Support Functions

Reframing Conventional Statistical Methods

Mikel Aickin
Minneapolis, Minnesota, USA
mikelaickin@gmail.com
Version 18 Oct 2025

Abstract. Conventional statistical inference is awkward, which has the consequence that it is difficult to explain and even difficult to use properly. The problem is not with its fundamental elements (confidence intervals and hypothesis tests), but instead with how they are packaged. Here support functions are offered as an alternative package, which very much improves real-world practice and provides conceptual clarity to the process of scientific inference.

Introduction

The idea that probability can describe real world events goes back to Girolamo Cardano, who wrote a little book for gamblers in the 1520's. For about 200 years it remained an esoteric topic for natural philosophers, until astronomers found that they could use probability to explain the unwanted variability that they inevitably saw in their measurements of star and planet positions. But the modern era of probability began in 1837 with Poisson's book on the probabilities of judgments, in which he invented confidence intervals for binomial proportions and differences between binomial proportions. Curiously, although the 19th century was an explosion of progress in science generally, it was a statistical desert. The first rains fell in 1895 and again in 1900 when Karl Pearson wrote about "non-normal" distributions in biology, and proposed the chi-square test. This initiated the rapid development of modern statistics in the first quarter of the 20th century.

Based on his work at the Rothamsted Experimental Station in the early 1920's, R.A. Fisher promoted the null hypothesis test, which he pushed

forward very strongly as the fundamental method of statistics in his 1934 book. Due to his compelling arguments, and perhaps also his disagreeable personality, statisticians were persuaded to adopt it as the defining contribution of their discipline to empirical science. It was soon accepted across a wide swath of data-oriented fields, and it is even today arguably the dominant method of statistics.

In 1934 and again in 1937, Jerzy Neyman re-invented the confidence interval, evidently ignorant of Poisson's priority. Neyman's second article was remarkable; in one publication he presented essentially the complete modern theory of confidence intervals. His personality was, however, just short of Fisher's level of disagreeability, and so his method came to take second place in the statistical armamentarium.

So the situation in the mid-20th century was that data analysts had two basic procedures to choose from; confidence intervals and null hypothesis tests. Near the end of this period the American Statistical Association began its series of complaints about how the craft of statistical analysis was too seldom used, and when it was employed, it was too frequently abused. For more than half a century the continuing claim was that the fault was in statistics education and the quality of professional presentations. It was not seriously considered to be a problem of how statistical methods had come to be packaged. Only recently has it been entertained by some in the statistics profession that packaging might be the main problem.

The purpose of this note is to suggest that the concept of *support functions* provides a potential solution. It is based on the observation that confidence intervals and hypothesis tests are not separate methods, but two sides of the same coin. That this is true has been known at least since Allan Birnbaum's 1961 article, in which he showed that both methods could be expressed with one curve. He did not, however, see the curve as doing anything more than showing that the two methods were one, and so he proposed no re-packaging of statistical inference. Consequently, his article had no influence, and it was not until Charles Poole re-introduced the idea in 1987 that it was seriously proposed as an alternative. But despite arguing that using his curves would provide an improved method of inference, Poole gave no interpretational metaphor for what he suggested. Part of the purpose of this note is to offer *empirical support* as the missing metaphor.

Support Functions

The setting is the usual one for statistical inference. A collection of probability distributions is proposed for the description of the mechanism producing observed values. The general notation is $\mathrm{pr}(A{:}\theta)$, the

probability that observed variables will fall into a set A, assuming that θ is the true value of a parameter which determines the distribution. The concept of θ as an unknown variable permits the consideration of a collection of possible distributions for the observations, with the tacit assumption that one of them is correct.

The probably distributions relate outcome sets A to the parameter θ, and the inverse problem is to use sample observations to make inferential statements about the true θ. To do this, a *support function* has the following characteristics:

1. It is a function defined over the parameter space which assumes values between 0 and 1.
2. It is defined in terms of a sample, or more frequently in terms of a small number of statistics computed from a sample.
3. At a particular parameter value, it is the p-value for a test of the hypothesis that the parameter value is the true one.
4. The set of parameter values for which the function is above a cutoff is a confidence interval.

