On Frequentist Testing: revisiting widely held confusions and misinterpretations

Aris Spanos, Virginia Tech

After reading chapter 13.2 of the 2022 book "**Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives**", by Michael G. Titelbaum, I decided to write a few comments relating to his discussion in an attempt to delineate certain key concepts in frequentist testing with a view to shed light on several long-standing confusions and misinterpretations of these testing procedures. The key concepts include 'what is a frequentist test', 'what is a test statistic and how it is chosen', and 'how the hypotheses of interest are framed'.

The first thing that needs to be brought out immediately is that frequentist testing is framed in terms of applied mathematics and its proper framing is of utmost importance in forefending confusions and misinterpretations. Regrettably, philosophers of science tend to undervalue that formalism as cumbersome and needless, which rings hollow when one compares the statistical framing with that of first order logic and other parts of epistemology!

To understand the frequentist testing one needs to place it in its proper context which is a particular **statistical model** comprising several probabilistic assumptions whose validity of the particular data is of paramount importance.

The simplest statistical model is **a simple Bernoulli model** denoted by:

$$X_k \backsim \mathsf{BerIID}(\theta, \theta(1-\theta)),\ \ 0<\theta<1,\ \ x_k=0,1,\ k=1,2,...n,..., \tag{1}$$

where 'BerIID' stands for Bernoulli (Ber), Independent and Identically Distributed (IID). The underlying random variable X takes only two values, $X=1$, say Head (H) and $X=0$ Tails (T), with $\mathbb{P}(X=1)=\theta$ and $\mathbb{P}(X=0)=1-\theta$. In the context of the statistical model in (1), the relevant data $\mathbf{x}_0:=(x_1, x_2, ..., x_n)$ are viewed as a single realization of the *sample* $\mathbf{X}:=(X_1, X_2, ..., X_n)$. Note that a random variable is denoted by a capital letter (X_k) and the corresponding observation by a small letter (x_k); this minor notational point will avert numerous confusions in practice!

The first issue that arises in practice is how to frame the hypothesis of interest in frequentist testing. This arises because there are a number of confusions between R.A. Fisher's (1922) framing of the null hypothesis (H_0) and the Neyman-Pearson (N-P) (1933) framing that includes both the null (H_0) and the alternative (H_1) hypothesis. The truth is that the objective is identical for both framings: **learn from data $\mathbf{x}_0:=(x_1, x_2, ..., x_n)$ about the 'true' value, say θ^*, of the unknown parameter** θ. In light of that, the entire parameter space $(0, 1)$ is relevant for frequentist testing since theoretically θ^* can take any one of the values in this interval. Hence, the framing of the hypotheses needs to cover the whole of the parameter space. This was first stated in the Neyman-Pearson (N-P) lemma that provided the cornerstone of frequentist testing and included Fisher's framing as a special case; see Note 1 below.

In summary, the proper way to specify the hypotheses of interest in frequentist testing is to framed them in terms of the model's unknown parameter(s) θ, and ensure that they constitute a partition of the parameter space. For the statistical model in (1), partitioning can take various forms, including:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0, \text{ where } \theta_0{=}.5. \tag{2}$$

What is a frequentist test? It is not just a statistic, say $\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k$, whose distribution in this case is Binomial (Bin):

$$\overline{X}_n \backsim \mathsf{Bin}(\theta, \tfrac{\theta(1-\theta)}{n}),\ 0{<}\theta{<}1, \tag{3}$$

derived by assuming the validity of the statistical model assumptions 'BerIID'. Any attempt to use (3) to define any error probabilities, including the p-value (Titelbaum (2022), p. 465), is improper and will give rise to the wrong inference.

A frequentist test comprises two equally important components. For the hypotheses in (2), the optimal N-P test takes the form:

$$\text{(a) test statistic:} d(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}, \quad \text{(b) rejection region:} C_1(\alpha){=}\{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\}. \tag{4}$$

In frequentist testing, the test statistic is always a distance function $d(\mathbf{X})$ framed in terms of a statistic and a frequentist test also includes a rejection region whose choices are neither arbitrary or whimsical. Indeed, the choice of $d(\mathbf{X})$ and $C_1(\alpha)$ is interrelated and needs to satisfy two conditions. The *first* is that the distribution of $d(\mathbf{X})$ can be evaluated under both H_0 and H_1, and the *second* is that it has to give rise to an optimal test in terms of learning from data about θ^*, and there are good choices for $d(\mathbf{X})$ and $C_1(\alpha)$ and bad ones, which are evaluated in terms of their capacity to approximate θ^* framed in terms of their pre-data type I and II error probabilities.

