Nonparametric Bayesian Spatial Statistics

By **Robert L. Wolpert**
Duke University, USA

1 Invitation to Bayesian Analysis

2 Example: Rare Poisson Events

3 Hierarchical and Spatial Models

4 Continuous Models: Remote Sensing

5 Moving Average Spatial Models

6 Nonparametric Spatial Epidemiology

1. Invitation to Bayesian Analysis

- What *is* Bayesian analysis?
- Hypothesis Tests: P-Values Are Misleading
- Sequential Analysis: Stopping Rule Principle
- ◦ Estimation & Testing: Admissibility
- ◦ Model Elaboration and Synthesis (Hierarchical Models)

Setup:

Outcome Space	$\mathcal{X}$	(Might be product space $\mathcal{X}^n$)
Parameter Space	Θ	(Might be subset of $\mathbb{R}^p$)
Likelihood Function	$f_\theta(X)$	(Might be product $\prod f_\theta(x_i)$)

Examples:

Bi: $f_\theta(X) = \prod \binom{n_i}{X_i} (p)^{X_i} \, (1-p)^{n_i - X_i} \quad \propto (p)^{S_n} \, (1-p)^{F_n}$

No: $f_\theta(X) = \prod \left[e^{-(X_i-\mu)^2/2\sigma^2} \big/ \sqrt{2\pi\sigma^2} \right] \quad \propto \sigma^{-n} \, e^{-\frac{n}{2\sigma^2}\left[S_n^2 + (\bar{X}_n - \mu)^2\right]}$

Po: $f_\theta(X) = \prod \left[\lambda^{X_i} \, e^{-\lambda} / X_i! \right] \quad \propto \lambda^{S_n} \, e^{-n\lambda}$

What *is* Bayesian Analysis?

Bayesian analysis is the construction of a probability model for the joint distribution of *all* uncertain quantities—
both **parameters** ($\theta \in \Theta$) and **observables** ($X \in \mathcal{X}$).

Inference is based on the *conditional* distribution of the unknown quantities of interest (sometimes the parameter vector or some function of it, sometimes future or missing observations), **given** the values of any observed quantities (usually $X = x \in \mathcal{X}$, but sometimes only censored or noisy observations are available).

What *is* Bayesian Analysis? (Cont.)

The **Likelihood Function** $f_\theta(x)$ is the density function for X with respect to an (arbitrary) reference measure $m(dx)$. It is evaluated at the observed value of $X = x \in \mathcal{X}$ and regarded as a function of $\theta \in \Theta$.

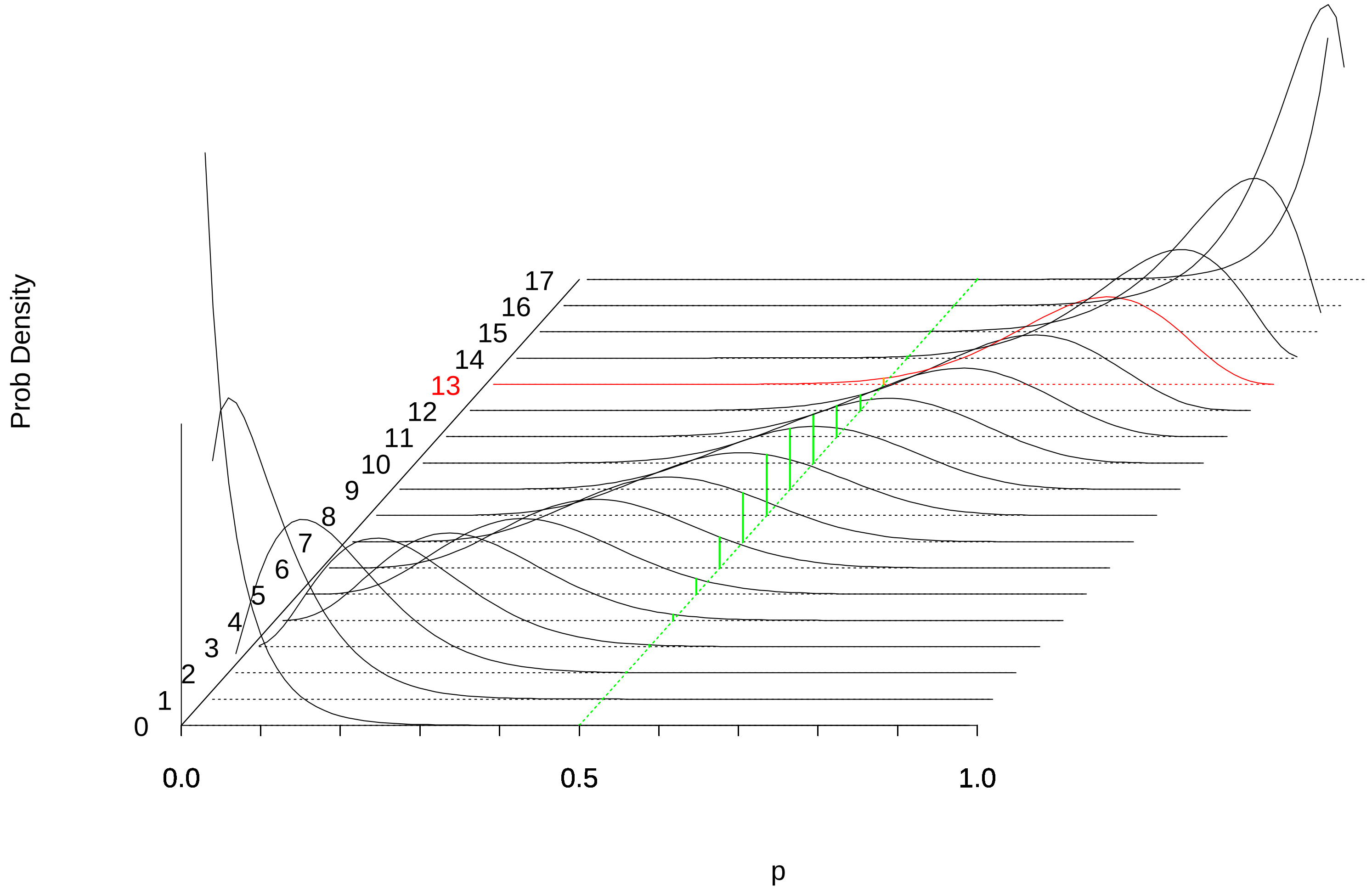
In Bayesian analysis it is thought of as the *conditional* density of X, given θ, and is often written in conditional form

