Just call it a "p-value"!

(not a hypothesis probability,
not a false alarm probability)

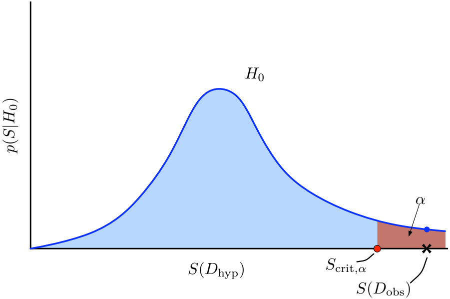
Tom Loredo
Dept. of Astronomy, Cornell University
`http://www.astro.cornell.edu/staff/loredo/`

CASt Summer School — 5 June 2014

Significance Testing and p-values

Neyman-Pearson testing

- Specify simple null hypothesis H_0 such that rejecting it implies an interesting effect is present
- Devise statistic $S(D)$ measuring departure from null
- Divide sample space into probable and improbable parts (for H_0); $p(\text{improbable}|H_0) = \alpha$ (Type I error rate), with α specified a priori
- If $S(D_{\rm obs})$ lies in improbable region, reject H_0; otherwise accept it
- Report: "H_0 was rejected (or not) with a procedure with false-alarm frequency α"



Neyman and Pearson devised this approach guided by Neyman's *frequentist principle*:

> *In repeated practical use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error. (Berger 2003)*

A *confidence region* is an example of a familiar procedure satisfying the frequentist principle.

They insisted that one also specify an alternative, and find the error rate for falsely rejecting it (Type II error).

Fisher's p-value testing

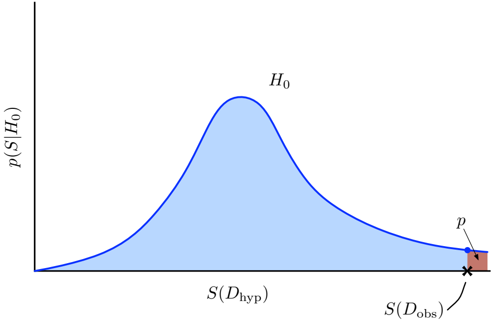
Fisher (and others) felt reporting a rejection frequency of α no matter where $S(D_{\rm obs})$ lies in the rejection region does not accurately communicate the strength of evidence against H_0.

He advocated reporting the *p-value*:

$$p = P(S(D) > S(D_{\rm obs})|H_0)$$

Smaller p-values indicate stronger evidence against H_0.

Astronomers call this the *significance level* or (sometimes) *false-alarm probability*. Statisticians don't—for good reason!



p-value complications

Fisherian testing does not have the straightforward frequentist properties of NP testing, but everyone uses it anyway.

E.g., rejections of H_0 with p-value$= 0.05$ are *not* "wrong 5% of the time under the null" or "with 5% false-alarm probability." They are wrong *100% of the time* under the null. To quantify the conditional error rate (i.e., the error rate among datasets with the same p-value), you *must* say something about the alternative.

Even NP tests have unpleasant frequentist properties; e.g., the strength of the evidence against the null (e.g., quantified by a conditional false alarm rate) for a fixed-α test grows weaker as N increases. NP themselves advocated decreasing α with N, but there are no general rules for this.

False alarm rates

Berger (2003) discusses the relationship between p-values, false alarm rates, and Bayesian posterior probabilities (or odds and Bayes factors).

In simple settings where one can easily bound false alarm rates, he shows the p-value significantly underestimates the false alarm rate among datasets sharing a given p-value.

This gives insight into why we've come to consider apparently small p-values—like "2σ" ($p \approx 0.05$) or "3σ" ($p \approx 0.003$)—to represent only weak evidence against the null. Typically, datasets with such p-values are not much more probable under alternatives than under the null.

He also shows that a "conditional frequentist" calculation of the false alarm rate in some settings amounts to computation of a Bayes factor.

Entries to the literature

- "402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies" (Thompson 2001) [web site]
- *The significance test controversy: a reader* (ed. Morrison & Henkel 1970, 2006) [Google Books]
- "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" (Berger 2003 with discussion; 2001 Fisher Lecture), *Statistical Science*, **18**, 1–32
- "Odds Are, It's Wrong: Science fails to face the shortcomings of statistics" (By Tom Siegfried 2010) [*Science News*, March 2010]

Example based on Berger (2003)

Model: $x_i = \mu + \epsilon_i$, $(i = 1 \text{ to } n)$ $\epsilon_i \sim N(0, \sigma^2)$

Null hypothesis, H_0: $\mu = \mu_0 = 0$

Test statistic:

$$t(x) = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$$

p-value:

$$\begin{aligned} p(t|H_0) &= \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \\ p\text{-value} &= P(t \geq t_{\rm obs}) \end{aligned}$$

t	p-value
1	0.317
2	0.046
3	0.003

$p = .05 \rightarrow$ "significant"

$p = .01 \rightarrow$ "highly significant"

Collect the p-values from a large number of tests in situations where the truth eventually became known, and determine how often H_0 is true at various p-value levels.

- Suppose that, overall, H_0 was true about half of the time.
- Focus on the subset with $t \approx 2$ (say, $[1.95, 2.05]$ so $p \in [.04, .05]$, so that H_0 was rejected at the 0.05 level.
- Find out how many times in that subset H_0 turned out to be true.
- Do the same for other significance levels.

