5.8 Neyman's Performance and Fisher's Fiducial Probability

Many say fiducial probability was Fisher's biggest blunder; others suggest it still hasn't been understood. Most discussions avoid a side trip to the Fiducial Islands altogether, finding the surrounding brambles too thorny to negotiate. I now think this is a mistake, and it is a mistake at the heart of the consensus interpretation of the N-P vs. Fisher debate. We don't need to solve the problems of fiducial inference, fortunately, to avoid taking the words of the Fisher–Neyman dispute out of context. Although the Fiducial Islands are fraught with minefields, new bridges are being built connecting some of the islands to Power Peninsula and the general statistical mainland.

So what is fiducial inference? I begin with Cox's contemporary treatment, distilled from much controversy. The following passages swap his upper limit for the lower limit to keep to the example Fisher uses:

We take the simplest example, . . . the normal mean when the variance is known, but the considerations are fairly general. The lower limit

$$\overline{x} - z_c\sigma/\sqrt{n}$$

derived here from the probability statement

$$\Pr(\mu > \overline{X} - z_c\sigma/\sqrt{n}) = 1 - c$$

is a particular instance of a *hypothetical* long run of statements a proportion $1 - c$ of which will be true, . . . assuming our model is sound. We can, at least in principle, make such a statement for each c and thereby generate a collection of statements, sometimes called a *confidence distribution*. (Cox 2006a, p. 66; $\overline{x}$ for $\overline{y}$, $\overline{X}$ for $\overline{Y}$, and z_c for k_c^*)

Once $\overline{x}$ is observed, $\overline{x} - z_c\sigma/\sqrt{n}$ is what Fisher calls the *fiducial c percent limit* for μ. It is, of course, the *specific* 1–c lower confidence interval estimate $\hat{\mu}_{1-c}(\overline{x})$ (Section 3.7).

Here's Fisher in the earliest paper on fiducial inference in 1930. He sets $1 - c$ as 0.95. Starting from the significance test of a specific μ, he identifies the corresponding *95 percent value* $\overline{x}_{.05}$, such that in 95% of samples $\overline{X} < \overline{x}_{.05}$. In the normal testing example, $\overline{x}_{.05} = \mu + 1.65\sigma/\sqrt{n}$. Notice $\overline{x}_{.05}$ is the cut-off for a 0.05 one-sided test T+ (of $\mu \leq \mu_0$ vs. $\mu > \mu_0$).

[W]e have a relationship between the statistic [$\overline{X}$] and the parameter μ, such that [$\overline{x}_{.05}$] is the 95 per cent. value corresponding to a given μ, and this relationship implies the perfectly objective fact that in 5 per cent. of samples [$\overline{X} > \overline{x}_{.05}$. That is, $Pr(\overline{X} \leq \mu + 1.65\sigma/\sqrt{n}) = 0.95$] (Fisher 1930, p. 533; substituting μ for θ and $\overline{X}$ for T.)

$\overline{X} > \overline{x}_{.05}$ occurs whenever $\mu < \overline{X} - 1.65\sigma/\sqrt{n}$ the *generic* $\hat{\mu}_{.95}(\overline{X})$. For a particular observed $\overline{x}$, $\overline{x} - 1.65\sigma/\sqrt{n}$ is the "fiducial 5 per cent. value of μ."

We may know as soon as $\overline{X}$ is calculated what is the fiducial 5 per cent. value of μ, and that the true value of μ will be less than this value in just 5 per cent. of trials. This then is a definite probability statement about the unknown parameter μ which is true irrespective of any assumption as to its *a priori* distribution. (ibid.)[3]

This seductively suggests $\mu < \hat{\mu}_{.95}(\overline{x})$ gets the probability 0.05 – a fallacious probabilistic instantiation.

However, there's a kosher probabilistic statement about $\overline{X}$, it's just not a probabilistic assignment to a parameter. Instead, a particular substitution is, to paraphrase Cox, "a particular instance of a hypothetical long run of statements 95% of which will be true." After all, Fisher was abundantly clear that the fiducial bound should not be regarded as an inverse inference to a posterior probability. We could only obtain an inverse inference by considering μ to have been selected from a superpopulation of μ's, with known distribution. The posterior probability would then be a deductive inference and not properly inductive. In that case, says Fisher, we're not doing inverse or Bayesian inference.

