

# Lecture 1: Reproducibility of Science: Impact of Bayesian Testing and Multiplicity

**Jim Berger**  
Duke University

*CBMS Conference on Model Uncertainty and Multiplicity*  
*July 23-28, 2012*

# Outline

- I. Reproducibility of Science
  - Evidence of an increasing lack of reproducibility of science
  - Some reasons for the lack of reproducibility
    - \* Publication bias
    - \* Experimental biases, including programming errors
    - \* The very considerable rewards for ‘positive’ results
    - \* Statistical biases
    - \* Egregiously bad statistics
    - \* The incorrect way in which  $p$ -values are used
    - \* Failure to adjust for multiplicities
      - Multiple testing
      - Multiple looks at the data
      - Multiple statistical analyses
  - How Bayesian analysis can help
- II. A Brief History of Bayesian Statistics
- III. Could Fisher, Jeffreys and Neyman have agreed on testing?

# I. Reproducibility of Science

## A. Evidence for a Lack of Reproducibility

- “The reliability of results from observational studies has been called into question many times in the recent past, with several analyses showing that well over half of the reported findings are subsequently refuted.” JNCI, 2007
- The NIH funded randomized clinical trials to follow up exciting results from 20 observational studies. Only 1 replicated.
- Bayer Healthcare reviewed 67 in-house attempts at replicating the findings in published research.
  - Less than 1/4 were viewed as having been essentially replicated.
  - Over 2/3 had major inconsistencies leading to project termination.
- John P. A. Ioannidis, JAMA-2005, 218-28: Five of 6 highly cited nonrandomized studies were contradicted or had found stronger effects than were established by later studies.

Even the best studies often fail to replicate.

- Ioannidis looked at the 49 most famous medical publications from 1990-2003 resulting from randomized trials; 45 claimed successful intervention.
  - 7 (16%) were contradicted by subsequent studies
  - 7 others (16%) had found effects that were stronger than those of subsequent studies
  - 20 (44%) were replicated
  - 11 (24%) remained largely unchallenged.
- Phase II drug trials success rates are falling (28% 5 years ago, 18% now) (Arrowsmith (2011) Nature Reviews Drug Discovery 10)
- 50% phase III drug trial failure rates are now being reported, versus a 20% failure rate 10 years ago (Arrowsmith (2011) Nature Reviews Drug Discovery 10); 70% phase III cancer drug failure rate
- Reports that 30% of phase III drug trial successes fail to replicate



## B. Some Reasons for a Lack of Reproducibility

### 1. Publication bias:

- Negative (and especially small negative) studies are often never reported or, if they are, can have publication delays of up to 3 years.

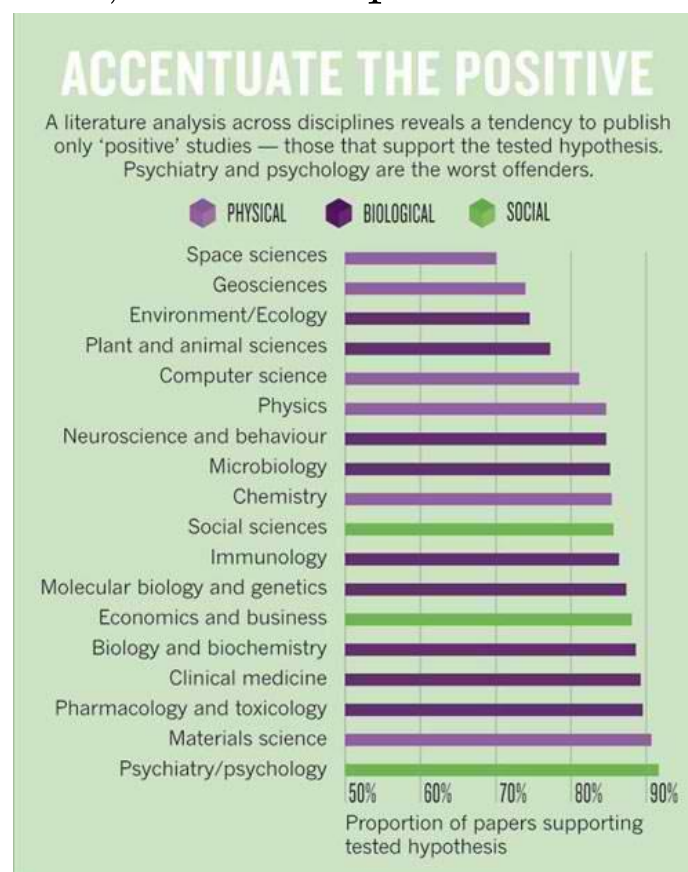


Figure 1: From Fanelli, D. *Scientometrics* 90, 891–904 (2011).

- Ioannides: Looked at a meta-analysis of a widely studied head and neck cancer:
  - the meta-analysis reported on 80 published studies;
  - they found 13 additional published studies not in the meta-analysis;
  - they found 10 non-published studies, but were able to get the data;
  - they found another 38 studies where data could not be obtained;
  - who knows how many other studies were done leaving no record.

The original 80 provided significance at 0.05 in the meta-analysis; the 80+13 were barely significant; the 80+13+10 did not yield significance.

- Effect sizes for observational studies with small sample sizes tend to be much larger than effect sizes for studies with large sample sizes.

- Interesting approach to detecting bias (Ioannides and Trikalinos, 2007):
  - look at the available studies, focusing on sample sizes and effect sizes;
  - perform a meta-analysis to estimate the overall effect;
  - simulate studies from this assumed population effect and the same sample sizes to determine how many studies should be nonsignificant, and compare this with the actual number of nonsignificant studies.

*Examples (Gregory Francis, Psychon Bull Rev, 2012):*

- The probability having one or fewer non-significant studies in the ten Bem (2011) psi\* experiments is 0.058.
- Meissner & Brigham (2001) performed a meta-analysis of 18 experiments on verbal overshadowing\*\*, nine of which were significant. The probability of nine or fewer non-significant experiments is 0.022 (1:5 odds)

*Note:* If there were publication bias, this would *underestimate* its extent, because of using the published studies to determine the overall effect.

\*sensing future events and using that information to judge the present

\*\*visual memory is impaired after subjects give a verbal description of the stimuli

