PHIL6334/ECON6614-Lecture Notes 4: Hypothesis Testing 1: Foundational Issues

Aris Spanos [Spring 2019]

1 Introduction

Neyman and Pearson (1933) proposed a recasting of Fisher's significance testing by bringing the whole of the parameter space Θ of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ into play and partitioning it into:

$$H_0\text{: } \theta\in\Theta_0 \text{ vs. } H_0\text{: } \theta\in\Theta_1,$$

with a view to address three perceived weaknesses of Fisher's approach to testing:

(i) The ad hoc **choice of a test statistic**. How does one construct an optimal test statistic?

(ii) the **vulnerability** of the **p-value** to **abuse**. One can evaluate $p(\mathbf{x}_0)$ and then choose a threshold $p(\mathbf{x}_0) \gtrless c_0$ that yields the inference one prefers.

(iii) Fisher's **falsificationist** stance. Scientists would like to know if there is evidence *for* (not just against) a particular substantive claim.

The N-P recast framework succeeded in addressing (i). The choice of a N-P test statistic in $\mathcal{T}_\alpha{:=}\{d(\mathbf{X}), C_1(\alpha)\}$ in the context of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathfrak{X}{:=}\mathbb{R}^n_X,$$

stems from selecting a particular significance level $\alpha{=}\mathbb{P}(\mathbf{x}_0\in C_1(\alpha);\ H_0)$ and then choosing a test statistic $d(\mathbf{X})$ in conjunction with a rejection region $C_1(\alpha)$ that dominates every other possible test in terms of its power for all $\theta\in\Theta_1$ [UMP]. Note that the partitioning of the parameter space Θ, into H_0: $\theta\in\Theta_0$ vs. H_1: $\theta\in\Theta_1$, corresponds to the partitioning of the sample space $\mathbb{R}^n_X$ into an acceptance ($C_0(\alpha)$) and a rejection region ($C_1(\alpha)$). In this sense, Neyman and Pearson (1933) did put forward a theory of optimal testing to supplement Fisher's theory of optimal estimation.

In relation to (ii) and (iii), Neyman and Pearson replaced Fisher's post-data p-value rejection of H_0 based on a threshold c:

$$\text{if } p(\mathbf{x}_0){=}\mathbb{P}(d(\mathbf{X}){>}d(\mathbf{x}_0);\ H_0) < c, \text{ reject } H_0,$$

with the **accept/reject** H_0 decision rules:

$$\text{[i] if } \mathbf{x}_0\in C_0(\alpha)\text{, accept } H_0, \qquad \text{[ii] if } \mathbf{x}_0\in C_1(\alpha)\text{, reject } H_0. \tag{1}$$

based on the rejection region defined in terms of a pre-data threshold stemming from the type I error probability $\alpha{=}\mathbb{P}(d(\mathbf{X}){>}c_\alpha;\ H_0)$, also known as the significance level.

The N-P recasting of frequentist testing, however, raised several foundational issues because one could not interpret the N-P results 'Accept H_0' ('Reject H_0') as data $\mathbf{x}_0$ provide evidence *for (against)* H_0 (for H_1). Why?

(i) A particular test $\mathcal{T}_{\alpha}$:=$\{d(\mathbf{X}), \mathcal{C}_1(\alpha)\}$ could have led to 'Accept H_0' because the power of the particular test to detect an existing discrepancy $\gamma \neq 0$ was very low. This can easily happen when the sample size n is small.

(ii) A particular test $\mathcal{T}_{\alpha}$:=$\{d(\mathbf{X}), \mathcal{C}_1(\alpha)\}$ could have led to 'Reject H_0' simply because the power of that test was high enough to detect 'trivial' discrepancies from H_0. This can easily happen when the sample size n is very large.

As Mayo (1996) argued, neither the Fisherian nor the N-P approach to frequentist testing offered a *coherent evidential account* that could answer unambiguously the basic question:

When do data $\mathbf{x}_0$ provide *evidence for or against* a hypothesis H (H_0 or H_1)?

In light of that major weakness, it should come as no surprise to learn that most practitioners in different fields sought such answers by trying unsuccessfully (and misleadingly) to distill an evidential interpretation out of Fisher's p-value because it is more informative as well as data-specific than the accept/reject H_0 rules; see Lehmann and Romano (2005). In their eyes the pre-designated α is equally vulnerable to manipulation [pre-designation to keep practitioners 'honest' is unenforceable in practice]. This search for an evidential interpretation gave rise to the current chaotic situation of numerous misinterpretations and abuses of the p-value, including ad hoc rejection thresholds (<.05), cherry picking, data-dredging, multiple testing and p-hacking.

A crucial problem with the p-value is that a small (large) p-value could not be interpreted as evidence for the presence (absence) of a substantive discrepancy γ for the same reasons as (i)-(ii). The power of the test affects the p-value. For instance, a very small p-value can easily arise in the case of a very large sample size n.

1.1 The large n problem

The large n problem, initially raised by Berkson (1938), is that as n increases, $p(\mathbf{x}_0)$ decreases, and thus there is always a large enough n to reject H_0, however small $(\theta^* - \theta_0) \neq 0$ and the adopted threshold $c>0$, i.e., when $(\theta^* - \theta_0) \neq 0$, $p(\mathbf{x}_0) \underset{n\rightarrow\infty}{\rightarrow} 0$. Hence, a rejection of H_0 with $p(\mathbf{x}_0)$=.03 and n=50, does not have the same evidential weight for the falsity of H_0 as a rejection with $p(\mathbf{x}_0)$=.03 and n=20000. This questions the strategy of evaluating 'significance' using $p(\mathbf{x}_0)$<.05 and ignoring n.

Example 1. Consider the **simple (**one parameter**) Normal model**:

$$\boxed{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}):\ X_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ x_k \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma^2 > 0,\ k \in \mathbb{N}:=(1, 2, ..., n, ...),} \tag{2}$$

where for simplicity we assume that σ^2 is **known**. The optimal test for the hypotheses:

$$H_0\text{: } \mu \leq \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0, \tag{3}$$

is $T_{\alpha} := \{d(\mathbf{X}), C_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_{\alpha}\}$, where $d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma}$ and $C_1(\alpha) = \{\mathbf{x}$: $d(\mathbf{x}) > c_{\alpha}\}$. The relevant sampling distribution for the type I error probability is:

$$d(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma} \overset{\mu=\mu_0}{\backsim} \mathsf{N}(0, 1),$$

and that for the power is:

$$d(\mathbf{X}){=}\tfrac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma}\overset{\mu=\mu_1}{\backsim}\mathsf{N}(\delta,1),\ \ \delta{=}\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\ \text{for all } \mu_1{>}\mu_0. \tag{4}$$

The problem arises when the p-value is detached from the particular context in (5) and is treated as providing the same evidence for a particular alternative H_1, regardless of the power of the test in question. For instance, in (4) the power increases with $\sqrt{n}$ since $\delta_1{=}\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, for all $\mu_1{\in}\Theta_1$, rendering the test more and more capable of detecting smaller and smaller discrepancies from H_0 with the same probability; see figures 1a-b. When viewed as *testing within* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, a significance test has a well-defined power function that becomes relevant when the rejection (acceptance) of H_0 with a small (large) enough p-value.

A typical graph for the power of this test is given in figure 1 with $n{=}100$.

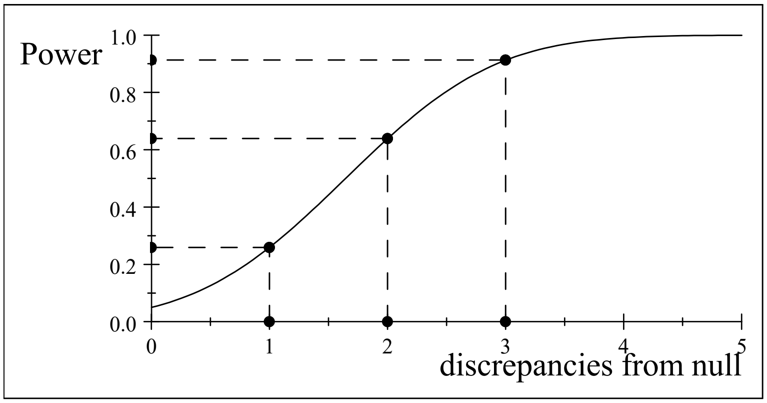

Fig. 1: Power curve

When the sample size is increased to $n{=}100,000$ (figures 1a), the power for tiny discrepancies γ from the null indicate significance since the generic capacity of the test increases dramatically: $\pi(.0075){=}.766\,30$, $\pi(.01){=}.966\,64$, $\pi(.012){=}.984$. The problem becomes worse as n increases to 1000000; see Figure 1b.

