The Progress on the Foundations of Bayesian - Frequentist Unification

Jim Berger

Duke University

B.G. Greenberg Distinguished Lectures

Department of Biostatistics

University of North Carolina at Chapel Hill

May 12, 2016

Outline

- Status of the unification for estimation problems
- Status of the unification for testing problems
  - Unification via conditional frequentist testing and rejection odds
  - Situations of supposed conflict in testing
    - Sequential testing (can be unified)
    - The Jeffreys-Lindley Paradox (not a conflict)
    - Sequential endpoint testing (an unavoidable conflict)
- Weaker unification: establishing that a Bayes procedure has satisfactory frequentist properties

Why should we care about this?

- Without an agreed upon foundation, statistics seems to outsiders like just a collection of methods, not a serious science.
- We cannot resolve many problems involving the misuse of statistics because we cannot agree on their resolution.
  Example: Almost everyone in the statistics profession decries the way p-values are used, yet the abuse continues unabated because we offer a confusing plethora of alternatives.
- In some sense, we only "have it right" when the answer is the same when the data is examined from different perspectives.
- Regulatory statistics (e.g., FDA) would be much easier if there was statistical agreement.

My view: I approach this as both a Bayesian and a frequentist, and try to find the common ground.

Reviewing the situation for estimation problems

A *small sample* estimation example: joint Bayesian-frequentist confidence intervals for medical diagnosis (Berger and Mossman, 2001)

The Medical Problem:

- Within a population, $p_0 = Pr(\text{Disease } D)$.
- A diagnosic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = Pr(P \mid \text{patient has } D)$.
- $p_2 = Pr(P \mid \text{patient does not have } D)$.

It follows from Bayes theorem that

$$\theta = Pr(D \mid P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}.$$

The Statistical Problem: The p_i are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical studies), find a $100(1-\alpha)\%$ confidence set for θ.

Suggested Solution: Assign p_i the Jeffreys-rule prior

$$\pi(p_i) \propto p_i^{-1/2}(1-p_i)^{-1/2}$$

(more or less the same answer would arise from the uniform prior $\pi(p_i) = 1$). By Bayes theorem, the posterior distribution of p_i given the data, x_i, is

$$\pi(p_i \mid x_i) = \frac{p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}}{\int p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}\, dp_i},$$

which is the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distribution.

Finally, compute the desired confidence set (formally, the $100(1-\alpha)\%$ equal-tailed posterior credible set) by simulation:

- drawing random p_i from the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distributions, $i = 0, 1, 2$;
- computing the associated $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$;
- repeating this process $10,000$ times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated θ to form the desired confidence limits.

$n_0 = n_1 = n_2$	(x_0, x_1, x_2)	95% confidence interval
20	(2,18,2)	(0.107, 0.872)
20	(10,18,0)	(0.857, 1.000)
80	(20,60,20)	(0.346, 0.658)
80	(40,72,8)	(0.808, 0.952)

Table 1: The 95% equal-tailed posterior credible interval for $\theta = p_0 p_1/[p_0 p_1 + (1 - p_0)p_2]$, for various values of the n_i and x_i.

The actual goal of the scientist was to find *frequentist* confidence intervals for

$$\theta = Pr(D \mid P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2} .$$

Consider the frequentist percentage of the time that the 95% Bayesian credible sets miss on the left and on the right (ideal would be 2.5% each) for the indicated parameter values when $n_0 = n_1 = n_2 = 20$.

(p_0, p_1, p_2)	O-Bayes	Log Odds	Gart-Nam	Delta
$(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$	2.86,2.71	1.53,1.55	2.77,2.57	2.68,2.45
$(\frac{1}{10}, \frac{9}{10}, \frac{1}{10})$	2.23,2.47	0.17,0.03	1.58,2.14	0.83,0.41
$(\frac{1}{2}, \frac{9}{10}, \frac{1}{10})$	2.81,2.40	0.04,4.40	2.40,2.12	1.25,1.91

Conclusion: The objective Bayes confidence sets had superior performance as frequentist confidence sets than did any of the confidence sets derived using non-Bayesian methods. (The O-Bayes sets were usually also smaller.)

