Reproducibility of Science: P-values and Multiplicity

Jim Berger

Duke University

8^{th} International Purdue Symposium on Statistics
June 21, 2012

Outline

- 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

I. 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

II. 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.

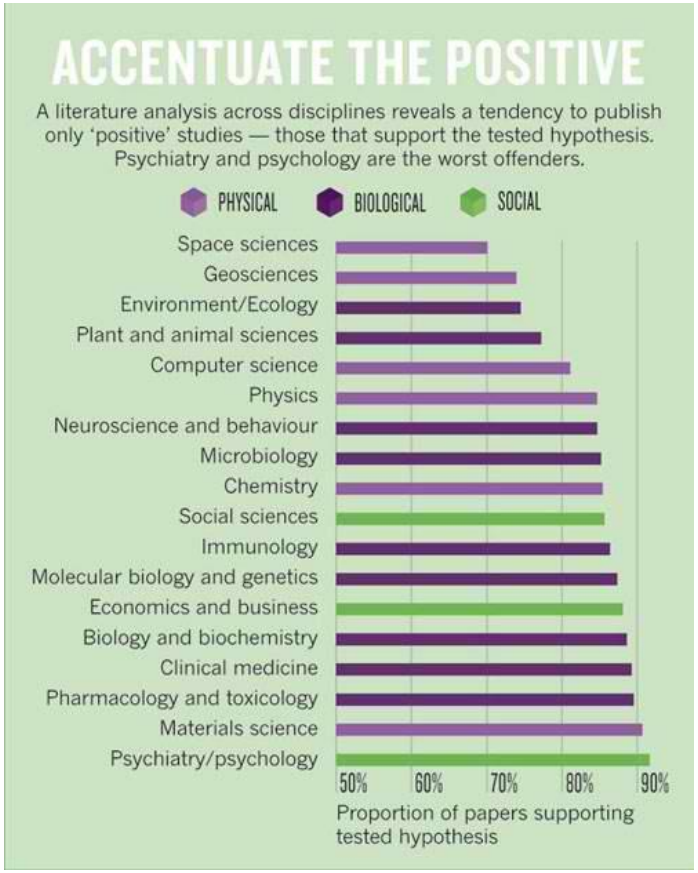


Figure 1: From Fanelli, D. Scientometrics 90, 891–904 (2011).

- 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.