## 2. Experimental biases:

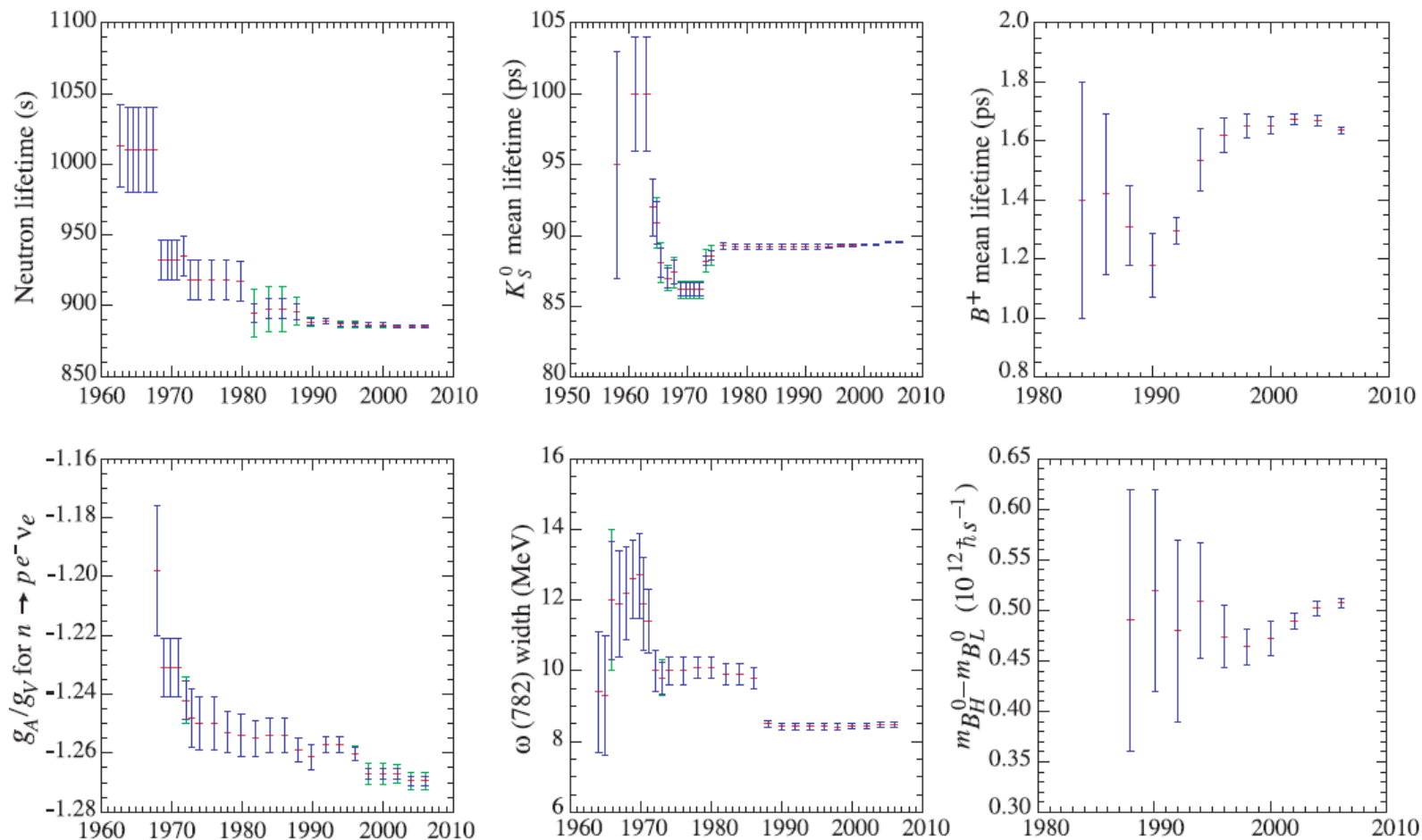


Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

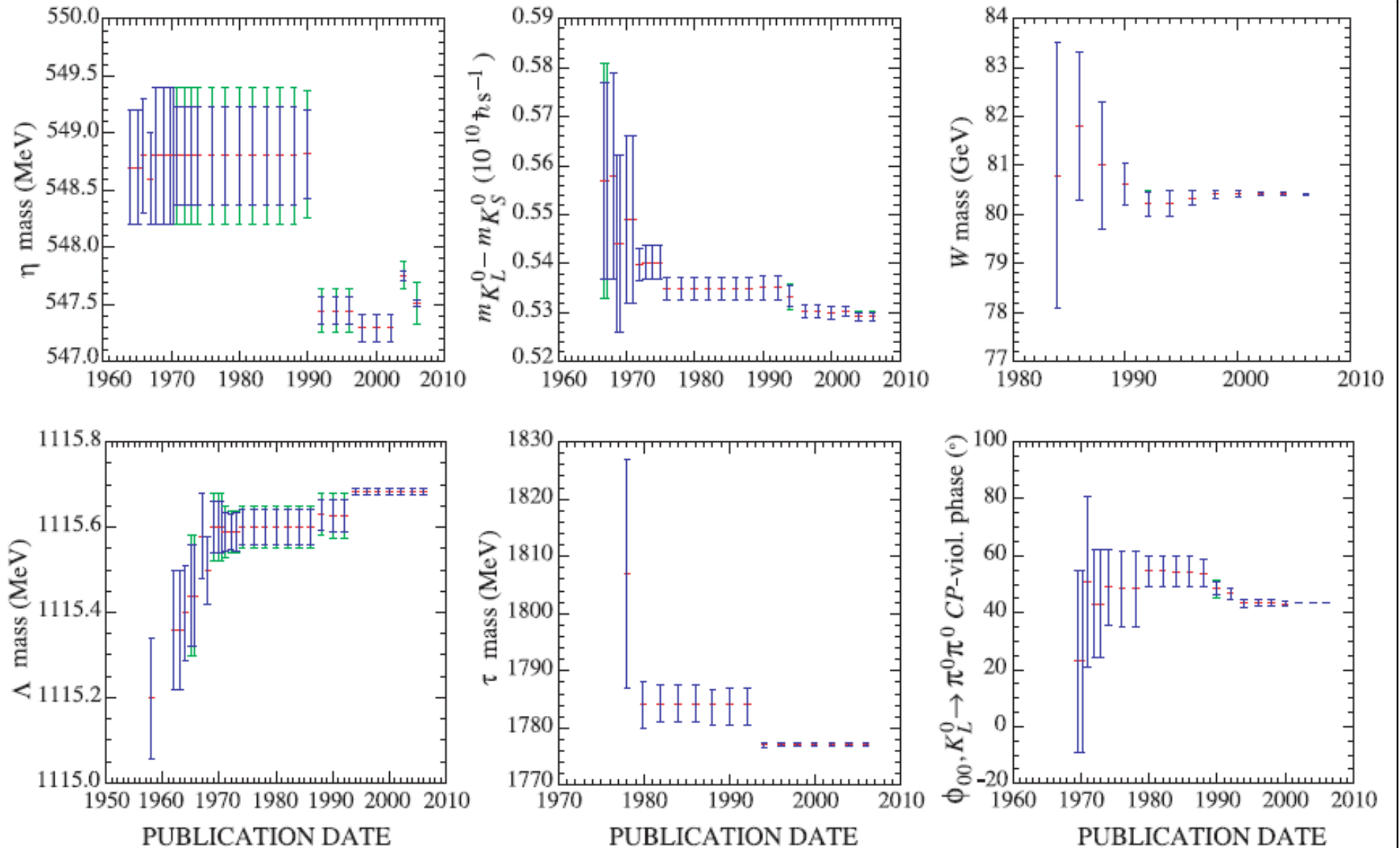


Figure 3: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

### 3. The very considerable rewards for 'positive' results

- Money and fame
  - “There is nothing wrong with cancer research that a little less money wouldn't cure.” (Nathan Mantel, NCI)
- Promotion and tenure
- Journals want high impact factors
- ...
- And, except perhaps for physics, there seems to be little to no *scientific* penalty for having a positive finding later refuted.

## 4. Statistical biases

- Confounding, especially in observational studies
  - Worse with large sample sizes

- Programming errors

“To err is human, but to really foul things up requires a computer.”

Farmers’ Almanac (1978)

- ...

## 5. Use of egregiously bad statistics

### 5.1. Using statistics ‘as a language’:

Sander Nieuwenhuis, Birte U Forstmann & Eric-Jan Wagenmakers, *Nature Neuroscience* 14, 1105–1107 (2011).

- Reviewed 513 neuroscience articles in five top-ranking journals.
- Found 157 comparing ‘Treatment A’ and ‘Treatment B.’
  - 78 correctly looked at the mean difference of effects for significance.
  - 79 had at least one instance of incorrectly concluding that there was a significant difference between the treatments if one was ‘significant at the 0.05 level against a control’ and the other was not (for instance, if  $z_A = 1.97$  and  $z_B = 1.95$  ).

### 5.2. Purposely ignoring statistical principles:

- The tradition in epidemiology is to ignore multiple testing.
- The tradition in psychology is to ignore optional stopping.

“You cannot ask us to take sides against arithmetic.” Winston Churchill



## 6. The incorrect way in which $p$ -values are used:

“To  $p$ , or not to  $p$ , that is the question?”

- Few non-statisticians understand  $p$ -values, most erroneously thinking they are some type of error probability (Bayesian or frequentist).
  - A survey 30 years ago:
    - \* “What would you conclude if a properly conducted, randomized clinical trial of a treatment was reported to have resulted in a beneficial response ( $p < 0.05$ )?”
      1. Having obtained the observed response, the chances are less than 5% that the therapy is not effective.
      2. The chances are less than 5% of not having obtained the observed response if the therapy is effective.
      3. The chances are less than 5% of having obtained the observed response if the therapy is not effective.
      4. None of the above.
    - \* We asked this question of 24 physicians ... Half ... answered incorrectly, and all had difficulty distinguishing the subtle differences...
    - \* The correct answer to our test question, then, is 3.”

“This isn’t right. This isn’t even wrong.” –Wolfgang Pauli, on a submitted paper

- \* **Actual correct answer:** The chances are less than 5% of having obtained the observed response *or any more extreme response* if the therapy is not effective.
- But, is it fair to count ‘possible data more extreme than the actual data’ in the evidence against the null hypothesis?  
Jeffreys (1961): “An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.”
- Matthews (1998): “The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding.”
- When testing precise hypotheses, true error probabilities (Bayesian or frequentist) are much larger than  $p$ -values.
  - Later examples.
  - *Applet* (of German Molina) available at [www.stat.duke.edu/~berger](http://www.stat.duke.edu/~berger)

## 7. Failure to adjust for multiplicities:

- Failure to properly account for multiple testing:
  - In a recent talk about the drug discovery process, the following numbers were given in illustration.
    - \* 10,000 relevant compounds were screened for biological activity.
    - \* 500 passed the initial screen and were studied in vitro.
    - \* 25 passed this screening and were studied in Phase I animal trials.
    - \* 1 passed this screening and was studied in a Phase II human trial.

This could be nothing but noise, if screening was done based on ‘significance at the 0.05 level.’

“Basic research is like shooting an arrow in the air and, where it lands, painting a target.” Homer Adkins

  - *Multiple Multiple Testing* (e.g., plasma samples are sent to separate genomic, protein, and metabolic labs for ‘discovery’.)
  - *Serial Studies* (e.g., there have been 16 large Phase III Alzheimer’s trials - all failing; the probability of that is 0.44)
  - The tradition in epidemiology is to ignore multiple testing,
    - \* usually arguing that the purpose is to find anomalies for further study.

- The tradition in psychology is to ignore optional stopping; if one is close to  $p = 0.05$ , go get more data to try get there (with no adjustment).
  - *Example:* Suppose one has  $p = 0.08$  on a sample of size  $n$ . If one takes up to four additional samples of size  $\frac{n}{4}$ , the probability of reaching  $p = 0.05$  is  $\frac{2}{3}$ .
  - When bias is present, one can often quickly reach  $p = 0.05$ .
- Multiple statistical analyses
  - Data selection “Torture the data long enough and they will confess to anything.”
    - \* Removing ‘outliers’ (that don’t seem ‘reasonable’)
    - \* Removing unfavorable data (e.g., because psychic powers come and go)
  - Trying out multiple models until ‘one works.’
  - Trying out multiple statistical procedures until ‘one reveals the signal.’ (At CERN  $10^{12}$  ‘cuts’ can potentially be applied to each particle track.)
  - Subgroup analysis
  - ...

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011), Psychological Science, 22, 1359–1366: show ‘significant evidence’ that listening to the song ‘When I’m Sixty-four’ by the Beatles can reduce a listener’s age by 1.5 years.



## C. How Bayesian Analysis Can Help

### 1. Bayesian Hypothesis Testing

**San Jose Mercury News**

mercurynews.com WEST VALLEY 102 Friday, September 25, 2009 THE NEWSPAPER OF SILICON VALLEY 75 cents

**AIDS MILESTONE**

# New path for HIV vaccine

Some in study protected from infection, but trial raises more questions

By Karen Kaplan and Thomas H. Maugh II  
*Los Angeles Times*

Hours after HIV researchers announced the achievement of a milestone that had eluded them for a quarter of a century, reality began to set in: Tangible progress could take another decade.

A Thai and American team announced early Thursday in Bangkok that they had found a combination of vaccines providing modest protection against infection with the virus that causes AIDS, unleashing excitement worldwide. The idea of a vaccine to prevent infection with the human immunodeficiency virus, HIV, had long been frustrating and fruitless.


But by Thursday afternoon, initial euphoria gave way to a more sober assessment. There is still a very long way to go before reaching the goal of producing a vaccine that reliably shields people from HIV.

Some researchers questioned whether the apparent 31 percent reduction in infections was a sta-

A researcher during the Thai phase III HIV Vaccine Trial, also known as RV 144, tests the "prime-boost" combination of two vaccines.

ASSOCIATED PRESS

See **VACCINE**, Page 14



## Hypotheses and data:

- Alvac had shown no effect
- Aidsvax had shown no effect

*Question:* Would Alvac as a primer and Aidsvax as a booster work?

*The Study:* Conducted in Thailand with 16,395 individuals from the general (not high-risk) population:

- 74 HIV cases reported in the 8198 individuals receiving placebos
- 51 HIV cases reported in the 8197 individuals receiving the treatment

## The test that was performed:

- Let  $p_1$  and  $p_2$  denote the probability of HIV in the placebo and treatment populations, respectively.
- Test  $H_0 : p_1 = p_2$  versus  $H_1 : p_1 \neq p_2$
- Normal approximation okay, so

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{\sigma}_{\{\hat{p}_1 - \hat{p}_2\}}}} = \frac{.009027 - .006222}{.001359} = 2.06$$

is approximately  $N(\theta, 1)$ , where  $\theta = (p_1 - p_2)/(.001359)$ .

We thus test  $H_0 : \theta = 0$  versus  $H_1 : \theta \neq 0$ , based on  $z$ .

- Observed  $z = 2.06$ , so the  $p$ -value is 0.04.
- If test  $H_0 : \theta = 0$  versus  $H_1 : \theta > 0$ , the  $p$ -value is 0.02.