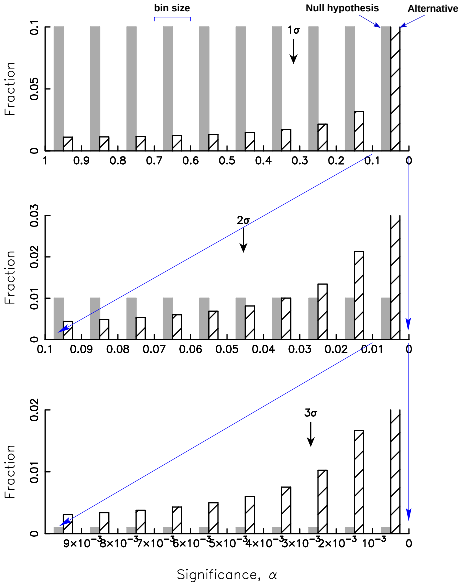
A Monte Carlo experiment

- Choose $\mu = 0$ OR $\mu \sim N(0, 4\sigma^2)$ with a fair coin flip*
- Simulate $n = 20$ data, $x_i \sim N(\mu, \sigma^2)$
- Calculate $t_{\rm obs} = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$ and $p(t_{\rm obs}) = P(t > t_{\rm obs}|\mu = 0)$
- Bin $p(t)$ separately for each hypothesis; repeat

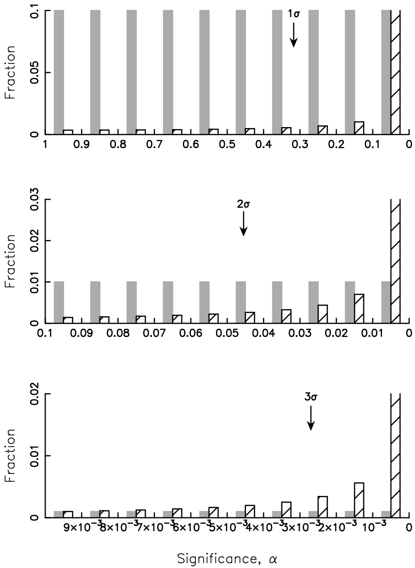
Compare how often the two hypotheses produce data with a 2– or 3–σ effect.

*A neutral assumption that gives alternatives a "fair" chance and may *over*estimate the evidence against H_0 in real settings

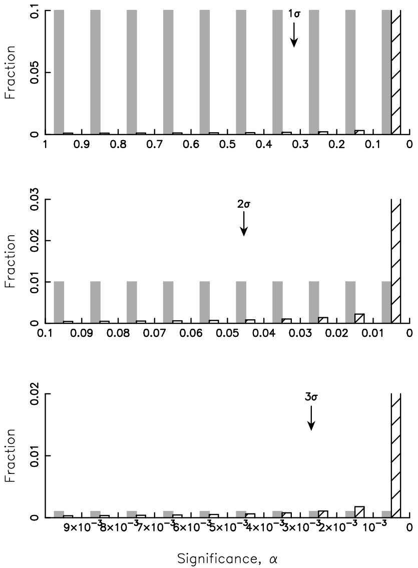
Significance Level Frequencies, $n = 20$



Significance Level Frequencies, $n = 200$



Significance Level Frequencies, $n = 2000$



What about another μ prior?

- For data sets with H_0 rejected at $p \approx 0.05$, H_0 will be true *at least* 23% of the time (and typically close to 50%). (Edwards et al. 1963; Berger and Selke 1987)
- At $p \approx 0.01$, H_0 will be true *at least* 7% of the time (and typically close to 15%).

What about a different "true" null frequency?

- If the null is initially true 90% of the time (as has been estimated in some disciplines), for data producing $p \approx 0.05$, the null is true at least 72% of the time, and typically over 90%.

In addition . . .

- At a fixed p, the proportion of the time H_0 is falsely rejected grows as $\sqrt{n}$. (Jeffreys 1939; Lindley 1957)
- Similar results hold generically; e.g., for χ^2. (Delampady & Berger 1990)

A p-value is not an easily interpretable measure of the weight of evidence against the null.

- It does not measure how often the null will be wrongly rejected among similar data sets
- A naive false alarm interpretation typically overestimates the evidence
- For fixed p-value, the weight of the evidence decreases with increasing sample size

Bayesian view of false-alarm rate

$$B \equiv \frac{p(\{x_i\}|H_1)}{p(\{x_i\}|H_0)} = \frac{p(p_{\rm obs}|H_1)}{p(p_{\rm obs}|H_0)}$$

$\rightarrow$ B here is just the ratio calculated in the Monte Carlo!

Why is p-value a poor measure of the weight of evidence?

- We should be *comparing hypotheses*, not trying to identify rare/surprising events—an observation surprising under the null motivates rejection only if it is not surprising under reasonable alternatives
- Comparison should use the *actual data*, not merely membership of the data in some larger set. A p-value conditions on incomplete information.

Harold Jeffreys, addressing an audience of statisticians:

> For n from about 10 to 500 the usual result is that $K = 1$ when $(a - \alpha_0)/s_\alpha$ is about 2. . . not far from the rough rule long known to astronomers, i.e. that differences up to twice the standard error usually disappear when more or better observations become available. . . I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the [observed] trial was improbable; that is, that it has not predicted something that has not happened. As an argument astronomer's experience is far better. (Jeffreys 1980)