2. Experimental biases:

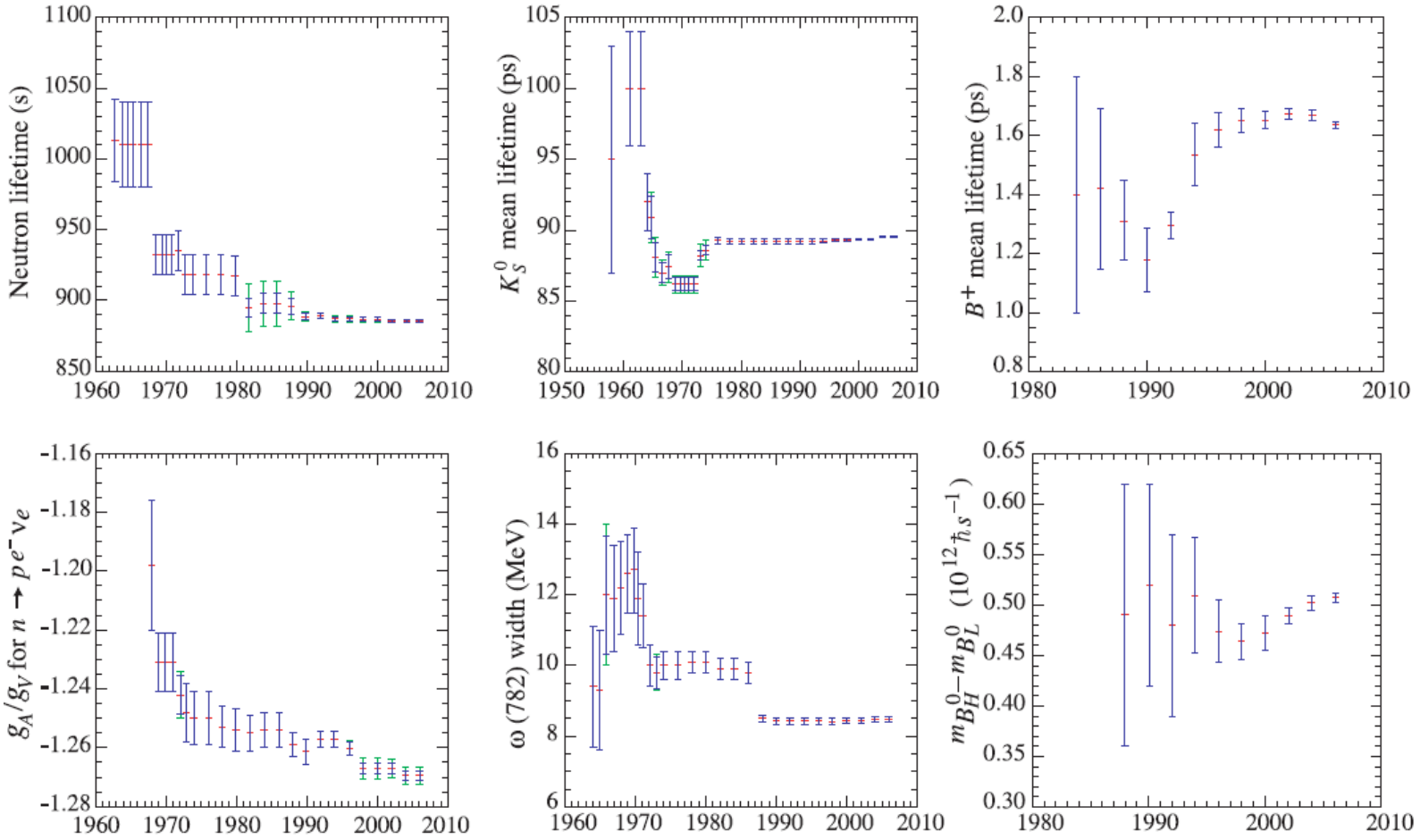


Figure 2: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

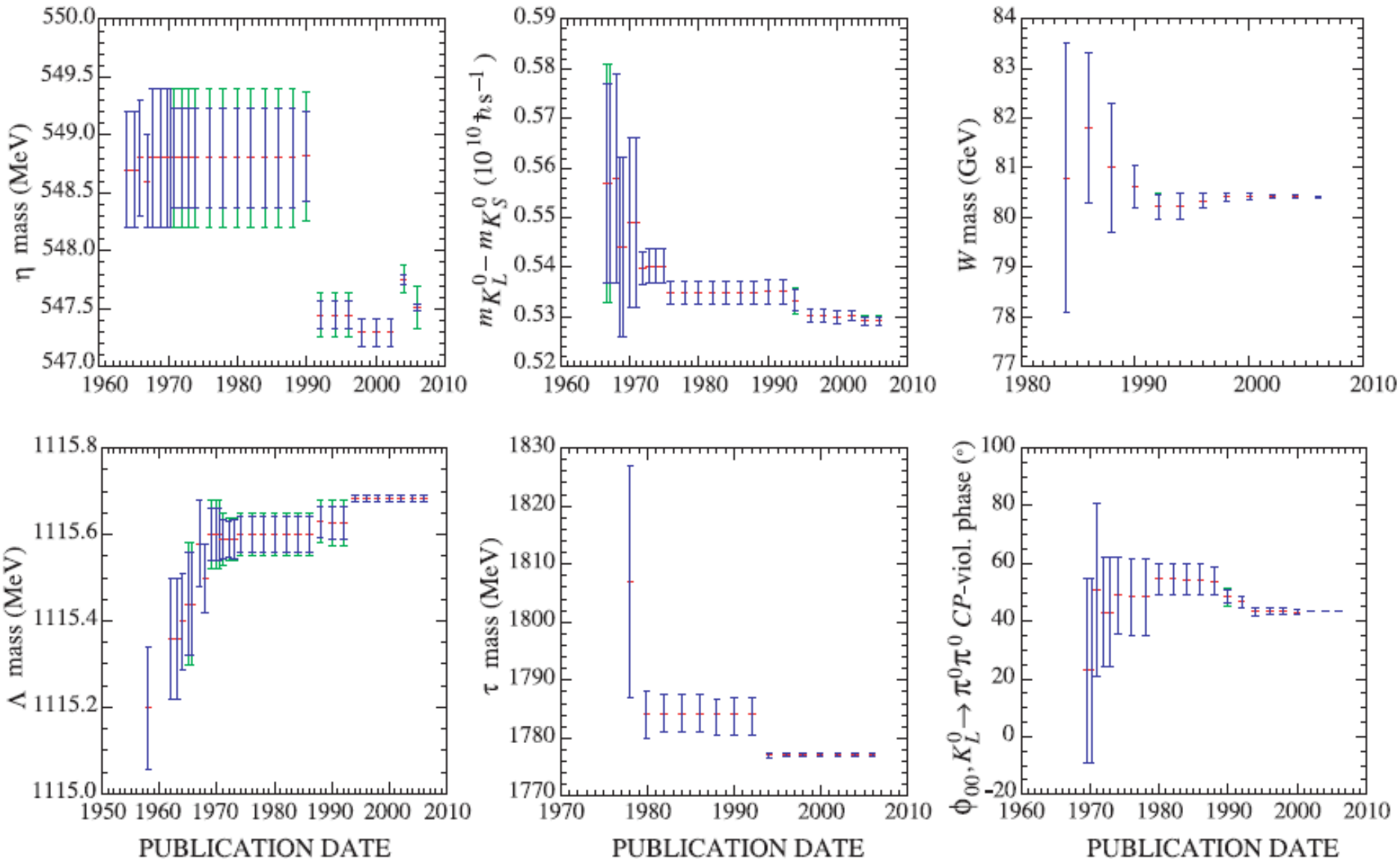


Figure 3: Historical record of values of some particle properties published over time, with quoted error bars (Particle Data Group).

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 *professional* penalty for having a positive finding later refuted.
  - What is the situation in this regard with statistics?

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.
  - See the *applet* (of German Molina) available at www.stat.duke.edu/∼berger.

7. Failure to adjust for multiplicities:

- Failure to properly account for multiple testing:
  "Basic research is like shooting an arrow in the air and, where it lands, painting a target." Homer Adkins
  - 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.'
  - *Multiple Multiple Testing* (e.g., the same 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 under the null is only 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.

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

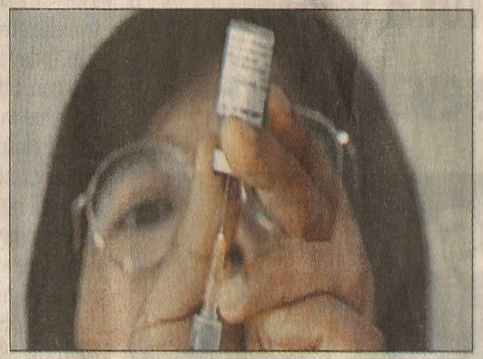
Hours after HIV researchers announced the achievement of a milestone that had eluded them for a quarter of a century, reality began to set in: Tangible progress could take another decade.

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-

See **VACCINE**, *Page 14*



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

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 infection in the placebo and treatment populations, respectively.
- Test $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$
- 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 $\mathrm{N}(\theta, 1)$, where $\theta = (p_1 - p_2)/(.001359)$.
  Test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, based on z.
- Observed $z = 2.06$, so the p-value is 0.02.

Bayesian analysis:

Posterior odds of H_1 to H_0 = [Prior odds of H_1 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,

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

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

Frequentist perspective for odds of correct to incorrect rejection:

Let α and $(1-\beta(\theta))$ be the Type I error and power for testing H_0 versus H_1 with, say, rejection region $\mathcal{R} = \{z : z > 1.645\}$. Then

$$\begin{aligned} O &= \text{Odds of } \textit{correct rejection} \text{ to } \textit{incorrect rejection} \\ &= [\text{prior odds of } H_1 \text{ to } H_0] \times \frac{(1-\bar{\beta})}{\alpha}, \end{aligned}$$