The conditioning issue that must be addressed in frequentist foundations:

Basic question for a frequentist: What is the sequence of possible data over which the frequentist averages are taken? (Fisher:"relevant subset;" Lehmann: "frame of reference;")

When the average is over all the possible data, we call that the *unconditional frequentist* average.

Artificial example: Observe independent X_1 and X_2, where

$$X_i = \begin{cases} \theta + 1 \text{ with probability } 1/2 \\ \theta - 1 \text{ with probability } 1/2. \end{cases}$$

Consider the confidence set (a point) for θ

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2. \end{cases}$$

Unconditional coverage:

$$P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75.$$

This is correct but silly: if $x_1 \neq x_2$, we know $C(x_1, x_2) = \theta$;
if $x_1 = x_2$, $C(X_1, X_2)$ equals θ only with probability $1/2$.

The correct conditional answer:

- Define the statistic $S = |X_1 - X_2|$, measuring the "strength of evidence" in the data (either 0 or 2). This is the *maximal ancillary* statistic.
- Compute frequentist coverage conditional on the strength of evidence ancillary statistic S:

$$\begin{aligned} P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, s = 2) &= 1 \\ P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, s = 0) &= \tfrac{1}{2}. \end{aligned}$$

Note: This answer would have arisen from *objective Bayes* analysis with the objective prior $\pi(\theta) = 1$.

This is part of a much bigger picture:

Theorem 1 *(Stein and others): If the statistical model is invariant under a group operation $g \in G$, with G being isomorphic to a continuous parameter space, then the right-Haar prior (guaranteed to exist) yields O-Bayes credible intervals that are exact conditional confidence intervals, conditional on a maximal invariant statistic S, for any parameter for which the Bayesian credible sets are invariant, i.e., for which $C(g \circ x) = \bar{g} \circ C(x)$.*

- Most textbook statistical problems satisfy the conditions of this theorem.
- Many non-textbook problems satisfy the conditions of this theorem:
  - Augie Kong story: Mapping genes for complex traits based on affected half-sib data X, with the goal being to find $\theta =$ *the true location of a susceptibility gene.*
  - As θ is a location parameter, use $\pi(\theta) = 1$.

The conclusions for estimation type problems

- For most of the standard estimation problems in statistics, optimal objective Bayesian answers are identical to the frequentist answers (although they might be interpreted differently).
- For difficult estimation problems, it is often the case that the best frequentist answers are obtained through objective Bayesian analysis.
  - Using the Bayesian methodology assures that the conditioning problem is addressed.
  - Objective Bayesian answers typically have the best small-sample size frequentist performance.
- Thus unification of frequentist and Bayesian statistics in estimation problems is possible, if objective Bayesian methodology is employed.

Towards unifying the Bayesian and frequentist approaches to testing

Possible unifications of frequentist and Bayesian testing

- Conditional frequentist testing (Kiefer (1976), Brown (1977), Berger, Brown and Wolpert (1996),...)
  - Find a conditioning statistic S measuring the 'strength of evidence' in the data, and
  - report the conditional Type I error probability $\alpha(S) = Pr(\mathcal{R} \mid S, H_0)$.
  - This is fully frequentist and, for a certain sensible choice of S, $\alpha(S)$ typically equals the objective Bayesian posterior probability of H_0.
- Odds of true rejections to false rejections
  - Discussed in the first talk.

Situations of supposed conflict in testing

Interim or sequential analysis

In *interim or sequential analysis*, one periodically looks at the accumulated data during a study, with the option of stopping the study and drawing a conclusion at each look.

- Unconditional frequentists increase the error probability with each look at the data (since each of the analysis stages increases the probability of having an incorrect rejection).
- Bayesians do not adjust the error probability (the *stopping rule principle*).
- Conditional frequentists and users of Bayes factors (post-experimental rejection odds) also do not adjust, so agreement is possible here.

