Lecture 1; September 14, 1999

1. Overview of the course (see also the outline on the web page or in the grad brochure):

   (a) Outline

   - Basics: likelihood, sufficiency, ancillarity, exponential models, transformation models
   - Likelihood inference: maximum likelihood estimation, likelihood ratio tests, score function, Fisher information, marginal and conditional likelihoods, profile likelihood
   - Classical theory of testing: Neyman-Pearson lemma, size and power, uniformly most powerful tests, ...
   - Estimating functions, quasi-likelihood, generalized linear models

   (b) Requirements: There will be weekly homework problems and a three hour exam in December. We will take turns writing up the lecture notes for general distribution.

   (c) Office Hours: NR's office is 6019, and she will have office hours Tuesday from 2 to 3, other times by appointment

   (d) Books:

   i. TPE: Theory of Point Estimation, Lehmann and Casella. This is the 2nd edition of TPE by Lehmann. Fall term will mainly use only Chapter 1, which is new in the 2nd edition.

   ii. CH: Theoretical Statistics, Cox and Hinkley. I will mainly cover Chs 1 through 5, with some of 9.

   iii. TSH: Testing Statistical Hypotheses (2nd ed.) Lehmann. More advanced treatment of measure theory, and a great deal on the classical theory of estimation.

   iv. BNC: Inference and Asymptotics, Barndorff-Nielsen and Cox. First four chapters update some material in CH.

   v. P&S: Probability and Statistics, Fraser. May be helpful for background.

   vi. Azz: Statistical Inference, Azzalini. Good discussion of likelihood inference, approximately MSc level.

   vii. CB: Statistical Inference, Casella and Berger. Text for Math Stat I&II. Excellent reference.

2. General discussion of statistical theory and methods.

   Chapter 1.1 and 1.2 of CH is a very well written account of much of what I discussed in class, and see also pages 1–4 of TPE.

3. "X is a random variable with density function $f(x;\theta)$ or $f(x)$".

   This is the starting point for most of the material to be covered. Typically X will be a scalar random variable or a vector random variable of length n, and $f(x;\theta)$ will be a density function with respect to counting measure or Lebesgue measure. The problem is to reason from observed data x back to θ or $f(\cdot)$. In more complicated problems, such as observations taken in continuous time, the definitions of X and its density may not be obvious.

   A measure-theoretic formulation can help to make this much more precise. From this point of view we have a "probability triple" $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{X}$ is the sample space, typically identified with R^n or R, $\mathcal{A}$ is the Borel σ-field, and $\mathcal{P}$ is a probability measure which is absolutely continuous with respect to Lebesgue or counting measure. A very concise but helpful account is given in Chapter 1.2 and 1.3 of TPE. (I found two typos: on p.8, l.-2 "Example 2.1" should be "Example 2.2", p.16, 2nd display should be $E(T) = \int T(x) dP_X(x)$.)

   A simple set of examples of such models might be the following:

   (a) $X_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1, \ldots n$, where z_i are known constants, and $\epsilon_1, \ldots \epsilon_n$ are assumed to be independent, identically distributed, with each following a $N(0,1)$ distribution. Then

   $$f(x;\theta) = \frac{1}{\sqrt{(2\pi)^n}} \exp -\frac{1}{2} \sum (x_i - \alpha - \beta z_i)^2$$

   and $\theta = (\alpha, \beta)$.

   (b) $X_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1, \ldots n$, where z_i are known constants, and $\epsilon_1, \ldots \epsilon_n$ are assumed to be independent, identically distributed, with each following the density $f_0(e)$ or $f_0(e;\nu)$ where the form of f_0 is known (possibly up to an unknown number of parameters).

   (c) $X_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1, \ldots n$, where z_i are known constants, and $\epsilon_1, \ldots \epsilon_n$ are assumed to be independent, identically distributed, with each following an unknown density f that satisfies some smoothness conditions

   (d) $X_i = \alpha + \beta z_i + \epsilon_i, \quad i = 1, \ldots n$, where z_i are known constants, and $\epsilon = (\epsilon_1, \ldots \epsilon_n)$ follows a known (up to parameters) joint distribution $f_0(\epsilon;\nu)$

   (e) $X_i = \mu(z_i) + \epsilon_i, \quad i = 1, \ldots n$, where z_i are known constants, and $\epsilon_1, \ldots \epsilon_n$ are assumed to be independent, identically distributed, with each following an unknown density f that satisfies some smoothness conditions

   In general we would expect to require that the dimension of the data, n, be greater than the dimension of the parameter space, in order to construct inference about the parameter. This is not strictly true, as for example in cases (c) and (e) the dimension of the parameter space is a space of unknown functions, and hence effectively infinite, yet progress can be made via smoothing

techniques. Model (c) is an example of a semiparametric model, and Model (e) a nonparametric model. Most models considered in this course will be parametric models, with a parameter θ taking values in a p-dimensional space Θ, typically (a subset of) R^p.

