Classical Statistics: Smoke and Mirrors

George Gabor
Department of Mathematics and Statistics
Dalhousie University
Halifax, NS
Canada

Motto:
Of all fields of scientific endeavor the one that is most bewitched by its own language is classical statistics.

PART 1
Classical Statistics

Classical (frequentist, orthodox) statistics is (among other faults)

irrelevant & incoherent & incorrect

The reasons that such a faulty vehicle enjoys almost universal acceptance are

historical, psychological, sociological, political.

IRRELEVANCE

Consider the simplest of all possible problems of inference:

> Given a finite population of known size a certain unknown number of which has a certain attribute, say, A. What can we say about this unknown number if we have observed a given segment (the sample) of the population?

Now this is a straightforward problem of inference, i.e. a question of

"**how to reason consistently and honestly from incomplete information so that we take fully into account what is known, but avoid assuming what is not known**" [6]

It can be solved by applying simple inferential logic and the result is a probability distribution of the unknown given all that is known.

As shocking as it may sound, classical statistics is helpless in the face of even this simplest of all imaginable problems. As a way of dealing with it, it offers a solution to a completely **different** and utterly **irrelevant** problem (without, of course, coming clean what it is doing) that sounds like this:

> If we consider **all possible samples** of the same size, what quantity would **on average** be closest (in some specified sense) to the unknown number (the estimate)? and how far is it expected, again **on average**, to wander away from it (the variance)?

The altered problem is not a problem of inference. It calls for a speculation on the behavior of certain averages in which what is known (the sample plus a minimum of background information) plays only a minor part. Unknown and unobserved data, **all that could have happened but didn't**, are actually given equal weight to that of the observed, i.e. actual evidence carries very little weight.

Non-discrimination between what is known and what is not is perhaps politically very correct, but is it good inference? Is it inference at all?

But that is not all.

The surreptitious alteration of the original problem is accompanied by arias about the sample being selected "randomly". A never defined, and in fact meaningless word designed to bamboozle the natives. And it does. Just as it fools the statisticians themselves.

By a similar bait and switch scam,

classical statistics turns every problem of inference into an irrelevant problem of averages over all possible outcomes in which <u>all that could have happened but didn't</u> play just as important role as what actually happened, and in which a good part of prior knowledge is simply ignored.

In other words, what is unobserved (and therefore can't be part of the evidence) is just as essential as the what is observed, i.e. the evidence.

And that is still not all.

The error then is always compounded by another error:

Conveniently ignoring the fact that in the altered problem probabilities belong exclusively to the collective of all possible samples, the solution to the altered problem is applied to a single observation hoping that the two wrongs - altering the problem and transferring probability inexplicably from the collective to a single occurrence - somehow make one right.

In fact, the nowhere stated but absolutely essential

fundamental dogma of classical statistics (FDCS)

is that

the composite effect of the two errors is nil, and the result is inferentially correct.

In what sense it might be correct is a mystery since inference is never defined. When objected to, what on gets is an aria: It works. Again, in what sense it does (if indeed it does) is never explained. In fact FDCS is profoundly wrong.

Example 1

Suppose you have to measure the weight μ of a chemical compound. You have two measuring instrument of widely different precisions: σ_0=1, and σ_1=10 (units). It is your habit to flip a coin to decide which one to use before you make the measurement (this is called "randomization" in the parlance of orthodoxy). After you observe the result of the coin toss y (which can be either 0 or 1), you make the measurement. Based on that measurement x you need to infer the value μ.

This is a straightforward problem of inference:
Based on what is known - the data D=(x,y), and the background information I - what can we say about μ? I.e. What is the probability P(μ|DI)?

But you are trained (brainwashed) to think in different terms: frequencies. You think about μ as an unknown constant not entitled to a probability distribution. Only the "random" data D=(x,y) does because it presumably "produces" frequencies (which you think is the same as probability) in repeated trials (never mind that you may never repeat the trial).

So instead of P(μ|DI), you speculate about P(D| μ) (without actually calling it P(D|μ) because that would already confer a probabilistic status on μ). Quite a different thing, isn't it? But let's see the result of that speculation.

Assigning Gaussian error distribution to the measurements (why?), i.e. $x \sim N(\mu,\sigma)$, we have

$$\frac{x-\mu}{\sigma} = z \sim N(0,1) \text{, or}$$

$$P(\mu - z\sigma < x < \mu + z\sigma) = \Phi(z) - \Phi(-z) .$$

If you want this "coverage probability" to be, say, .95, then $\Phi(z) - \Phi(-z) = .95$, from which $z \approx 2$ follows. Thus you know that $P(\mu - 2\sigma < x < \mu + 2\sigma) = .95$.

Now please note: **This quantity is meaningless as soon as the measurement x is observed!**

But you are not even aware of that. You are used to the double somersaults of frequentist thinking. It has become your second nature.
So in blissful ignorance you proceed to perform the second somersault:

Reorder the simple inequality thus: $P(x-2\sigma<\mu<x+2\sigma)=.95$, get fooled by it, and compute it for the observed x where it does not make sense anymore. (E.g. for x=1.2, you'll happily declare that the probability is about .95 that….well, here you'll have to do a bit of fudging… is$1.2\pm 2\sigma$.) Despite the double whammy, you declare (according to the FDCS) that the result "makes sense" (as you have been told anyway), so you accept it.

But wait, there are other flies in the ointment. What happened to the coin? The result of the coin toss is, after all, part of the data. Indeed,

for y=0, when you use the more precise instrument, a 95% confidence interval for μ is about $x\pm 2$;
for y=1, when the measurement is much less precise, a 95% confidence interval for μ is about $x\pm 20$.

