PHIL 6014: Spring 2023
Special Topics in Philosophy
Philosophy of Inductive-Statistical Inference
Wed. 4-6:30 McBryde 223
Prof. Deborah Mayo

This course is an introduction to the philosophy of inductive-statistical inference in relation to general problems of philosophy of science (e.g., falsification vs confirmation, underdetermination, science vs pseudoscience) and to current controversies regarding uncertain inference in scientific practice (e.g., statistical significance tests, Bayesian vs frequentist methods, replication crisis, and science and values in evidence policy). We will study examples of statistical evidence in the law, psychology, medicine and physics. You do not need to have a statistical or a philosophical background, only an interest in learning about philosophy of statistics (PhilStat) in its relations to problems of philosophy of science and statistical epistemology.

- Courses in research methods in the *social sciences* allow for an impressive array of statistical methods and models,
- but using them successfully requires reacting to challenges regarding their legitimate use and interpretation.
- —often subject of philosophical controversies (although it is typically not recognized)

The reverse problem often arises in courses in philosophy of science:

- Without statistical understanding, tackling problems about uncertain evidence are often out of touch with tools actually used

- Practitioners consulting texts in philosophy of science are typically at a loss to see how they are relevant to their problems.

Why is it important to address these issues to those without (as well as with) a statistical background?

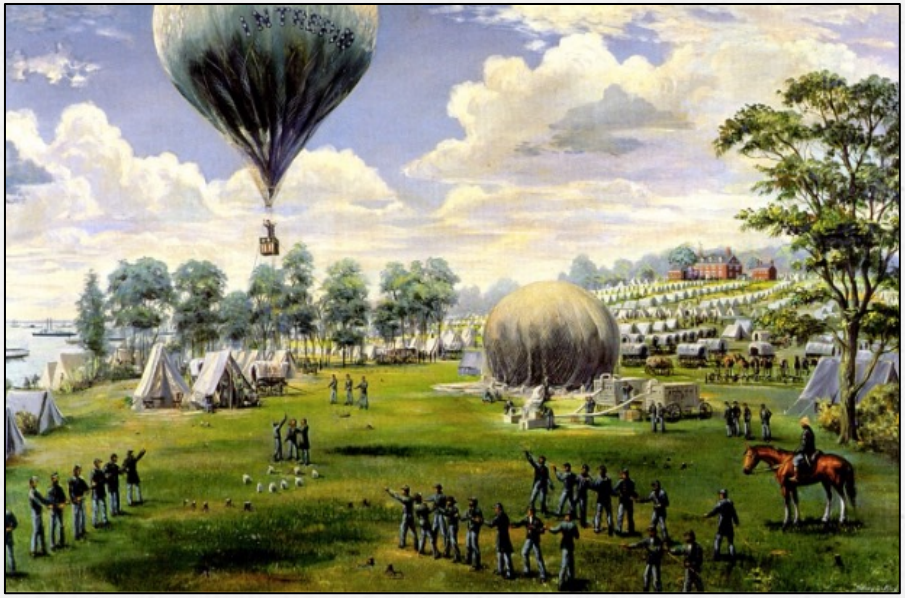
- High powered methods make the computations invisible to most users
- Methodological advocacy is directed at being nontechnical or requiring very minimal understanding of technical complexities
- We need to be able to critically analyze the debates about methods

- By the time you finish this course, you will be able to comprehend and critically evaluate the debates, disagreements, and controversies now taking place

- You will be beyond the typical audience to which the popular, “non-technical” arguments are directed.

- You will also understand the historical, statistical and personality backdrop to the issues
- I'm keen to write a new edition and/or companion to the book—with your help
- I am posting outlines of the chapters ("tours") to help you cover the material

Start from the preface:
The Statistics Wars



Bayesian-Frequentist Wars

The Goal of Frequentist Inference: Construct procedures with frequentist guarantees

good long-run performance

The Goal of Bayesian Inference: Quantify and manipulate your degrees of beliefs

belief probabilism

Larry Wasserman (p. 24)