An example of sequential testing with rejection odds

(Berger, Boukai and Wang, 1998)

Data: $X_1, X_2, \ldots$ are i.i.d. $N(\theta, \sigma^2)$, θ and σ^2 unknown, and arrive sequentially.

To test: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

A default Bayes test (Jeffreys, 1961):

Prior distribution : $Pr(H_0) = Pr(H_1) = 1/2$

Under H_0, prior on σ is $\pi_0(\sigma) = 1/\sigma$.

Under H_1, prior on (θ, σ) is $\pi_1(\theta, \sigma) = \frac{1}{\sigma}\,\pi_1(\theta \mid \sigma)$, where $\pi_1(\theta \mid \sigma)$ is Cauchy(θ_0, σ).

Bayes factor of H_0 to H_1, if one stops after observing $X_1, X_2, \ldots X_n$ $(n \geq 2)$, is

$$B_n = \frac{1}{\sqrt{2\pi}} \left[\int_0^\infty \left(1 + \frac{(n-1)n\xi}{n-1+t_n^2}\right)^{-\frac{n}{2}} (1+n\xi)^{\frac{n-1}{2}} e^{\frac{1}{2\xi}} \xi^{-\frac{3}{2}}\, d\xi \right]^{-1},$$

where t_n is the usual t-statistic.

A common sequential stopping rule (any other could also be used):
If $B_n \leq R$, $B_n \geq A$ or $n = M$, then stop the experiment.

Intuition:

R = "odds of H_0 to H_1" at which one would wish to stop and reject H_0.

A = "odds of H_0 to H_1" at which one would wish to stop and accept H_0.

M = maximum number of observations that can be taken

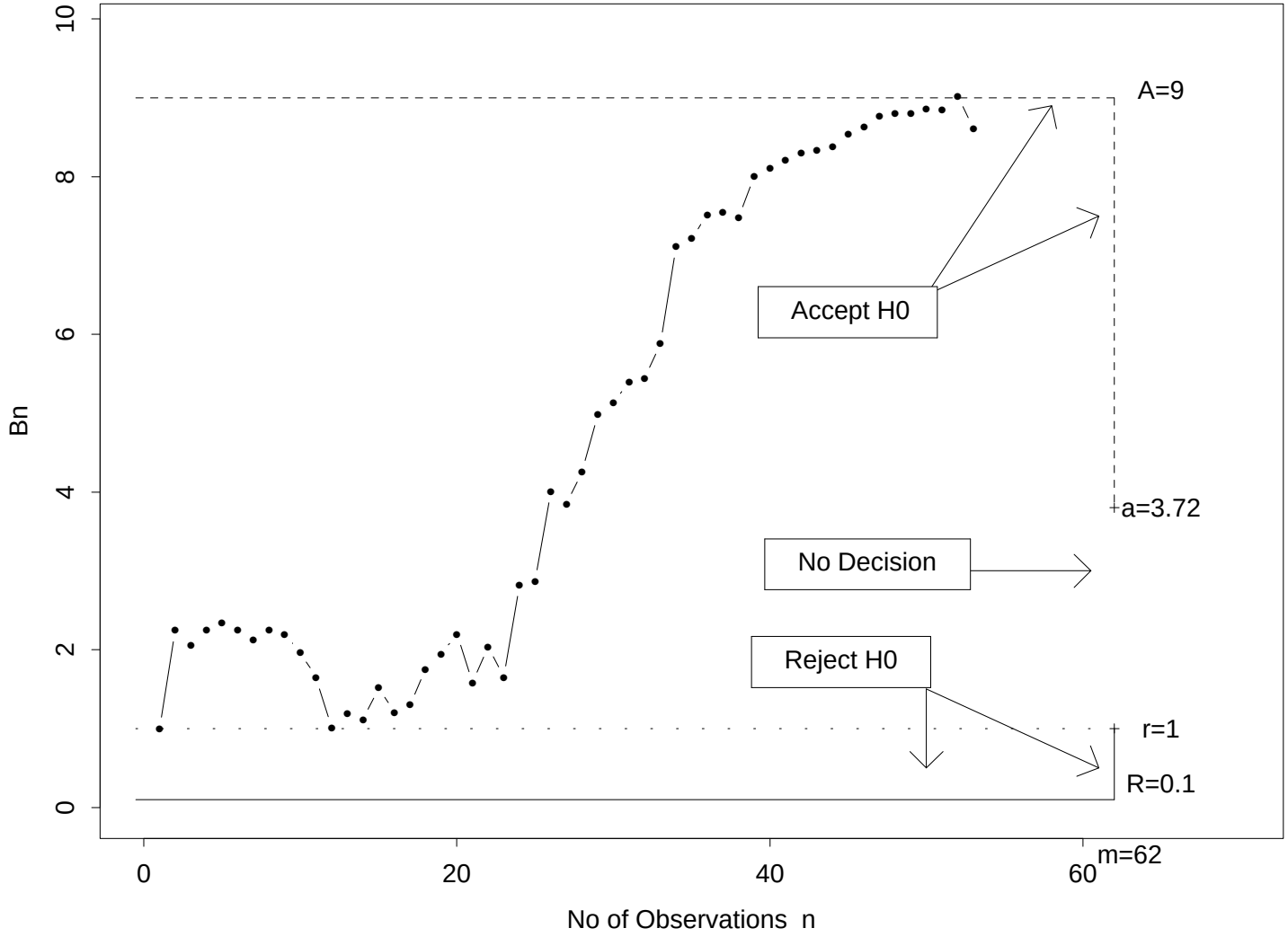
Example:

