Lecture 7.
The Likelihood Principle and Stopping Rule Principle

- George Barnard, 1915-2002.
- An inventor (popularizer?) of sequential analysis
- An inventor (popularizer?) of pivotal analysis
- Perhaps the first to espouse the 'likelihood principle' and 'stopping rule principle.'

- Allan Birnbaum, 1923-1976.
- Worked on foundations of statistics.
- Was a frequentist.
- Famous 1962 paper on 'proving' the Likelihood Principle
  - but this 'invalidated' frequentist statistics

Focus on the likelihood function $\mathcal{L}(\boldsymbol{\theta}) = f(\boldsymbol{x} \mid \boldsymbol{\theta})$, for the observed data $\boldsymbol{x}$.

Likelihood Principle (LP):

- All the information about $\boldsymbol{\theta}$ obtainable from an experiment is contained in $\mathcal{L}(\boldsymbol{\theta})$.
- Two likelihood functions $\mathcal{L}_1(\boldsymbol{\theta})$ and $\mathcal{L}_2(\boldsymbol{\theta})$ (from the same or different experiments but about the same $\boldsymbol{\theta}$) contain the same information about $\boldsymbol{\theta}$ if they are proportional to one another.

In a sequential experiment, the observations $X_1, X_2, \ldots$ arrive sequentially; examples include a series of products coming off an assembly line, a series of missiles being tested, and a series of patients participating in a clinical trial.

Stopping Rule Principle (SRP): In a sequential experiment, the information from the experiment about $\boldsymbol{\theta}$ is not affected by the reason for stopping experimentation (although it is, of course, dependent on the stopping point).

- Robert Wolpert, 1950 – .
- Coauthor of *The Likelihood Principle*, from which examples here are taken.
- Work on conditional frequentist inference.
- Work on Uncertainty Quantification of computer models.

Example 1: Two scientists are collaborating on an experiment. They have a joint graduate student who is conducting the experiment, and they both are watching her work.

- The observations $X_1, X_2, \ldots$ are i.i.d. $Bernoulli(\theta)$ random variables.
- The scientists agree they are testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$.
- After the ninth observation, they simultaneously say "That's enough data," and they tell the student to stop experimenting.
- The data consisted of 9 successes and 3 failures.
- Each scientist analyzes the data; when getting back together, they are shocked that they disagree.
  - Scientist 1 claims that there is not significant evidence at the 0.05 level.
  - Scientist 2 says there is significant evidence at the 0.05 level.
- How did this disagreement happen?

Scientist 1's analysis: He had planned from the beginning to take just 12 observations (but had not communicated this). Thus the number of successes, X, is $Binomial(12, \theta)$, and the p-value for the observed $x = 9$ is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{12} \binom{12}{x} 0.5^x (1 - 0.5)^{(12-x)} = .0730\,.$$

Scientist 2's analysis: She had planned to continue taking observations until observing 3 failures. Thus, for her, X has a $Negative - binomial(3, \theta)$ distribution, and the p-value is

$$p = Pr(X \geq 9 \mid \theta = 0.5) = \sum_{x=9}^{\infty} \binom{x+2}{x} 0.5^x (1 - 0.5)^3 = .0338\,.$$

- The two scientists had different *stopping rules*.
- But note that these were just thoughts in their heads; these thoughts had no effect on the actual experiment that was conducted or the results.
- The stopping rule principle says such thoughts should not matter; the stopping rule should not affect the analysis.

For Scientist 1 the observed likelihood function was

$$\mathcal{L}_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3\,;$$

For Scientist 2 it was

$$\mathcal{L}_2(\theta) = \binom{11}{9}\theta^9(1-\theta)^3\,.$$

Since $\mathcal{L}_1(\theta) \propto \mathcal{L}_2(\theta)$, the Likelihood Principle also says that the evidence about θ from either viewpoint is the same.

Example 2: A scientist enters the statistician's office with $n = 100$ observations, assumed to be independent and from a $N(\theta, 1)$ distribution, with the desire to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. She reports that $\bar{x}_n = 0.2$, so the standardized test statistic is $z_{100} = \sqrt{n}\,|\bar{x}_{100} - 0| = 2$.

- A careless classical statistician might simply conclude that there is significant evidence against H_0 at the 0.05 level.
- A careful classical statistician will ask the scientist "Why did you cease experimentation after 100 observations?"
  - If the scientist replies, "I just decided to take a batch of 100 observations," there would seem to be no problem.
  - But there is another important question that should be asked (from the classical perspective), namely: "What would you have done had the first 100 observations not yielded significance?"