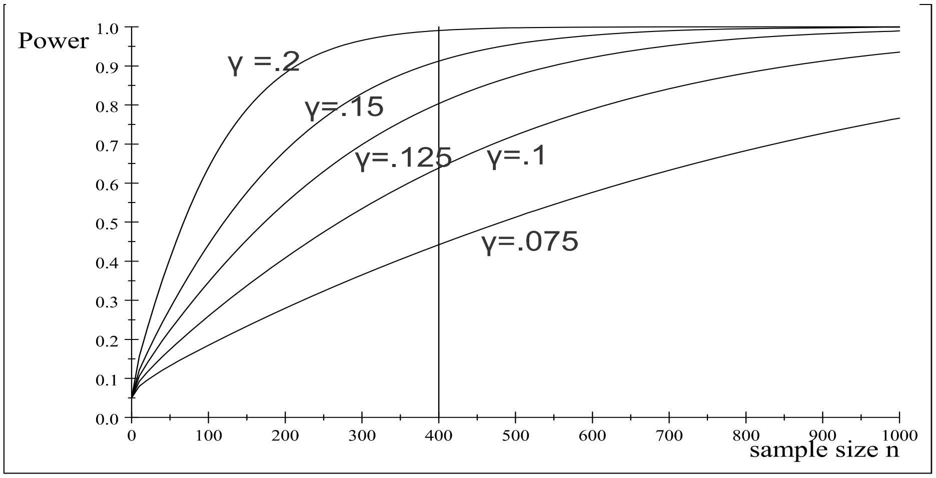

Fig. 2a: Power for different n and discrepancies $\gamma\sigma$ from the null

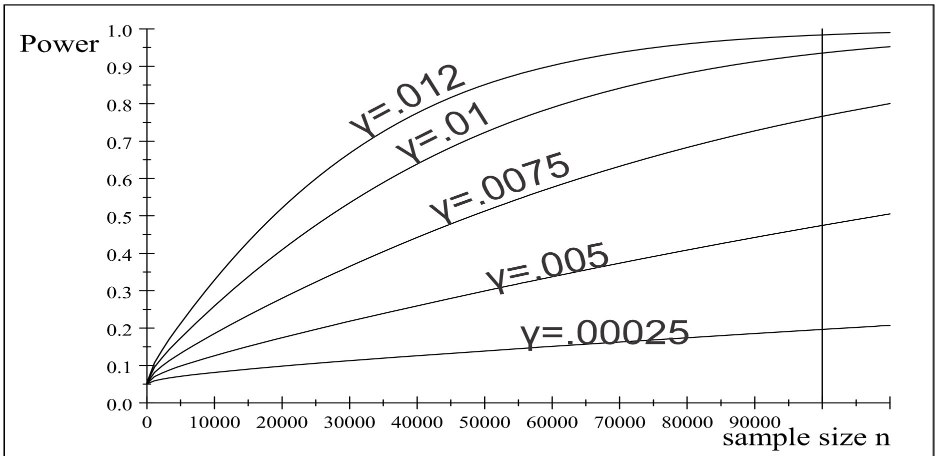

Fig. 2b: Power for different n and discrepancies $\gamma\sigma$ from the null

Don't take my word on this, here is Fisher (1935) arguing: "By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or ... quantitatively smaller departures from the null hypothesis." (pp. 21-22).

This 'sensitivity' renders a rejection of H_0 with a large n (high power) very different in *evidential terms* from a rejection of H_0 with a small n (low power). That is, the p-value and the accept/reject rules were never meant to provide evidence for or against particular hypotheses beyond the coarse accept/reject H_0; see Mayo (2018), pp. 240-5.

The practice of detaching the p-value from the *particular context*, as specified by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ H_0\colon\ \theta\in\Theta_0 \text{ vs. } H_1\colon\ \theta\in\Theta_1,\ \mathcal{T}_\alpha:=\{d(\mathbf{X}),C_1(\alpha)\} \text{ and } \mathbf{x}_0, \qquad (5)$$

and instead reporting statistical significance by attaching **asterisks** to the estimated parameters (or their standard errors):

$$(^{*})\Leftrightarrow p(\mathbf{x}_0)\leq .05,\ \ (^{**})\Leftrightarrow p(\mathbf{x}_0)\leq .01,\ \ (^{***})\Leftrightarrow p(\mathbf{x}_0)\leq .001,\ \ (^{****})\Leftrightarrow p(\mathbf{x}_0)\leq .0001,$$

is a **bad strategy** that gave rise to two main fallacies.

The fallacy of rejection: (mis)interpreting reject H_0 [evidence against H_0] as evidence *for* a particular H_1; this can easily arise when a test has high enough power (e.g. large n).

The **fallacy of acceptance**: (mis)interpreting a large p-value or accept H_0 [no evidence against H_0] as evidence for H_0; this can easily arise when a test has very low power (e.g. small n).

To counter the decrease in the p-value as n increases, some textbooks advise practitioners to use rules of thumb based on decreasing α as n increases; see Lehmann and Romano (2005). Good (1988) suggests standardizing the p-value $p(\mathbf{x}_0)$ to n=100 using the formula: $p_{100}(\mathbf{x}_0)=\min\left(.5,\left[p(\mathbf{x}_0)\cdot\sqrt{n/100}\right]\right)$, $n>40$; see table 1.

Table 1: Actual $[p_n(\mathbf{x}_0)=.01]$ vs. standardized p-value

n	50	100	500	1000	10000	100000	1000000
$p_{100}(\mathbf{x}_0)$	.007	.010	.022	.032	.100	.300	.5

Can we do better than rules of thumb?

The above fallacies arise in statistics in a variety of different guises, including the distinction between *statistical and substantive significance*. A few textbooks in statistics warn readers that one should not conflate the two, but they offer no principled way to address the problem head on. In the statistics literature, as well as in the secondary literatures in several applied fields, there have been numerous attempts to circumvent these two fallacies, but none succeeded. These fallacies can be addressed using the **post-data severity evaluation** of inference results (accept/reject, p-values) by offering an evidential account in the form of the discrepancy from the null warranted by the data; see Mayo and Spanos (2006, 2011), Mayo (2018), excursion 3, tour III, excursion 5.

2 Post-data severity evaluation

On reflection the above fallacies stem primarily from the fact that there *is* a problem when the p-value and the accept/reject H_0 results are detached from the test itself. That is, the results are viewed as providing the *same evidence* for a particular hypothesis H (H_0 or H_1), regardless of the generic capacity (the power) of the test in question to detect discrepancies from H_0. The intuition behind this reflection is that a small p-value or a rejection of H_0 based on a test with low power (e.g. a small n) for detecting a particular discrepancy γ *provides stronger evidence* for the presence of a particular discrepancy γ than using a test with much higher power (e.g. a large n). Mayo (1996) proposed a frequentist evidential account based on harnessing this intuition in the form of a **post-data severity evaluation** of the accept/reject results. This is based on custom-tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{x}_0$. This evidential account can be used to circumvent the above fallacies, as well as other charges against frequentist testing.

The severity evaluation is a post-data appraisal of the accept/reject and p-value results that revolves around the discrepancy γ from H_0 warranted by data $\mathbf{x}_0$.

■ A hypothesis H passes a *severe test* $\mathcal{T}_\alpha$ with data $\mathbf{x}_0$ if:

(S-1) $\mathbf{x}_0$ accords with H, and

(S-2) with very high probability, test $\mathcal{T}_\alpha$ would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

Severity can be viewed as an feature of a test $\mathcal{T}_\alpha$ as it relates to a particular data $\mathbf{x}_0$ *and* a specific inferential claim H being considered. Hence, the severity function has three arguments, $SEV(\mathcal{T}_\alpha, \mathbf{x}_0, H)$, denoting the severity with which H passes $\mathcal{T}_\alpha$ with $\mathbf{x}_0$; see Mayo and Spanos (2006), Mayo (2018).

