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

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