R = 0.1 (i.e., stop when 1 to 10 odds of H_0 to H_1)

A = 9 (i.e., stop when 9 to 1 odds of H_0 to H_1)

M = 62

An Application: The data arose as differences in time to recovery between paired patients who were administered different hypotensive agents. Testing $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$ is thus a test to detect a mean difference in treatment effects.



Comments about this sequential test:

- For the actual data, the stopping boundary would have been reached at time $n = 52$ ($B_{52} = 9.017 > A = 9$), and the conclusion would have been to accept H_0, and report the odds of correct to incorrect acceptance of H_0 of $B_{52} = 9.017$.
- From (a generalization) of the theorem in talk 1, this has the same frequentist justification as reporting the pre-experimental acceptance odds for a correct to incorrect acceptance of H_0.
  - Curiously, the pre-experimental acceptance odds will depend on the stopping rule used, but the Bayes factor does not.
- Computation is easy :
  - No stochastic process computations are needed (as are needed in unconditional frequentist testing).
  - Computations do not change as the stopping rule changes.
  - Sequential testing is as easy as fixed sample size testing.

The reason the Bayes factor does not depend on the stopping rule:
Writing the normal density of the X_i as $f(x_i \mid \theta, \sigma)$, optional stopping alters the data density to be

$$\tau_N(x_1, x_2, \ldots, x_N) \prod_{i=1}^{N} f(x_i \mid \theta, \sigma)\,,$$

where N is the (random) time at which one stops taking data and $\tau_N(x_1, x_2, \ldots, x_N)$ gives the probability (often 0 or 1) of stopping sampling.

Then

$$\begin{aligned} B_n &= \frac{\int \frac{1}{\sigma}\tau_N(x_1, x_2, \ldots, x_N) \prod_{i=1}^{N} f(x_i \mid \theta_0, \sigma) d\sigma}{\int\int \pi_1(\theta \mid \sigma)\frac{1}{\sigma}\tau_N(x_1, x_2, \ldots, x_N) \prod_{i=1}^{N} f(x_i \mid \theta, \sigma) d\sigma d\theta} \\ &= \frac{\int \frac{1}{\sigma} \prod_{i=1}^{N} f(x_i \mid \theta_0, \sigma) d\sigma}{\int\int \pi_1(\theta \mid \sigma)\frac{1}{\sigma} \prod_{i=1}^{N} f(x_i \mid \theta, \sigma) d\sigma d\theta}\,. \end{aligned}$$

There are two consequences of this result:

1. Use of the Bayes factor gives experimenters the freedom to employ optional stopping without penalty.

2. There is no harm if 'undisclosed optional stopping' is used (common in some areas of psychology), as long as the Bayes factor is used to assess significance. In particular, it is a consequence that an experimenter cannot fool someone through use of undisclosed optional stopping.