Example 2. Consider the **simple Bernoulli model** specified by:

$$\boxed{\mathcal{M}_\theta(\mathbf{x}):\ X_k \backsim \mathsf{BerIID}(\theta, \theta(1-\theta)),\ x_k{=}0,1,\ \theta{\in}[0,1],\ k{=}1,2,\ldots n,...} \tag{6}$$

Arbuthnot's 1710 conjecture: the ratio of males to females in newborns might *not* be 50-50 ('fair'). This can be tested by framing it as a statistical null hypothesis in terms of θ:

$$H_0\text{: } \theta{=}\theta_0, \quad \text{where } \theta_0{=}.5 \text{ denotes 'fair'},$$

in the context of (6) based on the random variable X defined by:

$$\{X{=}1\}{=}\{\text{male}\},\ \{X{=}0\}{=}\{\text{female}\}.$$

As shown in Lecture Notes (LN) 2 $\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^{n}X_i$, is the best estimator of θ with a sampling distribution:

$$\widehat{\theta}_n{:=}\overline{X}_n \backsim \mathsf{Bin}\left(\theta, \tfrac{\theta(1-\theta)}{n}; n\right). \tag{7}$$

Using (7) one can derive the *test statistic*:

$$d(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0,1;n\right), \tag{8}$$

whose Binomial distribution can be approximated accurately using a $\mathsf{N}(0,1)$ distribution for $n > 20$.

Consider probing the Arbuthnot value $\theta_A = .5$ using the hypotheses:

$$\text{(i)} \quad H_0\text{: } \theta \leq \theta_A \text{ vs. } H_1\text{: } \theta > \theta_A,$$
$$\text{(ii)} \quad H_0\text{: } \theta = \theta_A \text{ vs. } H_1\text{: } \theta \neq \theta_A,$$

using data of $n{=}30762$ newborns during the period 1993-5 in Cyprus, 16029 boys and 14833 girls. In view of the sample size it is advisable to choose a smaller significance level, say $\alpha{=}.01 \Rightarrow c_\alpha{=}2.326,\ c_{\frac{\alpha}{2}}{=}2.575$. The test statistic based on $\theta_A{=}\theta_0{=}.5$ yields $\widehat{\theta}_n(\mathbf{x}_0){=}\frac{16029}{30762}{=}.521$:

$$d_A(\mathbf{x}_0){=}\frac{\sqrt{30762}(\frac{16029}{30762}-.5)}{\sqrt{.5(.5)}}{=}7.389,$$

with the associated p-values:

$$\text{(i) } p_{A>}(\mathbf{x}_0){=}\mathbb{P}(d_A(\mathbf{X}) > 7.389;\ H_0){=}7.4{\times}10^{-14} < \alpha{=}.01,$$
$$\text{(ii) } p_{A\neq}(\mathbf{x}_0){=}\mathbb{P}(|d_A(\mathbf{X})| {>}7.389;\ H_0){=}1.48{\times}10^{-13}{<}\tfrac{\alpha}{2}{=}.005.$$

As argued above, there is no such a thing as a two-sided p-value, irrespective of whether the N-P formulation of the alternative hypothesis is two-sided, and thus the only relevant p-value is $p_{A>}(\mathbf{x}_0)$.

An important feature of the severity evaluation is that it is *post-data*, and thus the sign of the observed test statistic $d(\mathbf{x}_0)$ provides information that indicates the *directional* inferential claims that 'passed'. In relation to the above example, the severity 'accordance' condition (S-1) implies that the rejection of $\theta_0{=}.5$ with $d(\mathbf{x}_0){=}7.389{>}0$, indicates that the form of the inferential claim that 'passed' is of the generic form:

$$\theta > \theta_1{=}\theta_0{+}\gamma, \quad \text{for some } \gamma \geq 0. \tag{9}$$

To establish the particular discrepancy γ warranted by data $\mathbf{x}_0$, the severity post-data 'discordance' condition (S-2) calls for evaluating the probability of the event: "outcomes $\mathbf{x}$ that accord less well with $\theta{>}\theta_1$ than $\mathbf{x}_0$ does", i.e. $[\mathbf{x}\text{: } d(\mathbf{x}) \leq d(\mathbf{x}_0)]$:

$$SEV(T_\alpha^{>}; \theta > \theta_1){=}\mathbb{P}(\mathbf{x}\text{: } d_A(\mathbf{x}) \leq d_A(\mathbf{x}_0); \theta > \theta_1 \text{ is false}). \tag{10}$$

Note that $\mathsf{Sev}(T_{\alpha}^{>};\mathbf{x}_0;\theta{>}\theta_1)$ is evaluated at $\theta{=}\theta_1$:

$$SEV(T_{\alpha}^{>};\theta > \theta_1){=}\mathbb{P}(\mathbf{x}\colon\ d_A(\mathbf{x}) \leq d_A(\mathbf{x}_0);\ \theta{=}\theta_1),\ \text{for } \theta_1{=}\theta_0{+}\gamma, \tag{11}$$

because the probability decreases with γ and in the case of reject one is seeking the 'largest' discrepancy warranted by $\mathbf{x}_0$. The evaluation of SEV is based on:

$$\begin{array}{c} d(\mathbf{X}){=}\frac{\sqrt{n}(\hat{\theta}_n-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}\left(\delta(\theta_1), V(\theta_1); n\right),\ \text{for } \theta_1 > \theta_0, \\ \delta(\theta_1){=}\frac{\sqrt{n}(\theta_1-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \geq 0,\ V(\theta_1){=}\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)},\ 0 < V(\theta_1) \leq 1. \end{array} \tag{12}$$

Since the hypothesis that 'passed' is of the form $\theta{>}\theta_1{=}\theta_0{+}\gamma$, the objective of SEV$(T_{\alpha}^{>};\theta > \theta_1)$ is to determine the *largest* discrepancy $\gamma{\geq}0$ warranted by data $\mathbf{x}_0$.

Example 2 (continued). For the observed test statistic $d_A(\mathbf{x}_0){=}7.389$, table 2 evaluates SEV$(\mathcal{T}_{\alpha};\theta > \theta_1)$ for different values of γ, with the evaluations based on:

$$\frac{[d_A(\mathbf{X})-\delta(\theta_1)]}{\sqrt{V(\theta_1)}} \overset{\theta=\theta_1}{\backsim} \mathsf{Bin}\left(0, 1; n\right) \simeq \mathsf{N}(0,1). \tag{13}$$

Note that for the above data, the scaling $\sqrt{V(\theta_1)}$ takes values within the range:

$$\sqrt{V(.51)}{=}\sqrt{\tfrac{.51(1-.51)}{.5(1-.5)}}{=}.9998 \text{ and} \sqrt{V(.525)}{=}\sqrt{\tfrac{.525(1-.525)}{.5(1-.5)}}{=}.9988,$$

and thus close enough to 1 to ignore $\sqrt{V(\theta_1)}$ and use $\left[d_A(\mathbf{X})-\delta(\theta_1)\right]$.

The evaluation of $SEV(T_{\alpha}^{>};\gamma{>}\gamma_1)$, like that of the power, is based on the distribution of the test statistic under the alternative (13), but unlike power, the SEV uses $d_A(\mathbf{x}_0)$ as the threshold instead of c_{α} and $SEV(T_{\alpha}^{>};\gamma{>}\gamma_1)$ takes the form (11) with the tail areas probability from $\mathsf{N}(0,1)$. For instance, when $\gamma{:=}(\theta_1{-}\theta_0){=}.01$, the evaluation of severity components yields:

$$d_A(\mathbf{x}_0){=}7.389,\ \delta(\theta_1){=}\frac{\sqrt{30762}(.51-.5)}{\sqrt{.5(.5)}}{=}3.508,\ d_A(\mathbf{x}_0){-}\delta(\theta_1){=}7.389{-}3.508{=}3.881 \rightarrow$$
$$\rightarrow SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.51){=}\mathbb{P}(\mathbf{x}\colon\ d_A(\mathbf{x}) \leq 7.389;\ \theta{=}\theta_1{=}.51){=}.999.$$

For $\gamma = .013,\ 7.389{-}\frac{\sqrt{30762}(.513-.5)}{\sqrt{.5(.5)}}{=}2.828\,8,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.513){=}.997,$
for $\gamma = .0143,\ 7.389{-}\frac{\sqrt{30762}(.5143-.5)}{\sqrt{.5(.5)}}{=}2.373,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.5143){=}.991,$
for $\gamma = .015,\ 7.389{-}\frac{\sqrt{30762}(.515-.5)}{\sqrt{.5(.5)}}{=}2.127,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.515){=}.983,$
for $\gamma = .016,\ 7.389{-}\frac{\sqrt{30762}(.516-.5)}{\sqrt{.5(.5)}}{=}1.776,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.516){=}.962,$
$\vdots$
for $\gamma = .021,\ 7.389{-}\frac{\sqrt{30762}(.521-.5)}{\sqrt{.5(.5)}}{=}0,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}.521){=}.500,$
for $\gamma = .025,\ 7.389{-}\frac{\sqrt{30762}(.525-.5)}{\sqrt{.5(.5)}}{=}-1.381,\ SEV(T_{\alpha}^{>};\theta{>}\theta_1{=}-1.381){=}.084.$

Table 2 reports the severity curve for various discrepancies, plotted in fig. 3.