$$L(\theta) = f(x \mid \theta)$$

To complete the joint distribution of $(\theta, X) \in \Theta \times \mathcal{X}$, all we need is

the marginal distribution of $\theta \in \Theta$, or **Prior Distribution** $\pi(d\theta)$.

Example: $\theta \sim \mathsf{Be}(.5, .5)$, $X|\theta \sim \mathsf{Bi}(17, \theta)$:


Prob Density
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
0.0
0.5
1.0
p

Thus we have:

$$
\begin{array}{llcll}
\text{Prior} & \theta & \sim & \pi(d\theta) & = 1/\pi\sqrt{\theta(1-\theta)} \\
\text{Likelihood} & X|\theta & \sim & f(x \mid \theta)\, m(dx) & = \binom{17}{x}\theta^x(1-\theta)^{17-x} \\
\text{Joint} & X,\theta & \sim & \pi(d\theta)\, f(x \mid \theta)\, m(dx) & = \binom{17}{x}\frac{\theta^x(1-\theta)^{17-x}}{\pi\sqrt{\theta(1-\theta)}} \\
\text{Posterior} & \theta|X & \sim & \frac{\pi(d\theta)\, f(x|\theta)}{\int_\Theta \pi(d\theta)\, f(x|\theta)} & \sim \mathsf{Be}(x+.5, 17.5-x) \\
 & & \propto & \pi(d\theta)\, f(x \mid \theta) & \propto \theta^{x-.5}(1-\theta)^{16.5-x}
\end{array}
$$

Bayesian Analysis is inference based on this conditional or, as it is better known, **posterior distribution**— hypothesis tests, estimation, decision theory, *etc.*

Examples of Bayesian Inference:

Estimation:	$T^\pi:$	$\mathcal{X} \to \Theta$	$T^\pi(x) = \mathsf{E}[\theta \mid X = x]$
	$T^\pi:$	$\mathcal{X} \to \Theta$	$\mathsf{P}[\theta \le T^\pi \mid X = x] = 1/2$
Intervals:	$U^\pi:$	$\mathcal{X} \to 2^\Theta$	$\mathsf{P}[\theta \in U^\pi \mid X = x] \ge 0.95$
Hypotheses:	$P_0^\pi:$	$\mathcal{X} \to [0,1]$	$P_0^\pi = \mathsf{P}[H_0 \mid X = x]$
Decisions:	$\delta^\pi:$	$\mathcal{X} \to \mathcal{A}$	$R(\delta^\pi, \theta) \equiv \mathsf{E}[L(\delta^\pi, \theta) \mid \theta]$
			$r(\delta^\pi) \equiv \mathsf{E}[L(\delta^\pi, \theta)]$

Hypothesis Testing: P-Values are Misleading

Consider the question of testing

$$H_0 : \theta = \theta_0 \ \mathit{vs.}\ H_1 : \theta = \theta_1,$$

for two points $\theta_0, \theta_1 \in \Theta$. Set

$$\lambda(X) \equiv \frac{L(\theta_1 \mid X)}{L(\theta_0 \mid X)},$$

the **Likelihood Ratio** *against* H_0, and let $x \in \mathcal{X}$ be the observed value of X. All statistical paradigms (Bayesian, Frequentist, Likelihoodist) use $\lambda(X)$ to quantify the evidence against H_0:

