Advances in Social Sciences Research Journal - Vol.5, No.8
Publication Date: Aug. 25, 2018
DoI: 10.14738/assrj.58.4967.

Saint-Mont, U. (2018). Where Fisher, Neyman and Pearson went astray: On the logic (plus some history and philosophy) of statistical tests. *Advances in Social Sciences Research Journal,* 5*(8)* 672-691.

Where Fisher, Neyman and Pearson went astray: On the logic (plus some history and philosophy) of Statistical Tests

Uwe Saint-Mont
Professor at Nordhausen University of Applied Sciences, Germany

ABSTRACT

Statistical tests, nowadays used throughout the sciences, have had considerable impact and a colourful history. This article studies the underlying logic(s) and associated philosophies, in particular Neyman's and Pearson's point of view.

Keywords: Statistical testing; tests of hypotheses; scientific thinking; modes of inference; history of statistics; foundational issues

INTRODUCTION

Every scientific endeavour consists of (at least) two components: A hypothesis on the one hand and data on the other. There is always a more or less abstract level - some theory, a set of concepts, certain relations of ideas - and a concrete level, i.e., empirical evidence, experiments or some observations which constitute matters of fact.

The focus of this contribution is on elementary models connecting both levels that have been very popular in the social sciences - statistical tests. Going from simple to complex we will examine four paradigms of statistical testing (Fisher, Likelihood, Bayes, Neyman & Pearson) and an elegant contemporary treatment.

In a nutshell, testing is an easy problem that has a straightforward mathematical solution. However, it is rather surprising that the statistical mainstream has pursued a different line of argument. The application of the latter theory in psychology and other fields has brought some progress but has also impaired scientific thinking.

FISHER: ONE HYPOTHESIS

> Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis. (Fisher (1935), p. 16)

The simplest and oldest formal model is Fisher's test of significance. There is just one distribution, called the "hypothesis" H (or P_H), and a sample from this population. Formally, the random variables $X, X_1, \ldots, X_n$ are iid, $X \sim P_H$, and $X_i = x_i$ are the observations subsequently encountered ($i = 1, \ldots, n$). Thus $\mathbf{x} = (x_1, \ldots, x_n)$ is the vector of data at hand.

In the continuous case, X has a density $f(x)$, whereas in the discrete case, X assumes values $x_1, x_2, \ldots$ with corresponding probabilities $p_1, p_2, \ldots$ Information theory often restricts attention to random variables assuming values in a finite alphabet $\mathcal{X} = \{x_1, \ldots, x_k\}$. In the following, in order to keep technical issues to a minimum, a random variable will be discrete if not otherwise stated.

Given this setting, suppose one observes a single x for which $p = P_H(X = x) = 0$. That is, this observation should not have occurred since the hypothesis does not account for it. It is simply impossible to see x if P_H is the case. In other words, this concrete observation x falsifies the hypothesis P_H, it is a counterexample to the law P_H. In philosophical jargon, this is a strict, logical conclusion (the modus tollens). One concludes without any doubt, although formalized by the probability statement $(p = 0)$, that the hypothesis in question is not the case.

Now, what if p is "small"? Obviously, no matter how small the probability, as long as $p > 0$, the observation x is possible and we cannot infer with rigour that $X \sim P_H$ did not produce it. All one may say is that

> *Either* an exceptionally rare chance has occurred, *or* the theory [hypothesis] of random distribution is not true. (Fisher (1956/73), p. 42, emphasis in the original)

Of course, such an "inductive statistic" (IS) argument is much weaker than "deductive nomological" (DN) conclusions, like the one considered before, and consequently a lot of discussion has been spawned by Fisher's dichotomy (see, e.g., Salmon (1989), Fetzer (2001)). More important to our reasoning is the observation that no general statistical theory evolved from Fisher's dichotomy. Here are two reasons why: First, if X assumes k distinct values $x_1, \ldots, x_k$ with probabilities $p_1, \ldots, p_k$, "small" crucially depends on the number of possible observations k. (A probability of $p = 1/100$ is small if $k = 10$, however the same probability is rather large if $k = 10^6$ or $k = 10^{20}$, say.) Second, if P_H is the uniform distribution, we have no reason whatsoever to discard the hypothesis no matter which value occurs. Each and every x_j is equally (un)likely, but possible nonetheless. If P_H has a geometric distribution, i.e., if X assumes the natural number j $(j \geq 1)$ with probability 2^{-j} it would be very difficult to tell beyond which number j_0 the probability p_{j_0} could be said to be "small". Finally, for any continuous distribution, in particular the standard normal, we have $P(X = x) = 0$ for every x. However, since some realization must show up, some x will occur nevertheless.

Perhaps for reasons such as these, Fisher came up with a more sophisticated idea. Typically, most values observed are rather "moderate", and only a few are "extreme" outliers (very large or very small). Suppose large values of X are suspicious. Then, having encountered x, it is straightforward to calculate $p = P_H(X \geq x)$, the probability of observing a value at least as large as x. If this so-called p-value is small, we have reason to reject the hypothesis. Of course, if small values of X are suspicious, it is $P_H(X \leq x)$ that should be considered, and in the case of outliers to the left and to the right of the origin, $P(X \geq |x|)$ is of interest. Thus we have a general rule: Calculate the probability of the value observed and of all "more extreme" events. This kind of evaluation may be crude, but it is also a straightforward way to quantify the evidence in the data x about the hypothesis P_H.

In the earliest test of this kind recorded, Arbuthnot (1710) looked at London births. His hypothesis was that it is equally likely to have a boy or a girl. (Why should one of the sexes be preferred?) Considering a moderate number n of years altogether, it would not be astonishing if the boys outnumbered the girls in about $n/2$ years, but by the laws of probability it would also not be surprising if there were more girls in perhaps 20 out of 30 years. However, it would be very surprising if, over a longer period of time, one sex outnumbered the other permanently. As a matter of fact, Arbuthnot checked $n = 82$ successive years and learned that in each and every year more boys than girls were born. If $P_H(boy) = P_H(girl) = 1/2$, the probability of the event "boys always outnumbering girls" happening by chance is 2^{-82}. Thus he concluded that some force "made" more boys than girls.

Suppose Arbuthnot had found eighty years with more boys than girls. Then Fisher's advice is to calculate

$$\begin{aligned} p &= P_H(X \geq 80) = P_H(X = 82) + P_H(X = 81) + P_H(X = 80) \\ &= \binom{82}{82}2^{-82} + \binom{82}{81}2^{-82} + \binom{82}{80}2^{-82} = \frac{1 + 82 + 3321}{2^{82}} = \frac{3404}{2^{82}} \approx 7 \cdot 10^{-22}. \end{aligned} \tag{1}$$

Since all probabilities sum up to one, this seems to be a small "probability value", and thus a remarkable result. Fisher (1929), p. 191, writes:

> It is a common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator [...]