So if, say, y=0, i.e. you take the more precise interval $x\pm 2$, its coverage probability is nowhere near .95 because you **could have had y=1** with actually equal probability (why? - a simple question orthodoxy can't answer). You didn't, but you could have. Indeed, computing the conditional probabilities

$$P(\mu-2<x<\mu+2\,|\,y=0)=.95 \text{ , but } P(\mu-2<x<\mu+2\,|\,y=1)=.16 \text{ .}$$

Thus combining the two

$$P(\mu-2<x<\mu+2)=.95P(y=0)+.16P(y=1)=\frac{1}{2}(.95+.16)=.55$$

And conversely, if y=1, i.e. your measurement less precise, and you take $x\pm 20$ as your interval, its coverage probability is higher than .95 because you **could have had y=0** with equal probability. You didn't, but you could have. Indeed,

$$P(\mu-20<x<\mu+20\,|\,y=1)=.95\text{, but } P(\mu-20<x<\mu+20\,|\,y=0)\approx 1\text{.}$$

Thus

$$P(\mu-20<x<\mu+20)=.95P(y=1)+1P(y=0)=\frac{1}{2}(.95+1)=.975$$

Let there be no mistake about it: These results are correct **coverage probabilities** - only they are doubly **irrelevant** because

a) why would anyone knowingly (that is the crux, isn't it) care about coverage probabilities?

b) why would anyone **vitiate a precise result because one could have had the less precise one; and overvalue a less precise result because one could have had the more precise one.**

INCOHERENCE

Example 2

Suppose you have series of urns of identical content: the same fixed number of red and white balls. To make inference about the unknown ratio$\theta$of the red balls in the urns, you conduct two experiments.

In Experiment 1 you take a single ball out of each of a fixed number, say 12, of the urns.
In Experiment 2 you keep on picking a single ball from each urn until you have three red ones.

Now suppose that

Exp. 1 resulted in **3 red balls out of the pre-fixed 12** (data D_1);
Exp. 2 brought the desired **3 red balls at the 12th selection** (data D_2).

Simple questions orthodoxy can't answer or answers incorrectly.

Q1. Are the trials independent?

The routine orthodox answer to this is yes. All computation is based on that assumption, though a short reflection would demonstrate the opposite.

Q2. In the long run (whatever that means), would the frequency of red balls be the same as the probability of drawing a red ball from a single urn?

The routine orthodox answer to this is yes, though there is no way probability theory alone could possibly answer this question. This goes to the very heart of what probability really is.

Q3. Would or should the inference be different in the two experiments above?

The routine orthodox answer to this is yes, because the collectives (the sample spaces of all that could have been observed) are widely different. Indeed, in Exp.1 the sample space is finite with 2^{12} elements, but it is infinite in Exp.2.

To do, for example, an orthodox hypothesis test to test the pair of hypotheses

H_0: $\theta=1/2$ as opposed to, say,

H_a: $\theta<1/2$,

the **p-value** of the test in

Exp.1 would be $p_1 \approx .073$, while in

Exp.2, it would be $p_2 \approx .033$.

The difference is due to the sample spaces (the set of **all that could have been observed**) being widely different.

If one believes the frequentist claim that the p-value of a test is a kind of weight of evidence against the null-hypothesis, than the evidence against H_0 in Exp.2 is more than twice as strong as in Exp.1.

Very spooky if you think about it:

In both experiments we observed 3 red balls out of 12. Only what **could have been** observed are different, i.e. the *intention*, in Exp.2, of the experimenter to go on indefinitely if necessary greatly influences the inference.

Q4. Does such a procedure make any sense? is the p-value relevant?

To assess the plausibility of a hypothesis H , a normal person would ask this: What is the plausibility of H in light of the evidence, i.e. What is P(H|DI) ?

But that would mean defining and assigning probabilities on a hypothesis space, a definite no-no for frequentists.

Remember: H is either true or false, and unknown constant. Only the data is "random". And in the frequentist mind set this nebulous, never defined, and in fact meaningless buzz-word "random" is the only permitted "source" of probability (which in turn is thought of as identical with frequency). That the reason to use inference and assign probabilities is that one's knowledge is incomplete does not enter the orthodox mind.

So, the frequentist is stuck with P(D|H) - and by not calling it as such he is compelled to ignore I. Not quite the same thing.

In fact it is a logical error, and simply **false** to say that P(H|DI)=P(D|HI); much less it is true that P(H|DI)=P(D|H). One can be quite different from the other.

And yet, using the well tried bait and switch scam, classical statistics sells the p-value as if it were something like P(H|DI).

Thus, goes the advise, if the p-value, the probability of the data (and a host of other values that did' not occur!! - the p-value is a strange beast) upon the hypothesis, is small, then the hypothesis can be rejected.

Let's see how this stands up in practice.

INCORRECTNESS

Example 3

Suppose a heinous crime is committed in a town of a million people. The police, under tremendous pressure to produce results, arrests the first passer by at one of precincts. This person is subjected to a very accurate test that, according to its specifications, produces false results (negative or positive) about one in a million. The result of the test turns out to be positive.

The prosecution, trained in classical statistics, argues thus.
Since, given innocence, the probability of obtaining positive test result is extremely small, it follows that the accused is overwhelmingly more likely to be guilty than not.

This sounds soo reasonable and sooo scientific to the court and the jury (and since no one around is trained in the logic of inference) that they find the accused guilty and hang him by the neck until dead.

In more precise terms the argument runs like this.
If G and $\overline{G}$ denote, resp., the guilt and innocence of the accused, and the datum is D={the test is positive}, then we know that the p-value of the test of the hypothesis of innocence $P(D|\overline{G}) \approx 10^{-6}$.

$$\Rightarrow \qquad P(\overline{G}|D) \approx 10^{-6} \quad \Rightarrow \qquad P(G|D) = 1 - P(\overline{G}|D) \approx 1$$

QED

Except that the first implication, an incorrect logical *non sequitur* of course, went unstated (as in all classical statistical hypothesis test) and thus remained unnoticed.

