Multiplicity and Unification in Frequentist (Error) Statistics: Learning from D. R. Cox

Deborah G. Mayo

- Long standing controversies in the foundations of frequentist statistics stem from failure to appreciate the <u>multiplicity</u> of frequentist methods, and erroneous conceptions of the <u>unifying</u> core that links together and directs their interpretation and justification
- Learning from the work of statistician D.R. Cox serves to explicate the <u>multiplicity</u> of frequentist methods and goals, and identify the <u>unification</u> of principles that direct their valid use in realistic scientific inquiries.
- Moving away from the oversimplified caricatures of the standard methods (e.g., significance tests, confidence intervals) on which critics focus, the tools emerge as satisfying piecemeal learning goals within a multiplicity of methods, models, and experimental designs, and interconnected checks, and reports, of error.
- Denying the assumption that the overall unified principle for these methods must be merely controlling the low-long run error probabilities of methods (behavioristic rationale), we advance an epistemological principle that renders hypothetical error probabilities relevant for learning about particular phenomena.
- We contrast it with current Bayesian attempts at unifications of frequentist and Bayesian methods.

  Note: Cox also distinguishes uses of statistical methods for inference and for decision, but even focusing on inference is to identify a huge territory marked by conflicts of personalities and philosophies…

Multiple Roles of Probability in Induction

Inductive inference: the premises (background and evidence statements) can be true while the conclusion inferred may be false without a logical contradiction:
the conclusion is "evidence transcending".

probability naturally arises in capturing induction, but there is more than one way this can occur:
To measure:

1. how much confidence or belief in a hypothesis
2. how reliable or how well tested is a hypothesis

- too often presupposed it can only be the former—leading to well-known difficulties with using frequentist ideas (universes are not as plenty as blackberries—Peirce)
- our interest is in how frequentist probability may be used for the latter…

For instance, significance testing uses probability to characterize the proportion of cases in which a null hypothesis H_0 would be erroneously rejected in a hypothetical long-run of sampling, an error probability.

We may even refer to frequentist statistical inference as error probability statistics, or just **error statistics.**

His analogy is with calibrating instruments:
"Arguments involving probability only via its (hypothetical) long-run frequency interpretation are called frequentist. That is, we define procedures for assessing evidence that are calibrated by how they would perform were they used repeatedly. In that sense they do not differ from other measuring instrument." (Cox 2006, p. 8)

As with the use of measuring instruments, <u>we employ the performance features to make inferences about aspects of the particular thing that is measured,</u> (aspects that the measuring tool is appropriately capable of revealing).

It's not hard to see how we do so, if we remember the central problem of inductive inference

The **argument** from:

H entails data **y, (*H* "fits" y)**
<u>y is observed,</u>
therefore *H* is correct,

is, of course, *deductively invalid*.

A central problem is to be able nevertheless to warrant inferring *H*, or regarding y as good evidence for *H*.

Although **y** accords with *H*, even in deterministic cases, many rival hypotheses (some would say infinitely many) would also fit or predict **y**, and would pass as well as *H*.

In order for a test to be <u>probative</u>, one wants the prediction from *H* to be something that would be very difficult to achieve, and not easily accounted for, were *H* false and rivals to *H* correct (whatever they may be)

i.e., the test should have high probability of <u>falsifying</u> *H*, if *H* is false….

Statistical Significance Tests

While taking elements from Fisherian and Neyman-Pearson approaches, Cox's conception differs from both

1. We have empirical data ***y*** treated as observed values of a random variable (vector) ***Y***.

   y is of interest in so far as it provides information about the probability distribution of **Y** as defined by the relevant **statistical model**—often an abstract and idealized representation of the underlying data generating process.

2. We have a hypothesis framed in terms of parameters of the distribution, <u>the hypothesis under test</u> or the <u>null hypothesis</u> H_0.

Note: statistical hypotheses are not events
An elementary example:

> **Y** consists of *n* Independent and Identically Distributed (IID) components (r.v's), Normally distributed with unknown mean μ and possibly unknown standard deviation σ :

$$Y_i \sim \text{NIID}(\mu,\sigma^2), \quad i = 1,2,\ldots,n.$$

- A <u>simple hypothesis</u> is obtained if the value of σ is known and H_0 *asserts that* $\mu = \mu_0$, a given constant.

We find a function $T = t(\mathbf{Y})$ of the sample, the test statistic, such that:

(i) The larger the value of t the more inconsistent or *discordant* the data are with H_0 in the respect tested

(ii) The statistic $T = t(\mathbf{Y})$ has a (numerically) known probability distribution when H_0 is true.

The *p-value (observed statistical significance level)* corresponding to t(**y**) is:

$$\mathbf{P(t(Y) \geq t(y); H_0):=P(T \geq t; H_0)}$$

regarded *as a measure of accordance with H_0* in the respect tested.

