The Nature of Evidence

Bret Larget
large t@stat.wisc.edu

Departments of Botany and of Statistics
University of Wisconsin—Madison

Botany 940—January 31, 2006

Definitions of Evidence

What is evidence?

According to *The Merriam Webster Dictionary*,

Evidence is PROOF or TESTIMONY; matter submitted in court to determine the truth of alleged facts,

A statistical definition according to Goodman and Royall (1988),

Evidence is a property of data that makes us alter our beliefs about how the world around us is working.

Definitions of Evidence

What is evidence?

According to *The Merriam Webster Dictionary*,

Evidence is PROOF or TESTIMONY; matter submitted in court to determine the truth of alleged facts,

A statistical definition according to Goodman and Royall (1988),

Evidence is a property of data that makes us alter our beliefs about how the world around us is working.

Definitions of Evidence

What is evidence?
According to *The Merriam Webster Dictionary*,

Evidence is PROOF or TESTIMONY; matter submitted in court to determine the truth of alleged facts,

A statistical definition according to Goodman and Royall (1988),

Evidence is a property of data that makes us alter our beliefs about how the world around us is working.

The Debate in Statistics over Evidence

- The debate over what statistical inference methods ought to be used in science extends back to the 1920s.
- The debate continues today.
- However, almost all introductory courses in statistics for scientists do not discuss the debate within the statistics community about the related philosophical issues.
- You may be surprised that the primary inferential procedures taught in most statistics courses is a combination of two schools of thought, and was disliked greatly by the founders of each school.

The Debate in Statistics over Evidence

- The debate over what statistical inference methods ought to be used in science extends back to the 1920s.
- The debate continues today.
- However, almost all introductory courses in statistics for scientists do not discuss the debate within the statistics community about the related philosophical issues.
- You may be surprised that the primary inferential procedures taught in most statistics courses is a combination of two schools of thought, and was disliked greatly by the founders of each school.

The Debate in Statistics over Evidence

- The debate over what statistical inference methods ought to be used in science extends back to the 1920s.
- The debate continues today.
- However, almost all introductory courses in statistics for scientists do not discuss the debate within the statistics community about the related philosophical issues.
- You may be surprised that the primary inferential procedures taught in most statistics courses is a combination of two schools of thought, and was disliked greatly by the founders of each school.

Testing

Now you have the chance to tell me how to carry out a statistical test. . . .

Schools of Thought

I will describe four separate schools of thought on how to do statistical inference.

1. Fisher and Significance tests;
2. Neyman and Pearson and Hypothesis tests;
3. Likelihood Ratios;
4. Bayesian Inference.

The Fisher School

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. — R. A. FISHER *(1937).*

- According to Fisher, the necessary elements of a significance test were
  - a null hypothesis;
  - and a test statistic with a null distribution.
- This results in a p-value.
- The p-value is interpreted as the probability of obtaining data at least as extreme as the observed data assuming that the null hypothesis is true.
- Data for which the p-value is less than an arbitrary threshold such as 0.05 is called significant.

The Neyman-Pearson School

- According to Neyman and Pearson, the necessary elements of a hypothesis test were
  - null and alternative hypotheses;
  - a test statistic with a null distribution;
  - and a rejection region.
- The decision to accept or reject the null hypothesis in favor of the alternative hypothesis is based on whether or not the test statistic falls into the rejection region.

A Revealing Quote

. . . no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.

But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong.
— J. NEYMAN AND E. PEARSON *(1933)*

Discussion of Quote

- Neyman and Pearson regarded hypothesis testing as a process which guaranteed a long-run error rate of rejecting false null hypotheses.
- The price of this objective method to make decisions is that we abandon our ability to measure evidence or judge truth in individual experiments.
- As the inability to make judgments in individual experiments is clearly undesirable, standard practice evolved to “fix” the Neyman-Pearson procedure.
- Fisher’s p-value was added as a measure of the strength of evidence against the null hypothesis.

The Combined Approach

- The combined approach became standard practice despite the vehement arguments against each other’s methods from the founders of each school.
- The combined method is lauded by many as being scientific because of its objectivity.
- However, the combined method is an automatic procedure for drawing inferences that does not allow for the inclusion of judgment or knowledge of the underlying scientific processes.
- There is no mechanism to include any prior evidence.

P-values as Measures of Evidence

Over the decades, many authors have criticized p-values and hypothesis testing procedures for science.

- P-values depend on how the data is collected.
- P-values measure the probability of data that is unobserved.
- P-values do not measure the size of the effect.
- P-values can lead to rejection of a hypothesis without an alternative hypothesis that better explains the data.

Criticism 1

P-values depend on how the data is collected.

- Consider an experiment comparing treatments A and B.
- The first five cases do better with A and the 6th does better with B.
- If the plan was to test six cases,

$$\text{p-value} = \mathbb{P}\,(\text{5 or more successes}) = 6\left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = 0.11.$$

- If the plan was to test cases until B was better,

$$\text{p-value} = \mathbb{P}\,(\text{stop at 6th or later trial}) = \sum_{k=6}^{\infty}\left(\frac{1}{2}\right)^k = 0.03.$$

- Identical data can lead to different p-values (and conclusions) depending on which unobserved realizations are included in the probability calculation.