The interpretation of a value s at the parameter value θ is that it is the *support* which the sample gives to the parameter. A support of zero means that the sample effectively rules-out the parameter value. Small supports mean that there is very little support for the parameter value. A support of one means that the sample gives the maximum possible support to the parameter value, and values near one mean that the parameter value is very well-supported by the sample.

Based on the duality of confidence intervals and hypothesis tests, the support of s for a parameter value θ has two characteristics:

1. s is the p-value of the test of the hypothesis that θ is the true parameter value.
2. θ is on the boundary of a 1-s confidence interval for the parameter.

The second property is the one that gives meaning to the term “support”. The support which a sample gives to a parameter value is one minus the maximum confidence with which one can rule it out. That is, one interprets the confidence coefficient of a confidence interval as the confidence with which one can rule-out all parameter values which it does not contain. The support for a parameter value is then the complement of its maximal rule-out confidence.

If θ just lies on the boundary of a 0.95 confidence interval, then it can be ruled-out with confidence 0.95, and so it has support 0.05. Due to the duality of confidence intervals and hypothesis tests, it also has a p-value of 0.05. On the other hand, if θ just lies on the boundary of a 0.05 confidence interval, then it can *at most* be ruled-out with confidence 0.05,

and so it has support 0.95. In the first case there is a convincing argument that θ could be ruled-out, but in the second case the strongest argument against the parameter value is only very weak.

A typical support function is shown in Fig. 1. For each value s on the vertical axis, the interval containing values with support above s forms a 1-s confidence interval. For each parameter value θ the value of the support is its p-value.