Today, the standard levels of significance are $5\%, 1\%$, and 0.1%. Although, "surely God loves the 0.06 nearly as much as the 0.05?" (Rosnow und Rosenthal 1989)

Objections

Despite the above rather natural derivation, problems with p-values and their proper interpretation turned out to be almost endless:

The smaller the p-value, the larger the evidence against some hypothesis H, an idea already stated explicitly in Berkson (1942). Thus one should be able to compare p-values or combine p-values of different studies. Unfortunately, if two experiments produce the same p-value, they do not provide the same amount of evidence, since other factors, in particular the total number of observations n, also play a considerable role (Cornfield 1966: 19).

Johnstone (1986), p. 496, elaborates: "Thus, as Jeffreys explained in 1939, if the sample is very large, the level of significance P tends to exaggerate the evidence against the null hypothesis, i.e. P tends to be smaller than it ought to be. But in practice, if the sample is very large, a good orthodox statistician will 'deflate' intuitively the level of significance P accordingly."McPherson (1989) comments on this: "This is very likely true, but it is an inadequate base for presenting the p value approach to scientists."

The best one can do seems to be rules of thumb. For example, Efron und Gous (2001), p. 212, consider the normal distribution and sample size n in order to translate p-values into evidence. However, Royall (1986) demonstrates that contradictory statements are possible: "A given P-value in a large trial is usually stronger evidence that the treatments really differ than the same P-value in a small trial of the same treatments would be" (Peto et al. (1976), p. 593). But also "The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect [...] if the p-value is the same, the probability of committing a Type I error remains the same. Thus one can be more confident with a small N than a large N" (Bakan (1970), p. 241) is a reasonable line of argument.

In a nutshell, it is very difficult to interpret and combine p-values in a logically satisfactory way (see Greenland et al. (2016), Hubbard and Lindsay (2008) for recent overviews). Schmidt (1996), p. 126, also collects common ideas, in particular,

> If my findings are not significant, then I know that they probably just occurred by chance and that the true difference is probably zero. If the result is significant, then I

> know I have a reliable finding. The p values from the significance test tell me whether the relationship in my data are large enough to be important or not. I can also determine from the p value what the chances are that these findings would replicate if I conducted a new study

and then concludes that "every one of these thoughts about the benefits of significance testing is false." The most devastating point, however, seems to be the following consideration.

The Observed and the Unobserved

The distinction between the observed and the unobserved is fundamental to science. Science is built on facts, not speculation. Why have eminent statisticians confounded these two areas?

> It is not difficult to see how 'Student' and Fisher found themselves defending the use of the P integral. For if one accepts that it is possible to test a null hypothesis without specifying an alternative, and that the test must be based on the value of a test statistic in conjunction with its known sampling distribution on the null hypothesis, then the integral of the distribution between specified limits is the only measure which is invariant to transformation of the statistic. It follows that one is virtually forced to consider the area between the realized value of the statistic and a boundary as the rejection area - the P integral, in fact. (Edwards (1992), p. 178)

In other words, although the last paragraph can be interpreted as an invariance argument in favour of p-values (even if the measuring process is rather arbitrary, and only the ordering of the values recorded corresponds to something real, the p-value makes sense, since $P(X \geq x) = P(f(X) \geq f(x))$ for any monotone transformation f); Fisher, considering a single hypothesis, simply had no other choice but to calculate P integrals such as (1). He knew that this way to proceed was not really sound:

> Objection has sometimes been made that the method of calculating confidence limits by setting an assigned value such as 1% on the frequency of observing 3 or less [...] is unrealistic treating values less than 3, which have not been observed, in exactly the same manner as 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation. (Fisher (1956/73), p. 71)

However, a rather straightforward example illustrates that even the roundabout idea of "approximation" is difficult to defend. Suppose $P_H(X < x) = 0.01$ and $P_H(X = x) = 0.02$, small values of X being suspicious. If x is observed, the one-sided test may reject P_H since $P_H(X \leq x) = 0.03$. Now look at the (modified) hypothesis K where $P_K(X = x) = 0.02$, but $P_K(X < x) = 0.4$. In this case $P_K(X \leq x) = 0.42$ and no test would reject K. Yet the probability of the observed value x is the same for both hypotheses! The conclusion differs tremendously just because of values that were *not* observed:

> An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it. (Jeffreys (1939), p. 316)

Altogether, Fisher's paradigm seems to be too coarse. What is needed are more elaborated models, able to distinguish between observed and merely possible values, and explicitly formalizing other relevant aspects, such as the probability of committing an error or the strength of some effect.

TWO HYPOTHESES

In order to keep things as simple as possible, E. S. Pearson (1938), p. 242, proposed the following move:

> [...] the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the observed events with a greater degree of probability.[1]

Given (at least) two hypotheses H and K, it is of fundamental importance to understand that there are *two completely different ways to generalize* Fisher's approach. Either one sticks with integrals, which is the main feature of the Neyman-Pearson theory, or one directly compares $P_H(x)$ with $P_K(x)$. We will start with the latter idea:

Likelihood Ratio Tests

Given two hypotheses, it is perhaps most obvious to study the ratio $P_K(x)/P_H(x)$. In particular, since "... a proper measure of strength of evidence should not depend on probabilities of unobserved values" (Royall (1997), p. 69). Obviously, a ratio larger than one is evidence in favour of K, and a ratio that is smaller than one provides evidence in favour of H.

With successive observations $x_1, x_2, \ldots$ evidence for (and against) some hypothesis should build up. Mathematically, it is straightforward to consider the likelihood ratio, i.e., the product

$$r_n = r_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{P_K(x_i)}{P_H(x_i)}. \tag{2}$$

With every observation, the odds change in favour of one of the hypotheses (and thus, simultaneously, against the other). Let P_{X^n} be the empirical distribution of a sample of size n. Due to the law of large numbers, $P_{X^n}(x) \to P_H(x)$ for every $x \in \mathcal{X}$ almost surely, if P_H is the true distribution. This basic result almost immediately implies the likelihood convergence theorem: That is, (2) converges almost surely to zero if H is true, and to $+\infty$ if K is true. (See Royall (1997), p. 32, for discrete probability distributions and Chow and Teicher (1997), p. 257, for densities.)

It thus seems to be justified to decide in favour of K if the likelihood ratio exceeds some preassigned threshold s ($s > 1$). As Robbins (1970) showed, if H is correct, the probability that the ratio at one point of time exceeds s is just $1/s$. Formally:

$$P\left(\prod_{i=1}^{n} \frac{P_K(X_i)}{P_H(X_i)} \geq s \ \text{ for some } \ n = 1, 2, \ldots\right) \leq \frac{1}{s}$$

Notice that even "if an unscrupulous researcher sets out deliberately to find evidence supporting his favourite hypothesis [K] over his rival's [H], which happens to be correct, by a factor of at least [s], then the chances are good that he will be eternally frustrated" (Royall (1997), p. 7).

[1] As early as 1926, Gosset wrote to E.S. Pearson: "[...] if there is any alternative hypothesis [...] you will be much more inclined to consider that the original hypothesis is not true [...]" (See Royall (1997), p. 68, and the discussion in Hodges (1990), pp. 76.) It may be mentioned that Laplace had improved upon Arbuthnot in the 1770s, i.e., he had compared sex ratios at birth of several cities (cf. Stigler (1986), pp. 134).