To see the reasons for this question, suppose the scientist replies: "I would then have taken another batch of 100 observations." This reply does not completely specify a stopping rule, but the scientist might agree that she was implicitly considering a stopping rule of the form

- take the first 100 observations;
  - if $z_{100} > k$ for some critical value k, then stop and reject H_0,
  - if $z_{100} < k$ then take another 100 observations and reject if $z_{200} > k$.
- For this procedure to have level $\alpha = 0.05$, k must be chosen to be 2.18 (Pocock, 1977). Since the actual data had $z_{100} = 2 < 2.18$, the scientist could not actually conclude significance, and hence would have to take the next 100 observations.

This strikes many people as peculiar. The interpretation of the results of an experiment depends not only on the data obtained and the way it was obtained, but also upon thoughts of the experimenter concerning plans for the future.

The puzzled scientist scientist leaves and gets the next 100 observations.

- She reports that $z_{200} = 2.2 > 2.18$; has significance now been obtained?
- No, again the statistician asks what the scientist would have done had the results not been significant.
  - The scientist says, "If my grant renewal were to be approved, I would then take another 100 observations;
  - If the grant renewal were rejected, I would have no more funds and would have to stop the experiment at 200 observations."
  - The classical statistician *must* then advise her that, if the grant is rejected, significance has been obtained, but otherwise another 100 observations must be taken.

Note: This is not at all fanciful; the standard practice in psychology experiments is to do precisely the above: keep taking observations in batches until $p < 0.05$, ignoring the issue of the stopping rule.

Formalizing the stopping rule in a sequential experiment:

Definition: A *stopping rule* is a function $\boldsymbol{\tau}$ of the sequential data $x_1, x_2, \ldots$ such that, at each time j, $\tau(x_1, x_2, \ldots, x_j)$ is the probability (usually 0 or 1) of stopping the experiment.

Example 1. The stopping rule $\tau(x_1, x_2, \ldots, x_n) = 1$, with all other $\tau(x_1, x_2, \ldots, x_j) = 0$, defines the fixed sample size experiment of sample size n.

Example 2. Consider the stopping rule

- $\tau(x_1, x_2, \ldots, x_{50}) = 1$ if the p-value < 0.01 and is 0 otherwise;
- $\tau(x_1, x_2, \ldots, x_{100}) = 1$,
- all other $\tau(x_1, x_2, \ldots, x_j) = 0$.

So take the first 50 samples, stopping if the p-value is less than 0.01, and otherwise take another 50 samples and then stop.

Example 3. After each observation x_i, check whether the p-value < 0.01; if so, stop experimentation and otherwise continue. The stopping rule is then, for every j, $\tau(x_1, x_2, \ldots, x_j) = 1$ if the p-value is less than 0.01; else $\tau(x_1, x_2, \ldots, x_j) = 0$.

The LP implies the SRP: In a sequential experiment, for every j, the density of $X_1, X_2, \ldots, X_j$ is

$$\tau(x_1, x_2, \ldots, x_j) f(x_1, x_2, \ldots, x_j \mid \boldsymbol{\theta}) \, .$$

Hence, for two different stopping rules $\tau_1(\cdot)$ and $\tau_2(\cdot)$ and final data of sample size N (the *stopping time*)

$$\tau_1(x_1, x_2, \ldots, x_N) f(x_1, x_2, \ldots, x_N \mid \boldsymbol{\theta}) \propto \tau_2(x_1, x_2, \ldots, x_N) f(x_1, x_2, \ldots, x_N \mid \boldsymbol{\theta}),$$

as functions of $\boldsymbol{\theta}$.

