

# Applied Statistics Qualifying Exam

May 24, 2004  
12:00pm-3:30pm

This exam consists of three parts. You must answer two questions from PART A, two questions from PART B, and one question from PART C. Make sure to clearly indicate which problems you are attempting. Begin each problem on a new sheet of paper, and write on only one side of the paper.

## PART A. You must answer two of these three questions.

A1. Consider the following experimental design:

**Objective:** This was a randomized, double-blind, crossover study of 30 children with attention-deficit/hyperactivity disorder (ADHD) that evaluated the time course effects of four doses of Adderall (5, 10, 15, and 20 mg), an inactive control (placebo), and a positive control (clinical dose of methylphenidate).

**Method:** For each treatment condition, a capsule was administered in the morning and assessments were performed in an analog classroom setting every 1.5 hours across the day. Subjective (teacher ratings of deportment and attention) and objective (scores on math tests) measures were obtained for each classroom session, and these measures were used to evaluate time-response and dose-response effects of Adderall.

**Results:** For doses of Adderall greater than 5 mg, significant time course effects were observed. Rapid improvements on teacher ratings and math performance were observed by 1.5 hours after administration, and these effects dissipated by the end of the day. The specific pattern of time course effects depended on dose: the time of peak effects and the duration of action increased with dose of Adderall.

Factor/ $df_n, df_d$	SKAMP Attention	SKAMP Deportment	PERMP: No. Attempted	PERMP: No. Correct	SSE Item Average
Time/5,140	28.7/.001	31.2/.001	23.2/.001	30.6/.001	0.80/NS
Dose/4,112	6.8/.001	31.0/.001	7.7/.001	7.2/.001	1.07/NS
T×D/20,560	2.1/.005	6.4/.001	3.9/.001	4.6/.001	0.96/NS

*Note:* SKAMP = teacher rating scale developed by Swanson, Kotkin, Agler, M-Flynn, Pelham; PERMP = permanent product measure from math questions; SSE = Stimulant Side Effects rating scale; T×D = time by dose interaction; NS = non-significant.

Discuss the ANOVA design used for this experiment. Explain the fault in the design, and give a correct alternative.

---

**A2.** Consider the fixed effects model

$$Y_{ij} \sim \text{independent Normal}(\mu_i, \sigma^2),$$

$j = 1, \dots, n_i$  and  $i = 1, \dots, m$  where  $\mu_i, i = 1, \dots, m$ , and  $\sigma^2$  are unknown constants.

(a) Using the fact that a  $\text{Normal}(\mu, \sigma^2)$  density has the form

$$n(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\},$$

write out the log-likelihood function

$$\ell(\mu_1, \dots, \mu_m, \sigma^2) = \ln f(\mathbf{y}|\mu_1, \dots, \mu_m, \sigma^2)$$

where  $\mathbf{y} = (y_{11}, \dots, y_{1n_1}, \dots, y_{m1}, \dots, y_{mn_m})'$  and  $f$  is the joint density of  $\mathbf{y}$ .

(b) Using your answer for part (a), find the maximum likelihood estimators of  $\mu_1, \dots, \mu_m$ , and  $\sigma^2$ . Justify that this estimator is a maximizer. Also, give the maximum likelihood estimator of  $\mu_1/\sigma$ .

(c) Consider the  $k$ th group  $Y_{k1}, \dots, Y_{kn_k}$ . It can be shown that

$$\frac{\bar{Y}_{k\cdot} - \mu_k}{S/\sqrt{n_k}} \sim t_{N-m}$$

where  $\bar{Y}_{k\cdot}$  is the sample mean of the  $k$ th group,  $S^2$  is an unbiased estimator of  $\sigma^2$ ,  $N = \sum_{i=1}^m n_i$ , and  $t_n$  is the  $t$ -distribution with  $n$  degrees of freedom. Find a 95% confidence interval for  $\mu_k$ , letting  $t_{\alpha, n}$  denote the  $100(1 - \alpha)$ th percentile of the  $t$ -distribution with  $n$  degrees of freedom.

---

**A3.** For testing  $H_0: \mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$ , the likelihood-based  $F$ -statistic has the forms

$$F = \frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0)\|^2/q}{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/(N-p)} = \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/q}{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(N-p)}.$$

Consider the quadratic model

$$\mathbf{E}[Y_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

for  $i = 1, \dots, N$  where  $Y_i$  is a normal random variable with unknown constant variance  $\sigma^2$  and  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$  is unknown. Now suppose we observe the following four data points  $(x_i, y_i)$ :

$$(-1, 5), (0, 1), (1, -1), (2, 0).$$

- (a) Find the maximum likelihood estimator of  $\boldsymbol{\beta}$ .  
 (b) Test the hypothesis that  $H_0 : \beta_1 = \beta_2 = 0$  at a .05 level of significance. The tables below gives the 95th and 97.5th percentile of the  $F$  distribution with  $df1$  and  $df2$  degrees of freedom.