Since the normal distribution is particularly important, Royall (1997), p. 52, considers it in much detail and finds that $s = 8$ and $s = 16$, or $s = 1/8 = 0.125$ and $s = 1/16 = 0.0625$, respectively, are reasonable choices. For more details see Royall (2000), Goodman und Royall (1988), and Bookstein (2014), p. 194, who reproduces Jeffreys' rule of thumb: $r_n > 1$ supports K, $1 > r_n > 0.3$ supports H, "but not worth more than a bare comment." However, the evidence in favour of H (and thus, equivalently, against K) is

substantial	if	$0.3 > r_n > 0.1$	very strong	if	$0.03 > r_n > 0.01$
strong	if	$0.1 > r_n > 0.03$	decisive	if	$0.01 > r_n$

Bayesian Tests

The likelihood ratio may serve as the core piece of a Bayesian analysis. To this end let π_H be the prior probability of the first hypothesis, and $\pi_K = 1 - \pi_H$ the prior probability of the second. Having observed $\mathbf{x} = (x_1, \ldots, x_n)$, Bayes' theorem states that the odds ratio of the posterior probabilities of the hypotheses is

$$\frac{\pi(K|x_1, \ldots, x_n)}{\pi(H|x_1, \ldots, x_n)} = r_n(x_1, \ldots, x_n) \cdot \frac{\pi_K}{\pi_H} = \prod_{i=1}^{n} \frac{P_K(x_i)}{P_H(x_i)} \cdot \frac{\pi_K}{\pi_H}. \tag{3}$$

If $0 < \pi_H < 1$, i.e., if both hypotheses are considered possible at the beginning, there are convergence results of a very general nature that guarantee that the true hypothesis will be found almost surely (e.g., Walker (2003, 2004)).

Moreover, it is possible to emulate Fisher's idea of a *single* explicit hypothesis. (For an example, see Bookstein (2014), pp. 197.)

Neyman and Pearson

Mathematicians J. Neyman and E.S. Pearson also improved upon Fisher's initial idea. In theory as well as in applications, their line of reasoning has become standard. Like Fisher, they used integrals, i.e., probabilities like $P(X \geq x)$. However, in order to avoid confounding the observed with the unobserved, they insisted that such probabilities be computed in advance, i.e., *before* recording empirical data.

Their paradigm situation is as follows: Denote by $N(\mu, \sigma)$ the normal distribution with expected value μ and standard deviation σ. Let $P_H \sim N(\mu_H, \sigma)$, $P_K \sim N(\mu_K, \sigma)$, and suppose without loss of generality that the absolute effect size $\eta = \mu_K - \mu_H$ is non-negative. Since for both hypothesis and each x the densities $\varphi_H(x)$ and $\varphi_K(x)$ are positive, we can never be sure which hypothesis is the case. All we can do is try to minimize the error of the first kind (a decision in favour of K, although H is true) and the error of the second kind (a decision in favour of H, although K is true).

Given population H or K, the mean $\bar{X}_n = \sum X_j/n$ of the observations is also normally distributed with parameters $\mu^{'}$, the correct hypothesis' expected value, and standard deviation $\sigma/\sqrt{n}$. (Thus, the larger the sample, the smaller the mean's standard deviation.) A rather straightforward treatment of this situation would look for the point m where $\varphi_H(x) = \varphi_K(x)$ which, due to symmetry, is just $m = (\mu_H + \mu_K)/2$, and decide in favour of H if $x < m$, and in favour of K if $x \geq m$. This leads to the total probability of error

$$P_e(n) = \alpha_n + \beta_n = P(\bar{X}_n \geq m|H) + P(\bar{X}_n < m|K) \tag{4}$$

which can be made arbitrarily small with growing n, for any fixed $\eta = \mu_K - \mu_H > 0$.

However, perhaps since the errors of the first and of the second kind have different consequences, Neyman and Pearson decided to treat the null hypothesis H (typically representing arbitrary fluctuations, i.e., "no effect") and the alternative K (representing a substantial effect) *asymmetrically*. With n and the effect size η thus given, Neyman und Pearson (1933), pp. 79, advised as follows:

> From the point of view of mathematical theory all that we can do is to show how the risk of the errors $[\alpha, \beta]$ may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance [between the two kinds of errors] should be struck, must be left to the investigator.

They also fixed α (i.e, the level of error of the first kind, meaning that an effect is detected although there is none). Now they could look for the optimum decision procedure, minimizing β, which they determined in Neyman und Pearson (1933).

Knowing the best test, one can also control for the errors (e.g., by fixing α to 0.01, and assuming $\beta = 0.3$, say), and set out to detect an effect of a certain size η with the minimum number of observations n necessary. E. S. Pearson (1955), p. 207, explains:

> The appropriate test is one which, while involving (through the choice of its significance level $[\alpha]$) only a very small risk of discarding my working hypothesis $[H]$ prematurely will enable me to demonstrate with assurance $[1 - \beta]$ (but without any unnecessary amount of experimentation) the reality of the influences which I suspect may be present $[K]$.

In this view, every observation comes with a cost and a major goal of the statistical design of experiments is to make just enough observations in order to convincingly demonstrate a certain effect - n is just as large as necessary, not as large as possible.

SOME CONSEQUENCES

The standard style of inference

Suppose there is an effect η of a certain size, and the sample size n is fixed. Then the investigation hinges strongly on the asymmetry between α and β, being treated differently. Cornfield (1966), p. 21, wasn't the only one to question this choice:

> It is clear that the entire basis for sequential analysis [and much of received testing theory] depends upon nothing more profound than a preference for minimizing β for given α rather than minimizing their linear combination. Rarely has so mighty a structure and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference.

In practice, researchers did not use the additional degree of freedom introduced by Neyman und Pearson (1933) either. Despite their and Fisher's advice, rather coarse standards such as $\alpha = 0.05$, or Cohen's (1988, 1992) classification of effects (small, medium, large) caught on, until testing became a "ritual" (Gigerenzer et al. 2004).

With all parameters set in advance, a test is indeed a strict decision procedure, and "the basic objection to this program is that it is too rigid..." (Lehmann (1993), p. 70). In fact, all one gets is an (asymmetric) dichotomous decision against or in favour of K, and the procedure is so tight that it cannot be extended at all. Cornfield (1966), p. 19, writes (see also Royall (1997), p. 111 on these matters):

> An experimenter, having made n observations in the expectation that they would permit the rejection of a particular hypothesis, at some predesignated significance level, say .05, finds that he has not quite attained his critical level. He still believes that the hypothesis is false and asks how many more observations would be required to have reasonable certainty of rejecting the hypothesis [...]
>
> Under these circumstances it is evident that there is no amount of additional information, no matter how large, which would permit rejection at the .05 level. If the hypothesis being tested is true, there is a .05 of its having been rejected after the first round of observations. To this chance must be added the probability of rejecting after the second round, given failure to reject after the first, and this increases the total chance of erroneous rejection to above .05 [...] Thus no amount of additional evidence can be collected which would provide evidence against the hypothesis equivalent to rejection at the $P = 0.05$ level [...]