An example of sequential testing using Bayes factors

(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

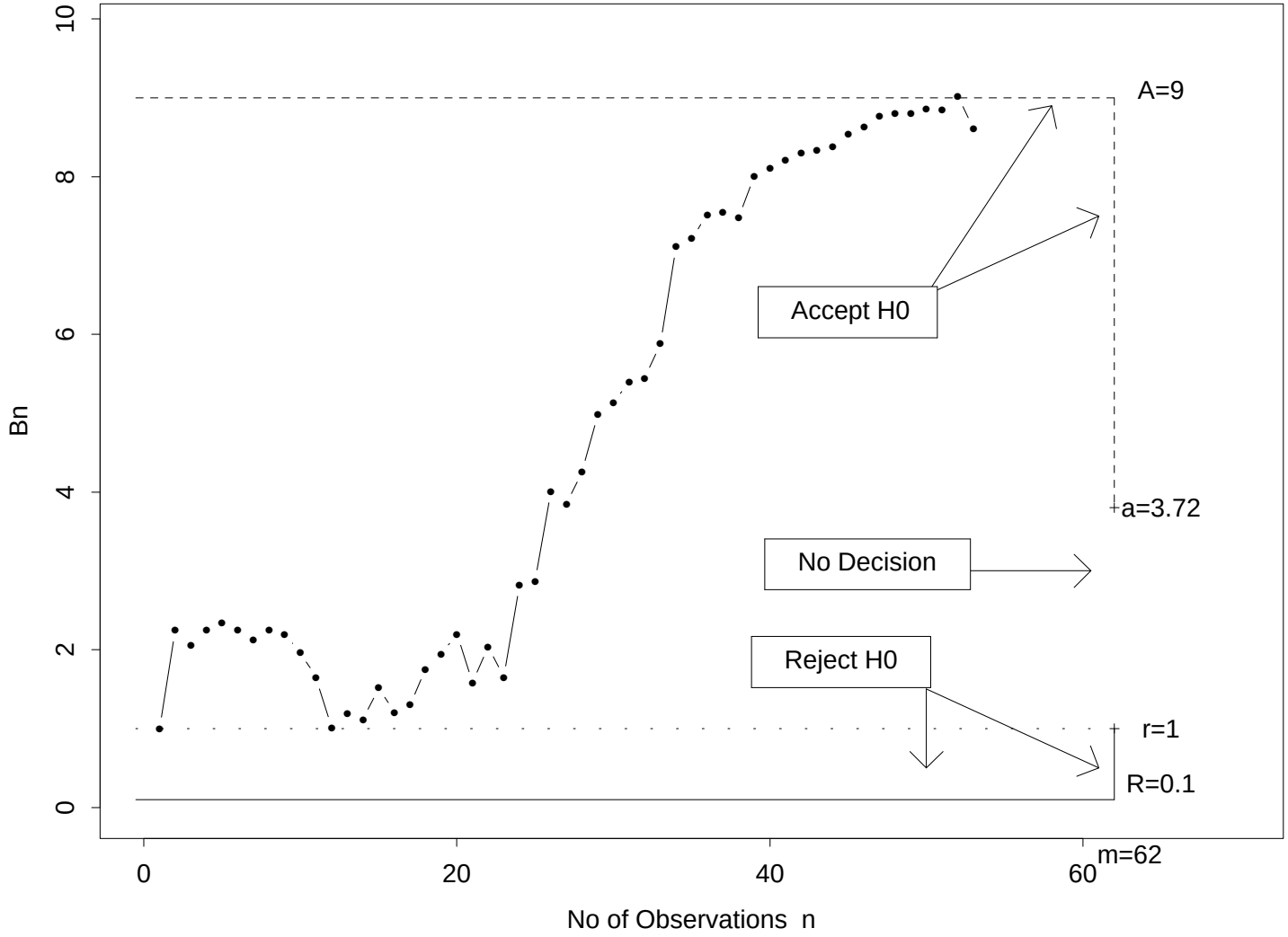
Example:

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 the Bayarri et. al. theorem in Lecture 6, 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.

Birnbaum's 'Proof' of the Likelihood Principle

Two other principles:

- **Sufficiency Principle:** All the evidence about $\boldsymbol{\theta}$ from an experiment is conveyed by a sufficient statistic (if one exists).
  (A pre-experimental random sufficient statistic is allowed, so this is completely compatible with frequentist statistics.)
- **Conditionality Principle:** Suppose you can conduct either experiment E_1 concerning $\boldsymbol{\theta}$ or E_2 concerning (the same) $\boldsymbol{\theta}$, the experiment being conducted being determined by the flip of a fair coin. Then the evidence about $\boldsymbol{\theta}$ is only that obtained from the experiment actually conducted.

In 1962, essentially everyone agreed with these two principles.

In 1962, essentially everyone (except Bayesians) thought the LP was ridiculous, because it ruled out p-values, frequentist methods, fiducial methods, ...

Birnbaum's Theorem: The Sufficiency Principle and Conditionality Principle imply the LP.