Table 2: Severity of 'Reject H_0: θ=.5 vs. H_1: $\theta > .5$' with $(T_\alpha^>; d(\mathbf{x}_0)>0)$

$\theta>\theta_1=\theta_0+\gamma,\ \gamma=$	.01	.013	.0143	.015	.016	.017	.018	.02	.021	.025
$\mathsf{Sev}(\theta>\theta_1)=$	.999	.997	.991	.983	.962	.923	.859	.645	.500	.084

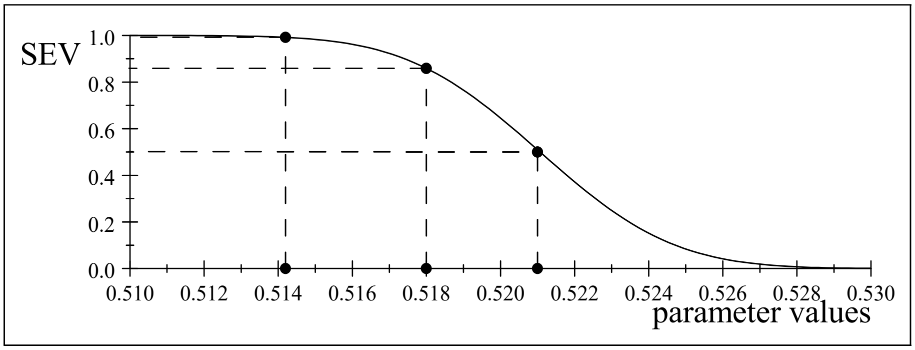

Fig. 3: The severity curve (table 2)

It is important to note that the value θ_1=.521 coincides with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$=.521, and the severity evaluation associated with the inferential claim $\theta>\widehat{\theta}_n(\mathbf{x}_0)$, always has probability .5, i.e. $SEV(T_\alpha^>; \theta>\widehat{\theta}_n(\mathbf{x}_0))$=.5. This lends credence to the argument that it is never a good idea to interpret point estimation as giving rise to the inferential claim that $\widehat{\theta}(\mathbf{x}_0) \simeq \theta^*$, since it represents just a single value from the sampling distribution of $\widehat{\theta}(\mathbf{X})$, with many other values equally or more probable than $\widehat{\theta}(\mathbf{x}_0)$.

Severity and evidential interpretation. Taking a very high probability, say .95, as a threshold, the largest discrepancy from the null warranted by this data is:

$$\gamma \leq .01637, \text{ since } SEV(T_\alpha^>; \theta > .51637)=.950.$$

Is this discrepancy substantively significant? In general, to answer this question one needs to appeal to substantive subject matter information to assess the warranted discrepancy on substantive grounds. In human biology it is commonly accepted that the sex ratio at birth is approximately $\theta^\star$=.5122; see Hardy (2002). Hence, the above warranted discrepancy $\gamma \geq .01637$ is *substantively significant* since it outputs $\theta \geq .5173 > \theta^\star$=.5122.

We can go even further and evaluate a substantive discrepancy of interest $(\theta_B - \theta_A)$, where θ_B=(18/35) is the value put forward by Nicholas Bernoulli:

$$\gamma^*=(18/35) - .5=.0142857,$$

The severity evaluation yields SEV$(T_\alpha^>; \theta > .5143)$=.991, indicating that there is *excellent evidence* for the inferential claim $\theta > \theta_1=\theta_0+\gamma^*$.

In terms of the ultimate objective of statistical inference, learning from data, this evidential interpretation seems highly effective because it narrows down an infinite set Θ to a very small subset!

2.1 Revisiting issues bedeviling frequentist testing

The above post-data evidential interpretation based on the severity assessment can also be used to shed light on a number of issues raised in the previous sections.

2.1.1 Addressing the large n problem

The post-data severity evaluation of the accept/reject H_0 result, addresses the large n problem by taking into consideration the generic capacity (power) of the test in evaluating the warranted discrepancy γ^* from H_0.

Example 1 (continued)**.** In the context of the simple Normal model (2), consider the hypotheses: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$ for $\mu_0=0$, $\alpha = .025, \sigma=2$. The severity curves shown below are associated with test T_α and are based on the same outcome $\kappa(\mathbf{x}_0)=1.96$ but different sample sizes ($n=25, n=100, n=400$), indicating that the severity for inferring $\mu > .2$ decreases as n increases: (i) $n=50$, $SEV(\mu > 0.2)=.895$, (i) $n=150$, $SEV(\mu > 0.2)=.769$, (iii) $n=400$, $SEV(\mu > 0.2)=.49$; see figure 4.

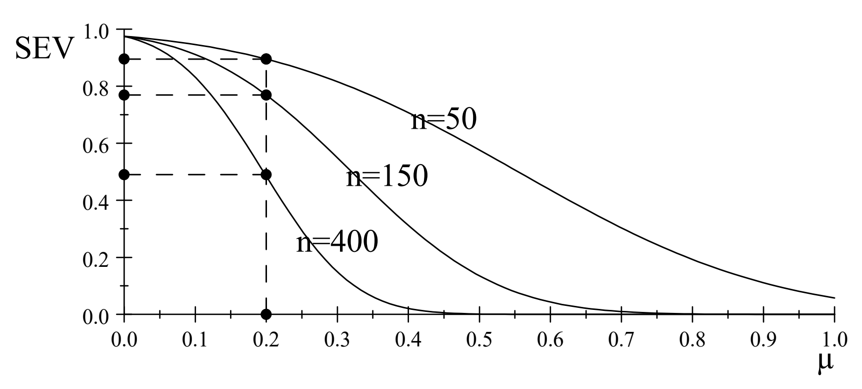

Fig. 4: Severity for $\mu > .2$ with different sample sizes

2.1.2 The arbitrariness of the N-P specification of hypotheses

The post-data information pertaining to the sign of $d(\mathbf{x}_0)$ is very important in addressing several criticisms of N-P testing using the severity evaluation, including: (i) reframing the null and alternative hypotheses, and (ii) manipulating α in an attempt to get the desired result.

[**a**] Does the evidential interpretation depends on the framing of H_0 and H_1, and not $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n_X$, test T_α and data $\mathbf{x}_0$.

Example 2 (continued). Recall the data pertaining to $n=30762$ newborns, 16029 boys and 14833 girls, with $\alpha=.01 \Rightarrow c_\alpha=2.326$, $c_{\frac{\alpha}{2}}=2.575$. Consider the following N-P specifications of H_0 and H_1 revolving around the Bernoulli value $\theta_B=\frac{18}{35}$ using three different formulations:

$$\begin{array}{lll} \text{(i)} & H_0\text{: } \theta \leq \theta_B \text{ vs. } H_1\text{: } \theta > \theta_B, & \mathcal{C}_1^{>}(\alpha)=\{\mathbf{x}\text{: } d_B(\mathbf{x}) > c_\alpha\} \\ \text{(ii)} & H_0\text{: } \theta \geq \theta_B \text{ vs. } H_1\text{: } \theta < \theta_B, & \mathcal{C}_1^{<}(\alpha)=\{\mathbf{x}\text{: } d_B(\mathbf{x}) < c_\alpha\} \\ \text{(iii)} & H_0\text{: } \theta = \theta_B \text{ vs. } H_1\text{: } \theta \neq \theta_B, & \mathcal{C}_1^{\neq}(\alpha)=\{\mathbf{x}\text{: } |d_B(\mathbf{x})| > c_{\frac{\alpha}{2}}\} \end{array} \tag{14}$$

Using the test statistic in example 2 yields:

$$d_B(\mathbf{x}_0) = \frac{\sqrt{30762}(\frac{16029}{30762}-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}} = 2.379. \tag{15}$$

The framing in (i) leads to *rejecting* H_0 at α=.01, confirmed by the p-value:

$$\text{(i) } p_{>}(\mathbf{x}_0) = \mathbb{P}(d(\mathbf{X}) > 2.379; H_0) = .009. \tag{16}$$

However, when the formulation (ii) and (iii) are used H_0 is accepted at α=.01 since $d_B(\mathbf{x}_0)=2.379 < c_{\frac{\alpha}{2}}=2.575$, and the traditional definition of the corresponding p-values are:

$$\text{(ii) } p_{<}(\mathbf{x}_0) = \mathbb{P}(d(\mathbf{X}) < 2.379; H_0) = .991,$$
$$\text{(iii) } p_{\neq}(\mathbf{x}_0) = \mathbb{P}(|d(\mathbf{X})| > 2.379; H_0) = .0174.$$

The three p-values based on the same data $\mathbf{x}_0$ confirm the absurdity of the traditional definition where the clause "equal to or more extreme" is interpreted using the form of the alternative hypothesis. On the other hand, the definition stemming from the post-data nature of the p-value and framed in terms of "all sample realizations $\mathbf{x}\in\mathfrak{X}$ such that $d(\mathbf{x})$ accords less well with H_0 than $\mathbf{x}_0$ does", indicates clearly that the only relevant p-value is (16).

