$,

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Demonstrate that

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \text{ is distributed as } \chi^2_p.$$

iii) What is the distribution of

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

where $\boldsymbol{S}$ is the sample covariance matrix.

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

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

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

Applied Statistics Comprehensive Exam

January 2015

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let X_1 and X_2 be a random sample from $X \sim \text{Unif}(0, 1)$. Find the pdf of $Y_1 = X_1 + X_2$ using the pdf-to-pdf technique. Hint: Let $Y_2 = X_2$.

2.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Pois}(q)$ and we want to estimate $\tau(\theta) = \theta$.

i. Find an unbiased estimator T or $\tau(\theta)$.

ii. Find the variance of T.

iii. Find the CRLB of $\tau(\theta) = \theta$.

iv. Is T the UMVUE of $\tau(\theta) = \theta$? Why or why not?

3.) If X and Y have the joint pdf given by:

$$f(x, y) = \begin{cases} x + y, & 0 < x < 1, 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find

i. $f(x|y)$

ii. $\text{E}[3X|y]$

iii. $\text{E}[XY]$

4.) Compare and contrast the Mann-Whitney (aka Wilcoxon Rank Sum) Test with the Median Test. Discuss their similarities/differences.

5.) Consider a situation where a two-stage nested design is applicable. Suppose a company wishes to determine whether the purity of raw material is the same from each supplier. There are four batches of raw material available from each supplier, and three determinations of purity are to be taken from each batch.

i. Explain in words and using a schematic diagram why it is a two-stage nested design. How does the design differ from a block design?

ii. Suppose, there are a levels of factor A, b levels of factor B, and r replicates. Consider that the jth level of factor B is nested under the ith level of factor A. Write the statistical model for the above two-stage nested design. Concisely describe all the terms used in the model.

iii. Assuming both A and B as random factors, sketch the appropriate ANOVA table showing the sources of variation, degrees of freedom, Sum of Squares (SS), Mean Squares (MS), Expected Mean Squares (EMS), and the F_0 statistic under the null hypotheses. You do not need to show the formulas for calculating the sum of squares or mean squares. However, you need to show the expressions for EMS. Always write the null and alternative hypotheses.

iv. Explain how you would use such an ANOVA table to test your hypotheses?

6.) For the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, prove that $\hat{\boldsymbol{\beta}}$ is a minimizer of the quadratic form $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ *if and only if* $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_X\mathbf{Y}$, where $\mathbf{P}_X$ is the projection onto the column space $\mathcal{C}(\mathbf{X})$.

7.) Consider the *balanced* one-factor random effect model,

$$Y_{ij} = \mu + a_i + \epsilon_{ij},$$

where $i = 1, \ldots, I$, $j = 1, \ldots, n$, $a_i \sim \mathcal{N}(0, \sigma_a^2)$, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ with all random effects independent.

i. Derive an expression for the response variance-covariance structure, $\mathbf{V}$.

ii. Show that the GLS estimator for μ is equivalent to the OLS estimator, $\bar{Y}_{\cdot\cdot}$.

(HINT: Use $(\mathbf{I} + \gamma\mathbf{J})^{-1} = (\mathbf{I} - \frac{\gamma}{1+n\gamma}\mathbf{J})$.)

8.) Show that a Generalized Linear Model (GLM) response has variance that depends on the mean. Specifically, consider the log-likelihood corresponding to a response from the exponential family,

$$l(\theta, \phi; Y) = \frac{Y\theta - b(\theta)}{a(\phi)} + c(Y, \phi),$$

where $b()$, $a()$, and $c()$ are scalar functions.

i. Show that $\mathrm{E}[Y] = b'(\theta)$. (HINT: Use $\mathrm{E}\left[\frac{\partial l}{\partial \theta}\right] = 0$.)

ii. Show that $\mathrm{Var}(Y) = b''(\theta)a(\phi)$. (HINT: Use $\mathrm{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right]$.)

