Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2001

February 13, 2001

1. Introduction

In **Probability Theory** we agree about the probability distribution of some random quantity X, usually with a continuous distribution with known probability density function (**pdf**) $f(x)$ or a discrete distribution with probability mass function (**pmf**) $p(x)$, and we compute probabilities with summations or integrals

$$\mathsf{P}[X \in A] = \int_A f(x)\,dx \qquad \text{or} \qquad \mathsf{P}[X \in A] = \sum_{x \in A} f(x)$$

or (more generally, since $g(x) \equiv I_A(x)$ gives the previous case) expectations

$$\mathsf{E}[g(X)] = \int_{\mathcal{X}} g(x)\,f(x)\,dx \qquad \text{or} \qquad \mathsf{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\,f(x),$$

where in both cases $\mathcal{X}$ denotes the **outcome space** of possible values of the random variable X, $A \subset \mathcal{X}$ is any suitable subset of outcomes of interest, and $g : \mathcal{X} \to \mathbb{R}$ is any function for which the indicated integral or sum makes sense. The common situation in which we have some number $n > 1$ of independent observations that comprise a "random sample" is really just a special case, where $\mathcal{X}$ is all or part of $\mathbb{R}^n$, so our setting already includes the possibility of multiple observations.

In **Statistical Inference** we are faced with the inverse problem: we observe an event $X \in A$ or value $g(X)$ or, to keep things simple, the actual outcome

$X = x \in \mathcal{X}$, and we try to learn about the probability distribution of X — i.e., learn what the pdf or pmf $f(x)$ must be, after observing $X = x$.

This means we must have more than one distribution in mind. One way to keep track of these is to *parameterize* the possible pdf's or pmf's by writing $f^\theta(x)$ or $f(x|\theta)$ for each possible index θ in some set Θ of labels for the distributions under consideration.

In much of what follows we will be concerned primarily or exclusively with the pdf or pmf only at the *observed* value $X = x$ of the data, which (we will argue) embodies all the evidence about θ contained in the observation $X = x$. It turns out (we'll see why later) that any positive multiple of the pdf or pmf reflects exactly the same information — so we introduce the new name and notation of **Likelihood Function**,

$$L(\theta) \propto f(x|\theta),$$

with the observed value of $X = x$ implicit and the proportionality constant completely arbitrary (so two likelihoods that are proportional to one another will be regarded as identical).

2. Example 1

Imagine that a new medical treatment is effective in treating some adverse condition for an uncertain fraction p of the population; we try to learn about p by selecting some sample size N (perhaps ten), finding N "randomly selected individuals from the population" (surely this would be a challenging task!), treating these N individuals, and counting the number X for whom the treatment is effective. Under the usual conditions of independence etc. (about which we should probably be more sceptical), X will have the **binomial distribution** with pmf

$$f(x|N,p) = \binom{N}{x} p^x \, (1-p)^{N-x}.$$

Here the outcome and parameter spaces are $\mathcal{X} = \mathbb{Z}_+$ and $\Theta = \mathbb{Z}_+ \times (0,1)$, if we treat N as uncertain, or $\mathcal{X} = \{0, \cdots, N\}$ and $\Theta = (0,1)$, if (as is more usual) we treat N as known. The **probabilist** will want to compute $\mathsf{E}[g(X)|N,p]$ for various functions $g : \mathcal{X} \to \mathbb{R}$ and known $0 < p < 1$, while the **statistician** will want to make statements about p after observing the value $X = x \in \mathcal{X}$ (the number of successes among the $N = 10$ subjects,

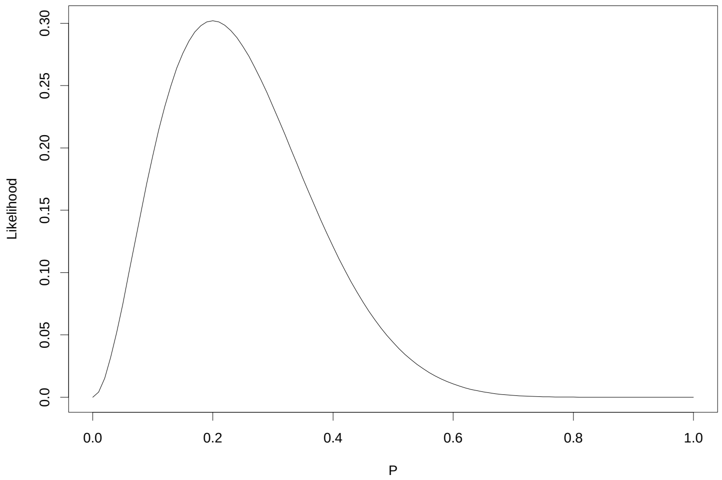

Figure 1. Binomial likelihood for $x = 2$, $n = 10$.