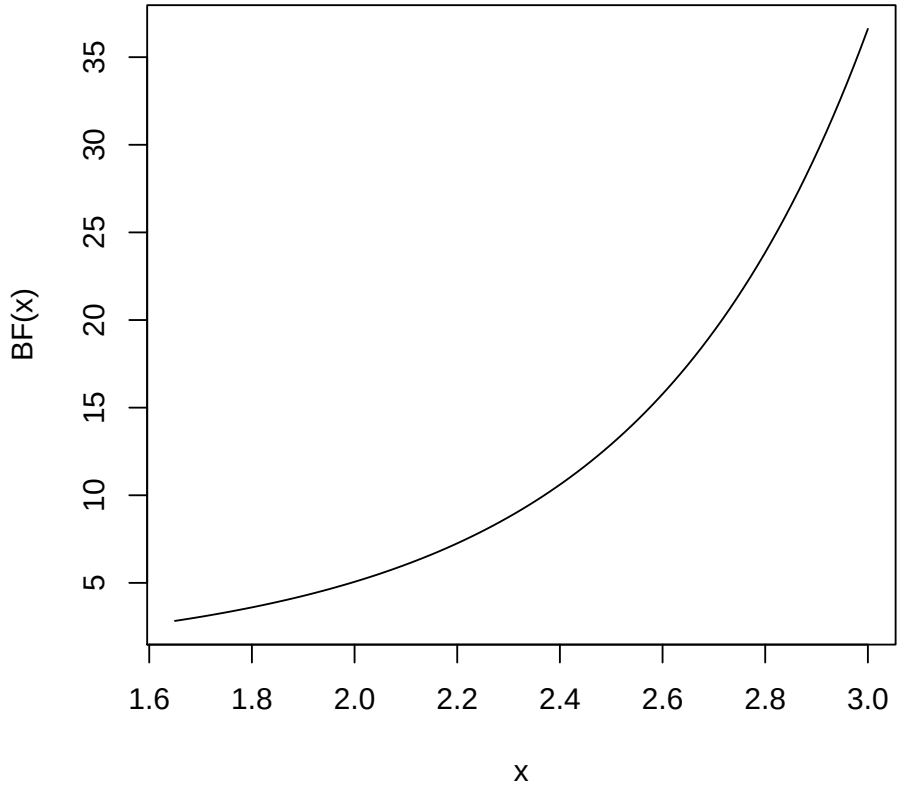
where $(1-\bar{\beta}) = \int (1-\beta(\theta))\pi(\theta)d\theta$ is average power wrt the prior $\pi(\theta)$.

- $\frac{(1-\bar{\beta})}{\alpha} = \frac{\text{average power}}{\text{type 1 error}}$ is the *experimental odds* of correct rejection to incorrect rejection.
- For vaccine example, $(1-\bar{\beta}) = 0.45$ and $\alpha = 0.05$ (the error probability corresponding to $\mathcal{R}$), so $\frac{(1-\bar{\beta})}{\alpha} = 9$.

average power	0.05	0.25	0.50	0.75	1.0	0.01	0.25	0.50	0.75	1.0
type I error	0.05	0.05	0.05	0.05	0.05	0.01	0.01	0.01	0.01	0.01
correct/incorrect	1	5	10	15	20	1	25	50	75	100

But that is pre-experimental; better is to report the actual data-based odds of correct rejection to incorrect rejection, namely the Bayes factor $B_{10}(z)$.

- For vaccine example, here is $B_{10}(z)$ (recall $\frac{(1-\bar{\beta})}{\alpha} = 9$):



- Reporting the Bayes factor is a valid *conditional frequentist* procedure (Kiefer, 1977 JASA, Brown, 1978 AOS) because

$$E[B_{10}(Z) \mid H_0, \mathcal{R}] = \frac{(1-\bar{\beta})}{\alpha} \quad \text{and} \quad E[B_{01}(Z) \mid H_1^*, \mathcal{R}] = \frac{\alpha}{(1-\bar{\beta})}.$$

A General Bound

Robust Bayesian theory suggests a general and simple way to calibrate p-values. (Sellke, 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* $p < e^{-1}$, $B_{01} \geq -e\, p\, \log(p)$.
- 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
$-ep\log(p)$	.879	.629	.409	.123	.072	.0189	.0025	.00031
$\alpha(p)$	.465	.385	.289	.111	.067	.0184	.0025	.00031