In the case of the test in (4), the N-P optimal theory of testing renders it Uniformly Most Powerful (UMP) whose relevant sampling distributions are:

$$d(\mathbf{X}) \overset{\theta=\theta_0}{\approx} \mathsf{N}(0,1), \quad d(\mathbf{X}) \overset{\theta=\theta_1}{\approx} \mathsf{N}(\delta_1, 1),\ \delta_1{=}\frac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}},\ \text{for all } \theta_1{>}\theta_0, \tag{5}$$

where '$\approx$' indicates an approximation of the scaled Binomial by the Normal distribution; see Spanos (2019).

Example 1. Consider particular data $\mathbf{x}_0$ representing 17 Hs out of $n{=}20$ flips (Titelbaum, 2022, p. 465):

$$\mathsf{HHHTHHHHHTHHHHTHHHHH} \tag{6}$$

This implies that $\overline{x}_n{=}\frac{17}{20}{=}.85$, and thus the observed test statistic yields:

$$d(\mathbf{x}_0){=}\frac{\sqrt{n}(\overline{x}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}{=}\frac{\sqrt{20}(.85-.5)}{\sqrt{.5(1-.5)}}{=}3.131 \tag{7}$$

The evaluation of the p-value yields:

$$p(\mathbf{x}_0){=}\mathbb{P}(\mathbf{x}\text{: } d(\mathbf{X}){>}d(\mathbf{x}_0); \theta{=}\theta_0){=}\mathbb{P}(\mathbf{x}\text{: } d(\mathbf{X}){>}3.131; \theta{=}.5){=}.00087. \tag{8}$$

In light of the fact that this p-value is based on $n{=}20$ observations, this indicates a clear departure from $\theta_0{=}.5$. It is important to emphasize that the p-value in (8) is

NOT the probability of the particular configuration in (6) occurring as a realization of the sample!

The question that arises at this stage is '**what about the validity of the probabilistic assumptions comprising the invoked statistical model in (1) for the particular data $\mathbf{x}_0$**?' If any of these assumptions are invalid for $\mathbf{x}_0$ the above inference results are unreliable! The Bernoulli assumption is innocuous since any data with two outcomes can always be framed with such a distribution. The IID assumptions, however, are often invalid with real data. When IID is invalid, the assumed distributions under both H_0 and H_1 in (5) are invalid, inducing sizeable discrepancies between the **actual error probabilities** and the **nominal** ones derived by assuming the validity of IID! This renders any evaluations based on their tail areas highly misleading; see Spanos (2019), ch. 15. Hence, in practice one needs to test the IID assumptions before using $\mathbf{x}_0$ in the context of the statistical model in question to draw inferences about θ^*.

Example 2. Let us return to the particular data $\mathbf{x}_0$ representing 17 Hs out of n=20 flips, and ask the question: Are the IID assumptions valid for data $\mathbf{x}_0$? Although there are many ways to test the IID assumptions (Spanos, 2019, ch.15), a particularly simple misspecification test is **the runs test**. A 'run' is a segment of the sequence of outcomes consisting of adjacent identical elements which are followed and proceeded by a different symbol. For the observed sequence in (6), the sequence of 7 runs is shown below:

$$\underbrace{\mathsf{HHH}}_{1}\,\underbrace{\mathsf{T}}_{2}\,\underbrace{\mathsf{HHHHH}}_{3}\,\underbrace{\mathsf{T}}_{4}\,\underbrace{\mathsf{HHHH}}_{5}\,\underbrace{\mathsf{T}}_{6}\,\underbrace{\mathsf{HHHHH}}_{7} \tag{9}$$

The runs test compares the *actual* number of runs R with the number of *expected* runs $E(R)$ – assuming that the sample is IID process – to construct **the runs test**:

$$d_R(\mathbf{X})=\tfrac{[E(R)-R]}{\sqrt{Var(R)}},\;\; C_1(\alpha)=\{\mathbf{x}\text{: } |d_R(\mathbf{x})|>c_{\frac{\alpha}{2}}\},\;\; E(R)=\tfrac{2n-1}{3},\;\; Var(R)=\tfrac{16n-29}{90}. \tag{10}$$

One can show that the distribution of $d_R(\mathbf{X})$, for $n\geq 40$, can be approximated by:

$$d_R(\mathbf{X})=[E(R)-R]\,/\sqrt{Var(R)}\overset{\mathsf{IID}}{\approx}\mathsf{N}(0,1).$$

Note that the above runs test can be extended to a more sophisticated test which accounts, not only for the number of runs, but also their different lengths, e.g. run 1 has length 3, run 2 has length 1 and run 3 has length 5.