The correct argument should have gone like this. What we want is not P(D|$\overline{G}$), but P($\overline{G}$|DI). Using Bayes' formula we have

$$P(\overline{G} \mid DI) = \frac{P(D \mid \overline{G}I)p(\overline{G} \mid I)}{P(D \mid \overline{G}I)p(\overline{G} \mid I) + P(D \mid GI)p(G \mid I)} = \frac{1}{1 + \dfrac{P(D \mid GI)p(G \mid I)}{P(D \mid \overline{G}I)p(\overline{G} \mid I)}}$$

$$= \frac{1}{1 + \dfrac{p(G \mid I)}{10^{-6}\, p(\overline{G} \mid I)}} \approx \frac{1}{1 + \dfrac{10^{-6}}{10^{-6}}} = \frac{1}{2}$$

So because of willful ignorance and faulty logic, a person with no more than a toss-up chance to be guilty or innocent was hanged.

Note that there is no way to decide the probability of a hypothesis H upon data D without knowing what that probability was prior to the data. The prior information, ignored by frequentist procedures, turns out to be vital.

Remark

The **p-value** also suffers from being an **incoherent** measure of support in the following sense.

In testing two hypotheses H_1 and H_2 such that H_1 implies H_2 , one would expect that the rejection of H_2 would entail the rejection of H_1. And conversely, support for H_1 would entail support for H_2 .

The p-value violates this natural and fundamental requirement of **coherence**.

PART 2
INFERENCE

In classical statistics you'll never find a definition of inference. It operates with a hodge-podge of ad hoc, sometimes contradictory, and frequently violated methods, principles, etc. The tacit assumption is, of course, that the FDCS acrobatics will always land it in an inferential heaven. As the examples indicate, either this heaven is not much to be wished for, or classical statistics has lead us astray and we ended up in an inferential purgatory instead. To see which, we have to define what we mean by inference, and what we wish it to accomplish.

We wish to assign plausibilities to **propositions** (on the Boolean algebra of propositions) given a certain state of information.. The plausibility of proposition A given that some other proposition B is true is indicated by (A|B).

Desiderata [6]:

1) Degrees of Plausibility are represented by real numbers

2)
   a) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
   b) All relevant evidence is taken into account, i.e. evidence is not taken into account or ignored arbitrarily.
   c) Equivalent states of knowledge are represented by equivalent plausibility assignments.

3) Qualitative correspondence to common sense.

   This is a very reasonable technical monotonicity requirement which says that if (A|C) changes so that (A|C')>(A|C) when C is updated to C', but (B|AC')=(B|AC), then $(\bar{A}|C')<(\bar{A}|C)$ and $(AB|C')\geq(AB|C)$.

1), 2a), and 3) are structural requirements; 2b) and 2c) are interface conditions for relating to the outer world.

<u>Cox's Theorem</u> (J.of Phys., 1946; *The algebra of probable inference*, 1962)

A measure of plausibility that satisfy the desiderata above must be a monotonic function of a function $P(\bullet)$ that satisfies the following rules (the product and the sum rule, respectively).

$$i) \quad P(AB \mid C) = P(A \mid BC)P(B \mid C) = P(B \mid AC)P(A \mid C)$$

$$ii) \quad P(A \mid C) + P(\overline{A} \mid C) = 1$$

We shall use function $P(\cdot)$ itself to measure plausibility and call it

PROBABILITY

That is it. These are the rules of inference.

To do inference consistently, one must follow these two simple rules. No further principles, criteria, etc. are needed. Just as deductive logic ensures that truth and falsity flow unimpeded through a series of deductions, the rules of inference *mutatis mutandis* ensure that the probabilities arrived at the end of an inferential process are the ones, and the only ones consistent with the initial probabilities.

Suppose H is an hypothesis, D is some data, and I is the background information. Then from the sum rule we have

$$P(HD \mid I) = P(H \mid DI)P(D \mid I) = P(D \mid HI)P(H \mid I)$$

from the second equation it follows that

$$P(H \mid DI) = \frac{P(D \mid HI)P(H \mid I)}{P(D \mid I)}$$

Bayes rule.

Bingo!

Classical statistics does not follow these rules. Or any rules, for that matter. It has no conception of inference and operates with a host of ad hoc "methods" and "principles" some of which are mutually contradictory, and most of which are violated.

Therefore, its results are without any logical status and, thus, strictly speaking meaningless. But in fact, after the usual double somersault of FDCS, they can be given a formal status which is either

a) wrong, or
b) identical with the results derived by proper inferential logic.

To say that classical statistics (with the help of FDCS, of course) is just another way of doing inference is like saying that there are many ways of arriving at truth in a series of deductions (such as mathematics): logic, and some other ways. The alternative must justify itself every time by showing that its result agrees with the logical one. If so, it is not needed; if not, it is wrong. Thus

Classical statistics is either superfluous or wrong.

Probability is an extension of deductive logic (and in fact the only extension consistent with our desiderata).

The two fundamental **(strong) syllogisms** of deductive logic are:

I. $$\frac{\begin{array}{c}A \Rightarrow B \\ A\ true\end{array}}{B\ true}$$

II. $$\frac{\begin{array}{c}A \Rightarrow B \\ B\ false\end{array}}{A\ false}$$

Let $C \equiv \{A \Rightarrow B\}$. Then the two strong syllogisms correspond to the product rule in the form

$$P(B \mid AC) = \frac{P(AB \mid C)}{P(A \mid C)}, \qquad and \qquad P(A \mid \overline{B}C) = \frac{P(A\overline{B} \mid C)}{P(\overline{B} \mid C)}$$

From I. we have P(AB|C)=P(A|C), and from II. we have $P(A\overline{B} \mid C) = 0$.
Therefore P(B|AC)=1 and $P(A \mid \overline{B}C) = 0$.

However, the two rules of inference also contain more: **weak "syllogisms"** widely used in **inductive reasoning**, but whose existence were dismissed by Carl Popper and his followers.

i. $$\frac{\begin{array}{c}A \Rightarrow B \\ B\ true\end{array}}{A\ is\ more\ plausible}$$

and

ii. $$\frac{\begin{array}{c}A \Rightarrow B \\ A\ false\end{array}}{B\ is\ less\ plausible}$$

Let's see i.

