Bayes versus Frequentist

This lecture combines three blog posts that I wrote on this topic.

1 Adventures in FlatLand: Stone's Paradox

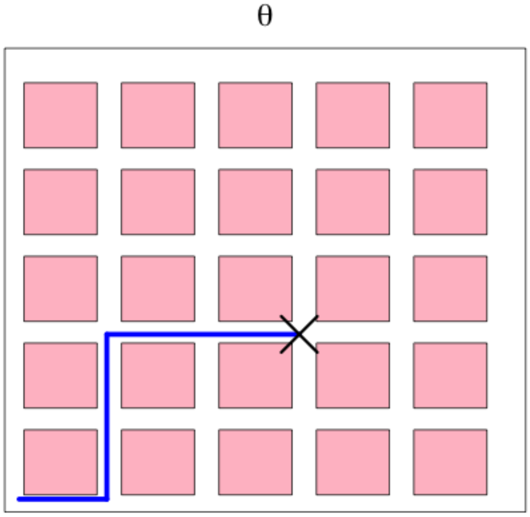
Mervyn Stone is Emeritus Professor at University College London. He is famous for his deep work on Bayesian inference as well as pioneering work on cross-validation, coordinate-free multivariate analysis, as well as many other topics.

Today I want to discuss a famous example of his, described in Stone (1970, 1976, 1982). In technical jargon, he shows that "a finitely additive measure on the free group with two generators is nonconglomerable." In English: even for a simple problem with a discrete parameters space, flat priors can lead to surprises. Fortunately, we don't need to know anything about free groups to understand this example.

1.1 Hunting For a Treasure In Flatland

I wonder randomly in a two dimensional grid-world. I drag an elastic string with me. The string is taut: if I back up, the string leaves no slack. I can only move in four directions: North, South, West, East.

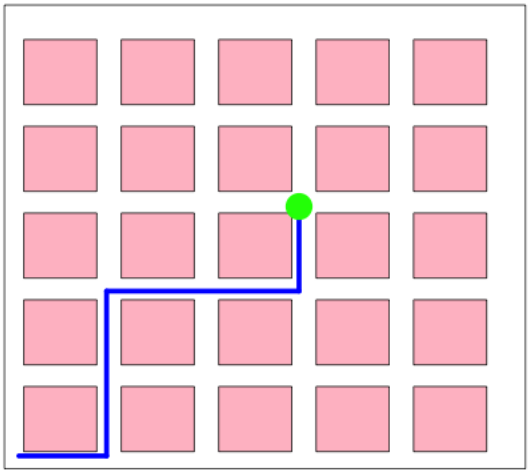
I wander around for a while then I stop and bury a treasure. Call this path θ. Here is an example:



My Path. The X is the treasure.

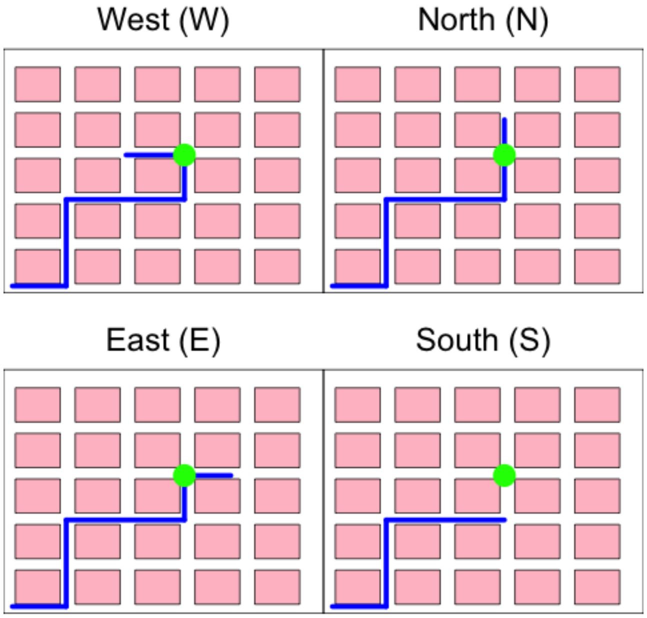
Now I take one more random step. Each direction has equal probability. Call this path x. So it might look like this:

x



One more random step.