Measures of Evidence Against H_0:

1. **Bayesian** (T Bayes, 1763): Begin with marginal probabilities $\pi_j = \Pr[\theta = \theta_j]$, compute conditional (*posterior*) probabilities

$$\mathsf{P}[H_0 \mid X = x] = \frac{\pi_0 L(\theta_0 \mid x)}{\sum \pi_j \, L(\theta_j \mid x)} = \frac{\frac{\pi_0}{\pi_1}}{\frac{\pi_0}{\pi_1} + \lambda(x)}$$

2. **Fisherian** (RA Fisher, 1934): Compute the P**-value**

$$P = \mathsf{P}[\lambda(X) \geq \lambda(x) \mid H_0]$$

3. **Likelihoodist** (I Hacking, 1965): Treat $\lambda(x)$ itself as the measure of relative support for H_0 and H_1.

Usual Procedures:

1. **Bayesian** (P-S Laplace, 1774): For objectivity, choose $\pi_0 = \pi_1 = 1/2$; Reject H_0 in favor of H_1 if $P[H_0 \mid X] = (1+\lambda)^{-1}$ is small (say, < 0.05);
2. **Fisherian** (RA Fisher, 1934): Reject H_0 in favor of H_1 if $P = \mathsf{P}[\lambda(X) \geq \lambda(x) \mid \theta = \theta_0]$ is small (say, < 0.05);
3. **Likelihoodist** (I Hacking, 1965): Reject H_0 in favor of H_1 if λ is large (say, > 20).

Question:

How often do we wrongly reject **TRUE** hypotheses?

Frequentist Guarantee:

IF H_0 is true, **THEN** the Type-I Error probability is $\leq 5\%$, *i.e.*,

$$\mathsf{P}[\textbf{Reject} \mid H_0] \leq 0.05.$$

But is *that* really the question we want answered?

Suppose an experiment leads to $P \approx 0.05$; what is the chance the (rejected!) hypothesis H_0 is actually true?

$$\mathsf{P}[H_0 \mid \textbf{Reject w/P} = \textbf{.05}] = ????$$

Most naïve users imagine the answer is 5%, more or less.
They are wrong.

Binomial Example:

Our clinic studies 17 matched pairs of subjects; one subject in each pair is given the Treatment, the other the Control. The Treated subject does better in 13 of the 17 pairs.

Do Treated and Control differ?

The P-value for testing the two-sided hypothesis

$$H_0 : [p = 1/2] \qquad vs. \qquad H_1 : [p \neq 1/2]$$

is

$$P = \mathsf{P}[X \geq 13 \text{ or } X \leq 4 \mid p = 1/2] = 0.04904,$$

leading us to **REJECT** H_0 at level 0.05.

What is the probability we are making an error? Is it about 5%?

Repeated Use:

Imagine we use similar statistical methods in a long stream of trials:

- in about half of them, H_0 is true and $p = 1/2$
- in about half of them, H_0 is false and $p \neq 1/2$.

We have to model what p will be in the false-H_0 cases; one choice is to let p be random, perhaps with a symmetric $\mathsf{Be}(\alpha, \alpha)$ distribution for some $\alpha \geq 0$.

The numbers S of successes and $F = n - S$ of failures in a trial of n subjects would satisfy:

Repeated Use (Continued)

$$\begin{aligned}\mathsf{P}[S=s] &= \frac{1}{2}\binom{n}{s}\left(\frac{1}{2}\right)^n + \frac{1}{2}\int_0^1 \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\binom{n}{s}p^{\alpha+s-1}(1-p)^{\alpha+f-1}\,dp \\ &= \frac{1}{2}\binom{n}{s}\left[2^{-n} + \frac{\Gamma(2\alpha)\Gamma(\alpha+s)\Gamma(\alpha+f)}{\Gamma(\alpha)^2\Gamma(2\alpha+n)}\right]\end{aligned}$$