Low p-values, if properly computed, count as evidence against H_0 (evidence of a specified *discrepancy* or *inconsistency* with H_0).

Cox's treatment contrasts with what is often taken as the strict Neyman-Pearson (N-P) formulation:

- Rejects *recipe-like view of tests*, e.g., set a preassigned threshold value α and "reject" H_0 if and only if $p \leq \alpha$.

However, such "behavioristic" construals have roles in Cox's conception: While a relatively mechanical use of p-values is widely lampooned, there are contexts where it serves as a screening device, decreasing the rates of publishing misleading results (e.g., microarrays)

Virtues: impartiality, relative independence from manipulation, gives protection of known and desirable long-run properties.

But the main interest and novelty here is developing an *inferential or evidential rationale.*

$$\mathbf{P(t(Y) \geq t(y);}\ \boldsymbol{H_0}\mathbf{):=P(T \geq t;}\ \boldsymbol{H_0}\mathbf{)}$$

For example: low p-values, if properly computed, count as evidence against H_0 (evidence of a specified *discrepancy* or *inconsistency* with H_0).

Why?

It’s true that such a rule provides low error rates (i.e., erroneous rejections) in the long run when H_0 is true, a *behavioristic argument.*

Cox’s formulation:

> Suppose that we were to treat the data as just decisive evidence against H_0, then, in hypothetical repetitions, H_0 would be rejected in a long-run proportion p of the cases in which it is actually true.

But what matters for us, is that such a rule also provides a way to determine whether a *specific data set* provides evidence of <u>inconsistency with or discrepancy</u> from H_0.

…along the lines of the probative demand for tests…

FEV: Frequentist Inductive Inference Principle

The reasoning is based on the following frequentist inference principle (with respect to H_0):

(FEV) y fails to count as evidence against H_0, if, such discordant results are fairly frequent even if H_0 is correct.

y counts as evidence of a discrepancy only if (and to the extent that) a less discordant result would probably have occurred, if H_0 correctly described the distribution generating **y**.

So if the p-value is not small, it is fairly easy (frequent) to generate such discordant results even if the null is true, so this is not good evidence of a discrepancy from the null…

Where "such discordant results" are those as or even further from H_0 than the outcome observed.

Weight Gain Example.

To distinguish between this "evidential" use of significance test reasoning, and the familiar appeal to “low long-run erroneous behavior” (N-P), consider a very informal example:

Suppose that weight gain is measured by a multiplicity of well-calibrated and stable methods, and the results show negligible change over a test period (e.g., before and after England).
This is grounds for inferring that my weight gain is negligible within limits set by the sensitivity of the scales.

Why?

While it is true that by following such a procedure in the long run one would rarely report weight gains erroneously, that is not the rationale for the particular inference.

The justification is rather that the error probabilistic properties of the weighing procedure reflect what is actually the case in the *specific instance*.

(It informs about the *specific cause* of the lack of a recorded increase in weight).

Low long run error, while *necessary*, is <u>not</u> *sufficient* for warranted evidence in a particular case.

General Severity Account of Evidence for statistical and non-statistical cases: FEV falls under a more general account of evidence or inductive inference, extending beyond statistical hypotheses:

Data* y *provide evidence for a claim H, just to the extent that H passes a <u>severe test</u> with y

I am not hanging this extension on Cox, unless…..

The intuition behind requiring **severity** is that:

> *Data* $\boldsymbol{y}_0$ *in test T provide good evidence for inferring H (just) to the extent that H passes severely with* $\boldsymbol{y}_0$*, i.e., to the extent that H would (very probably) not have survived the test so well were H false.*
>
> *This may be a quantitative or a qualitative assessment.*
>
> *(Popper, Peirce)*

Frequentist statistical methods can (and often do) supply tools for inductive inference by providing methods for evaluating the *severity* or *probativeness* of tests —although they don't directly do so, the severity rationale gives guidance in using them in this way.

> Just as one makes inferences about changes in body mass based on performance characteristics of various scales, *error probabilities* of tests indicate the *capacity of the particular test to have revealed inconsistencies in the respects probed*, and this in turn allows relating p-values to statements about the underlying process.

Peirce: shall we assume instead the scales read my mind and mislead me just when I don't know the weight, but doo fine with items of known weight, e.g., 5 lb potatoes.

—itself a highly unreliable way to proceed in reasoning from instruments…

A Multiplicity of Types of Null Hypotheses (Cox's taxonomy)

Considerable criticisms and abuses of significance tests would have been avoided had Cox's taxonomy in (1958) been an integral part of significance testing.