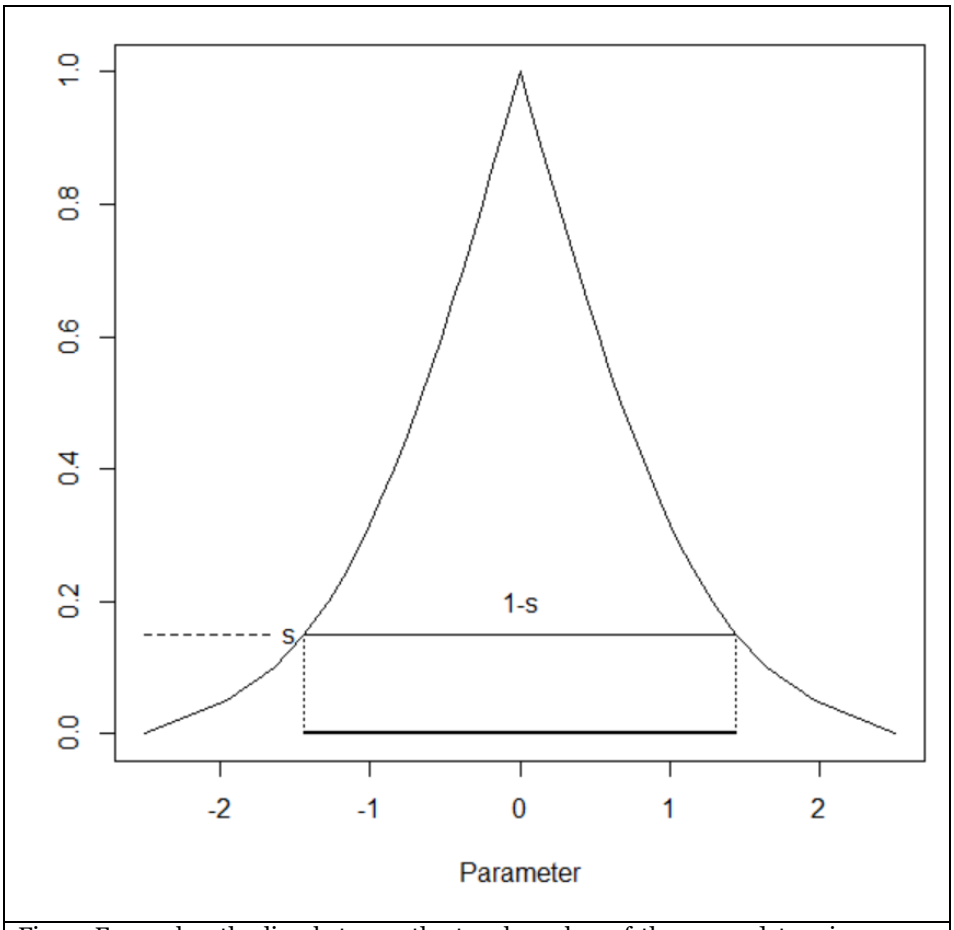

Fig. 1. For each s the line between the two branches of the curve determines a 1-s confidence interval (the thick line).

The form of Fig. 1 seems strange because we are not used to seeing it. Its justification is that it is based on the twin concepts of confidence intervals and hypothesis tests, which we are used to seeing. In fact, Francis Galton was the first to realize that this was an alternative form for

representing probability distributions. He presented it in 1885, but it did not catch on.

Support functions are easy to compute in the modern era. All that is required in a given situation is a computer program that will either test any parametric hypothesis, or will compute confidence intervals for any confidence coefficient. In the first case there must be a p-value for each parameter value, and in the second case an interval for any possible confidence coefficient. Such programs are very widely available.

Extensions

The basic logic of support allows support functions to be extended beyond single parameter values. Given a subset T of the parameter space, if supp(θ:x) is the support for θ from a sample x, then the support for T is

$$\mathrm{supp}(T: x) = \max\{\mathrm{supp}(\theta : x) : \theta \in T\}$$

The support for a parameter subset is the maximum (actually supremum) of the supports of the values it contains. This follows since the support for the set should be at least as large as the support for any parameter it contains, and there is no reason for it to be any larger. This has the advantage that (when T is closed) there is an actual member of T which has the maximal support.

Given a collection of supports $\mathrm{supp}(\theta_i : x_i)$ the support for all of them simultaneously is

$$\mathrm{supp}(\theta_1, \theta_2, \ldots : x_1, x_2, \ldots) = \min\{\mathrm{supp}(\theta_i : x_i)\}$$

Again the logic dictates that the support should not be any less than the minimum, and there is no reason why it should be larger. It is worth pointing out that this does not depend on the joint probability distribution of the observations x_i, and thus it completely resolves the historically thorny multiple-testing problem.

If each of the θ_i parameters could actually be the same parameter, then the support for the contention that they are equal to θ is

$$\mathrm{supp}(\theta, \theta, \ldots : x_1, x_2, \ldots)$$

from the previous computation, and thus the minimum of all the individual support functions at θ. The maximum of this expression over all θ is the support that the true parameters θ_i are all the same.

An Example

One of the most famous experiments in physics was instigated by the astronomer royal Frank Dyson in 1919. In order to keep his friend and colleague Arthur Eddington out of the military draft in WWI, he obtained funding for an expedition to measure the bending of starlight near the rim of the sun during an eclipse. The nominal reason for the expedition was the belief that the bending angle following from Einstein's recently-announced theory of general relativity was twice that predicted from Newton's theory of light. Twin forays to Solari in Brazil and Principe (an island off the west coast of Africa) were to get the measurements that would settle the Einstein-Newton contest.

There were altogether photographic plates from three telescopes; two astrographics (state of the art), one at each site (SA and PA), and a small 4-inch telescope at Solari (S). Getting accurate bending angles was an exceedingly difficult task, and there is considerable controversy about whether Eddington was actually able to do it. The results were published at the end of 1919, and they were declared to confirm Einstein and reject Newton. This was the commencement of Einstein's otherwise inexplicable rise to fame.