From the product rule we have

$$P(A \mid BC) = P(A \mid C)\frac{P(B \mid AC)}{P(B \mid C)}$$

But from I. we know that P(B|AC)=1; and since P(B|C)≤ 1, it follows that

$$P(A \mid BC) \geq P(A \mid C)$$

i.e. A becomes more plausible.

Even weaker "syllogisms" can be derived from the rules of inference.

Probability is therefore not

physical science, not the study of frequencies as classical statistics would make you believe (though how it could possibly be even thought to do that based on such simple, **purely logical** axioms as the Kolmogorov axioms that forms the basis of mathematical probability theory is a mystery).

Probability is

an extension of deductive logic for cases when deductive inference is impossible due to incomplete knowledge, i.e. it is **the logic of science** (and quite possibly all reasoning).

When the background information consists of frequencies (as in our opening problem), extended logic (Bayesian analysis) automatically takes them into account, and by a built in updating feature allows the prediction of future expected frequencies from past ones. Classical inference has no procedures to do such elementary inductive reasoning. Strictly speaking, classical statistics can't even get off the ground.

Qualitative weak syllogisms and the quantitative connection it establishes between past and future make extended logic the **logic of induction**.

Hume's skeptical argument against induction and Popper's ideas of scientific reasoning were brilliantly refuted by the late **D.C.Stove** (see *The rationality of induction; Popper and after*). After finding the flaw in Hume's argument and unmasking Popper's objections he puts forward compelling arguments for, and identifies probability as the proper vehicle of induction. However, not being aware of Cox's theorem and extended logic, he makes a futile attempt to forge some frequentist ideas into some sort of logic. This unfortunately vitiates the second half of his book on induction which still awaits to be put aright.

A comprehensive treatise on the **historical, psychological, sociological and political** reasons such deeply, indeed fatally flawed paradigm as classical statistics has come to be enjoying almost universal acceptance also awaits to be written.

Miscellana

Confidence sets or intervals are concerned with coverage frequencies, i.e. confidence sets are constructed to provide a given frequency of coverage of an unknown constant, say, θ in repeated sampling. Only the false FDCS lends inferential value to such sets in the orthodox mind. Consider, however, the following

Example 4 [1]

Suppose it is known that a certain unknown parameter θ of an experiment is between 0 and 1. In order to gain some insight as to the value of θ one collects an observation X from a uniform distribution on the 0,1 interval in manner that is totally unconnected to the experiment and θ (e.g. produced by a favorite toy of statisticians: a "randomizer"). Now you would think that such an "observation", coming from a source so utterly unrelated to anything relevant to the experiment at hand, could produce no insight whatsoever as to the value of θ. (If you do think that, then you are not yet hopelessly infected by orthodoxy.) However, if coverage frequency you are after, X could easily give you that as follows.

Let B be any subset of the parameter space (known to be the (0,1) interval), and define

$$I_X = \begin{cases} B & \mathit{if} \;\; 0 < X \le .05 \\ (0,1) & \mathit{if} \;\; .05 < X < .95 \\ \overline{B} & \mathit{if} \;\; .95 \le X < 1 \end{cases}$$

The coverage probability of I_X is

$$\begin{aligned} &P(I_X \text{ covers } \theta) = \\ &\quad P(I_X \text{ covers } \theta \,|\, 0 < X \le .05) P(0 < X \le .05) \\ &+ P(I_X \text{ covers } \theta \,|\, .05 < X < .95) P(.05 < X < .95) \\ &+ P(I_X \text{ covers } \theta \,|\, .95 < X \le 1) P(.95 < X \le 1) \\ &\qquad\qquad = p \times .05 + 1 \times .9 + (1-p) \times .05 \\ &\qquad\qquad = .05[p + (1-p)] + .9 = .95 \end{aligned}$$

I_X is indeed a 95% confidence set that does the coverage trick required from such sets. And yet it is totally useless as a tool of inference.

The example above not only sheds light on the nature of confidence intervals and the difference between coverage probability and inferential value, it also raises several important questions: What is the role and what is the inferential value (if any) of **randomization**? And: What sort of property is **"randomness"**, and whose property is it, if indeed it is something real?

Lets revisit the standard finite survey example we had began this (series) of talk(s) with:

> Given a finite population of known size N a certain unknown number of which has a certain attribute, say, A. What can we say about this unknown number if we have observed a given segment (the sample) of the population?

For the orthodox mind the problem is ill-defined until the manner in which the sample was taken is specified (and when it is specified, it would proceed by solving an altered and irrelevant problem of averages). So let's imagine how a Socratic dialogue might develop between an orthodoxian (O) and a Bayesian (B).

B: Given the information above, the problem can be solved. In fact, what is required here is the solution of THIS problem, with the state of information given above, not some other problem where more or different information is given.

O: What state of information has got to do with it?

B: Well, let's see. Suppose you know that all population units are identical. In this case the link between sample and population is so strong that a single observation, however selected, would reveal the whole population. In our problem the state of information is different, and, as a result, the link between sample and population is much weaker. However, the link is still strong enough to produce an elegant and meaningful Bayesian solution using a Hypergeometric sampling distribution for the number of units with attribute A in the sample, and a uniform prior (on the integers between 0 and N) for the number of units with attribute A in the population (as dictated by the state of information given).

O: Aha! Now you are trying to pull a fast one. For the Hypergeometric distribution to be valid, the sample has to be random, i.e. it needs to be picked in a random fashion?

B: Though I would be curious to know if by valid you mean correspondence to some physical state, or valid description of a certain state of knowledge, I'd rather you elaborate on what you mean by "random fashion".

O: Textbook stuff. Such a way that all possible samples have the same probability to be selected. (And there are $\binom{N}{n}$ different samples.)