in Example 1). Let us now play the role of a statistician who has observed $x = 2$ successes in $N = 10$ tries. The likelihood function is any or all of

$$\begin{aligned} L(p) &= \binom{10}{2} p^2 (1-p)^8 \\ &= 45 p^2 (1-p)^8 \\ &\propto p^2 (1-p)^8 \end{aligned}$$

The plot was generated by the S-Plus function

```
eg1 <- function(x=2, n=10) {
    p  <- seq(0, 1,, 101);
    y  <- dbinom(x,n,p);
    plot(p,y,xlab="P",ylab="Likelihood",type="l")
}
```

3. Three Questions, Three Paradigms

Historically *Statistical Inference* has been viewed as answering one or more of these questions:

- Find a **Point Estimate** for θ, i.e. a Θ-valued function $S : \mathcal{X} \to \Theta$ which in some sense (we'll be more precise later) is intended to satisfy

"$S(X) \approx \theta$"; in Example 1, we'll want an estimator $S(2) \approx p$ of the success probability for the new treatment.

- Find an **Interval Estimate** for θ, i.e. a set-valued function $U : \mathcal{X} \to \mathfrak{P}(\Theta)$ intended to satisfy $\theta \in U(X)$ "frequently" and also to be "small"; in Example 1, find a subset $U(2) \subset (0,1)$ that is both short and still seems likely to contain p.

- **Test Hypotheses** about θ, i.e. for subsets $H \subset \Theta$ report on the plausibility of the assertion "$\theta \in H$". For the hypothesis $H = [\frac{1}{2}, 1)$, for example, report how plausible is the assertion "$p \geq \frac{1}{2}$" after observing $X = 2$.

Over time three different paradigms have arisen to guide the use of data to answer these questions. We will look at how each of these paradigms would answer the three questions. Two other questions we will occasionally consider are

- **Predict** the value z of some as-yet unobserved variable $Z \sim f_z(z|\theta)$ with some predictor $\hat{Z} : \mathcal{X} \to \mathcal{Z}$ intended to satisfy $\hat{Z}(X) \approx z$.

- Make a **Decision**, choosing an **Action** $a \in \mathcal{A}$ intended to minimize **Loss** $L(a, \theta)$ (or maximize Utility $U(a, \theta)$); this is effected by a **Decision Function** $\delta : \mathcal{X} \to \mathcal{A}$ whose **Risk** $R(\delta, \theta) = \mathsf{E}[L(\delta(X), \theta)]$ is small in some sense.

Note that the last of these subsumes all its predecessors, for suitable choices of action space $\mathcal{A}$ and loss $L(a, \theta)$.

3.1. The Bayesian Paradigm

Historically the first of these is the **Bayesian** paradigm, usually attributed to Thomas Bayes (1763) and Pierre-Simon Laplace (1774). In their original approach the uncertain quantity θ is treated as if it were random, with pdf (proportional to) the Likelihood Function $L(\theta)$; in the present example we may evaluate the integral $\int_0^1 p^2(1-p)^8\, dp = 1/495$, so in the Bayes-Laplace approach we would treat p as uncertain with pdf $495\, p^2(1-p)^8$, $0 < p < 1$, and answer the three questions by something like:

- **Point Estimate**: $S(2) = \mathsf{E}[p \mid X = 2] = \int_0^1 p\, 495 p^2(1-p)^8\, dp = 1/4$; other choices might include the mode $S'(2) = \operatorname{argmax} 495 p^2(1-p)^8 = 2/10$ or median $S''(2) = 0.2358$, with $\mathsf{P}[p \leq S''(x) | X = x] = 1/2$;

- **Interval Estimate**: A "central" 90% interval $U = (u_-, u_+)$, for example, that satisfies $\mathsf{P}[p \in U \mid X = 2] = 0.90$, could be found by locating the points $u_-, u_+ \in (0,1)$ satisfying $0.05 = \int_0^{u_-} 495p^2(1-p)^8\,dp = \int_{u_+}^1 495p^2(1-p)^8\,dp$, or $U = [0.08, 0.47]$; and

- **Hypotheses Test**: $\mathsf{P}\big[p \in [\frac{1}{2}, 1) \mid X = 2\big] = \int_{0.5}^1 495p^2(1-p)^8\,dp = 0.0327$.