Types of Null Hypothesis (multiple uses in piecemeal steps involved in linking data and models to learn about modeled phenomena)

1. **Embedded null hypotheses**
2. **Dividing null hypotheses**:
3. **Null hypotheses of absence of structure**
4. **Null hypotheses of model adequacy**
5. **Substantively-based null hypotheses.**

Evaluating such local hypotheses for these piecemeal learning goals demands its own criteria

1. Embedded null hypotheses

Consider a parametric family of distributions f(**y**; θ) indexed by unknown (vector of) parameters $\theta = (\mu, \lambda)$, where λ denotes unknown *nuisance* parameter(s).

The **null** and **alternative hypotheses** of interest are:

$$H_0\text{: } \mu = \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0$$

Central tasks, often ignored, take central stage in Cox's discussion and indeed are an important part of the justification of this approach (though must be skimpy here):

(a) —need to choose appropriate test statistic $T(\boldsymbol{Y})$

(b) —need to be able to compute the probability distribution of *T* under the null (and perhaps alternative) hypotheses (it should not depend on nuisance parameters)

(c) —need to collect data ***y*** to so that they satisfy adequately the assumptions of the relevant probability model.

If we succeed with all that….

p-value not small—evidence the data are consistent with the null…

But no evidence against is not automatically evidence for the null…

but we can infer the *absence of a discrepancy* from (or concordance with) H_0

To infer the *absence of a discrepancy* from H_0 as large as δ, examine the **probability** $\beta(\delta)$:

$\beta(\delta)$ = Prob[a result as or more discordant from H_0 (than is our data) would have occurred; evaluated under the assumption that $\mu_1 = \mu_0 + \delta$],
i.e.

$$\beta(\delta) = P(t(\mathbf{Y}) > t(\mathbf{y}) \ ; \ \mu_1 = \mu_0 + \delta)$$

If $\beta(\delta)$ is near 1, then, following FEV, the data are *good evidence* that $\mu < \mu_0 + \delta$.

Interpreting $\beta(\delta)$:

$\beta(\delta)$ may be regarded as *stringency or severity* with which the test has probed the discrepancy δ;
that is, $\mu < \mu_0+\delta$ has passed a <u>severe test:</u>
(other terms: precision, sensitivity)

- Avoids unwarranted interpretations of "consistency with H_0" with insensitive tests ("fallacies of acceptance").
- More relevant to specific data than is a test's <u>power</u>, which is calculated relative to a predesignated critical value c_α beyond which the test "rejects" the null.

Power at $\mu_1 = \mu_0 + \delta$: $\pi(\mu_1) = P(t(\mathbf{Y}) > c_\alpha \ ; \ \mu_1=\mu_0+\delta)$

In contrast: $\beta(\delta) = P(t(\mathbf{Y}) > t(\mathbf{y}) \ ; \ \mu_1=\mu_0+\delta)$

An important passage in Cox 2068, p. 25)
"In the Neyman-Pearson theory of tests, sensitivity of a test is assessed by the notion of power….. In the approach adopted here the assessment is via the distribution of the random variable P, again considered for various alternatives"

P here is the significance probability regarded as a random variable—which it is…

Cox makes an equivalent *move referring to confidence intervals—I'll come back to…*

<u>small p-values</u> (evidence of *some* discrepancy in direction of alternative).
Critics correctly note that the p-value ALONE doesn't tell you the size of the discrepancy indicated "effect size"), but it, together with sample size, can be used to determine this:

- If a test is so sensitive that so small a p-value is probable, even when $\mu < \mu_0 + \delta$,
  then a small p-value is *not evidence* of a discrepancy from H_0 in excess of δ .

  So by this reasoning we also avoid "fallacies of rejection"

5. Substantively-based null hypotheses.

A theory ***T*** predicts that H_0 is at least a very close approximation to true situation (perhaps T has already passed several theoretical and empirical tests)

Rival theory ***T**** predicts a specified discrepancy from H_0 and the test is designed to discriminate between ***T*** and ***T**** in a thus far untested domain.

Focus on interpreting
p-value not small.

H_0 may be formed deliberately to let ***T*** "stick its neck out"

Discrepancies from ***T*** in the direction of ***T**** are given a very good chance to be detected, so *if no significant departure is found, this constitutes evidence for T in the respect tested*

Famous "null results" take this form
(e.g., set the GTR predicted values at 0)
e.g., rivals to the GTR predicted a breakdown of the Weak Equivalence Principle (WEP) for massive self-gravitating bodies, the earth-moon system: this effect, the Nordvedt effect would be 0 for GTR (identified with the null hypothesis) and non-0 for rivals.

Measurements of the round trip travel times between the earth and moon (between 1969 and 1975) set upper bounds to the possible violation of the WEP, and because the tests were sufficiently sensitive, these measurements provided good evidence that the Nordvedt effect is absent, i.e., evidence for the null hypothesis

Such a negative result does not provide evidence for all of GTR (in all its areas of prediction), but it does provide evidence for its correctness with respect to this effect.