## Bayesian analysis:

Posterior odds of  $H_1$  to  $H_0 = [\text{Prior odds of } H_1 \text{ to } H_0] \times B_{10}(z)$ ,

where

$$\begin{aligned}
 B_{10}(z) &= \text{Bayes factor of } H_1 \text{ to } H_0 = \text{'data-based odds of } H_1 \text{ to } H_0\text{'} \\
 &= \frac{\text{average likelihood of } H_1}{\text{likelihood of } H_0 \text{ for observed data}} = \frac{\int \frac{1}{\sqrt{2\pi}} e^{-(z-\theta)^2/2} \pi(\theta) d\theta}{\frac{1}{\sqrt{2\pi}} e^{-(z-0)^2/2}},
 \end{aligned}$$

For  $z = 2.06$  and  $\pi(\theta) = \text{Uniform}(0, 2.95)$ , the nonincreasing prior *most favorable* to  $H_1$ ,

$$B_{10}(z) = 5.63 \quad (\text{recall, the one-sided p-value is } 0.020)$$

(The actual subjective 'study team' prior yielded  $B_{10}^*(2.06) = 4.0$ .)



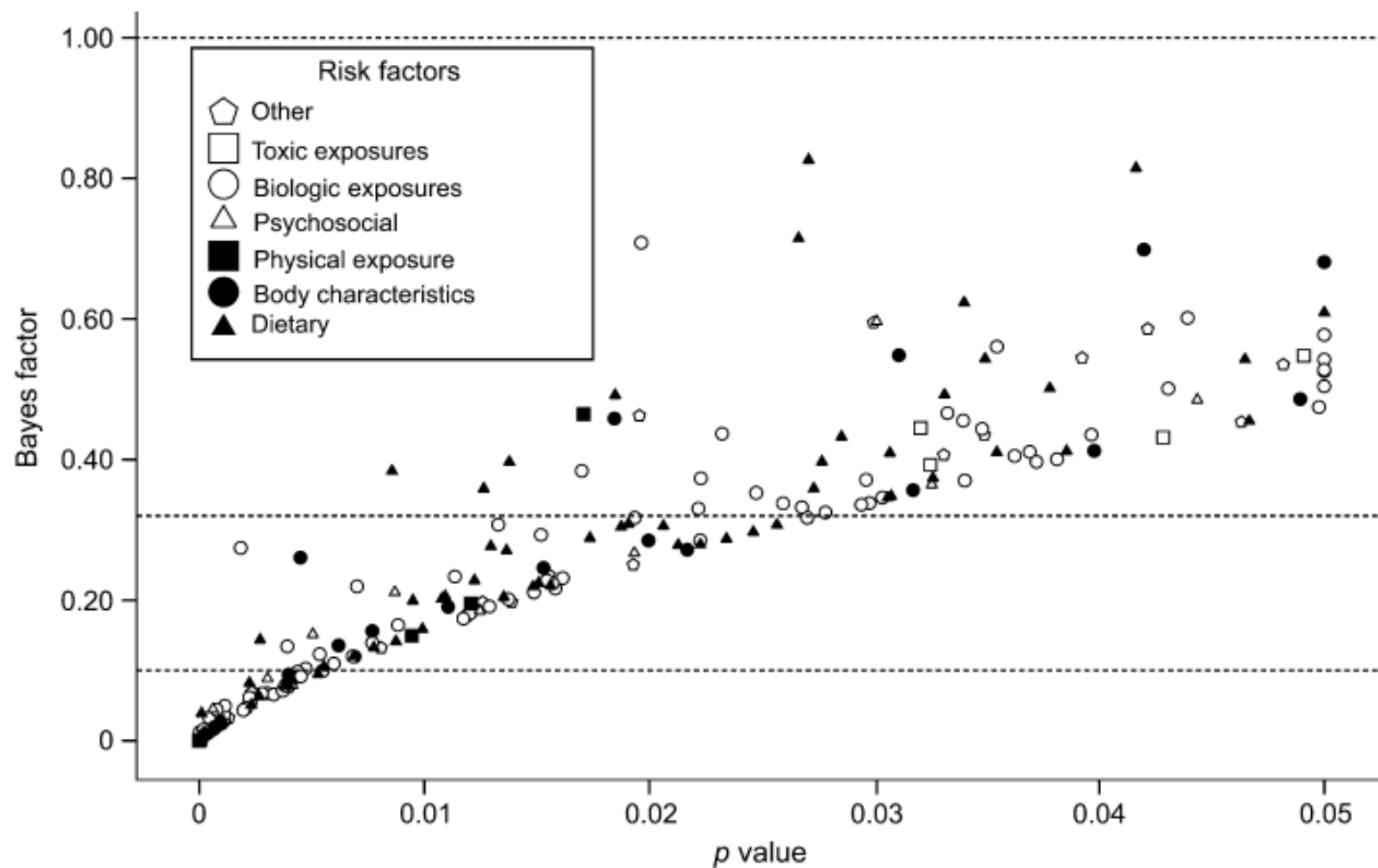
## General Conversion of $p$ -values to Bayes Factors

Robust Bayesian theory suggests a general and simple way to calibrate  $p$ -values. (Selke, Bayarri and Berger, 2001 Am. Stat.).

- A *proper*  $p$ -value satisfies  $H_0 : p(X) \sim \text{Uniform}(0, 1)$ .
- Consider testing this versus  $H_1 : p \sim f(p)$ , where  $Y = -\log(p)$  has a decreasing failure rate (a natural non-parametric alternative).
- **Theorem 1** *If  $p < e^{-1}$ ,  $B_{01} \geq -e p \log(p)$ .* (Vovk, 1993 JRSSB, derived this from a parametric alternative.)
- An analogous lower bound on the conditional Type I frequentist error is

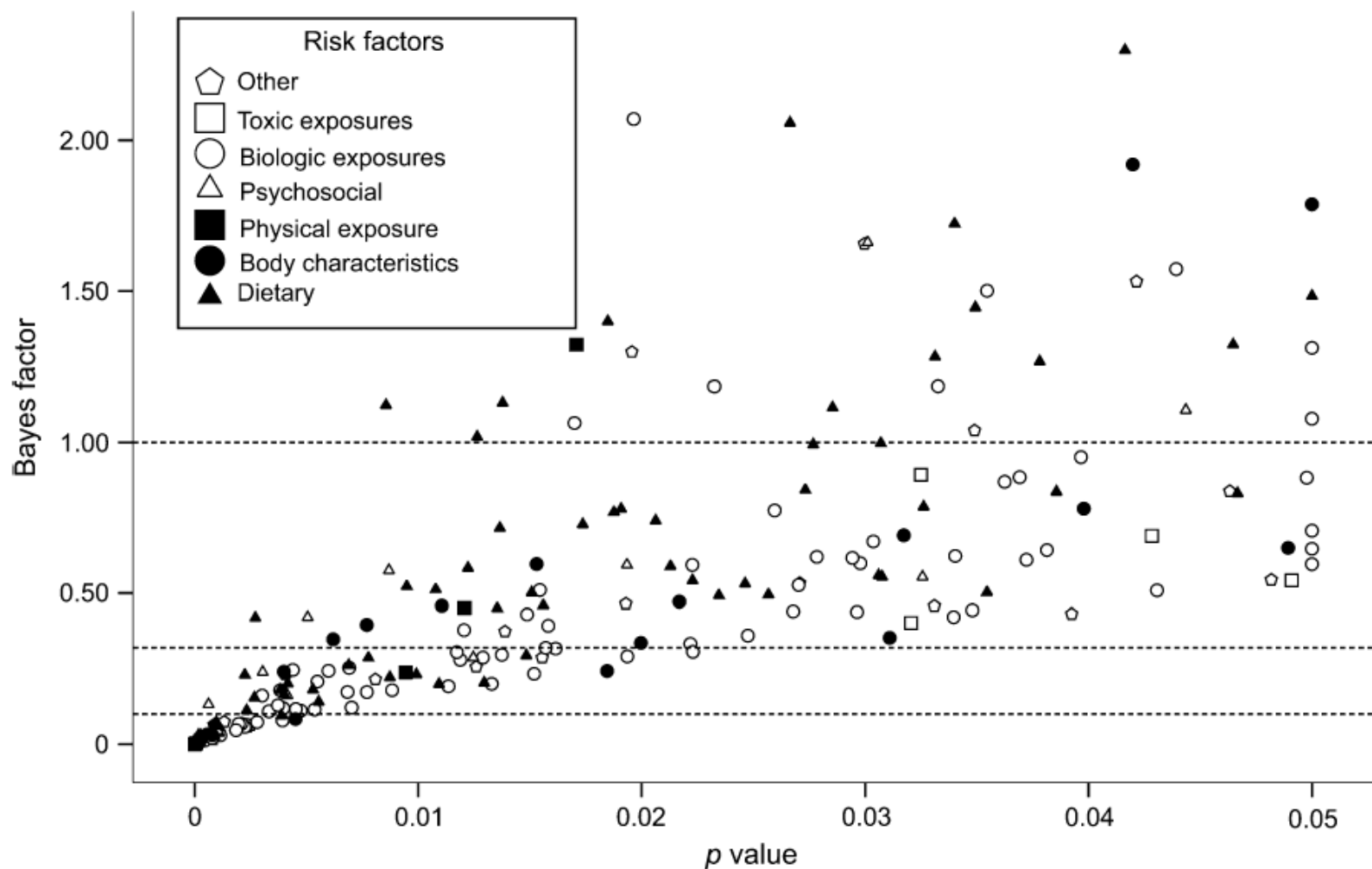
$$\alpha(p) \geq (1 + [-e p \log(p)]^{-1})^{-1}.$$

$p$	.2	.1	.05	.01	.005	.001	.0001	.00001
$-e p \log(p)$	.879	.629	.409	.123	.072	.0189	.0025	.00031
$\alpha(p)$	.465	.385	.289	.111	.067	.0184	.0025	.00031