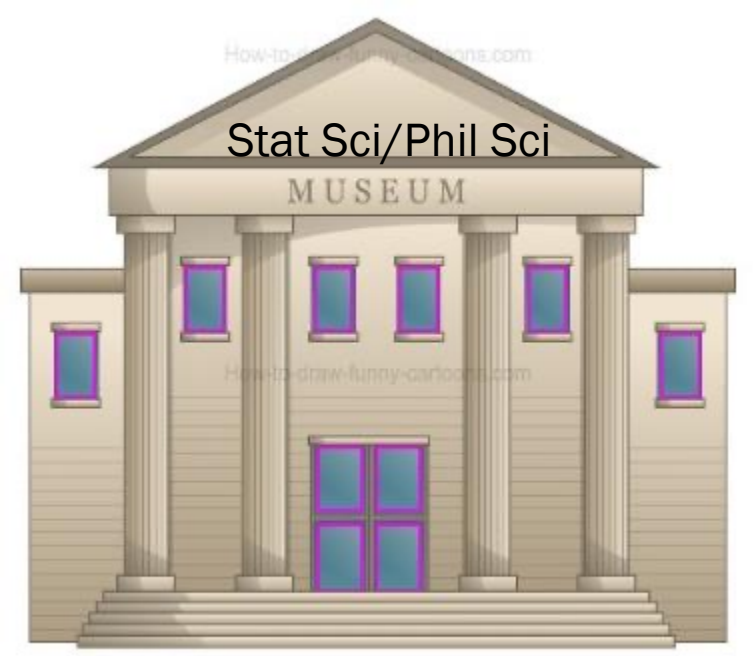
But now we have marriages and reconciliations (pp. 25-8)

- End of foundations? (Unifications and Eclecticism—”we use whatever works”)
- Long-standing battles still simmer below the surface (agreement on numbers)
- We wouldn’t have American Statistical Association Task Forces, as in 2022, were it not that concepts are more confused than ever

- I will not be proselytizing for a given school; they all have shortcomings …

- The goal is to unlock the mysteries that are leaving many skeptical statistical consumers in the dark about a crucial portion of science (12)

Let’s brush the dust off the pivotal debates, walk into the museums to hear the founders: Fisher, Neyman, Pearson, Savage and many others in relation to today’s statistical crisis in science (xi)



A metastatistical tool: Statistical inference as severe testing

- Main source of the statistical crisis in science?
- We set sail with a simple tool: you don't have evidence for a claim if little or nothing has been done to rule out how it can be false
- You needn't accept this principle to use it to excavate the statistics wars

A claim is warranted to the extent it passes severely

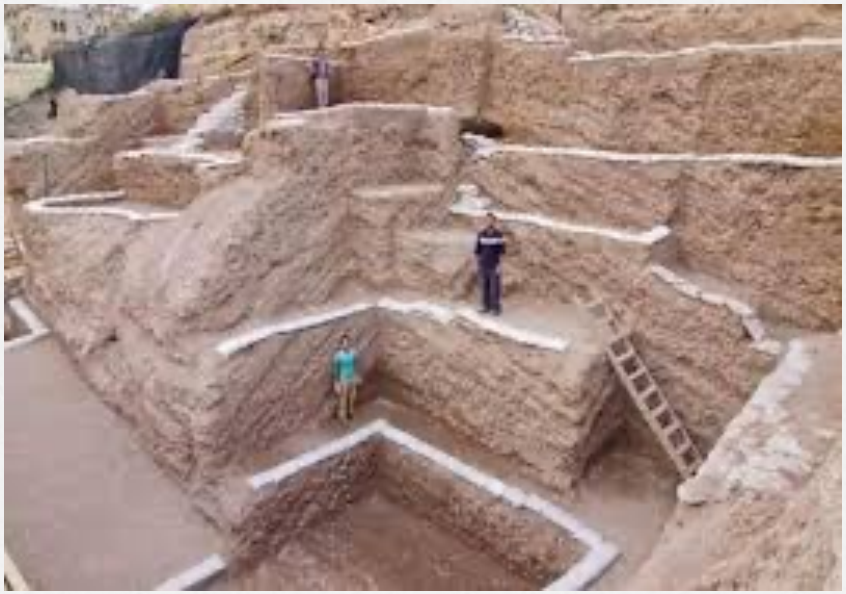
- We have evidence for a claim only to the extent that it has been subjected to and passes a test that would probably have found it flawed or specifiably false, just if it is
- This probability is the stringency or severity with which it has passed the test

A philosophical excursion

"Taking the severity principle, along with the aim that we desire to find things out… let's set sail on a philosophical excursion to illuminate statistical inference." (8)

"…and engage with a host of tribes marked by family quarrels, peace treaties, and shifting alliances" (xiv)

Revisit some taboos: problems of induction & falsification, science vs. pseudoscience



Excursion 1 How to Tell What's True About Statistical inference

Tour I: Beyond Probabilism and Performance

Most findings are false?