Some argue I should allow inferring all of GTR (Chalmers, Laudan, Musgrave) —but I think it is this piecemeal way that inquiries enable progress (realizing that NOT all of GTR has been warranted provoked developing rivals…)

Unlike the large-scale theory testers, we are not after an account of "theory choice" but rather learning about aspects of theoretical phenomena by local probes…

Statistical reasoning is at the heart of inquiries which are broken down into questions about parameters in statistical distributions intermediate between the full theory and the actual data.

My conception is to view them as probing for piecemeal errors—and note, an error can be any mistaken claim or flawed understanding—both empirical and theoretical!

- A fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that *inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results*.

- Although the complexity of the story makes it more difficult to set out neatly, as, for example, if a single algorithm is the whole of inductive inference, the payoff is an account that approaches the kind of full-bodied arguments that scientists build up in order to obtain reliable knowledge.

- The goal is not representing beliefs or opinions (as in personalist Bayesian accounts) but avoiding being misled by beliefs and opinions

Series of Confidence Intervals

(goal: to distinguish inferential and behaviorist justifications while avoiding an infamous fallacy of R.A. Fisher)

Consider our Normal mean "embedded" example, rather than running a significance test we may form the **$1 - \alpha$ upper confidence bound, $\mathrm{CI}^{U}(\boldsymbol{Y}; \alpha)$**

for estimating mean μ, let σ be known, $\sigma_y = \sigma n^{-.5}$.

The upper $1 - \alpha$ limit is $\bar{Y} + k(\alpha)\sigma_y$

$k(\alpha)$ the upper α-value of the standard Normal distribution.

Pick a small α, say .025

The upper .975 limit is $\bar{Y} + 1.96\sigma_y$

Consider the inference:

Infer (or regard the data as evidence for)

$\mu < \bar{y} + 1.96\sigma_y$

$\mathrm{CI}^{U}(\boldsymbol{Y}; .025)$

One rationale is that it instantiates an inference rule that yields true claims with high probability (.975) because

$$P(\mu < \bar{Y} + 1.96\sigma_y) = .975$$

Whatever the true value of μ

The procedure has high long-run “coverage probabilities.”

They "rub-off" on the particular case.

Instead we might view $\mu < \bar{y} + 1.96s_y$as an inference from a type of *reductio ad absurdum* argument:

suppose in fact that this inference is false and the true mean is μ^*, where $\mu^* > \bar{y} + 1.96\sigma_y$.

Then it is very probable that we would have observed a larger sample mean:

$$P(\bar{Y} > \bar{y}; \mu^*) > .975.$$

Therefore, one can reason, $\bar{y}$ is inconsistent at level .975 (or .025) with having been generated from a population with μ in excess of the upper limit.

This reasoning is captured in FEV (Mayo and Cox).

Aside: This statistic is directly related to a test of $\mu = \mu_0$ against $\mu < \mu_0$.

In particular, $\bar{Y}$ is statistically significantly smaller than values of μ in excess of $CI^U(\boldsymbol{Y}; \alpha)$ at level α.

Fisher's fiducial error

This may make μ look like a random variable----but it is not; these claims do not hold once a specific $\bar{y}$ is plugged in for $\bar{Y}$

$P_{\mu}(\mu < \bar{Y} + 0\sigma_x) = .5$

$P_{\mu}(\mu < \bar{Y} + .5\sigma_x) = .7$

$P_{\mu}(\mu < \bar{Y} + 1\sigma_x) = .84$

$P_{\mu}(\mu < \bar{Y} + 1.5\sigma_x) = .93$

$P_{\mu}(\mu < \bar{Y} + 1.96\, \sigma_x) = .975$

"Our attitude toward the statement [$\mu < \bar{Y} + 1.96\, \sigma_x$] might then be taken to be the same as that to other uncertain statements for which the probability is [.975] of their being true, and hence…is virtually indistinguishable from a probability statement about [μ]. However, this attitude is in general incorrect, in our view because the confidence statement can be known to be true or untrue…… The system of confidence limits simply summarizes what the data tell us about μ, given the model….It is wrong to combine confidence limit statements about different parameters as though the parameters were random variables." (Cox and Hinkley 1974, p. 227)

But what exactly is it telling us about μ?

