2002 Honors Examination in Statistics

Department of Mathematics and Statistics
Swarthmore College

Instructions: This examination consists of a total of four questions. Number the questions clearly in your work and start each question on a new page. You must make it clear how you arrived at your answer. Answers without any work may lose credit even if they are correct.

This is a closed-book three-hour examination. You may not refer to notes or textbooks.

Question 1:

In many marketing data sets, for instance those involving web purchase behavior, there are n_i individuals in a given household, where we know that at least one member in the household has purchased from a given site. Suppose our goal is to estimate θ_i, the proportion of members in the household, who have purchased from the site. A common model for such a process, is the truncated binomial model where we model the conditional distribution of X, the number of persons in the household who have purchased at the site, given that we know $X \geq 1$. Based on this set-up:

(1a) Show that $P(X=x|X \geq 1) = C(n_i, x)\theta_i^x(1-\theta_i)^{n_i-x}/[1-(1-\theta_i)^{n_i}]$ where $C(n_i, x)$ is n_i choose x, the number of ways in which x objects can be chosen from n_i

(1b) Do you think the truncated binomial model is reasonable for this data, given that the assumptions behind the truncated binomial model are the same as that for ordinary binary random variables? Justify your answer by explicitly commenting on the believability of each of the binomial distribution assumptions.

(1c) Show that X is complete and sufficient for estimating θ_i.

(1d) Show that the expected number of family members that have purchased from the site, given that at least one has, is given by $n_i\theta_i/[1-(1-\theta_i)^{n_i}]$.

Suppose that the number of persons n_i in each household is given by a Poisson random variable with rate λ_i, and rates vary across the population according to a gamma distribution with parameters r and s.

(1e) What is the marginal distribution of n_i, that you would observe in the population? You need not give the actual density, just state what distribution it is.

(1f) Utilizing your answer to (1e) and the results in (1d), if you pick a household at random, what is the expected number of persons who have purchased at the given website?

(1g) Say how you would compute the variance of the number of persons. You need not actually compute the variance, just say how you would do it.

Question 2:

The sale of products by auctions is very common today. In an auction, the seller puts a product up for sale, sets an asking price P, and a reservation price $R < P$, for which he or she will not sell the product below (note that the buyers only know P and not R). Furthermore, assume that this is a "silent bid final auction" in which each consumer gets to put in one and only one bid, and does NOT see the bids of other potential buyers. One common assumption in auctions is that the distribution of consumer's willingness to pay is uniformly distributed between (L, U) where L represents the lower bound and U the upper bound. For simplicity you may assume that $L = 0$ and $U = 1$. Based on this information:

(2a) What is the expected bid of a given consumer, assuming he or she bids?

(2b) What is the distribution of the largest bid of n bidders?

(2c) What is the expected value of the largest bid of n bidders?

(2d) What is the expected amount of additional revenue a seller should expect if one additional bidder enters the auction?

(2e) Assuming only 1 bidder, and a fixed R such that $0<R<1$, what is the probability that a randomly chosen bidder will bid enough to buy the product?

Question 3:

A standard approach used in the design of new products is a technique called "conjoint" analysis. In conjoint analysis, a subject is shown a set of T profiles, where each profile is a product made up of A categorical attributes (e.g. name, shape of box, color of box, etc...). For simplicity, assume that each of the A attributes has exactly 2 levels, call the two levels L and H. Furthermore, assume that when the subject is shown a profile (e.g. LLHH = L on attributes 1 and 2 and H on attributes 3 and 4), he or she provides a continuous rating score on a 1-100 scale where 1 indicates total dislike for the product and 100 a strong positive preference.

(3a) How many different possible product configurations are there?

(3b) Assuming that each attribute is coded as a dummy variable (1=H) and (0=L), show that the conjoint model with Y = rating score and $X_1, \ldots, X_A$ = independent variables = attributes, is equivalent to a dummy variable regression model (assume a Gaussian error term).

(3c) What is the interpretation of the intercept in the model given by (3b)?

(3d) Describe how the conjoint model in (3b) can be used to determine the predicted rating of any of the possible product configurations.

(3e) Prove that $(X'X)^{-1}X'Y$ is an unbiased estimate of the conjoint analysis regression coefficients, β, called partworths, if the model in (3b) is written in its standard form $Y=X'\beta+\varepsilon$.

(3f) What is the variance of the estimated coefficients, $\hat{\beta}$, from (3e)?

A Bayesian version of the conjoint model exists when we write the two-stage model:

$$Y_i=X'_i\beta_i+\varepsilon_i \qquad \text{(STAGE 1)}$$
$$\beta_i \sim MVN(\bar{\beta}, \Sigma) \qquad \text{(STAGE 2)}$$

where $\varepsilon_i \sim MVN(0,\Lambda)$ and MVN(a,b) denotes a multivariate normal distribution with mean vector a and variance covariance matrix b.

(3g) Prove that the individual conjoint model (i.e. STAGE 1 only) and the aggregate conjoint model, $\beta_i = \bar{\beta}$, are special cases of the Bayesian model and are hence nested within this two-stage model.

(3h) What is the marginal distribution of Y_i given prior parameters ($\bar{\beta}$, Σ)?

(3i) What is the mean of the posterior distribution of β_i conditional on Y_i, $\bar{\beta}$, Σ?

Question 4:

In many behavioral experiments with two treatments, there is always a discussion of whether to run a within-subjects design where each subject receives both treatments, or a between-subjects design in which each subject is randomly assigned to a treatment condition, and receives one and only one treatment.

(4a) Write down the two-sample t-test statistic and the paired-sample t-test statistic (with appropriate degrees of freedom) corresponding to the between subjects and within-subjects designs in terms of $X_1,\ldots, X_N$ and $Y_1,\ldots Y_N$, the sample outcomes of the two groups.

(4b) Explain why the within-subjects design is typically preferable, even though the number of degrees of freedom is less.

(4c) Describe how you would generate a permutation distribution to test for the difference in means and the according p-value associated with the test, assuming a between-subjects design.

(4d) Assuming that the number of subjects in each of the two groups will be the same, write down an equation for the necessary sample size N for the between-subjects two sample t-test that will lead to a $100 * (1-\alpha)\%$ confidence interval of width w.

End of Examination