Dyson and Eddington published some results from individual stars, although they did so somewhat vaguely, in remarkable distinction to the sprawling tables they presented as arguments for the intricate data adjustments that were required to get any results at all. From them we can compute support functions, based on conventional t-tests, shown in Fig. 2. We can plainly see from this figure how disconcerting the results were. The SA telescope completely supported Newton, the PA results supported values between Newton and Einstein, but ruled out Newton while giving decent support to Einstein. The S results supported values higher than Einstein's, giving almost no support to him while firmly ruling out Newton. It was a subjective judgment on the part of Eddington to claim that the S telescope provided the definitive results, and on this basis to declare Einstein the winner. Clearly this would not have been possible if he had had support functions available. He was, in fact, only able to do it by having cast the issue as Einstein vs. Newton, and not allowing any consideration that they might both be wrong. This was a fundamental inferential error, the form of which continues in the present day because hypothesis testing makes it an easy one to commit.

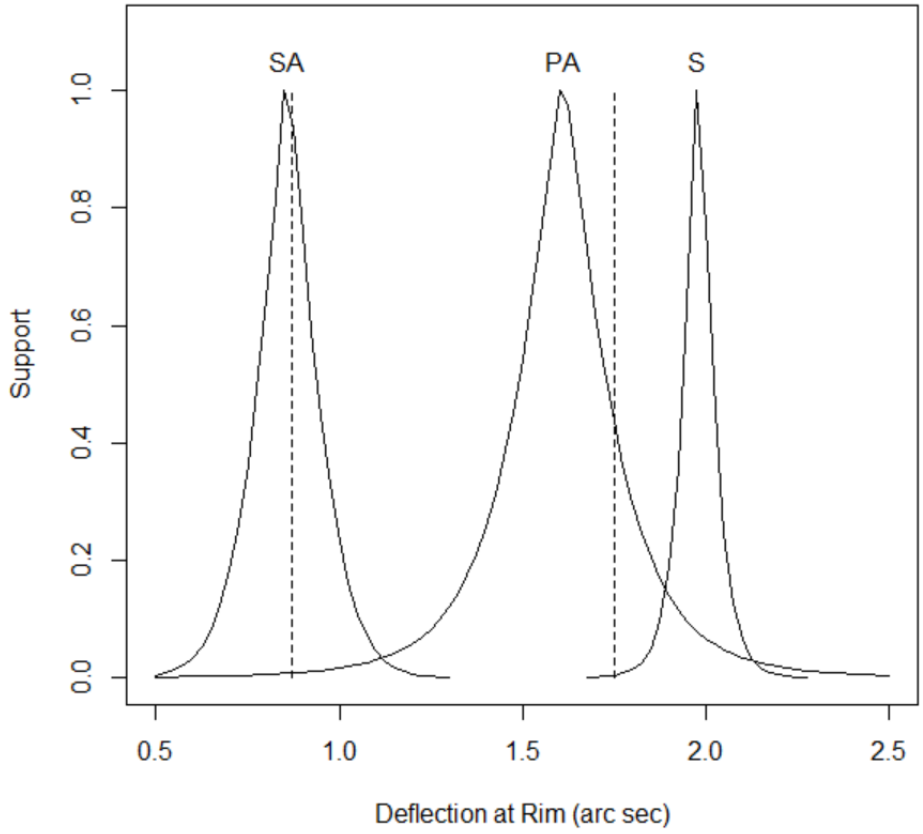

Fig. 2. Support functions for starlight bending from the three expedition telescopes, Sobral 4-inch (S), Sobral astrographic (SA) and Principe astrographic (PA). The left dashed line is Newton's prediction, and the right dashed line is Einstein's prediction.