$P_{\mu}(\mu < \bar{Y} + 0\sigma_x) = .5$

$P_{\mu}(\mu < \bar{Y} + .5\sigma_x) = .7$

$P_{\mu}(\mu < \bar{Y} + 1\sigma_x) = .84$

$P_{\mu}(\mu < \bar{Y} + 1.5\sigma_x) = .93$

$P_{\mu}(\mu < \bar{Y} + 1.96\,\sigma_x) = .975$

Answer: it tells you which discrepancies are and are not indicated (at various degrees of severity)

If we replace P_{μ} with SEV (the degree of severity or even "corroboration") the claims are true even after substituting the observed sample mean
(the severity with which the inference has passed the test with the data)

Selection Effects, double-Counting, Violations of Use-Novelty

It can happens that the null hypothesis or the test statistic are constructed, or selected for testing, by preliminary inspection of the data, so as to yield agreement (or disagreement) between data and hypotheses
In some cases the actual procedure generating the final test result is altered:

The general point involved has been discussed extensively in both philosophical and statistical literatures.

- In the former under such headings as *requiring novelty or avoiding ad hoc hypotheses* (use-constructions, etc.)

  Some of us, notably, Worrall and I, have long debated this issue….(he calls it the UN Charter)

- Under the latter, as rules against peeking at the data, shopping for significance, data mining, etc., for taking *selection effects* into account.

Rather than take an “always” or “never” view, Cox set out a taxonomy of cases, some where adjustments were called for, but others not.
One of my goals has been to provide criteria for *when* various data dependent selections matter and *how* to take account of their influence on error probabilities.

It is well-known that certain types of double counting can lead to unreliable inferences: For example, if one is allowed to search through several factors and selectively reports just those that show (apparently) impressive correlations, there is a high probability of erroneously inferring a real correlation.

However, it is equally clear that there are reliable procedures for using data both to identify and test hypotheses: the use of a DNA match to identify a criminal, radiointerferometry data to estimate the deflection of light, and in using a ruler to measure the length of a table.

Here, although the inferences (about the criminal, the deflection effect, the table length) were constructed to fit the data, they were deliberately constrained to reflect what is correct, at least approximately.

Critics mistakingly assume that any case involving data-dependent specifications or double-counting must be similarly altered by the frequentist error statistician!

I have examples of each which I will not be able to discuss….

Example 1: **Hunting for Statistical Significance**

Investigators have 20 independent sets of data, each reporting on different but closely related effects. After doing all 20 tests, with 20 nulls, H_{0i}, i = 1, …20 they report only the smallest p-value, e.g., 0.05, and its corresponding null hypothesis, say $H_{013.}$

e.g., there is *no difference between some treatment* (a childhood training regimen*) and a factor*, f_{13} (some personality characteristic later in life).

In particular, if the null hypothesis chosen for testing just because it's a factor that yields a large test statistic, the probability of finding *some such discordance or other* may be high even under the null.

Thus, following FEV, we would not have genuine evidence of inconsistency with the null, and unless the p-value is modified accordingly, the inference would be misleading.

The "hunting procedure" does a very poor job in alerting us to, in effect, *temper our enthusiasm (Cox),* even where such tempering is warranted.

This "hunting" procedure should be compared with a case where H_{013} was preset as the single hypothesis to test, and the small p-value found.

- In the hunting case, the possible results are the possible statistically significant factors that might be found to show a "calculated" statistical significant departure from the null. The relevant type 1 error probability is the probability of finding at least one such significant difference out of 20, even though the global null is true (i.e., all twenty observed differences are due to chance).
- The probability that this procedure yields erroneous rejection differs from, and will be much greater than, 0.05 (and is approximately 0.64).
- There are different, and indeed many more, ways one can err in this example than when one null is preset, and this is reflected in the adjusted p-value.

Critics often assume the frequentist error statistician always gives less weight to an inference based on data snooping and searching…
Admittedly statistical texts do not explicitly draw needed distinctions…

Example 2. Hunting for a Murderer

(hunting for the source of a known effect by eliminative induction)
Testing for a DNA *match* with a given specimen, known to be that of the murderer, a search through a data-base of possible matches is done one at a time.

We are told, in a fairly well-known presentation of this case, that:
P(DNA match; not murderer) = very small
P(DNA match; murderer) ~ 1

The first individual, if any, from the data-base for which a match is found is declared to truly match the criminal, i.e., *to be the murderer*.

(The null hypothesis, in effect, asserts that the person tested does NOT "match the criminal"; so the null is rejected iff there is an observed DNA match.)

Example 2 is superficially similar to Example 1, finding a DNA match being somewhat akin to finding a statistically significant departure from a null hypothesis: one searches through data and concentrates on the one case where a "match" with the criminal's DNA is found, ignoring the non-matches.

If one adjusts for "hunting" in Example 1, shouldn't one do so in broadly the same way in Example 2?

No!

(Although some have erroneously supposed frequentists say “yes”)

In Example 1 the concern is inferring a genuine, “reproducible" effect, when in fact no such effect exists; in Example 2, there is a known effect or specific event, the criminal's DNA, and reliable procedures are used to track down the specific cause or source (as conveyed by the low "erroneous-match" rate.)

