Statistical Inference

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

1. Likelihood Principle

The "Likelihood Principle" asserts that for *any* inferential purpose, *all* of the evidence from any observation $X = x^*$ about the parameter θ governing the distribution of $X \sim f(x \mid \theta)$ lies in the Likelihood Function

$$L(\theta) \propto f(x^* \mid \theta).$$

Bayesian and Classical statistics are both concerned with the function $f(x \mid \theta)$, but use it in different ways— in Classical Analysis, measures of estimator precision and of evidence against an hypothesis are based on the probabilities with which X might take on various values "more extreme" than those observed; this violates the LHP by relying on $f(x \mid \theta)$ for values of x other than x^*. Bayesian analysis with a preselected prior distribution $\pi(d\theta)$ is based on the posterior distribution

$$\pi(d\theta \mid x^*) \propto L(\theta)\pi(d\theta)$$

and so *is* consistent with LHP, but Objective Bayesian analysis in which $\pi(d\theta)$ is selected by some formal rule (*e.g.*, Jeffreys' rule), once again uses $f(x \mid \theta)$ for values of $x \neq x^*$ through dependence on the Fisher Information $I(\theta) \equiv -\mathsf{E}[\nabla^2 f(X \mid \theta)]$.

In this section we will see what the LHP means, why it is appealing, and why it is violated by both Classical and Object Bayesian analysis.

1.1. Example 0

A frequently cited example of LHP violation is actually the first glimpse of the Stopping Rule Principle (SRP). Imagine two experimentors considering the question of $H_0 : \theta \le 0.5$ with alternative $H_1 : \theta > 0.5$ on the basis of Bernoulli trials $\xi_j = 1$ w/prob θ, 0 w/prob $1-\theta$. One of them chooses Binomial sampling with a fixed $n = 10$ and observes $X_1 = 7$ successes; the other employs Negative Binomial sampling until 3 failures are observed, which happens to occur after $X_2 = 7$ successes. Their likelihood functions are

$$\begin{aligned} L_1(\theta) &= \tbinom{n}{x}\theta^x(1-\theta)^{n-x} & L_2(\theta) &= \tbinom{\alpha+x-1}{x}\theta^x(1-\theta)^{\alpha} \\ &= \tbinom{10}{7}\theta^7(1-\theta)^3 & &= \tbinom{7+3-1}{7}\theta^7(1-\theta)^3 \\ &= 120\theta^7(1-\theta)^3 & &= 36\theta^7(1-\theta)^3 \end{aligned}$$

so both have the same likelihood; but the p-values against H_0 are

$$\begin{aligned} p_1 &= \Pr[X_1 \ge 7 \mid \theta = 1/2] = \texttt{1-pbinom(6,10,0.5)} = 0.171875 \\ p_2 &= \Pr[X_2 \ge 7 \mid \theta = 1/2] = \texttt{1-pnbinom(6,3,0.5)} = 0.089844 \end{aligned}$$

so H_0 would be rejected at level $\alpha = 0.10$ by the second experimentor but not by the first.

A Bayesian with uniform prior $\theta \sim \mathsf{Be}(1,1)$ would have in each case a posterior $\theta \sim \mathsf{Be}(8,4)$ distribution and so would find in each case $\mathsf{P}[\theta \le 0.5 \mid X = 7] = \texttt{pbeta(0.5,8,4)} = 0.1132813$, but an Objective Bayesian would find different Jeffreys prior distributions

$$\begin{aligned} I_1(\theta) &= \mathsf{E}\,\tfrac{-\partial^2}{\partial\theta^2}[X\ln\theta + (1-X)\ln(1-\theta)] & I_2(\theta) &= \mathsf{E}\,\tfrac{-\partial^2}{\partial\theta^2}[X\ln\theta + \alpha\ln(1-\theta)] \\ &= \mathsf{E}[X/\theta^2 + (1-X)/(1-\theta)^2] & &= \mathsf{E}[X/\theta^2 + \alpha/(1-\theta)^2] \\ &= 1/\theta(1-\theta), \text{ so} & &= \alpha/\theta(1-\theta)^2, \text{ so} \\ \pi_1(\theta) &= \sqrt{I_1(\theta)} & \pi_2(\theta) &= \sqrt{I_2(\theta)} \\ &\propto \theta^{-1/2}(1-\theta)^{-1/2} \sim \mathsf{Be}(0.5,0.5) & &\propto \theta^{-1/2}(1-\theta)^{-1} \sim \mathsf{Be}(0.5,0.0) \end{aligned}$$