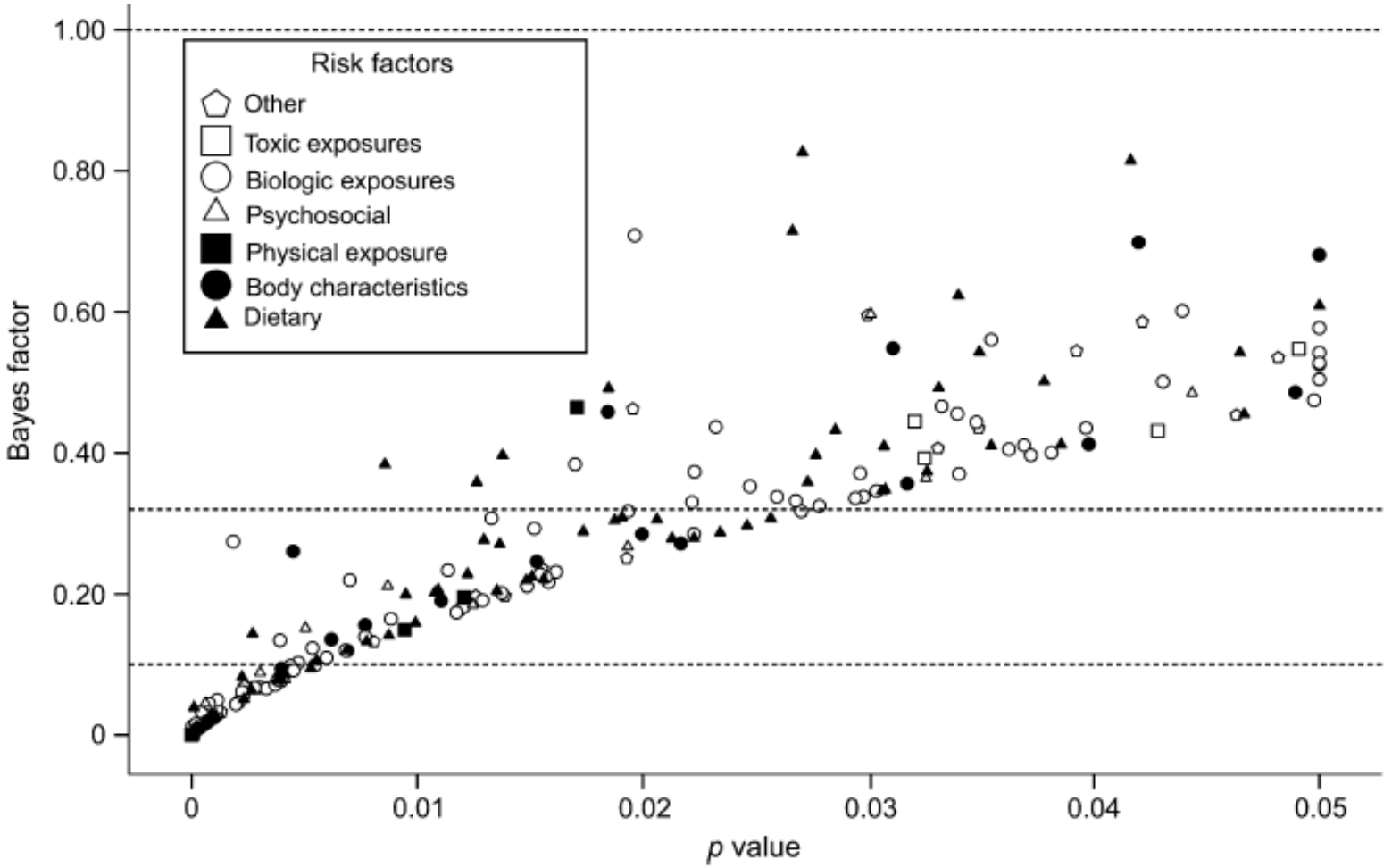


FIGURE 1. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 0.50 (relative risk = 1.65). The dashed lines correspond to threshold values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

