1 Estimation and Beyond in the Bayes Universe.

1.1 Estimation

No Bayes estimate can be unbiased but Bayesians are not upset! No Bayes estimate with respect to the squared error loss can be unbiased, except in a trivial case when its Bayes' risk is 0.

Suppose that for a proper prior π the Bayes estimator $\delta_\pi(X)$ is unbiased,

$$(\forall\theta)E^{X|\theta}\delta_\pi(X)=\theta.$$

This implies that the Bayes risk is 0.

The Bayes risk of $\delta_\pi(X)$ can be calculated as repeated expectation in two ways,

$$r(\pi,\delta_\pi)=E^\theta E^{X|\theta}(\theta-\delta_\pi(X))^2=E^X E^{\theta|X}(\theta-\delta_\pi(X))^2.$$

Thus, conveniently choosing either $E^\theta E^{X|\theta}$ or $E^X E^{\theta|X}$ and using the properties of conditional expectation we have,

$$\begin{aligned} r(\pi,\delta_\pi) &= E^\theta E^{X|\theta}\theta^2-E^\theta E^{X|\theta}\theta\delta_\pi(X)-E^X E^{\theta|X}\theta\delta_\pi(X)+E^X E^{\theta|X}\delta_\pi^2(X)\\ &= E^\theta E^{X|\theta}\theta^2-E^\theta\theta[E^{X|\theta}\delta_\pi(X)]-E^X\delta_\pi(X)E^{\theta|X}\theta+E^X E^{\theta|X}\delta_\pi^2(X)\\ &= E^\theta E^{X|\theta}\theta^2-E^\theta\theta\cdot\theta-E^X\delta_\pi(X)\delta_\pi(X)+E^X E^{\theta|X}\delta_\pi^2(X)=0. \end{aligned}$$

Bayesians are not upset. To check for its unbiasedness, the Bayes estimator is averaged with respect to the model measure $(X|\theta)$, and one of the Bayesian commandments is: *Thou shall not average with respect to sample space, unless you have Bayesian design in mind.* Even frequentist agree that insisting on unbiasedness can lead to bad estimators, and that in their quest to minimize the risk by trading off between variance and bias-squared a small dosage of bias can help. The relationship between Bayes estimators and unbiasedness is discussed in Lehmann (1951), Girshick (1954), Bickel and Blackwell (1967), Noorbaloochi and Meeden (1983) and OHagan (1994).

Here is an interesting Bayes estimation problem. For the solution I admittedly used Wolfram's MATHEMATICA software since my operational knowledge of special functions is not to write home about.

Binomial n from a single observation! Two graduate students from GaTech were conducting a survey of what percentage p of Atlanteans will vote for reelection of the President in November 2004. The student who did the actual survey left the value $X=10$ on the answering machine of the other student but did not say what sample size n was used, and left for China, while the project was due in a few days. What n and p should be reported?

The problem can be formalized to estimation of Binomial proportion and sample size on the basis of a single measurement. This is a problem where one wishes to be a Bayesian since the frequentist solutions involve lots of hand waiving!

Let X be a single observation from $\mathcal{B}(n,p)$ and let $p\sim Beta(\alpha,\beta)$ and $n\sim Poi(\lambda)$.

Then the Bayes rule for n is given by

$$\delta(x)=x+\lambda\left[\frac{\partial}{\partial z}\ \ln{}_1F_1(\beta,\alpha+\beta+x;z)\right]_{z=\lambda} \tag{1}$$

	$X=10$	$X=15$	$X=20$	$X=30$	$X=100$	$X=1000$
I	19.4342	21.214	24.2752	32.329	100.472	1000.04
II	28.6666	29.2974	30.4852	35.4996	100.798	1000
III	17.3773	23.626	29.5563	40.888	114.437	1017.49
IV	21.6113	28.8745	35.6381	48.2978	126.481	1035.95

Table 1: Estimators $\hat{n}$ when X is observed. Prior on p is Beta(2,2) and prior on n is: (I) Poisson $\lambda = 20$; (II) Poisson $\lambda = 20$; (III) geometric $\lambda = 0.9$; and (IV) geometric $\lambda = 0.95$.

where

$$ {}_1F_1(\alpha,\gamma;z) = 1 + \frac{\alpha}{\gamma}\frac{z}{1} + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \frac{\alpha(\alpha+1)((\alpha+2)}{\gamma(\gamma+1)(\gamma+2)}\frac{z^3}{3!} + \cdots \tag{2} $$

is the Kummer confluent function.
The marginal likelihood for n [after integrating out p] is

$$ f(x|n) = \frac{\Gamma(\alpha+x)}{\Gamma(1+x)B(\alpha,\beta)} \cdot \frac{\Gamma(n+1)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)\Gamma(n+1-x)}. $$

Under the Poisson prior $\pi(n) = \frac{\lambda^n}{n!}e^{-\lambda},\ n = 0,1,2,\ldots$ the posterior mean (Bayes rule under s.e.l.) is

$$ \begin{aligned} \delta(x) &= \frac{\sum_{n\geq x} n f(x|n)\pi(n)}{\sum_{n\geq x} f(x|n)\pi(n)} \\ &= x + \lambda\frac{\Gamma(\beta+1)}{\Gamma(\beta)}\frac{\Gamma(\alpha+\beta+x)}{\Gamma(\alpha+\beta+x+1)}\frac{{}_1F_1(\beta+1,1+\alpha+\beta+x;\lambda)}{{}_1F_1(\beta,\alpha+\beta+x;\lambda)}. \end{aligned} $$

Since

$$ \frac{\partial}{\partial z}\,{}_1F_1(\alpha,\gamma;z) = \frac{\alpha}{\gamma}\,{}_1F_1(\alpha+1,\gamma+1;z) $$

the statement in (1) follows.