Why Bayesian Analysis Automatically Follows the LP and SRP

Satisfying the LP: Assuming the prior distribution for $\boldsymbol{\theta}$ is developed independently of the model and data,

(i) the posterior only depends on the observed likelihood and prior and so is utilizing only the observed likelihood from the model and data;

(ii) proportional likelihoods will imply proportional posteriors (since the same prior will be used) and hence the same conclusions.

Satisfying the SRP: In a sequential experiment, for every j, the density of $X_1, X_2, \ldots, X_j$ is $\tau(x_1, x_2, \ldots, x_j) f(x_1, x_2, \ldots, x_j \mid \boldsymbol{\theta})$, which is proportional to $f(x_1, x_2, \ldots, x_j \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ and will hence lead to the same posterior.

There are two consequences of this SRP result:

1. Use of the Bayesian methods 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 Bayesian methodology is employed. In particular, it is a consequence that an experimenter cannot fool someone through use of undisclosed optional stopping.

A 'Counterexample' to the LP and SP:

Example 1: Suppose $X_1, X_2, \ldots$ are independent $N(\theta, 1)$ random variables and that a confidence set for θ is desired. Let $\bar{X}_j$ be the sample average, and $Z_j = \sqrt{j}\bar{X}_j$.

- Consider the stopping rule which stops sampling at the first $|Z_j| > 2$.
- The Law of the Iterated Logarithm implies that this is sure to happen, so this is a proper stopping rule (i.e., is guaranteed to stop with probability 1).

The objective Bayesian analysis utilizes the constant prior for θ and, if the sequential procedure stops at observation n, would produce the 95% credible interval for θ

$$C_n = \left(\bar{x}_n - \frac{1.96}{\sqrt{n}}, \bar{x}_n + \frac{1.96}{\sqrt{n}} \right) .$$

The seemingly paradoxical feature of this example is that the credible set can never contain $\theta = 0$ (since, because of the stopping rule, $\bar{x}_n < -2/\sqrt{n}$ if $\bar{x}_n$ is negative, and $\bar{x}_n > 2/\sqrt{n}$ if $\bar{x}_n$ is positive); thus the procedure has frequentist coverage probability of 0 at $\theta = 0$.

It thus seems that the experimenter can, through choice of the stopping rule, "trick" the Bayesian into believing that θ is not zero.

- If the objective prior $\pi(\theta) = 1$ is appropriate, the objective Bayesian will insist that $\theta = 0$ is not a likely value.
- If the experimenter knew that $\theta = 0$ is a real possibility, i.e., has positive prior probability p, then it was scientific fraud not to reveal that to the statistician. If that had been revealed and, for instance, the $N(0, A)$ density is then used for θ, conditional on $\theta \neq 0$

$$P(\theta = 0 \mid \bar{x}_n) = \left[1 + \frac{(1-p)}{p} \; \frac{\exp\{\frac{1}{2} Z_n^2 / [1 + (nA)^{-1}]\}}{\sqrt{1+nA}}\right]^{-1}$$

  and the posterior density for θ, conditional on $\theta \neq 0$, is

$$[1 - P(\theta = 0 \mid \bar{x}_n)] \times N\left(\frac{A}{[A + n^{-1}]} \; \bar{x}, \frac{A}{[nA+1]}\right).$$

- In the Example, suppose $Z_n \geq 2$ at $n = 110$ with $Z_{110} = 2.1$. If $p = 0.1$ and $A = 1$, $P(\theta = 0 \mid \bar{x}_{110}) = 0.116$, so the credible set will be the union of the point 0 and the 88.4 % credible set from the conditional density.

