M4056 Review. **October 8–10, 2010**

Section I. Provide a precise mathematical definition for each of the following basic notions of mathematical statistics, and then provide an example or illustration:

- Bernoulli random variable
- Binomial random variable
- Gamma random variable
- Moment generating function
- Random sample of size n from a population with distribution $f_X(x)$
- Sample *pdf* (sample *pmf*)
- Statistic
- Sampling distribution of a statistic
- Sample mean
- Sample variance
- Normal random variable
- Chi squared random variable
- Parametric family of distributions $f_X(x|\theta)$
- Sufficient statistic
- Likelihood function
- Estimator
- Method of moments estimator
- Maximum likelihood estimator
- Mean squared error (of an estimator)
- Bias (of an estimator)

Section II.

1. Let M be the mean of a sample of size n from a population described by a random variable X with distribution $f_X(x)$. Explain the difference in meaning between:
   - the expected value of X,
   - the sample mean (i.e., M itself),
   - the expected value of M.
2. Let S^2 be the sample variance of a sample of size n from a population described by a random variable X with distribution $f_X(x)$. Explain the difference in meaning between:
   - the variance of X,
   - the sample variance (i.e., S^2 itself),
   - the expected value of S^2,
   - the variance of the sample mean ($\operatorname{Var} M$),
   - the variance of S^2 ($\operatorname{Var} S^2$).
3. Let Z be normal$(0, 1)$ random variable. Compute the *pdf* of Z^2, and find the moment generating of this distribution.
4. Let $(X_1, \ldots, X_n)$ be a sample of size n from a normal(μ, σ^2) population. Prove that the sample mean $\overline{X}$ has a normal$(\mu, \sigma^2/n)$ distribution.
5. The beta$(\theta, 1)$ density is given by $f_X(x|\theta) = \theta x^{\theta-1}$, for $x \in (0, 1)$ and $\theta > 0$.
   (a) Find a sufficient statistic for θ.
   (b) Find the maximum likelihood estimator for θ in terms of the statistic in (a).

Section III.

The test will include one or two longer problems, which will involve the ideas in the following (mainly sufficient statistics and MLEs).

1. (a) Suppose $X_1, \ldots, X_n$ are independent, and each uniformly distributed on $[-\theta, \theta]$. Show that
$$T = \frac{3}{n}(X_1^2 + \cdots + X_n^2)$$
is an unbiased estimator of θ^2. (b) Is $\sqrt{T}$ an unbiased estimator of θ?

   Hints:

   (a) We must show that $\mathrm{E}_\theta(T) = \theta^2$. Now,
$$\mathrm{E}_\theta(T) = \frac{3}{n}(\mathrm{E}_\theta(X_1^2) + \cdots + \mathrm{E}_\theta(X_n^2)),$$
and if $X = X_i$, then $\mathrm{E}_\theta(X^2) = \int_{-\theta}^{\theta} x^2 \frac{1}{2\theta}\, dx = \frac{\theta^2}{3}$.

   (b) The expectation of $\sqrt{T}$ is difficult to compute. For example, if $n = 2$,
$$E_\theta(\sqrt{T}) = \int_{-\theta}^{\theta}\int_{-\theta}^{\theta} \frac{\sqrt{x_1^2 + x_2^2}}{4\theta^2}\, dx_1 dx_2.$$
But we know that for any random variable X, $(\mathrm{E}\, X)^2 \leq \mathrm{E}(X^2)$, with equality only if there is a constant c such that $1 = P(X = c)$. (One way to see this is by noting that the variance of X is positive unless there is a constant c such that $1 = P(X = c)$, and $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2$). Thus, $(\mathrm{E}\sqrt{T})^2 < \mathrm{E}\, T$, and so $\mathrm{E}\sqrt{T} < \sqrt{\mathrm{E}\, T} = \theta$.

2. Suppose X is discrete, with $P(X = 0) = \alpha\beta$, $P(X = 1) = \alpha(1 - \beta)$, $P(X = 2) = (1 - \alpha)\beta$, $P(X = 3) = (1 - \alpha)(1 - \beta)$.

   (a) Find a sufficient statistic for (α, β).

   (b) Find the maximum likelihood estimator for (α, β) in terms of the statistic in (a).

   (c) Given the data
$$3\ 3\ 2\ 0\ 0\ 3\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 2\ 0\ 3\ 3\ 0\ 0\ 0\ 1\ 1\ 0\ 2\ 0\ 1\ 1\ 2\ 0,$$
find the MLE of (α, β).

The next problem is a little fanciful. I was interested in exploring generalizations of the problem of estimating θ as a parameter of the uniform distribution on $(0, \theta)$. Problem 3 is too involved to appear on a 45-minute midterm, but it reviews some important ideas in a novel situation.

Definition. If $\vec{X} := (X_1, \ldots, X_n)$ is a sample, then the *order statistics* of $\vec{X}$, denoted $X_{(i)}$, $i = 1, \ldots, n$ are defined as follows: $X_{(1)}$ is the minimum of the X_i, $X_{(2)}$ is the next smallest, etc., and (finally) $X_{(n)}$ is the largest.

3. Let the family of distributions $f_X(x \,|\, p, \alpha)$, $0 < p < 1$, $0 < \alpha < 1$ be defined as follows:
$$f_X(x \,|\, p, \alpha) = \begin{cases} \frac{p}{\alpha} & \text{if } 0 < x < \alpha; \\ \frac{1-p}{1-\alpha} & \text{if } \alpha < x < 1; \\ 0 & \text{if } x \notin (0, 1). \end{cases}$$

In other words, X is in $(0, \alpha)$ with probability p, and if X is in this interval then it's uniformly distributed here; otherwise, X is in $(\alpha, 1)$ and uniformly distributed on $(\alpha, 1)$. Let $\vec{x}$ be an instance of a sample of size n from this distribution.

(a) Show that the likelihood function $L(p, \alpha \,|\, \vec{x})$ is defined by:

$$L(p, \alpha \,|\, \vec{x}) = \left(\frac{p}{a}\right)^k \left(\frac{1-p}{1-\alpha}\right)^{n-k}, \quad \text{for } a \in (x_{(k)}, x_{(k+1)}).$$

(b) Find the log-likelihood function $l(p, \alpha \,|\, \vec{x})$, and show that as a function of a, it is concave up.

(c) Explain why (b) shows that the maximum likelihood estimate for a must belong to the set $\{0, x_{(1)}, x_{(2)}, \ldots x_{(n)}, 1\}$. (More properly, the MLE of a should be "just to one side" of one of these numbers.)

(d) Might the MLE for a depend on p? Might the MLE for p depend on a?

(e) *The following question does not have an answer that you can compute, and it raises issues that we have yet to address. I ask only in the spirit of curiosity.* What properties should the data exhibit to make it reasonable to try to fit this model to the data?

4. Suppose we modify Problem 7.19 (which you did as homework) by adding a parameter α to the model:

$$Y_i = \alpha + \beta\, x_i + \epsilon_i.$$

(a) Find a three-dimensional sufficient statistic for $(\alpha, \beta, \sigma^2)$.

(b) Write out the equations that you would need to solve to find the MLE of (α, β).