Opinionated Lessons in Statistics

by Bill Press

#12 P-Value Tests

Surprise! Now we become frequentists for a while…

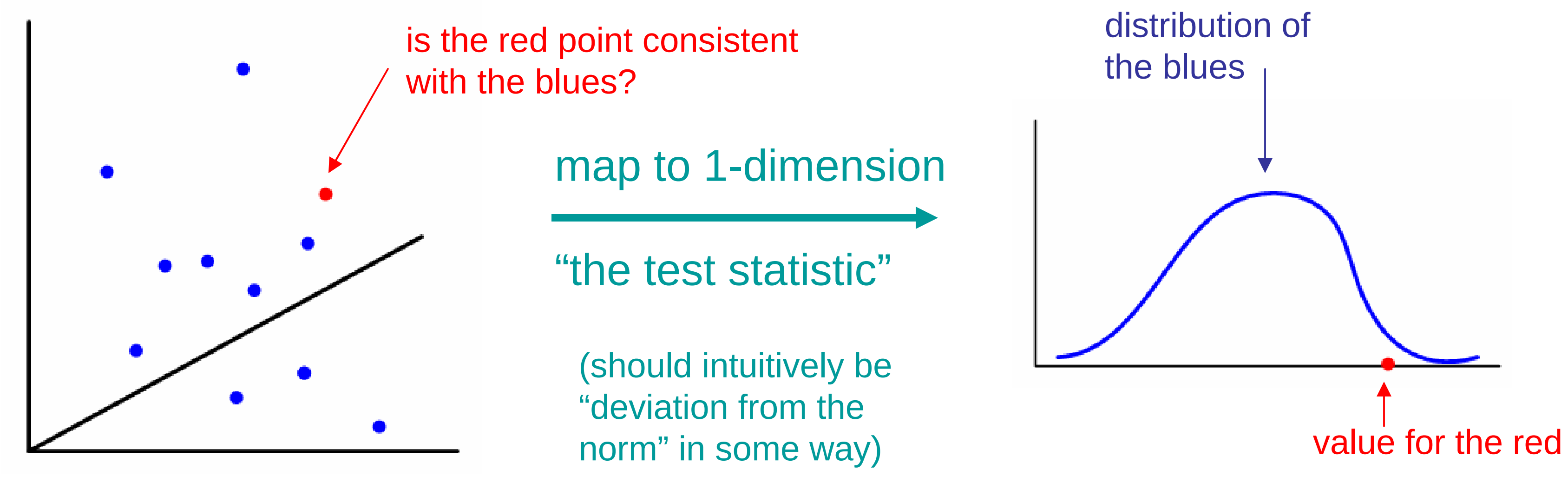
What if we want to know whether a single model or hypothesis H_0 is right or not, given some data D? Put better: Is D consistent with H_0 or inconsistent?

Bayesian: *Hmm. I know how to compute $P(D|H_0)$, because I know what the H_0 model is. But how the heck do I compute P(D| "not H0")? "Not H0" isn't a model. It's the absence of a model! I'm stuck.*

Frequentist (Ronald Fisher): *No problem! You imagine that your H_0 model produces a whole population of data sets. Then, you see whether your actual data is consistent with this imaginary population, using a very clever technique that I invented, called p-value.*

Bayesians have one D (the actual) and an EME set of known hypotheses. Frequentists have only one hypothesis (at a time), the "**null hypothesis H_0**", but they can imagine many data sets drawn randomly from it.

A data set can be a point in some high-dimensional space.
How do we decide if a point (the actual data) is “consistent” with a population of other points (the imagined data sets drawn from the null hypothesis)?



is the red point consistent with the blues?

map to 1-dimension

“the test statistic”

(should intuitively be “deviation from the norm” in some way)