> In reality the statements with which we are concerned differ materially in logical content from inverse probability statements, and it is to distinguish them from these that we speak of the distribution derived as a *fiducial* frequency distribution, and of the working limits, at any required level of significance, . . . as the *fiducial limits* at this level. (Fisher 1936, p. 253)

So, what is being assigned the fiducial probability? It's the method of reaching claims to which the probability attaches. This is even clearer in his 1936 discussion where σ is unknown and must be estimated. Because $\overline{X}$ and S (using the Student's t pivot) are sufficient statistics "we may infer, without any use of probabilities *a priori*, a frequency distribution for μ which shall correspond with the aggregate of all such statements . . . to the effect that the probability μ is less than $\overline{x} - 2.145s/\sqrt{n}$ is exactly one in forty" (ibid., p. 253). This uses Student's t distribution with $n = 15$. It's plausible, at that point, to suppose Fisher means for $\overline{x}$ to be a random variable.

Suppose you're Neyman and Pearson working in the early 1930s aiming to clarify and justify Fisher's methods. 'I see what's going on,' we can imagine Neyman declaring. There's a method for outputting statements such as would take the general form

$$\mu > \overline{X} - 2.145\, s/\sqrt{n}.$$

Some would be in error, others not. The method outputs statements with a probability (some might say a propensity) of 0.975 of being correct. "We may

[3] It's correct that $(\mu \leq \overline{X} - z_c\, \sigma/\sqrt{n})$ iff $(\overline{X} > \mu + z_c\sigma/\sqrt{n})$.

look at the purpose of tests from another viewpoint": probability ensures us of the performance of a method (it's methodological).

At the time, Neyman thought his development of confidence intervals (in 1930) was essentially the same as Fisher's fiducial intervals. There was evidence for this. Recall the historical side trip of Section 3.7. When Neyman gave a (1934) paper to the Royal Statistical Society discussing confidence intervals, seeking to generalize fiducial limits, he made it clear that the term confidence coefficient refers to "probability of our being right when applying a certain rule" for making statements set out in advance (p. 140). Much to Egon Pearson's relief, Fisher called Neyman's generalization "a wide and very handsome one," even though it didn't achieve the uniqueness Fisher had wanted (Fisher 1934c, p. 137). There was even a bit of a mutual admiration society, with Fisher saying "Dr Neyman did him too much honour" in crediting him for the revolutionary insight of Student's t pivotal, giving the credit to Student. Neyman (1934, p. 141) responds that of course in calling it Student's t he is crediting Student, but "this does not prevent me from recognizing and appreciating the work of Professor Fisher concerning the same distribution."

In terms of our famous passage, we may extract this reading: In struggling to extricate Fisher's fiducial limits, without slipping into fallacy, they are led to the N-P construal. Since fiducial probability was to apply to significance testing as well as estimation, it stands to reason that the performance notion would find its way into the N-P 1933 paper.[4] So the error probability applies to the method, but the question is whether it's intended to qualify a given inference, or only to express future long-run assurance (performance).

N-P and Fisher Dovetail: It's Interpretation, not Mathematics

David Cox shows that the Neyman–Pearson theory of tests and confidence intervals arrive at the same place as the Fisherian, even though in a sense they proceed in the opposite direction. Suppose that there is a full model covering both null and alternative possibilities. To establish a significance test, we need to have an appropriate test statistic d($\boldsymbol{X}$) such that the larger the d($\boldsymbol{X}$) the greater the discrepancy with the null hypothesis in the respect of interest. But it is also required that the probability distribution of d($\boldsymbol{X}$) be known under the assumption of the null hypothesis. In focusing on the logic, we've mostly

[4] "[C]onsider that variation of the unknown parameter, μ, generates a continuum of hypotheses each of which might be regarded as a null hypothesis . . . [T]he data of the experiment, and the test of significance based upon them, have divided the continuum into two portions." One a region in which μ lies between the fixed fiducial limits, "is accepted by the test of significance, in the sense that values of μ within this region are not contradicted by the data at the level of significance chosen. The remainder… is rejected" (Fisher 1935a, p. 192).

considered just one unknown parameter, e.g., the mean of a Normal distribution. In most realistic cases there are additional parameters required to compute the *P*-value, sometimes called "nuisance" parameters λ, although they are just as legitimate as the parameter we happen to be interested in. We'd like to free the computation of the *P*-value from these other unknown parameters. This is the error statistician's way to ensure as far as possible that observed discordances may be blamed on discrepancies between the null and what's actually bringing about the data. We want to solve the classic Duhemian problems of falsification.