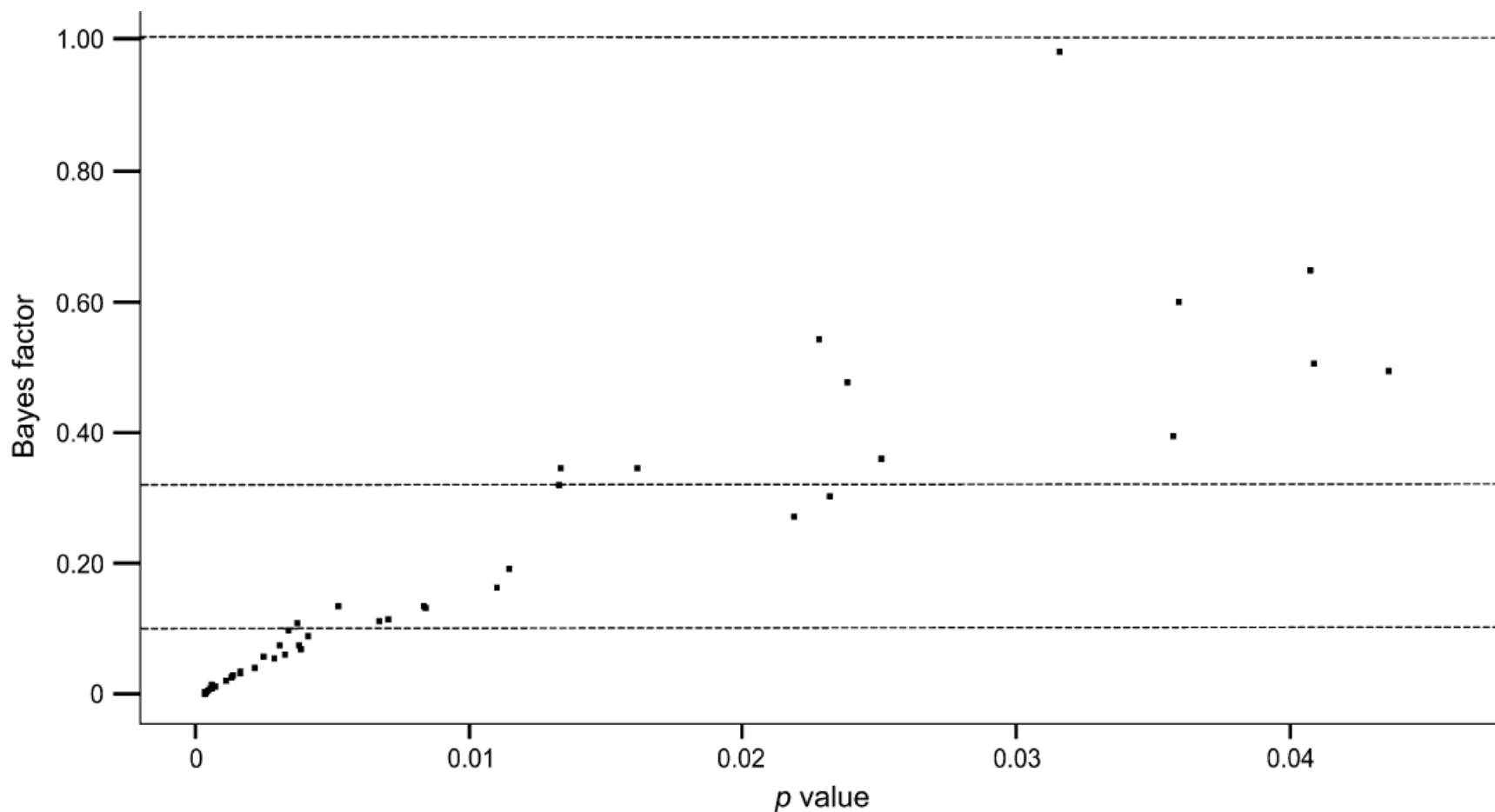


**FIGURE 1.** Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed  $p$  value in each study. Shown are calculations assuming  $\theta_A$  of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

Figure 4: J.P. Ioannides: Am J Epidemiol 2008;168:374–383



**FIGURE 2.** Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed  $p$  value in each study. Shown are calculations assuming  $\theta_A$  of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.



**FIGURE 3.** Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed  $p$  value in each meta-analysis. Calculations assume  $\theta_A$  equal to the median relative risk observed in the 50 genetic associations (relative risk = 1.44). The dashed lines correspond to threshold values (1.00, 0.32, 0.10) separating different Bayes factor categories.

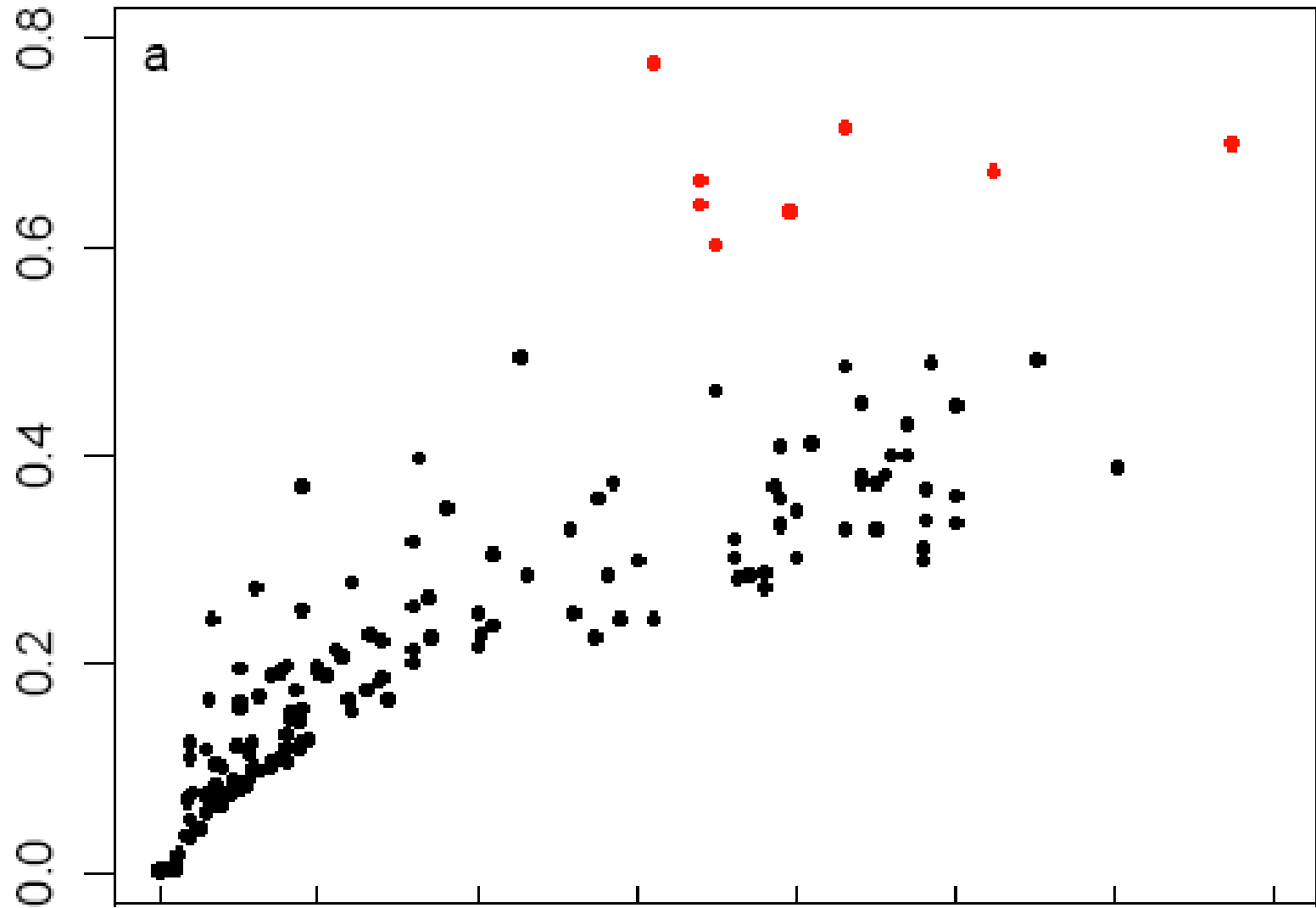


Figure 5: Elgersma and Green (2011):  $\alpha(p)$  versus observed  $p$ -values for 314 articles in Ecology in 2009.

## 2. The Bayesian Approach to Multiple Testing

**Key Fact:** Bayesian analysis deals with multiplicity testing solely through the assignment of prior probabilities to models or hypotheses.

### Example: Multiple Testing under Exclusivity

Suppose one is testing mutually exclusive hypotheses  $H_i$ ,  $i = 1, \dots, m$ , so each hypothesis is a separate model.

If the hypotheses are viewed as exchangeable, choose  $P(H_i) = 1/m$ .

**Example:** 1000 energy channels are searched for a signal:

- if the signal is known to exist and occupy only one channel, but no channel is theoretically preferred, each channel can be assigned prior probability 0.001.
- if the signal is not known to exist (e.g., it is the prediction of a non-standard physics theory) prior probability 1/2 should be given to ‘no signal,’ and probability 0.0005 to each channel.

*This is the Bayesian solution regardless of the structure of the data. In contrast, frequentist solutions depend on the structure of the data.*

## Example: Genome-wide Association Studies (GWAS)

- Early genomic epidemiological studies almost universally failed to replicate (estimates of the replication rate are as low as 1%), because they were doing multiple testing at ‘ordinary p-values’.
- A very influential paper in Nature (2007), by the Wellcome Trust Case Control Consortium, argues for a cutoff of  $p < 5 \times 10^{-7}$  ( $-ep \log(p) = 2.0 \times 10^{-5}$ ).
  - Derived from a Bayesian argument, with prior odds of an association set at 1/100,000.
  - Found 21 genome/disease associations; all but one have been replicated.
  - Later studies in GWAS have recommended cutoffs as low as  $5 \times 10^{-8}$  ( $-ep \log(p) = 2.3 \times 10^{-6}$ ).

## Summary 1. There is a lack of recognition that better statistics is the solution to much of the reproducibility problem

The extent of the problem:

- Dozens (hundreds) of articles addressing the problem; few say much about statistics (except those written by statisticians).
- Few journals adequately police the statistical analyses in their papers.  
“What’s the difference between ignorance and apathy?”  
“I don’t know and I don’t care.”
- An extreme illustration - *The Decline Effect* (see “The Truth Wears Off,” by Jonah Lehrer in the New Yorker, 2010):
  - This is the well-observed phenomenon that as more studies come in on something, the effect size declines.
  - This has been hypothesized to be a law of nature, like the uncertainty principle; scientists observing nature change nature.



## Summary 2: How Bayesian analysis can help

- While it may not be possible to replace  $p$ -values with Bayes factors, one can at least replace them with
  - $-ep \log(p)$ , termed the lower bound on the odds of no effect to there being an effect; or
  - $[1 + (-ep \log(p))^{-1}]^{-1}$ , termed the lower bound on the conditional frequentist Type 1 error.
- With Bayesian analysis there is no debate about a penalty for multiple tests, since prior probabilities are transparent.
- There is then no optional stopping issue; formal Bayesian answers do not depend on the stopping rule (although  $-ep \log(p)$  might).
- There is then a systematic way to deal with multiple statistical analyses, through Bayesian model averaging.