so

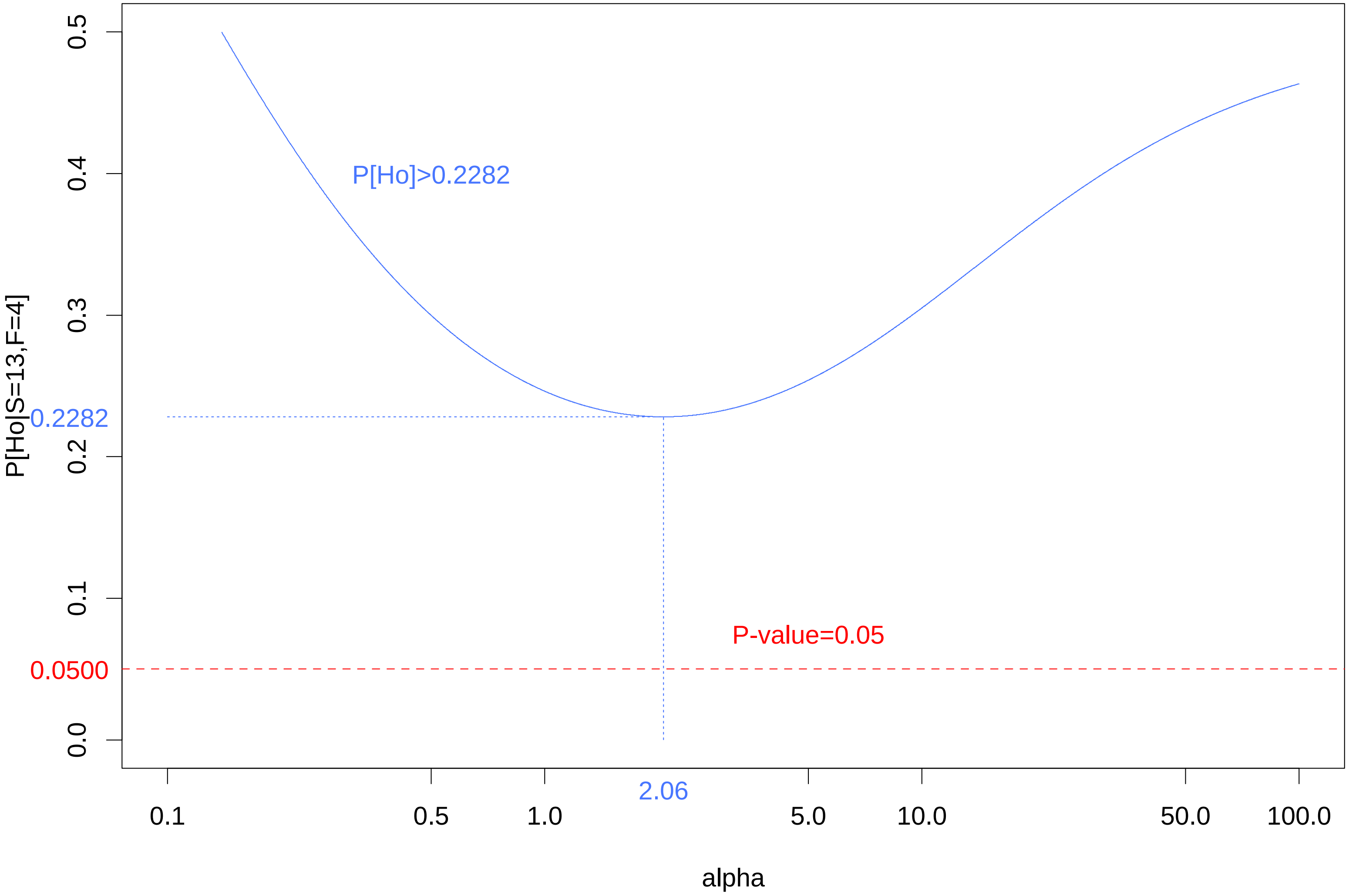
$$\mathsf{P}[H_0 \mid S=s] = \frac{1}{1 + 2^n\,\frac{\Gamma(2\alpha)\Gamma(\alpha+s)\Gamma(\alpha+f)}{\Gamma(\alpha)^2\Gamma(2\alpha+n)}}$$

Back to the Example:

With $S = 13$ and $F = 4$ in $n = S + F = 17$ trials,

$$\begin{aligned} \mathsf{P}[H_0 \mid S, F = 4, 13] &= \frac{1}{1 + 2^{17} \frac{\Gamma(2\alpha)\Gamma(\alpha+13)\Gamma(\alpha+4)}{\Gamma(\alpha)^2\Gamma(2\alpha+17)}} \\ &\geq 0.2282 \qquad \text{for any } \alpha > 0 \end{aligned}$$

(The minimum occurs at about $\alpha \approx 2.0642$). Thus for **any** $\alpha > 0$, the evidence of $S = 13$ with $P-$value of 0.04904 gives only about 3:1 odds against H_0, *not* the 19:1 most people imagine— weak evidence indeed! Should we really reject H_0?