As Cox puts it, we want a test statistic with a distribution that is split off from the unknown nuisance parameters, which we can abbreviate as λ. The full parameter space Θ is partitioned into components $\Theta = (\psi, \lambda)$, such that the null hypothesis is that $\psi = \psi_0$, with λ an unknown nuisance parameter. Interest may focus on alternatives $\psi > \psi_0$. We do have information in the data about the unknown parameters, and the natural move is to estimate them using the data. The twin goals of computing the *P*-value, $\Pr(d > d_0; H_0)$, free of unknowns, and constructing tests that are appropriately sensitive, produce the same tests entailed by N-P theory, namely replacing the nuisance parameter by a sufficient statistic V. A statistic V, a sufficient statistic for nuisance parameter λ, means that the probability of the $d(X)$ conditional on the estimate V depends only on the parameter of interest ψ_0. So we are back to the simple situation with a null having just a single parameter ψ. This "largely determines the appropriate test statistic by the requirement of producing the most sensitive test possible with the data at hand" (Cox and Mayo 2010, p. 292). Cox calls this "conditioning for separation from nuisance parameters" (ibid.). I draw from Cox and Mayo (2010).

In the most familiar class of cases, this strategy for constructing appropriately sensitive or powerful tests, separate from nuisance parameters, produces the same tests as N-P theory. In fact, when statistic V is a special kind of sufficient statistic for nuisance parameter λ (called *complete*), there is no other way of achieving the N-P goal of an exactly α-level test that is fixed regardless of nuisance parameters – these are called *similar* tests.[5] Thus, replacing the nuisance parameter with a sufficient statistic "may be regarded as an outgrowth of the aim of calculating the relevant *P*-value independent of unknowns, or alternatively, as a byproduct of seeking to obtain most powerful

[5] The goal of exactly similar tests leads to tests that ensure

$$\Pr(d(\boldsymbol{X}) \text{ is significant at level } \alpha | v; H_0) = \alpha,$$

where v is the value of the statistic V used to estimate the nuisance parameter. A good summary may be found in Lehmann (1981).

similar tests." These dual ways of generating tests reveal the underpinnings of a substantial part of standard, elementary statistical methods, including key problems about Binomial, Poisson, and Normal distributions, the method of least squares, and linear models.[6] (ibid., p. 293)

If you begin from the "three steps" in test generation described by E. Pearson in the opening to Section 3.2, rather than the later N-P–Wald approach, they're already starting from the same point. The only difference is in making the alternative explicit. Fisher (1934b) made the connection to the N-P (1933) result on uniformly most powerful tests:

> . . . where a sufficient statistic exists, the likelihood, apart from a factor independent of the parameter to be estimated, is a function only of the parameter and the sufficient statistic, explains the principal result obtained by Neyman and Pearson in discussing the efficacy of tests of significance. Neyman and Pearson introduce the notion that any chosen test of a hypothesis H_0 is more powerful than any other equivalent test, with regard to an alternative hypothesis H_1, when it rejects H_0 in a set of samples having an assigned aggregate frequency ε when H_0 is true, and the greatest possible aggregate frequency when H_1 is true. . . (pp. 294–5)

> It is inevitable, therefore, that if such a statistic exists it should uniquely define the contours best suited to discriminate among hypotheses differing only in respect of this parameter; . . . When tests are considered only in relation to sets of hypotheses specified by one or more variable parameters, the efficacy of the tests can be treated directly as the problem of estimation of these parameters. Regard for what has been established in that theory, apart from the light it throws on the results already obtained by their own interesting line of approach, should also aid in treating the difficulties inherent in cases in which no sufficient statistics exists. (ibid., p. 296)

This article may be seen to mark the point after which Fisher's attitude changes because of the dust-up with Neyman.

Neyman and Pearson come to Fisher's Rescue

Neyman and Pearson entered the fray on Fisher's side as against the old guard (led by K. Pearson) regarding the key point of contention: showing statistical inference is possible without the sin of "inverse inference". Fisher denounced the *principle* of *indifference*: "We do not know the function . . . specifying the super-population, but in view of our ignorance of the actual values of θ we may" take it that all values are equally probable (Fisher 1930, p. 531). "[B]ut

[6] Requiring exactly similar rejection regions, "precludes tests that merely satisfy the weaker requirement of being able to calculate P approximately, with only minimal dependence on nuisance parameters," which could be preferable especially when best tests are absent. (Ibid.)

however we might disguise it, the choice of this particular a priori distribution for the θ is just as arbitrary as any other. . .” (ibid.).