"Several methodologists have pointed out that the high rate of nonreplication of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. …


OMG!

Simple significance tests (Fisher)

P-value. …to test the conformity of the data under analysis with H_0 in some respect:

…we find a function $d(\boldsymbol{x})$ of the data, the **test statistic**, such that

- the larger the value of $d(\boldsymbol{x})$ the more inconsistent are the data with H_0;

- Small P-value indicates *some* underlying discrepancy from H_0 because **very probably you would have seen a less impressive** difference *d* than observed d_{0bs} were H_0 true.

- Usually require .05, .025, .01
- Tool to avoid being fooled by randomness
- Still not evidence of a substantive scientific hypothesis H*

Neyman-Pearson (N-P) tests:

A null* and alternative hypotheses H_0, H_1 that are exhaustive*

H_0: “no effect” vs. H_1: “some positive effect”

Type 1 error (mistakenly rejecting) and
Type 2 error (mistakenly failing to reject)

*test hypothesis

Despite personality conflicts & jealousies

- They both fall under tools for "appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, 1033)–*error probabilities*
- Can place all under the rubric of ***error statistics***
- Confidence intervals, N-P and Fisherian tests, resampling, randomization

Both Fisher and N-P methods: it's easy to lie with statistics by selective reporting

- Sufficient finagling—cherry-picking, significance seeking, multiple testing, post-data subgroups, trying and trying again—may practically guarantee a preferred claim *H* gets support, even if it's unwarranted by evidence

Severity Requirement (weak):

- Such a test fails a *minimal requirement* for a stringent or severe test
- N-P and Fisher did not put it in these terms but our severe tester does

This concern alters the role of probability (typically just two):

Probabilism. To assign a degree of probability, confirmation, support or belief in a hypothesis, given data $\boldsymbol{x}_0$ (absolute or comparative)

(e.g., Bayesian, likelihoodist, Fisher (at times))

Performance (more apt than frequentist*). Ensure long-run reliability of methods, coverage probabilities (frequentist, behavioristic Neyman-Pearson, Fisher (at times))

- There are roles for both, but neither “probabilism” nor “performance” directly captures the idea of error probing capacity
- high degree of belief (in the sense of personalism) can’t suffice: even where a claim is known to be true, it can be poorly tested

- Good long-run performance is a necessary, not a sufficient, condition for severity

    — example: 2 weighing machines

- What bothers you with selective reporting, cherry picking, stopping when the data look good, P-hacking
- Not problems about long-runs—

We cannot say the case at hand has done a good job of avoiding the sources of misinterpreting data

A claim *C* is not warranted _____

- ***Probabilism****:* unless *C* is true or probable (gets a probability boost, made comparatively firmer)

- ***Performance***: unless it stems from a method with low long-run error

- ***Probativism (severe testing)*** unless something (a fair amount) has been done to probe ways we can be wrong about *C*

Popper vs logics of induction/ confirmation

Severity was Popper's term, and (though he never cashed it out adequately), the debate between Popperian falsificationism and inductive logics of confirmation/ support parallel those in statistics.

Excursion 1 Tour II: Error Probing Tools vs. Logics of Evidence (p. 30)

To understand the stat wars, start with the holy grail–a purely formal (syntactical) logic of evidence

It should be like deductive logic but with probabilities

Engine behind probabilisms (e.g., Carnapian confirmation theories, Likelihood accounts, Bayesian posteriors)

Comparative Logic of Support

- **Ian Hacking (1965)** "Law of Likelihood": ***x*** support hypothesis H_0 less well than H_1 if,

  $\Pr(\boldsymbol{x};H_0) < \Pr(\boldsymbol{x};H_1)$

  (rejects in 1980)

The data support H_0 less well than they support H_1 if ***x*** *is is less* probable under H_0 *than under* H_1

- The "**likelihood"** of H_0 is less than H_1

Likelihood Principle (LP)

In probabilisms, the import of the data is via the ratios of *likelihoods* of hypotheses

$$\Pr(\boldsymbol{x}_0;H_0)/\Pr(\boldsymbol{x}_0;H_1)$$

The data $\boldsymbol{x}_0$ are fixed, while the hypotheses vary

A pivotal disagreement in the philosophy of statistics battles

Why N-P Introduced error-probabilities

"there *always* is such a rival hypothesis [H_1] *viz.*, that things just had to turn out the way they actually did" (Barnard 1972, 129).