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

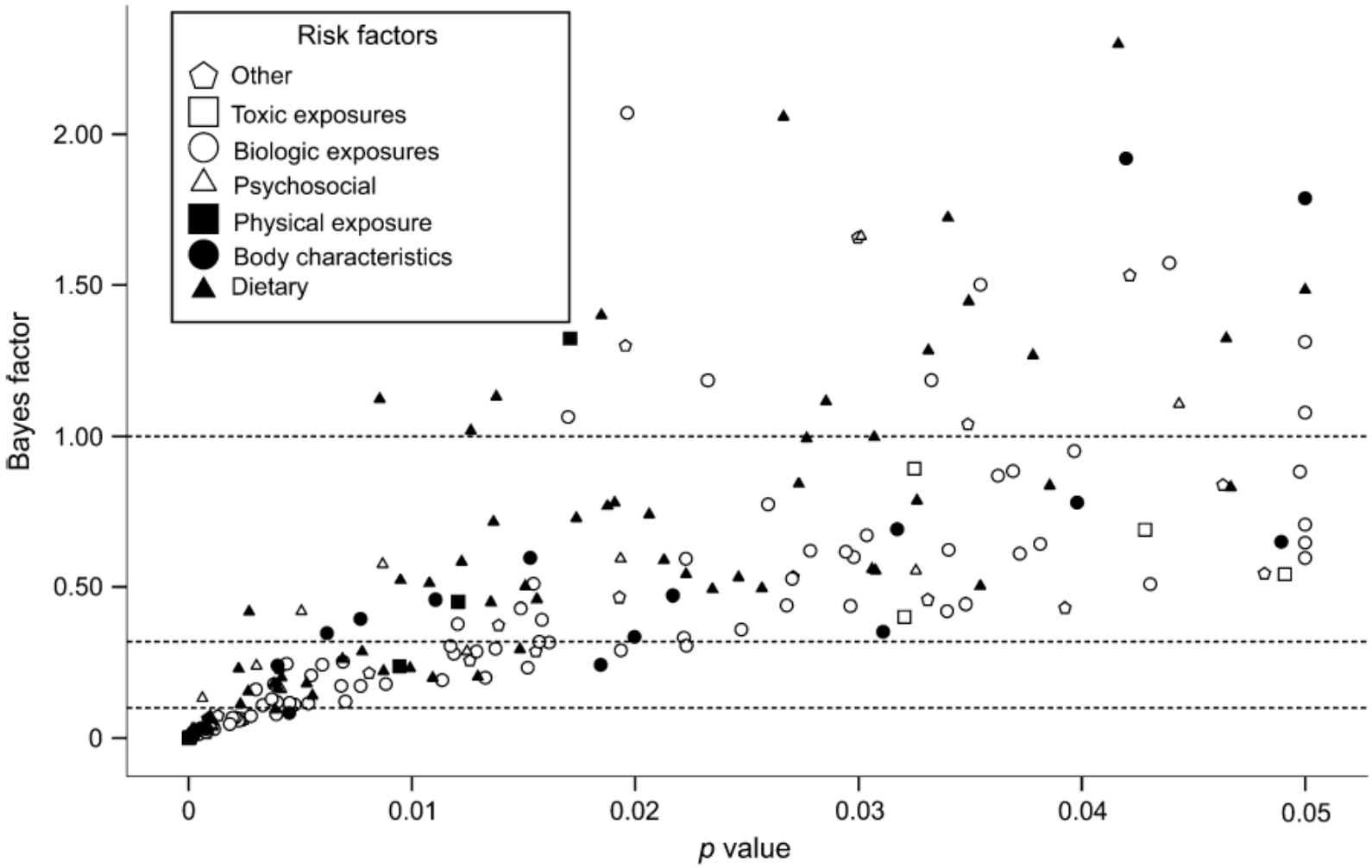


FIGURE 2. Estimated Bayes factors for 272 epidemiologic studies with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each study. Shown are calculations assuming θ_A of 1.50 (relative risk = 4.48). The dashed lines correspond to threshold values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