### Summary 3. Other efforts to address the reproducibility issue\*

- There have been a variety of efforts to establish protocols for scientific investigation:
  - Pre-experimental statements of intent and plan.
  - Documentation of all manipulations of data and all analyses attempted (e.g. Sweave); at a minimum, give the data.
  - Protocols for allowed methods of analysis.
- Efforts to allow publication of all results, positive or not.
- Optimal solution is to convince the science funding agencies to include statisticians on research teams, or at least provide funds for the data analysis, but this would require a radical expansion of statistics.
- Should statistical societies (as opposed to individual statisticians) police systemic bad statistical practice?

\*“I was going to buy a copy of *The Power of Positive Thinking*, and then I thought: What the hell good would that do?” –Ronnie Shakes

## II. A Brief History of (Objective) Bayesian Statistics

## The Reverend Thomas Bayes, began the objective Bayesian theory, by solving a particular problem

- Suppose  $X$  is Binomial  $(n,p)$ ; an 'objective' belief would be that each value of  $X$  occurs equally often.
- The only prior distribution on  $p$  consistent with this is the uniform distribution.
- Along the way, he codified Bayes theorem.
- Alas, he died before the work was finally published in 1763.



REV. T. BAYES

The real inventor of Objective Bayes was Simon Laplace (also a great mathematician, astronomer and civil servant) who wrote *Théorie Analytique des Probabilité* in 1812

- He virtually always utilized a 'constant' prior density (and clearly said why he did so).
- He established the 'central limit theorem' showing that, for large amounts of data, the posterior distribution is asymptotically normal (and the prior does not matter).
- He solved very many applications, especially in physical sciences.
- He had numerous methodological developments, e.g., a version of the Fisher exact test.



*Académie des Sciences*

6. Laplace in his robes as Chancellor of the Senate.

## What's in a name, part I

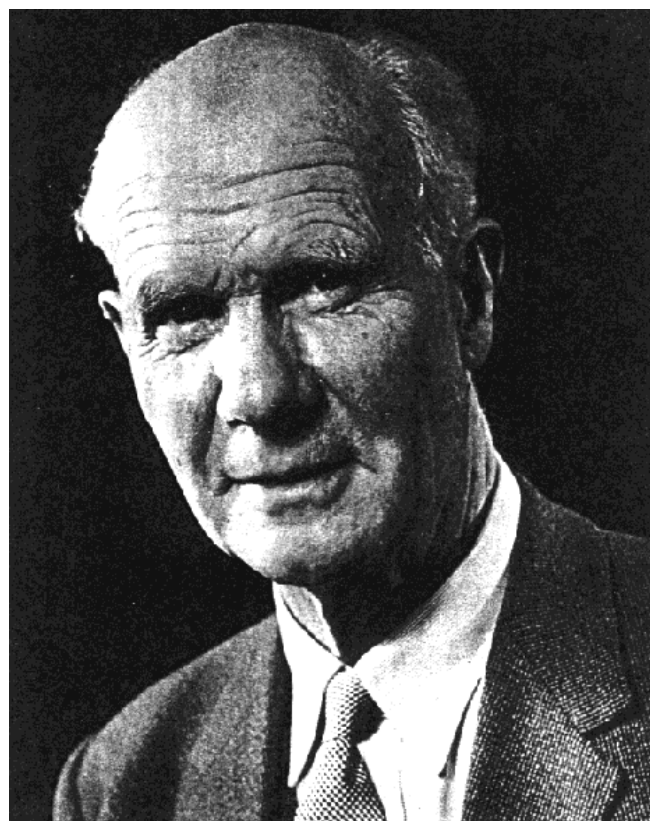
- It was called *probability theory* until 1838.
- From 1838-1950, it was called *inverse probability*, apparently so named by Augustus de Morgan.
- From 1950 on it was called *Bayesian analysis* (as well as the other names).



AUGUSTUS DE MORGAN

The importance of inverse probability b.f. (before Fisher): as an example, Egon Pearson in 1925 finding the 'right' objective prior for a binomial proportion

- Gathered a large number of estimates of proportions  $p_i$  from different binomial experiments
- Treated these as arising from the predictive distribution corresponding to a fixed prior.
- Estimated the underlying prior distribution (an early empirical Bayes analysis).
- Recommended something close to the currently recommended 'Jeffreys prior'  $p^{-1/2}(1-p)^{-1/2}$ .



EGON SHARPE PEARSON

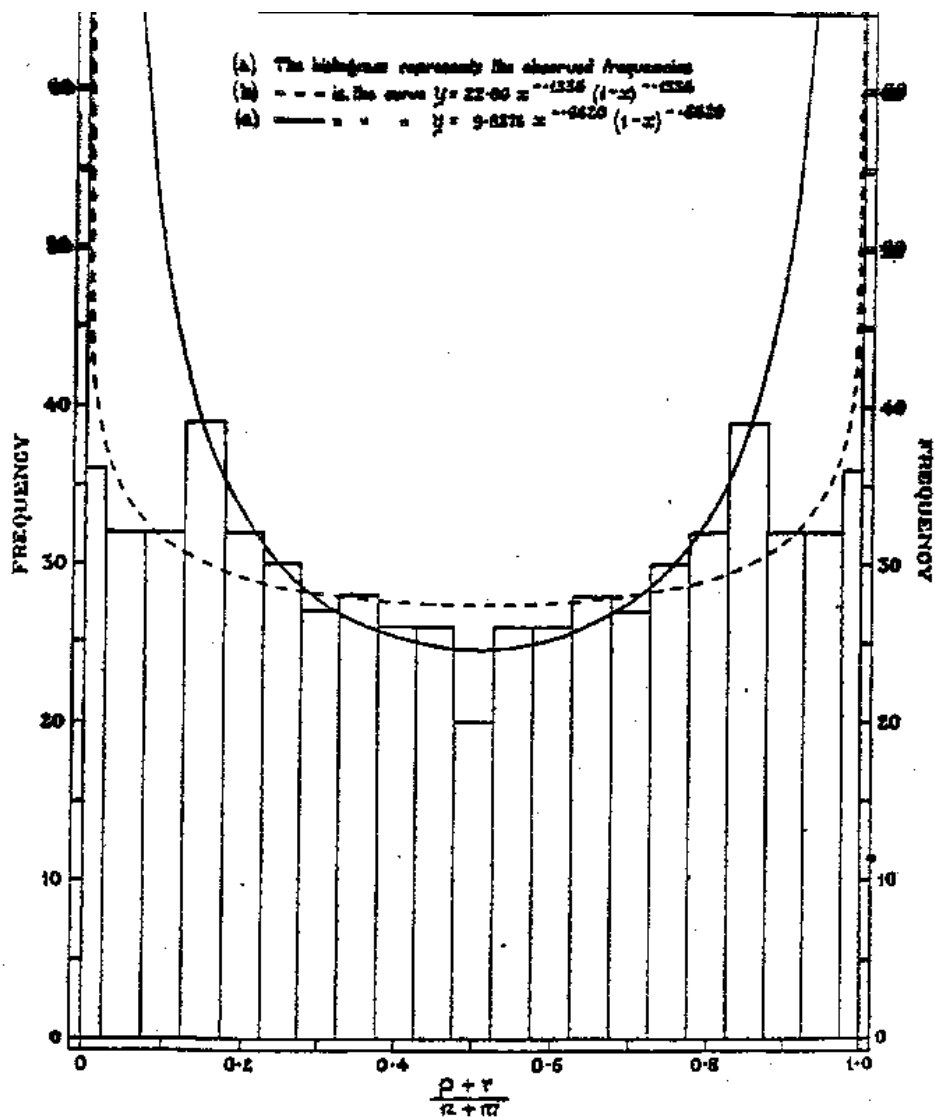
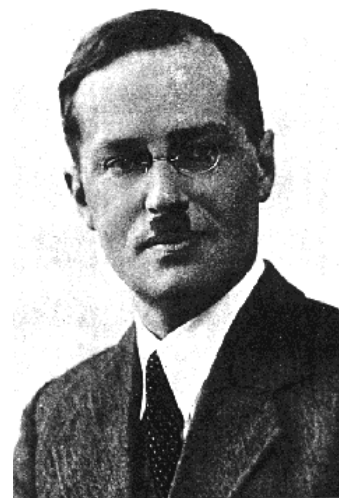


Fig. 3. Distribution of Frequencies of  $\frac{p+r}{n+m}$  in 300 samples (made symmetrical).



## 1930's: 'inverse probability' gets 'replaced' in mainstream statistics by two alternatives

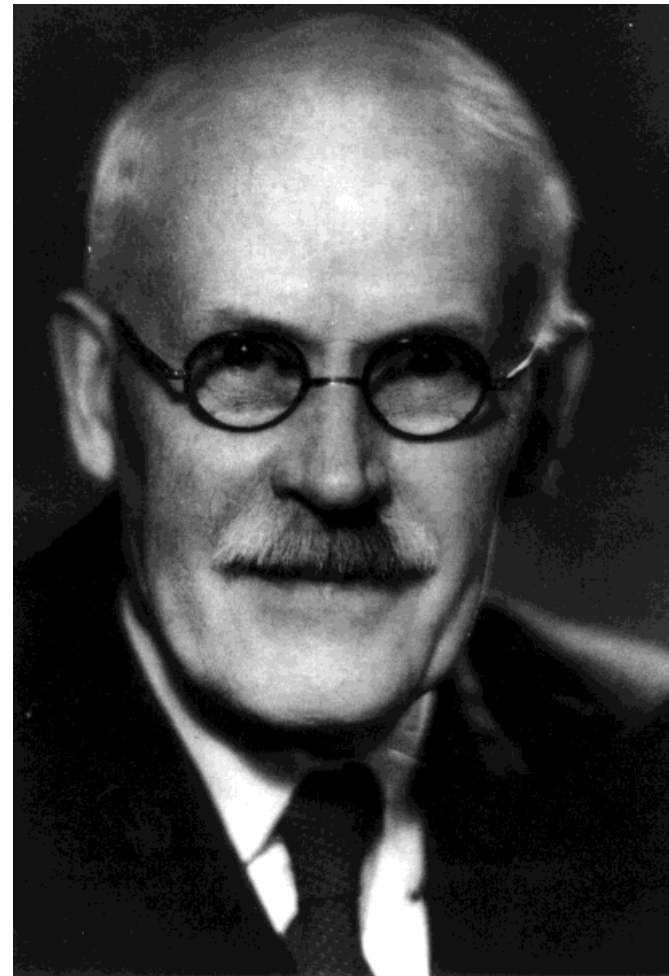
- For 50 years, Boole, Venn and others had been calling use of a constant prior logically unsound (since the answer depended on the choice of the parameter), so alternatives were desired.
- R.A. Fisher's developments of 'likelihood methods,' 'fiducial inference,' ... appealed to many.
- Jerzy Neyman's development of the frequentist philosophy appealed to many others.



