# Opinionated Lessons in Statistics

by Bill Press

## #14 Bayesian Criticism of P-Values

**Here are three Bayesian criticisms of tail tests:**

(1) Their result depends on the choice of test or (more argumentatively) what was in the mind of the experimenter

These are called “stopping rule paradoxes”.

Hypothesis $H_0$: a coin is fair with P(heads)=0.5

Data: in 10 flips, the first 9 are heads, then 1 tail.

Analysis Method I. Data this extreme, or more so, should occur under $H_0$ only

$$\frac{1+10+10+1}{2^{10}} = 0.0214$$

(you lose: referee wants p<0.01 and tells you to get more data)

## Analysis method II.

*"I forgot to tell you," says the experimenter, "my protocol was to flip <u>until a tail</u> and record N (=9), the number of heads."*

Under $H_0$ $\quad p(N) = 2^{-(N+1)}$

$$p(\geq N) = 2^{-(N+1)}(1 + \tfrac{1}{2} + \tfrac{1}{4} + \cdots) = 2^{-N}$$

$$P(\geq 9) = 2^{-9} = 0.00195$$

(You win. Your paper get's published.)

Stopping rule effects are a serious methodological issue in biomedical research, where for ethical reasons stopping criteria may depend on outcomes in complicated and unpredictable ways, or be ad hoc after the experiment starts (and rightly so – see next slide!)

April 8, 2006

# British Rethinking Rules After Ill-Fated Drug Trial

By ELISABETH ROSENTHAL,
International Herald Tribune

In February, when Rob O. saw the text message from Parexel International pop up on his cellphone in London — "healthy males needed for a drug trial" for £2,000, about $3,500 — it seemed like a harmless opportunity to make some much-needed cash. Parexel, based in Waltham, Mass., contracts with drug makers to test new medicines.

Just weeks later, the previously healthy 31-year-old was in intensive care at London's Northwick Park Hospital — wires running directly into his heart and arteries, on dialysis, his immune system, liver, kidneys and lungs all failing — the victim of a drug trial gone disastrously bad.

One of six healthy young men to receive TGN1412, a novel type of immune stimulant that had never before been tried in humans, Rob O. took part in a study that is sending shock waves through the research world and causing regulators to rethink procedures for testing certain powerful new drugs.

Although tests of TGN1412 in monkeys showed no significant trouble, all six human subjects nearly died. One is still hospitalized and the others, though discharged, still have impaired immune systems, their future health uncertain.

On Wednesday, after releasing its interim report on the trial as well as previously confidential scientific documents that were part of the application for a trial permit, the British government announced it was convening an international panel of experts to "consider what necessary changes to clinical trials may be required" for such novel compounds.

The outcome "could potentially affect clinical trials regulation worldwide," the announcement said. In statements this week, both Parexel and the drug's manufacturer, TeGenero, emphasized that they had complied with all regulatory requirements and conducted the trial according to the approved protocol. But they declined to answer questions e-mailed to them about the specifics of the science involved.

"The companies have worked according to strict standards applicable for such type of studies," said Kristin Kaufmann, a spokeswoman for TeGenero.

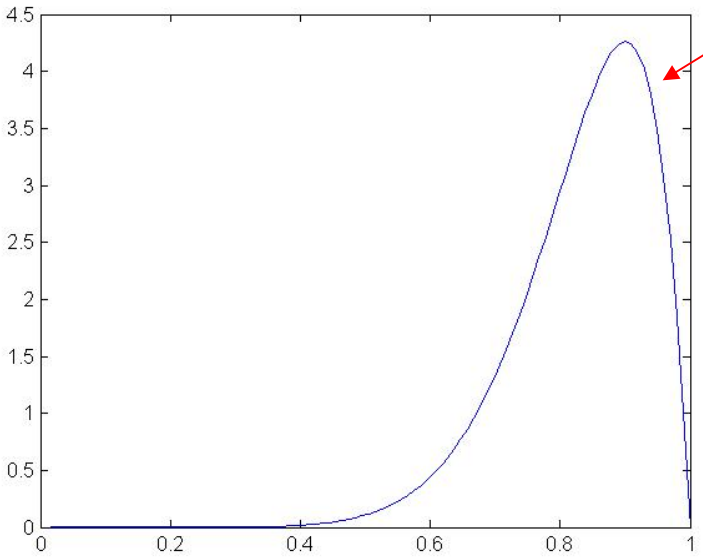
What would be a Bayesian approach?

$H_p$ is the hypothesis that prob = $p$.

$P(H_p)$ is its probability.

$$P(H_p|\text{data}) \propto P(\text{data}|H_p)P(H_p) \propto p^9(1-p)$$

$$P(H_p|\text{data}) = \frac{p^9(1-p)}{\int_0^1 p^9(1-p)dp}$$




The curve is the answer.
We might, however, summarize it in various ways:

Likelihood (or posterior probability) ratio:

$$\frac{P(H_{0.5}|\text{data})}{P(H_{\max}|\text{data})} = \frac{0.1074}{4.2616} = 0.0252$$

Bayes tail probability:

$$\int_0^{0.5} P(H_p|\text{data})dp = 0.0059$$

For an example in which we might use a more sophisticated prior, suppose the data is 10 heads in a row.

*"Hmm. When people make me watch them flip coins, 95% of the time it's a (nearly) fair coin [A], 4% of the time it's a double-headed [B] or double-tailed coin [C], and 1% of the time something else weird is happening [D]."*

| | | |
|---|---|---|
| Case A: | $0.95 \times (0.5)^{10} = 0.00093$ | $0.043$ |
| Case B | $0.02 \times 1^{10} = 0.02$ | $0.915$ |
| Case C | $0.02 \times 0^{10} = 0$ | $0.000$ |
| Case D | $0.01 \times \int_0^1 p^{10} dp = 0.00091$ | $0.042$ |

This kind of analysis can be dignified by the term "meta-analysis" if you can justify your choice of priors on the basis of already published data. (Somewhat more rigorously than the above.) However, it is also a good way to live your life, especially if you are a person who likes to make bets!

(Can you remember that we were listing three Bayesian criticisms of tail tests?)

(2) Not suitable for comparing hypotheses quantitatively. Best you can do is rule one out, leaving the other viable. Ratio of p-values is not anything meaningful!

you should go learn about Likelihood Ratio tests, but I personally think that Bayes odds ratio is easier to compute and easier to interpret

(3) The sanctification of certain p-values (e.g., the magic p=0.05 value) is naïve and misleading.

(on the one hand) 1 in 20 results are wrong! Imagine if we built nuclear power plants to this low a standard.

(on the other hand) the large majority of results with p=0.10 are in fact correct. These could sometimes be acted on.

Slavish adherence to p=0.05 is largely due to the young Fisher (who became arguably the greatest statistician to ever have lived).

Ronald Aylmer Fisher (1890-1962)

Fisher studied with Gossert (Student) as a young man. Gossert never approved of the p=0.05 rule, and understood as the Master Brewer that no single p-value was suitable for optimizing economic return: it depends on the relative costs of success and failure (origins of decision theory).

There is a fun article on this posted in the course web site:

*Journal of Economic Perspectives—Volume 22, Number 4—Fall 2008—Pages 199–216*

**Retrospectives**

Guinnessometrics: The Economic Foundation of "Student's" *t*

Stephen T. Ziliak