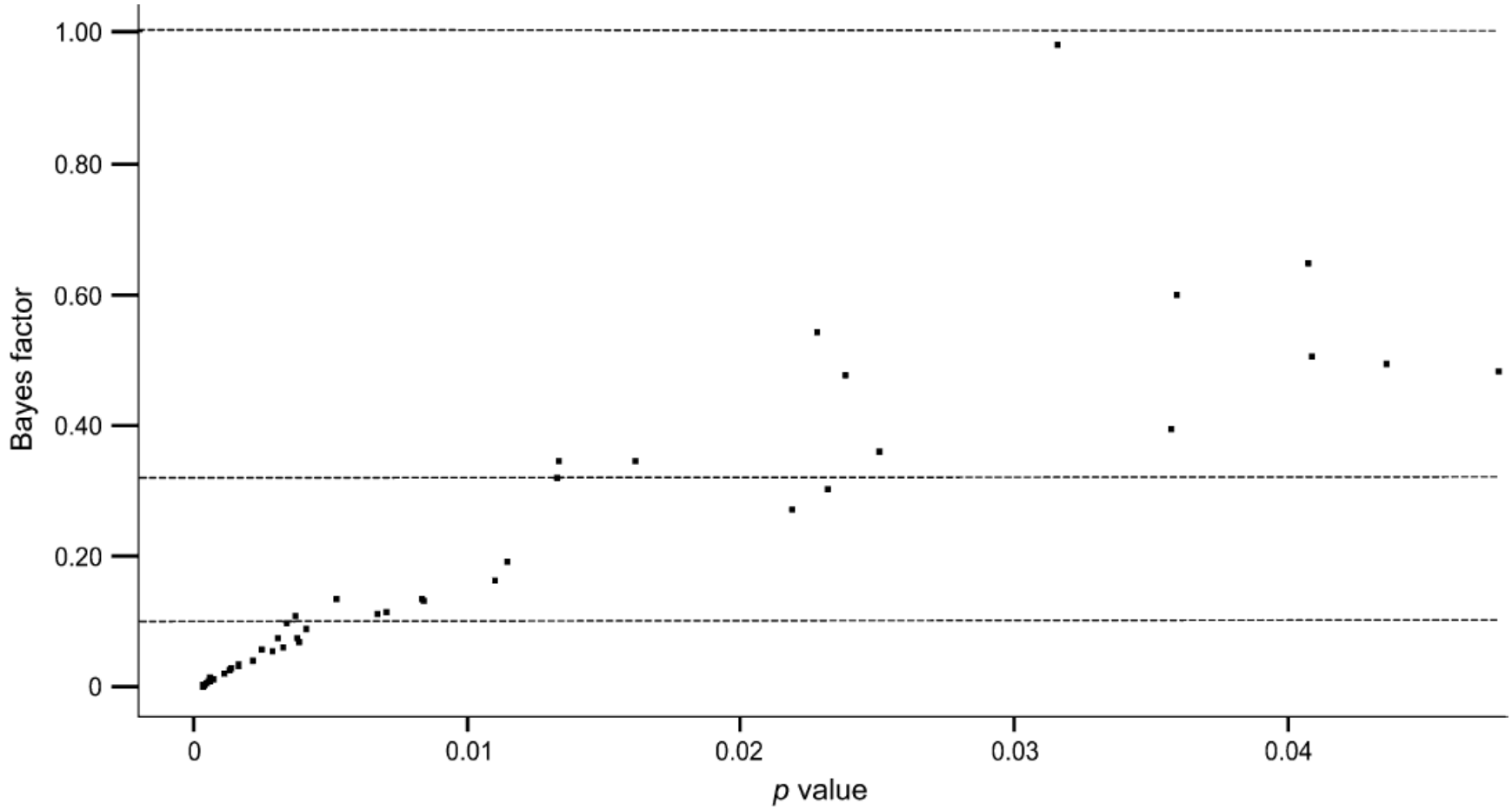


FIGURE 3. Estimated Bayes factors for 50 meta-analyses of genetic associations with formally statistically significant results. The Bayes factor is plotted against the observed *p* value in each meta-analysis. Calculations assume θ_A equal to the median relative risk observed in the 50 genetic associations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