In this case the sample size n=20, is rather small, but we can apply the test anyway for illustration purposes:

$$d_R(\mathbf{x}_0)=\tfrac{(13-7)}{3.233}=1.856,\quad p_R(\mathbf{x}_0)=\mathbb{P}(\mathbf{x}\text{: } |d_R(\mathbf{x})|>1.856)=.0317,$$

where the p-value indicates departures from the IID assumptions at any significance level $\alpha\geq$.032. One can dispute this particular threshold α, but in light of the small sample size n=20, this threshold ensures that the test has sufficient power to detect departures from IID. On the issue of why the p-value should always be one-sided is because the data render irrelevant one of the two tails post-data (when $d_R(\mathbf{x}_0)$ is revealed). This is the key difference between the p-value and type I and II error probabilities that are pre-data; see Spanos (2019). ch. 13.

It is important to emphasize that the runs test in (10) is probing the validity of the IID assumptions underlying the invoked statistical model in (1), and thus it poses very different questions to the data when compared to the N-P test in (5) which assumes the validity of the assumptions and probes for θ^*! Indeed, misspecification testing should not be framed in terms of the parameter(s) of the underlying statistical model because it probes outside the boundaries of the given model, as opposed to N-P testing that probes within its boundaries. In practice, misspecification testing pre-dates N-P testing to secure the validity of the invoked statistical model and thus the reliability of the ensuing inference; see Mayo and Spanos (2004).

More broadly, the accept/reject H_0 results and small/large p-values do not provide evidence for or against particular hypotheses since the sample size n in conjunction with the pre-specified α play a crucial role in transforming such results into evidence using their post-data severity evaluation that outputs the warranted discrepancy from the null value; see Mayo and Spanos (2006). For a given α, ignoring the sample size n is likely to give rise to the fallacies of acceptance and rejection. This is due to the inherent trade-off between the type I and II error probabilities, which implies that for a given α the power of the test increases with n. The post-data severity evaluation provides an evidential account of the accept/reject H_0 results by taking fully into account the sample size n; see Mayo (2018), Spanos (2023).

Note 1. The widely held impression that Fisher's significance testing and N-P testing are two very different approaches is just another misconstrual of frequentist testing. They are not so different! Stating just a point null hypothesis H_0: $\theta{=}\theta_0$ and a threshold α, Fisher brings into play the type I and II error probabilities (and power) indirectly into his significance testing. Don't take my word for this claim, read Fisher (1935), pp. 21-22 describing how the power of the test increases with the sample size n but calling it the 'sensitivity' of a test. How could one explain the acerbic Fisher vs. Neyman-Pearson exchanges? They were talking passed each other since the type I and II error probabilities are pre-data – framing the capacity of the test –, and Fisher's p-value is post-data evaluation indicating potential departures from H_0 in light of the observed test statistic. Fisher can get away without specifying an alternative hypothesis H_1 since the sign of the observed test statistic indicates the direction of departure, eliminating one of the two tails; see Spanos (2019), ch. 13. It should be noted that Fisher constructed his numerous test statistics using intuition, but they turned out to define optimal frequentist tests when supplemented by an appropriate rejection region, and Neyman and Pearson (1933) give him credit for that.

Note 2. For further discussions on frequentist testing, its misinterpretations and misuses, including the base-rate fallacy, the Jeffreys-Lindley Paradox, Akaike type selection criteria, etc., see the following papers:

Spanos, Aris (2007) "Curve-Fitting, the Reliability of Inductive Inference and the

Error-Statistical Approach," *Philosophy of Science*, 74(5): 357-381.

Spanos, Aris (2010) "Is Frequentist Testing Vulnerable to the Base-Rate Fallacy?" *Philosophy of Science*, 77: 565-583.

Spanos, Aris (2013) "Who Should Be Afraid of the Jeffreys-Lindley Paradox?" *Philosophy of Science*, 80: 73-93.

Spanos, Aris (2022) "Severity and Trustworthy Evidence: Foundational Problems versus Misuses of Frequentist Testing." *Philosophy of Science* 89(2): 378-397.

References

[1] Fisher, R.A. (1922) "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society A*, 222: 309-368.

[2] Fisher, R.A. (1935) *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[3] Mayo, Deborah G. (2018) *Statistical inference as severe testing: How to get beyond the statistics wars*, Cambridge University Press.

[4] Mayo, D.G. and A. Spanos (2004) "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science*, **71**: 1007-1025.

[5] Mayo, D.G. and A. Spanos. (2006) "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *The British Journal for the Philosophy of Science,* **57**: 323-357.

[6] Neyman, J. and E.S. Pearson (1933) "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society, A,* **231**, 289-337.

[7] Spanos, Aris (2019) *Introduction to Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*, 2nd edition, Cambridge University Press, Cambridge.

[8] Spanos, Aris (2023) "Revisiting the Large n (Sample Size) Problem: How to Avert Spurious Significance Results." *Stats* 6(4): 1323-1338.