JERZY NEYMAN

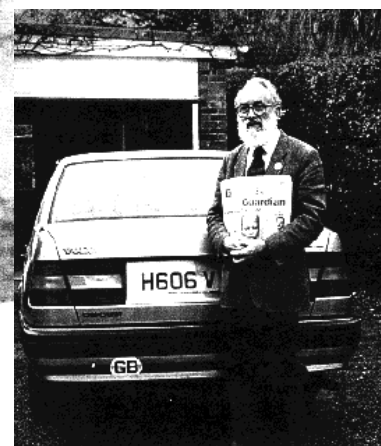
Harold Jeffreys (also a leading geophysicist) revived the Objective Bayesian viewpoint through his work, especially the *Theory of Probability* (1937, 1949, 1963)

- The now famous *Jeffreys prior* yielded the same answer no matter what parameterization was used.
- His priors yielded the ‘accepted’ procedures in all of the standard statistical situations.
- He began to subject Fisherian and frequentist philosophies to critical examination, including his famous critique of p-values: “An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.”



## What's in a name, part II

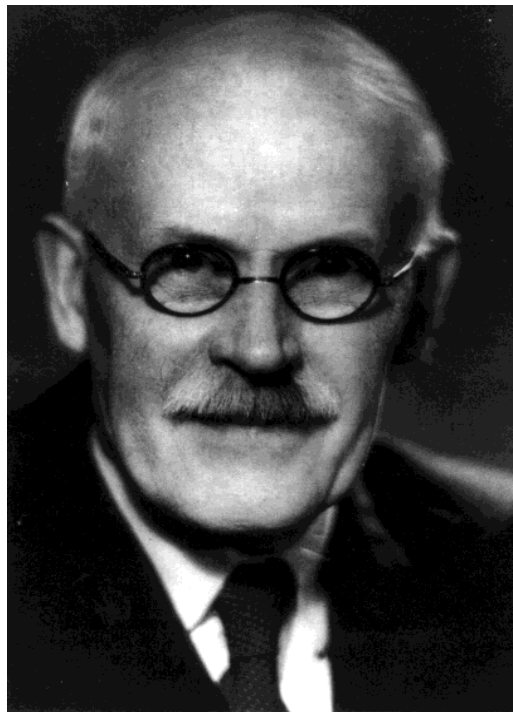
- In the 50's and 60's the *subjective* Bayesian approach was popularized (de Finetti, Rubin, Savage, Lindley, ...)
- At the same time, the *objective* Bayesian approach was being revived by Jeffreys, but Bayesianism became incorrectly associated with the subjective viewpoint. Indeed,
  - only a small fraction of Bayesian analyses done today heavily utilize subjective priors;
  - objective Bayesian methodology dominates entire fields of application today.



### **III. Could Fisher, Jeffreys and Neyman Have Agreed on Testing?**



Ronald Fisher



Harold Jeffreys

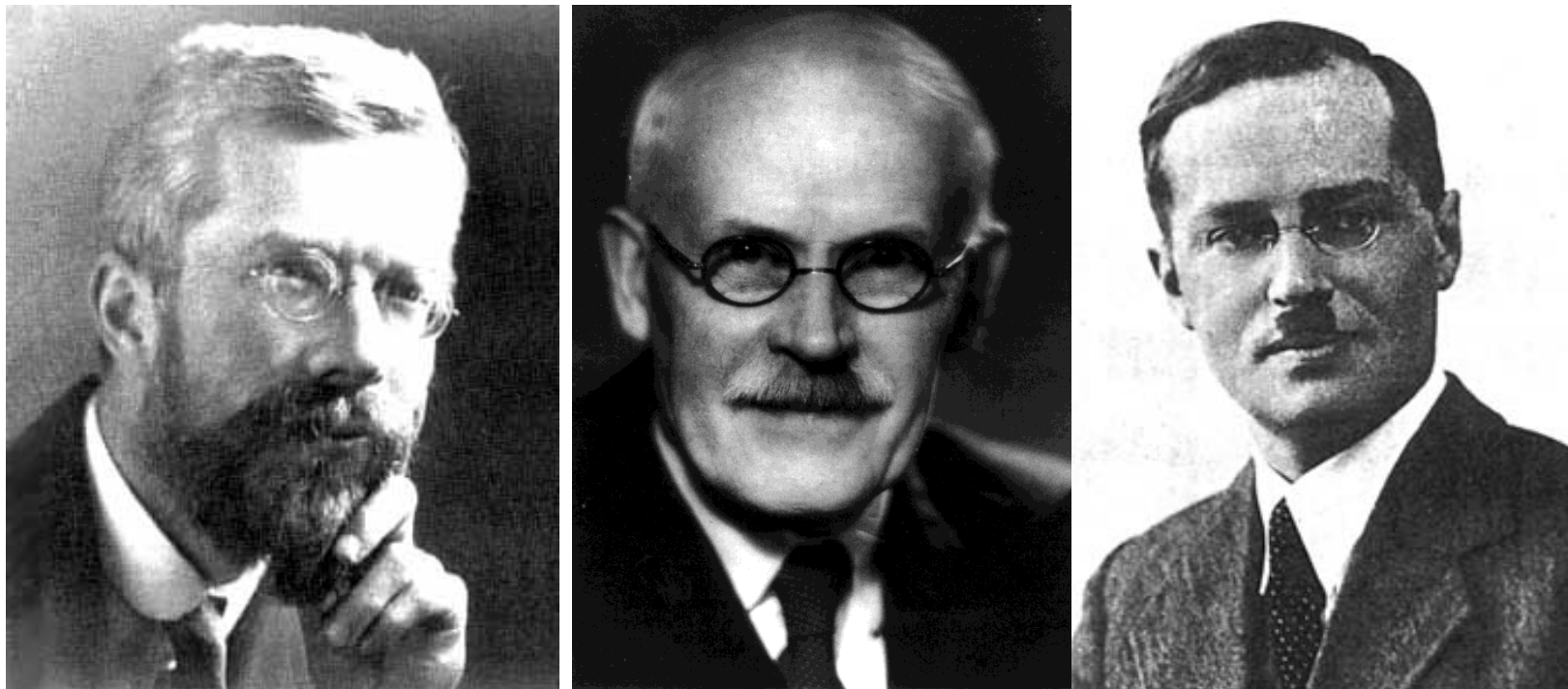


Jerzy Neyman

NO

NO

NO



Is there a statistical testing methodology that is compatible with the core elements of the statistical approaches of Fisher, Jeffreys, and Neyman?

## *Disagreements* and **Disagreements**

- Fisher, Jeffreys and Neyman *Disagreed* as to the correct foundations of statistics, but often agreed on the statistical procedure to use.
  - All supported use of the same estimation and confidence procedures for the normal linear model
  - *Disagreeing* only on the interpretation (e.g., to be assigned to the statement  $(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$  is a 95% confidence interval for a normal mean).
- In testing, however, they **Disagreed** as to the basic numbers to be reported. Was this unavoidable?

## Fisher's *significance testing*

Experiment yields  $X \sim f(x|\theta)$ ; test  $H_0 : \theta = \theta_0$ .



- Choose a test statistic  $T = t(X)$ , so that large values of  $T$  reflect evidence against  $H_0$ .
- Compute the  $p$ -value, for the *observed data*  $x$ ,

$$p = P_0(t(X) \geq t(x)),$$

rejecting  $H_0$  if  $p$  is small (e.g.,  $p \leq 0.05$ ).

- The justification is that the  $p$ -value can be viewed as an index of the ‘strength of evidence’ against  $H_0$ .

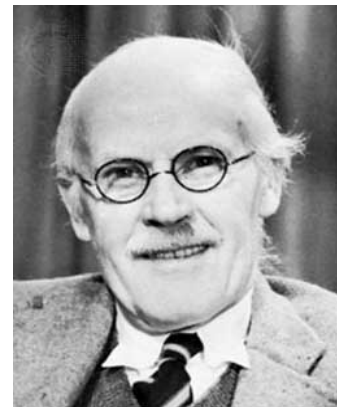


## Neyman-Pearson *hypothesis testing*



- Test  $H_0 : \theta = \theta_0$  versus *alternative*  $H_1 : \theta = \theta_1$ .
- Reject  $H_0$  if  $T > c$ ;  $c$  is a pre-chosen critical value.
- Compute Type I and Type II error probabilities  $\alpha = P_0(\text{rejecting } H_0)$  and  $\beta = P_1(\text{accepting } H_0)$ , where  $P_i$  refers to probability under  $H_i$ .
- The justification is the *Frequentist Principle*:  
In repeated actual use of a statistical procedure, the average actual error should not be greater than the average reported error.

# The Jeffreys approach to testing



- Define the Bayes factor (or likelihood ratio)

$$B(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

- Reject  $H_0$  (accept  $H_0$ ) as  $B(x) < 1$  ( $B(x) > 1$ ).
- Report the (objective) posterior probabilities

$$\Pr(H_0|x) = \frac{B(x)}{1 + B(x)} \quad \left( \text{or } \Pr(H_1|x) = \frac{1}{1 + B(x)} \right).$$

## The Disagreement

Fisher, Jeffreys and Neyman would report considerably different numbers in actual practice.

*Example:* In testing whether a normal mean is zero or not, with  $z = \frac{\bar{x}}{\sigma/\sqrt{n}} = 2.3$  (or  $z = 2.9$ ) ( $\sigma$  known,  $n = 10$ ),

- Fisher would report  $p = 0.021$  (or  $p = .0037$ ).
- Jeffreys would report the objective posterior probability  $\Pr(H_0|x) = 0.30$  (or  $\Pr(H_0|x) = 0.10$ )  
(using equal prior probabilities of the hypotheses and a conventional Cauchy(0, $\sigma$ ) prior on the alternative).
- Neyman, had he pre-specified  $\alpha = 0.05$ , would report  $\alpha = 0.05$  in either case (and a Type II error  $\beta$ ).

## Their criticisms of the other approaches

- Criticisms of Bayesian analysis in general. (Fisher and Neyman rarely addressed Jeffreys particular version of Bayesian theory.)
  - Fisher and Neyman rejected pre-Jeffreys objective Bayesian analysis as logically flawed.
  - Fisher's subsequent position was that Bayesian analysis is based on prior information that is only rarely available.
  - Neyman, in addition, assumed that Bayesian analysis violated the Frequentist Principle.

- Criticisms of Neyman-Pearson testing
  - Fisher and Jeffreys rejected N-P tests because they report fixed  $(\alpha, \beta)$ ; i.e., do not provide evidence or error measures that vary with the data.
  - Fisher disliked alternative hypotheses and power.
- Criticisms of Fisher's significance testing
  - Neyman and Jeffreys: alternative hypotheses are essential.
  - Neyman:  $p$ -values violate the Frequentist Principle.
  - $p$ -values are commonly misinterpreted as error rates, resulting in a considerable overestimation of the actual evidence against  $H_0$ .