9.) Consider the $p \times 1$ vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ representing a random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are mutually independent and each with distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

i. Write down the joint density function of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

ii. Explain in plain language suitable for a layperson the concept of likelihood function with reference to the above problem. Are there any differences between a joint density function and likelihood function? Justify your answer.

iii. Derive the maximum likelihood estimates of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ for the problem stated above. You must show your work to receive credit.

10.) Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $Var(\epsilon_i) = \sigma^2 x_i^2$.

i. Suppose that we use the transformation $y' = y/x$, and $x' = 1/x$. Is this a variance stabilizing transformation?

ii. What are the relationship between the parameters in the original and transformed models?

iii. Suppose we use the method of weighted least squares with $w_i = 1/x_i^2$. Is this equivalent to the transformation introduced in part a? Explain why.

Applied Statistics Comprehensive Exam

August 2014

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, \ldots, X_n$ be a random sample from:

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{elsewhere} \end{cases}$$

If $Y_n = X_{1:n}$, does Y_n have a limiting distribution? If so, what is it?

2.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Unif}(0, \theta)$, $\theta > 0$.

i. Find the MME of θ. Is it unbiased?

ii. Find the MLE of θ. Is it unbiased?

3.) Let $X_1, \ldots, X_5$ be a random sample from $X \sim \text{Pois}(\theta)$, $\theta > 0$. Find the UMP size $\alpha = 0.05$ test for $H_0 : \theta = 1$ versus $H_1 : \theta > 1$. Identify the test statistics and the corresponding critical value for this test. Also, find $\pi_\phi(\theta = 3)$.

4.) In general terms, describe the procedure used to perform the Wilcoxons Signed Ranks Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$ where μ_D = mean of the difference "Pre-Post"). If needed, construct a sample table of data to explain the procedure.

5.) Consider a situation where two-stage nested design is applicable. Suppose a company wishes to determine whether the purity of raw material is the same from each supplier. There are four batches of raw material available from each supplier, and three determinations of purity are to be taken from each batch.

i. Explain in words and using a schematic diagram why it is a two-stage nested design. How does the design differ from a block design?

ii. Suppose, there are a levels of factor A, b levels of factor B, and r replicates. Consider that the jth level of factor B is nested under ith level of factor A.

Write the statistical model for the above two-staged nested design. Concisely describe all the terms used in the model.

iii. Assuming both A and B as random factors, sketch the appropriate ANOVA table showing the sources of variation, degrees of freedom, Sum of Squares (SS), Mean Squares (MS), Expected Mean Squares (EMS), and the F_0 statistic under the null hypotheses. You do not need to show the formulas for calculating the sum of squares or mean squares. However, you need to show the expressions for EMS. Always write the null and alternative hypotheses.

iv. Explain how would use such an ANOVA table to test your hypotheses?

6.) Suppose $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, and that $\mathbf{A}$ is a symmetric and idempotent matrix. Prove that:

$$\mathbf{Y}^T\mathbf{A}\mathbf{Y} \sim \chi^2\left(rank(\mathbf{A}), \frac{\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}}{2}\right).$$

(HINT: Define $\mathbf{Z} = \mathbf{V}^T\mathbf{Y}$, where $\mathbf{V}$ is defined using a basis for $\mathcal{C}(\mathbf{A})$.)

7.) Consider the General Linear Mixed Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, where $\text{Cov}(\boldsymbol{u}) = \mathbf{G} = \sigma_u^2\mathbf{I}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The BLUP for $\boldsymbol{u}$ is given by

$$\hat{\boldsymbol{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\mathbf{V} = \text{Cov}(\mathbf{Y})$ and $\hat{\boldsymbol{\beta}}$ indicates GLS estimators for $\boldsymbol{\beta}$. Prove the "U" in BLUP. That is, prove that

$$\text{E}[\hat{\boldsymbol{u}} - \boldsymbol{u}] = \mathbf{0}.$$

8.) For a $2 \times 2 \times 2$ contingency table, show that equal XY conditional odds ratios (homogeneous association) is equivalent to equal YZ conditional odds ratios.