and so the two Objective Bayesians find different posterior distributions $\pi_1(\theta \mid X = 7) \sim \mathsf{Be}(7.5,3.5)$ and $\pi_2(\theta \mid X = 7) \sim \mathsf{Be}(7.5,3.0)$ and different posterior probabilities $\mathsf{P}_1[H_0 \mid X = 7] = \texttt{pbeta(0.5,7.5,3.0)} = 0.07026183$ and $\mathsf{P}_2[H_0 \mid X = 7] = \texttt{pbeta(0.5,7.5,3.5)} = 0.1020155$, again disagreeing at level 0.10.

1.2. Example 1

Suppose X_1 and X_2 are independent with $\mathsf{P}[X_j = \theta \pm 1] = 1/2$ for some unknown $\theta \in \mathbb{R}$. The smallest 75% confidence interval for θ is

$$C(X_1, X_2) = \begin{cases} \text{the point } \frac{X_1+X_2}{2} & X_1 \ne X_2 \\ \text{the point } X_1 - 1 & X_1 = X_2 \end{cases}$$

Thus, $0.75 = \Pr[\theta \in C(X_1, X_2)]$ and with repeated use of this confidence interval we would have $\theta \in C(X_1, X_2)$ exactly three-quarters of the time.

BUT, once we observe X_1 and X_2 it seems absurd to report 75% as the confidence level— it should be 100%, if $X_1 \neq X_2$, or 50%, if $X_1 = X_2$. From a post-experimental view it seems silly not to condition on the observed values of X_j.

1.3. Example 2 (Cox, 1958)

A laboratory has two measuring instruments (voltmeters, perhaps); the blue one has an accuracy of ± 0.01, the red one has an accuracy of ± 0.05. The experimentor uses whichever voltmeter is available each day; each is available about half the time.

What accuracy should she report with her data?

1.4. Example 3

We observe a digital signal $X_i \sim \mathsf{No}(\theta, 0.25)$, where $\theta = \pm 1$. To test $H_0 : \theta = -1$ classically we might Reject whenever $X_i > 0$; this gives a test with error probabilities (both Type I and Type II) $\Phi(-2) = 0.02275$. If $X = 0$ is observed we can reject H_0 confidently, with a p-value of 0.0228, well below the conventional 0.05 cut-off. BUT— is this really strong evidence against H_0 and in favor of $H_1 : \theta = +1$??? Is it really fair to reject H_0 in favor of $H_1 : \theta = +1$ at level $\alpha = .0228$ when we observe $X = 0$?

1.5. Example 4

Suppose X is either one, two, or three, with probability distribution p_θ for $\theta = 0$ or $\theta = 1$, where p_θ is given by the following table:

	X		
	1	2	3
$\theta = 0$	0.009	0.001	0.99
$\theta = 1$	0.001	0.989	0.01

A test of $H_0 : \theta = 0$ at level $\alpha = 0.01$ would reject H_0 if $X = 1$ or $X = 2$, and would accept H_0 when $X = 3$; the Type-II error probabilities would be $\beta = .01$, making this a test with fine pre-experimental properties.

The outcome $X = 1$ is troubling, however; the rule above says we Reject H_0 for $X = 1$, with $\alpha = 0.01$, but the likelihood ratio is $9 : 1$ in *favor* of H_0 for this outcome! How can we reject H_0??

1.6. Example 5

Perhaps the most extreme example is due to Jiunn Hwang and George Casella (1982), based on the famous James-Stein estimators. Willard James and Charles Stein (1961) showed that in dimensions $p > 2$ the sample mean $\bar{X}$ is not an admissible estimator of the mean μ for data $X \sim \mathsf{No}(\mu, I_p)$; they found a better (for L^2 risk) estimator of the form

$$\delta^{\mathsf{JS}}(x) \equiv \left[1 - \frac{p-2}{|x|^2}\right] x$$

(here $|x|^2 = \Sigma {x_j}^2$ is the sqared length of our one observation). While having lower risk than x, this is still inadmissible and is downright silly when $\Sigma {x_j}^2$ is smaller than $p - 2$; others (Baranchik in 1964 mentions it) showed that the "positive-part James Stein estimator"

$$\delta^{\mathsf{JS}+}(x) \equiv \left[1 - \frac{p-2}{|x|^2}\right]^+ x$$

is even better (we just truncate to zero when $|x|^2 < p - 2$).