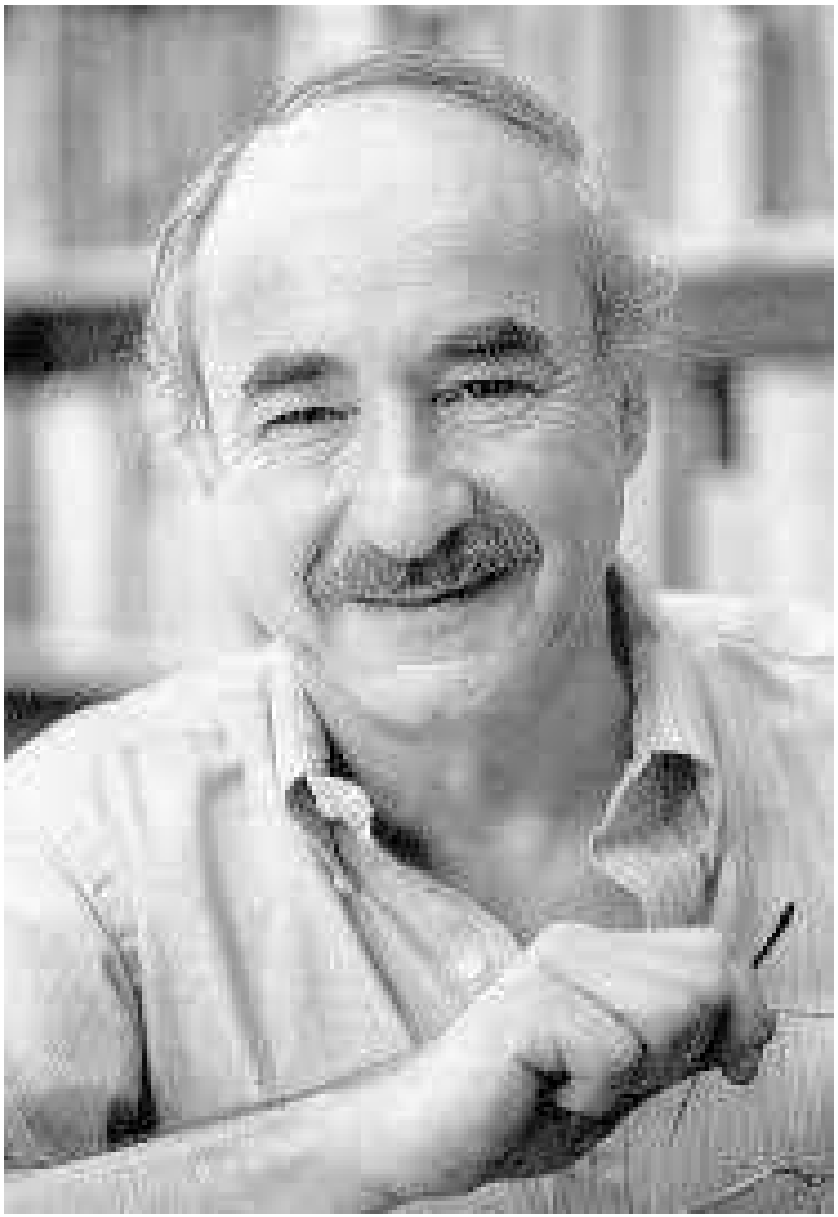
The Philosophical Puzzle: How can there be no penalty for optional stopping?

- Bayesian analysis is just probability theory and so cannot be wrong on this question.
- Remember the earlier examples where it seemed absurd to adjust the answer because of optional stopping.

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."

Other criticisms of the LP and SRP

- The LP seems to imply that one only needs the likelihood function to do the analysis, and there are many examples to show that this is not so.
  - All the examples are countered by using the LP through Bayesian analysis, but no other method of countering the examples is known.
- From the stopping rule, one may learn about the experimenter's prior beliefs (although these should have been queried directly).
  - True, but one is learning evidence external to the experiment.
- Implementation: the LP and SRP do not say how to use the likelihood function.
  - There is a 'likelihood school' of statistics, but it does not solve many problems.
  - Objective Bayesian analysis is typically used to implement the LP, but there is no definitive argument for that.
  - What is the likelihood function in complicated settings?



- Morris DeGroot, 1931-1989.
- The major US statistician in supporting Bayesian statistics after Savage.
- Built the first Bayesian department, at CMU
- His talk on the last point was titled *The most important question for a Bayesian: where is the bar?*
  - e.g., should the likelihood for the joint density $f(x, y, \theta)$ be $f(x \mid y, \theta)$ or $f(x, y \mid \theta)$, x being the data and y a future observation.

- M.J. (Susie) Bayarri, 1956-2014.
- The first great female Bayesian.
- Coauthor with de Groot on the above.
- We will see lots about her contributions to statistics.
- Won an award as the best restaurant critic in Valencia.

- And note that objective Bayesian analysis also has a problem with the LP.
  - The Jeffreys-rule prior for θ in a binomial problem is $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$.
  - The Jeffreys-rule prior for θ in a negative-binomial problem is $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1}$.

Hence, in the example of the two scientists with 9 successes and 3 failures, an objective Bayesian would have different posterior distributions and appears to be violating the LP.