The Philosophical Puzzle: How can there be no penalty for interim analysis?

- Bayesian analysis is just probability theory and so cannot be wrong foundationally.
- The 'statistician's client with a grant application example.'

But it is difficult; as Savage (1961) said "When I first heard the stopping rule principle from Barnard in the early 50's, I thought it was scandalous that anyone in the profession could espouse a principle so obviously wrong, even as today I find it scandalous that anyone could deny a principle so obviously right."

Jeffreys–Lindley Paradox

(perhaps the most famous perceived conflict between Bayesian and frequentist testing)

Illustration in the Normal Example:

- $X_i \mid \theta \overset{i.i.d.}{\sim} N(x_i \mid \theta, \sigma^2)$, σ^2 known.
- Test $H_0 : \theta = \theta_0 \quad$ versus $\quad H_1 : \theta \neq \theta_0$.
- Can reduce to sufficient statistic $\bar{x} \sim N(\bar{x} \mid \theta,\ \sigma^2/n)$.
- Prior on H_1: $\pi_1(\theta) = N(\theta \mid \theta_0, v_0^2)$
- Marginal likelihood under H_1: $m_1(\bar{x}) = N(\bar{x} \mid \theta_0, v_0^2 + \sigma^2/n)$.
- posterior probability:

$$\Pr(H_0 \mid \bar{x}) = \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\frac{1}{(2\pi(v_0^2+\sigma^2/n))^{1/2}} \exp\left\{-\frac{1}{2} \frac{1}{v_0^2+\sigma^2/n} (\bar{x}-\theta_0)^2\right\}}{\frac{1}{(2\pi\sigma^2/n)^{1/2}} \exp\left\{-\frac{1}{2} \frac{1}{\sigma^2/n} (\bar{x}-\theta_0)^2\right\}}\right]^{-1}$$

$$= \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \frac{\exp\{\frac{1}{2}\ z^2\ [1 + \frac{\sigma^2}{nv_0^2}]^{-1}\}}{\{1 + nv_0^2/\sigma^2\}^{1/2}}\right]^{-1},$$

where $z = \dfrac{|\bar{x} - \theta_0|}{\sigma/\sqrt{n}}$ is the usual (frequentist) test statistic for this problem.

In the normal testing example, for fixed z and large n,

$$\begin{aligned}\Pr(H_0 \mid \bar{x}) &= \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\exp\{\frac{1}{2}\, z^2 \, [1 + \frac{\sigma^2}{n v_0^2}]^{-1}\}}{\{1 + n v_0^2/\sigma^2\}^{1/2}}\right]^{-1} \\ &\approx 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\sigma}{\sqrt{n}\, v_0} \, \exp\{\frac{1}{2}\, z^2\} \longrightarrow 1 \quad \text{as } n \to \infty\,,\end{aligned}$$

so that a classical test can *strongly reject* H_0 (which happens when z is moderately large) and the Bayesian analysis can, at the same time, *strongly support* H_0 (if n is so large that $\exp\{\frac{1}{2}\, z^2\}/\sqrt{n}$ is small, even though z is moderately large); reaching opposite conclusions is the 'paradox.'

- This is not a paradox in the true sense, since it is just mathematics.
- *Robust Bayesian resolution:* $H_0 : \theta = \theta_0$ is just an approximation to $H_0 : |\theta - \theta_0| < \epsilon$, where ϵ can reflect reality or just experimental bias. The approximation is only accurate when $\epsilon < \sigma/(4\sqrt{n})$ (Berger and Delampady, 1987); thus, for very large n, it is typically not reasonable to use $H_0 : \theta = \theta_0$ as the null hypothesis, so the 'paradox' becomes vacuous.