In other words: In this perspective, α is a limited, non-renewable resource. "Once we have spent this error rate, it is gone" (Tukey (1991), pp. 104). Thus it has to be used with great care: "[...] a very few prespecified comparisons will be allowed to eat up the available error rate, and the remaining comparisons have the logical status of hints, no matter what statistical techniques may be used to study them." (Tukey (1991), pp. 104)

In order to avoid an "inflation" of error, it seems wise to distribute the error rate of 5% say, among all tests planned. The standard technique is to adjust α, a priori, by some scheme taking the whole family of tests into account. Salsburg (1985), p. 221, reports the consequences of such a consistent attitude:

> Finally, we should consider the subclass of practitioners who are 'more holy than the Pope,' so to speak. To these practitioners, the whole purpose of the religion of Statistics is to maintain the sanctity of the alpha level (which is another name for 0.05). No activity that appears to involve looking at data for sensible combinations of interesting effects is allowed. It is forbidden, in fact, to do anything more than to compute the p value using a method determined in advance of the experiment and fully documented at that time.

Note also that if only a small proportion of α is spent in every test, the overall procedure becomes very conservative: In the Neyman-Pearson framework, a very small α corresponds to an inflation of β and thus deteriorating power $1 - \beta$. Since research in the social sciences is generally plagued by low power, this attitude makes it even more difficult to detect effects. Ellis (2010), p. 79, concludes:

> Instead of dealing with the very credible threat of Type II errors, researchers have been imposing increasingly stringent controls to deal with the relatively unlikely threat of

> Type I errors (Schmidt 1992). In view of these trade-offs, adjusting alpha may be a bit like spending $1,000 to buy insurance for a $500 watch.

Royall (1991), p. 57, states another way to deal with the problem described by Tukey. Instead of lowering α for each test, one simply restricts the number of planned tests:

> ...do not allow those who are conducting the trial to look at the results as they accumulate. That is, [...] conceal the evidence from the physician until the trial is completed.

Altogether, the Neyman-Pearson framework gives some justification for minimizing the amount of information collected, and the number of looks at the data. This fits well with Popper's rationalistic view, who always emphasized the role of theory and deduction in the guise of falsification, downplaying the role of data, and rejecting induction firmly (Popper 1959, Popper and Miller 1983). However, scientific common sense and practice rather point in the opposite direction: If we are to learn from experience, an open-minded attitude and any reasonable analysis, be it hypothesis- or data-driven, should be encouraged. Keiding (1995), p. 242, admits that

> [...] it is indeed unsatisfactory to have to defend, perhaps in the face of senior, highly qualified substantive scientists, our mainstream statistical thinking which assumes that you are not supposed to look at the data when searching for methods of optimal analysis with the purpose of gaining new knowledge.

Confusion

Since there are several theories (at least two), each of them accompanied by a certain "logic", data analysis is a tricky business, and there is also lot of confusion.

In particular, despite their mathematical similarity, data-dependent p-values and error levels set in advance are completely different. It is against the grain of the Neyman-Pearson theory to calculate α-levels a posteriori (for example, one, two or three stars indicating that some empirical result has been significant at the 0.05, 0.01 or the 0.001-level), to report p-values instead of zero-one decisions, or to restrict attention to one hypothesis (typically the null, although two hypotheses might be mentioned). Nevertheless, practice and textbooks use p-values and α-levels almost interchangeably, thus creating an "alphabet soup" (Hubbard 2004).

On a less formal level, there is also much conceptual confusion, (inductive) evidence in Fisher's sense and (deductive) decisions in Neyman's and Pearson's being conflated:

> This hybrid is essentially Fisherian in its logic, but it plays lip service to the Neyman-Pearson theory of testing [...] Some researchers do use the Neyman-Pearson theory of testing in a pure form, but they constitute a small minority [...] Regardless of their terminology and verbal allegiance, most researchers in the fields mentioned above use and/or accept as valid a pattern of inductive reasoning that is characteristic for the Fisherian test of significance. (Spielman (1974), p. 211)

It is a crucial ingredient of the standard Neyman-Pearson theory to treat the hypotheses asymmetrically. Typically, the null hypothesis represents the idea that pure chance produced the data at hand, whereas its alternative claims that an interesting substantial effect has left its traces

in the data. Obviously, any "logic of empirical science" demands that the more data there is, the more difficult it should be for a substantial hypothesis to succeed: "... in physics and the related disciplines the parent theory is subjected to ever more critical examination as measurement techniques, in their broadest sense, improve. That is, as power increases the 'observational hurdle' that the theory must clear becomes greater." (Oakes (1986), pp. 40)

In other words, as information accrues, it becomes easier to detect if the data deviate from a particular hypothesis. For example, suppose your hypothesis (derived from basic theory) claims that about $6.6 \cdot 10^{10}$ neutrinos should hit the surface of the earth per second and cm^2. Then measurements should confirm this guess, i.e., the number of neutrinos actually counted should be close to $6.6 \cdot 10^{10}/s \cdot cm^2$. In the jargon of statistical tests this means that "[...] in the physical sciences the substantive theory is associated with the null hypothesis and to the extent that it defies rejection it commands respect" (cf. Oakes (1986), p. 41. For a contemporary example see van Dyk (2014).)

However (Oakes (1986), pp. 40), "the opposite is the case in the social and behavioural sciences [...] In psychology and the social sciences the substantive theory is associated with the alternative hypothesis and is corroborated as the null hypothesis is rejected. In this sense the observational hurdle which the theory must clear is lowered as power or experimental precision is increased. This is the great weakness of identifying a theory with the alternative hypothesis":

> Putting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory* (Meehl (1978), p. 822, emphasis in the original).

Perhaps it is quite telling that, although this phenomenon was described by an eminent psychophysicist 50 years ago (Meehl 1967), and has been decried many times ever since (e.g., Meehl (1990, 1997), Gelman et al. (2013), Bookstein (2014)), this kind of "mindless statistics" (Gigerenzer 2004) has thrived (Hubbard und Ryan 2000). Its "career" is quite similar and related to that of p-values which, despite their major shortcomings, have also become standard in many sciences.

The scientific style of inference

Apart from the consequences already described, the standard treatment, i.e.,

1. considering intervals like $P_H(X \leq x)$ instead of point probabilities,
2. dealing with the hypotheses (and thus, α, β) in an asymmetric manner, and
3. putting all parameters constituting a standard test on a par

has led to the following:

(a) $1 - \beta$ is identified with the importance or even with the "scientific power" of a certain study.

(b) Power analysis. For example, upon designing a clinical trial, it is now mandatory to calculate the number of patients n, given the level of significance α, power $1 - \beta$, and effect size η.

(c) The attitude that observations are "expensive" - since, given α, β and η, the above line of thought supposes that a small sample is optimum.