When the prior on n is $\mathcal{G}eom(1-\lambda)$ [$\pi(n|\lambda) = \lambda^n(1-\lambda),\ \lambda \geq 0,\ n = 0,1,\ldots$], and the prior on p is Beta $\mathcal{B}e(\alpha,\beta)$, the Bayes rule is

$$ \delta(x) = x + \frac{\lambda\beta(x+1)}{x+\alpha+\beta}\frac{{}_pF_q(\{\beta+1,x+2\},\{x+\alpha+\beta+1\},\lambda)}{{}_pF_q(\{\beta,x+1\},\{x+\alpha+\beta\},\lambda)}, $$

where ${}_pF_q$ is the generalized hypergeometric function.

Table 1 gives some values of $\hat{n}$ for different X. Note the "surprise effect" [if the prior expectation for n is small compared to X, the $\hat{n}$ is essentially X.]

Tramcar Problem. You have found yourself in a city of Belgrade (Serbia and Montenegro). Belgrade does not have underground system, instead it has developed tramcar traffic system. The tramcars are marked consecutively from 1 to N. You have no detailed prior information about the city (size, area, etc.) since

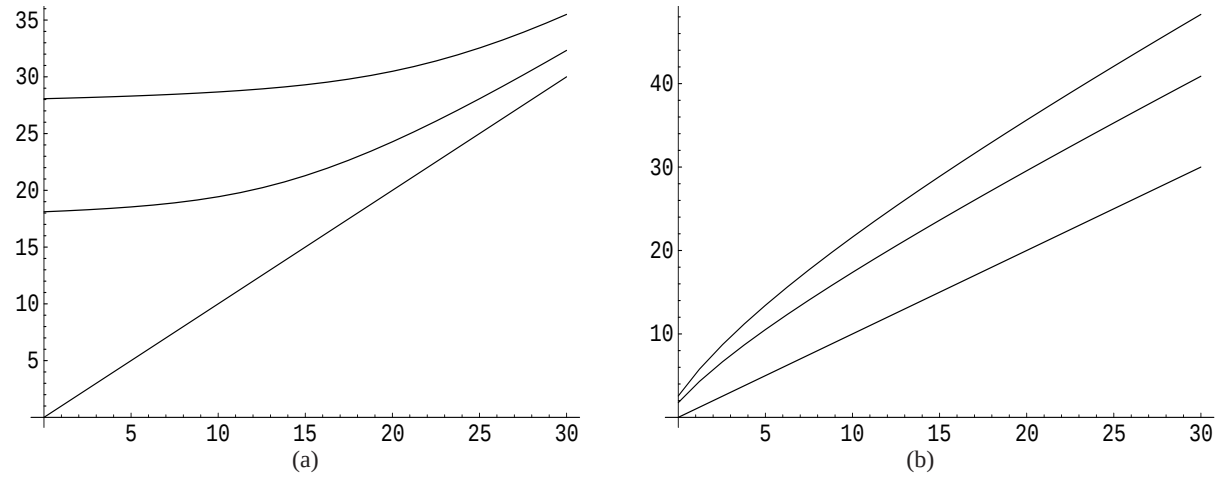


Figure 1: (a) Bayes rules for $\alpha = \beta = 2$ and Poisson prior on n with $\lambda = 20$ and $\lambda = 30$; (b) Bayes rules for $\alpha = \beta = 2$ and geometric prior on n with $\lambda = 0.9$ and $\lambda = 0.95$.

geography of Balkan is not your strong side, and old Yugoslavia guidebook you got talks only about restaurants and hotels at the Adriatic seaside. Estimate N if you have seen a single tramcar marked 100 on your way from the airport.

The tramcar problem was discussed by Jeffreys (1989) who attributes it to M. H. A. Newman. Some versions use San Francisco as a city. The tramcar problem is a special case of *coincidences problem* discussed by Diaconis and Mosteller (1989).

Obviously, given N, the observed tramcar is one out of N and its number X is distributed as discrete uniform on $\{1, 2, \ldots, N\}$. The unknown N apriory can take any value in $\mathbb{N} = 1, 2, 3, \ldots$. A flat prior $P(N = n) = 1,\ n \in \mathbb{N}$ would not lead to the proper posterior. Instead, we take the prior $P(N = n) = 1/n,\ n \in \mathbb{N}$. This is not a probability distribution but the posterior would be. Robert (2001) argues that N in this problem could be interpreted as "scale" parameter and the reciprocal prior is justified. The posterior, if X is observed is proportional to $\frac{1}{N^2}\mathbf{1}(N \geq X)$. The normalizing constant is $\psi(X) = \sum_{k=X}^{\infty} \frac{1}{k^2}$ – the second derivative of the logarithm of Gamma function, $\psi(x) = \frac{d^2\Gamma(x)}{dx^2}$, evaluated at X.

The posterior mean does not exist since $P(N \geq k|X) = \psi(k)/\psi(X) \approx \frac{\int_k^{\infty} 1/x^2 dx}{\int_X^{\infty} 1/x^2 dx} = X/k$, and $\sum_{k \geq X} P(N \geq k|X) = \infty$. But the posterior median is close to $2X$.

