Statistical Inference

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

March 27, 2001

Week 12. Testing and Kullback-Liebler Divergence

1. Likelihood Ratios

Let X_1, X_2, X_2, ... be independent, identically distributed random variables taking values in some space $\mathcal{X}$, whose distribution has a probability density function $f(x \mid \theta)$. If we believe that one of the two hypotheses

$$H_0 : [\theta = \theta_0] \qquad\qquad H_1 : [\theta = \theta_1]$$

is true but we are unsure which, we can compute the *likelihood ratio* against the null hypothesis H_0,

$$\lambda_n = \frac{f_1(x_1) f_1(x_2) \cdots f_1(x_n)}{f_0(x_1) f_0(x_2) \cdots f_0(x_n)}$$

where we have simplified the notation by writing $f_0(x)$ for $f(x \mid \theta_0)$ and $f_1(x)$ for $f(x \mid \theta_1)$.

A Likelihoodist Statistician would find the likelihood ratio λ_n to be the best direct measure of the relative support of the data for these two hypotheses; a Bayesian statistician with prior probabilities $\pi_0 = \mathsf{P}[H_0]$ and $\pi_1 = \mathsf{P}[H_1] = (1 - \pi_0)$ would find the best measure to be the posterior probabilities

$$
\begin{aligned}
\mathsf{P}[H_0 \mid X_n] &= \frac{\pi_0 f_0(x_1) f_0(x_2)\cdots f_0(x_n)}{\pi_0 f_0(x_1) f_0(x_2)\cdots f_0(x_n) + \pi_1 f_1(x_1) f_1(x_2)\cdots f_1(x_n)} \\
&= \frac{\pi_0}{\pi_0 + \pi_1 \lambda_n}
\end{aligned}
$$

or, equivalently, the posterior odds

$$
\frac{\mathsf{P}[H_1 \mid X_n]}{\mathsf{P}[H_0 \mid \mathbf{x_n}]} = \frac{\pi_1}{\pi_0}\,\lambda_n
$$

Finally, the Classical (or Frequentist) statistician would rely on the Neymann-Pearson Lemma, which asserts that the most powerful test of the hypothesis H_0 at level $0 < \alpha < 1$ upon observing $\lambda_n = \lambda^*$ is to Reject H_0 in favor of H_1 if $\lambda^* \geq c_\alpha$, where

$$
c_\alpha \equiv \inf\{c < \infty :\ \mathsf{P}[\lambda_n \geq c \mid \theta_0] \leq \alpha\}
$$

and otherwise fail to reject, or equivalently to reject if and only if the "P-value"

$$
P_n \equiv \mathsf{P}[\lambda_n \geq \lambda^* \mid \theta_0]
$$

satisfies $P_n \leq \alpha$. This is "most powerful" in the sense that the "power" $\mathsf{P}[\text{ Reject } H_0 \ \mid H_1] = \mathsf{P}[\lambda_n \geq c_\alpha \mid \theta_1]$ is larger than for any other test of "size" $\mathsf{P}[\text{ Reject } H_0 \ \mid H_0] = \mathsf{P}[\lambda_n \geq c_\alpha \mid \theta_0] \leq \alpha$.

Let's explore how λ_n behaves as $n \to \infty$, and from that see what inference Likelihoodists, Bayesians, and Classical statisticians are likely to make.

1.1. Log Likelihood Ratio as a Random Walk

For any $\theta \in \Theta$ the log likelihood ratio is a sum of independent random variables

$$
\log \lambda_n = \sum_{j=1}^{n} \log\left[f_1(X_j)/f_0(X_j)\right],
$$

the j^{th} of which has mean

$$\begin{aligned}
\mu &= \mathsf{E}[\log \frac{f_1(X)}{f_0(X)} \mid \theta] \\
&= \int_{\mathcal{X}} [\log f_1(x) - \log f_0(x)]\, f(x)\, dx \\
&= \int_{\mathcal{X}} [\log \frac{f_1(x)}{f(x)} - \log \frac{f_0(x)}{f(x)}]\, f(x)\, dx \\
&= K(f, f_0) - K(f, f_1)
\end{aligned}$$

where $K(f, g)$ is the "Kullback-Liebler Divergence" defined by

$$K(f, g) \equiv \int_{\mathcal{X}} -\log \frac{g(x)}{f(x)}\, f(x)\, dx.$$

This "divergence" satisfies $K(f, f) = 0$ for all $f(x)$ and $K(f, g) > 0$ for all other $g(x)$, but is not symmetric and so is not (quite!) a distance metric. It has many interesting properties, some described in the course text by *Bickel & Doksom* (§2.2, 3.2) and others we will encounter below. By the Law of Large Numbers, $(1/n) \log \lambda_n \to \mu$ as $n \to \infty$ and hence $\lambda_n \approx e^{n\mu}$, so

$$\lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} e^{n\mu} = \begin{cases} 0 & \text{if } K(f, f_0) < K(f, f_1), \\ \infty & \text{if } K(f, f_0) > K(f, f_1) \end{cases}$$

and, in particular, statisticians of all three paradigms will be lead to the right conclusion in the limit as $n \to \infty$ if either H_0 is true, in which case $\mu = -K(f_0, f_1) < 0$ and $\lambda_n \to 0$, or H_1 is true, so $\mu = K(f_1, f_0) > 0$ and $\lambda_n \to \infty$.