If, then, we follow writers like Boole, Venn, . . . in rejecting the inverse argument as devoid of foundation and incapable even of consistent application, how are we to avoid the staggering falsity of saying that however extensive our knowledge of the values of x . . . we know nothing and can know nothing about the values of θ? (ibid.)

When Fisher gave his paper in December 1934 (“The Logic of Inductive Inference”), the old guard were ready with talons drawn to attack his ideas, which challenged the overall philosophy of statistics they embraced. The opening thanks (by Arthur Bowley), which is typically a flowery, flattering affair, was couched in scathing, sarcastic terms (see Fisher 1935b, pp. 55–7). To Fisher’s support came Egon Pearson and Jerzy Neyman. Neyman dismissed “Bowley’s reaction to Fisher’s critical review of the traditional view of statistics as an understandable attachment to old ideas (1935, p. 73)” (Spanos 2008b, p. 16). Fisher agreed: “However true it may be that Professor Bowley is left very much where he was, the quotations show at least that Dr. Neyman and myself have not been left in his company” (1935a, p. 77).

So What Happened in 1935?

A pivotal event was a paper Neyman gave in which he suggested a different way of analyzing one of Fisher’s experimental designs. Then there was a meet-up in the hallway a few months later. Fisher stops by Neyman’s office at University College, on his way to a meeting which was to decide on Neyman’s reappointment in 1935:

And he said to me that he and I are in the same building . . . That, as I know, he has published a book – and that’s *Statistical Methods for Research Workers* – and he is upstairs from me so he knows something about my lectures – that from time to time I mention his ideas, this and that – and that this would be quite appropriate if I were not here in the College but, say, in California … but if I am going to be at University College, then this is not acceptable to him. And then I said, ‘Do you mean that if I am here, I should just lecture using your book?’ And then he gave an affirmative answer. . . . And I said, ‘Sorry, no. I cannot promise that.’ And then he said, ‘Well, if so, then from now on I shall oppose you in all my capacities.’ And then he enumerated – member of the Royal Society and so forth. There were quite a few. Then he left. Banged the door. (Neyman in C. Reid 1998, p. 126)

Imagine if Neyman had replied: ‘I’d be very pleased to use *Statistical Methods for Research Workers* in my class.’ Or what if Fisher had said: ‘Of course you’ll want to use your own notes in your class, but I hope you will use a portion of my text when mentioning some of its key ideas.’ Never mind. That was it. Fisher went on

to a meeting wherein he attempted to get others to refuse Neyman a permanent position, but was unsuccessful. It wasn't just Fisher who seemed to need some anger management training, by the way. Erich Lehmann (in conversation and in 2011) points to a number of incidences wherein Neyman is the instigator of gratuitous ill-will. I find it hard to believe, however, that Fisher would have thrown Neyman's wooden models onto the floor.

> One evening, late that spring, Neyman and Pearson returned to their department after dinner to do some work. Entering they were startled to find strewn on the floor the wooden models which Neyman had used to illustrate his talk . . . Both Neyman and Pearson always believed that the models were removed by Fisher in a fit of anger (C. Reid 124, noted in Lehmann 2011, p. 59).

Neyman left soon after to start the program at Berkeley (1939), and Fisher didn't remain long either, moving in 1943 to Cambridge and retiring in 1957 to Adelaide. I've already been disabusing you of the origins of the popular Fisher–N-P conflict (Souvenir L). In fact, it really only made an appearance long after the 1933 paper!

1955–6 Triad: Telling What's True About the Fisher–Neyman Conflict

If you want to get an idea of what transpired in the ensuing years, look at Fisher's charges and Neyman's and Pearson's responses 20 years later. This forms our triad: Fisher (1955), Pearson (1955), and Neyman (1956). Even at the height of mudslinging, Fisher said, "There is no difference to matter in the field of mathematical analysis . . . but in logical point of view" (1955, p. 70).

> I owe to Professor Barnard . . . the penetrating observation that this difference in point of view originated when Neyman, thinking he was correcting and improving my own early work on tests of significance as a means to the 'improvement of natural knowledge,' in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure. . . . Russians are made familiar with the ideal that research in pure science can and should be geared to technological performance. (ibid., pp. 69–70)

Pearson's (1955) response: "To dispel the picture of the Russian technological bogey, I might recall how certain early ideas came into my head as I sat on a gate overlooking an experimental blackcurrant plot . . .!" (Pearson 1955, p. 204). He was "smitten" by an absence of logical justification for some of Fisher's tests, and turned to Neyman to help him solve the problem. This takes us to where we began with our miserable passages, leading them to pin down the required character for the test statistic, the need for the alternative and power considerations.