If $X = 100$, as in the problem, the posterior is given in Figure 2, The normalizing constant is $\psi(100) = 0.0100502$. Let $p_k = P(N = k|X)$.

From $\sum_{k=100}^{198} p_k = 0.4987$ and $\sum_{k=100}^{199} p_k = 0.5012$ we conclude that the posterior median is between 198 and 199.

We could use the prior $P(N = n) = \frac{1}{n^{\alpha}},\ \alpha > 0, n \in \mathbb{N}$. The hyperparameter α could reflect our opinion on the size of the city. If we believe the city is large, α should be small. The solution would involve Riemann's generalized zeta function, but numerical solution is straightforward. If $\alpha > 1$ the posterior mean for N would exist.

A Tiny Dosage of Bayesian Hypocrisy. Assume that given θ, an observation X is distributed as uniform

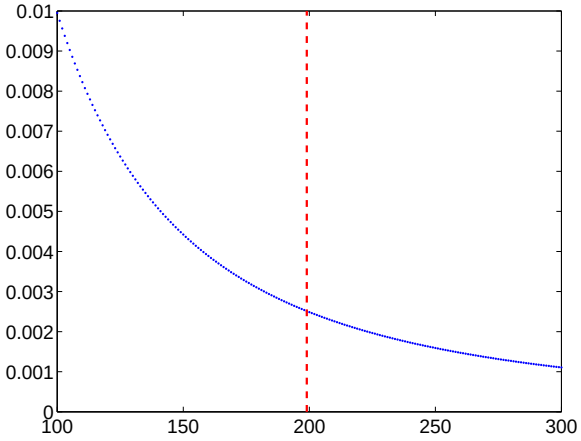


Figure 2: Tramcar posterior probabilities of $N = k,\ k = 100, 101, \ldots$ for $X = 100$. Vertical line denotes the posterior median, the posterior mean does not exist.

$\mathcal{U}(0, |\theta|)$. The prior for θ is uniform $\mathcal{U}(-10, 10)$. If $X = 9$ is observed, what are the estimators of θ and $|\theta|$?

The posterior is proportional to $\frac{1}{|\theta|}\mathbf{1}(-10 \le \theta \le 10)\mathbf{1}(0 \le x \le |\theta|)$, and the normalizing constant C satisfies

$$1 = C \int_x^{10} \frac{d\theta}{\theta} = C(\log \frac{10}{x}),$$

from which $C = \frac{1}{\log 10/x}$. Thus, $\pi(\theta|x) = \frac{1}{2|\theta| \log 10/x}\mathbf{1}(x \le |\theta| \le 10)$.

If $X = 9$ is observed, $\pi(\theta|9) = \frac{1}{2|\theta| \log 10/9}\mathbf{1}(9 \le |\theta| \le 10)$.

The Bayes estimator for $|\theta|$ is $E^{\theta|x}|\theta| = \int_9^{10} \theta \frac{1}{\theta \log 10/9} d\theta = \frac{1}{\log 10/9} = 9.4912$.

The Bayes estimator of θ is at best useless; $X = 9$ is observed, and $\hat{\theta} = E^{\theta|x}\theta = 0$, since the posterior is symmetric. Now, I am reporting an estimator/action similar to averaged frequentist estimators I used to ridicule in the conditionality principle. Am I a Bayesian hypocrite? To my defense, the frequentist could observe and be conditional, but θ's are not observable. Should I report a randomized action for θ, $9.4912 \cdot X + (-9.4912) \cdot (1 - X)$, where X is 1 if my favorite coin flips heads up?

1.2 Interval Estimation: Credible Sets

Bayesians call interval estimators of model parameters *credible sets.* Of course, the measure used to assess the credibility of an interval estimator is the posterior measure, if available. If not, the prior will do. Students learning concept of classical confidence intervals (CI's) often make an error by stating that *the probability that the CI interval* $[L, U]$ *contains parameter* θ *is* $1 - \alpha$. The right statement seems convoluted, one needs to generate data from such model many times and for each data set to exhibit the CI. Now, the proportion of CI's covering the unknown parameter is "tends to" $1 - \alpha$. Bayesian interpretation of a credible set C is natural: The probability of a parameter belonging to the set C is $1 - \alpha$. A formal definition follows.

Assume the set C is a subset of Θ. Then, C is credible set with credibility $(1 - \alpha) \cdot 100\%$ if

$$P^{\theta|X}(\theta \in C) = E^{\theta|X}\mathbf{1}(\theta \in C) = \int_C \pi(\theta|x)d\theta \ge 1 - \alpha.$$

If the posterior is discrete, then the integral becomes sum (counting measure) and

$$P^{\theta|X}(\theta \in C) = \sum_{\theta_i \in C} \pi(\theta_i|x) \geq 1 - \alpha.$$

This is the definition of *a* $(1-\alpha)100\%$ credible set, and of course for a given posterior function such set is not unique.

For a given credibility level $(1-\alpha)100\%$, the shortest credible set is of interest. To minimize size the sets should correspond to highest posterior probability (density) areas. Thus the acronym HPD.