Two people, Bob (a Bayesian) and Carla (a classical statistician) want to find the treasure. There are only four possible paths that could have yielded x, namely:



West (W)

North (N)

East (E)

South (S)

Let us call these four paths N, S, W, E. The likelihood is the same for each of these. That is, $p(x|\theta) = 1/4$ for $\theta \in \{N, S, W, E\}$. Suppose Bob uses a flat prior. Since the likelihood is also flat, his posterior is

$$P(\theta = N|x) = P(\theta = S|x) = P(\theta = W|x) = P(\theta = E|x) = \frac{1}{4}.$$

Let B be the three paths that extend x. In this example, $B = \{N, W, E\}$. Then $P(\theta \in B|x) = 3/4$.

Now Carla is very confident and selects a confidence set with only one path, namely, the path that shortens x. In other words, Carla's confidence set is $C = B^c$.

Notice the following strange thing: no matter what θ is, Carla gets the treasure with probability 3/4 while Bob gets the treasure with probability 1/4. That is, $P(\theta \in B|x) = 3/4$ but the coverage of B is 1/4. The coverage of C is 3/4.

Here is quote from Stone (1976): (except that I changed his B and C to Bob and Carla):

" ... it is clear that when Bob and Carla repeatedly engage in this treasure hunt, Bob will find that his posterior probability assignment becomes increasingly discrepant with his proportion of wins and that Carla is, somehow, doing better than [s]he ought. However, there is no message ... that will allow Bob to escape from his Promethean situation; he cannot learn from his experience because each hunt is independent of the other."

Stone is not criticizing Bayes (as far I can tell). He is just discussing the effect of using a flat prior.

1.2 More Trouble For Bob

Let A be the event that the final step reduced the length of the string. Using the posterior above, we see that Bob finds that $P(A|x) = 3/4$ for each x. Since this holds for each x, Bob deduces that $P(A) = 3/4$. On the other hand, Bob notes that $P(A|\theta) = 1/4$ for every θ. Hence, $P(A) = 1/4$.

Bob has just proved that $3/4 = 1/4$.

1.3 The Source of The Problem

The apparent contradiction stems from the fact that the prior is improper. Technically this is an example of the non-conglomerability of finitely additive measures. For a rigorous explanation of why this happens you should read Stone's papers. Here is an abbreviated explanation, from Kass and Wasserman (1996, Section 4.2.1).

Let π denotes Bob's improper flat prior and let $\pi(\theta|x)$ denote his posterior distribution. Let π_p denote the prior that is uniform on the set of all paths of length p. This is of course a proper prior. For any fixed x, $\pi_p(A|x) \to 3/4$ as $p \to \infty$. So Bob can claim that his posterior distribution is a limit of well-defined posterior distributions. However, we need to look at this more closely. Let $m_p(x) = \sum_\theta f(x|\theta)\pi_p(\theta)$ be the marginal of x induced by π_p. Let X_p denote all x's of length p or $p+1$. When $x \in X_p$, $\pi_p(\theta|x)$ is a poor approximation to $\pi(\theta|x)$ since the former is concentrated on a single point while the latter is concentrated on four points. In fact, the total variation distance between $\pi_p(\theta|x)$ and $\pi(\theta|x)$ is $3/4$ for $x \in X_p$. (Recall that the total variation distance between two probability measures P and Q is $d(P,Q) = \sup_A |P(A) - Q(A)|$.) Furthermore, X_p is a set with high probability: $m_p(X_p) \to 2/3$ as $p \to \infty$.

While $\pi_p(\theta|x)$ converges to $\pi(\theta|x)$ as $p \to \infty$ for any fixed x, they are not close with high probability.