Alas, "inventing virtuous-sounding terms" (cf. Jaynes (2003), p. 514) like *power* does not solve problems. Rather, emphasis on $1-\beta$ obscures the fact that the "real" impact of a trial consists in its contribution to a series of experiments, all investigating the same phenomenon (Ottenbacher 1996). To this end, the effect size η is much more important:

> ...the emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance. Regardless of sample size, the size of an effect in one study is a reasonable estimate of the size of an effect in replication (Tversky and Kahneman (1971), p. 110).

Guttman (1985), pp. 3, adds: "The emphasis on statistical significance over scientific significance in education and research represents a corrupt form of the scientific method..."

A priori power analysis (Cohen 1988, Ellis 2010) hinges on the idea that α, $1-\beta$, η and n "... are so related that any one of them is a function of the other three, which means that when any three of them are fixed, the fourth is completely determined" (Cohen (1988), p. 14). Although it surely is a good idea to think hard about one's hypotheses before collecting data, formulae such as (4) indicate that α and β had better depend on n. In particular, $\alpha(n)$ should be a decreasing function in n (see, e.g., subsection "objections," Lindley (1957), Hurlbert and Lombardi (2009), pp. 333, and Naaman (2016)).

Finally, and most importantly, since information accrues with data, the overall attitude toward n should be quite the opposite to that of Pearson:

> There are no inferential grounds whatsoever for preferring a small sample [...] the larger the sample the better [...] The larger the sample size the more stable the estimate of effect size; the better the information, the sounder the basis from which to make a decision [...] (Oakes (1986), pp. 29, 32)

In everyday life, this often means collecting data until the evidence has accumulated sufficiently:

> An experiment involving an image-producing apparatus often ends appropriately with a 'golden event', that is, a picture or image of something whose existence has been conjectured, but possibly questioned. An experiment involving a counting apparatus often ends appropriately when a decision based on some probability model suggests that enough counts have been taken for some purpose. (Ackermann (1989), p. 189)

If some insight thus occurs all of a sudden, the crucial last step, has, with a wink, been called the *interocular traumatic test*: "You know what the data mean when the conclusion hits you between the eyes" (Edwards et al. (1962), also see Bookstein (2014)).

Altogether, η and n seem to be much more important than α and β. It is also no coincidence that any philosophy based on a suboptimal formal treatment yields opinions that are at variance with common sense (Neyman 1977, Mayo 1996). For an important example see the next section.

A blurred view

Neyman und Pearson (1933), p. 74, state:

> If x is a continuous variable ... then any value of x is a singularity of relative probability equal to zero. We are inclined to think that as far as a particular hypothesis is

> concerned, no test based upon a theory of probability (Footnote: cases will of course, arise where the verdict of a test is based on certainty...) can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

In the light of the above discussion, this statement - still very popular today - permutes rule and exception. It is much too pessimistic, since, owing to the (very) general convergence results, no matter whether the variables are discrete or continuous, given enough observations, H and K can be distinguished with hardly any doubt. For example, just a few throws suffice to decide between a cube with the numbers $\{0,\dots,5\}$, and a cube with the numbers $\{1,\dots,6\}$. More generally speaking, if the support of H and K is not the same (i.e., if there exists some x such that $P_H(x) = 0$ and $P_K(x) > 0$, or vice versa), one is able to discriminate deterministically between the hypotheses after just a finite number of observations.

Of course, for any continuous random variable X, and any realization x, $P(X = x) = 0$. Therefore, given two hypotheses, one has to consider their densities, $f_H(x)$ and $f_K(x)$ say. In the case of the normal family (and many others), the support of any two densities coincides. Thus, rather trivially, no matter which x is observed, one cannot decide for sure if H or K is the case. However, the ratio $f_K(x)/f_H(x)$ gives valuable evidence and much more so will $r_n(x_1,\dots,x_n)$ if n is not too small. Asymptotically, any doubt vanishes completely. Thus in a nutshell, a statistical test is a powerful tool.

TESTING NEED NOT BE COMPLICATED

Statisticians, philosophers and scientists have written much about (styles of) inference and statistical philosophies (e.g., Neyman (1955, 1961, 1977), Jones (1986), Good (1988), Barnett (1999), Fisher (2003), Jaynes (2003), Dienes (2011), Cumming (2014), Spanos (2014), Haig (2016)). Instead of adding another opinion, it may be wiser to go back to the original issue:

First, since the basic problem is rather elementary, one expects an elegant, satisfactory answer. Second, since Fisher's treatment is too coarse and leads immediately to almost inextricable problems, there is a consensus that two hypotheses should be considered. Third, the last section shows that Neyman's and Pearson's treatment has led to disappointment. Why?

Looking at their model from a mathematical point of view, the P integral springs to mind. Introduced by Fisher - faute de mieux - it is given the leading part in Neyman's and Pearson's two-hypotheses setting, and is at the root of all subsequent trouble. More precisely: To keep up the basic distinction between the observed and the unobserved, one has to stick to a strict prior viewpoint. Since this is hardly possible and has curious consequences, it is no coincidence that Neyman's and Pearson's stance has merged with Fisher's position (and other ideas), almost inevitably creating confusion and endless discussion.

The good news is that a large part of the scientific and philosophical turmoil is due to a *particular mathematical treatment* - Neyman and Pearson have made testing more complicated than it needed to have been. Therefore a better, more elegant treatment should be able to rectify most of the defects since, due to the law of large numbers, testing is an easy problem:

If n is not too small, the empirical distribution of the data P_{X^n} is (in any reasonable sense) close to the true distribution P_H. The test of one hypothesis gives a formalized answer to the simple question: Is the data I have observed compatible with my hypothesis? If there are two or several hypotheses, the question becomes: Given my set of data, which hypothesis should I

choose? Qualitatively speaking, it is reasonable to choose the hypothesis which is closest to the data, and to reject a hypothesis if the data is "far away" from P_H.

A contemporary treatment, focussing on information and (generalized) distance of distributions, therefore starts with the likelihood ratio. Suppose there are two hypotheses H, K, represented by their distributions P_H, P_K. Define their KL-divergence (Kullback and Leibler 1951):

$$D(P_K||P_H) = \sum_{x\in\mathcal{X}} P_K(x) \log \frac{P_K(x)}{P_H(x)},$$

where $D(P_K||P_H) < \infty$, and, for the sake of mathematical simplicity, $\mathcal{X}$ is a finite set. $D(P_K||P_H)$ may be interpreted as a "generalized distance" between distributions. Since its introduction, it has become a core concept of information theory and beyond (e.g., see Cover und Thomas (2006), pp. 377, and their pointers to the literature).

The key result, connecting the likelihood ratio and KL-divergence is

$$\log \frac{P_K(x_1,\ldots,x_n)}{P_H(x_1,\ldots,x_n)} = \sum_i \log \frac{P_K(x_i)}{P_H(x_i)} = n\left(D(P_{X^n}||P_H) - D(P_{X^n}||P_K)\right), \tag{5}$$

where P_{X^n} is the empirical distribution of the data.