Definition: The $(1-\alpha)100\%$ HPD credible set for parameter θ is a set C, subset of Θ of the form

$$C = \{\theta \in \Theta | \pi(\theta|x) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant for which

$$P^{\theta|X}(\theta \in C) \geq 1 - \alpha.$$

Geometrically, if the posterior density is cut by a horizontal line at the hight $k(\alpha)$, the set C is projection on the θ axis of the part of line inside the density, i.e., the part that lies below the density.

Result: The HPD set C minimizes the size among all sets $D \in \Theta$ for which

$$P^{\theta|X}(\theta \in D) = 1 - \alpha.$$

The proof is essentially a special case of Neyman-Pearson lemma. If $I_C(\theta) = \mathbf{1}(\theta \in C)$ and $I_D(\theta) = \mathbf{1}(\theta \in D)$, then the key observation is

$$(\forall \theta) \quad (\pi(\theta|x) - k(\alpha)) \cdot (I_C(\theta) - I_D(\theta)) \geq 0.$$

Indeed, for θ's in $C \cap D$ and $(C \cup D)^c$, the factor $I_C(\theta) - I_D(\theta) = 0$. If $\theta \in C \cap D^c$, then $I_C(\theta) - I_D(\theta) = 1$ and $\pi(\theta|x) - k(\alpha) \geq 0$. If, on the other hand, $\theta \in D \cap C^c$, then $I_C(\theta) - I_D(\theta) = -1$ and $\pi(\theta|x) - k(\alpha) \leq 0$. Thus,

$$\int_\Theta (\pi(\theta|x) - k(\alpha))[I_C(\theta) - I_D(\theta)]d\theta \geq 0.$$

The statement of the theorem now follows from the chain of inequalities,

$$\begin{aligned}
&\int_C (\pi(\theta|x) - k(\alpha))d\theta \geq \int_D (\pi(\theta|x) - k(\alpha))d\theta \\
&(1-\alpha) - k(\alpha) \cdot \text{ size}(C) \geq (1-\alpha) - k(\alpha) \cdot \text{ size}(D) \\
&\text{size}(C) \leq \text{ size}(D).
\end{aligned}$$

The size of a set is simply its total length if the parameter space Θ is one dimensional, total area, if Θ is two dimensional, and so on.

IQ Example (Continued from Handout 3). Recall Jeremy, the enthusiastic GaTech student, that that used Bayesian inference in modeling his IQ test scores. For a score $X|\theta$ he was using normal $\mathcal{N}(\theta, 80)$ likelihood,

while the prior was $\mathcal{N}(110, 120)$. After a score of $X = 98$ was recorded, the resulting posterior was normal $\mathcal{N}(102.8, 48)$.

Frequentist point estimator is $\hat{\theta} = 98$, and 95% confidence interval is $[98 - 1.96\sqrt{80},\ 98 + 1.96\sqrt{80}] = [80.4692, 115.5308]$. The length of this interval is approximately 35.

Bayesian counterparts are $\hat{\theta} = 102.8$, and $[102.8 - 1.96\sqrt{48},\ 102.8 + 1.96\sqrt{48}] = [89.2207, 116.3793]$ which has the length of approximately 27.

Notice that the Bayesian credible set is shorter. This is a consequence of the fact that posterior variance is smaller than the likelihood variance, because of the incorporation of information.

Cauchy Example. Augustine Cauchy[1] did not dream that the distribution that is named after him would serve as a source of many counterexamples and peculiarities in probability and statistics. Assume that $X_1 = 2, X_2 = -7, X_3 = 4$, and $X_4 = -6$ are sampled from Cauchy $\mathcal{C}a(\theta, 1)$ distribution with parameter of interest θ. The likelihood is

$$f(x|\theta) \propto \frac{1}{1+(2-\theta)^2} \cdot \frac{1}{1+(7+\theta)^2} \cdot \frac{1}{1+(4-\theta)^2} \cdot \frac{1}{1+(6+\theta)^2},$$

which, unfortunately, does not simplify. For the flat prior $\pi(\theta) = 1$, the posterior is proportional to the likelihood. Figure 3 shows the posterior. Notice that it is bimodal.

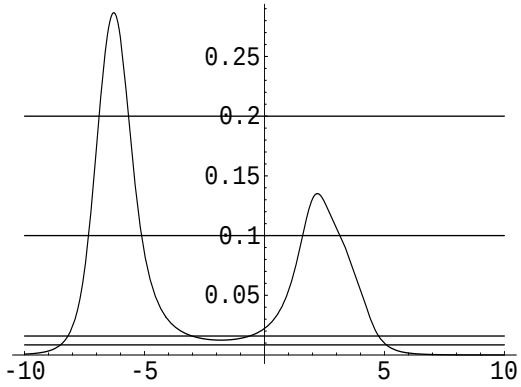


Figure 3: The Cauchy Example. Horizontal lines $k = 0.008475, 0.0159, 0.1$, and 0.2 are shown. The first two determine 99% and 95% credible sets.

Four horizontal lines at levels $k = 0.008475, 0.0159, 0.1$, and 0.2 are shown. These lines determine four credible sets,

- $k(0.01) = 0.008475:\ \ [-8.498, 5.077]$ with $P^{\theta|X}(-8.498 \le \theta \le 5.077) = 99\%$;
- $k(0.05) = 0.0159:\ \ [-8.189, -3.022] \cup [-0.615, 4.755]$ with posterior credibility of 95%;
- $k = 0.1:\ \ [-7.328, -5.124] \cup [1.591, 3.120]$ with posterior credibility of 0.642; and
- $k = 0.2:\ \ [-6.893, -5.667]$ with posterior credibility of 0.313.