Later Bayesians (including Reverend Bayes) generalized this idea somewhat and put it onto stronger foundation grounds by regarding the pdf $f(x \mid \theta)$ as a **conditional** pdf for X **given** θ; once we identify some **marginal** pdf $\pi(\theta)$ for θ, with their product we can form the joint pdf $f(x,\theta) = f(x|\theta)\pi(\theta)$ and, from it, the *other* conditional pdf, here called the **posterior pdf**,

$$\pi(\theta \mid x) = \frac{f(x,\theta)}{f_1(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta')\pi(\theta')\,d\theta'} \propto L(\theta)\pi(\theta).$$

The presentation above would be the special case of a uniform prior distribution with density $\pi(\theta) \equiv 1$ for $\theta \in (0,1)$; we'll consider (and motivate) other possibilities later.

3.1.1. Computational Notes

For any $\alpha > 0$ and $\beta > 0$, $\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. Here $\Gamma(z) \equiv \int_0^\infty t^{z-1}e^{-t}\,dt$ is Euler's **Gamma function**, satisfying $\Gamma(n+1) = n!$ for integers $n \in \mathbb{Z}_+$, but well-defined by the integral given for all $z \in \mathbb{C}$ with $\Re(z) > 0$ and, by analytic continuation, for all $z \in \mathbb{C}$ other than the non-positive integers $\mathbb{Z}_-$. In particular, this lets us evaluate $\int_0^1 p^2(1-p)^8\,dp = \Gamma(3)\Gamma(9)/\Gamma(12) = \frac{2!}{9\cdot10\cdot11} = 1/495$ and $\int_0^1 p\,495p^2(1-p)^8\,dp = 495 * \Gamma(4)\Gamma(9)/\Gamma(13) = 1/4$; we may also recognize $c\,p^2(1-p)^8$ as the pdf for the $\mathsf{Be}(3,9)$ distribution with mean $3/12 = 1/4$. S-Plus lets us evaluate (u_-, u_+) = `qbeta(c(.05,.95), 3, 9)` = $(0.07882005, 0.47008680)$ and $\mathsf{P}[p \geq \frac{1}{2} \mid X = 2]$ = `1-pbeta(0.5, 3, 9)` = `pbeta(0.5, 9, 3)` = 0.03271484.

3.2. The Classical Paradigm

Midway through last century Ronald Fisher, Jersey Neyman, Karl Pearson, and others offered another way of approaching the three questions. Instead of using probability theory to describe uncertainty about a parameter $\theta \in \Theta$, they considered for each fixed θ how likely it would be to observe values X that were equal to the observed x, or that offered even more "extreme" evidence against some hypothesis $H \subset \Theta$ or assertion $\theta \in U$. In this approach the usual estimate for θ is is the **Maximum Likelihood Estimator** or **MLE**

$$\hat{\theta}(x) \equiv \text{argmax}\,[L(\theta)]\,,$$

the value $\hat{\theta} \in \Theta$ where $L(\theta)$ (or, equivalently, $\ell(\theta) \equiv \log L(\theta)$) attains its maximum, if in fact the maximum is attained. In Example 1 this estimate is easily shown to be $\hat{p} = \frac{x}{N} = 2/10$, as is evident from the plot.

In the Classical paradigm the evidence against an hypothesis like $p \in H = [\frac{1}{2}, 1)$ would be quantified by the probability $\mathsf{P}[X \le 2 \mid p] = \texttt{pbinom(2,10,p)}$ for $p \in H$, the probability of observing the actual data $x = 2$ *or other outcomes regarded as more extreme* (here the "more extreme" outcomes would be $X < 2$, since under H one would expect large numbers of success). The maximum over $p \in H$ is attained at the boundary, $\texttt{pbinom(2,10,0.5)} = (1 + 10 + 45)/2^{10} = 56/1024 = 0.0546875$.

A symmetric 90% interval estimate in the Classical paradigm is given by a random interval $U(x) = \big(U_-(x), U_+(x)\big)$ with endpoints that must satisfy $\mathsf{P}[p < U_-(X)] \le 0.05$ and $\mathsf{P}[U_+(X) < p] \le 0.05$, for every $0 < p < 1$; the shortest interval is given by $U =$ `[qbeta(0.05,x,n-x+1), qbeta(0.95,x+1,n-x)` in general, or `qbeta(.05,2,9), qbeta(.95,3,8)` in Example 1, leading to

- **Point Estimate**: $\hat{p}(2) = x/N = 2/10$;
- **Interval Estimate**: Central 90% interval $U = (u_-, u_+) = [0.037, 0.507]$;
- **Hypotheses Test**: $\mathsf{P}\big[X \le 2 \mid p\big] \le \texttt{pbinom(2,10,0.5)} = 0.0547$.