distribution of the blues

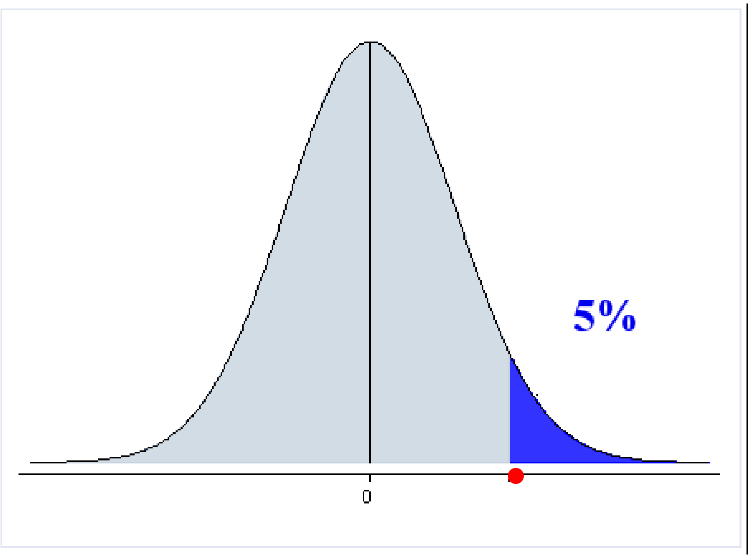
value for the red

If the red dot is very unlikely (if drawn from the blue distribution) then the actual data is inconsistent with the null hypothesis and disproves the null hypothesis.

But now we need to know what does unlikely mean?

Fisher (and others) decided that a good way to measure “unlikely” is by area (CDF) on the blue distribution.

The red dot’s p-value is the fraction of the time that a statistic drawn from the blue distribution (the null hypothesis) would be **as extreme or more extreme** than the red dot’s (actual data’s) observed value.

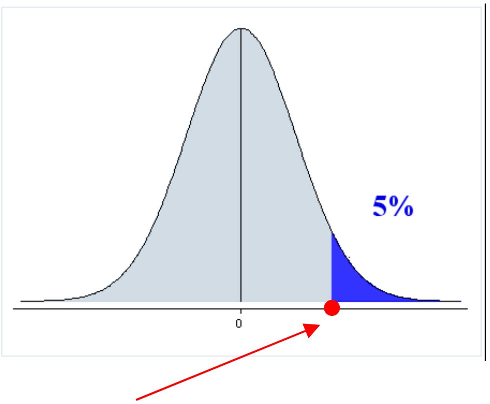


This definition has the nice property that a p-value of 1% will occur by chance (if the null hypothesis is actually true) only 1% of the time; p-value of 0.1%, 0.1% of the time; and so on.

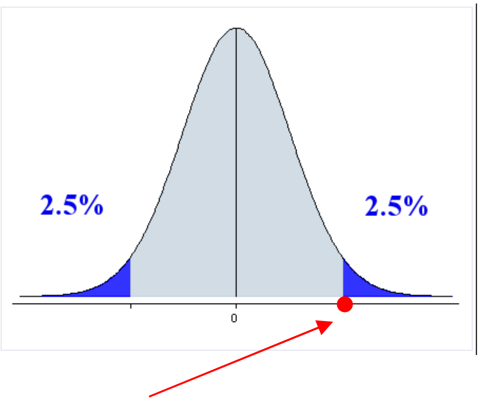
In other words, under the null hypothesis, the p-values are uniformly distributed in (0,1).

One remaining ambiguity: What does “as extreme or more extreme” mean?

one- versus two-tailed tests



p=0.05 if + is extreme, but –
has an innocuous explanation



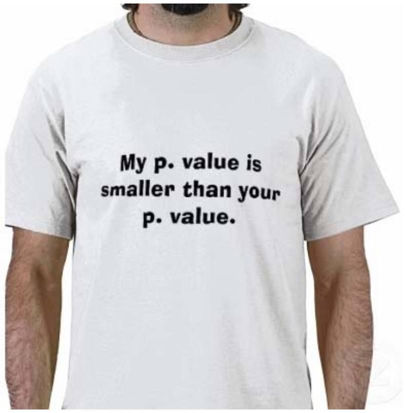
p=0.05 if + and – would be
considered equally extreme

There is an element of subjectivity in deciding whether a one- or two-tailed test is appropriate.

If the observed data is sufficiently extreme, the null hypothesis is disproved. (It can never be proved.)

Tactics: The idea is to pick a null hypothesis that is uninteresting, so that if you rule it out you have discovered something interesting.

