Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2005

1. Introduction

In Statistical Inference we observe the value x of a random variable or vector X taking values in some space $\mathcal{X}$, and try to learn about the probability distribution from which X was drawn. We can always index the set of possible distributions for X by elements θ from a set Θ. Usually X will come from some parametric family (normal distributions, gamma, Poisson, etc.), so that Θ will be a subset of $\mathbb{R}^d$ for some small integer d (often one or two), but the formulation will allow us to consider even nonparametric analysis where Θ might be a very large space, like all continuous unimodal density functions on $\mathbb{R}$.

We will always take X to have a probability density function (**pdf**) $f(x \mid \theta)$ with respect to some arbitrary reference measure $m(dx)$ on $\mathcal{X}$, so that for every suitable set $A \subset \mathcal{X}$ and function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathsf{P}^\theta[X \in A] = \int_A f(x \mid \theta)\, m(dx) \tag{1}$$

$$\mathsf{E}^\theta[g(X)] = \int_{\mathcal{X}} g(x)\, f(x \mid \theta)\, m(dx). \tag{2}$$

Usually for us the space $\mathcal{X}$ where X takes its values will be either some countable set like the integers $\mathbb{Z}$ or integer vectors $\mathbb{Z}^p$, or else some interval or other simple geometric set in Euclidean space $\mathbb{R}^p$. For discrete distributions we can always take $m(dx)$ to be the so-called *counting measure* that assigns $m(A) =$ the number of points in A, so that $f(x \mid \theta)$ is just the probability mass function and the integrals in (1) and (2) are just sums. For absolutely continuous distributions we can let $m(dx) = dx$ be the usual (Lebesgue) volume measure in Euclidean space, so that $f(x \mid \theta)$ is just the usual pdf and the integrals in (1) and (2) are just Riemann integrals.

Probability theory (as taught in Duke's STA 205 or, less formally, in STA 213 or STA 104=MTH 135) tells us how to calculate $\mathsf{P}^\theta[X \in A]$ and $\mathsf{E}^\theta[g(X)]$ for various families of distributions $f(x \mid \theta)$, like normal or binomial or exponential or Poisson distributions, when we know the parameters θ. **Statistical Inference** is concerned with the inverse problem— after observing $X = x \in \mathcal{X}$, try to learn about $\theta \in \Theta$. In probability there is usually only one distribution under study (so there isn't any need for the θ superscript on the expectation symbol E^θ or probability symbol P^θ), but many possible values x for X, but in statistics many θ's will be often be considered.

Inference is the art or science of learning about $\theta \in \Theta$ from observing $x \in \mathcal{X}$. While it is possible to look systematically at all forms of inference as "decision problems," here we will instead focus on two traditional inferential aims: **Estimation**, in which we try to guess the value of θ after observing $X = x$ with some **Statistic** $T(X)$ (a *statistic* is any function on $\mathcal{X}$; an *estimator* is a statistic taking values in Θ, with the intention that $T(X) \approx \theta$), and **Hypothesis Testing**, in which we try to assess the plausibility of a *hypothesis* of the form $H : \theta \in A$ for sets $A \subset \Theta$. We will learn about a variety of desirable properties that estimators might have, and will learn how to construct and assess them; we will see how different statistical traditions assess hypotheses. All of this "inference" is based, in one way or another, on the same function $f(x \mid \theta)$.

1.1. The Likelihood Function

The function $f(x \mid \theta)$ depends on both $x \in \mathcal{X}$ and $\theta \in \Theta$, and so could be thought of as function (the pdf) of $x \in \mathcal{X}$, indexed by $\theta \in \Theta$— but for statisticians the opposite view is more useful: as a function of $\theta \in \Theta$, called the **Likelihood Function**, for fixed $x \in \mathcal{X}$. The observation of $X = x$ lends evidence in favor of values of θ where $f(x \mid \theta)$ is high, and evidence against values of θ where $f(x \mid \theta)$ is low.

It is only the *ratio* of the likelihood function at different points that is meaningful. Since multiplying $f(x \mid \theta)$ by a constant doesn't alter those ratios, the likelihood is only defined up to an arbitrary multiplicative factor that is "constant" in the sense that it doesn't depend on θ (it might depend on x). The reference measure $m(dx)$ was arbitrary, so different choices should lead to the same likelihood function; for any $c(x) > 0$ the density function for X with respect to $c(x)^{-1}m(dx)$ will be $c(x)f(x \mid \theta)$, so again the likelihood is only defined up to a (maybe x-dependent) constant factor. Sometimes we use the notation $L(\theta)$ or $L_x(\theta)$ for the likelihood $f(x \mid \theta)$, or

$\ell(\theta)$ or $\ell(\theta \mid x)$ for its logarithm $\ln f(x \mid \theta)$.