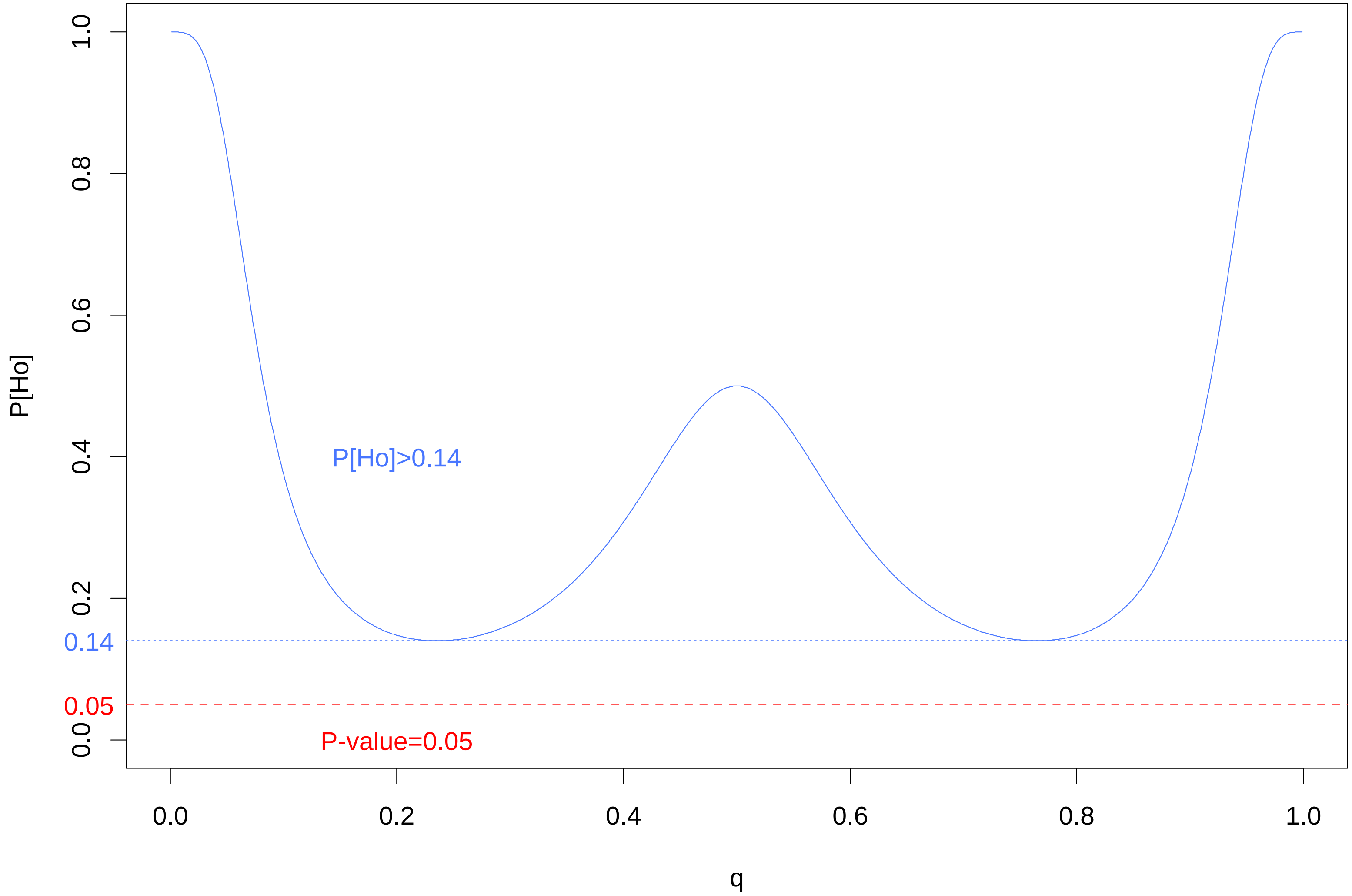

P[Ho]>0.2282
P-value=0.05
0.2282
0.0500
0.0
0.1
0.2
0.3
0.4
0.5
P[Ho|S=13,F=4]
0.1
0.5
1.0
2.06
5.0
10.0
50.0
100.0
alpha

Repeated Use (Continued)

A more extreme way to model what p will be in our false-H_0 cases would be to let p take some fixed value (say, q) when H_0 fails, leading to

$$
\begin{aligned}
\mathsf{P}[P\text{-value} = .04904] &= \mathsf{P}[S = 4 \text{ or } S = 13] \\
&= \tfrac{1}{2}\binom{17}{4}\left(\tfrac{1}{2}\right)^4\left(\tfrac{1}{2}\right)^{13} + \tfrac{1}{2}\binom{17}{13}\left(\tfrac{1}{2}\right)^{13}\left(\tfrac{1}{2}\right)^4 + \\
&\quad \tfrac{1}{2}\binom{17}{4}q^4(1-q)^{13} + \tfrac{1}{2}\binom{17}{13}q^{13}(1-q)^4 \\
&= 1190\left[2 \times 2^{-17} + q^4(1-q)^{13} + q^{13}(1-q)^4\right] \\
\mathsf{P}[H_0 \mid S = 4 \text{ or } 13] &= \frac{2}{2 + 2^{17}\left[q^4(1-q)^{13} + q^{13}(1-q)^4\right]} \\
&\geq 0.1400 \text{ for all } 0 < q < 1.
\end{aligned}
$$

This is the extremum over *all possible* distributions of p under H_1.