- Pr(H_0 is less well supported than H_1; H_0) is high for some H_1 or other

- Pr(Test would yield such good support for some H_1; even though H_0) = high

All error probabilities violate the LP

(even without selection effects):

Sampling distributions, significance levels, power, all depend on something more [than the likelihood function]–something that is irrelevant in Bayesian inference–namely the sample space

(Lindley 1971, 436)

Souvenir B Likelihood vs error statistics

The LP implies…the irrelevance of predesignation, of whether a hypothesis was thought of before hand or was introduced to explain known effects (Rosenkrantz 1977, 122)

Likelihood Principle (LP)

If the statistical model is correct, then all the information from the data (for inference about a parameter in that model) comes through the likelihood ratio.

Held by Bayesians and Likelihoodists

(qualifications to arise)

Optional Stopping:

Error probing capacities are altered not just by data dredging, but also via data dependent stopping rules:

Testing claims about the mean μ *of a normal distribution*

H_0: no effect vs. H_1: some effect

2-sided H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$.

Instead of fixing the sample size *n* in advance, in some tests, *n* is determined by a *stopping rule*:

- Keep sampling until H_0 is rejected at ("nominal") 0.05 level

Keep sampling until sample mean M differs from 0 from some amount (2SE)

- *Trying and trying again*: Having failed to rack up a statistically significant difference after 10 trials, go to 20, 30 and so on until obtaining a 2 SE difference

Table 1.1 The effect of repeated significance tests (the "try and try again" method)

Number of trials n	Probability of rejecting H_0 with a result nominally significant at the 0.05 level at or before n trials, given H_0 is true
1	0.05
2	0.083
10	0.193
20	0.238
30	0.280
40	0.303
50	0.320
60	0.334
80	0.357
100	0.375
200	0.425
500	0.487
750	0.512
1000	0.531
Infinity	1.000

In testing the mean of a standard normal distribution

Why do stopping rules drop out?

Go back to simple likelihoodist

Bernoulli trials p. 33

$\boldsymbol{x}$ = <S,S,F,S>

H_0 : Pr(S) = .2 (so Pr(F) = .8)

H_1 : Pr(S) = .8 (so Pr(F) = .2)

Pr(S) under the two hypotheses need not sum to 1, e.g., H_1 could have assigned Pr(S) = .3 or anything other than .2

LIK(H_0) = Pr(X=1)Pr(X=1)Pr(X=0)Pr(X=1)

Bernoulli trials p. 33

x = <S,S,F,S>

H_0 : Pr(S) = .2 (so Pr(F) = .8)

H_1 : Pr(S) = .8 (so Pr(F) = .2)

LIK(H_0) = Pr(X=1)Pr(X=1)Pr(X=0)Pr(X=1)

We should write "; H_0 " for each, e.g., Pr(X=1; H_0)

(.2)(.2)(.8)(.2) = .0064

What's the LIK(H_1)?

$\boldsymbol{x}$ = <S,S,F,S>

H_0 : Pr(S) = .2 (so Pr(F) = .8)

H_1 : Pr(S) = .8 (so Pr(F) = .2)

What's the LIK(H_1)=

Pr(X=1)Pr(X=1)Pr(X=0)Pr(X=1)

We should write "; H_1 " for each e.g Pr(X=1 ;H_0)

(.8)(.8)(.2)(.8) = .1024

Bernouilli trials p. 34

$\boldsymbol{x}$ = <S,S,F,S>

Likelihood Ratio

LIK(H_0) = .0064

LIK(H_1) = .1024

LIK(H_1)/ LIK(H_0) = 16

Binomial distribution

x = <S,S,F,S>

To have a Binomial probability distribution need to consider different ways x could occur

ways of getting 3 S's out of 4 trials = 4C3 = 4

S,S,F,S; S,S,S,F; S,F,S,S, F,S,S,S

nCk $p^k(1 - p)^{n-k}$

is another way of writing $\binom{n}{k}p^k(1 - p)^{n-k}$

Binomial distribution

$\boldsymbol{x}$ = <S,S,F,S>

So

LIK(H_0) = Pr(x; H_0) = (4C3).0064

LIK(H_1) = Pr(x; H_1) = (4C3).1024

But the Likelihood Ratio in favor of H_1 is still .1024/.0064 = 16

Negative Binomial trials

$\boldsymbol{x}$ = <S,S,F,S>

The result could have occurred in another way. Perhaps, instead of fixing 4 trials, we sample until the 3rd success

How many different ways could this have happened (resulting in **x**)?