B: Hm. I seem to recall that your kind of texts call random every sample entitled to a probability distribution? Not just the ones obtained be "random fashion". But never mind. Let me ask you this. How would you accomplish what you just said, i.e. how would you ensure that in your selection of a single sample, all samples, even the ones not selected have the same probability to be the lucky one?

O: There are many ways. An obvious and widely used one is to give the pot holding the names or serial numbers of the units a good shake or stir.

B: And how much stir would you deem enough? Or rather, how much stir would be deemed enough by the statistical theory you subscribe to?

O: (thinking hard) Wee…ell, this does not really seem to be part of statistical theory….but I would say when one is reasonably sure that it is beyond human capacity to utilize any structure that might have existed in the pot and linked them together.

B: Hm. Might have existed and linked them together physically?, or might have existed and linked them together logically? For an interesting configuration might exist in the pot (physically, that is) you know nothing about. Does that matter? And alternatively, if you are in possession of such knowledge, wouldn't you be better off (i.e. come up with a more precise inference) by making use of that information rather than messing it up?

O: (hesitating) Weeeell….actually….now that you point this out….

B: And in the absence of such knowledge, what would be accomplished by the whole stirring and shaking hocus-pocus? In what way would it change your state of knowledge?

O: (hesitating) …Well, it wouldn't…

B: Moreover, though you may be successful in messing up some structure that might exist in the pot, would you not just create another structure by the whole process? A Turkish text is gibberish, i.e. "random" configuration of letters to me, but not for a Turk. Every pot has some "structure".

O: Yes, of course. Only I, or anyone else, would not know about it.

B: Bingo! Then you agree: What matters is not what goes on in the pot, but what you know, or suspected to know about it. The sampling distribution then is not the result or product of some physical stirring and shaking process nearly impossible to describe, but merely the description of a certain knowledge: equal share to all possible results in the absence of such knowledge (as in our case), or some other distribution in the presence of some knowledge. Now you may find that your knowledge, though far from ignorance, is so difficult to model that you rather opt for destroying it. Or others may suspect that you have some knowledge which you may abuse and demand that you destroy it. But that's about it.

O: (pondering) I'm not sure…

B: Now suppose the sample has already been selected. And though you are told it was done in a random fashion (whatever that means), you are suspicious. Could you, just by inspecting the sample, decide if that particular sample is indeed random? Could you reject it and say: No, this is not random for this and this reason.

O: (surprised by the question) I don't think so. Every sample had a chance, the same chance, to be the one. There are no telling signs.

B: Then the expression "random sample" is doubly meaningless since that mysterious property of "randomness" belongs to the process at best, not the sample. And we are uncertain, as we have just concluded, as to what that process would accomplish - statistically, that is. Physically of course it does something awfully complicated and difficult to describe.

O: Yeeees…so it seems…

B: And yet you ortodoxians use "randomness" as some magic property of a single sample. As if some properties (and pretty vague, mostly hoped for properties at that) of the process would, by some magic glue, stick to the selected individual transferring to it those properties that belonged to the collective before (and we don't really know what those properties are). As a result of that magic, "random" samples are suddenly entitled to own probabilities which they were not entitled before.

O: Well, if the pot is not stirred or shaken, I could be accused of cooking the sample.

B: That is a valid point as I myself have pointed out when I said that you may be suspected to have some knowledge which you may abuse, so you may be asked to destroy it. But what has this got to do with probability in general, and the Hypergeometric distribution in particular?

O: (with a sudden sneering smile) Let's bring in frequencies.

B: This seems to be one of those strange obsessions of yours. But, alright, let's bring them in.

O: In repeated trials an unstirred, or not sufficiently stirred, pot may not produce the long run frequencies we expect.

B: And what might those frequencies be?

O: The probabilities, of course.

B: So you agree: Probabilities and frequencies are different beasts. I am glad we cleared this up. But let's stay with frequencies. Would a sufficiently(?) stirred pot produce them as you would expect?

O: (the sneer fading) So I am told…

B: Is this an empirical observation? - because we are talking about a complex physical procedure the result of which depends on host of physical parameters such as the style, amount, etc. of stirring, and its frequency producing capacity (if any) would need to be reestablished on each occasion. To my knowledge this is no only not done routinely, but it has never been done. Or has someone actually taken the trouble and solved a nearly intractable dynamical problem, or rather a huge family of problems of stirring and shaking various pots in various ways (and what a waste of time would that be)? - in which case I would need the appropriate references. At any rate, we are already outside of the realm of statistics and in the realm physics.

O: (discouraged) Well…

B: But more to the point: Do those frequencies really matter? After all, most statistical experiments in general, and surveys in particular are unique events. At any rate, frequencies were not, and very seldom are part of the question.

O: (in a sudden outburst) No matter what you say, I still believe that a stirred pot has better chance to produce a representative sample than an unstirred one.

B: Though we have not even settled the question of how much stirring would be sufficient, I must point out that this is the third time you have shifted the debate. Initially randomization was supposed to be THE source of the sampling distribution; then it was supposed to safeguard your (or the survey's) integrity; then the emphasis shifted to the production of certain frequencies; and now it is the representativeness of the sample. So which one is it?

O: I'm confused…

And as confused though as he is, he leaves the scene to teach (or brainwash?) a class on the wonderful properties of random samples.