The Jeffreys-Lindley 'Paradox' and Experimental Bias

Suppose that H_0 is truly precise (e.g. 0 psychic effect), but that the experiment has some bias $b \sim N(b \mid 0, \delta^2)$. Then

$$\begin{aligned} \Pr(H_0 \mid \bar{x}) &= \left[1 + \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\exp\{\frac{1}{2}\, z_b^2 \, [1 + \frac{(\delta^2 + \sigma^2/n)}{v_0^2}]^{-1}\}}{\{1 + \frac{v_0^2}{\delta^2 + \sigma^2/n}\}^{1/2}} \right]^{-1} \\ &\approx 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\sqrt{\delta^2 + \sigma^2/n}}{v_0} \, \exp\{\frac{1}{2}\, z_b^2\}, \end{aligned}$$

when $\sqrt{\delta^2 + \sigma^2/n}$ is small, and where $z_b = |\bar{x} - \theta_0| / \sqrt{\delta^2 + \sigma^2/n}$ can be thought of as standard normal under H_0 in the presence of the bias. Then

$$\lim_{n \to \infty} \Pr(H_0 \mid \bar{x}) = 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\delta}{v_0} \, \exp\{\frac{(\bar{x} - \theta_0)^2}{2\delta^2}\}$$

which does not go to 1. Also

$$\Pr(H_0 \mid \bar{x}) \approx 1 - \frac{\Pr(H_1)}{\Pr(H_0)} \, \frac{\delta}{v_0} \, \exp\{\frac{z^2 \sigma^2}{2n\delta^2}\},$$

for the interesting range of $z^2\sigma^2/[n\delta^2]$.

A situation where Bayesians and frequentists seemingly cannot agree: Sequential Endpoint Testing

A sequence of null and alternative hypotheses $\{H_0^1, H_1^1\}, \ldots, \{H_0^m, H_1^m\}$ are to be tested sequentially. In the unconditional frequentist world,

- the ordering is important and must be pre-specified;
- one picks the same Type I error α for each hypothesis test;
- one is allowed to continue rejecting hypotheses until there is a failure to reject (at which point you stop), and state that the probability that the overall Type I error (the probability that *any* rejection is in error) is α.

To a Bayesian, this is not logical:

- Suppose one rejects 1000 completely independent hypotheses; is there really only a 0.05 chance that at least one of the rejections is wrong?
  - Indeed, then $P(H_1^1, H_1^2, \ldots, H_1^{1000} \mid data) = \prod_{i=1}^{1000} P(H_1^i \mid data) \approx 0$.

We have not yet found a good conditional frequentist solution to this problem (i.e., one that agrees with the objective Bayesian answer).

A Weaker Form of Bayesian-Frequentist Unification: utilizing the Bayesian procedure but also giving its frequentist properties

Pedagogical example: testing exclusive hypotheses and the problem of test statistic dependency

Suppose one is testing mutually exclusive hypotheses $H_i,\ i = 1, \ldots, m$, so that exactly one and only one of the H_i is true.

Bayesian analysis: If the hypotheses are viewed as exchangeable, choose $P(H_i) = 1/m$ and analyze the data $\boldsymbol{x}$.

- Let $m_i(\boldsymbol{x})$ denote the marginal density of the data under H_i. (The data density integrated over the prior density for unknown parameters under H_i.) This is often called the *likelihood* of H_i.
- The posterior probability of H_i is

$$Pr(H_i \mid \boldsymbol{x}) = \frac{m_i(\boldsymbol{x})}{\sum_{j=1}^{m} m_j(\boldsymbol{x})}\ .$$

- Thus the likelihood $m_i(\boldsymbol{x})$ for H_i is 'penalized' by a factor of $O(\frac{1}{m})$, resulting in multiplicity control.

Null control: If there is a good possibility of 'no effect' use, e.g.,

- $Pr(H_0) \equiv Pr(\text{no effect}) = 1/2$,
- $Pr(H_i) = 1/(2m)$.