There were multiple attempts to repeat the Eddington experiment in the half century afterwards, which were not collected for joint publication until Goldoni and Stefanini in 2020. Fig. 3 shows support functions for their Eddington data, together with support based on eight subsequent repetitions. The obvious implication is that the later studies support values at and above what Eddington found, and that they effectively rule-out Einstein (and Newton even moreso). Physicists generally claim that the subsequent data confirm Eddington, which is true, but they avoid making the observation that the data eliminate Einstein. They can do this because it is not standard practice to display a support function, so that the more obvious conclusion is hidden. They can also avoid dealing with Fig. 2, thereby escaping an even more obvious conclusion – that starlight bending experiments provide at best flimsy evidence for distinguishing Einstein from Newton. This raises the issue of whether experiments with

significant flaws should be relied on for deciding fundamental issues in physics.

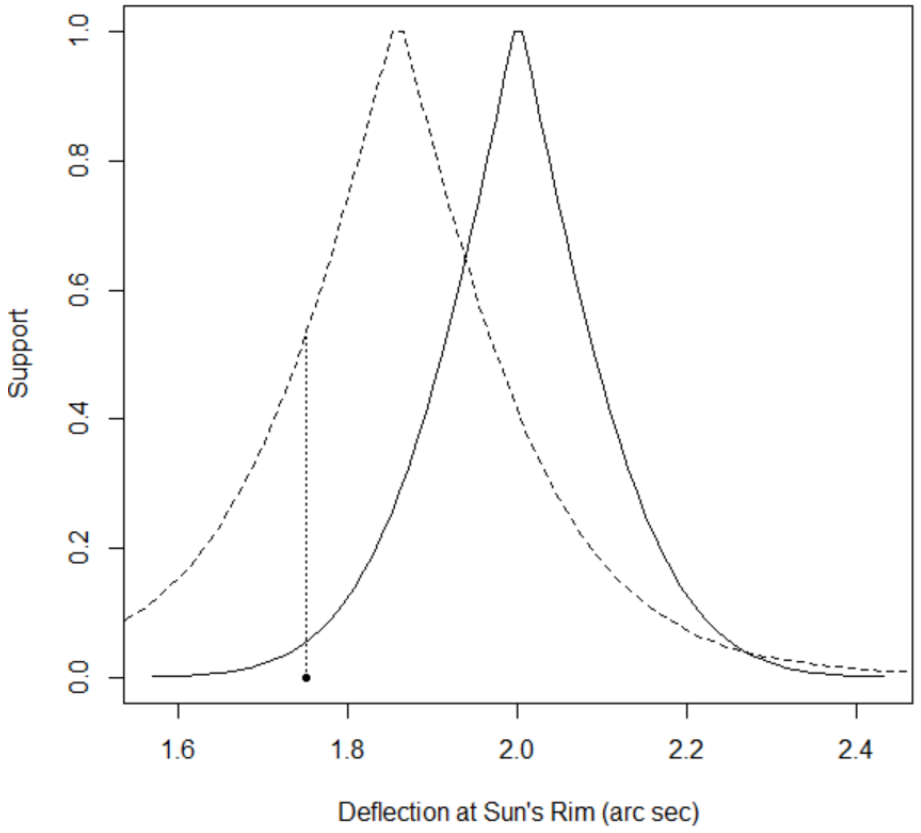

Fig. 3. Support from 8 starlight-bending experiments (solid) compared with Eddington's support function (dashed) as given by Goldoni & Stefanini.

Fig. 3 illustrates the combination of support functions for what may be the same parameter. If we assume that all the experiments pertain to the same parameter value, then the minimum of the two support functions is the sub-support function for their common value. Thus combining Eddington with subsequent experiments gives highest support (about 0.63) to values around 1.87, and effectively rules out Einstein.