Negative Binomial trials

x = <S,S,F,S>

The first 3 trials must have 2 successes in some order: SSF, SFS, FSS

3C2 ways

This differs from the Binomial coefficient

Negative Binomial trials

x = <S,S,F,S>

kth success on nth trial n-1Ck-1

3C2 in our case

Likelihood ratio is still the same
LIK(H_1) / LIK(H_0)

LIK(H_0) = Pr(x; H_0) = (3C2).0064

LIK(H_1) = Pr(x; H_1) = (3C3).1024

So, should it matter which way x resulted?

Not for accounts that hold the LP

x = <S,S,F,S> p. 303

Box gives the example of stopping rules. Stopping rules don't alter the posterior distribution, as we learned from the extreme example in Excursion 1 (Section 1.5). For a simple example, he considers four Bernoulli trials: ⟨S, S, F, S⟩. The same string could have come about if $n = 4$ was fixed in advance, or if the plan was to sample until the third success is observed. The latter are called negative Binomial trials, the former Binomial. The string enters the likelihood ratio the same way, $\binom{4}{3}\theta^3(1-\theta)$ and $\binom{3}{2}\theta^3(1-\theta)$ respectively: the only difference is the coefficients, which cancel. But the significance tester distinguishes them, because the sample space, and corresponding error probabilities, differ.[1] When it comes to model testing, Box contends, this LP violation is altogether reasonable, since "we are considering whether, given A, the sample is likely to have occurred at all" (ibid., p. 75).

Negative Binomial trials

$\boldsymbol{x}$ = <S,S,F,S>

kth success on nth trial n-1Ck-1

3C2 in our case

Likelihood ratio is the same LIK(H_1) / LIK(H_0) = 16

LIK(H_0) = Pr(x; H_0) = (3C2).0064

LIK(H_1) = Pr(x; H_1) = (3C3).1024

Negative Binomial and Binomial trials

$\boldsymbol{x}$ = <S,S,F,S>

The likelihoods are said to be the “same” when they are proportional for all hypotheses

The coefficient drops out (in the ratio)

(even though they have different sufficient statistics)

This is why stopping rules drop out

Keep sampling until sufficiently many more S's than F's

Keep sampling until the P-value is .05 (in 2-sided testing)

Exhibit (ii): How Stopping Rules Drop Out. Our question remains: by what magic can such considerations disappear? Formally, the answer is straightforward. Consider two versions of the above experiment: In the first, 1.96 is reached via fixed sample size ($n = 169$); in the second, by means of optional stopping that ended at 169. While $d(\boldsymbol{x}) = d(\boldsymbol{y})$, because of the stopping rule, the likelihood of $\boldsymbol{y}$ differs from that of $\boldsymbol{x}$ by a constant k, that is,

$$\Pr(\boldsymbol{x}|H_i) = k\Pr(\boldsymbol{y}|H_i) \text{ for constant } k.$$

Given that likelihoods enter as ratios, such proportional likelihoods are often said to be the "same." Now suppose inference is by Bayes' Theorem. Since likelihoods enter as ratios, the constant k drops out. This is easily shown. I follow E, L, & S; p. 237.

For simplicity, suppose the possible hypotheses are exhausted by two, H_0 and H_1, neither with probability of 0.

p. 45 x and y refer to two ways a sample could come about

To show $\Pr(H_0|\boldsymbol{y}) = \Pr(H_0|\boldsymbol{x})$:

(1) We are given the proportionality of likelihoods, for an arbitrary value of k:

$$\Pr(\boldsymbol{y}|H_0) = k\Pr(\boldsymbol{x}|H_0),$$

$$\Pr(\boldsymbol{y}|H_1) = k\Pr(\boldsymbol{x}|H_1).$$

(2) By definition:

$$\Pr(H_0|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{y}|H_0)\Pr(H_0)}{\Pr(\boldsymbol{y})}.$$

The denominator $\Pr(\boldsymbol{y}) = \Pr(\boldsymbol{y}|H_0)\ \Pr(H_0) + \Pr(\boldsymbol{y}|H_1)\ \Pr(H_1)$.