One indication of the non-frequentist nature of  $p$ -values can be seen from the *applet* (of German Molina) at

WWW.STAT.DUKE.EDU/~BERGER

The situation considered follows Neyman's proposals to evaluate testing methods by simulations on a long series of *different* testing problems.

Suppose the  $i^{th}$  test consists of

- normal data with unknown mean  $\theta_i$  and known variance;
- testing of  $H_0 : \theta_i = 0$  versus  $H_1 : \theta_i \neq 0$ .

The applet simulates a long series of such tests, and records how often  $H_0$  is true for  $p$ -values in given ranges.

## Basis for Unification in Testing: the Conditional Frequentist Approach

**Basic question:** What is the sequence of possible data for which to consider frequentist evaluations?

(Fisher: “relevant subset;” Lehmann: “frame of reference.”)

**Artificial example:** Observe  $X_1$  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 for  $\theta$

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

**Unconditional coverage:**

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

This is 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  $\theta$  only with probability  $1/2$ .

**One must use conditional coverage:**

- Define the conditioning statistic  $S = |X_1 - X_2|$ , measuring the “strength of evidence” in the data (here ranging from  $s = 0$  to  $s = 2$ );
- Compute frequentist coverage conditional on the strength of evidence  $S$ .

$$P_{\theta}(C(X_1, X_2) \text{ contains } \theta \mid s = 2) = 1$$

$$P_{\theta}(C(X_1, X_2) \text{ contains } \theta \mid s = 0) = \frac{1}{2}.$$

*Note:* The unconditional coverage arises as the expected value of the conditional coverage.



## Conditional frequentist testing

- Find a statistic,  $S(x)$ , whose magnitude indicates the “strength of evidence” in  $x$ .
- Compute error probabilities as

$$\begin{aligned}\alpha(s) &= \text{Type I error prob, given } S(x) = s \\ &= P_0(\text{Reject } H_0 \mid S(x) = s)\end{aligned}$$

$$\begin{aligned}\beta(s) &= \text{Type II error prob, given } S(x) = s \\ &= P_1(\text{Accept } H_0 \mid S(x) = s)\end{aligned}$$

## History of conditional frequentist testing

- Many Fisherian precursors based on conditioning on ancillary statistics and on other statistics (e.g., the Fisher exact test, conditioning on marginal totals).
- General theory in Kiefer (1977 JASA); the key step was in understanding that almost *any* conditioning is formally allowed within the frequentist paradigm (ancillarity is not required).
- Brown (1978) found optimal conditioning frequentist tests for ‘symmetric’, simple hypotheses.
- Berger, Brown and Wolpert (1994 AOS) developed the theory discussed herein for testing simple hypotheses.

## Our suggested choice of the conditioning statistic

- Let  $p_i$  be the  $p$ -value from testing  $H_i$  against the other hypothesis; use the  $p_i$  as measures of ‘strength of evidence,’ as Fisher suggested.
- Define the conditioning statistic  $S = \max\{p_0, p_1\}$ ; its use is based on deciding that data (in either the rejection or acceptance regions) with the same  $p$ -value has the same ‘strength of evidence.’

## The conditional frequentist test ( $T^C$ )

- Accept  $H_0$  when  $p_0 > p_1$ , and reject otherwise.
- Compute Type I and Type II conditional error probabilities (CEPs) as

$$\alpha(s) = P_0(\text{rejecting } H_0 \mid S = s) \equiv P_0(p_0 \leq p_1 \mid S(X) = s)$$

$$\beta(s) = P_1(\text{accepting } H_0 \mid S = s) \equiv P_1(p_0 > p_1 \mid S(X) = s).$$

- The resulting test is given by

$$T^C = \begin{cases} \text{if } B(\mathbf{x}) \leq c, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x})); \\ \text{if } B(\mathbf{x}) > c, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})), \end{cases}$$

where  $c$  is the critical value at which the two  $p$ -values are equal.

## The potential *Agreement*

- The evidentiary content of  $p$ -values is acknowledged, but ‘converted’ to error probabilities by conditioning.
- The conditional error probabilities  $\alpha(s)$  and  $\beta(s)$  are fully data-dependent, yet fully frequentist.
- $\alpha(s)$  and  $\beta(s)$  are exactly equal to the (objective) posterior probabilities of  $H_0$  and  $H_1$ , respectively, so the conditional frequentists and Bayesians report the same error (Berger, Brown and Wolpert, 1994 AOS).

## A simple example

Sellke, Bayarri and Berger (2001 American Statistician)

- $H_0 : X \sim \text{Uniform}(0, 1)$  vs.  $H_1 : X \sim \text{Beta}(0.5, 1)$ .
- $B(x) = \frac{1}{(2\sqrt{x})^{-1}} = 2\sqrt{x}$  is the likelihood ratio (and Bayes factor).
- $p_0 = P_0(X \leq x) = x$  and  $p_1 = P_1(X \geq x) = 1 - \sqrt{x}$ .
- Accept  $H_0$  when  $p_0 > p_1$  (i.e., when  $x > .382$ ) and reject otherwise.
- Define  $S = \max\{p_0, p_1\} = \max\{x, 1 - \sqrt{x}\}$  (so it is declared that, say,  $x = \frac{3}{4}$  has the same “strength of evidence” as  $x = \frac{1}{16}$ ).
- The conditional test is

$$T^C = \begin{cases} \text{if } x \leq 0.382, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(x) = (1 + \frac{1}{2}x^{-1/2})^{-1}; \\ \text{if } x > 0.382, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(x) = (1 + 2x^{1/2})^{-1}. \end{cases}$$

- $\alpha(x)$  and  $\beta(x)$  are objective Bayes posterior probabilities of  $H_0$  and  $H_1$

## An extreme example of the difference between $p$ -values and CEPs

Observe  $X \sim N(x \mid 0, \sigma^2)$ . Test

$$H_0 : \sigma^2 = 1 \quad \text{versus} \quad H_1 : \sigma^2 = 1.1 .$$

The  $p$ -value is just the usual one, e.g.  $p = 0.05$  if  $x = 1.96$ .

The conditional frequentist error probability when rejecting  $H_0$  is

$$\alpha(x) = \left( 1 + (0.953)e^{-x^2/22} \right)^{-1} .$$

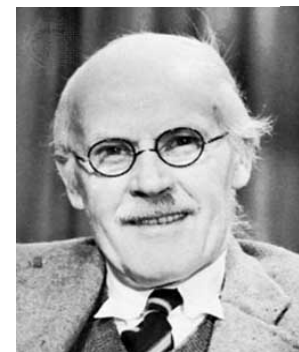
$x$	1.65	1.96	2.58	2.81	3.29	3.89	4.42
$p$	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
$\alpha(x)$	0.481	0.468	0.437	0.411	0.391	0.345	0.302

## An Aside: Other types of conditioning

- Ancillary  $S$  rarely exist and, when they exist, can result in unnatural conditional error probabilities (e.g., in the above example,  $\beta(x)$  equals the constant  $1/2$  as the likelihood ratio  $B(x)$  varies from 1 to 2).
- Birnbaum suggested ‘intrinsic significance,’ conditioning defined through likelihood concepts. This rarely works (e.g., in the above example  $\alpha(x) = 1$  when  $B(x)$  varies between 0 and  $1/2$ ).
- Kiefer (1977) suggested ‘equal probability continuum’ conditioning, which also fails in the above example.



Could Fisher, Jeffreys and Neyman have agreed on use of  $T^C$ ?



## Possible views of Jeffreys

- Since  $T^C$  can be interpreted as his objective Bayesian test, he certainly would have approved of its use from a methodological perspective.
- He would also have approved, for a composite alternative hypothesis, of reporting the ‘average’ Type II error, as  $T^C$  encourages. (Jeffreys felt that the dependence of the N-P power function on unknown parameters greatly limited its usefulness.)

# Possible views of Neyman



- He would not have been a priori hostile to  $T^C$ , it is fully frequentist within the N-P framework
- Another positive would be that one can use a classical rejection region (instead of the generic region  $p_0 < p_1$ ), if primary interest lies in the unconditional error – useful in settings, such as acceptance sampling, with contractual unconditional error rates; the unconditional error probabilities can certainly also be reported.
- But, while Neyman used conditioning as a technical tool for similar tests, he never embraced the concept.

# Possible views of Fisher



- It would be a positive that  $T^C$  is based on conditioning, with a reported error probability that varies continuously with the data.
- Use of  $p$ -values to measure the ‘strength of evidence’ in the data would be a positive, but using them to define a conditioning statistic is unusual. Still, Fisher quite often suggested unusual (non-ancillary) conditioning statistics, when they resulted in a simplified analysis (as here).
- The need for an alternative hypothesis to determine  $T^C$  would have been viewed as the biggest negative.

## Generalizations of Conditional Frequentist Testing

- Berger, Boukai and Wang (1997a, 1997b) generalized to simple versus composite hypothesis testing, including sequential settings.
- Dass (1998) generalized to discrete settings.
- Dass and Berger (2001) generalized to two composite hypotheses, under partial invariance.
- Berger and Guglielmi (2001) used the ideas to create a finite sample exact nonparametric test of fit.

## General Formulation (Dass and Berger, 2003 Scand. J. Stat.)

- Test  $H_0 : X$  has density  $p_0(x | \eta)$   
versus  $H_1 : X$  has density  $p_1(x | \eta, \xi)$   
on  $\mathcal{X}$ , with respect to a  $\sigma$ -finite dominating measure  $\lambda$ .
- $\{p_0(x | \eta) : \eta \in \Omega\}$  is invariant with respect to a group operation  $g \in \mathcal{G}$ .
- $g \circ (\eta, \xi) = (g \circ \eta, \xi)$  and  $\{p_1(x | \eta, \xi) : \eta \in \Omega\}$  is invariant with respect to  $g$  for each given  $\xi$ .
- Utilize the right-haar prior,  $\nu_{\bar{G}}(\eta)$ , on  $\eta$  and a proper prior  $\pi(\xi)$  on  $\xi$ .
- The original problem is reduced to testing distributions of the maximal invariant,  $T$ :

$$H_0 : T \sim f_0 \quad \text{versus} \quad H_1 : T \sim f_1,$$

where  $f_0(t) = \int p_0(x | \eta) d\nu_{\bar{G}}(\eta)$  and  $f_1(t) = \int p_1(x | \eta, \xi) d\nu_{\bar{G}}(\eta) d\pi(\xi)$ .

Since these are simple distributions, the rest of the analysis proceeds as before, and the equivalence between posterior probabilities and conditional frequentist error probabilities holds.