In the most complete (i.e., Bayesian) setting, H and K are endowed with prior probabilities, π_H and π_K, respectively. Given an iid sample $x_1,\ldots,x_n$ from either P_H or P_K, let $A_n \subseteq \mathcal{X}_n$ be the acceptance region for H, depending on n. Thus one obtains the error probabilities $\alpha_n = P_H(\bar{A}_n)$ and $\beta_n = P_K(A_n)$, where $\bar{A}_n$ denotes the complement of A_n, i.e., the acceptance region of K. Finally, it is straightforward to minimize the total probability of error $P_e(n) = \pi_H\alpha_n + \pi_K\beta_n$.

Given this symmetric treatment of the hypotheses, and error probabilities depending on n, it turns out (Cover und Thomas (2006), p. 388) that "the optimum decision rule is to choose the hypothesis with the maximum a posteriori probability," which means to choose K if $\pi_K P_K(X_1,\ldots,X_n) > \pi_H P_H(X_1,\ldots,X_n)$, and H, if the inequality is in the other direction. Equivalently, the best strategy is a decision in favour of K if

$$\log \frac{\pi_K}{\pi_H} + \sum_i \log \frac{P_K(X_i)}{P_H(X_i)} > 0,$$

and in favour of H otherwise. Because of (5), the latter inequality is tantamount to a decision in favour of K if and only if

$$nD(P_{X^n}||P_H) - \log \pi_H > nD(P_{X^n}||P_K) - \log \pi_K.$$

Without prior probabilities, it is best to decide in favour of K if the empirical distribution is "closer" to P_K, i.e., if $D(P_{X^n}||P_H) > D(P_{X^n}||P_K)$. More generally speaking, because of (5), a decision in favour of K if $P_K(x_1,\ldots,x_n)/P_H(x_1,\ldots,x_n) \geq s$, i.e., if the likelihood ratio exceeds a certain threshold ($s \geq 1$), is equivalent to a decision in favour of K if $D(P_{X^n}||P_H) - \frac{1}{n}\log s \geq D(P_{X^n}||P_K)$. In other words, the likelihood ratio test advises choosing K if the divergence $D(P_{X^n}||P_K)$ is smaller than $D(P_{X^n}||P_H)$ minus the asymptotically vanishing "safety margin" $(\log s)/n \geq 0$. Moreover, if K is true,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P_K(X_1,\ldots,X_n)}{P_H(X_1,\ldots,X_n)} \to D(P_K||P_H) \quad \text{in probability.}$$

The test closest to Fisher's original idea is Hoeffding's "universal test", which merely compares the data with a fixed hypothesis. It decides in favour of H if the empirical distribution is within a certain acceptance region A_n about P_H, i.e., if

$$D(P_{X^n}||P_H) \leq c_n.$$

Because of the law of large numbers, the sequence c_n decreases rapidly with increasing n. (For details see Hoeffding (1965), theorems 7.1 and 5.1.)

DISCUSSION AND CONCLUSIONS

At first glance, it seems to be a drawback that KL-divergence is not a proper metric. In particular, given two distributions P_H and P_K, in general, $D(P_H||P_K) \neq D(P_K||P_H)$. However, in the case of data and hypotheses this is an advantage, since there is a striking asymmetry between moving from specific observations to general laws (induction) and the opposite direction (deduction).

Hoeffding's test starts with a hypothesis H and asks if the data lies within a circle of radius c_n about P_H. An even more straightforward way to proceed would be to start with data $\mathbf{x}_n = (x_1, \ldots, x_n)$ and ask if P_H lies within a circle of radius $c_n^{'}$ about the empirical distribution. There is a real difference: Hoeffeding looks for data compatible with some conjectured hypothesis, whereas the second approach conditions on the data and looks for hypotheses that are compatible with the observations.

At least from a mathematical point of view, asymptotically, these preferences do not matter, since, if H is the true hypothesis, by the law of large numbers $P_{X^n}(a) \to P_H(a)$ for every $a \in \mathcal{X}$ in probability (and almost surely). Thus, for every a with $P_H(a) > 0$,

$$\lim_{n\to\infty} \frac{P_{X^n}(a)}{P_H(a)} = \lim_{n\to\infty} \frac{P_H(a)}{P_{X^n}(a)} = 1$$

which implies $\lim_{n\to\infty} D(P_{X^n}||P_H) = D(P_H||P_{X^n}) = 0$ with probability one.

In total generality, i.e., in philosophy, deduction is regarded as rather unproblematic. However, the problem of induction has haunted statistics, philosophy, and perhaps also the sciences at least since David Hume's time (Howson 2003). A standard statistical test is a particularly simple model to study these matters - a "test bed" if you allow the play on words. Any such test considers hypotheses (typically two), collects an iid sample $x_1, \ldots, x_n$, and finally decides in favour of or against a hypothesis. Schematically,

$$\begin{array}{ccccc} P_H & & \text{Potential distributions (laws)} & & P_K \\ & \nwarrow & & \nearrow & \\ & & \text{Data} \quad (x_1, \ldots, x_n) & & \end{array}$$

Philosophers named this kind of reasoning "inference to the best explanation" (Lipton 2004), but also leading statisticians have always been well aware of the basic issue involved. While Fisher (1935, 1955) thought in terms of inductive inference, Neyman (1977) sided with the deductive line of argument. Considering a single experiment, Fisher thus calculates the p-value, expressing the *evidence* in the data against a hypothesis. Starting with hypotheses, instead, Neyman and Pearson focus on *probabilities of error* and how to control them. (Quite similarly, Bayesians focus on the data at hand, whereas orthodox theory is much more concerned with the process producing the data.)

In the end, the strong link between Neyman's Frequentist school and Popper's critical rationalism strengthened both points of view, with the consequence that their positions succeeded after the death of their major opponents (Fisher died in 1962, and Carnap in 1970). In particular, induction was banned, and *mathematical* statistics superseded semantic reasoning to an extent that even analyzing given sets of data became suspicious.

Since the 1970s, many statisticians, scientists and philosophers have worked on overcoming this distorted view (e.g., Tukey (1977), Hedges und Olkin (1985), Ghosh (1988), Schmidt (1992, 1996), Berthold und Hand (2003), Jaynes (2003), Heckman (2005), Howson und Urbach (2006), Rissanen (2007), Hurlbert and Lombardi (2009), Pearl (2009), Ellis (2010), Bookstein (2014)). Perhaps since the "big questions" thus demanded much attention, rather elementary facts like those pointed out in this contribution could be overlooked easily.

It is most significant that due to the law of large numbers, the distance between sample and population shrinks (quickly) when n gets larger. This basic insight makes testing an easy problem: Given enough data - and thus information - the true distribution comes into focus almost inevitably. Therefore, mathematically, all approaches based on the straightforward ratio $P_K(X = x)/P_H(X = x)$ lead to unequivocal and strong convergence results. In other words, sufficiently precise and distinct hypotheses can be tested efficiently, at least, if the hypotheses are treated in a rather symmetric way (Royall 1997, Robert 2007).