3.3. The Likelihoodist Paradigm

Bayesian statisticians integrate over $\theta \in \Theta$, holding $X = x \in \mathcal{X}$ fixed; Classical statisticians sum over $x \in \mathcal{X}$, holding $\theta \in \Theta$ fixed; Likelihoodists don't integrate or add at all. They also use the MLE as a point estimate, but for a 90% interval estimate would offer $U = [u_-, u_+]$ for the two solutions $u_- < u_+$ to the equation $L(p) = 0.90\,L(\hat{p})$, and for a report on an hypothesis $H \subset \Theta$ would offer $\sup_{\theta \in H} L(\theta)/L(\hat{\theta})$, or here

- **Point Estimate**: $\hat{p}(2) = x/N = 2/10$;

- **Interval Estimate**: 90% interval $U = (u_-, u_+) = [0.146, 0.262]$;
- **Hypotheses Test**: $\sup_{p \geq .5} L(p)/L(\hat{p}) = L(.5)/L(.2) = 0.146$

3.4. Not really a paradigm, but...

Encourged by the Central Limit Theorem, and perhaps uninformed about the subtle points of probability and statistics, naïve investigators often treat all data as if it were normally distributed and use *ad hoc* procedures that would be defensible if the data really did follow a normal distribution; in the present example they would treat the data as Bernoulli indicator variables $Y_i \sim \mathsf{Bi}(1, p)$ with observed mean $\bar{Y}_{10} = X/N = 0.20$, with approximate distribution $\bar{Y}_{10} \approx \mathsf{No}(p, p(1-p)/N)$. Since the Gaussian or normal distribution satisfies $0.90 = \mathsf{P}[\mu - 1.645\sigma \leq Z \leq \mu + 1.645\sigma]$, and since one can hope that the sample mean $\bar{Y} = X/N = 0.20$ and variance $s^2 = X(N-X)/(N-1) = 0.17778$ of the $\{Y_i\}$ will be close to their population mean $\mu = p$ and variance $\sigma^2 = p(1-p)$, this leads to naïve inference of:

- **Point Estimate**: $\hat{p}(2) = x/N = 2/10$;
- **Interval Estimate**: 90% interval $U = \hat{p} \pm 1.645\sqrt{s^2/N} = 0.20 \pm 1.645\sqrt{(16/900)} = [-0.019, 0.419]$;
- **Hypotheses Test**: $\mathsf{P}[\bar{Y}_{10} \leq 0.2 \mid p \geq .5] \approx \mathsf{P}[(Z-\mu)/\sigma \leq (.20 - .50)/\sqrt{16/900} = \Phi(-2.25) = 0.0122$.

For large samples this will give a decent approximation to the Classical (Fisherian/Frequentist) approach; for small samples (as in the present case) the approach is indefensible, but still surprisingly common.

3.5. Example 1b

Now imagine that another new medical treatment is undergoing testing at another laboratory, this time with a different scheme: subjects are given the treatment until $\alpha = 2$ successes are found, and the number Y of failures is counted. Upon observing $Y = 8$ failures before the second success, the likelihood function in this negative binomial sampling scheme is:

$$
\begin{aligned}
f(y|\alpha,p) &= \binom{-\alpha}{y} p^\alpha\,(p-1)^y \\
&\propto p^2\,(1-p)^8
\end{aligned}
$$

exactly the same as in Example 1. Both Bayesian and Likelihoodist analyses, which depend on the data only through the Likelihood Function, will give exactly the same answers as before, but not the Classical approach: here the "more extreme" outcomes than the observed $Y = 2$ under the hypothesis $H : p \geq \frac{1}{2}$ would be $Y > 2$, giving

- **Point Estimate**: $\hat{p}(2) = x/N = 2/10$ (as before);
- **Interval Estimate**: Central 90% interval $U = (u_-, u_+) = [0.04, 0.43]$;
- **Hypotheses Test**: $\mathsf{P}\big[Y \geq 8 \mid p\big] \leq$ `1-pnbinom(7,2,0.5)` $= 0.02$.

Later we will discuss arguments why inference *should* in general depend only on $L(\theta)$, a proposition known as the **Likelihood Principle**; evidently Bayesian and Likelihoodist analyses are consistent with the LP, while Classical analysis is not.