## Example: nonparametric testing of fit

Berger and Guglielmi (2001, JASA)

To test:  $H_0 : X \sim \mathcal{N}(\mu, \sigma)$  vs.  $H_1 : X \sim F(\mu, \sigma)$ ,  
where  $F$  is an unknown location-scale distribution.

- Define a weighted likelihood ratio (or Bayes factor)  $B(x)$ , by
  - Choosing a Polya tree prior for  $F$ , centered at  $H_0$ .
  - Choosing the right-Haar priors  $\pi(\mu, \sigma) = 1/\sigma$ .
- Choose  $S$  based on  $p$ -value conditioning.
- Define the conditional frequentist test as before, obtaining

$$T^C = \begin{cases} \text{if } B(\mathbf{x}) \leq c, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x})); \\ \text{if } B(\mathbf{x}) > c, & \text{accept } H_0 \text{ and report the 'average'} \\ & \text{Type II CEP } \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})), \end{cases}$$

where  $c$  is the critical value at which the two  $p$ -values are equal.

*Notes:*

- The Type I CEP again exactly equals the objective Bayesian posterior probability of  $H_0$ .
- The Type II CEP depends on  $(\mu, \sigma)$ . However, the ‘average’ Type II CEP is the objective Bayesian posterior probability of  $H_1$ .
- This gives *exact* conditional frequentist error probabilities, even for small samples.
- Computation of  $B(x)$  uses the (simple) exact formula for a Polya tree marginal distribution, together with importance sampling to deal with  $(\mu, \sigma)$ .

## Other features of conditional frequentist testing with $p$ -value conditioning

- Since the CEPs are also the objective Bayesian posterior probabilities of the hypotheses,
  - they do not depend on the stopping rule in sequential analysis, so
    - \* their computation is much easier,
    - \* one does not ‘spend  $\alpha$ ’ to look at the data;
  - there is no danger of misinterpretation of a  $p$ -value as a frequentist error probability, or misinterpretation of either as the posterior probability that the hypothesis is true.
- One has *exact* frequentist tests for numerous situations where only approximations were readily available before;
  - computation of a CEP is easier than computation of an unconditional error probability.



- One does not need to compute (or even mention)  $S$ .
  - The computation of the CEPs can be done directly, using the (often routine) objective Bayesian approach; the theory ensures the validity of the conditional frequentist interpretation.
  - Likewise, in elementary courses, one need not introduce  $S$ . Indeed, there is only *one* test

$$T^C = \begin{cases} \text{if } B(\mathbf{x}) \leq c, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha(\mathbf{x}) = B(\mathbf{x})/(1 + B(\mathbf{x})); \\ \text{if } B(\mathbf{x}) > c, & \text{accept } H_0 \text{ and report Type II CEP} \\ & \beta(\mathbf{x}) = 1/(1 + B(\mathbf{x})), \end{cases}$$

where  $c$  is the critical value at which the two  $p$ -values are equal.

- For  $T^C$ , the rejection and acceptance regions together span the entire sample space, so that one might ‘reject’ and report CEP 0.40.

One could, instead:

- Specify an ordinary rejection region  $R$  (say, at the  $\alpha = 0.05$  level); note that then  $\alpha = E[\alpha(s)1_R]$ .
- Find the ‘matching’ acceptance region, calling the region in the middle a *no decision* region;
- Construct the corresponding conditional test.
- Note that the CEPs would not change.

## An example of sequential conditional frequentist testing

(Berger, Boukai and Wang, 1998)

*Data:*  $X_1, X_2, \dots$  are i.i.d.  $N(\theta, \sigma^2)$ , both unknown, and arrive sequentially.

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

Note: Exact unconditional frequentist tests do not exist for this problem

**A default Bayes test** (Jeffreys, 1961):

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

Under  $H_0$ , prior on  $\sigma^2$  is  $g_0(\sigma^2) = 1/\sigma^2$ .

Under  $H_1$ , prior on  $(\mu, \sigma^2)$  is  $g_1(\mu, \sigma^2) = \frac{1}{\sigma^2} g_1(\mu | \sigma^2)$ , where  $g_1(\mu | \sigma^2)$  is Cauchy( $\theta_0, \sigma$ ).

**Motivation:** Under  $H_1$ , the additional parameter  $\theta$  is given a Cauchy default testing prior (centered at  $\theta_0$  and scaled by  $\sigma$ ), while the common (orthogonal) parameter  $\sigma^2$  is given the usual noninformative prior (see Jeffreys, 1961)

**Bayes factor** of  $H_0$  to  $H_1$ , if one stops after observing  $X_1, X_2, \dots, X_n$  ( $n \geq 2$ ), 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 simplification:** It is convenient to consider the sequential test in terms of the sequence  $B_1, B_2, \dots$ , instead of the original data. We monitor this sequence as data arrive, deciding when to stop the experiment and make a decision.

**Note:** If  $H_0$  and  $H_1$  have equal prior probabilities of  $1/2$ , then the posterior probability of  $H_0$  is  $B_n/(1 + B_n)$ .

**The conditional frequentist test:** A slight modification of this Bayesian  $t$ -test is a conditional frequentist test such that

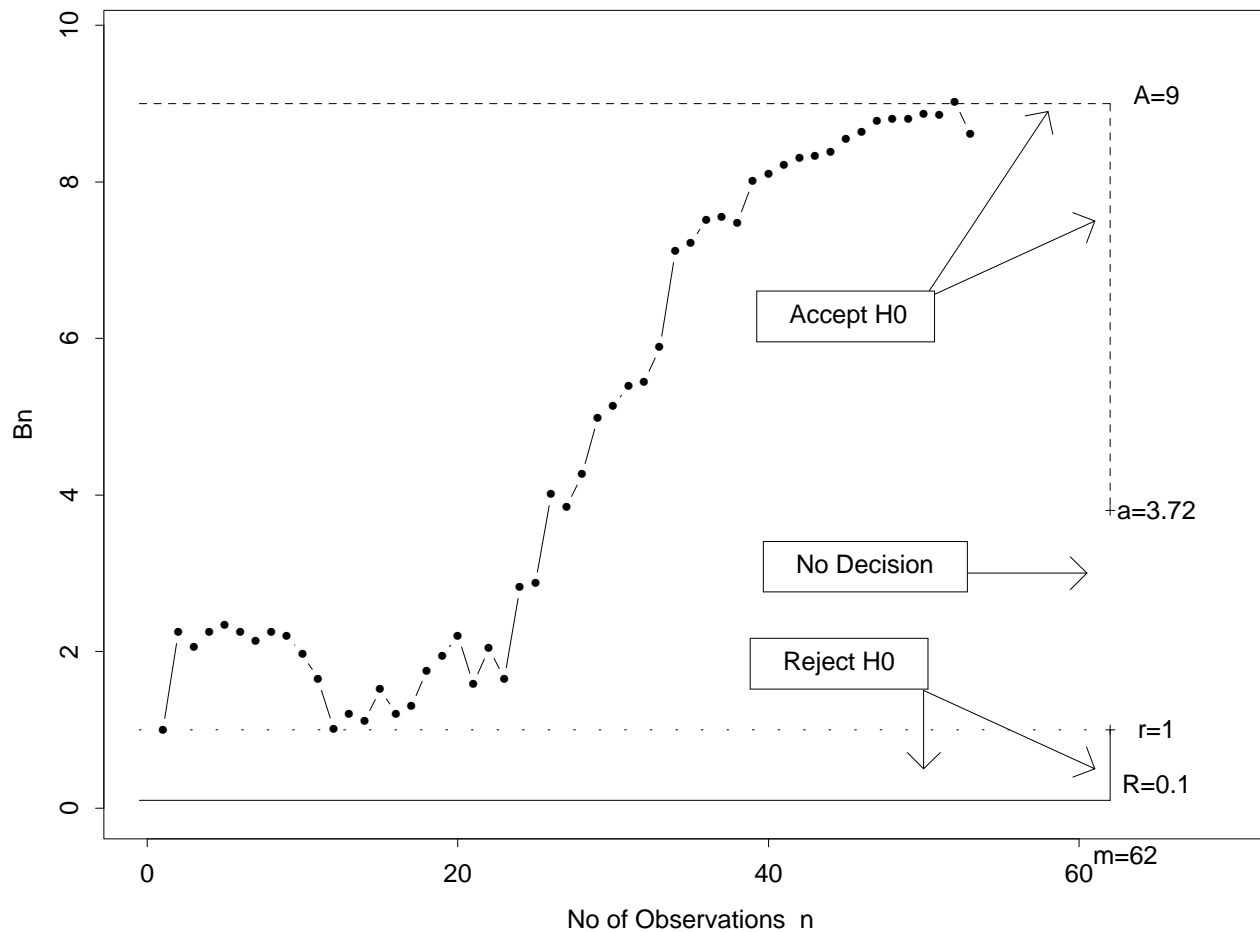
$$\begin{aligned}\alpha(s) &= P_{\theta_0}(\text{Type I error} \mid s) \\ &= \frac{B_N}{1+B_N} \quad (\text{also the posterior probability of } H_0),\end{aligned}$$

and the “average” Type II error is  $\frac{1}{1+B_N}$ , also the posterior probability of  $H_1$ .

(The modification,  $T_1^*$ , is the introduction of a “no decision” region in the acceptance region, where if  $1 < B_N < a$ , one states “no decision”.)

(Berger, Boukai and Wang, *Stat.Sci* 1997, *Biometrika* 1999)

*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.



**The 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  $H_1$ )

$M = 62$

**The test  $T_1^*$**  : If  $N$  denotes the stopping time,

$$T_1^* = \begin{cases} \text{if } B_N \leq 1, & \text{reject } H_0 \text{ and report Type I CEP} \\ & \alpha^*(B_N) = B_N/(1 + B_N); \\ \text{if } 1 < B_N < a, & \text{make no decision;} \\ \text{if } B_N \geq a, & \text{accept } H_0 \text{ and report the 'average'} \\ & \text{Type II CEP } \beta^*(B_N) = 1/(1 + B_N). \end{cases}$$

**Example:**  $a = 3.72$  (found by simulation). 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 error probability  $\beta^*(B_{52}) = 1/(1 + 9.017) \approx 0.100$ .



*Comments About the Sequential Test:*

- $T_1^*$  is fully frequentist (and Bayesian).
- The conclusions and stopping boundary all have simple intuitive interpretations.
- Computation is easy (except possibly computing  $a$ , but it is rarely needed in practice):
  - No stochastic process computations are needed.
  - Computations do not change as the stopping rule changes.
  - Sequential testing is as easy as fixed sample size testing.
- $T_1^*$  essentially follows the Stopping Rule Principle, which states that, upon stopping experimentation, inference should not depend on the reason experimentation stopped.
- This equivalence of conditional frequentist and Bayesian testing applies to most classical testing scenarios.