This problem disappears if you use a proper prior.

1.4 The Four Sided Die

Here is another description of the problem. Consider a four sided die whose sides are labeled with the symbols $\{a, b, a^{-1}, b^{-1}\}$. We roll the die several times and we record the label on the lowermost face (there is a no uppermost face on a four-sided die). A typical outcome might look like this string of symbols:

$$a \;\; a \; b \; a^{-1} \; b \; b^{-1} \; b \; a \; a^{-1} \; b$$

Now we apply an annihilation rule. If a and a^{-1} appear next to each other, we eliminate these two symbols. Similarly, if b and b^{-1} appear next to each other, we eliminate those two symbols. So the sequence above gets reduced to:

$$a \;\; a \; b \; a^{-1} \; b \; b$$

Let us denote the resulting string of symbols, after removing annihilations, by θ. Now we toss the die one more time. We add this last symbol to θ and we apply the annihilation rule once more. This results in a string which we will denote by x.

You get to see x and you want to infer θ.

Having observed x, there are four possible values of θ and each has the same likelihood. For example, suppose $x = (a, a)$. Then θ has to be one of the following:

$$(a), \quad (a\,a\,a), \quad (a\,a\,b^{-1}), \quad (a\,a\,b)$$

The likelihood function is constant over these four values.

Suppose we use a flat prior on θ. Then the posterior is uniform on these four possibilities. Let $C = C(x)$ denote the three values of θ that are longer than x. Then the posterior satisfies

$$P(\theta \in C|x) = 3/4.$$

Thus $C(x)$ is a 75 percent posterior confidence set.

However, the frequentist coverage of $C(x)$ is 1/4. To see this, fix any θ. Now note that $C(x)$ contains θ if and only if θ concatenated with x is smaller than θ. This happens only if the last symbol is annihilated, which occurs with probability 1/4.

1.5 Likelihood

Another consequence of Stone's example is that, in my opinion, it shows that the Likelihood Principle is bogus. According to the likelihood principle, the observed likelihood function contains all the useful information in the data. In this example, the likelihood does not distinguish the four possible parameter values.

The direction of the string from the current position — which does not affect the likelihood — has lots of information.

1.6 Proper Priors

If you want to have some fun, try coming up with proper priors on the set of paths. Then simulate the example, find the posterior and try to find the treasure. If you try this, I'd be interested to hear about the results.

Another question this example raises is: should one use improper priors? Flat priors that do not have a finite sum can be interpreted as finitely additive priors. The father of Bayesian inference, Bruno DeFinetti, was adamant in rejecting the axiom of countable additivity. He thought flat priors like Bob's were fine.

It seems to me that in modern Bayesian inference, there is not universal agreement on whether flat priors are evil or not. But in this example, I think that most statisticians would reject Bob's flat prior-based Bayesian inference.

1.7 Conclusion

I have always found this example to be interesting because it seems very simple and, at least at first, one doesn't expect there to be a problem with using a flat prior. Technically the problems arise because there is group structure and the group is not amenable. Hidden beneath this seemingly simple example is some rather deep group theory.

Many of Stone's papers are gems. They are not easy reading (with the exception of the 1976 paper) but they are worth the effort.

2 Robins and Wasserman Respond to a Nobel Prize Winner

This section written by James Robins and Larry Wasserman

Chris Sims is a Nobel prize winning economist who is well known for his work on macroeconomics, Bayesian statistics, vector autoregressions among other things. One of us (LW) had the good fortune to meet Chris at a conference and can attest that he is also a very nice guy.

Chris has a paper called *On an An Example of Larry Wasserman.* This post is a response to Chris' paper.

The example in question is actually due to Robins and Ritov (1997). A simplified version appeared in Wasserman (2004) and Robins and Wasserman (2000). The example is related to ideas from the foundations of survey sampling (Basu 1969, Godambe and Thompson 1976) and also to ancillarity paradoxes (Brown 1990, Foster and George 1996).