- The probability is high that we would not obtain a match with person i, if i were not the criminal; so, by FEV, finding the match is excellent evidence that i is the criminal. Moreover, each non-match found, by the stipulations of the example, virtually excludes that person;

Note: the contrast in hunting for a DNA match is finding a match with the first person tested, as opposed to hunting through a data base

In example 2, the more negative results found, the more the inferred "match" is fortified; whereas in Example 1 this is not so.

Quick Summary so far….

1. The FEV is the basis for the "inferential" construal of frequentist statistics that emerges in the work of Cox—piecing together discussions over the years
(Mayo and Cox, 2006, "Frequentist Statistics as a Theory of Inductive Inference)

The quantitative aspects are in the form of degree of stringency and sizes of discrepancies detected; not degrees of belief/confirmation in hypotheses

2. FEV is post-data (sensitive to the actual data), in contrast to the pre-data criteria of strict Neyman-Pearson account: (avoids its coarse appraisals, fallacies of significant, and of non-significant, results; is the basis for identifying the rationale for adjusting error probabilities to take account of "selection effects" —just when this is warranted)

…connects to the more general and thorny issue that Cox has done more than any other error statistician to address.

3. Relevance of the sampling distribution for the learning goal at hand.

~the reference class problem for frequentists (with important differences)
Cox stresses the need "for some assurance that with our particular data currently under analysis sound conclusions are drawn. This raises important issues of ensuring, as far as is feasible, the relevance of the long run to the specific instance" (Cox 2006, p. 8)

A Mixture of Tests of different precisions
(Cox 1958, 360)

Among the most widely cited example in statistics ever!

"Suppose that we are interested in the mean of a normal population and that, by an objective randomization device, we draw either (i) with probability .5, one observation, y, from a normal population of mean μ and standard deviation σ_1 or (ii) with probability .5, one observation y, from a normal population of mean μ and standard deviation σ_2, where σ_1, σ_2 are known, $\sigma_1 << \sigma_2$and where we know in any particular instance which population has been sampled."

Flip fair coin to decide which of two experimental tests (or instruments) to run: E', and E": e.g., E' might use a much larger sample size.

E': Y is normal $N(\mu, 10^{-4})$

E'' :(Y is normal $N(\mu, 10^{4})$

Perhaps we are testing a null hypothesis $\mu = 0$, and reporting the p-value.

We have a set of data **y**,

if **y** had come from E', the p-value is $p'(\mathbf{y})$ but

if **y** came from E" its p-value is $p''(\mathbf{y})$.

Which hypothetical series of repetitions should be used to determine the associated error probabilities?

A critic of frequentist tests might allege:
The overall type 1 error probability for the mixture of tests is the average: $\{p'(\mathbf{y}) + p''(\mathbf{y})\}/2$

After all, we're to consider all possible repetitions, which could include different outcomes of the coin toss….
But once you know the result of the randomizer, say that it was E", it seems you should report the p-value as is $p''(\mathbf{y})$.

Were it true that the frequentist tester would have to report the average p-value, these tests would be open to criticism. But would they have to? No.

- highly misleading of the stringency of the actual test.
- an experimenter who used a very imprecise tool gets credit from the fact that he could have done a better job and used a precise tool enough to have the randomizer choose a far less precise tool.
- From the perspective of interpreting the specific data actually available, this makes no sense.

(My version: we use a randomizer that nearly always directs us to use a very precise and reliable instrument, but with small probability tells us to just ask Isaac to guess. Now spoze data y in front of me resulted in one of the rare cases that allows Isaac to just guess…. Overall, this disjunctive test might rarely err but that is irrelevant in appraising Isaac's guess.)

> **Cox: Weak Conditionality Principle (WCP):** In a mixture experiment, if it is known which experiment produced the data, inferences about θ *are appropriately drawn in terms of the sampling distribution* of the experiment known to have been performed.

Once we know the data have been generated by E_j, given that our inference is about some aspect of E_j, our inference should not be influenced by whether a coin was tossed to decide which of two experiments to perform, and the result was to perform E_j.

Although Cox (1958) made this point long ago, it is still given as one of 2-3 central examples against frequentist tests.

Worse, it is supposed that to avoid it, we must condition on the actual data

But, as Cox pointed out long ago, conditioning on the experiment actually performed (i.e., using the correct sampling distribution) does not entail conditioning on the specific data observed.

Conditioning on the data would entail no use of error probabilities since these always invoke outcomes other than the one actually observed!

Such irrelevance of the sampling distribution is embodied by the *likelihood principle*…

It's here where Bayesians often find an opening for a suggested bridge….