Now for Hwang and Casella: let $\alpha \in (0, 1)$ and find the $1 - \alpha^{\text{th}}$ percentile of the χ^2_p distribution, $\chi^2_p(1 - \alpha) \equiv$ `qchisq(1-alpha,p)`. For sufficiently small $\epsilon > 0$, the confidence set

$$C^{\mathsf{HC}}(x) = \begin{cases} \{\theta : |\theta - \delta^{\mathsf{JS}+}(x)|^2 < \chi^2_p(1-\alpha)\} & \text{if } |x|^2 > \epsilon^2 \\ \emptyset & \text{if } |x|^2 \le \epsilon^2 \end{cases}$$

is never larger than the classical confidence sphere

$$C(x) = \{\theta : |\theta - x|^2 < \chi^2_p(1-\alpha)\},$$

has pre-experimental coverage probability exceeding $1 - \alpha$, and is *empty* if $|x| \le \epsilon$! Imagine reporting an empty confidence set with positive probability.

1.7. Birnbaum

In his 1962 JASA article, Alan Birnbaum proved the astonishing result that the LHP is equivalent to the following two (rather benign-looking) principles:

WCP (Weak Conditionality Principle): Suppose there are two experiments E_1 and E_2 where the only unknown is the parameter θ, common to the two problems. Consider the **mixed** experiment E_* in which we select $i = 1$ or $i = 2$ with equal probabilities, then perform experiment E_i; then the resulting evidence about θ is that from experiment E_i, and we can ignore the existence of the other (unperformed) experiment.

WSP (Weak Sufficiency Principle): Consider an experiment E and a sufficient statistic T. Then if $T(x_1) = T(x_2)$, the evidence about θ from observing x_1 is the same as the evidence about θ from observing x_2.

Most statisticians agree with WCP and WSP but use methods inconsistent with the LHP. Go figure.

1.7.1. Proof

To *prove* Birnbaum's assertion we need to be more formal about exactly what is meant by the terms. We begin by defining an "Experiment" to be a triple $E = (\mathcal{X}, \Theta, f_\theta(x))$ consisting of an outcome space $\mathcal{X}$, a parameter space Θ, and a family of probability density functions $f_\theta(\cdot)$ on $\mathcal{X}$ (with some implicit reference measure— typically counting measure when $\mathcal{X}$ is a discrete set and Lebesgue measure when $\mathcal{X}$ is a subset of Euclidean space), indexed by $\theta \in \Theta$. We do not *define* the "evidence about $\theta \in \Theta$ from observing $x \in \mathcal{X}$ in experiment E", but we introduce notation $\mathsf{Ev}(x, E)$ for this concept and write the three principles above more formally as:

WCP (Weak Conditionality Principle): Suppose there are two experiments $E_1 = (\mathcal{X}_1, \Theta, f_\theta^1)$ and $E_2 = (\mathcal{X}_2, \Theta, f_\theta^2)$ where the only unknown is the parameter $\theta \in \Theta$, common to the two problems. Consider the **mixed** experiment $E_* = (\mathcal{X}_*, \Theta, f_\theta^*)$ given by

$$\begin{aligned} \mathcal{X}_* &\equiv (1, 2) \times (\mathcal{X}_1 \cup \mathcal{X}_2) \\ f_\theta^*\big((i, x_i)\big) &\equiv \frac{1}{2} f_\theta^i(x_i) \end{aligned}$$

in which we select $i = 1$ or $i = 2$ with equal probabilities 1/2, then perform experiment E_i. Then

$$\mathsf{Ev}\big((i, x_i), E_*\big) = \mathsf{Ev}(x_i, E_i)$$

WSP (Weak Sufficiency Principle): Consider an experiment $E = (\mathcal{X}, \Theta, f_\theta)$ and a sufficient statistic T. Then if $T(x_1) = T(x_2)$, $\mathsf{Ev}(x_1, E) = \mathsf{Ev}(x_2, E)$.

LHP (Likelihood Principle): Suppose there are two experiments $E_1 = (\mathcal{X}_1, \Theta, f^1_\theta)$ and $E_2 = (\mathcal{X}_2, \Theta, f^2_\theta)$ where the only unknown is the parameter $\theta \in \Theta$, common to the two problems, and that there are two points $x_1^* \in \mathcal{X}_1$ and $x_2^* \in \mathcal{X}_2$ and a number $c > 0$ for which

$$f^1_\theta(x_1^*) = c f^2_\theta(x_2^*) \qquad \forall \theta \in \Theta$$

Then $\mathsf{Ev}(x_1^*, E_1) = \mathsf{Ev}(x_2^*, E_2)$.