9.) The spectral decomposition of a $k \times k$ symmetric matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2' + \cdots + \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k'$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ are the associated normalized eigenvectors.

i. Let $\boldsymbol{X}$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| > 0$. Then, by spectral decomposition with $\boldsymbol{A} = \boldsymbol{\Sigma}$, we get

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i',$$

where $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_i = (1/\lambda_i)\boldsymbol{e}_i$. Show that $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ is distributed as χ^2_p, where χ^2_p denotes the chi-square distribution with p degrees of freedom.

ii. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent observations from any population with mean vector $\boldsymbol{\mu}$ and finite covariance matrix $\boldsymbol{\Sigma}$. Then by central limit theorem, for large sample sizes and $n >> p$,

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Demonstrate that

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \text{ is distributed as } \chi^2_p.$$

iii. What is the distribution of

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

where $\boldsymbol{S}$ is the sample covariance matrix.

10.) In a multiple linear regression model, show

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1)s_{x_j}^2},$$

where $s_{x_j}^2$ is the variance of x_j and R_j^2 denotes the coefficient of determination from the regression of x_j on the other regressors. Use this result to show that multicolinearity causes inflated variance for parameter estimates.

Applied Statistics Comprehensive Exam

January 2014

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, \ldots, X_n$ be a random sample from:

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{elsewhere} \end{cases}$$

If $Y_n = X_{1:n}$, does Y_n have a limiting distribution? If so, what is it?

2.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Unif}(0, \theta)$, $\theta > 0$.

i. Find the MME of θ. Is it unbiased?

ii. Find the MLE of θ. Is it unbiased?

3.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Pois}(\theta)$, $\theta > 0$. Find the UMVUE of θ.

4.) In general terms, describe the procedure used to perform the Sign Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ vs $H_1 : \mu_d \neq 0$ where $\mu_D =$ mean of the difference "Pre-Post"). If needed, construct a sample table of data to explain the procedure.

5.) Consider a single-factor fixed effect analysis of variance (ANOVA) model given by

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad i = 1, 2, \ldots, a \quad j = 1, 2, \ldots, n$$

Let $y_{i\cdot}$ represent the total of the observations under ith treatment, $\bar{y}_{i\cdot}$ represent the average of the observations under the ith treatment, $y_{\cdot\cdot}$ represent grand total for all the observations, and $N = an$ is the total number of observations.

i. If the total corrected sum of squares is

$$SS_{\text{Total}} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\cdot\cdot})$$

derive the fundamental ANOVA identity

$$SS_{\text{Total}} = SS_{\text{Treatments}} + SS_{\text{Error}}$$

ii. We are interested in testing the hypothesis of no difference in the treatments:

$$H_0 : \tau_i = \tau_2 = \ldots = \tau_a = 0$$
$$H_1 : \tau_i \neq 0 \quad \text{for at least one } i$$

The appropriate test statistic

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_{\text{Error}/(N-a)}} = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}}$$

follows an F-distribution with $a - 1$ and $N - a$ d.f.

Explain why and when does the ratio of mean squares for treatments and mean squares for error follow an F-distribution.

6.) Suppose $\mathbf{P}_V$ projects onto a vector space V of dimension k.

i. Show that the column space $\mathcal{C}(\mathbf{P}_V) = V$.

ii. Show that the rank of $\mathbf{P}_V$ is k.

7.) Consider a multiple linear regression situation with two predictors x_1 and x_2. Let $\mathbf{P}_x$ denote the projection onto the column space of $\mathbf{X} = [\mathbf{1}|\mathbf{x}_1|\mathbf{x}_2]$, and let $\mathbf{P}_{x_i}$ denote the projection onto the column space of $\mathbf{X}_i = [\mathbf{1}|\mathbf{x}_i]$. Show that the mean-corrected projections associated with each predictor are orthogonal if either mean-corrected projection is equivalent to a mean- and predictor-corrected projection:

$$\mathbf{P}_{x_i} - \mathbf{J}/n = \mathbf{P}_x - \mathbf{P}_{x_j} \Rightarrow (\mathbf{P}_{x_i} - \mathbf{J}/n)(\mathbf{P}_{x_j} - \mathbf{J}/n) = \mathbf{0}$$

8.) Consider a contingency table of dimension $I \times J$.

i. Describe what is meant by "multinomial sampling" for such a contingency table.

ii. Under the assumption of multinomial sampling, derive the likelihood ratio statistic for the test of independence, including degrees of freedom.