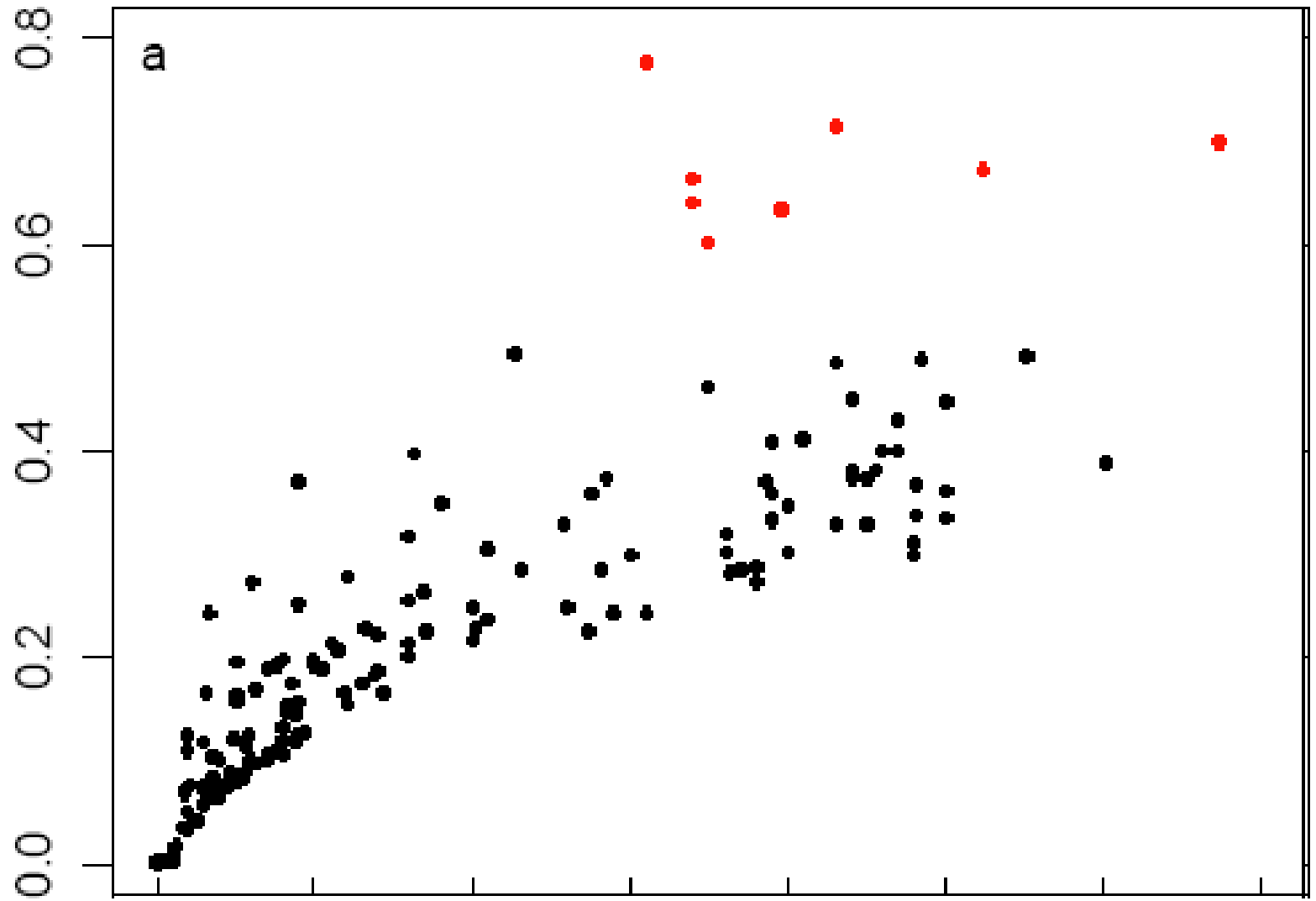


Figure 5: Elgersma and Green (2011): $\alpha(p)$ versus observed p-values for 314 articles in Ecology in 2009.

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 $H_i,\ i = 1, \ldots, m$, 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, 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 strongly on feature of the data such as their dependence structure, making them challenging to implement.

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 proposed cutoff $p < 5 \times 10^{-7}$ $(-ep\log(p) = 2 \times 10^{-5})$
  - Found 21 genome/disease associations; 20 have been replicated.
- Bayes argument for the cutoff:
  - *Pre-experimental 'odds of true positive to false positive'*
    $$= \textit{prior odds} \times \frac{(1-\bar{\beta})}{\alpha}.$$
  - For the GWAS study, they choose $\textit{prior odds} = \frac{1}{100{,}000}$ and $(1-\bar{\beta}) = 0.5$, giving odds of $10:1$ in favor of a true positive if $\alpha = 5 \times 10^{-7}$. (They stated the prior odds could vary by a factor of 10.)
  - The article also argued that it is better to just compute the Bayes factors $B_{10}(z)$, and the posterior odds $= \textit{prior odds} \times B_{10}(z)$. These ranged between $\frac{1}{10}$ and 10^{68} for the 21 claimed associations.

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

George Casella, 1951-2012

- Being at the forefront of promoting good statistical practice
- Testing and p-values: papers on
  - understanding and proper use of p-values
  - conditional frequentist methods
  - objective Bayesian alternatives
- Bayesian multiplicity control
  - "Objective Bayesian analysis of multiple changepoints for linear models" (*Bayesian Statistics,* 2007)
  - "Consistency of objective Bayes factors as the model dimension grows" (*AOS*, 2010)
  - Many genomics papers
  - Software (BAMD)