Notice that for $k = 0.0159$ and $k = 0.1$ the credible set for θ is not a compact. This shows that two separate intervals "clash" for the ownership of θ and this is a useful information. This non-compactness can also point out that the prior is not agreeing with the data.

There is no frequentist counterpart for the CI for θ in the above model.

Lack of Invariance. One undesirable property of credible sets is the lack of invariance with respect to

[1] Statistical Joke: **Question:** What was Cauchy's least favorable question? **Answer:** Got a moment?

monotone transformations. Here is an example form Berger (1985) presented as an exercise.

Let $X|\theta$ be the shifted exponential with density $f(x|\theta) = e^{-(x-\theta)} \cdot \mathbf{1}(\theta \leq x < \infty)$. Let θ be half-Cauchy, $\pi(\theta) = \frac{2}{\pi(1+\theta^2)}$, $\theta \geq 0$.

(i) Show that the posterior is proportional to

$$\pi(\theta|x) \propto \frac{e^{\theta}}{1+\theta^2}\mathbf{1}(0 \leq \theta \leq x).$$

(ii) Demonstrate that the posterior is increasing in θ and that $(1-\alpha)100\%$ HPD credible set is of the form $[\beta, x]$, for some $\beta \in (0, x)$.

(iii) Transform the posterior to $\pi^*(\eta|x)$ by transformation $\eta = e^{\theta}$. You will obtain

$$\pi^*(\eta|x) \propto \frac{1}{1+(\log(\eta))^2}\mathbf{1}(1 \leq \eta \leq \log x).$$

Show that $\pi^*(\eta|x)$ is decreasing in η and that credible set for η is of the form $[1, \gamma]$, for some $\gamma < e^x$.

(iv) Transform this interval back to the space of θ's, obtain $[\log 1, \log \gamma] = [0, \beta'] \neq [\beta, x]$.

1.3 Testing Statistical Hypotheses in Bayesian Fashion

Bayesians test statistical hypotheses [2] by evaluating prior or posterior probabilities of regions in the parameter space Θ. For example, the hypothesis H_i is the statement that parameter θ belongs to a particular subset $\Theta_i \subset \Theta$. Often, the sets Θ_i partition the parameter space Θ, but in the case of model comparisons and model selection can be singletons.

Two approaches to Bayesian testing are often considered, although they are closely related: Decision Theoretic Approach and Bayes Factors.

1.4 Decision Theoretic Approach

Let $X|\theta$ be distributed as $f(x|\theta)$ and let $\theta \in \Theta$ be the parameter of interest with prior distribution $\pi(\theta)$.

Assume that Θ_0 and Θ_1 are two nonoverlapping subsets of Θ. We assume that Θ_0 and Θ_1 partition Θ, that is, $\Theta_1 = \Theta_0^c$, although formulations in which $\Theta_1 \neq \Theta_0^c$ are easily incorporated.

Call the statement $\theta \in \Theta_0$ the null hypothesis H_0 and the statement $\theta \in \Theta_1 = \Theta_0^c$ the alternative hypothesis H_1,

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1.$$

From the decision theoretic standpoint it is natural to make inference about the indicator $\mathbf{1}(\theta \in \Theta_0)$ and have actions in the space $\mathcal{A} = \{0, 1\}$. If $a = 1$ the hypothesis H_0 is accepted, if $a = 0$ it is rejected.

The traditional loss is 0-1 loss is

$$L(\theta, a) = \mathbf{1}(a \neq \mathbf{1}(\theta \in \Theta_0)),$$

[2] Teacher's Disclaimer: Whenever I hear *test H_0, reject H_0,p-value is 0.075, ...* and similar testing statistical jargon, I cannot help thinking of how the concept of testing and especially its methodology is unnatural and how the general population of users was coerced by a few to use it. The jargon of empirical testing and quality control from World War II-time factories found its place in statistics. Researchers could not publish in scientific journals (especially medical) if their results had not been significant at the level $\alpha = 0.05$. Now, the times changed, but the battle for sensible inference could be lost. An ill conceived concept – testing statistical hypotheses – is so prevalent that its eradication is impossible. Strictly speaking, Bayesians do not test hypotheses in a rigid accept/reject fashion. They summarize posterior distributions. Bayesians evaluate probabilities of regions in the parameter space with respect to prior and posterior measures. They compare and select models. I believe that in some Bayesian texts the testing vocabulary remained traditional more as a communication convenience, and smooth transition from classical to Bayesian point of view, and we will use it as is (at places).

This loss is 1 when the action and the indicator $\mathbf{1}(\theta \in \Theta_0)$ differ, and no loss is incurred if the action matches the true state of nature, i.e., when $a = \mathbf{1}(\theta \in \Theta_0)$.

The optimal action a^*, minimizing the posterior expected loss, is 1 if the posterior probability of hypothesis H_0 exceeds 1/2, and 0 else,

$$a^* = \mathbf{1}(\ P^{\theta|X}(\theta \in \Theta_0) > P^{\theta|X}(\theta \in \Theta_1)\ ) = \mathbf{1}(\ P^{\theta|X}(\theta \in \Theta_0) > 1/2).$$

Thus, the hypothesis with higher posterior probability is selected.