Criticism 2

P-values measure the probability of data that is unobserved.

> *A hypothesis that may be true may be rejected because it has not predicted . . . results which have not occured.*
> *— H. Jeffreys (1961)*

Criticism 3

P-values do not measure the size of the effect.

- Consider a paired test with known variance 1.
- $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 > \mu_2$.
- Experiment 1:
  $n = 25$, $\bar{x}_1 - \bar{x}_2 = 0.5$, $z = 0.5/(1/\sqrt{25}) = 2.50$,
  $p = 0.0062$.
- Experiment 2:
  $n = 2500$, $\bar{x}_1 - \bar{x}_2 = 0.05$, $z = 0.05/(1/\sqrt{2500}) = 2.50$,
  $p = 0.0062$.
- Large experiments give the same quantitative measure of "evidence" to small, possibly scientifically unimportant results that smaller experiments might give to larger estimated effects that could be of true scientific importance.

Criticism 4

P-values can lead to rejection of a hypothesis without an alternative hypothesis that better explains the data.

> *Nor do you find experimentalists typically engaged in disproving things. They are looking for appropriate evidence for affirmative conclusions. Even if the mediate purpose is the disestablishment of some current idea, the immediate objective of a working scientist is likely to be to gain affirmative evidence in favor of something that will refute the allegation that is under attack.*
>
> — J. BERKSON *(1942)*

Likelihood Ratios

- Fisher developed the concept of likelihood in the 1920s along with the principle of maximum likelihood to estimate unknown parameters.
- The likelihood has the same equation as the formula for the probability of the data, except that what is considered known and what is considered unknown are reversed.
- In our previous example, consider a sequence of six independent trials with success probability p.
- The probability of x successes is

$$f(x \mid p) = \binom{6}{5} p^x (1-p)^{6-x} \quad \text{for } x = 0, \ldots, 6$$

- Fixing p, this function is a probability mass function with argument x and sums to one.
- Fixing x, this is a continuous function of p and is called a likelihood function.

The Likelihood Principle

> *Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data, and the likelihood ratio is to be interpreted as the degree to which the data support one hypothesis against the other.*
>
> — A. W. F. EDWARDS

Likelihood Ratios

- In this school of thought, likelihoods are never interpreted except in comparison to other likelihoods.
- Typically, the logarithm of the likelihood ratio is examined so that the scale of evidence in each direction is the same.
- If the probability of the data is twice as large for one hypothesis as another, then the likelihood ratio is either 2 or 0.5 depending on which is placed in the numerator.
- The numbers 2 and 0.5 are not the same distance from 1, but the evidence in favor of one hypothesis or the other is the same.
- On the log scale, $\log 2 = -\log 0.5$ and distance from 0 is a symmetric measure of relative evidence.

Criticisms of Likelihood Ratios

- Comparisons of simple hypotheses (one parameter problems where the two competing hypotheses assign the parameter different values) are straightforward theoretically, but are too simple for many practical problems.
- Likelihood models for hypotheses of interest often involve composite hypotheses with likelihood models that include both structural parameters of interest as well as possibly many nuisance parameters.
- Inference from likelihood ratios poses both practical numerical problems and potentially theoretical problems when the parameter space is complicated.
- What does a likelihood ratio mean in practice as a measure of the strength of evidence in favor or against a hypothesis?

Bayesian Inference

- Bayesian inference is distinguished from the other methods we have discussed in that probability is used to directly measure the belief an observer has in hypotheses.
- Bayesian inference also uses probability to calculate the chances of possible outcomes of data given certain hypotheses as in likelihood ratios.
- The Bayesian paradigm is that the beliefs of an observer are updated with data.
- This paradigm has a real appeal, but is controversial as it depends on a subjective assessment of prior belief.
- Different observers with different prior beliefs can reach different conclusions based on the same data.

Bayesian Inference and Likelihood Ratios

Board doodles for now. . . .

Controversy over Bayesian Inference

- Bayesian inference is controversial primarily because of prior distributions.

One point of view:

> *I do not trust Bayesian inference because Bayes' Theorem states that posterior probabilities of hypotheses depend on a subjective prior distribution.*

An alternative point of view:

> *Probability is the only reasonable way to specify belief in a hypothesis. A consequence of Bayes' Theorem says that if you do not specify a prior distribution on hypotheses, then you cannot directly measure the overall strength of evidence for a hypothesis.*

Controversy over Bayesian Inference

- Bayesian inference is controversial primarily because of prior distributions.

One point of view:

> *I do not trust Bayesian inference because Bayes' Theorem states that posterior probabilities of hypotheses depend on a subjective prior distribution.*

An alternative point of view:

> *Probability is the only reasonable way to specify belief in a hypothesis. A consequence of Bayes' Theorem says that if you do not specify a prior distribution on hypotheses, then you cannot directly measure the overall strength of evidence for a hypothesis.*

Bibliography

BERKSON, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Society* **37**: 325–335.

GOODMAN, S.N. AND R. ROYALL (1988). Evidence and Scientific Research. *American Journal of Public Health* **78**: 1568–1574.

GOODMAN, S.N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine* **130**: 995–1004.

GOODMAN, S.N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* **130**: 1005–1013.