Example: 1000 energy channels are searched for the Higgs boson. In each, one observes $X_i \sim N(x_i \mid \theta_i, 1)$, and at most one of $H_i : \theta_i \neq 0$ is true.

Suppose $x_5 = 3$, and the other 999 of the X_i are standard normal variates.

- If testing in isolation $H_5^0 : \theta_5 = 0$ versus $H_5^1 : \theta_5 \neq 0$, with prior probabilities of 1/2 each and a standard unit information Cauchy prior on θ_i under H_5^1, then $Pr(H_5^1 \mid x_5 = 3) = \frac{m_5^1(3)}{m_5^1(3)+m_5^0(3)} = 0.96$.
- With multiplicity control, assigning $Pr(H_i) = 1/1000$, this becomes (on average over the 999 standard normal variates) $Pr(H_5^1 \mid \boldsymbol{x}) = \frac{m_5(\boldsymbol{x})}{\sum_{j=1}^{1000} m_j(\boldsymbol{x})} = 0.019$ (and 0.38 for $x_5 = 4$; and 0.94 for $x_5 = 5$)
- With null control in addition to multiplicity control, ($Pr(\text{no effect}) = 1/2$ and $Pr(H_i) = 1/(2000)$), this becomes $Pr(H_5^1 \mid \boldsymbol{x}) = 0.019$.
- If null control was employed but *pre-experimentally* the physicist decided to use all of the non-null mass on H_5, the answer would have *legitimately* been $Pr(H_5^1 \mid \boldsymbol{x}) = 0.96$.

Main issue: This is the Bayesian solution regardless of the structure of the data; in contrast, frequentist solutions depend on the structure of the data.

Example: For each channel, test $H_{0i} : \theta_i = 0$ versus $H_{1i} : \theta_i \neq 0$.

Data: $X_i, i = 1, ..., m$, are $N(x_i \mid \theta_i, 1, \rho)$, ρ being the correlation.

If $\rho = 0$, one can just do individual tests at level α/m (Bonferroni) to obtain an overall error probability of α.

If $\rho > 0$, harder work is needed:

- Choose an overall decision rule, e.g., "declare channel i to have the signal if X_i is the largest value and $X_i > K$."
- Compute the corresponding error probability, which can be shown to be

$$\alpha = \Pr(\max_i X_i > K \mid \theta_1 = \ldots = \theta_m = 0) = E^Z \left[1 - \Phi \left(\frac{K - \sqrt{\rho} Z}{\sqrt{1 - \rho}} \right)^m \right],$$

  where Φ is the standard normal cdf and Z is standard normal.

Note that this gives (essentially) the Bonferroni correction when $\rho = 0$, and converges to $1 - \Phi[K]$ as $\rho \to 1$ (the one-dimensional solution).

A Bayesian solution: (with Shih-Han Chang)

Suppose the prior, $\pi_i(\theta_i)$, for the signal magnitude under M_i is

$$\theta_i \sim N(0, \tau^2)\,.$$

Lemma 1 *The posterior probability of the i^{th} model, $i \geq 1$, is*

$$P(M_i \mid \boldsymbol{X}) = \left[\sqrt{1+a\tau^2}(m)\exp\left\{\frac{-\tau^2}{2(1+\tau^2 a)}\left(\frac{x_i}{1-\rho}+m\bar{x}b\right)^2\right\} + \sum_{k=1}^{m}\exp\left\{\frac{-\tau^2}{2(1+\tau^2 a)}\left(\frac{x_i^2-x_k^2}{(1-\rho)^2}+\frac{2bm\bar{x}}{1-\rho}(x_i-x_k)\right)\right\}\right]^{-1}$$

where

$$\begin{cases} a = \frac{1+(m-2)\rho}{(1+(m-1)\rho)(1-\rho)} \\ b = \frac{-\rho}{(1+(m-1)\rho)(1-\rho)} \end{cases} \quad \textit{and} \quad \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}^{-1} = \begin{pmatrix} a & b & \cdots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & a \end{pmatrix}$$