2.1 The Model

Here is (a version of) the example. Consider iid random variables

$$(X_1, Y_1, R_1), \ldots, (X_n, Y_n, R_n).$$

The random variables take values as follows:

$$X_i \in [0,1]^d, \quad Y_i \in \{0,1\}, \quad R_i \in \{0,1\}.$$

Think of d as being very, very large. For example, $d = 100,000$ and $n = 1,000$.

The idea is this: we observe X_i. Then we flip a biased coin R_i. If $R_i = 1$ then you get to see Y_i. If $R_i = 0$ then you don't get to see Y_i. The goal is to estimate

$$\psi = P(Y_i = 1).$$

Here are the details. The distribution takes the form

$$p(x, y, r) = p_X(x) p_{Y|X}(y|x) p_{R|X}(r|x).$$

Note that Y and R are independent, given X. For simplicity, we will take $p(x)$ to be uniform on $[0,1]^d$. Next, let

$$\theta(x) \equiv p_{Y|X}(1|x) = P(Y = 1|X = x)$$

where $\theta(x)$ is a function. That is, $\theta : [0,1]^d \to [0,1]$. Of course,

$$p_{Y|X}(0|x) = P(Y = 0|X = x) = 1 - \theta(x).$$

Similarly, let

$$\pi(x) \equiv p_{R|X}(1|x) = P(R = 1|X = x)$$

where $\pi(x)$ is a function. That is, $\pi : [0,1]^d \to [0,1]$. Of course,

$$p_{R|X}(0|x) = P(R = 0|X = x) = 1 - \pi(x).$$

The function π is **known.** We construct it. Remember that $\pi(x) = P(R = 1|X = x)$ is the probability that we get to observe Y given that $X = x$. Think of Y as something that is expensive to measure. We don't always want to measure it. So we make a random decision about whether to measure it. And we let the probability of measuring Y be a function $\pi(x)$ of x. And we get to construct this function.

Let $\delta > 0$ be a known, small, positive number. We will assume that

$$\pi(x) \geq \delta$$

for all x.

The only thing in the the model we don't know is the function $\theta(x)$. Again, we will assume that

$$\delta \leq \theta(x) \leq 1 - \delta.$$

Let Θ denote all measurable functions on $[0,1]^d$ that satisfy the above conditions. The parameter space is the set of functions Θ.

Let $\mathcal{P}$ be the set of joint distributions of the form

$$p(x)\,\pi(x)^r(1-\pi(x))^{1-r}\,\theta(x)^y(1-\theta(x))^{1-y}$$

where $p(x) = 1$, and $\pi(\cdot)$ and $\theta(\cdot)$ satisfy the conditions above. So far, we are considering the sub-model $\mathcal{P}_\pi$ in which π is known.

The parameter of interest is $\psi = P(Y = 1)$. We can write this as

$$\psi = P(Y = 1) = \int_{[0,1]^d} P(Y = 1|X = x)p(x)dx = \int_{[0,1]^d} \theta(x)dx.$$

Hence, ψ is a function of θ. If we know $\theta(\cdot)$ then we can compute ψ.

2.2 Frequentist Analysis

The usual frequentist estimator is the Horwitz-Thompson estimator

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\pi(X_i)}.$$

It is easy to verify that $\widehat{\psi}$ is unbiased and consistent. Furthermore, $\widehat{\psi}-\psi = O_P(n^{-\frac{1}{2}})$. In fact, let us define

$$I_n = [\widehat{\psi} - \epsilon_n,\ \widehat{\psi} + \epsilon_n]$$

where

$$\epsilon_n = \sqrt{\frac{1}{2n\delta^2}\log\left(\frac{2}{\alpha}\right)}.$$

It follows from Hoeffding's inequality that

$$\sup_{P\in\mathcal{P}_\pi} P(\psi \in I_n) \geq 1-\alpha$$

Thus we have a finite sample, $1-\alpha$ confidence interval with length $O(1/\sqrt{n})$.