For a frequentist error statistician it's a bridge too far….

“It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma” (Cox and Mayo 2009)

—Such texts are seriously misleading!

So, what is this likelihood principle?

Perhaps THE key issue in the philosophy of statistics battles

The likelihood function is central in all the accounts but for those who endorse the (strong) *likelihood principle*, likelihoods suffice to convey “all that the data have to say”

> According to Bayes’s theorem, $P(\boldsymbol{x}|\mu)$ ... constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, if *y* is the datum of some other experiment, and if it happens that $P(\boldsymbol{x}|\mu)$ and $P(\boldsymbol{y}|\mu)$ are proportional functions of μ (that is, constant multiples of each other), then each of the two data ***x*** and ***y*** have exactly the same thing to say about the values of μ… (Savage 1962, p. 17.)

—the error probabilist needs to consider, *in addition*, the sampling distribution of the likelihoods.

—significance levels and other error probabilities all violate the likelihood principle (Savage 1962).

Note: likelihood fixes the *actual* outcome, while error statistics considers outcomes *other than the one observed* in order to assess the error properties.

Can illustrate using an example from Cox's taxonomy of **Null hypotheses of absence of structure:** no effect, difference "due to chance"

$$H_0\text{: } \mu = 0 \text{ vs. } H_1\text{: } \mu \neq 0$$

in Normal testing, 2-sided. Instead of fixing the same size *n* in advance, n is determined by a ***stopping rule*:**

Keep sampling until $\bar{Y}$ reaches the .05 "significance level", i.e., until

$$|\bar{Y}| \geq 1.96\sigma/\sqrt{n}).$$

This stopping rule is guaranteed to end (it's "proper").

Whereas with n fixed in advance the type 1 error probability is .05, this stopping rule leads to an actual significance level that would differ from, and be greater than .05.

By contrast, likelihoods are unaffected by this stopping rule, "The likelihood principle emphasized in Bayesian statistics implies, … that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proved or disproved (p. 193)…This 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, Savage 1963, p. 239).

For Cox, (e.g., Cox and Hinkley 1974,p. 51) this example is itself a reductio of the LP, because it allows

getting the wrong answer with high or even maximal probabilities.

"Even if in fact $\mu = 0$, there always appears to be strong evidence against $\mu = 0$ (p. 51) (contradicts *weak repeated sampling principle*—method should not allow errors with high probability)

This happens because because:
The likelihood at $\mu = 0$ can be made arbitrarily smaller than $\mu = \bar{y}$ (the observed sample mean).

So there's a conflict between error statistical principles and the Likelihood Principle at a very basic level---

Peter Armitage (Savage Forum, 1962) shows that optional stopping allows the same kind of problem for Bayesians with "objective" or "uninformative" priors: a .05 significant difference corresponds to the null hypothesis getting a .05 posterior probability.

The Bayesian, using this stopping rule, is guaranteed to give a low probability to the null, even though it is true.

It's especially telling to see how the identical point can be made in terms of confidence intervals:

Optional Stopping with (2-sided) Confidence Intervals

Keep sampling until the (usual) 95% confidence interval excludes 0

Berger and Wolpert (the Likelihood Principle, 1988) concede that using this stopping rule

"has thus succeeded in getting the [Bayesian] conditionalist to perceive that $\mu \neq 0$, and has done so honestly. (pp. 80-81)

This seems a striking admission—especially as they will assign a probability of .95 to the truth of the interval estimate:

$$\mu = \bar{y} \pm 1.96\sigma/\sqrt{n}$$

How then can there be ongoing movements toward unifications between frequentists and objective Bayesian approaches?

Some "unificationists" (cleverly) co-op the error probability term:

<u>Frequentist error statistician</u>: there is evidence against the null when the p-value is low (given evidence)

<u>Bayesian unifier</u>: there is evidence against the null when the posterior probability to the null is low (given evidence)

In the latter, the posterior probability IS the error probability (associated with the null)

Assuming as they do that the latter are "what we really want", Howson (1997), and others, declare the frequentist approach <u>unsound</u> because <u>a low p-value need not correspond to a low posterior to the null.</u>

This could only be so if frequentists claimed p-values were posteriors, which of course they do not…

As for whether it's what we really want, this has not been shown for any of the priors (subjective, reference, other)

Still, it's very interesting for our goals to understand how this leads Berger and other "reference" Bayesians to offer their Bayesian test as something the frequentist should embrace: using certain priors, the p-value and the posterior agree.

("Could Fisher, Jeffreys, and Neyman have Agreed on Testing?")

The conflict between p-values and Bayesian posteriors: common example is the 2- sided test of what Cox calls

Dividing null hypotheses:

The null of zero difference divides the possible situations into two qualitatively different regions,

e.g., compared with a standard a new drug increases decreases survival rate.