The Bayesian answer when the correlation goes to 1:

When $m = 2$ and $\rho \to 1$ under model M_i,

$$P(M_i \mid \boldsymbol{X}) \to \begin{cases} \frac{1}{1+\exp\left\{-\frac{1}{2}\left(x_i^2 - x_{(-i)}^2\right)\right\}} & i = 1, 2 \\ 1 & i = 0 \end{cases}$$

When $m > 2$ and $\rho \to 1$ under model M_i,

$$P(M_i \mid \boldsymbol{X}) \to 1 \quad \text{(and correctly so)}\,.$$

Note that the adhoc frequentist test does not have this behavior. As $\rho \to 1$, the frequentist test

- still has probability α of incorrectly rejecting a true M_0;
- still has positive probability of not detecting a signal when M_i is true.

To obtain a fully-powered frequentist procedure, try the strategy of utilizing the Bayes procedure, and verifying that it has satisfactory frequentist properties:

- Use $\max_{i \geq 1} P(M_i \mid X)$ as the test statistic, i.e., *accept model M_i if its posterior probability, $P(M_i \mid \mathbf{X})$, is the largest and is greater than a specified threshold p.*
- Find its frequentist calibration, i.e. choose the threshold p so that the *false positive probability* (FPP)

$$P(\max_{i \geq 1} P(M_i \mid X) > p \mid M_0) = \alpha \, .$$

Note: All Bayesian procedures have FPP that go to zero under the null hypothesis at a rate of $O\left(\frac{1}{\log m}\right)$ or faster.

A curious aside: When $m \to \infty$ and $\rho \in [0, 1)$, under the null model:

$$P(M_0 \mid \boldsymbol{X}) \to P(M_0) = \frac{1}{2} \, .$$

A robust Bayes version (with $\hat{\tau}^2$ estimated) of the resulting powerful frequentist procedure at level α

- Find the Bayesian acceptance threshold probability p that yields *false positive probability* $= \alpha$:
  - Solve for k^* in

$$\alpha = \mathbf{P}(false\ positive \mid null\ model\ , \hat{\tau}^2) \approx \frac{1}{log\, m}\left(\frac{1}{k^*} - \frac{1}{2}\right).$$

  - Solve for p in

$$-2\, log\left(\sqrt{\pi}\left(\frac{1}{k^*} - \frac{1}{2}\right)\right) = log\, k^* + 2\, log\left(\frac{p}{1-r}\right) + 2\left(\frac{1}{k^*}\right).$$

- Then choose model M_i if it has the largest $P(M_i \mid \boldsymbol{X})$ which exceeds p; else choose M_0.

To reiterate here the potential advantages of Bayesian multiplicity control:

- Its implementation depends only on the prior probability assignment, and not on the structure of the data. Hence it is potentially computationally feasible in scenarios of high dependence amongst test statistics.
- It is 'fully powered,' whereas adhoc frequentist procedures which achieve multiplicity control need not be. For instance, in the previous example of correlated data and as $\rho \to 1$,
  - the adhoc frequentist procedure with error probability 0.05 declares a discovery if $\max_i |X_i| > 1.96$, which could be right or wrong;
  - the Bayesian procedure has the rather remarkable property that, for more than two observations, the posterior probability of the true hypothesis (correctly) goes to 1.

So where are we on the foundations of Bayesian - frequentist unification?

- Unification is pretty much possible for estimation type problems, utilizing objective Bayesian analysis.
  - It might also be possible using the bootstrap, but the Bayesian properties of the bootstrap are not clear.
- For testing, utilization of *odds of correct rejection to incorrect rejection* seems to lead to a simple potential unification using Bayes factors.
  - We still have a ways to go, however, in agreeing on objective Bayesian testing methods, unless we just go with $1/[-ep\log p]$.
- Of the areas where it is often thought that unification is not possible, only multiple endpoint testing seems to be a problem.
- Areas in which the verdict is still out:
  - Model uncertainty
  - Nonparametrics and semiparametrics

Thanks!