COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 26, 2005, 9:00 A.M.–1:00 P.M.

STATISTICS 175 QUESTIONS

1. **Call center service time:**

   Call centers are modern service networks in which customer service agents (or servers) provide services to customers via telephones. Usually servers are cross trained to handle multiple types of service. It is of managerial and operational interest to understand factors that affect service times (i.e. times needed to finish serving calls).

   The following data are a subset of a call center database collected at a major US northeastern bank. Three servers are randomly selected from the thousand or so servers. Each server can handle two types of calls, and for each service type, two service times are recorded. It is known that the service times in this call center can be very well modelled by a lognormal distribution. The table below reports the Log(Service Time)s instead, which can be assumed to be normally distributed.

   We want to investigate whether **server** and **service type** are significant factors affecting Log(Service Time).

	Server	SERTYPE	Log(ServTime)
1	Avi	OT	4.37
2	Avi	OT	6.02
3	Avi	RE	4.30
4	Avi	RE	4.81
5	Ed	OT	3.93
6	Ed	OT	4.33
7	Ed	RE	2.71
8	Ed	RE	3.04
9	Tom	OT	4.91
10	Tom	OT	5.56
11	Tom	RE	3.74
12	Tom	RE	3.93

   (a) **[5 points]** State an appropriate model for this data with an interpretation for different terms.

   (b) **[10 points]** Derive the corresponding ANOVA table with the degree of freedom, sum of squares, and mean squares.

   (c) **[5 points]** Calculate the ANOVA estimates of the variance components involved in the model.

(d) **[10 points]** The call center manager wants to implement a bonus system to award servers with short Log(Service Time)s. To decide whether such a system is plausible, she needs to know whether **server** is a significant factor for Log(Service Time). As a summer intern student, you are asked to come up with a statistical test for that purpose. You also need to state your conclusion based on the sample provided.

(e) **[20 points]** Using the sample, derive your best predictor for Avi's Log(Service Time) of the two service types, respectively. Justify your result.

2. **Survival in the Donner Party**

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

An anthropologist used the ages and sexes of the adult (over 15 years) survivors and non-survivors of the party, to study the theory that females are better able to withstand harsh conditions than are males. For any given age, were the odds of survival greater for women than for men?

There are 45 adults in the party, and the following table shows 10 of them.

	AGE	FEMALE	SURVIVAL
1	23	0	0
2	40	1	1
3	40	0	1
4	30	0	0
5	28	0	0
6	40	0	0
7	45	1	0
8	62	0	0
9	65	0	0
10	45	1	0

The variables recorded are:

```
AGE         the age of the adult.
FEMALE      a binary indicator of being a female (1) or not (0).
SURVIVAL    a binary indicator of being a survivor (1) or not (0).
```

To determine the relationship between the odds of survival and AGE and FEMALE, a generalized linear model is fitted with the following commands in R.

```
> donner <- read.table("05hs175-donner.TXT", header=T, sep="\t")
> donner$FEMALE <- as.factor(donner$FEMALE)
> donner$SURVIVAL <- as.factor(donner$SURVIVAL)
> attach(donner)
> options(contrasts=c("contr.treatment", "contr.treatment"))
> donner.glm <- glm(SURVIVAL~AGE+FEMALE, data=donner, family=binomial)
```

The output is as follows.

```
> summary(donner.glm)

Call: glm(formula = SURVIVAL ~ AGE + FEMALE, family = binomial,
data = donner)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.63312    1.11018   1.471   0.1413
AGE         -0.07820    0.03728  -2.097   0.0359 *
FEMALE1      1.59729    0.75547   2.114   0.0345 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
AIC: 57.256

Number of Fisher Scoring iterations: 4
```

(a) **[5 points]** Write down the fitted model and interpret the intercept and the coefficient of AGE.

(b) **[5 points]** Report the odds ratio of survival for females over males along with its 95% confidence interval. Interpret the odds ratio.

(c) **[5 points]** Answer the following question:

For any given age, were the odds of survival greater for women than for men?

(d) **[5 points]** Report the odds ratio of survival for a 50-year-old woman over a 20-year-old woman along with its 95% confidence interval. Interpret the odds ratio.

(e) **[5 points]** Sam is a 23-year-old male. Suppose the fitted logistic regression model is plausible. Predict his probability of surviving.

(f) **[10 points]** Write down the estimated equation for the log-odds of survival if the indicator variable for sex were 1 for males and 0 for females. Write down the estimated equation for the log-odds of perishing if the binary response were 1 for a non-survivor and 0 for a survivor.

(g) **[15 points]** In logistic regression problems, log-odds are modelled prospectively as functions of explanatory variables. In some studies, particularly those in which the probabilities of positive responses are very small, independent samples are drawn retrospectively from the population. Prove the following claim:

In a logistic regression model for a retrospective study, the estimated intercept is not an estimate of the prospective study; while estimates of the coefficients of explanatory variables estimate the corresponding coefficients in the prospective model and may be interpreted in the same way. Hence, the same happens for odds ratio.