So which one is it? The role of randomization (in any form) is

a) to produce the (or a?) proper distribution? Nonsense. Probability (distributions) are not physical entities, products or byproducts of any process. Besides, that distribution is supposedly defined on all possible samples. In what way, by what magic glue would this collectively owned entity stick to a single sample selected thus? At any rate, how does one ensure and/or check that, due to the randomization process, the resulting sample is indeed random?, that probability is indeed produced?, and that what was produced have indeed been divided equally among the recipients?

b) to produce some frequencies in the long run? Nonsense. To begin with, it may or may not produce them - a nearly intractable question that belongs more to physics than statistics. And whether it does or does not produce them is utterly irrelevant due to the uniqueness of each survey (and our inevitable death in the not-so-long run). In any case, while probability assignments may change due to a change in background information, frequencies, being physical entities, are not influenced by such changes. They are what they are. Past frequencies, if known, form part of the background information and incorporated as such into the Bayesian inferential process. Future frequencies are simply unknown. We can merely speculate about their expected values which, incidentally, orthodoxy can't do (it merely equates probability with frequency) but poses no difficulty to Bayesian inference. But again, frequencies are simply irrelevant and usually get into the picture only as an orthodox contraband.

c) to ensure the representativeness of the sample? Nonsense. If anything, the opposite is the case. A perfect, or ideal representative sample, the one to approximate, is a miniature replica of the population with resp. to the objectives of the survey. The role of a sampling plan is to ensure that the sample is as representative (informative) as possible. That is indeed the holy grail of sampling surveys. Is it reasonable to believe that some randomizer, a deliberately dumb and blind mechanism guaranteed to be utterly unrelated to the problem at hand is a good way, let alone the best way to ensure that (remember Example 4)? If such were the case, why would we need statisticians (if indeed we do)? In fact, to create a sample as representative as possible, i.e. establish a connection between sample and population, one needs to rely on prior information. Randomization, far from being a good vehicle to establish such a connection, is an almost perfect vehicle to destroy any information and severe any connection that might exist between sample and population. A strange procedure indeed.

d) to protect the integrity of the survey(or)? Well, this, at last, is indeed a legitimate, though much abused role - the only one in my opinion. For the surveyor could be accused to have injected in the selection of the sample information he might have possessed or thought to have possessed, i.e. that he cooked the sample. The easiest (though not necessarily the wisest) way to protect the integrity (though not necessarily the quality) of the survey is to **destroy** this **information** in a compelling manner, i.e. by randomization. That is why such a process is deliberately dumb and blind, that is why we shuffle cards, and that is why teams in a sports tournament are paired "randomly" (though random matching makes it reasonably certain that not the best two teams will play in the final). So it all comes down to this: Randomization is willful destruction of information. Having this clarified, the next step is the realization that

"behind any randomized scheme there is a non-randomized one which is better but require more thought".

Not all samples are created equal

The belief that they are lies in the heart of orthodoxy. That this belief is untenable is well illustrated by Example 1. Even orthodoxians have realized that in such cases the inclusion of the whole sample space in the calculations would be absurd. It is sometimes argued that the example is artificial and the problem is easily remedied by breaking up the sample space into two subsets (as indicated by the precise and the imprecise instruments, whichever the case might be) not to be mixed by averaging. When its procedures fail, such ad hocery is characteristic of orthodoxy: instead of recognizing the fundamental flaws in its logic, it resorts to ad hoc devices to fix, or hope to fix problems (without realizing, of course, that the wrong problem is being addressed to begin with). However, it is easy to come up with less "artificial" examples as the difficulty easily arise naturally in simple problems.

Example 5

Suppose the results of an experiment is known to lie in a unit length interval with an unknown center θ, i.e. all measurements are known to lie between θ-1/2, and θ+1/2 (and that is all we know). From two observations x_1 and x_2, we need to infer to the value of θ.

Consider the two extreme configurations:

1) The distance between x_1 and x_2 is 1.

In such a case we would actually know that θ is just the average of the two observations. The inference is deductive, the resulting knowledge of θ is exact.

2) The distance between x_1 and x_2 is 0.

In such a case we would be as uncertain about value of θ is as we can be with only a single observation, say, x_1. All we could infer is that θ lies somewhere between x_1-1/2 and x_1+1/2.

There are an infinity of possibilities between these two extreme configurations resulting in various degrees of uncertainty about θ ranging from the smallest (Case 1) to the largest (Case 2). It is easy to see that mixing all these possibilities, so different in the quality of inference they allow, into a single average soup would be no less foolish and misleading than it was in Example 1. Any inferential process worthy of its name ought naturally to reflect these differences in information content of various samples.

It would also allow inductive inference to merge seamlessly into deductive inference as one moves from the most uninformative Case 2 to the fully informative Case 1 - as Bayesian inference indeed does all these.

That samples are not created equal had been known and keenly felt already by the dedicated frequentist Sir Ronald Fisher who, in his later years, devoted a large part of his remarkable genius to trying to fix this and other problems of frequentism. He can be credited with two of the most ingenious but, alas, failed attempts to "eat the Bayesian omelet without breaking the Bayesian egg": fiducial probability and ancillary statistics. The former was supposed to solve the problem of magic transference of probability not only from the collective to the individual but also from the sample space to the parameter space, and turn frequentist probabilities (the real thing in his view) into some sort of logical (possibly Bayesian?) ones.

The latter, ancillary statistics, addressed the question of the reference set, that subset of the sample space within which an average soup would be palatable. Fisher apparently believed that in many (most?, every?) problems the sample space of all possible samples can be resolved into mutually exclusive, inferentially incompatible subsets within each of which, however, a sort of inferential unity prevails in the sense, that within each subset all samples would be equal in their information content and precision of inference they would allow. According to the **conditionality principle** (one of dozens of such ad hoc, sometimes contradictory, and often violated frequentist devices to keep it on track), it would be improper to use the whole sample space as reference set. The proper thing to do is to restrict the sample space to (condition on) an inferentially more homogenous subset as indicated by the so called ancillary statistic of the sample, a statistic whose parameter-free distribution is fully known in advance (in Example 5, $|x_1-x_2|$ would be such a statistic; in Example 1, the result of the coin toss y would do the job).

A beautiful idea - if only it had worked. But it did not. Far from being universally applicable, ancillary statistics exist only in a narrow family of relatively simple problems. And then they may not be unique. And when they do exist and unique, the distribution restricted to the proper reference set by the ancillary statistic agrees (formally that is) with the Bayesian result (with an uninformative prior) which can be obtained painlessly by the two simple inferential rules without resorting to ad hoc principles. Yet again: When it works, it is superfluous.