Similarly, the post-data severity argument suggests that the sign of the observed test statistic $d_B(\mathbf{x}_0)=2.379 > 0$ indicates that the inferential *claim* that 'passed' on the basis of data $\mathbf{x}_0$ is of the form:

$$\theta > \theta_1 = \theta_0 + \gamma.$$

To evaluate the particular discrepancy γ warranted by data $\mathbf{x}_0$, condition (S-2) calls for the evaluation of the probability of the event: 'outcomes $\mathbf{x}$ that accord less well with $\theta > \theta_1$ than $\mathbf{x}_0$ does', i.e. ($\mathbf{x}$: $d_B(\mathbf{x}) \leq d_B(\mathbf{x}_0)$) , giving rise to (10), but with a different θ_0. Table 3 lists several such evaluations of:

$$SEV(T_{\alpha}^{>}; \theta>\theta_1) = \mathbb{P}(d_B(\mathbf{X}) \leq 2.379;\ \theta_1 = \tfrac{18}{35}+\gamma),$$

for different values of γ. The evaluations take the form:

for $\gamma = -.0043$, $2.379 - \frac{\sqrt{30762}(.51-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}} = 3.882$, $SEV(T_{\alpha}^{>}; \theta > \theta_1 = .51) = .999$,

for $\gamma = -.0013$, $2.379 - \frac{\sqrt{30762}(.513-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}} = 2.830$, $SEV(T_{\alpha}^{>}; \theta>\theta_1 = .513) = .997$,

$\vdots$

for $\gamma = .0007$, $2.379 - \frac{\sqrt{30762}(.515-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}} = 2.128$, $SEV(T_{\alpha}^{>}; \theta>\theta_1 = .515) = .983$,

$\vdots$

for $\gamma = .0107$, $2.379 - \frac{\sqrt{30762}(.525-\frac{18}{35})}{\sqrt{\frac{18}{35}(1-\frac{18}{35})}} = -1.381$, $SEV(T_{\alpha}^{>}; \theta>\theta_1 = .525) = .084$.

NOTE that the third decimal differences in evaluating $[d_B(\mathbf{X}) - \delta(\theta_1)]$ stem from the fact that $\sqrt{V(\theta_1)}$ is different in the two cases since the value θ_0 is different.

A number of negative values of γ are included to bring out the fact that the results of table 2 and 3 are identical when viewed as inferential claims pertaining to θ; they simply have a different null values θ_0, and thus different discrepancies, but identical θ_1 values and severity evaluation.

Table 3: Severity of N-P Accept H_0: $\theta{\geq}\frac{18}{35}$ vs. H_1: $\theta{<}\frac{18}{35}$ with $(T_{\alpha}^{>}; d_B(\mathbf{x}_0){>}0)$

$\theta{>}\theta_0{+}\gamma,\ \gamma{=}$	-.0043	-.0013	.000	.0007	.0017	.0027	.0037	.0057	.0067	.0107
θ_1	.51	.513	.5143	.515	.516	.517	.518	.520	.521	.525
$\mathsf{Sev}(\theta{>}\theta_1)=$	.999	.997	.991	.983	.962	.923	.859	.645	.500	.084

The severity evaluations in tables 2 and 3, not only render the choice between (i)-(iii) irrelevant, they also scotch the widely used argument pertaining to the *asymmetry* between the null and alternative hypotheses. They demonstrate that the post-data severity evaluation addresses such asymmetries, actual or perceived.

[b] manipulating α in an attempt to get the desired result. It is worth noting that the severity evaluation also addresses this problem. Choosing a larger significance level, say $\alpha{=}.02 \Rightarrow c_{\alpha}{=}2.053$ and $c_{\frac{\alpha}{2}}{=}2.326$ would lead to rejecting the null in both specifications (i) and (iii) since $d_B(\mathbf{x}_0){=}2.379$. The post-data severity evaluation renders that irrelevant since it is driven entirely by $d_B(\mathbf{x}_0){=}2.379{>}0$.

2.1.3 Addressing the fallacy of rejection

The potential arbitrariness of the N-P specification of the null and alternative hypotheses and the associated p-values is brought out in probing the Bernoulli value $\theta_B{=}\frac{18}{35}$ using the different formulations of the hypotheses in (14). Can the severity evaluation explain away these confusing results?

The choice of the N-P framing in (iii), was driven solely by substantive information suggested by the fact that $\theta_A{=}\frac{1}{2}$ and $\theta_B{=}\frac{18}{35}$ are the substantive values of interest. The framing choice (iii) H_1: $\theta \neq \theta_B$ was based on lack of information about the direction of departure, which makes sense pre-data, but not post-data. The framing choice (i) where H_1: $\theta > \frac{18}{35}$ reflects the *post-data* direction of departure indicated by $d_B(\mathbf{x}_0){=}2.379 > 0$. In light of that, the severity evaluation confirms that $p_{>}(\mathbf{x}_0){=}.009$ is the only relevant p-value.

Recall that from a post-data perspective the definition of the p-value is: the p-value is the probability of all possible outcomes $\mathbf{x}{\in}\mathbb{R}_X^n$ that accord less well with H_0 than $\mathbf{x}_0$ does, when H_0 is true.

This perspective brings out the vulnerability of both the p-value and the N-P reject H_0 rule to the fallacy of rejection in cases where n is large. Viewing the above relevant p-value ($p_{B>}(\mathbf{x}_0){=}.009$) from the severity vantage point, it is directly related to H_1 passing a severe test. This is because the probability that test $\mathcal{T}_{\alpha}$ would have produced a result that accords less well with H_1 than $\mathbf{x}_0$ does ($\mathbf{x}$**:** $d_B(\mathbf{x}){<}d_B(\mathbf{x}_0)$), if H_1 were false (H_0 true) is very high since:

$$\mathsf{Sev}(T_{\alpha}^{>}; \mathbf{x}_0; \theta{>}\theta_0) \quad {=}\mathbb{P}(d_B(\mathbf{X}){<}d(\mathbf{x}_0); \theta \leq \theta_0){=}1{-}\mathbb{P}(d_B(\mathbf{X}){>}d_B(\mathbf{x}_0); \theta{=}\theta_0){=}.991$$

■ **What's wrong with the p-value when used as *evidence against* H_0?** A small p-value indicates the existence of *some* discrepancy $\gamma \geq 0$, but provides no information concerning the magnitude warranted by $\mathbf{x}_0$. The severity evaluation remedies that by relating $\mathsf{Sev}(T_\alpha^{>};\mathbf{x}_0;\theta > \theta_0)$=.991 to the discrepancy γ warranted by data $\mathbf{x}_0$ that revolves around the inferential claim $\theta > \theta_0 + \gamma$. Given that the p-value is evaluated at θ=θ_0, the implicit discrepancy associated with the p-value is γ=0. This ignores the power of the test that gave rise to the rejection of H_0.