Theorem 1 ((Birnbaum, 1962)) . *WCP + WSP $\Rightarrow$ LHP*

Proof:

Construct E_* as before and define a statistic $T : X^* \to X^*$ by:

$$T\big((j, x_j)\big) \equiv \begin{cases} (1, x_1^*) & \text{if } j = 2 \text{ and } x_j = x_2^* \\ (j, x_j) & \text{otherwise,} \end{cases}$$

so $T(x_1^*) = T(x_2^*)$ but otherwise $T(x^*)$ leaves each x_j fixed. To show that T is *sufficient* we must show that the conditional distribution of x^* given $T(x^*) = t$ does not depend on θ; that follows from direct calculation:

$$\mathsf{P}\big[x^* = (j, x_j) \mid T\big(x^*\big) = t\big] = \begin{cases} \frac{c}{c+1} & \text{if } t = (1, x_1^*) \text{ and } j = 1 \text{ and } x_j = x_1^* \\ \frac{1}{c+1} & \text{if } t = (1, x_1^*) \text{ and } j = 2 \text{ and } x_j = x_2^* \\ 1 & \text{if } t \neq (1, x_1^*) \text{ and } t = (j, x_j) \\ 0 & \text{if } t \neq (1, x_1^*) \text{ and } t \neq (j, x_j) \end{cases}$$

Now Birnbaum's theorem follows by noting that

$$\begin{aligned} \mathsf{Ev}(x_1^*, E_1) &= \mathsf{Ev}\big((1, x_1^*), E_*\big) && \text{by WCP} \\ &= \mathsf{Ev}\big((2, x_2^*), E_*\big) && \text{by WSP} \\ &= \mathsf{Ev}(x_2^*, E_2) && \text{by WCP,} \end{aligned}$$

as claimed.

1.8. Stopping Rules

Probably the most celebrated consequence of LHP is the irrelevence of stopping rules for making inference in sequential procedures. As (Edwards et al. 1963) wrote,

> "The irrelevence of stopping rules to statistical inference restores a simplicity and freedom to experimental design. Many experimentors would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience."

First we illustrate the problem:

Imagine that a client enters your statistical consulting office reporting that she has taken $n = 100$ observations from $X_j \sim \mathsf{No}(\theta, 1)$, and wants to test $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$ at level $\alpha = 0.05$. The classical procedure gives a p-value of $p = 2\Phi(-\sqrt{n}|\bar{x}_n|)$, and rejects H_0 whenever $p \leq \alpha$ or, equivalently, when $\sqrt{n}|\bar{x}_n| \geq z_{\alpha/2}$.

When you learn that her data show $\bar{x}_{100} = 0.20$, the problem seems easy—evidently the p-value is $p = 2\Phi(-2.00) = 0.0455 < \alpha$, leading to rejection. But when by chance you ask "Why did you take $n = 100$ observations?" and learn that the answer is "Because that was enough to get significance," your answer has to change.

If her intention was to reject if $\sqrt{100}|\bar{x}_{100}| \geq k$ for $k = z_{.025} = 1.96$, and otherwise to take another 100 observations and see if that leads to significance, *i.e.*, to $\sqrt{200}|\bar{x}_{200}| \geq k$, then the true probability of a Type-I error is

$$p = \Pr\left[|Z_1| > k \;\; or \;\; |Z_1 + Z_2| > k\sqrt{2}\right]$$

or about 0.0768 for $k = 1.960$, so her test does not have its nominal size $\alpha = 0.05$. To achieve this size she would have to reject when either $\sqrt{100}|\bar{x}_{100}|$ or $\sqrt{200}|\bar{x}_{200}|$ exceeds $k = 2.12$. Since hers do not, we now must change our advice and say she cannot reject H_0!

It is (or should be!) disturbing that the evidential import of her results should depend on her intentions, and not on the data and experiment. Even more alarming, *most* experiments are begun without a clear picture of when to stop taking data, so this "silly example" is in fact the usual situation. Let's describe sequential experiments more precisely.