For example, let $X = (X_1, \cdots, X_p)$ be a vector of independent Poisson-distributed random variables $X_j \sim \mathsf{Po}(\lambda)$, all with the same mean parameter $\lambda \geq 0$. In this case we can take $\theta = \lambda \in \Theta = [0, \infty)$ and, for $x \in \mathcal{X} = \mathbb{Z}_+^p$ (p-vectors of nonnegative integers), the likelihood function is

$$f(x \mid \theta) = \frac{\theta^{\sum x_j}\, e^{-p\theta}}{\prod x_j!} \propto \theta^{\sum x_j}\, e^{-p\theta} = \exp\left(p\overline{X}_p \ln(\theta) - p\theta\right).$$

Notice that $f(x \mid \theta)$ depends on $x = (x_1, \cdots, x_p) \in \mathbb{Z}_+^p$ only through the average $\overline{X}_p$ or sum $S_p(X) = \sum_1^p X_j$; these are our first examples of **Sufficient Statistics**, functions of the data that embody all evidence about θ. Of course any monotone function of S_p or $e^{-\overline{X}_p}$ is also sufficient. What inference can be made if we observe a total of $S_{10} = 1$ in $p = 10$ independent observations? The likelihood function may be displayed using the `R` code:

```
eg1 <- function(s=1, p=10) {
    mu <- s/p;
    x  <- seq(0, 4*mu,, 101);
    y  <- dpois(s,p*x);
    plot(x,y,xlab="Theta",ylab="Likelihood",type="l");
}
```

A plot of this function appears in Figure 1.

1.2. Three Paradigms

There are three main inferential paradigms in common use today: **Bayesian**, **Classical** (also called Frequentist or Fisherian), and **Likelihoodist**. In the Bayesian paradigm the uncertain quantity $\theta \in \Theta$ is thought of as a random variable with some probability distribution $\pi(d\theta)$ (the "prior distribution"), and the likelihood function $L_x(\theta) = f(x \mid \theta)$ is thought of as the conditional probability density function for x, given θ; statistical analysis is just probability theory based on the joint distribution $\pi(d\theta) f(x \mid \theta)$ for θ and x, or on the **Posterior Distribution** $\pi(d\theta \mid x) = c\pi(d\theta) f(x \mid \theta)$ (the normalizing constant $c = 1/\int_\Theta \pi(d\theta) f(x \mid \theta)$ is seldom important).

The language and methods of probability theory are not used to describe uncertainty about $\theta \in \Theta$ in the Classical paradigm; instead we compute for each possible value of θ the probability of observing different values $X' = x'$,

and base inference on how probable the observed x and certain others would be under different possible θ's.

The Likelihoodist paradigm is a minimilist school, with little use of probability. The likelihood function is used to measure the relative evidence the data offer for different values of θ, but "evidence" is never defined and no probabilistic predictions or statements are made.

In this course we will be occupied primarily with presenting and comparing Bayesian and Classical methods for statistical inference.

Likelihood	$L(\theta)$	$f(x \mid \theta)$, as a function of θ
Statistic	$S = S(X)$, $T, ...$	A function of the data, X
Sufficient	S, $T, ...$	A statistic $S(x)$ such that $f(x \mid \theta) = \phi(S)$
Parameter	$\theta \in \Theta$	The possible "States of Nature"
Outcome	$x \in \mathcal{X}$	The possible values of X

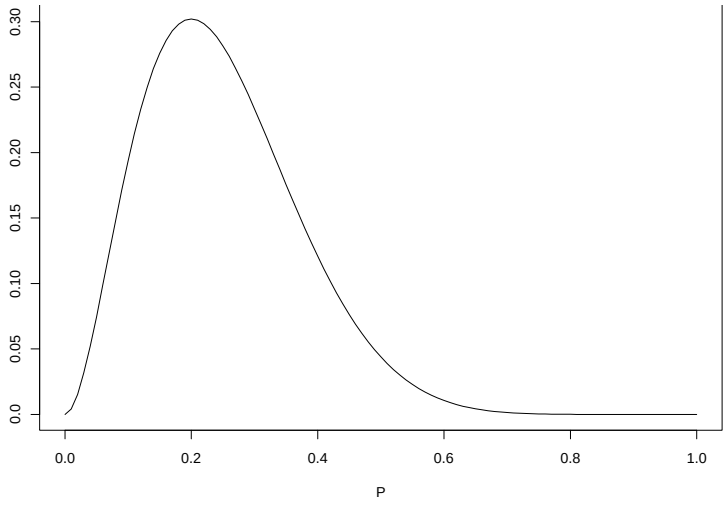

Figure 1. Poisson likelihood for $S_{10} = 1$, $p = 10$.