		95% <i>df1</i>			
		1	2	3	4
<i>df2</i>	1	161.447	199.500	215.707	224.583
	2	18.513	19.000	19.164	19.247
	3	10.128	9.552	9.277	9.117
	4	7.709	6.944	6.591	6.388

		97.5% <i>df1</i>			
		1	2	3	4
<i>df2</i>	1	647.789	799.500	864.163	899.583
	2	38.506	39.000	39.165	39.248
	3	17.443	16.044	15.439	15.101
	4	12.218	10.649	9.979	9.605

**PART B.** You must answer two of these three questions.

**B1.** Consider the no-intercept regression model

$$Y = \beta x + \epsilon$$

where  $x$  is a known real-valued input variable,  $\beta$  is an unknown constant, and  $\epsilon$  is a random error term with a mean of zero and an unknown constant variance  $\sigma^2$ .

(a) Suppose we observe  $N$  data points  $(x_i, y_i), i = 1, \dots, n$  and we wish to minimize the penalized residual sum of squares

$$V(\alpha, \beta) = \sum_{i=1}^N (y_i - \alpha - \beta x_i)^2 + \beta^2.$$

What are the values of  $\alpha$  and  $\beta$  which minimize  $V$ ?

(b) Denote the respective solutions to (a) as  $\hat{\alpha}$  and  $\hat{\beta}$ . Find the bias and variance of  $\hat{\beta}$ . What is the bias of  $\hat{\alpha}$ ?

---

**B2.** Suppose we have a real-valued input variable  $X$  that we wish to use for classification. The input vector is either from class 0 or class 1; the output variable is  $G = 0$  or  $G = 1$  in the two respective cases. Now consider the logistic regression model

$$\ln \frac{P(G = 1|X = x)}{P(G = 0|X = x)} = \beta x.$$

Given training data  $(x_1, g_1), \dots, (x_N, g_N)$  where  $g_1, \dots, g_N \in \{0, 1\}$ , the log-likelihood function for the  $N$  observations is given by

$$\ell(\beta) = \sum_{i=1}^N \ln P(G = g_i|X = x_i).$$

(a) Show that

$$\frac{d\ell}{d\beta} = \sum_{i=1}^N x_i \left( g_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right).$$

Now suppose we observe the following four observations:  $(-1, 1), (0, 0), (1, 1), (2, 0)$ .

(b) Evaluate  $\frac{d\ell}{d\beta} \Big|_{\beta=-1}$  and  $\frac{d\ell}{d\beta} \Big|_{\beta=0}$ .

(c) Show that the solution of  $\frac{d\ell}{d\beta} = 0$  is the unique maximizer of  $\ell$ .

(d) Find a value of  $\beta$  which is within 0.1 of the maximum likelihood estimator of  $\beta$ . Justify your answer.

**B3.** There are many different methods to classify data (logistic regression, neural networks, discriminant analysis, etc.). How do you determine the best technique for classifying?

## **PART C.** You must answer one of these two questions.

**C1.** Consider the following experimental design and determine whether it is a reasonable method of clustering. Explain why or why not.

Study Design. A k-means cluster analysis of patients with spinal and radicular pain based on the SF-36 Health Survey scales.

**Objective.** The aim was to determine whether spine patients fall into clusters according to self-reported health status as measured by the SF-36 and to determine if clustering is similar across four common diagnostic categories: herniated disc, spinal stenosis, spondylosis, and chronic pain syndrome. The grouping of patients (mean age of 50 years, 50% male) was accomplished by “k-means” cluster analysis based on each patient’s scores on the eight scales of the SF-36 Health Survey. In order to reduce bias in the selection of clusters, we standardized scores so that those variables with higher variability and higher absolute values were not disproportionately represented in the solutions. Using the technique suggested by the authors of the SF-36, all scores were standardized relative to the general U.S. population and were scaled to have a mean of 50 and standard deviation of 10. [12](#) A score of 50 represents the average score (the “norm”) for the U.S. population. Any score below 30 (i.e., below two standard deviations from normal) can be interpreted as a significant health deficiency.

In order to conduct such an analysis, a prespecified number of clusters are assigned, and the k-means algorithm attempts to place patients into one of the clusters so as to minimize the total variation among the individual profiles within each of the groups. This is accomplished as follows. Patient profiles are sequentially moved from group to group, and the total variation within each group is measured. If total variation is reduced by a move, then the patient stays in its new group. Otherwise, the move is reversed. The process continues until there is no switch of a patient from one group to another that will reduce the overall within-group variation. In the end, the goal of the analysis is satisfied insofar as patients with similar profiles of scores on the SF-36 reside in the same cluster. Because the k-means algorithm will produce some kind of solution for any number of prespecified groups, it is critical to determine the most appropriate number of groups (clusters) for a given data set. This is achieved by running the cluster analysis for different numbers of prespecified groups and observing when the benefit of adding additional groups begins to diminish. In general, each time an additional group is added, the solution will better “fit” the data; the perfect fit occurring when the number of “groups” is equal to the number of patients. However, after the addition of a certain number of groups, the benefits (as indicated by a reduction in the within group variability) will be very small, and a large number of groups also makes it difficult to interpret results. In our case, we found that three groups (clusters) yielded a good “fit” but that the addition of a fourth and fifth group only had a marginal impact on the within-group variability. In addition, the groupings arrived at by the analysis appeared each to have reasonable clinical interpretations—an important goal that is more difficult to

achieve with a large number of groups, some of which may not contain many members (patients).

---

**C2.** What information will clustering provide about the data? How can you determine that the result gives good clusters? Explain the pros and cons of k-means clustering versus hierarchical clustering.