P[Ho]
1.0
0.8
0.6
0.4
0.2
0.14
0.05
0.0
P[Ho]>0.14
P-value=0.05
0.0
0.2
0.4
0.6
0.8
1.0
q

Example:

For larger sample-sizes things are more extreme; for example, a trial with 480 successes in 900 matched subject pairs will lead to a P-value for the hypothesis $H_0 : [p = 1/2]$ with two-sided alternative $H_1 : [p \neq 1/2]$ of

$$P = \mathsf{P}[X \geq 480 \text{ or } X \leq 420 \mid p = 1/2] = 0.04916,$$

even though in repeated testing of many hypotheses, half true and half false, the fraction of true hypotheses among those with $P = .04916$ exceeds

$$\begin{aligned} \mathsf{P}[H_0 \mid P = .049] &\geq \inf_{0<q<1} \frac{2}{2 + 2^{900}\left[q^{420}(1-q)^{480} + q^{480}(1-q)^{420}\right]} \\ &= 0.21274. \end{aligned}$$

Normal Examples:

The same thing happens in two-sided normal-distribution examples, and in all regular statistical problems where asymptotic normality holds. The normal calculations are easier and give some idea of how we might "recalibrate" P-values.

Frequentist testing of $H_0 : \mu = \mu_0$ against two-sided alternative $H_1 : \mu \neq \mu_0$ for normally-distributed data $X_i \sim \mathsf{No}(\mu, \sigma^2)$ is an elementary problem, whose well-known answer is given by

$$P = 2\,\Phi\left(\frac{-|\bar{X}_n - \mu_0|}{\sigma/\sqrt{n}}\right),$$

usually leading to rejection whenever $\bar{X}_n$ differs from μ_0 by more than two standard deviations or so. How safe are we in rejecting H_0 when $P = .05$, *i.e.*, when $Z = \sqrt{n}\,(\bar{X}_n - \mu_0)/\sigma = \pm 1.96$?

Normal Examples (Cont'd):

It's an easy calculation; without losing any generality we can take $n = 1$, $\mu_0 = 0$, and $z = x = 1.96$. Again in a stream of 50% true and 50% false hypotheses, perhaps with $\mu = \mu_1$ in all the false ones, we find

$$\begin{aligned}
\mathsf{P}[H_0 \mid P = 0.05] &= \frac{e^{-1.96^2/2}}{e^{-1.96^2/2} + \frac{1}{2}e^{-(1.96+\mu_1)^2/2} + \frac{1}{2}e^{-(1.96-\mu_1)^2/2}} \\
&= \frac{2}{2 + e^{-1.96\mu_1 - \mu_1^2/2} + e^{+1.96\mu_1 - \mu_1^2/2}} \\
&\geq 0.2265
\end{aligned}$$

(The minimum occurs at about $\mu_1 \approx 2$). Thus for **any** μ_1, the evidence of $Z = 1.96$ with $P-$value of 0.05 gives only about 4:1 odds against H_0, again *not* the 19:1 most people imagine.