Independence revisited

It is worth revisiting the question of independence in connection of Example 5: Are the two observations x_1 and x_2 independent? In my experience the problem is instantly metamorphosed into a question of irrelevant physical independence in the traditionally trained mind: If the observations were obtained in a physically unconnected manner, then the answer is yes, otherwise it is no. And yet, the observations could be separated by years and miles and still not be independent. What is wrong?

Suppose the first observation x_1 is about to be observed. Since, according to the specifications, nothing is known about θ, and since x_1 must be within half unit of θ, all we know is that x_1 could end up anywhere on the real line. Having x_1 observed, what can be said about the second observation x_2 ? Can that also be anywhere? Not by a long shot. Since x_1 must be within half unit of θ, θ must be within half unit of x_1, i.e. by jut one observation our knowledge has already increased infinite fold. We can't expect a similar improvement in our knowledge by the next observation. Indeed, as a little reflection would show, x_2 must be within one unit of x_1, only a minor improvement compared to the improvement on our total cluelessness as to the value of x_1; in comparison the location of the next observation x_2 can be found next to x_1 with pinpoint accuracy. In short, the correlation between x_1 and x_2 is probably very strong. Not only x_1 and x_2 are not independent, but their correlation is as strong as possible as it is indeed 1. Were the whereabouts of θ be given more precisely, the corresponding correlation between the first two observations would be lower and approach 0 as exact knowledge of θ is approached. Only when θ is known exactly would the observations be independent. In a similar manner, as the observations accumulate, new observations would add only a diminishing amount to our knowledge already accumulated, and the correlation between every two new observations given the past would weaken and slowly approach 0. (These qualitative arguments can, of course, be confirmed by easy Bayesian calculations.)

It seems that the strength of dependence between two observations is not only not a question of physical connections but of logical ones, but, as it hinges on what one knows, it is not even an absolute measure.

How could orthodoxy have gone so wrong with such fundamentally important concept discussed at the beginning in every introductory text? The answer lies in orthodoxy's mistaken physical conception of probability and virtually everything related to it, and its consequent denying of a probabilistic status from anything that is not "random". How this have come to pass is somewhat of a mystery since the Kolmogorov axioms, the point of departure in virtually every text on probability, are clearly logical in nature; and though they fail to specify the nature of probability (unlike the desiderata that imply Cox's theorem), they could not possibly have served as the foundation of any physical or natural science.

Be as it may, following the textbook definition of independence, x_1 is independent from x_2 if $p(x_2|x_1)=p(x_2)$. As it was demonstrated above, this is manifestly not the case here: while $p(x_2)= p(x_1)$ is a distribution on the whole real line, $p(x_2|x_1)$ is concentrated on the interval (x_1-1, x_1+1). For an orthodoxian, however, $p(x_1)$ is a distribution concentrated on the interval $(\theta -1/2, \theta +1/2)$, and so is $p(x_2|x_1)$. It is a declaration, usually based on the belief of physical disconnectedness of the observations, that x_1 does not change (physically?) the distribution of x_2 (to be precise, it is in fact the *knowledge* of x_1 that presumably does or does not change the distribution of x_2). But, as it has just been demonstrated, this is manifestly false. It would only be true if the value of θ were known. But it isn't. What is going on?

Well, orthodoxy considers θ an unknown constant. Nothing "random" about it. It is thought to be "out there" in the physical world to be measured, however imperfectly, like one measures physical variables such as mass or heat. This is not only irrelevant and possibly misleading (e.g. the unknown fraction of balls with attribute A in the urn does have physical status and fully knowable in principle, but the "probability of heads" in a coin tossing experiment is merely a mathematical artifact that is not knowable even in principle), but it confers on θ a rather ambivalent status: on the one hand, it is clearly unknown (to be estimated, tested, etc.); on the other hand, we pretend as if it were fully known, otherwise independence, for example, could be declared. And it is all or nothing. θ is either fully known, or completely unknown - no grades of knowledge allowed. This ambivalence of the extremes is never resolved; whichever aspect is needed is used opportunistically. Is there any way to resolve the ambivalence? Not if the logical nature of probability is rejected and a probability distribution is denied from θ on the basis of irrelevant physicalistic arguments. Only if θ is allowed to take its proper place after the vertical line can the ambivalence be resolved. Only then can differences in background information I, and the degree of dependence between the observations as a function of that information be quantified that would confirm the heuristic arguments presented above. But then we would venture on forbidden Bayesian territory, wouldn't we. Well, not so forbidden, as we shall see presently.

The stock response from orthodoxians after hearing arguments like the above is a shrug: "So what? The observations are *conditionally* independent. What we meant was that $p(x_2|x_1,\theta)= p(x_2|\theta)$. No big deal."

Such hypocritical opportunism, such readiness to take nourishment from Bayesian ideas without, of course, taking the consequences, is characteristic of orthodox thinking. Orthodoxy always hopes to get away with it - and mostly, and somewhat mysteriously, it does. It is seldom taken to task for using the profoundly false FDCS as its fundamental *modus operandi* because it has never been openly declared and few of its clients see through the fog of technicalities (physicists, due to their technical prowess, are notable exceptions - and they are instinctively Bayesians). And it is seldom taken to task for meaning conditional independence when saying independence because few are aware of the logical nature of the concept, and because the conceptual confusion is thought not to cause any trouble. But it does.

For, as we have seen, what one knows matters. In the presence of even the minimal uncertainty as to the value of the parameter the observations are no longer independent - conditionally or unconditionally. In testing composite hypotheses (such as H_a in Example 2) the very composite nature of the hypothesis is an indication of uncertainty as to value of the parameter. Upon such hypotheses, the data are no longer independent, and the conditional nature of the concept can no longer be swept under the carpet. Since testing composite hypotheses is the daily bread and butter of statistics (even though the classical hypothesis test does a botch job of it), it is interesting to see how orthodoxy saves its skin in this case. Even a cursory browsing through chapters devoted to such tests in standard texts would reveal that, despite the promising heading, the delivered product is not what has been promised (isn't it always the case with orthodoxy?): there is not a single composite hypothesis tested in those chapters. Instead what one finds there is a host of ways to reduce composite hypotheses to simple ones (monotone likelihoods, maximized likelihood ratios, etc. are the main vehicles to do just that).