Similarly we can compute

$$\sigma^2 = \mathsf{E}[\big(\log \frac{f_1(X)}{f_0(X)} - \mu\big)^2 \mid \theta] = \int_{\mathcal{X}} [\log f_1(x) - \log f_0(x) - \mu]^2\, f(x)\, dx$$

and, by the Central Limit Theorem,

$$\lim_{n\to\infty} \mathsf{P}\left[\frac{\log \lambda_n - n\mu}{\sigma\sqrt{n}} \geq z\right] \quad = \quad \Phi(-z)$$

and hence the Likelihoodist, Bayesian, and Frequentist reports for large n and a true hypothesis $H_0 : [\theta = \theta_0]$ will be approximately

$$\begin{aligned}
\lambda_n &\approx e^{-n\,K(f_0, f_1)} \to 0 \\
\mathsf{P}[H_0 \mid x] &\approx \frac{\pi_0}{\pi_0 + \pi_1 e^{-n\,K(f_0, f_1)}} \to 1 \\
P_n &\approx \Phi\left(-\frac{\log \lambda_n + nK(f_0, f_1)}{\sigma\sqrt{n}}\right) \sim \mathsf{Un}(0, 1)
\end{aligned}$$

while if $H_1 : [\theta = \theta_1]$ is true they will be approximately

$$\begin{aligned} \lambda_n &\approx e^{n\,K(f_1,f_0)} \to \infty \\ \mathsf{P}[H_0 \mid x] &\approx \frac{\pi_0}{\pi_0 + \pi_1 e^{n\,K(f_1,f_0)}} \to 0 \\ P_n &\approx \Phi\left(\frac{-nK(f_1,f_0) - nK(f_0,f_1)}{\sigma\sqrt{n}}\right) \\ &= \Phi\big(-[K(f_1,f_0) + K(f_0,f_1)]\sqrt{n/\sigma^2}\big) \to 0 \end{aligned}$$

1.2. Martingales and Interated Logarithms

Suppose that $H_0 : [\theta = \theta_0]$ is true. At any fixed sample-size n, we have seen how Classical and Bayesian testing methods behave. What happens as we observe successively larger samples?

The Law of the Iterated Logarithm for random walks states that, with probability one, a random walk like $\log \lambda_n$ with *i.i.d.* steps of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ will satisfy, with probability one,

$$\begin{aligned} \limsup_{n\to\infty} \frac{\log \lambda_n - n\mu}{\sigma\sqrt{2n \log\log n}} &= +1 \\ \liminf_{n\to\infty} \frac{\log \lambda_n - n\mu}{\sigma\sqrt{2n \log\log n}} &= -1 \end{aligned}$$

and, in particular, that if H_0 is true (so that $\mu = -K(f_0, f_1) < 0$), then with probability one there will be infinitely many numbers $n \in \mathbb{N}$ for which $\log \lambda_n + n\,K(f_0, f_1) > \sigma\sqrt{n \log\log n}$, so that

$$\begin{aligned} P_n &\approx \Phi\left(-\frac{\log \lambda_n + nK(f_0,f_1)}{\sigma\sqrt{n}}\right) \\ &< \Phi(-\sqrt{\log\log n}) \to 0. \end{aligned}$$

Thus the P-value will fall arbitrarily close to zero, as $n \to \infty$, lending arbitrarily strong evidence against even a true hypothesis.

If H_0 is true then λ_n is a positive martingale starting at $\lambda_0 = 1$, since

$$\mathsf{E}[\lambda_{n+1} \mid X_1, ..., X_n] = \lambda_n \int_{\mathcal{X}} \frac{f_1(x)}{f_0(x)}\, f_0(x)\, dx = \lambda_n \int_{\mathcal{X}} f_1(x)\, dx = \lambda_n$$

and so, by Doob's Martingale Maximal Inequality, for any $b < \infty$ and $N > 0$,

$$\mathsf{P}[\sup_{n \le N} \lambda_n > b] \le \frac{\mathsf{E}[\lambda_0]}{b} = \frac{1}{b}$$

and hence if H_0 is true the probability that the Bayesian posterior probability of H_0 *ever* falls below any $0 < p < 1$ is bounded above by

$$\begin{aligned} \mathsf{P}\big[\inf_{n \le N} \mathsf{P}[H_0 \mid \mathbf{x_n}] \le p\big] &= \mathsf{P}\big[\inf_{n \le N} \frac{\pi_0}{\pi_0 + \pi_1 \lambda_n} \le p\big] \\ &= \mathsf{P}\big[\sup_{n \le N} \lambda_n \ge \frac{\pi_0(1-p)}{\pi_1\, p}\big] \le \frac{\pi_1}{\pi_0} \frac{p}{1-p} \end{aligned}$$

as $n \to \infty$ so, unlike Classical methods, Bayesian methods will not give arbitrarily strong evidence against a true hypothesis.