9.) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Then if $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j$ and $\boldsymbol{S} = \frac{1}{(n-1)}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$, the test statistic for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ is Hotelling's T^2. At α level of significance we reject the null hypothesis in favor of the alternative if observed

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$$

where $F_{p,n-p}$ is a random-variable with an F-distribution with p and $n-p$ degrees of freedom. Show that T^2 is equivalent to the likelihood ratio test for testing H_0 against H_1 because

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$$

where $\Lambda^{2/n}$ is called Wilk's lambda.

10.) Explain what multicollinearity is and identify three ways that we can detect its presence in a data set.

Applied Statistics Comprehensive Exam

August 2013

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) If $Z \sim \mathcal{N}(0,1)$, then show $Y = Z^2 \sim \chi^2(1)$.

2.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \mathcal{N}(\theta, \sigma^2)$. Find the UMVUE for $\theta(1-\theta)$.

3.) If $X_1, \ldots, X_n$ is a random sample from $f(x)$ with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$ then show: $\mathrm{E}[\bar{X}] = \mu$ and $\mathrm{E}[s^2] = \sigma^2$.

4.) In general terms, describe the procedure used to perform the Sign Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ versus $H_1 : \mu_D \neq 0$ where μ_D = mean of the difference "Pre-Post). If needed, construct a sample table of data to explain the procedure.

5.) Consider the use of a blocking factor, such as farms or batches of material, in an experiment comparing several treatments. A blocking variable is frequently treated as a random effects factor in the analysis. Answer the following questions related to this practice.

i. Conceptually, what does it mean to treat blocks effects as random rather than fixed?

ii. What impact will this have on tests of fixed effects and confidence intervals for treatment means?

iii. What is gained by treating blocks as a random factor?

6.) Prove the following, for the General Linear Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$:

i If $\mathbf{c}^T\boldsymbol{\beta}$ is estimable, then $\mathrm{E}[\mathbf{c}^T\hat{\boldsymbol{\beta}}] = \mathbf{c}^T\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}$ is the Least Squares estimator.

ii. If $\mathbf{c}^T\boldsymbol{\beta}$ is estimable, then $\mathrm{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-}\mathbf{c}$, where $\hat{\boldsymbol{\beta}}$ is the Least Squares estimator.

iii. Using i. and ii., state two important properties of estimable functions of linear model parameters.

7.) 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

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

8.) Consider a random vector $\mathbf{Y}$ with conditional Poisson response distribution, $Y_i|\lambda_i \sim \text{Poi}(\lambda_i)$, and a prior for λ_i of unknown distribution but with moments assumed as: mean $= \Lambda$ and variance $= \tau^2\Lambda$, so that $\lambda_i \sim (\Lambda, \tau^2\Lambda)$. Show that, for $\tau^2 > 0$, this corresponds to a marginal overdispersed Poisson distribution for Y_i with a variance multiplier that is greater than 1.

9.) Consider the orthogonal factor model $\mathbf{x} = \mathbf{LF} + \mathbf{e}$.

i. Derive the model-implied covariance structure of $\mathbf{x}$

ii. Verify that $\mathbf{L}$ is not a unique solution by showing that both factor analysis equations above still hold when $\mathbf{L}$ is transformed (i.e. rotated) by any orthogonal $m \times m$ matrix $\mathbf{T}$.

10.) Prove that in the multiple linear regression, R^2 is the square of correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Applied Statistics Comprehensive Exam

January 2013

Ph.D Theory Exam

This comprehensive exam consists of 10 questions pertaining to theoretical 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.) Let $X_1, \ldots, X_n$ be a random sample from:

$$f(x) = \begin{array}{ll} \frac{1}{\theta} & , \quad 0 < x < \theta \\ 0 & , \quad \text{elsewhere} \end{array}$$

If $Y_n = X_{1:n}$, does Y_n have a limiting distribution? If so, what is it?