Until you disinter the underlying source of the problem – fiducial inference – the "he said/he said" appears to be all about something that it's not. The reason Neyman adopts a performance formulation, Fisher (1955) charges, is that he denies the soundness of fiducial inference. Fisher thinks Neyman is wrong because he "seems to claim that the statement (a) 'μ has a probability of 5 per cent. of exceeding $\overline{X}$' is a different statement from (b) '$\overline{X}$ has a probability of 5 per cent. of falling short of μ'" (p. 74, replacing θ and T with μ and $\overline{X}$). There's no problem about equating these two so long as $\overline{X}$ is a random variable. But watch what happens in the next sentence. According to Fisher, Neyman violates

> . . . the principles of deductive logic [by accepting a] general symbolical statement such as
>
> [1] $\Pr\{(\overline{x} - ts) < \mu < (\overline{x} - ts)\} = \alpha$,
>
> as rigorously demonstrated, and yet, when numerical values are available for the statistics $\overline{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read
>
> [2] Pr {92.99 < μ < 93.01} = 95 per cent.,
>
> to *deny* to this *numerical* statement any validity. This evidently is to deny the syllogistic process. (Fisher 1955, p. 75, in Neyman 1956, p. 291)

But the move from (1) to (2) is fallacious! Is Fisher committing this fallacious probabilistic instantiation (and still defending it in 1955)? I. J. Good describes how many felt, and still feel:

> It seems almost inconceivable that Fisher should have made the error which he did in fact make. [That is why] . . . so many people assumed for so long that the argument was correct. They lacked the *daring* to question it. (Good 1971a, p. 138).

Neyman (1956) declares himself at his wit's end in trying to convince Fisher of the inconsistencies in moving from (1) to (2). "Thus if $\boldsymbol{X}$ is a normal random variable with mean zero and an arbitrary variance greater than zero, then I expect" we may agree that $\Pr(\boldsymbol{X} < 0) = 0.5$ (ibid., p. 292). But observing, say, $\boldsymbol{X} = 1.7$ yields $\Pr(1.7 < 0) = 0.5$, which is clearly illicit. "It is doubtful whether the chaos and confusion now reigning in the field of fiducial argument were ever equaled in any other doctrine. The source of this confusion is the lack of realization that equation (1) does not imply (2)" (ibid., p. 293). It took the more complex example of Bartlett to demonstrate the problem: "Bartlett's revelation [1936, 1939] that the frequencies in repeated sampling [from the same or different populations] need not agree with Fisher's solution" in the case of a difference between two Normal means with different variances, "brought

about an avalanche of rebuttals by Fisher and by Yates" (ibid., p. 292).[7] Some think it was only the collapse of Fisher's rebuttals that led Fisher to castigate N-P for assuming error probabilities and fiducial probabilities *ought* to agree, and begin to declare the idea "foreign to the development of tests of significance." As statistician Sandy Zabell (1992 p. 378) remarks, "such a statement is curiously inconsistent with Fisher's own earlier work" as in Fisher's (1934b) endorsement of UMP tests, and his initial attitude toward Neyman's confidence intervals. Because of Fisher's stubbornness "he engaged in a futile and unproductive battle with Neyman which had a largely destructive effect on the statistical profession" (ibid., p. 382).[8]

Fisher (1955) is spot on about one thing: When "Neyman denies the existence of inductive reasoning, he is merely expressing a verbal preference. For him 'reasoning' means what 'deductive reasoning' means to others" (p. 74). Nothing earth-shaking turns on the choice to dub every inference "an act of making an inference." Neyman, much like Popper, had a good reason for drawing a bright red line between the use of probability (for corroboration or probativeness) and the probabilists' use of confirmation: Fisher was blurring them.