1.3. Martingales and Sample Size

If H_0 is true we have seen that λ_n is a positive martingale starting at $\lambda_0 = 1$. For any $0 < a < 1 < b$ let τ_{ab} be the stopping time

$$\tau_{ab} \equiv \inf\{n \ge 0 : \lambda_n \notin (a, b)\}$$

(the first time λ_n either exceeds $b > 1$ or falls below $a < 1$) and denote by p_{ab} the probability that λ_n exceeds b before falling below a, $p_{ab} = \mathsf{P}[\lambda_{\tau_{ab}} \ge b]$. Then Doob's theorem applied to the two martingales λ_n and $\log \lambda_n - n\mu$ give us

$$\begin{aligned} \mathsf{E}[\lambda_{\tau_{ab}} \mid H_0] &\approx (1 - p_{ab})a + (p_{ab})b \\ &= 1 \qquad \Longrightarrow \\ p_{ab} &= \frac{1-a}{b-a} \\ \mathsf{E}[\log \lambda_{\tau_{ab}} - \mu\tau_{ab} \mid H_0] &\approx (1 - p_{ab}) \log a + (p_{ab}) \log b - \mu \mathsf{E}[\tau_{ab} \mid H_0] \\ &= \frac{(b-1)\log a + (1-a)\log b}{b-a} - \mu \mathsf{E}[\tau_{ab} \mid H_0] \\ &= 0 \qquad \Longrightarrow \\ \mathsf{E}[\tau_{ab} \mid H_0] &= -\frac{(b-1)\log a + (1-a)\log b}{(b-a)K(f_0, f_1)} \end{aligned}$$

and, in the limit as $b \to \infty$,

$$\mathsf{E}[\inf\{n \ge 0 : \lambda_n < a\} \mid H_0] = \frac{\log(1/a)}{K(f_0, f_1)}$$

so the sample-size needed to reach a posterior probability of $\mathsf{P}[H_0 \mid \vec{x}_n] = \frac{\pi_0}{\pi_0+\pi_1\lambda_n} > 1-\epsilon$ for a true hypothesis $H_0 : \{x_j \sim f_0(x)\}$ has expectation

$$n \ge \frac{\log\frac{\pi_1}{\pi_0} + \log\frac{1-\epsilon}{\epsilon}}{K(f_0, f_1)}$$

and, similarly, if H_0 is false and $x_j \sim f_1(x)$, then

$$\mathsf{E}[\inf\{n \ge 0 : \lambda_n > b\} \mid H_1] = \frac{\log b}{K(f_1, f_0)}$$

and the sample-size needed to achieve a posterior probability of $\mathsf{P}[H_0 \mid \vec{x}_n] = \frac{\pi_0}{\pi_0+\pi_1\lambda_n} < \epsilon$ for a false hypothesis is

$$n \ge \frac{\log\frac{\pi_0}{\pi_1} + \log\frac{1-\epsilon}{\epsilon}}{K(f_1, f_0)}.$$

Evidently the sample-size needed varies directly with the logistic of the desired posterior probability, and inversely as the Kullback-Liebler discrepancy between f_0 and f_1.

2. Kullback-Liebler and Fisher Information

For small $\epsilon > 0$, a second-order Taylor-series approximation of the KL divergence from $f(x \mid \theta)$ to $f(x \mid \theta+\epsilon)$ gives

$$\begin{aligned} K\big(f(x \mid \theta), f(x \mid \theta+\epsilon)\big) &= \int_{\mathcal{X}} -\log\frac{f(x \mid \theta+\epsilon)}{f(x \mid \theta)}\, f(x \mid \theta)\, dx \\ &\approx \int_{\mathcal{X}} \left[-\epsilon\nabla\log f(x \mid \theta) - \frac{\epsilon^2}{2}\nabla^2\log f(x \mid \theta)\right] f(x \mid \theta)\, dx \\ &= I(\theta)\epsilon^2/2 \end{aligned}$$

or, in $q > 1$ dimensions, $\epsilon^{\mathsf{T}} I(\theta)\epsilon/2$. This suggests a close link between KL divergence and the "Information Metric" notion of the distance between different distributions,

$$\begin{aligned} d_I(\theta_0, \theta_1) &= \int_{\theta_0}^{\theta_1} \sqrt{I(\theta)}\, d\theta \qquad \text{(in } q=1 \text{ dimension)} \\ &= \inf_{\gamma} \int_0^1 \sqrt{\gamma'(t)^{\mathsf{T}} I(\theta)\gamma'(t)}\, dt \end{aligned}$$

where the infimum is over all differentiable paths γ from $\gamma(0) = \theta_0$ to $\gamma(1) = \theta_1$; evidently, for near-by θ_0, θ_1,

$$K\big(f(x \mid \theta_0), f(x \mid \theta_1)\big) \approx d_I(\theta_0, \theta_1)^2/2.$$