Fig. 3 also illustrates one of the most pernicious side-effects of null hypothesis testing. It is conventional to test a null hypothesis (such as two medical treatments being equally effective) by seeing whether it is confirmed or rejected, *without considering any alternatives*. This leads to the practice of confirming equal treatment effectiveness when in fact considerably important effects have as much support as the null value. It

also leads to the reverse error of rejecting treatment equality to favor one of the treatments, when the largest reasonably supported beneficial effect of the winning treatment is clinically negligible. In Fig. 3 we could say that Eddington's data give support 0.50 to Einstein, but then we would be required to observe that they also give support 0.50 to about 1.95, rather far from Einstein's value. The conclusion from Eddington's data would then be consistent with the later repetitions; his results did not have enough precision to decisively confirm Einstein.

A Default Support Function

Even in a specific situation, support functions are not uniquely defined. All that is required is a complete set of confidence intervals (one for each confidence coefficient, and so that they are nested) or a complete set of hypothesis tests (a p-value for every parameter value). This suggests that there can be a choice among multiple support functions, with no issue of there being a "correct" one. This section derives one approach that may be attractive enough to be the default.

Start with the case of a real-valued statistic x, and suppose that for each possible θ we know the CDF of x when θ is true, which I denote $F(x{:}\theta)$. For now, think of both x and θ as being variable.

Fixing θ, the interval of x values for which

$$F(x{:}\,\theta) \geq \frac{s}{2} \quad 1 - F(x{:}\,\theta) \geq \frac{s}{2}$$

is a 1-s central probability interval. The values of θ for which the above inequalities hold, for the observed x, is a 1-s confidence interval. Using the fact that s will be the support for θ when the latter is on the boundary of the 1-s confidence interval, we must have either $2F(x{:}\,\theta) = s$ or $2\big(1 - F(x{:}\,\theta)\big) = s$. For reasons that will become evident in a second, I express this as

$$s = \min\left\{\frac{F(x{:}\,\theta)}{1/2}, \frac{1 - F(x{:}\,\theta)}{1/2}\right\}$$

This shows how to compute the support s for any value of θ.

I now want to go beyond this practical example to give a more theoretical argument. To do this, I define the *Galton function* as

$$\Gamma(a, b) = \min\left\{\frac{a}{b}, \frac{1 - a}{1 - b}\right\}$$

for a and b in the unit line. This was the transformation of a CDF that Galton published in 1885. Let u have the uniform distribution on the unit line. Then

$$\Gamma(u, b) \geq s$$

is the same as

$$u \geq sb \text{ and } 1 - u \geq s(1 - b), \text{the second being } u \leq 1 - s + sb$$

The probability of these conditions is thus 1-s.

To apply this, if the CDF's are continuous and strictly increasing, then $F(x{:}\theta)$ has a uniform distribution when θ is true, and so the θ's for which $\Gamma(F(x{:}\theta), 1/2) \geq s$ form a 1-s confidence interval, and so finally the support for any particular θ is $\Gamma(F(x{:}\theta), 1/2)$. This reproduces the argument given above, but the point is that it very greatly generalizes.

A more general result depends on there being a $g(x,\theta)$ with the property that if θ is the true value, then its CDF is known to be $F(g{:}\theta)$. Again assuming continuity and strict increasingness, this is all that is required for $F(g(x,\theta){:}\theta)$ to be uniformly distributed, and this in turn implies that support is given by

$$s = \Gamma(F(g(x, \theta){:}\, \theta), F(\gamma(\theta){:}\, \theta))$$

where $\gamma(\theta)$ is what $g(x,\theta)$ estimates when θ is true, usually taken to be its expected value, $\gamma(\theta)=E[g(x,\theta){:}\theta]$

Historically, something like $g(x,\theta)$ was to have a distribution that did not depending on θ, when θ was true. It was called a "pivotal quantity", and its advantage was that only one table was necessary for its distribution. Most of the currently common examples are of this form. Tables are now irrelevant, which is why the generalization is useful. I would refer to this as a *generalized pivotal quantity*, and its use as above to obtain support would be a reasonable default computation. Again, I point out that other sensible computations are often possible.