The above decision theoretic formulation can be generalized to losses that penalize errors $a > \mathbf{1}(\theta \in \Theta_0)$ and $a < \mathbf{1}(\theta \in \Theta_0)$ differently. Let the penalty for $a = 0$ when $\mathbf{1}(\theta \in \Theta_0) = 1$ be K_0 and the penalty for $a = 1$ when $\mathbf{1}(\theta \in \Theta_0) = 0$ be K_1. If $a = \mathbf{1}(\theta \in \Theta_0)$ no loss is incurred. The optimal action is

$$a^* = \mathbf{1}\left[\ P^{\theta|X}(\theta \in \Theta_0) > \frac{K_1}{K_0 + K_1}\right],$$

that is, accept H_0 if its posterior probability exceeds $\frac{K_1}{K_0+K_1}$. Notice that scaling of K_0 and K_1 is unimportant since the "acceptance threshold" $\frac{K_1}{K_0+K_1} = \frac{1}{1+K_0/K_1}$ depends on the ratio K_0/K_1.

IQ Example (Continued from Handouts 3,7). Recall Jeremy, the enthusiastic GaTech student, that that used Bayesian inference in modeling his IQ test scores. For a score $X|\theta$ he was using normal $\mathcal{N}(\theta, 80)$ likelihood, while the prior was $\mathcal{N}(110, 120)$. After a score of $X = 98$ was recorded, the resulting posterior was normal $\mathcal{N}(102.8, 48)$.

Jeremy claims it was not his day and his genuine IQ is at least 105. After all he is at Tech!

The posterior probability of $\theta \geq 105$ is

$$p_0 = P^{\theta|X}(\theta \geq 105) = P\left(Z \geq \frac{105 - 102.8}{\sqrt{48}}\right) = 1 - \Phi(0.3175) = 0.3754,$$

less that 38%, and his claim is rejected. Posterior odds in favor of H_0 are 0.3754/(1-0.3754)=0.4652.

If underestimating Jeremy's IQ is two times more serious than overestimating it (Jeremy may be hurt and leave Tech, Jeremy's parents are litigious,...), that is if $K_0 = 2K_1$, then the "acceptance threshold" $\frac{1}{1+K_0/K_1} = 1/3 < 0.3754$, and Jeremy's claim is not rejected.

Notice, that if we did not know of Jeremy's score $X = 98$, the probabilities of the hypotheses would be evaluated with respect to prior distribution $\mathcal{N}(110, 120)$. The prior probability of $\theta \geq 105$ is

$$\pi_0 = P^{\theta}(\theta \geq 105) = P\left(Z \geq \frac{105 - 110}{\sqrt{120}}\right) = \Phi(0.4564) = 0.676.$$

Thus, the prior odds in favor of H_0 are 2.0864.

Exercise. Assume Normal/Normal model with known variance, $[X|\theta] \sim \mathcal{N}(\theta, \sigma^2)$, $[\theta] \sim \mathcal{N}(\mu, \tau^2)$, and $[\theta|X] \sim \mathcal{N}(\mu^*, \rho^2)$. Assume $K_0 - K_1$ loss. Show that for testing $H_0 : \theta \leq \theta_0$ the "rejection region" is $\mu^* > \theta_0 + z_{K_0,\,K_1}\ \rho$, where $z_{K_0,\,K_1} = \Phi^{-1}\left(\frac{K_1}{K_0+K_1}\right)$. In the classical α-level test, the rejection region is $X > \theta_0 + z_{1-\alpha}\sigma$.

$0 \le \log B_{10}(x) \le 0.5$	evidence against H_0 is **poor**
$0.5 \le \log B_{10}(x) \le 1$	evidence against H_0 is **substantial**
$1 \le \log B_{10}(x) \le 2$	evidence against H_0 is **strong**
$\log B_{10}(x) > 2$	evidence against H_0 is **decisive**

Table 2: Treatment of H_0 according to the value of log-Bayes factor.

1.5 Bayes Factor

Let $\frac{\pi_0}{\pi_1} = \frac{P^\theta(\theta\in\Theta_0)}{P^\theta(\theta\in\Theta_1)}$ and $\frac{p_0}{p_1} = \frac{P^{\theta|X}(\theta\in\Theta_0)}{P^{\theta|X}(\theta\in\Theta_1)}$ be the prior and posterior odds in favor of the hypothesis H_0, respectively.

Definition. The Bayes factor in favor of H_0 is the ratio of corresponding posterior to prior odds,

$$B^\pi_{01}(x) = \frac{\frac{P^{\theta|X}(\theta\in\Theta_0)}{P^{\theta|X}(\theta\in\Theta_1)}}{\frac{P^\theta(\theta\in\Theta_0)}{P^\theta(\theta\in\Theta_1)}} = \frac{p_0/p_1}{\pi_0/\pi_1}. \tag{3}$$

From the definition, $B^\pi_{10}(x) = 1/B^\pi_{01}(x)$.

Bayes Factor for Simple Hypotheses is Likelihood Ratio. When the hypotheses are simple $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, and the prior is two point distribution $\pi(\theta_0) = \pi_0$ and $\pi(\theta_1) = \pi_1 = 1 - \pi_0$, the Bayes factor in favor of H_0 becomes likelihood ratio. Indeed,

$$B^\pi_{01}(x) = \frac{\frac{P^{\theta|X}(\theta\in\Theta_0)}{P^{\theta|X}(\theta\in\Theta_1)}}{\frac{P^\theta(\theta\in\Theta_0)}{P^\theta(\theta\in\Theta_1)}} = \frac{f(x|\theta_0)\pi_0}{f(x|\theta_1)\pi_1}\Big/\frac{\pi_0}{\pi_1} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

If the prior is a mixture of two priors, ξ_0 under H_0 and ξ_1 under H_1, then Bayes factor is ratio of two marginal (priorpredictive) distributions generated by ξ_0 and ξ_1. The spirit is "likelihood" ratio, and importance of observations is emphasized although the priors are important as well since marginals depend on them.