2.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Unif}(0, \theta)$, $\theta > 0$.

i. Find the MME of θ. Is it unbiased?

ii. Find the MLE of θ. Is it unbiased?

3.) Let $X_1, \ldots, X_n$ be a random sample from $X \sim \text{Pois}(\theta)$, $\theta > 0$. Find the UMVUE of θ.

4.) In general terms, describe the procedure used to perform the Wilcoxons Signed Ranks Test when testing for equal means from two dependent populations (i.e. $H_0 : \mu_D = 0$ vs $H_1 : \mu_D \neq 0$ where μ_D = mean of the difference "Pre-Post"). If needed, construct a sample table of data to explain the procedure.

5.) For a completely crossed three-way model, data leads to the following ANOVA table:

Source	df	SS
A	2	82.7
B	3	218.7
C	2	58.7
AB	6	131.5
AC	4	36.1
BC	6	17.3
Residual	12	35.2
Total	35	580.2

i. If factor A is random and factors B and C are both fixed, write out the expected mean squares corresponding to each effect in the model.

ii. Compute a test statistic for the main effect of factor B. Provide a critical value for this test (0.05 level).

iii. Compute a test statistic for the main effect of factor C. Provide a critical value for this test (0.05 level).

6.) Prove that $rank(\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T) = rank(\mathbf{X})$. HINTS:

i. Show that $(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is a generalized inverse of $\mathbf{X}$, using the result: If $\mathbf{PA}^T\mathbf{A} = \mathbf{QA}^T\mathbf{A}$, then $\mathbf{PA}^T = \mathbf{QA}^T$.

ii. Show that $rank(\mathbf{AA}^-) = rank(\mathbf{A})$, using the result: $rank(\mathbf{AB}) \leq rank(\mathbf{A})$.

7.) Consider a linear regression model with three predictors,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

i. Construct a corresponding reduced model associated with the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$.

ii. Using appropriate matrix notation, present an expression for the sum of squares associated with the difference between the two models, $SS(x_2, x_3|x_1, \text{Mean})$.

iii. Show that $SS(x_2, x_3|x_1, \text{Mean})$ is distributed independently of the full model error sum of squares.

iv. Show that $SS(x_2, x_3|x_1, \text{Mean})$ is *not* distributed independently of the reduced model error sum of squares.

v. Construct an F-statistic to test $H_0 : \beta_2 = \beta_3 = 0$. Include the distribution of this statistic.

8.) Assume we have an $I \times J \times K$ 3-way contingency table with variables X, Y, and Z.

i. Write down the formula for loglinear model (XY, XZ).

ii. Show that for the model in i, Y and Z are conditionally independent given X.

9.) Multivariate normality

i. Provide a definition of multivariate normality

ii. Describe the steps used to construct a χ^2 plot to check for multivariate normality. Provide as much detail as possible. What would you look for in the finished plot? When would this plot be a trustworthy diagnostic tool?

10.) Consider the following equation

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \epsilon_{ij}, \quad i = 1, 2, 3, \ j = 1, \cdots, n,$$

which represents the regression model corresponding to an analysis of variance with three treatments and n observations per treatment. Suppose that the indictor variables x_1 and x_2 are defined as

$$x_1 = \begin{cases} 1 & \text{if observation is from treatment 1} \\ -1 & \text{if observation is from treatment 2} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if observation is from treatment 2} \\ -1 & \text{if observation is from treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

i. Show that the relationship between the parameters in the regression and analysis -of-variance model is

$$\begin{aligned} \beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3}{3} = \bar{\mu} \\ \beta_1 &= \mu_1 - \bar{\mu} \\ \beta_2 &= \mu_2 - \bar{\mu} \end{aligned}$$

ii. Write down the $\mathbf{y}$ and $\mathbf{X}$ matrix.

iii. Develop an appropriate sum of squares for testing the hypothesis $H_0 : \ \tau_1 = \tau_2 = \tau_3 = 0$.

Note that the model for the one-way analysis of variance is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \ j = 1, \cdots, n.$$