2.1.4 Addressing the fallacy of acceptance

Example 2 (continued). **Arbuthnot's 1710 conjecture reparameterized.**

Consider the example of the ratio of males to females in newborns. The equality of males and females can be tested in the context of a simple Bernoulli model (6) where $\{X$=$1\}$=$\{\text{female}\}$, $\{X$=$0\}$=$\{\text{male}\}$, using the statistical hypotheses:

$$\text{(i)}\quad H_0\text{: } \varphi \leq \varphi_A \text{ vs. } H_1\text{: } \varphi > \varphi_A,\ \ \varphi_A{=}.5,$$
$$\text{(ii)}\quad H_0\text{: } \varphi \geq \varphi_A \text{ vs. } H_1\text{: } \varphi < \varphi_A,$$

in terms of φ=$\mathbb{P}(X$=$1)$=$E(X)$; NOTE that φ=$1{-}\theta$ in terms of the notation in (6). The best (UMP) tests for (i)-(ii) take the form:

$$\text{(i)}\quad T_\alpha^{>}{:=}\{d_A(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}},\ C_1^{>}(\alpha){=}\{\mathbf{x}{:}d(\mathbf{X}) > c_\alpha\}\},$$
$$\text{(ii)}\quad T_\alpha^{<}{:=}\{d_A(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}},\ C_1^{<}(\alpha){=}\{\mathbf{x}{:}d(\mathbf{X}) < -c_\alpha\}\},$$

where $\widehat{\varphi}_n{:=}\overline{X}_n{=}\frac{1}{n}\sum_{i=1}^n X_i$ as the best estimator of φ and:

$$d_A(\mathbf{X}){=}\tfrac{\sqrt{n}(\overline{X}_n-\varphi_0)}{\sqrt{\varphi_0(1-\varphi_0)}} \overset{H_0}{\backsim} \mathsf{Bin}\left(0,1;n\right). \tag{17}$$

Using the data with n=30762 newborns during the period 1993-5 in Cyprus, 16029 boys and 14833 girls, α=.01 ⇒ c_α=±2.326, yields $\widehat{\varphi}_n(\mathbf{x}_0)$=$\frac{14833}{30762}$ = .48219:

$$\text{(i)}\quad d_A(\mathbf{x}_0){=}\frac{\sqrt{30762}(\frac{14833}{30762}-.5)}{\sqrt{.5(.5)}}{=}-6.249 < 2.326,$$
$$\text{(ii)}\quad d_A(\mathbf{x}_0){=}\frac{\sqrt{30762}(\frac{14833}{30762}-.5)}{\sqrt{.5(.5)}}{=}-6.249 < -2.326,$$

where (i) indicates accepting of H_0, but (ii) indicates rejecting H_0, confirmed by the associated p-values:

$$\text{(i)}\quad p_{A>}(\mathbf{x}_0){=}\mathbb{P}(d_A(\mathbf{X}) > -6.249;\ H_0){=}1.0,$$
$$\text{(ii)}\quad p_{A<}(\mathbf{x}_0){=}\mathbb{P}(d_A(\mathbf{X}) < -6.249;\ H_0){=}2.065 \times 10^{-10} < \alpha{=}.01.$$

At the *coarse accept/reject* H_0 level, there is nothing contradictory about the above N-P results of accepting $\varphi \leq \varphi_A$ and rejecting $\varphi > \varphi_A$. They agree that, with very high probability, the test procedure suggests that the true value of φ, $\varphi^{*}{\in}[0,.5)$ or equivalently $\varphi^{*}{\notin}[5,1]$.

The problem is that the coarseness of these results renders them largely uninformative with respect to the main objective of testing which is to learn from $\mathbf{x}_0$ about φ^*. This is the issue the post-data severity evaluation aims to address. It is achieved by supplementing the coarse accept/reject rules with effective probing of the null value φ_A=.5 with a view to output the discrepancy warranted by data $\mathbf{x}_0$ in the direction indicated by the sign of $d_A(\mathbf{x}_0)$=$-$6.249. Hence, as a post-data error probability, the severity evaluation is guided by $d_A(\mathbf{x}_0) \gtrless 0$ and not by the accept/reject H_0 results as such because the latter is driven by its pre-data specification; see Spanos (2013).

As argued above, the sign of $d_A(\mathbf{x}_0)= -6.249$ indicates the relevant direction of departure from φ_A=.5, and thus the p-value that makes sense as a post-data error probability is the one for case (ii) since the values of $\mathbf{x}{\in}\{0,1\}^n$ that accord less well with φ_A=.5 lie to the left of that value. Similarly, for evaluating the post-data severity the relevant inferential claim is:

$$\varphi \leq \varphi_1{=}\varphi_0{+}\gamma, \quad \text{for some } \gamma \leq 0, \tag{18}$$

$$SEV(T_{\alpha}^{<}, \varphi \leq \varphi_1){=}\mathbb{P}(\mathbf{x}\text{: } d_A(\mathbf{x}) > d_A(\mathbf{x}_0); \varphi_1{=}\varphi_0{+}\gamma). \tag{19}$$

where the evaluation of SEV is based on (12); note that $\sqrt{V(\varphi_1)} \simeq 1$. In light of (18), the objective of SEV$(T_{\alpha}^{>}; \varphi \leq \varphi_1)$ is to determine the *largest* discrepancy $\gamma{\leq}0$ warranted by data $\mathbf{x}_0$.

The post-data severity evaluations are as follows:

for $\gamma = -.01,\ -6.249-\frac{\sqrt{30762}(.49-.5)}{\sqrt{.5(.5)}} = -2.741,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.49){=}.997$

for $\gamma = -.0122,\ -6.249-\frac{\sqrt{30762}(.4878-.5)}{\sqrt{.5(.5)}}{=}-1.9695,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.487).976,$

for $\gamma = -.013,\ -6.249-\frac{\sqrt{30762}(.487-.5)}{\sqrt{.5(.5)}}{=}-1.6888,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.487){=}.954,$

for $\gamma = -.0143,\ -6.249-\frac{\sqrt{30762}(.4857-.5)}{\sqrt{.5(.5)}}{=}-1.232\,8,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.4857){=}.891,$

for $\gamma = -.015,\ -6.249-\frac{\sqrt{30762}(.485-.5)}{\sqrt{.5(.5)}}{=}-.987,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.485){=}.838,$

for $\gamma = -.0178,\ -6.249-\frac{\sqrt{30762}(.48219-.5)}{\sqrt{.5(.5)}}{=}0.0,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.48219){=}.5,$

for $\gamma = -.018,\ -6.249-\frac{\sqrt{30762}(.482-.5)}{\sqrt{.5(.5)}}{=}.0551,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.482){=}.478,$

for $\gamma = -.019,\ -6.249-\frac{\sqrt{30762}(.481-.5)}{\sqrt{.5(.5)}}{=}.41586,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.481){=}.339,$

for $\gamma = -.021,\ -6.249-\frac{\sqrt{30762}(.479-.5)}{\sqrt{.5(.5)}}{=}1.1174,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}.479){=}.132,$

for $\gamma = -.025,\ -6.249-\frac{\sqrt{30762}(.475-.5)}{\sqrt{.5(.5)}}{=}2.521,\ SEV(T_{\alpha}^{>}; \varphi{\leq}\varphi_1{=}\ .475){=}.058.$

Table 4 reports the severity curve for various discrepancies and plotted in fig. 5. Using the severity threshold of .95, the warranted discrepancy from φ_A=.5 can be derived using the equation:

$$-6.249-\frac{\sqrt{30762}(\gamma)}{\sqrt{.5(.5)}}{=}-1.645 \rightarrow \gamma^*{=}-.013125,$$

where the value -1.645 is chosen because $1-\Phi(-1.645)=.95$; Φ denotes the cumulative distribution function (cdf) of $\mathsf{N}(0,1)$. The inferential claim warranted by data $\mathbf{x}_0$ with at least .95 severity is: $\varphi \leq .486875$.

Similarly, the severity evaluation of the discrepancy associated with the substantive value of φ, $\varphi^{\star}$=.4878, has $SEV(T_{\alpha}^{>};\varphi{\leq}.4878)$=.976, which exceeds the .95 threswhold. This indicates that the test result based on $d_A(\mathbf{x}_0)=-6.249$ is both statistically and substantively significant with data $\mathbf{x}_0$.

Finally, it is important to note that the post-data severity evaluations in table 4 are in complete agreement with those in tables 2 and 3.

Table 4: Severity of N-P 'Accept H_0': $\varphi{\leq}.5$ vs. H_1: $\varphi{>}.5$' with $(T_{\alpha}^{>};d_A(\mathbf{x}_0){<}0)$

γ=	-.01	-.0122	-.013	-.0143	-.015	.0178	.018	-.019	-.021	-.025
$\varphi_1{=}\varphi_0{+}\gamma$,	.49	.4878	.487	.4857	.485	.4822	.482	.481	.479	.475
$\mathsf{Sev}(\varphi{>}\varphi_1)$ =	.997	.976	.954	.891	.838	.500	.859	.339	.132	.058

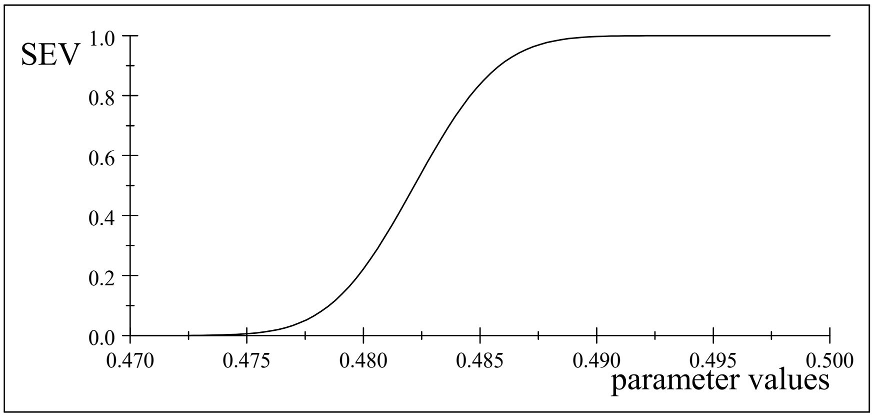

Fig. 5: Severity evaluation for N-P accept H_0

A comparison of Figures 5 and 3 indicates that their severity curves differ with respect to their slope since their evaluations pertain to the two different tails of the sampling distribution of $d_A(\mathbf{X})$ under H_1.

