Department of Mathematics and Statistics
UMass - Amherst
Basic Statistics Exam August 2011

Work all problems. Sixty points are needed to pass at the Master's level and seventy-five at the Ph.D. level.

1. **23 points.** Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with probability density function: $f(x|\beta) = \frac{1}{\beta}\exp(-x/\beta), x > 0, \beta > 0$.

   (a) Find the complete sufficient statistic for β.

   Let the prior distribution for the parameter β be the Gamma distribution with the following probability density function (i.e., $\beta \sim \text{Gamma}(\alpha, \theta)$):

   $$f(\beta|\alpha, \theta) = \frac{\beta^{\alpha-1}\exp(-\beta/\theta)}{\Gamma(\alpha)\theta^{\alpha}}, \quad \beta > 0, \alpha > 0, \theta > 0$$

   Note that for $\beta \sim \text{Gamma}(\alpha, \theta)$, the following are known: $\text{E}(\beta) = \alpha\theta$, and $\text{Var}(\beta) = \alpha\theta^2$.

   (b) Find the posterior distribution of β.

   (c) Find the posterior mean of β.

   (d) Describe how to construct a 90% posterior interval for β.

2. **31 points.** Many interesting problems involve comparing proportions between two groups. Consider the situation where the two groups correspond to cases (individuals with a disease) and controls (those without). Independent random samples of n_1 cases and n_2 controls are taken and for each individual a binary variable X is observed with $P(X = 1)$ equal to π_1 for the cases and π_2 for the controls. Let $X_{j1}, \ldots, X_{jn_j}$ denote the data for group j ($j = 1$ for controls, $j = 2$ for cases).

   (a) Assuming the population sizes can be treated as infinity, write down a model for the $X_{j1}, \ldots, X_{jn_j}$ and derive the MLE of π_j, call it $\hat{\pi}_j$.

   (b) Derive the mean and variance of $\hat{\pi}_j$ and an approximate large sample confidence interval for π_j.

   (c) Suppose n_j is small and an exact confidence set for π_j is desired, exact in the sense that if C_j is the confidence set, then $P(\pi_j \in C_j) \geq 1 - \alpha$. To derive such an interval

      i. first set-up how for any π_0 you would test $H_0 : \pi_j = \pi_0$ with a test of size at most α; that is, $P(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$.

      ii. Explain how you could construct a confidence set for π_j from the family of tests in the previous part (it is a family since there is a separate test for each π_0). Note that you won't have a simple formula for the confidence set but just need to explain how it could be constructed. Explain why if the test is of size $\leq \alpha$, for each π_0, then the confidence coefficient for the confidence set is at least $1 - \alpha$.

   (d) Find an estimate of $\Delta = \pi_1 - \pi_2$, say $\hat{\Delta}$, state how you would get an estimated standard error for $\hat{\Delta}$ and set-up an approximate large sample confidence interval for Δ.

   (e) In this situation, one quantity of interest is the odds-ratio, defined to be $\theta = \frac{\pi_1(1-\pi_2)}{(1-\pi_1)\pi_2}$.

      i. Explain how you can get an approximate standard error and confidence interval for $\hat{\theta}$, where $\hat{\theta}$ is obtained by replacing π_j in θ with $\hat{\pi}_j$

      ii. How could you use exact confidence sets for π_1 and π_2, say C_1 and C_2, respectively to construct a confidence set C such that $P(\theta \in C) \geq 1 - \alpha$? Again there won't be an explicit expression, but explain carefully how you would proceed and why C has a confidence coefficient $\geq 1 - \alpha$. HINT: Think Bonferonni.

3. **23 points.** The pdf of an exponential distribution with mean θ is:

$$f(x|\theta) = \frac{1}{\theta}\exp(-x/\theta), x > 0, \theta > 0$$

Let $X_1, \ldots, X_n$ be a random sample from this pdf.

(a) Derive the MLE of θ. It is required to justify that your answer is indeed an MLE.

(b) Give the MLE of θ^2 , with justification (Note that $\theta^2 = \text{Var}(X_i)$.)

(c) For a large n, find approximate 95% confidence intervals for θ and θ^2.

4. **23 points.** You are auditing the accounts of a major financial institution. A database contains records of the institution's every financial transaction in the past year. An unknown fraction of them, p, have errors. Your job is to estimate p. You randomly sample 1000 records from the database and examine each one to determine whether it contains an error. But you are not perfectly accurate; when you examine a record you have an unknown probability q of incorrectly judging whether it contains an error. For $i \in \{1, \ldots, 1000\}$, let X_i be 1 or 0 according to whether you judge the i'th record to contain an error and let $X = \sum X_i$. Answer the following questions; be explicit where you can.

   (a) What is the distribution of X?

   (b) You observe $X = 400$. Which values of (p, q) maximize the likelihood?

   (c) What would a 90% confidence region for (p, q) look like, approximately?

   (d) If you were confident that q is small, say $q < .01$, then what would a confidence region for p look like, approximately?