Let $\mathcal{X}$ be the outcome space for each observation $x_j \sim f_\theta(x)$, and let $\mathcal{X}^j$ be the set of j-tuples $\vec{x}_j = (x_1, ..., x_j)$ in the j-fold Cartesian product. Under the assumption of independence the joint pdf is

$$f_\theta^j(\vec{x}_j) \equiv \prod_{i=1}^{j} f_\theta(x_i).$$

Now we can define for each $m \ge 0$ the "fixed sample-size m experiment" by

$$E_m = \big(\mathcal{X}^m, \Theta, f_\theta^m\big).$$

A randomized *stopping rule* is a sequence of functions $\tau_j : \mathcal{X}^j \to [0,1]$ with the interpretation that we proceed sequentially, deciding at each stage $m \ge 0$ to stop with probability $\tau_m(\vec{x}_m)$ and otherwise to continue to continue to stage $m+1$, taking another observation $x_{m+1} \sim f_\theta(x)$. The rule is *proper* if it stops almost surely, and is *nonrandomized* if $\tau_j \in \{0,1\}$ for each j. For any proper scoring rule we can construct an experiment

$$\begin{aligned}
E_\tau &= \big(\mathcal{X}^\tau, \Theta, f_\theta^\tau\big)\\
\mathcal{X}^\tau &= \mathbb{N} \times \cup_{j=0}^{\infty} \mathcal{X}^j\\
f_\theta^\tau\big((m, \vec{x}_m)\big) &= \prod_{j=0}^{m-1} \big(1 - \tau_j(\vec{x}_j)\big)\tau_m(\vec{x}_m) \prod_{j=1}^{m} f_\theta(x_j)
\end{aligned}$$

One proper stopping rule is

$$\tau_j^1(\vec{x}_j) = \begin{cases} 0 & j < m \\ 1 & j \ge m, \end{cases}$$

leading to the fixed-sample-size experiment $E_{\tau^1} = E_m$; another is our client's,

$$\tau_j^2(\vec{x}_j) = \begin{cases} 0 & j \ne 100 \text{ and } j < 200 \\ 0 & j = 100 \text{ and } \sqrt{j}|\bar{x}_j| < z_{\alpha/2} \\ 1 & j = 100 \text{ and } \sqrt{j}|\bar{x}_j| \ge z_{\alpha/2} \\ 1 & j \ge 200 \end{cases}$$

These stopping rules gave different inference in the classical procedure above; evidently that procedure is not consistent with the:

SRP (Stopping Rule Principle): For any stopping rule $\{\tau_m\}$,

$$\mathsf{Ev}\big((m, \vec{x}_m), E_\tau\big) = \mathsf{Ev}(\vec{x}_m, E_m)$$

But clearly any procedure consistent with LHP automatically obeys the SRP, since the likelihoods in the two cases are

$$\begin{aligned} f_\theta^\tau\big((m, \vec{x}_m)\big) &= \prod_{j=0}^{m-1} \big(1 - \tau_j(\vec{x}_j)\big)\tau_m(\vec{x}_m) \prod_{j=1}^{m} f_\theta(x_j) \\ &\propto \prod_{j=1}^{m} f_\theta(x_j) \\ &= f_\theta^m(\vec{x}_m). \end{aligned}$$

Thus Bayesian analysis with any fixed prior leads to procedures in which the stopping rule is irrelevent– in our client's case, for example, a Bayesian with a uniform prior density would find the same credible interval

$$(1 - \alpha) = \mathsf{P}[\theta \in \bar{x}_m \pm z_{\alpha/2}/\sqrt{m}]$$

for either the sequential or the fixed-sample-size experiment, while the classical procedure and also Bayesian analysis with some "objective" priors would not lead to the same inference for both experiments.

References

Baranchik, A. J. (1964), "Multiple regression and estimation of the mean of a multivariate normal distribution," Technical Report 51, Department of Statistics, Stanford University.

Berger, J. O. and Wolpert, R. L. (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications (with discussion)*, *IMS Lecture Notes-Monograph Series*, volume 6, Hayward, CA: Institute of Mathematical Statistics, second edition.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference (with discussion)," *Journal of the American Statistical Association*, 57, 269–326.

Cox, D. R. (1958), "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–372.

Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian statistical inference for psychological research," *Psychological Review*, 70, 193–242.

Hwang, J. T. and Casella, G. (1982), "Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 10, 868–881.

James, W. and Stein, C. (1961), "Estimation with quadratic loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L. M. Le Cam, J. Neyman, and E. L. Scott, Berkeley, CA: University of California Press, volume 1.