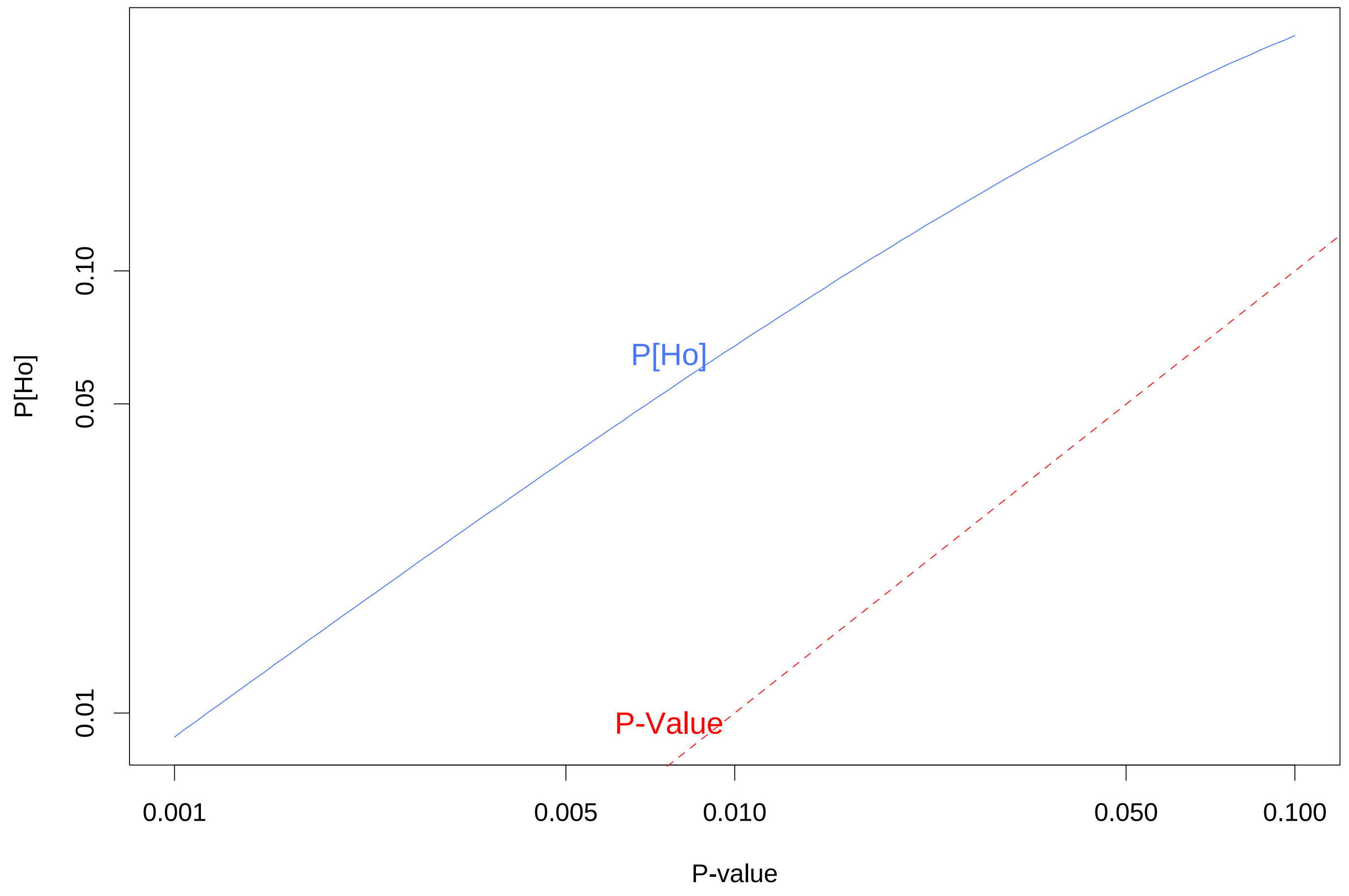
Normal Examples (Cont'd):

More generally, for any P-value $p = \Phi(z)$,

$$\mathsf{P}[H_0 \mid P = p] = \frac{2}{2 + e^{-\mu_1^2/2}[e^{-z\mu_1} + e^{+z\mu_1}]}$$

takes a minimum near $\mu_1 = z$, as much as an order of magnitude higher than p:

P-Value $=$	0.0010	0.0025	0.0050	0.0100	0.0250	0.0500
	999:1	399:1	199:1	99:1	39:1	19:1
Z-Score	3.291	3.023	2.807	2.576	2.241	1.960
$P[H_0] \geq$	0.0088	0.0203	0.0375	0.0676	0.1396	0.2265
	112:1	48:1	26:1	14:1	6:1	3:1


P[Ho]
0.10
0.05
0.01
P[Ho]
P-Value
0.001
0.005
0.010
0.050
0.100
P-value

Sequential Analysis: Stopping Rules

For testing a simple hypothesis $H_0 : X_i \sim f_0(x)\, m(dx)$ with simple alternative $H_1 : X_i \sim f_1(x)\, m(dx)$ with *i.i.d.* data X_i for $i \in \mathbb{N}$, virtually all statistical paradigms summarize the evidence from the first n observations through the **Likelihood Ratio** (*against* H_0),

$$\lambda_n \equiv \frac{f_1(x_1) \cdots f_1(x_n)}{f_0(x_1) \cdots f_0(x_n)}.$$

Likelihood Ratio: $\lambda_n \equiv \dfrac{f_1(x_1)\cdots f_1(x_n)}{f_0(x_1)\cdots f_0(x_n)}$

Bayesians base this choice on the calculation

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

while Frequentists base the choice on the Neyman-Pearson lemma which asserts that the most powerful possible statistical test of H_0 *vs.* H_1 will reject H_0 for large values of λ_n, with P-value

$$P = \mathsf{P}[\lambda_n(X) \geq \lambda_n(x) \mid X_j \sim f_0(x)\, m(dx)]$$

Asymptotics:

If H_0 is true then λ_n is a **positive martingale**:

$$\begin{aligned} E[\lambda_{n+1} \mid \mathcal{F}_n] &= \int_{\mathcal{X}} \frac{f_1(x_1)\cdots f_1(x_n)}{f_0(x_1)\cdots f_0(x_n)} \frac{f_1(x_{n+1})}{f_0(x_{n+1})}\, f_0(x_{n+1})\, d\, x_{n+1} \\ &= \int_{\mathcal{X}} \lambda_n\, f_1(x_{n+1})\, d\, x_{n+1} \\ &= \lambda_n \end{aligned}$$

and so by the Martingale convergence theorem,

$$\lim_{n\to\infty} \lambda_n = 0 \qquad a.s., \text{ if } H_0 \text{ is true}$$

while if H_1 is true then $1/\lambda_n \to 0$ is a positive martingale and

$$\lim_{n\to\infty} \lambda_n = \infty \qquad a.s., \text{ if } H_1 \text{ is true}$$

The logarithm of λ_n is an asymetric random walk under *any* possible distribution $f(x)$ for the $\{X_i\}$, even if *both* hypotheses are false:

$$\log \lambda_n = \sum_{i \le n} \big[\log f_1(X_i) - \log f_0(X_i) \big].$$

The steps have mean $\mu = \mathsf{E}\big[\log f_1(X_i) - \log f_0(X_i) \big]$ given by

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

the difference in the Kullback-Leibler divergences from the true distribution $f(x)$ to f_0 and f_1.