On a larger scale, putting information first gives sound answers to quite a few questions. For example, if there is an uncountable number of hypotheses, e.g., a parameterized family of distributions $P_\theta(x)$, Fisher's likelihood function $L_{\mathbf{x}}(\theta)$ provides the key to an elegant solution, which can be extended to an enormously general and powerful information-oriented approach that is perfectly compatible with scientific common sense (e.g., Aldrich (1997), Burnham and Anderson (2002), Li and Vitányi (2008)).

Finally, it should be mentioned that particular formal treatments are associated with certain schools of thought - and it is rather the detailed treatment that triggers the overall attitude than vice versa (methods first, philosophy second). Therefore, quite straightforwardly, an elegant mathematical treatment comes with a "moderate" and reasonable standpoint, whereas questionable decisions lead to rather "extremist" paradigms. The above exposition demonstrates that it may take decades - filled with excessive discussions ranging from formal minutiae to philosophical principles - to overcome popular, yet distorted, points of view.

References

Ackermann, R. (1989). The New Experimentalism. *Brit. J.Phil. Sci.* **40**, 185-190.

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood. *Stat. Sci.* **12(3)**, 162-176.

Arbuthnot, J. (1710). An Argument for Divine Providence, taken from the Constant Regularity Observ'd in the Births of Both Sexes. *Phil. Trans. R. Soc.* **27**, 186-190.

Bakan, D. (1970). The Test of Significance in Psychological Research. Chapter 25 in Morrision, D.E. and Henkel, R.E. (eds.). The Significance Test Controversy. *Aldine Publishing Company, Chicago, IL*, 231-251.

Barnett, V. (1999). Comparative Statistical Inference. (3rd ed.) *Wiley, New York*.

Berkson, J. (1942). Tests of Significance Considered as Evidence. *JASA* **37**, 325-335.

Berthold, M.R.; and D.J. Hand (eds., 2003). Intelligent Data Analysis. (2nd ed.) *Springer, Berlin*.

Bookstein, F.L. (2014). Measuring and reasoning. Numerical inference in the sciences. *Cambridge Univ. Press, New York*.

Burnham, K.P.; and D.R. Anderson (2002). Model Selection and Multimodel-Inference. A Practical Information-Theoretic Approach. (2. ed.) *Springer, New York*.

Chow, Y. S.; and Teicher, H. (1997). Probability Theory. Independence, Interchangeability, Martingales. *Springer, New York: Springer Texts in Statistics*.

Cohen, J. (1988). Statistical Power Analysis for the Social Sciences. (2. ed.) *Erlbaum, Hillsdale, NJ.*

Cohen J. (1992): A power primer. Quant. Meth. Psychol. **112**, 155–159.

Cornfield, J. (1966). Sequential Trials, Sequential Analysis and the Likelihood Principle. *American Statistician* **20(2)**, 18-23.

Cover, T.M. und J.A. Thomas (2006). Elements of Information Theory. *Wiley, New York*, 2nd ed.

Cumming, G. (2014). The new statistics: Why and how. Psychological Science, 25, 7-29.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, **6**, 274-290.

Edwards, A. W. F. (1922). Likelihood. *Johns Hopkins University Press, Baltimore, MD.*, 2nd ed.

Edwards, W.; Lindman, H.; and J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70**, 193-242.

Efron, B.; and Gous, A. (2001). Scales of Evidence for Model Selection: Fisher versus Jeffreys. In: Lahiri, P. (ed.) *IMS Lecture Notes* **38** on *Model Selection*, 210-256.

Ellis, P.D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. *Cambridge University Press*, New York.

Fetzer, J.H. (ed., 2001). The Philosophy of Carl G. Hempel. Studies in Science, Explanation, and Rationality. *Oxford University Press, New York*.

Fisher, R.A. (1929). The Statistical Method in Psychical Research. *Proceedings of the Society for Psychical Research* **39**, 189-192.

Fisher, R.A. (1935/1966). The Design of Experiments. Cited according to the 8th ed., *Hafner Publishing Company, New York*.

Fisher, R.A. (1935). The Logic of Inductive Inference. *J. of the Royal Stat. Soc.* **98**, 39-54.

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *J. of the Royal Stat. Soc., Ser. B* **17(1)**, 69-78.

Fisher, R.A. (1956/73). Statistical Methods and Scientific Inference. *Hafner Publishing Company, New York*. Citet according to the 3rd edition, 1973.

Fisher, R.A. (2003). Statistical methods, experimental design, and scientific inference: a re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference. *Oxford Univ. Press, Oxford*.

Gelman, A; Carlin, J.B.; Stern, H.S.; Dunson, D.; Vehtari, A.; and Rubin, D.B. (2013). Bayesian Data Analysis (revised ed.) *Chapman & Hall/CRC, New York*.

Ghosh, J.K. (ed., 1988). Statistical Information and Likelihood. A Collection of Critical Essays by Dr. D. Basu. *Springer, New York: Lecture Notes in Statistics*.

Gigerenzer, G. (2004). Mindless Statistics. *The J. of Socio-Economics* **33**, 587-606.

Gigerenzer, G.; Krauss, S.; and Vitouch, O. (2004). The Null Ritual. What you always wanted to know about significance testing but were afraid to ask. Chapter 21 in: Kaplan (ed.) The Sage Handbook of Quantitative Methodology for the Social Sciences. *Sage, Thoudsand Oaks*, 391-408.

Good, I.J. (1988). The Interface between Statistics and Philosophy of Science. *Statistical Science* **3(4)**, 386-412.

Goodman, S.N.; and Royall, R. (1988). Evidence and Scientific Research. *American J. of Public Health* **78(12)**, 1568-1574.

Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; and D.G. Altman (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31**, 337–350.

Guttman, L. (1985). The Illogic of Statistical Inference for Cumulative Science. *Applied Stochastic Models and Data Analysis* **1**, 3-9.

Haig, B.D. (2016). Tests of statistical significance made sound. *Educational and Psychological Measurement* **35**, 1-18.

Heckman, J.J. (2005). The Scientific Model of Causality. *Socio. Methodology* **35**, 1-162.

Hedges, L.V.; and Olkin, I. (1985). Statistical Methods for Meta-Analysis. *Academic Press, Orlando*.

Hodges, J. S. (1990). Can / May Bayesians Do Pure Tests of Significance? In: Geisser, S.; Hodges, and Press, J. (eds.) Bayesian and Likelihood Methods in Statistics and Econometrics (vol. 7). Essays in Honor of George A. Barnard. *North-Holland Publishing Company, Amsterdam: Studies in Bayesian Econometrics and Statistics*, 75-90.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Annals Math. Statistics* **36**, 369-408.

Howson, C. (2003). Hume's problem: Induction and the justification of belief. *Clarendon Press, Oxford.*

Howson, C.; and Urbach, P. (2006). Scientific Reasoning. The Bayesian Approach. (3rd ed.) *Open Court, Chicago and La Salle, IL.*

Hubbard, R. (2004). Alphabet soup. Blurring the distinctions between p's and α's in psychological research *Theory & Psychology* **14(3)**, 295-327.