2.2 Warranted discrepancy vs. effect sizes

Example 2 (continued). Consider testing the hypotheses:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0, \text{ where } \theta_0{=}.5, \tag{20}$$

in the context of the simple Bernoulli model (6) using a subset of the above data for 1993 only with n=10514 newborns 5442 boys and 5072 girls, assuming α=.01 $\Rightarrow$ c_{α}=2.326. The relevant test statistic yields:

$$d_A(\mathbf{x}_0)=\left(\sqrt{10514}(\tfrac{5442}{10514}-\tfrac{1}{2})/\sqrt{.5(.5)}\right)=3.608[.00015], \tag{21}$$

with the p-value, in square brackets, indicating *reject* H_0.

Severity evaluation. Since $d(\mathbf{x}_0)=3.608>0$, the relevant inferential claim is (Spanos, 2013):

$$\theta > \theta_1=\theta_0+\gamma, \quad \text{for some } \gamma \geq 0. \tag{22}$$

To establish the particular discrepancy γ warranted by data $\mathbf{x}_0$, the evaluation is:

$$SEV(T_\alpha;\theta>\theta_1)=\mathbb{P}(d_A(\mathbf{X}) \leq d_A(\mathbf{x}_0);\ \theta=\theta_1),\ \text{for } \theta_1=\theta_0+\gamma,\ \gamma\geq 0, \tag{23}$$

which is given in table 5.

Table 5: Severity Evaluation of 'Reject H_0' with $(T_\alpha;\mathbf{x}_0)$

γ	.01	.012	.0125	.013	.014	.015	.0176	.02	.025	.03
θ_1=.5+γ	.51	.512	.5125	.513	.514	.515	.5176	.52	.525	.53
Sev($\theta>\theta_1$)	.940	.874	.852	.827	.782	.703	.500	.311	.064	.005

Note that the objective of SEV($T_\alpha;\theta>\theta_1$) is to determine the *largest discrepancy* $\gamma\geq 0$ warranted by data $\mathbf{x}_0$ for a prespecified level of severity, say SEV($T_\alpha;\theta>\theta_1$)$\geq$.85. In this case:

$$\gamma^* \leq .01254,\ \text{since } SEV(T_\alpha;\theta > .51254)=.85. \tag{24}$$

That is, the testing-based *effect size* with data $\mathbf{x}_0$ and severity .85 is $\gamma^* \leq .01254$.

The warranted discrepancy γ^* should be contrasted with *estimation-based* effect size estimates whose primary aim is to get a more appropriate measure of the 'magnitude of the scientific effect'. Although there is no agreement in the literature about the most appropriate effect size estimate, Cohen's (1988) g is:

$$g=(.5176-.5)=.0176,$$

which is *much larger* than the substantively determined value γ^*=.01254 and $SEV(T_\alpha;\theta > .5176)=.50$.

This reveals the arbitrariness of grading such effects sizes as small, medium and large. Using Cohen's benchmarks, .0176 is tiny, since the benchmark for small is g_s=.05, despite the fact that $\gamma^*\leq$.01254 implies substantive significance.

As a general rule, it is never a good idea to view the point estimate, say $\widehat{\theta}(\mathbf{x}_0)=\overline{x}_n$, as (approximately) coinciding with θ^*, since it represents just a single value from the sampling distribution of $\widehat{\theta}(\mathbf{X})$. This calls into question the appropriateness of any estimation-based effect sizes, more generally, since they are based on a single realization.

Statistical vs. substantive significance. The warranted discrepancy γ^* from the null H_0 is as far as the statistical information in data $\mathbf{x}_0$ can take an inferential claim; γ^* cannot be proclaimed substantively significant. As mentioned above, there is a *substantively* determined value for θ, $\theta^\star\simeq$.5122, which indicates that the warranted inferential claim: $\theta > \theta_1=\theta_0+\gamma$, $\gamma^* \leq .01254$, implies that it is both *statistically* and *substantively significant.*

2.3 Addressing questionable practices in N-P testing

Example 3 (continued). [i] What happens if one were to replace the hypotheses in (20) with the simple-vs-simple hypotheses:

$$H_0\text{: } \theta{=}\tfrac{1}{2} \text{ vs. } H_1\text{: } \theta{=}\tfrac{18}{35}. \tag{25}$$

Given that H_0 and H_1 do not constitute a partition of $\Theta{:=}[0,1]$, the framing in (25) is *improper*; see Spanos (2013). Having said that, if one were interested in the discrepancy $(\frac{18}{35}{-}\frac{1}{2}){=}\frac{1}{70}{=}.014286$, then its post-data severity could be evaluated using table 5. The inferential claim $\gamma{\leq}\frac{1}{70}$ has severity: $SEV(T_\alpha;\theta{>}(.5{+}\frac{1}{70}){=}.5143){=}.75$.

[ii] What if one were to replace the hypotheses in (20) with:

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0, \text{ where } \theta_0{=}\tfrac{18}{35}, \tag{26}$$

with $\alpha{=}.01 \Rightarrow c_\alpha{=}2.326$? Since the test statistic yields:

$$d_B(\mathbf{x}_0){=}\left[\sqrt{10514}(\tfrac{5442}{10514}{-}\tfrac{18}{35})/\sqrt{.5(.5)}\right]{=}.679[.249], \tag{27}$$

the null will now be accepted, but the post-data severity evaluation will not change because $d_B(\mathbf{x}_0){=}.679{>}0$ indicates that θ_1 will remain the same; γ will change in table 3 since the relevant inferential claim is now $\theta{>}\theta_1{=}(\frac{18}{35}{+}\gamma)$.

Table 5A: Severity Evaluation of 'Reject H_0' with $(T_\alpha;\mathbf{x}_0)$

γ	-.0043	-.002	-.0018	-.013	-.0003	.0007	.003	.006	.011	.016
$\theta_1{=}\frac{18}{35}{+}\gamma$	.51	.512	.5125	.513	.514	.515	.5176	.52	.525	.53
Sev$(\theta{>}\theta_1)$	.940	.874	.852	.827	.782	.703	.500	.311	.064	.005

[iii] What if one were to change $\alpha{=}.01$ in case [ii] to $\alpha{=}.25 \Rightarrow c_\alpha{=}.674$? The null $\theta_0{=}\frac{18}{35}$ will now be rejected, but the severity evaluation remains the same since $d_B(\mathbf{x}_0){>}0$.

In summary, the severity evaluation remains invariant to the reframing of H_0 and H_1, and any changes in α, as long as the framing constitutes a partition of Θ.

2.4 Replicability of empirical evidence

The founder of modern statistics declared: "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to gives us a statistically significant result." (Fisher, 1935, p. 14)

In a non-experimental setting a particular phenomenon of interest is replicable when similar inference results are reached using data from different periods and different locations. To demonstrate this, let us return to the inferential results based on Cyprus data for 1993 comprising $n{=}10514$ newborns 5442 boys and 5072 girls. The above inferences will be compared to those based on Arbuthnot's data for 1706 for London (England) with n=15369 newborns 7952 boys and 7417 girls, assuming α=.01 $\Rightarrow$ c_{α}=2.326. The choice is based not only for their differences in time and place, but also on the fact that the sample sizes are similar.

Example 2. Consider testing the hypotheses (20) using Arbuthnot's data yielding $\widehat{\theta}(\mathbf{x}_0)=\frac{7952}{15369}$=.5174. The relevant test statistic yields:

$$d_A(\mathbf{x}_0)=\frac{\sqrt{15369}(.5174-.5)}{\sqrt{.5(.5)}}=4.316[.00008], \tag{28}$$

with the tiny p-value, in square brackets, indicating *reject* H_0. Since $d_A(\mathbf{x}_0)$=4.316>0, the relevant inferential claim is: $\theta>\theta_1=\theta_0+\gamma,\ \gamma\geq 0$, and the evaluation is based on (23); see table 6.