To discriminate H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

(The difference between p-values and posteriors are far less marked with one-sided tests).

"Assuming a prior of .5 to H_0, with n = 50 one can classically 'reject H_0 at significance level p = .05,' although $P(H_0|\boldsymbol{x}) = .52$ (which would actually indicate that the evidence favors H_0)." (Berger and Sellke)

As the sample size increases, the conflict becomes more noteworthy.

If *n* = 1000, a result statistically significant at the .05 level leads to a posterior to the null of .82!

What warrants such a prior?

		n						
p	***t***	***1***	***5***	***10***	***20***	***50***	***100***	***1,000***
.10	**1.645**	**.42**	**.44**	**.47**	**.56**	**.65**	**.72**	**.89**
.05	**1.960**	**.35**	**.33**	**.37**	**.42**	**.52**	**.60**	**.82**
.01	**2.576**	**.21**	**.13**	**.14**	**.16**	**.22**	**.27**	**.53**
.001	**3.291**	**.086**	**.026**	**.024**	**.026**	**.034**	**.045**	**.124**

Some claim the prior of .5 is a warranted frequentist assignment:

Hypothesis H_0 was randomly selected from an urn of null hypotheses in which 50% are true

(*) Therefore P(H_0) = p
What should go in the urn of hypotheses?

For the frequentist: either H_0 is true or false the probability in (*) is fallacious and results from an unsound instantiation: **Fallacy of Probabilistic Instantiation**

.

Some suggest an “impartial” or “uninformative” Bayesian prior gives .5 to H_0**,** the remaining .5 probability being spread out over the alternative parameter space,
Jeffreys, 1939.

This “spiked concentration of belief in the null” is at odds with the prevailing view “we know all nulls are false”.

Update Contemporary “Impersonal” Bayesianism: (the following comes from Cox's recent criticisms of impersonal Bayesians, included in Cox and Mayo 2007):

1. What do reference posteriors measure?

Because of the difficulty of eliciting subjective priors, and because of the reluctance among scientists to allow subjective beliefs to be conflated with the information provided by data, much current Bayesian work in practice favors conventional “default”, “uninformative,” or “reference”, priors .

- A classic conundrum: there is no unique “noninformative” prior. (Supposing there is one leads to inconsistencies in calculating posterior marginal probabilities).
- Any representation of ignorance or lack of information that succeeds for one parameterization will, under a different parameterization, entail having knowledge.

Giving up now on uninformative priors, contemporary “reference” Bayesian research seeks priors that are simply *conventions* to serve as weights for reference posteriors.

- The priors are not to be considered expressions of uncertainty, ignorance, or degree of belief.
- Conventional priors may not even be probabilities - a constant or flat prior for a parameter may not sum to one (improper prior). (Berger and Bernardo)

If priors are not probabilities, what then is the interpretation of a posterior?

2. Problems with nuisance parameters

We considered just a single parameter, but more commonly there are other “nuisance” parameters; the error statistician ensures that unknown nuisance parameters exert minimal threats to the validity of p-vale and other frequentist calculations.

By contrast, Bayesians require distributions for each unknowns.

Calculation of a reference prior is complicated and it depends on whether it is a parameter of interest or a nuisance, and on the “order of importance” in which nuisance parameters are arranged.

3. Priors depend on sampling distributions (choice of model chosen for inference)! Bayesian incoherent
The result is to:
(a) forfeit what is often considered a benefit of the Bayesian approach, and
(b) to violate the likelihood principle—often thought to be the cornerstone of Bayesian coherency.

If the prior is to represent information why should it be influenced by the sample space of a contemplated experiment?

Violating the likelihood principle introduces incoherency into the reference Bayesian account.

Reference Bayesians: it is “the price” that has to be paid for objectivity.

Default Bayesian priors seem to wreck havoc with basic Bayesian foundations, but without the payoff of an objective, interpretable output—**subjective Bayesians understandably are unhappy with this "O-Bayesian movement"**
— all this demands study by Bayesian philosophers

4. Reference posteriors with good frequentist properties

Reference priors are touted as having some good frequentist properties, at least in one-dimensional problems.

That a theory can be made to match known successes does not redound as strongly to that theory as were the successes to grow from first principles or basic foundations. (Especially where achieving this imposes violations to its initial basic theories or principles.)

If you want error probabilities, why not use techniques that provide them directly? (with objective checks on underlying assumptions?)

Recent O-Bayesian concessions, e.g., taking the stopping rule into account: " reinforces the idea that in reality that sort of approach is really just a technical trick to do frequentist inference by the back door" (Cox, 2008).

Philosophers wishing to champion Bayesian accounts need to grapple with thee questions and I hope to hear their thoughts on this during our conference.