Thus

$$
\begin{aligned}
\log \lambda_n &= \sum_{i \le n} \log \frac{f_1(X_i)}{f_0(X_i)} \text{ is a R.W.,} \\
&\approx n\left[K(f : f_0) - K(f : f_1)\right] \\
&\to -\infty \text{ if } f \text{ is closer to } f_0, \\
&\to +\infty \text{ if } f \text{ is closer to } f_1, \\
&\approx -n\,K(f_0 : f_1) \text{ if } H_0 \text{ true,} \\
&\approx +n\,K(f_1 : f_0) \text{ if } H_1 \text{ true.}
\end{aligned}
$$

Under H_0 the mean and variance are both available:

$$\begin{aligned} \mu_0 &= \int_{\mathcal{X}} \big[\log f_1(x) - \log f_0(x)\big]\, f_0(x)\, m(dx) \\ \sigma^2 &= \int_{\mathcal{X}} \big[\log f_1(x) - \log f_0(x)\big]^2 f_0(x)\, m(dx) - \mu_0^2, \end{aligned}$$

and by the Central Limit Theorem we have the approximate distribution

$$\frac{\log \lambda_n - n\mu_0}{\sqrt{n\sigma_0^2}} \approx \mathsf{No}(0,1)$$

and hence the Frequentist P-value is given approximately by

$$\begin{aligned} P &= \mathsf{P}[\lambda_n(X) \geq \lambda_n(x) \mid H_0] \\ &\approx \Phi\left(-\frac{\log \lambda_n - n\mu_0}{\sqrt{n\sigma_0^2}}\right). \end{aligned}$$

Question:

How often do we reject **TRUE** hypotheses?

If H_0 is **true**, then the **Central Limit Theorem**, and the **Law of the Iterated Logarithm** gives

$$\limsup_{n\to\infty} \pm \frac{\log \lambda_n - n\mu_0}{\sqrt{2n\sigma_0^2 \log\log n}} = 1$$

and so, for any $z \in \mathbb{R}$, there exists $n_z < \infty$ such that

$$\frac{\log \lambda_{n_z} - n_z\mu_0}{\sqrt{n_z\sigma_0^2}} > z$$

and hence $P = \Phi\left(-\frac{\log \lambda_{n_z} - n_z\mu_0}{\sqrt{n_z\sigma_0^2}}\right) \leq \Phi(z)$.

A Sequential Frequentist can reject any **true** H_0 at **any** level $p = \Phi(z)$!!!

How About Bayesian Analysis?

Even at stopping times T the Bayesian error report

$$P_T = P[H_0 \mid \mathcal{F}_T] = \frac{\frac{\pi_0}{\pi_1}}{\frac{\pi_0}{\pi_1} + \lambda_T}$$

satisfies the bound

$$\mathsf{P}[P_T \le p \mid H_0] \le p$$

for every $0 < p < 1$, so a Bayesian analysis allows one to stop sampling for any reason at all without distorting the results— ran out of money, time, or patience; achieved "significance"; computer crashed; coffee ran out...

A Few Incentives for Bayesian Analysis:

1. Does not distort evidence like P-values do
2. Freedom to use arbitrary stopping rules in sequential settings
3. An Estimator or Test Procedure is **Admissible** (*i.e.*, not worse than some other for all $\theta \in \Theta$) **if and only if** it is the Bayes Procedure for some prior $\pi(d\theta)$
4. Easy to "elaborate" models to add additional features, and to synthesize evidence from different sources, through **hierarchical models**

SO, let's look how we can use Bayesian methods in some specific problems... more next time!

Related Readings:

- Berger, J.O. and R.L. Wolpert (1984). *The Likelihood Principle*. Springer-Verlag, Berlin.
- Berger, J.O., L.D. Brown, and R.L. Wolpert (1994) A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.* **22**, 1787–1807.
- Wolpert, R.L. (1996). Testing simple hypotheses. In *Studies in Classification, Data Analysis, and Knowledge Organization*, H.H. Bock and W. Polasek, eds., 289–297. Springer-Verlag, Berlin.
- Berger, J.O. and B. Boukai and Y. Wang (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Stat. Sci.* **12**, 133–160.