For an example, if the model is that x is the average of a Normally distributed sample, with mean μ and standard deviation σ, then $\sqrt{n}(x\text{-}\mu)/s$ is a pivotal quantity (in the historical sense) where s is the sample standard deviation and n is the sample size, and its distribution is a t-distribution $F(x{:}n\text{-}1)$, with df=n-1. Then $\gamma(\mu, \sigma) = 0$ and $F(0{:}n\text{-}1)=1/2$, so that support for μ is

$$s = \Gamma\left(F(\sqrt{n}(x - \mu)/s{:}\, n - 1), 1/2\right)$$

For cases with large sample sizes, the t-distributions converge to the standard normal distribution, F_N. This happens not only for averages, but for many other kinds of estimates. In such cases, all that is needed is the

estimate x, and an estimate of its standard deviation (the standard deviation of the estimate, or SDE). This takes different forms in different cases, and is usually produced by computer routines. With s as the SDE, the support function becomes

$$s = \Gamma(F_N((x-\mu)/s),1/2)$$

This sort of approximation is used very frequently in statistics.

The important point here is that a default support function is often easy to construct. A considerable portion of any basic course in probability involves finding generalized pivotal quantities (although they are not called that) and their expectations, the critical issue being that the relevant CDF's are also presented. Thus conventional textbooks often contain all of the ingredients for making a support function for a one-dimensional parameter. The only step they do not take is the application of the Galton function and its interpretation in terms of support.

The above argument is oriented toward continuous, strictly increasing CDF's. This seems to leave out discrete cases. In these cases I would replace both $F(g(x,\theta):\theta)$ and especially $F(\gamma(\theta):\theta)$ with smoothed, continuous and increasing versions. This may seem questionable, but if done well it is quite accurate. In any case, procedures that use the actual, discrete CDF's tend to become complicated, and the usual practice is to smooth the results at the last step. I find it easier and more convincing to smooth at the first step.

Summary

The problem with conventional statistical inference is not that its fundamentals are somehow wrong, but that their packaging is suboptimal. Since confidence intervals and hypothesis tests are mirror-images of the same process, it seems sensible to package them as one inferential object, and this is the support function.

There is a sensible interpretation of support; a value is supported when the strongest argument against it is weak, and it is not supported when there is a strong argument against it. The only thing we can expect from an inferential principle is that it should tell us how much support a sample gives to various parameter values. Resting perhaps somewhat on our usual interpretation of "support", the concept could be relatively easily explained in the classroom. Support functions are easy to compute with modern software, and when it is lacking there is an attractive default version.

In 1987 Charles Poole made a prophecy; that if one consistently used support functions, one's inferences would differ significantly from the use

of conventional methods. In two decades of following his advice I have not found him to be mistaken.

References

Birnbaum A. Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association* 1961; 56(294): 246-249

Cardano G. *The Book on Games of Chance*. Sydney Henry Gould, trans. Holt, Rinehard and Winston: New York NY. 1961. Reprinted from Ore O. Cardano: *The Gambling Scholar.* Princeton University Press: Princeton NJ, 1953

Fisher RA. *Statistical Methods for Research Workers*. Oliver & Boyd: Edinburgh UK, Fifth Edition, 1934

Galton F. Application of a graphic method to fallible measures. *Jubilee of the Statistical Society*. Edward Stanford: London UK, 1885, pp. 262-265

Goldoni E, Stefanini L. A century of light-bending measurements: Bringing solar eclipses into the classroom. arXiv:2002.01179v1, 2020.

Neyman J. On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 1934;**97(4)**: 558-625.

Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A* 1937; **236**: 333-380.

Pearson K. Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London 1895; 186: 343-414

Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 5th Series 1900;Vol. L: 157–175

Poisson SD. *Recherches sur la Probabilité des Jugements*. Bachelier: Paris, France, 1837

Poole C. (1987) Beyond the confidence interval. *American Journal of Public Health* 1987;**77(2)**: 195-199.