> . . . the early term I introduced to designate the process of adjusting our actions to observations is 'inductive behavior'. It was meant to contrast with the term 'inductive reasoning' which R. A. Fisher used in connection with his 'new measure of confidence or diffidence' represented by the likelihood function and with 'fiducial argument'. Both these concepts or principles are foreign to me. (Neyman 1977, p. 100)

The Fisher–Neyman dispute is pathological: there's no disinterring the truth of the matter. Perhaps Fisher altered his position out of professional antagonisms toward the new optimality revolution. Fisher's stubbornness on fiducial intervals seems to lead Neyman to amplify the performance construal louder and louder; whereas Fisher grew to renounce performance goals he himself had held when it was found that fiducial solutions disagreed with them. Perhaps inability to identify conditions wherein the error probabilities "rubbed off" – where there are no "recognizable subsets" with a different probability of success – led Fisher to move to a type of default Bayesian stance. That Neyman (with the contributions of Wald, and later Robbins) might have gone overboard in his behaviorism, to the extent that even Egon wanted to divorce him – ending his 1955 reply to Fisher with the claim that "inductive behavior" was

[7] In that case, "the test rejects a smaller proportion of such repeated samples than the proportion specified by the level of significance" (Fisher 1939, p. 173a). Prakash Gorroochurn (2016) has a masterful historical discussion.

[8] Buehler and Feddersen (1963) showed there were recognizable subsets even for the t test.

Neyman's field, not his – is a different matter. Ironically, Pearson shared Neyman's antipathy to "inferential theory" as Neyman (1962) defines it in the following:

> In the present paper … the term 'inferential theory' … will be used to describe the attempts to solve the Bayes' problem with a reference to confidence, beliefs, etc., through some supplementation … either a substitute *a priori* distribution [exemplified by the so called principle of insufficient reason] or a new measure of uncertainty [such as Fisher's fiducial probability] (p. 16).

Fisher may have started out seeing fiducial probability as both a frequency of correct claims in an aggregate, and a rational degree of belief (1930, p. 532), but the difficulties in satisfying uniqueness led him to give up the former. Fisher always showed inductive logic leanings, seeking a single rational belief assignment. N-P were allergic to the idea. In the N-P philosophy, if there is a difference in problems or questions asked, we expect differences in which solutions are warranted. This is in sync with the view of the severe tester. In this sense, she is closer to Fisher's viewing the posterior distribution to be an answer to a different problem from the fiducial limits, where we expect the sample to change (Fisher 1930, p. 535).

Bridges to Fiducial Island: Neymanian Interpretation of Fiducial Inference?

For a long time Fiducial Island really was an island, with work on it sidestepped. A notable exception is Donald Fraser. Fraser will have no truck with those who dismiss fiducial inference as Fisher's "biggest blunder." "What? We still have to do a little bit of thinking! Tough!" (Fraser 2011, p. 330). Now, however, bridges are being built, despite minefields. Numerous programs are developing confidence distributions (CDs), and the impenetrable thickets are being penetrated. The word "fiducial" is even bandied about in these circles.[9] Singh, Xie, and Strawderman (2007) say, "a CD is in fact Neymanian interpretation of Fisher's fiducial distribution" (p. 132).

"[A]ny approach that can build confidence intervals for all levels, regardless of whether they are exact or asymptotically justified, can potentially be unified under the confidence distribution framework" (Xie and Singh 2013, p. 5). Moreover, "as a frequentist procedure, the CD-based method can bypass [the] difficult task of jointly modelling [nuisance parameters] and focus directly on the parameter of interest" (p. 28). This turns on what we've been

[9] Efron predicts "that the old Fisher will have a very good 21st century. The world of applied statistics seems to need an effective compromise between Bayesian and frequentist ideas" (Efron 1998, p. 112).

calling the piecemeal nature of error statistics. "The idea that statistical problems do not have to be solved as one coherent whole is anathema to Bayesians but is liberating for frequentists" (Wasserman 2007, p. 261).

I'm not in a position to evaluate these new methods, or whether they lend themselves to a severity interpretation. The CD program does at least seem to show the wide landscape for which the necessary mathematical computations are attainable. While CDs do not supply the uniqueness that Fisher sought, given that a severity assessment is always relative to the question or problem of interest, this is no drawback. Nancy Reid claims the literature on the new frequentist–fiducial "fusions" isn't yet clear on matters of interpretation.[10] What is clear, is that the frequentist paradigm is undergoing the "historical process of development . . . which is and will always go on" of which Pearson spoke (1962, p. 394).

Back to the ship!

[10] The 4th Bayesian, Fiducial and Frequentist workshop (BFF4), May 2017. Other examples are Fraser and Reid (2002), Hannig (2009), Martin and Liu (2013), Schweder and Hjort (2016).