4. The likelihood function

   Given a random variable X and its density function $f(x;\theta)$, the *likelihood function* is defined to be

$$L(\theta; x) = c(x) f(x; \theta),$$

   i.e. any function proportional to the density function evaluated at the observed value of the random variable. (There is really an equivalence class of likelihood functions, all differing by arbitrary multiplicative constants that may depend on x but may not depend on θ.)

   The likelihood function is invariant under one to one transformations of X not involving θ. Suppose $Z = g(X)$, then

$$f_Z(z;\theta) = f_X\{g^{-1}(z);\theta\}|dg^{-1}(z)/dz|$$

   showing that the likelihood function based on Z is in the same equivalence class as the likelihood function based on X.

   A handout on the likelihood function from STA 2162 was distributed.

   Exercises: The likelihood function for an i.i.d. sample of size n is the product of n quantities. The likelihood function for a sample of independent but not identically distributed random variables is as well. The construction of the likelihood function for a vector of dependent random variables is described in CH, Eg. 2.5. The likelihood function is invariant under one to one transformations of the parameter θ.

   Examples:

   (a) $Y_1, \ldots, Y_n$ i.i.d. Bernoulli (p):

$$L(p; y) = \prod p^{y_i}(1-p)^{1-y_i} = p^{\sum y_i}(1-p)^{n-\sum y_i}$$

   (b) R follows Binomial(n, p):

$$L(p; r) = \binom{n}{r} p^r (1-p)^{n-r}$$

   (c) Y follows a Negative Binomial(r, p):

$$L(p; n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Note that in (a) $\sum y_i$ is a sufficient statistic (determines the likelihood function), and that the likelihood functions in (b) and (c) are in the same equivalence class, showing in particular that the likelihood function is unaffected by the sampling rule. This is true in general, and has generated much discussion in the literature. However if we are interested in the distribution of the random function $L(\theta; X)$ or quantities derived from it, this distribution will depend on the sampling rule.

(d) $Y_1, \ldots Y_n$ i.i.d. $N(\mu, \sigma^2)$:

$$\begin{aligned} L(\mu, \sigma^2; y) &= \left(\frac{1}{\sqrt{(2\pi)}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum(y_i - \mu)/\sigma^2\} \\ &= \sigma^{-n} \exp -(\frac{\sum y_i^2}{2\sigma^2} + \frac{\mu \sum y_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}) \\ &= \sigma^{-n} \exp\{-\frac{\sum(y_i - \bar{y})^2}{2\sigma^2} - \frac{n(\bar{y} - \mu)^2}{2\sigma^2}\} \end{aligned}$$

Note that the likelihood function depends only on $(\sum y_i^2, \sum y_i)$, or equivalently $(\bar{y}, \sum(y_i - \bar{y})^2)$, and in particular if σ^2 is assumed to be known, that the likelihood function for μ depends only on $\bar{y}$.

(e) Nonhomogeneous Poisson process

Consider a Poisson process with rate function $\rho(t)$ observed over the interval $[0, t_0)$. Denoting the process by $\{N(t), 0 \le t \le t_0\}$, by definition

i. $\text{pr}[N(t+h) = N(t) + 1 | \{N(t), t \ge 0\}] = \rho(t)h + o(h)$
ii. $\text{pr}[N(t+h) = N(t) | \{N(t), t \ge 0\}] = 1 - \rho(t)h + o(h)$
iii. $\text{pr}[N(t+h) \ge N(t) + 2 | \{N(t), t \ge 0\}] = o(h)$

Thus if we observe $\{N(t)\}$ to increase by 1 at points $y_1, \ldots y_n$, and to be constant at all other points in the interval $[0, t_0)$ the joint probability of this is

$$\begin{aligned} L\{\rho(\cdot)\} &= \prod_{i=1}^{n} \rho(y_i)h \prod^{*}\{1 - \rho(a_j)h\}\{1 + o(h)\} \\ &= h^n \prod_{i=1}^{n} \rho(y_i) \exp \sum^{*} \log\{1 - \rho(a_j)h\}\{1 + o(h)\} \\ &= h^n \prod_{i=1}^{n} \rho(y_i) \exp \sum^{*}\{-\rho(a_j)h\}\{1 + o(h)\} \\ &\approx \prod_{i=1}^{n} \rho(y_i) \exp\{-\int_0^{t_0} \rho(y)dy\} \end{aligned}$$

where the approximation in the last line comes from taking the limit as $h \to 0$, and we drop the factor h^n without changing the equivalence class.

The properties i.–iii. are the defining properties of the nonhomogeneous Poisson process, and also define the rate function $\rho(\cdot)$. This function would normally be parametrized by a relatively small number of parameters, such as $\rho(t) = \exp(\theta_0 + \theta_1 t)$ or $\rho(t) = \theta$. In the latter case we get

$$L(\theta) = \theta^n \exp(-\theta t_0)$$

which is the same likelihood function as that for a sample of n independent observations from the exponential density

$$f(x;\theta) = \theta \exp(-\theta x)$$

with observed total failure time $t_0 = \sum x_i$.

Exercise: Verify the last remark. Give the expression for the likelihood function when $\rho(t) = \exp(\theta_0 + \theta_1 t)$.