Hubbard, R.; and Lindsay, R.M. (2008). Why P Values are Not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology* **18(1)**, 69-88.

Hubbard, R.; and Ryan, P.A. (2000). The Historical Growth of Statistical Testing in Psychology - and its Future Prospects. *Edu. & Psych. Measurement* **60(5)**, 661-681.

Hurlbert, S.H., and C.M. Lombardi (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, **46**, 311-349.

Jaynes, E.T. (2003). Probability Theory. The Logic of Science. Edited by Bretthorst, G. L. *Cambridge University Press, Cambridge.*

Jeffreys, H. (1939). Theory of Probability. *Clarendon Press, Oxford.*

Johnstone, D. J. (1986). Tests of Significance in Theory and Practice. *The Statistician* **35(5)**, 491-504.

Jones, L.V. (ed., 1986). The collected works of J.W. Tukey, Vol. III & IV: "Philosophy and Principles of Data Analysis." *Chapman & Hall, London.*

Keiding, N. (1995). *Test* **4(2)**, 241-242. Commentary on Cox, D.R. (1995). The Relation between Theory and Application in Statistics. (with discussion) *Test* **4(2)**, 207-261.

Kotz, S.; and Johnson, N.L. (1993). Breakthroughs in Statistics. Vol. I: Foundations and Basic Theory. *Spinger, New York.*

Kullback, S.; and Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics* **22(1)**, 79-86.

Lehman, E. L. (1993). Introduction to Neyman und Pearson (1933). In: Kotz and Johnson (1993), 67-72.

Li, M.; and Vitányi, P. (2008). An Introcudtion to Kolmogorov Complexity and its Applications. (3rd ed.) *Springer, New York.*

Lindley, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lipton, P. (2004). Inference to the Best Explanation. (2nd ed.) *Routledge, London.*

Mayo, D.G. (1996). Error and the Growth of Experimental Knowledge. *The University of Chicago Press, Chicago, IL.*

McPherson, G. (1989). The Scientist's View of Statistics - a Neglected Area. *J. of the Royal Stat. Society* **152**, 221-240.

Meehl, P.E. (1967). Theory-Testing in Psychology and Physiscs: a Methodological Paradox. *Philosophy of science* **34**, 103-115.

Meehl, P.E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *J. of Consulting and Clin. Psych.* **46**, 806-834.

Meehl, P.E. (1990). Apraising and Amending Theories: The Strategy of Lakatosian Defence and Two Principles that Warrant it. *Psychological Inquiry* **1(2)**, 108-141.

Meehl, P.E. (1997). The Problem is Epistemology, not Statistics: Replace Confidence Intervals and Quantify Accuracy of Risky numerical Predictions. In: Harlow, L.L.; Mulaik, S.A.; and Steiger, J.H. (eds., 1997). What if there were no Significance Tests? *Erlbaum, London*, 393-425.

Naaman, M. (2016). Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox. Electronic Journal of Statistics **10**, 1526–1550.

Neyman, J. (1955). The Problem of Inductive Inference. *Communications on Pure and Applied Mathematics* **VIII**, 13-46.

Neyman, J. (1961). The Silver Jubilee of My Dispute with Fisher, *Journal of the Operations Research Society of Japan*, **3**, 145-154.

Neyman, J. (1977). Frequentist Probability and Frequentist Statistics. *Synthese* **36**, 97-131.

Neyman, J.; and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosoph. Trans. Royal Soc. London A* **231**, 289-337. Cited according to Kotz and Johnson (1993), 73-108.

Oakes, M. (1986). Statistical Inference: A Commentary for the Social and Behavioral Sciences. *Wiley, New York.*

Ottenbacher, K.J. (1996). The Power of Replications and Replications of Power. *The American Statistician* **50(3)**, 271-275.

Pearl, J. (2009). Causality. Models, Reasoning and Inference. (2nd ed.) *Cambridge Univ. Press.*

Pearson, E. S. (1938). Student as Statistician. *Biometrika* **30**, 210-250.

Pearson, E. S. (1955). Statistical Concepts and their Relation to Reality. *J. of the Royal Statistical Society, Ser. B* **17(2)**, 204-207.

Peto, R.; Pike, M. C.; Armitage, P.; Breslow, N. E.; Cox, D. R.; Howard, S. V; Mantel, N.; McPherson, K.; Peto, J.; and Smith, P. G. (1976). Design and Analysis of Randomized Clinical Trials requiring prolonged Oberservation of each Patient, I: Introduction and Design. *British J. of Cancer* **34**, 585-612.

Popper, K.R. (1959). The Logic of Scientific Discovery. *Martino Fine Books (Reprint, 2014).*

Popper, K.R.; and Miller, D.W. (1983). A proof of the impossibility of inductive probability. *Nature* **302**, 687-688.

Rissanen, J. (2007). Information and Complexity in Statistical Modelling. *Springer, New York.*

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* **41**, 1397-1409.

Robert, C.P. (2007). The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation. (2nd ed.) *Springer, Berlin.*

Rosnow, R.L.; and Rosenthal, R. (1989). Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist* **44**, 1276-1284.

Royall, R. M. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *American Statistician* **40(4)**, 313-315.

Royall, R.M. (1991). Ethics and Statistics in Randomized Clinical Trials. *Stat. Sci.* **6**(1), 52-88.

Royall, R. M. (1997). Statistical Evidence. A Likelihood Paradigm. *Chapman & Hall, London.*

Royall, R. M. (2000). On the Probability of Observing Misleading Statistical Evidence (with discussion) *J. of the American Statistical Association* **95**, 760-780.

Salmon, W.C. (1989). Four Decades of Scientific Explanation. *University of Minnesota Press, Minnesota, MN.*

Salsburg, D.S. (1985). The Religion of Statistics as practiced in Medical Journals. *The American Statistician* **39(3)**, 220-223.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative konowledge in psychology. *American Psychologist* **47(10)**, 1173-1181.

Schmidt, F. L. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psych. Meth.* **1(2)**, 115-129.

Spanos, A. (2014). Recurring controversies about P values and confidence intervals revisited. *Ecology* **95**, 645-651.

Spielman, S. (1974). The Logic of Tests of Significance. *Phil. of Science* **41(3)**, 211-226.

Stigler, S.M. (1986). The History of Statistics. The Measurement of Uncertainty before 1900. *The Belknap Press of Harvard University Press, Cambridge, MA.*

Tukey, J.W. (1977). Exploratory Data Analysis. *Addison-Wesley: Behavioral Science; Quantitative Methods.*

Tukey, J.W. (1991). The Philosophy of Multiple Comparisons. *Stat. Sci.* **6(1)**, 100-116.

Tversky, A.; and Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin* **76**, 105-110.

van Dyk, D.A. (2014). The role of statistics in the discovery of a Higgs Boson *Annu. Rev. Stat. Appl.* **1**, 41-59.

Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90(2)**, 482-488.

Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Stat.* **32(5)**, 2028-2043.