4. Glossary

Hypothesis	$H \subset \Theta$	Any subset of parameters
Likelihood	$L(\theta)$	Any positive multiple $c\,f(x \mid \theta)$ of the pdf at the observed datum $X = x$, as a function of $\theta \in \Theta$
Outcome	$x \in \mathcal{X}$	A possible values of the observation X
Parameter	$\theta \in \Theta$	A possible "State of Nature"
Statistic	$S = S(X)$	Any function $S : \mathcal{X} \to \mathbb{R}$ of the data

5. Computational Notes

In a homework exercise you were asked to show that

$$
p = \int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z}\,dz = \sum_{y=0}^{\alpha-1} \frac{x^y}{y!} e^{-x}
$$

for nonnegative integers $\alpha \in \mathbb{Z}_+$ and positive $x > 0$. While the relation can be verified by differentiation, it is more illuminating to view it through the

Poisson process: the α^{th} event of a unit-rate Poisson process arrives later than time x if and only if fewer than α events have occurred by time x. These identical events have probabilities `1-pgamma(x,a)` and `ppois(a-1,x)`, respectively, since the time of the α^{th} event and the number of events by time x have the $\mathsf{Ga}(\alpha,1)$ and $\mathsf{Po}(x)$ distributions. In S-Plus this relation becomes `p = 1-pgamma(x, a) = ppois(a-1, x)`, leading to the following relations among x, p, and a:

```
p  =  1-pgamma(x,a)  =  ppois(a-1,x)
a  =  1+qpois(p,x)
x  =  qgamma(1-p,a)
```

A similar but subtler relation holds for the binomial and beta distributions:

$$q = \int_p^1 \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x)} t^x (1-t)^{n-x-1}\,dt = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

or `q = 1-pbeta(p,x+1,n-x)=pbinom(x,n,p)`, leading to the following relations among x, p, and q:

```
q  =  pbinom(x,n,p)  =  1-pbeta(p,x+1,n-x)
   =  pnbinom(x,n-x,1-p)
p  =  qbeta(1-q,x+1,n-x)
x  =  qbinom(q,n,p)
```

This relation may also be verified using calculus, or probabilistically by observing that at most x of n independent $\mathsf{Un}(0,1)$ random variables fall below $p \in (0,1)$ if and only if the $x+1^{\text{st}}$ smallest exceeds p; this order statistic has a $\mathsf{Be}(x+1, n-x)$ distribution.

This relation allows us to evaluate the endpoints of interval estimates for the uncertain success probability p after observing x successes in n tries for binomial data— for example, we can find an increasing sequence $u_+(x)$ satisfying $\mathsf{P}[p > u_+(X) \mid p] = 0.05$ by noting that the event $[p > u_+(X)]$ is the same as $[X \le x]$ for the largest x with $p > u_+(x)$, so

$$\begin{aligned}
u_+(x) &= \inf\{p : \mathsf{P}[X \le x|p] > 0.05\} \\
&= \inf\{p : \texttt{pbinom(x,n,p) > 0.05} \\
&= \inf\{p : \texttt{pbeta(p,x+1,n-x) < 0.95} \\
&= \texttt{qbeta(0.95,x+1,n-x)}
\end{aligned}$$

or $u_+ =$ `qbeta(0.95,3,8)` $= 0.5069$ in the example, while similar reasoning gives $u_-(x) =$ `qbeta(0.05,x,n-x+1)` in general or $u_- =$ `qbeta(0.05,2,9)` $=$ 0.03677 in the example.

The binomial coefficient is sometimes defined for integers $0 \le k \le n$ by $\binom{n}{k} \equiv \frac{n!}{k!\,(n-k)!}$, but it is better thought of as the fraction

$$\binom{n}{k} \equiv \frac{(n)_k}{k!}$$

where $(n)_k \equiv n \cdot (n-1) \cdots (n+1-k)$ (the first k terms of $n!$) is well-defined for all real $n \in \mathbb{R}$ and integers $k \in \mathbb{Z}_+$, even if n is non-integral or if $k > n$; for example, this leads to a stronger-than-usual version of the Binomial Theorem,

$$(a+b)^z = \sum_{k=0}^{\infty} \binom{z}{k} a^k b^{z-k}$$

valid for all $z \in \mathbb{R}$ (even negative or non-integral z), and to simple expressions for the Negative Binomial pmf, $X \sim \mathsf{NB}(\alpha, p)$:

$$\begin{aligned} f(x \mid \alpha, p) &= \binom{-\alpha}{x} p^\alpha \, (p-1)^x, \;\; x \in \mathbb{Z}_+ \\ &= \binom{x+\alpha-1}{x} p^\alpha \, (1-p)^x, \;\; x \in \mathbb{Z}_+ \end{aligned}$$

that are valid even for non-integral shape parameter $\alpha > 0$.