Thus, if $\pi(\theta) = \pi_0\xi_0(\theta) + \pi_1\xi_2(\theta)$ then,

$$B^\pi_{01}(x) = \frac{\frac{\int_{\Theta_0} f(x|\theta)\pi_0\xi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta)\pi_1\xi_1(\theta)d\theta}}{\frac{\pi_0}{\pi_1}} = \frac{m_0(x)}{m_1(x)}.$$

Bayes factor measures relative change in prior odds once the evidence is collected. Following Good, Jeffreys, and others, Table 2 gives guidelines for Bayesian testing of hypotheses depending on the value of log-Bayes factor. One could use $B^\pi_{01}(x)$ of course, but then $a \le \log B_{10}(x) \le b$ becomes $-b \le \log B_{01}(x) \le -a$. Negative values of log-Bayes factor are handled by symmetry and changed wording, in an obvious way.

Exercise. Jeremy again. What is the Bayes factor B^π_{01} in the Jeremy's case. Test H_0 is using the Bayes factor and wording from the Table 2. [Solution: $B^\pi_{01}(98) = 0.6882$, $\log B^\pi_{10}(98) = 0.3737$ and the evidence against H_0 is poor.]

Exercise. Assume $[X|\theta] \sim \mathcal{N}(\theta, \sigma^2)$ and $[\theta] \sim \pi(\theta) = 1$. Consider testing $H_0 : \theta \le \theta_0$ v.s. $H_1 : \theta > \theta_0$. Show that $p_0 = P^{\theta|X}(\theta \le \theta_0)$ is equal to classical p-value.

1.6 Bayesian Testing of Precise Hypotheses

Testing precise hypotheses in Bayesian fashion has a considerable body of research. Berger (1985), pages 148–157, has a comprehensive overview of the problem and provides wealth of references. See also Berger and Sellke (1984) and Berger and Delampady (1987).

If the priors are continuous, testing precise hypotheses in Bayesian fashion is impossible since for continuous priors and posteriors the probability of a singleton is 0.

Suppose $X|\theta \sim f(x|\theta)$ is observed and we are interested in testing

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta \neq \theta_0.$$

The answer is to have a prior that has a point mass at the value θ_0 with prior weight π_0 and a spread distribution $\xi(\theta)$ which is the prior under H_1 that has prior weight $\pi_1 = 1 - \pi_0$. Thus, the prior is

$$\pi(\theta) = \pi_0 \delta_{\theta_0} + \pi_1 \xi(\theta),$$

where δ_{θ_0} is Dirac mass at θ_0.

The marginal density for X is

$$m(x) = \pi_0 f(x|\theta_0) + \pi_1 \int f(x|\theta)\xi(\theta)d\theta = \pi_0 f(x|\theta_0) + \pi_1 m_1(x).$$

The posterior probability of $\theta = \theta_0$ is

$$\pi(\theta_0|x) = f(x|\theta_0)\pi_0/m(x) = \frac{f(x|\theta_0)\pi_0}{\pi_0 f(x|\theta_0) + \pi_1 m_1(x)} = \left[1 + \frac{\pi_1}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

Exercise. Show that the Bayes factor is $B^{\pi}_{01}(x) = \frac{f(x|\theta_0)}{m_1(x)}$.

Exercise. Show that $p_0 = \pi(\theta_0|x) \geq \left[1 + \frac{\pi_1}{\pi_0} \cdot \frac{r(x)}{f(x|\theta_0)}\right]^{-1}$, where $r(x) = \sup_{\theta \neq \theta_0} f(x|\theta)$. Usually, $r(x) = f(x|\hat{\theta}_{\text{MLE}})$, where $\hat{\theta}_{mle}$ is MLE estimator of θ. The Bayes factor $B^{\pi}_{01}(x)$ is bounded from below,

$$B^{\pi}_{01}(x) \geq \frac{f(x|\theta_0)}{r(x)} = \frac{f(x|\theta_0)}{f(x|\hat{\theta}_{\text{MLE}})}.$$

Application in Wavelet Thresholding. As we have discussed, testing a precise hypothesis in Bayesian fashion requires a prior that has a point mass component. Otherwise, the testing is impossible since any continuous prior density will give the prior (and hence the posterior) probability of 0 to the precise hypothesis. The following application of testing of precise hypotheses is in the area of wavelet shrinkage, for more details see Vidakovic (1998).

We start with the likelihood (it can be, for example, a marginal likelihood for $[d|\theta, \sigma^2] \sim \mathcal{N}(\theta, \sigma^2)$ when the parameter σ^2 is integrated out)

$$d|\theta \sim f(d|\theta),$$

where d and θ are observed noisy wavelet coefficient and the coefficient corresponding to a signal part. The model is $d = \theta + \epsilon$, $\epsilon \sim f$, and the goal is to estimate θ. As traditional in wavelet thresholding, the estimators are of the type *keep-or-kill* the observed wavelet coefficient.