Summing up

Classical statistics is a hopelessly flawed inferential tool whose operation is a based on bait and switch scams (FDCS, etc.). It is riddled with irrelevant, incoherent, and incorrect methods and concepts patched up by a host of contradictory ad hoc principles and criteria violated as often as adhered to. Its creaking edifice is held up not by its merits but by clients unaware of the bait and switch scam; clients who don't care because all they want is admission to the temple of science from its high priest, the statistician; and by sheer intellectual inertia maintained by vested interests in the status quo and that main source of orthodox indoctrination, the Intro.Stats. course offered to countless thousands of unsuspecting students every year.

Selected Bibliography

[1] Basu, D. (1988), *Statistical Information and Likelihood : A Collection of Critical Essays*, edited by Gosh, J.K., Springer Verlag, New York

This is a must read from the grand master of prickly examples that puncture the balloon of orthodoxy and demonstrate his dictum that nothing in classical statistics makes sense unless it has a Bayesian interpretation. Just listen to the music of the escaping air as sample surveys, randomization schemes, ancillary statistics, etc. all get their just deserts.

[2] Berger, J.O., Wolpert, R.L. (1988), *The Likelihood Principle,* 2nd ed., Inst. Math. Stat.

A thorough survey of the likelihood principle. According to this principle, identical likelihood functions must result in identical inferences. There are weaker and stronger versions of it. Bayesians don't need it since it is built into the Bayes formula; orthodoxians may or may not subscribe to one or another version (and they may not even though they may subscribe to other principles that imply it). Whichever the case, they often violate it (e.g. hypothesis test in Example 2).

[3] Edwards, A.W.F. (1972) *Likelihood*, Cambridge University Press

A survey of the neither-here-nor-there halfway house of likelihood methods which try to get by with just one, the supposedly objective ingredient in the Bayes formula. It is worth consulting to see how much blood one must sweat on the altar of the elusive god of objectivity for simple results produced with ease by Bayesian logic.

[4] Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge University Press

A classic from a distinguished philosopher. Should be mandatory reading for all statisticians. Discusses many of the orthodox fallacies emphasizing the fundamentally different nature of pre- and post-data arguments. Unwilling to go the Bayesian route, makes a heroic attempt to revive the dead and fix fiducial probability.

[5] Fisher, R.A. (1956), *Statistical Methods and Scientific Inference*, 2nd ed., Oliver&Boyd, London

Very deep, if cryptic at times, final testament from the greatest genius frequentist thinking has ever had showing how far his ideas were from rigid Neyman-style orthodoxy. As he came perilously close to Bayesian thinking, had it not been for his fixation on frequencies topped with a cathedral size ego, statistical inference would be taught and practiced very differently today.

[6] Jaynes, E.T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press

The most important work on inference since Jeffreys' [8] by the distinguished physicist(!). The breadth and depth of this towering masterpiece of style and substance is astonishing. Unfortunately it has remained unfinished due to the untimely death of its author. Larry Bretthorst, Jaynes' student, did the Gargantuan work of finishing it using Jaynes' notes and instructions. In a perfect world it would completely change the way statistical inference is taught and practiced. Don't hold your breath.

[7] Jaynes, E.T. (1983), *Papers on Probability, Statistics, and Statistical Physics*, edited by R.D.Rosenkrantz, D.Reidel Publ.Co.

A collection of important papers mapping the development of Jaynes' ideas on maximum entropy and inference. The paper Confidence Intervals vs. Bayesian Intervals is particularly recommended.

[8] Jeffryes, H. (1939), *Theory of Probability*, Clarendon Press, Oxford (latest edition 1988)

With Laplace's [9], the most important pre-Jaynes work on (Bayesian) inference. Probably would have defined the course of statistical inference had the field not been ruled by the strong personalities of Fisher and Neyman. Unlike his arguments, Jeffreys reserved, self-effacing style was no match for the overbearing Fisher's whose antics he watched with quiet amusement. So much for the triumph of truth in the sciences.

[9] Laplace, Pierre Simon (1812), *Théorie analytique des probabilités*, 2 vol., Courcier Imprimeur Paris, Third edition, Paris

A true classic in every sense of the word. Beware of misrepresentations by some nineteen century intellectual pigmies trying to cast long shadows.

[10] Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York

The Bible of the Neyman school of orthodoxy. Full of highly technical, mathematically correct but utterly irrelevant results (orthodoxy always mistakes mathematics for relevance) which established the Annals of Statistics style (and irrelevance). Generations of brainwashed graduate students make it arguably the single most damaging book to the cause of sound inference.

[11] Stove, D.C. (1986), *The Rationality of Induction*, Clarendon Press, Oxford

The master of the emperor-has-no-clothes school of philosophy takes apart and finds the fatal flaw in Hume's famous skeptical argument against induction. Shows, correctly, induction to be one of the main tools of science (and life) and proposes probability as its main vehicle. Not being aware of Cox's or Jaynes' work on inference, futile frequentism vitiates somewhat the last part of this brilliant work.

[12] Stove, D.C. (1982), *Popper and After: Four Modern Irrationalists*, Pergamon Press, New York

Contains a shorter version of the refutation of Hume's skeptical argument. Its chief merit, however, lies in the entertaining demolition of such anti-inductionist and/or anything goes relativist idols as Popper, Kuhn, Feyerabend, and Lakatos. (All of Stoves writings are profound and entertaining. His take on nineteen century German and English idealism in *The Plato cult and other philosophical follies* is priceless.)