Remark: We are mentioning the Horwitz-Thompson estimator because it is simple. In practice, it has three deficiencies:

1. It may exceed 1.
2. It ignores data on the multivariate vector X except for the one dimensional summary $\pi(X)$.
3. It can be very inefficient.

These problems are remedied by using an improved version of the Horwitz-Thompson estimator. One choice is the so-called *locally semiparametric efficient regression estimator* (Scharfstein et al., 1999):

$$\widehat{\psi} = \int \text{expit}\left(\sum_{m=1}^{k} \widehat{\eta}_m \phi_m(x) + \frac{\widehat{\omega}}{\pi(x)}\right) dx$$

where $\text{expit}(a) = e^a/(1+e^a)$, $\phi_m(x)$ are basis functions, and $\widehat{\eta}_1, \ldots, \widehat{\eta}_k, \widehat{\omega}$ are the mle's (among subjects with $R_i = 1$) in the model

$$\log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \sum_{m=1}^{k} \eta_m \phi_m(x) + \frac{\omega}{\pi(x)}.$$

Here k can increase slowly with n. Recently even more efficient estimators have been derived. Rotnitzky et al (2012) provides a review. In the rest of this post, when we refer to the Horwitz-Thompson estimator, the reader should think "improved Horwitz-Thompson estimator." **End Remark.**

2.3 Bayesian Analysis

To do a Bayesian analysis, we put some prior W on Θ. Next we compute the likelihood function. The likelihood for one observation takes the form $p(x)p(r|x)p(y|x)^r$. The reason for having r in the exponent is that, if $r = 0$, then y is not observed so the $p(y|x)$ gets left out. The likelihood for n observations is

$$\prod_{i=1}^{n} p(X_i)p(R_i|X_i)p(Y_i|X_i)^{R_i} = \prod_i \pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}\,\theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

where we used the fact that $p(x) = 1$. But remember, $\pi(x)$ is known. In other words, $\pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}$ is known. So, the likelihood is

$$\mathcal{L}(\theta) \propto \prod_i \theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

Combining this likelihood with the prior W creates a posterior distribution on Θ which we will denote by W_n. Since the parameter of interest ψ is a function of θ, the posterior W_n for θ defines a posterior distribution for ψ.

Now comes the interesting part. The likelihood has essentially no information in it.

To see that the likelihood has no information, consider a simpler case where $\theta(x)$ is a function on $[0, 1]$. Now discretize the interval into many small bins. Let B be the number of bins. We can then replace the function θ with a high-dimensional vector $\theta = (\theta_1, \ldots, \theta_B)$. With $n < B$, most bins are empty. The data contain no information for most of the θ_j's. (You might wonder about the effect of putting a smoothness assumption on $\theta(\cdot)$. We'll discuss this in Section 4.)

We should point out that if $\pi(x) = 1/2$ for all x, then Ericson (1969) showed that a certain exchangeable prior gives a posterior that, like the Horwitz-Thompson estimator, converges at rate $O(n^{-1/2})$. However we are interested in the case where $\pi(x)$ is a complex function of x; then the posterior will fail to concentrate around the true value of ψ. On the other hand, a flexible nonparametric prior will have a posterior essentially equal to the prior and, thus, not concentrate around ψ, whenever the prior W does not depend on the the known function $\pi(\cdot)$. Indeed, we have the following theorem from Robins and Ritov (1997):

Theorem. (Robins and Ritov 1997). Any estimator that is not a function of $\pi(\cdot)$ cannot be uniformly consistent.

This means that, at no finite sample size, will an estimator $\widehat{\psi}$ that is not a function of π be close to ψ for all distributions in $\mathcal{P}$. In fact, the theorem holds for a neighborhood around every pair (π, θ). Uniformity is important because it links asymptotic behavior to finite sample behavior. But when π is known and is used in the estimator (as in the Horwitz-Thompson estimator and its improved versions) we can have uniform consistency.