After observing d, we test the hypothesis $H_0 : \theta = 0$, versus $H_1 : \theta \neq 0$. If the hypothesis H_0 is rejected, d is kept, Otherwise, it is replaced by 0. Let the prior on signal part θ be

$$\theta \sim \pi(\theta) = \pi_0\delta_0 + \pi_1\xi(\theta),$$

where $\pi_0 + \pi_1 = 1,\ \delta_0$ is a point mass at 0, and $\xi(\theta)$ is a prior that describes distribution of θ when H_0 is false.

We threshold a wavelet coefficient d by the following rule:

$$\hat{\theta} = d\,\mathbf{1}\left(P^{\theta|d}(\theta = 0|d) < \frac{1}{2}\right),$$

where

$$P(H_0|d) = \left[1 + \frac{\pi_1}{\pi_0}\,\frac{1}{B^{\pi}_{01}(d)}\right]^{-1},$$

is the posterior probability of the null hypothesis, and

$$B^{\pi}_{01}(d) = \frac{f(d|0)}{\int_{\theta\neq 0} f(d|\theta)\xi(\theta)d\theta}$$

is the Bayes factor in favor of H_0. The constant $\frac{1}{2}$ in the thresholding rule can be replaced by more general $K_1/(K_0 + K_1)$ if the loss function is $K_0 - K_1$.

Exercise. In the Example in Handout 6 we obtained that the marginal likelihood for d was

$$d|\theta \sim \mathcal{DE}(\theta, \frac{1}{\sqrt{2\mu}})$$

Assume that the prior on θ is

$$\pi(\theta) = \pi_0\delta_0 + \pi_1\xi(\theta).$$

Show that the coefficient d will be "thresholded" if the posterior probability of H_0 is exceeds the posterior probability of H_1, i.e.,

$$\frac{\pi_0 e^{-c|d|}}{\pi_0 e^{-c|d|} + \pi_1(\Pi_1(c) + \Pi_2(c))} \geq \frac{1}{2},$$

where Π_1 and Π_2 are one-sided Laplace transformations of $\xi(\theta - d)$ and $\xi(\theta + d)$ and $c = \sqrt{2\mu}$.

TO BE ADDED!

References

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer Verlag.

[2] Berger, J. and Delampady, M. (1987). Testing precise hypothesis, *Statistical Science*, 3, 317–352.

[3] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.

[4] Bickel, P. J. and Blackwell, D. (1967). A note on Bayes estimates. *Annals of Mathematical Statistics*, **38**, 1907-1911.

[5] Blackwell, D. and Girshick, M. A. (1954). *Theory of games and statistical decisions.* Wiley, New York.

[6] Lehmann, E. L. (1951). A general concept of unbiasedness. *Annals of Mathematical Statistics,* **22**, 587-592.

[7] Noorbaloochi, S. and Meeden, G. (1983). Unbiasedness as the dual of being Bayes. *Journal of the American Statistical Association,* **78**, 619-623.

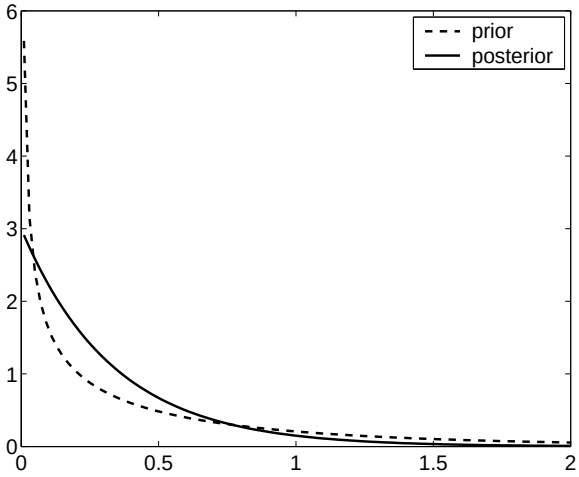
[8] OHagan, A. (1994) *Kendals Advanced Theory of Statistics, Volume 2b, Bayesian Inference,* Halsted Press, New York.

[9] Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association,* **93,** 173–179.

Exercises

1. Single Observation to Estimate Precision. Suppose you observed $X = -2$ from the population distributed as $N(0, 1/\theta)$ and wish to estimate the parameter θ. [θ is the reciprocal of variance σ^2 and is called the *precision parameter*].

A classical estimator of θ exists (say MLE), but one may be disturbed to estimate $1/\sigma^2$ by a single observation. As Bayesian, you believe that the prior on θ is Gamma $\mathcal{G}amma(1/2, 3)$. [3]



• What is the MLE of θ.

• Find the posterior distribution and the Bayes estimator of θ. If the prior on θ is $\mathcal{G}amma(\alpha, \beta)$, represent the Bayes estimator as weighted average (sum of weigths = 1) of the prior mean and the MLE.

• Find 95% HPD Credible set for θ.

• Test the hypothesis $H_0 : \theta \leq 1/4$ v.s. $H_1 : \theta > 1/4$.

[3] $Z \sim \mathcal{G}amma(\alpha, \beta)$ has a density $\frac{\beta^\alpha z^{\alpha-1}}{\Gamma(\alpha)} \exp\{-\beta z\}$. $EZ = \frac{\alpha}{\beta}$ and $VarZ = \frac{\alpha}{\beta^2}$.