Now substitute for each term in (2) the proportionality claims in (1). That is, replace $\Pr(\boldsymbol{y}|H_0)$ with $k\Pr(\boldsymbol{x}|H_0)$ and $\Pr(\boldsymbol{y}|H_1)$ with $k\Pr(\boldsymbol{x}|H_1)$.

(3) The result is

$$\Pr(H_0|\boldsymbol{y}) = \frac{k\Pr(\boldsymbol{x}|H_0)\ \Pr(H_0)}{k\Pr(\boldsymbol{x})} = \Pr(H_0|\boldsymbol{x}).$$

The posterior probabilities are the same whether the 1.96 result emerged from optional stopping, $\boldsymbol{Y}$, or fixed sample size, $\boldsymbol{X}$.

Optional Stopping

- “if an experimenter uses this [optional stopping] procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true”
  (Edwards, Lindman, and Savage 1963, 239)

- Understandably, they observe, the significance tester frowns on this, or at least requires adjustment of the P-values

“Imagine instead if an account advertised itself as ignoring stopping rules” (43)

- “[the] irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson).” (Edwards, Lindman, and Savage 1963, 239)
- “…these same authors who warn that to ignore stopping rules is to guarantee rejecting the null hypothesis even if it’s true” (43) declare it irrelevant.

What counts as bias (or even cheating) depends on statistical philosophy

- Are they contradicting themselves?
- “No. It is just that what looks to be, and indeed is, cheating from the significance testing perspective is not cheating from [*their*] Bayesian perspective.” (43)

Table 1.1 The effect of repeated significance tests (the "try and try again" method)

Number of trials n	Probability of rejecting H_0 with a result nominally significant at the 0.05 level at or before n trials, given H_0 is true
1	0.05
2	0.083
10	0.193
20	0.238
30	0.280
40	0.303
50	0.320
60	0.334
80	0.357
100	0.375
200	0.425
500	0.487
750	0.512
1000	0.531
Infinity	1.000

In testing the mean of a standard normal distribution

Nominal vs. *Actual* significance levels :

- With *n* fixed the Type 1 error probability is 0.05
- With this stopping rule the actual significance level differs from, and will be greater than 0.05

  (proper stopping rule)

At odds with reforms to block irreplication: 21 Word Solution

- Replication researchers (re)discovered that data-dependent hypotheses and stopping are a major source of spurious significance levels.
- Statistical critics, Simmons, Nelson, and Simonsohn (2011) place at the top of their list the need to block flexible stopping

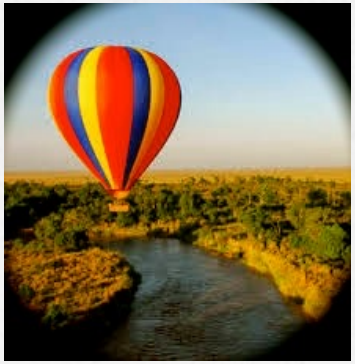
"Authors must decide the rule for terminating data collection before data collection begins and report this rule in the articles" (Simmons, Nelson, and Simonsohn 2011, 1362).

Competing Intuitions

- You've scarcely bent over backwards to block being fooled by chance by trying and trying again (at least using this test) and failing to report this
- On the other hand, why should intentions to stop alter the import of the evidence? (what if she always intended to go to 100 trials, say)
- Inference by Bayes Theorem says it should not: (as we derived). (stopping rule principle)

Competing Intuitions (Berger notes)

Intuition here has difficulty; as Savage (1961) said "When I first heard the stopping rule principle from Barnard in the early 50's, I thought it was scandalous that anyone in the profession could espouse a principle so obviously wrong, even as today I find it scandalous that anyone could deny a principle so obviously right."



Current State of Play in Bayesian-Frequentist Wars

1.3 View from a Hot-Air Balloon (p. 23)

How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantively different answers to the same problems? Is complacency in the face of contradiction acceptable for a central discipline of science? (Donald Fraser 2011, p. 329)

Decoupling

- Break off stat methods from their traditional philosophies
- Can Bayesian methods find a new foundation in error statistical ideas? (p. 27)
- Excursion 6: (probabilist) foundations lost; (probative) foundations found (432)