Note that a Bayesian will ignore π since the $\pi(X_i)'s$ are just constants in the likelihood. There is an exception: the Bayesian can make the posterior be a function of π by choosing a prior W that makes $\theta(\cdot)$ depend on $\pi(\cdot)$. But this seems very forced. Indeed, Robins and Ritov showed that, under certain conditions, any true subjective Bayesian prior W must be independent of $\pi(\cdot)$. Specifically, they showed that once a subjective Bayesian queries the randomizer (who selected π) about the randomizer's reasoned opinions concerning $\theta(\cdot)$ (but not $\pi(\cdot)$) the Bayesian will have independent priors. We note that a Bayesian can have independent priors even when he believes with probabilty 1 that $\pi(\cdot)$ and $\theta(\cdot)$ are positively correlated as functions of x i.e. $\int \theta(x)\pi(x)\,dx > \int \theta(x)\,dx \int \pi(x)\,dx$. Having independent priors only means that learning $\pi(\cdot)$ will not change one's beliefs about $\theta(\cdot)$.

So far, so good. As far as we know, Chris agrees with everything up to this point.

2.4 Some Bayesian Responses

Chris goes on to raise alternative Bayesian approaches.

The first is to define

$$Z_i = \frac{R_i Y_i}{\pi(X_i)}.$$

Note that $Z_i \in \{0\} \cup [1, \infty)$. Now we ignore (throw away) the original data. Chris shows that we can then construct a model for Z_i which results in a posterior for ψ that mimics the Horwitz-Thompson estimator. We'll comment on this below, but note two strange things. First, it is odd for a Bayesian to throw away data. Second, the new data are a function of $\pi(X_i)$ which forces the posterior to be a function of π. But as we noted earlier, when θ and π are a priori independent, the $\pi(X_i)'s$ do not appear in the posterior since they are known constants that drop out of the likelihood.

A second approach (not mentioned explicitly by Chris) which is related to the above idea, is to construct a prior W that depends on the known function π. It can be shown that if the prior is chosen just right then again the posterior for ψ mimics the (improved) Horwitz-Thompson estimator.

Lastly, Chris notes that the posterior contains no information because we have not enforced any smoothness on $\theta(x)$. Without smoothness, knowing $\theta(x)$ does not tell you anything about $\theta(x+\epsilon)$ (assuming the prior W does not depend on π).

This is true and better inferences would obtain if we used a prior that enforced smoothness. But this argument falls apart when d is large. (In fairness to Chris, he was referring to the version from Wasserman (2004) which did not invoke high dimensions.) When d is large, forcing $\theta(x)$ to be smooth does not help unless you make it very, very, very smooth. The larger d is, the more smoothness you need to get borrowing of information across different values of $\theta(x)$. But this introduces a huge bias which precludes uniform consistency.

2.5 Response to the Response

We have seen that response 3 (add smoothness conditions in the prior) doesn't work. What about response 1 and response 2? We agree that these work, in the sense that the Bayes answer has good frequentist behavior by mimicking the (improved) Horwitz-Thompson estimator.

But this is a Pyrrhic victory. If we manipulate the data to get a posterior that mimics the frequentist answer, is this really a success for Bayesian inference? Is it really Bayesian inference at all? Similarly, if we choose a carefully constructed prior just to mimic a frequentist answer, is it really Bayesian inference?

We call Bayesian inference which is carefully manipulated to force an answer with good frequentist behavior, **frequentist pursuit**. There is nothing wrong with it, but why bother?

If you want good frequentist properties just use the frequentist estimator. If you want to be a Bayesian, be a Bayesian but accept the fact that, in this example, your posterior will fail to concentrate around the true value.

2.6 Summary

In summary, we agree with Chris' analysis. But his fix is just frequentist pursuit; it is Bayesian analysis with unnatural manipulations aimed only at forcing the Bayesian answer to be the frequentist answer. This seems to us to be an admission that Bayes fails in this example.