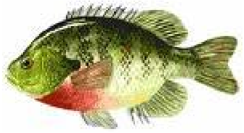
The classic p-value (or tail-) test terminology:



t-shirt for sale on the Web

- “null hypothesis”
- “the statistic” (e.g., t-value or χ^2)
  - calculable for the null hypothesis
  - intuitively should be “deviation from” in some way
- “the critical region” α
  - biologists use 0.05
  - physicists use 0.0026 (3 σ)
- one-sided or two?
  - somewhat subjective
  - use one-sided only when the other side has an understood and innocuous interpretation
- if the data is in the critical region, the null hypothesis is ruled out at the α significance level
- after seeing the data you
  - may adjust the significance level α
  - may not try a different statistic, because any statistic can rule out at the α level in $1/\alpha$ tries (“data dredging” for a significant result!)
- if you decided in advance to try N tests, then the critical region for α significance is α/N (Bonferroni correction).

There are some fishy aspects of tail tests, which we discuss later, but they have one big advantage over Bayesian methods: You don’t have to enumerate all the alternative hypotheses (“the unknown unknowns”).

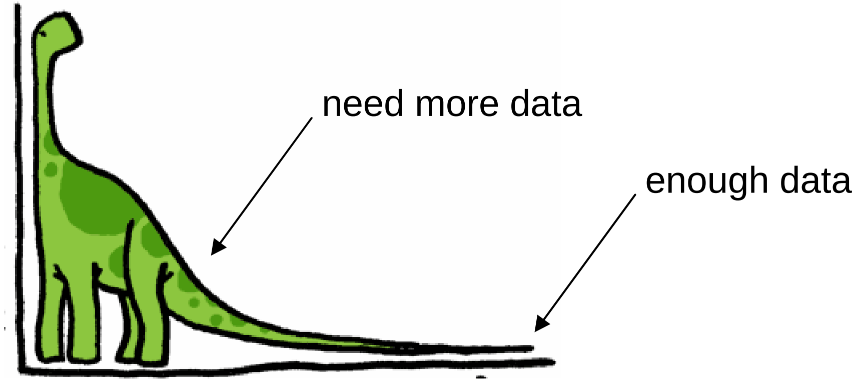


Tips on tail tests:

Don’t sweat a p-value like 0.06. If you really need to know, the only real test is to get significantly more data. Rejection of the null hypothesis is exponential in the amount of data.

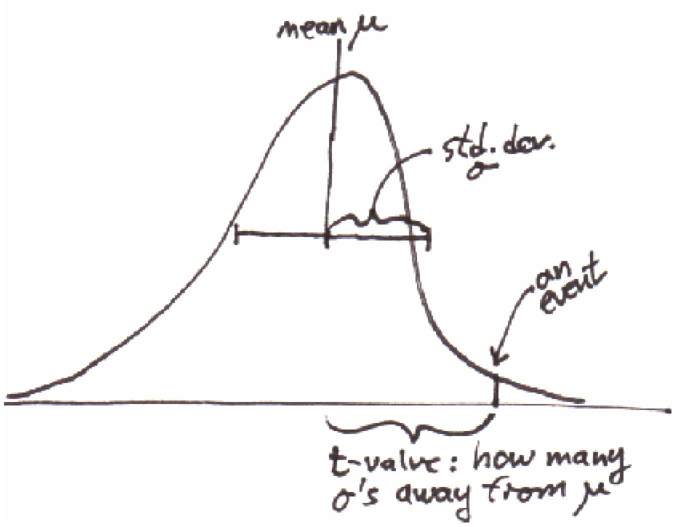
In principle, p-values from repeated tests s.b. exactly uniform in (0,1). In practice, this is often not quite true, because some “asymptotic” assumption will have crept in. All that really matters is that extreme tail values are being computed with moderate fractional accuracy. You can go crazy trying to track down not-exact-uniformity in p-values. (I have!)

Don’t **ever** interpret a p-value with language like “the null hypothesis has less than 5% probability of being correct”. That is just wrong! You **can** say, “a discrepancy this big would be seen by chance less than 5% of the time”. An equivalent statement is, “the model is ruled out with a 5% significance level”. (It’s confusing that numerically small significance levels are “highly significant”.)



Don’t confuse p-values with t-values (also sometimes named “Student”)

t-value = number of standard deviations from the mean



Intentionally drawn unsymmetric, not just sloppy drawing!

If the distribution of the test statistic is Normal (frequently the case because of the CLT), then t-values translate into p-values in a standard way. But don’t ever do this in the non-Normal case!

t-value (as “σ’s”)	one-tailed p-value	two-tailed p-value
1σ	≈16%	≈32%
2σ	≈2.3%	≈4.6%
3σ	≈0.13%	≈0.26%