Table 6: Severity Evaluation of 'Reject H_0' with $(T_{\alpha};\mathbf{x}_0)$

γ	.01	.012	.0125	.013	.014	.015	.0176	.02	.025	.03
θ_1=.5+γ	.51	.512	.5125	.513	.514	.515	.5176	.52	.525	.53
$\mathsf{Sev}(\theta>\theta_1)$	.967	.910	.888	.863	.801	.725	.481	.260	.030	.0009

For a prespecified level of severity, say SEV$(T_{\alpha};\theta>\theta_1)\geq$.85:

$$\gamma^{*}\leq .0132,\ \text{since}\ SEV(T_{\alpha};\theta>.5132)=.85. \tag{29}$$

which is very similar to the warranted discrepancy for the Cyprus 1993 data:

$$\gamma^{*}\leq .01254,\ \text{since}\ SEV(T_{\alpha};\theta>.51254)=.85. \tag{30}$$

Notice that in both cases, the substantive discrepancy of .0122 is with the above upper bounds.

2.5 Summary and conclusions

The *crucial* foundational issue discussed above concerns the form and nature of the evidence $\mathbf{x}_0$ can provide *for* $\theta\in\Theta_0$ or $\theta\in\Theta_1$. Neither the p-value, nor the N-P accept/reject rules provide an evidential interpretation, primarily because they are too coarse and highly vulnerable to the fallacies of acceptance and rejection.

These fallacies, however, can be circumvented using a post-data severity evaluation of the accept/reject results to output the discrepancy γ from the null warranted by data $\mathbf{x}_0$. This warranted inferential claim for a particular test T_{α} and data $\mathbf{x}_0$, gives rise to learning from data; see Mayo and Spanos (2006). The post-data severity evaluation was shown to address, not only the classic fallacies of acceptance and rejection, but several other foundational problems bedeviling frequentist testing since the 1930s, including several abuses of the p-value and questionable practices associated with N-P testing; see Mayo (2018).

Another foundational issue pertains to the presumption that the true $\mathcal{M}^{*}(\mathbf{x})$ lies within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. This can be addressed by securing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, vis-a-vis data $\mathbf{x}_0$, using trenchant *Mis-Specification (M-S) testing*, before applying frequentist testing. When $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified, the *nominal* and *actual error probabilities* can be very different, undermining the reliability of the test in question. This issue will be discussed in Lecture Notes 5.

3 Appendix: Misleading claims about CIs

The literature on the merits of using Confidence Intervals (CIs) often focuses on the observed CIs and ignores the fact that they are directly related to the p-value. In addition, attempts to discriminate among the different values of θ within an observed CI often invoke probabilities that have nothing to do with the relevant error probability, that pertaining the coverage of θ^*. These arguments often ignore the fact that detaching an observed interval within the parameter space from the relevant coverage probability error results in something that has nothing to do with the CI estimation as such.

3.1 Severity vs. observed Confidence Intervals (CIs)

The post-data severity evaluation can be used to address the issue of degenerate post-data coverage error probability, resulting the impossibility to distinguish between different values of μ within an observed CI; Mayo and Spanos (2006). This is achieved by replacing:

(i) the *factual* reasoning underlying CIs with the *hypothetical* reasoning, and

(ii) the inferential claims of overlaying the true μ^* with *post-data* severity-based inferential claims.

In the case of the simple Normal model (2), this takes the form of placing μ_1 on the relevant boundary (in light of $d(\mathbf{x}_0)>0$) of the observed CI $\overline{x}_n \pm c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ associated with:

$$\mathbb{P}(\overline{X}_n - c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \leq \mu^* \leq \overline{X}_n + c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})) = (1-\alpha)\,, \tag{31}$$

$$\mu > \mu_1 = \overline{x}_n - c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \text{ with } SEV(T_\alpha; \mu>\mu_1) = \mathbb{P}(d(\mathbf{X}) \leq d(\mathbf{x}_0);\ \mu=\mu_1). \tag{32}$$

A moment's reflection, however, suggests that the changes (i)-(ii) have eliminated any connection between on observed CI and (32). The severity assessment of $\mu > \mu_1$ does not assign probabilities to the observed CI or μ_1, but to the inferential claim in (32) to establish the discrepancy $\gamma^* = \mu_1 - \mu_0$ warranted by $\mathbf{x}_0$. The severity probability has *nothing* to do with the coverage probability. Moreover, the severity probabilities associated with the inferential claim $\mu > \mu_1 = \mu_0 + \gamma$ are graded, since they evaluated at different values of μ_1, but the coverage probabilities associated with $\overline{X}_n - c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ pertain to the whole interval.

What is not sufficiently appreciated by the statistical literature is that the equality of the tail areas in hypothesis testing and CIs stems from a mathematical duality, but that does *not* imply *inferential duality*. Indeed, when viewed in terms of the severity warranted inferential claims of the form $\mu \gtrless \mu_1 = \mu_0 + \gamma,$, many points μ_1 in any observed CI have very low severity; see Mayo and Spanos (2006).

3.2 Replacing the p-value with observed CIs

It is often claimed by the reformers that the p-value should be replaced by the analogous observed CI because: :

(i) an observed CI less vulnerable to the large n problem, and

(ii) more informative than the p-value since it provides an 'effect size'.

Cohen's (1994) recommendation to practitioners:

"routinely report effect sizes in the form of confidence intervals" (p. 1002).

A closer look at these claims reveals that they are false.

Claim (i) is false because a CI is equally vulnerable to the large n problem since the expected length of a consistent CI shrinks to zero as $n \to \infty$. For (31):

$$E\left([\overline{X}_n{+}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})]{-}[\overline{X}_n{-}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})]\right)=2c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \underset{n\to\infty}{\to} 0.$$

The fact that n plays a crucial role in defining the length of an observed CI also calls into question claim (ii) that it provides a reliable measure of the 'effect size'.

Further light can be shed on both claims, using the mapping between the p-value and the corresponding observed CI stemming from placing the null value μ_0 on its boundary. For the CI (31), using the relationship $c_\alpha{=}\Phi^{-1}(1{-}\alpha)$, Φ-cumulative distribution function (cdf) of $\mathsf{N}(0,1)$:

$$\mu_0{=}\overline{x}_n{\pm}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \to c_\alpha{=}|\tfrac{\sqrt{n}(\overline{x}_n-\mu_0)}{\sigma}| \to \alpha(\mu_0){=}\mathbb{P}(|d(\mathbf{X})|{>}|d(\mathbf{x}_0)|;\ \mu{=}\mu_0), \tag{33}$$

where $\alpha(\mu_0)$ denotes the smallest significance level at which H_0: $\mu{=}\mu_0$ is rejected.

It turns out that the mapping in (33) relates to another issue with observed CIs that pertain to attempts to assign probabilities that render different points within such an interval more or less likely. This move is fallacious because *post-data* the coverage probability $(1{-}\alpha)$, evaluated under $\mu{=}\mu^*$, makes no statistical sense. The observed CI either contains μ^* or it doesn't since post-data the factual scenario $\mu{=}\mu^*$ has played out. The apparent assignment of probabilities to different values of μ within the observed CI is attempted by extending (33) to all values $\mu_1{\neq}\mu_0$ and holding $\overline{x}_n$ constant:

$$\mu_1{=}\overline{x}_n{\pm}c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \to \alpha(\mu_1){=}\mathbb{P}(|d(\mathbf{X})|{>}|d(\mathbf{x}_0)|;\ \mu{=}\mu_1), \tag{34}$$

where $\mathbb{P}(|Z| \geq c_{\frac{\alpha}{2}}){=}(1{-}\alpha)$ for $Z \backsim \mathsf{N}(0,1)$. The family of curves in (34) was initially proposed by Birnbaum (1961), who called it *an omnibus confidence curve*, but is has been rediscovered by Kempthorne and Folks (1971) naming it *a consonance interval curve*, and more recently by Poole (1987), calling it a *p-value curve*.

The sleight of hand magic trick in the mappings in (33) and (34) is that $\overline{x}_n$ is (inadvertently) replaced half-stream with $\overline{X}_n$, and the factual reasoning is replaced with hypothetical reasoning.

Useful information. For the evaluations of severity in the examples above, the most convenient way to derive them is to use an app called 'Probability Distributions' that can be downloaded from both the Apple store as well as the Adroid app store. It is available courtesy of the department of Statistics, University of Iowa.