3 Freedman's Theorem

In this post I want to review an interesting result by David Freedman (Annals of Mathematical Statistics, Volume 36, Number 2 (1965), 454-456) available at projecteuclid.org.

The result gets very little attention. Most researchers in statistics and machine learning seem to be unaware of the result. The result says that, "almost all" Bayesian posterior distributions are inconsistent, in a sense we'll make precise below. The math is uncontroversial but, as you might imagine, the intepretation of the result is likely to be controversial.

Actually, I had planned to avoid "Bayesian versus frequentist" stuff on this blog because it has been argued to death. But this particular result is so neat and clean that I couldn't resist. I will, however, resist drawing any philosophical conclusions from the result. I will merely tell you what the result is. Don't shoot the messenger!

The paper is very short, barely more than two pages. My summary will be even shorter. (I'll use slightly different notation.)

Let $X_1, \ldots, X_n$ be an iid sample from a distribution P on the natural numbers $I = \{1, 2, 3, \ldots, \}$. Let $\mathcal{P}$ be the set of all such distributions. We endow $\mathcal{P}$ with the weak* topology. Hence, $P_n \to P$ iff $P_n(i) \to P(i)$ for all i.

Let μ denote a prior distribution on $\mathcal{P}$. (More precisely, a prior on an appropriate σ-field.) Let Π be all priors. Again, we endow the set with the weak* topology. Thus $\mu_n \to \mu$ iff $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous, real functions f.

Let μ_n be the posterior corresponding to the prior μ after n observations. We will say that the pair (P, μ) is consistent if

$$P^\infty(\lim_{n\to\infty} \mu_n = \delta_P) = 1$$

where P^∞ is the product measure corresponding to P and δ_P is a point mass at P.

Now we need to recall some topology. A set is nowhere dense if its closure has an empty interior. A set is meager (or first category) if it is a countable union of nowehere dense sets. Meager sets are small; think of a meager set as the topological version of a null set in measure theory.

Freedman's theorem is: the sets of consistent pairs (P, μ) is meager.

This means that, in a topological sense, consistency is rare for Bayesian procedures. From this result, it can also be shown that most pairs of priors lead to inferences that disagree. (The agreeing pairs are meager.) Or as Freedman says in his paper:

" ... it is easy to prove that for essentially any pair of Bayesians, each thinks the other is crazy."

As a postscript, let me add that David Freedman, who died in 2008, was a statistician at Berkeley. He was an impressive person whose work spanned from the very theoretical to the very applied. He was a bit of a curmudgeon, which perhaps lessened his influence a little. But he was a deep thinker with a healthy skepticism about the limits of statistical models, and I encourage any students reading this blog to seek out his work.

4 References

Basu, D. (1969). Role of the Sufficiency and Likelihood Principles in Sample Survey Theory. *Sankya*, 31, 441-454.

Brown, L.D. (1990). An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics*, 18, 471-493.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations.

Journal of the Royal Statistical Society. Series B, 195-233.

Foster, D.P. and George, E.I. (1996). A simple ancillarity paradox. *Scandinavian journal of statistics*, 233-242.

Godambe, V. P., and Thompson, M. E. (1976), Philosophy of Survey-Sampling Practice. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, eds. W.L.Harper and A.Hooker, Dordrecht: Reidel.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.

Robins, J.M. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models. *Statistics in Medicine*, 16, 285–319.

Robins, J. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: a review of some foundational concepts. *Journal of the American Statistical Association*, 95, 1340-1346.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J.M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 439-456.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 1096-1120.

Sims, Christopher. On An Example of Larry Wasserman. Available at: http://www.princeton.edu/ sims/.

Stone, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics*, 41, 1349-1353,

Stone, M. (1976). Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71, 114-116.

Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. *Studies in Logic and the Foundations of Mathematics*, 104, 413-426.

Wasserman, L. (2004). *All of Statistics: a Concise Course in Statistical Inference.* Springer Verlag.