PHIL 6334 - Probability/Statistics Lecture Notes 7:
Statistical Inferences vs. Probabilistic Deductions:
Diagnostic screening and their false positive/negative rates

Aris Spanos [Spring 2019]

1 Introduction

The statistics and philosophy of science literatures are prone to **conflate** two very different type of inferences.

(a) **Probabilistic inferences** stemming from premises that are *taken at face* value and the inferences are purely *deductive* in nature; given these probabilities we can deduce the following probabilities relating to the same events. Such inferences are usually at the level of a probability space $(S, \mathfrak{F}, \mathbb{P}(.))$ where S -the set of all possible outcomes, $\mathfrak{F}$-a field of events of interest and related events, and $\mathbb{P}(.)$ assigns probabilities to events in $\mathfrak{F}$.

(b) **Statistical inferences** stemming from model-based:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})=\{f(\mathbf{z};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{z}\in\mathbb{R}^n_Z,$$

premises whose *validity needs to be established* vis-a-vis particular data $\mathbf{z}_0$, and the resulting inferences are *inductive* in nature.

The discussion that follows distinguishes between the above two types of inferences and calls into question the statisticians and philosophers who ground their criticisms against **error statistics** and the *post-data severity* firmly on this particular confusion.

The source of the confusion is the analogical reasoning used to blur what is known as **false positive and false negative rates** associated with medical devises or procedures for detecting different diseases with **legitimate error probabilities**, the type I and II, associated with the Neyman-Pearson testing.

The main argument is that the apparent similarity between examples like the **Harvard Medical School test** and frequentist testing within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is **more apparent than real**. On closer examination, such examples have *none* of the basic features of a proper frequentist test:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}),\ \text{hypotheses, test statistics, data } \mathbf{z}_0,\ \text{sampling distributions, etc.}$$

Worse, *error probabilities* are misconstrued as **conditional probabilities** relating to events of interest, and not as probabilities stemming from $f(\mathbf{z};\boldsymbol{\theta}), \mathbf{z}\in\mathbb{R}^n_Z$ that determine the capacity of the test. In fact error probabilities, not only are not conditional, they are invariably assigned to inference procedures and *not* to θ. Hence, in such examples the *ampliative* dimension of frequentist testing is totally absent. Indeed, no learning from data can take place using medical screening devices or procedures because their false positive/negative rates represent simple deductive calculations based on known conditional probabilities among events.

2 A probabilistic result: the base-rate fallacy

Consider the case of a *medical diagnostic test* to detect a particular disease. It is well-known that such tests are almost never 100% accurate. Let us assume that for this particular diagnostic test the device has been calibrated to have the follow rates:

(a) **False negative.** If a patient has the disease, the test will likely detect it (give a positive result) with .95 probability, i.e. its *false negative* probability is .05.

(b) **False positive.** If a patient does *not* have the disease, the test will likely incorrectly give a positive result with .1 probability (*false positive*).

The question of interest is:

▶ when a person from a particular population, say Joan, tests positive using one these devices, what is the probability that she actually has the disease?

A common sense but naive answer is that since the test is 95% accurate, the probability that Joan has the disease is .95. Probabilistic reasoning, however, is more subtle than common sense, and a proper evaluation requires one to think through this question more systematically, beginning with specifying the *events of interest*. In this case, there are two such events one should be considering for Joan:

$$A \text{ - has the disease, } B \text{ - tests positive.}$$

Note that these events A and B are defined on a probability space $(S, \mathfrak{F}, \mathbb{P}(.))$, ensuring that $A \in \mathfrak{F}$ and $B \in \mathfrak{F}$, together with the related event to form a field; close under the set theoretic operations of $\cup, \cap, .$

Whether Joan has the disease or not are evaluated by the conditional probabilities:

$$\mathbb{P}(A|B) \text{ and } \mathbb{P}(\overline{A}|B),$$

and not the *accuracy* of the particular medical diagnostic test which pertains to the conditional probability $\mathbb{P}(B|A)$=.95 [correct diagnosis]and $\mathbb{P}(B|\overline{A})$=.1 [false positive]. How do we relate these probabilities?

Formal probabilistic reasoning relates the two via the conditional probability formula:

$$\mathbb{P}(A|B)=\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}=\frac{\mathbb{P}(A)\cdot\mathbb{P}(B|A)}{\mathbb{P}(A)\cdot\mathbb{P}(B|A)+\mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A})}, \tag{1}$$

where the second equality stems from the fact that:

$$\mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B|A),$$

and since A and $\overline{A}$ constitute a partition of the relevant outcomes set S :

$$\mathbb{P}(B)=\mathbb{P}(A)\cdot\mathbb{P}(B|A)+\mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A}),$$

known as the *total probability rule*. Note that (1) is often (misleadingly) called Bayes formula.

Since, $\mathbb{P}(B|A)$=.95, $\mathbb{P}(B|\overline{A})$=.05, $\mathbb{P}(B|\overline{A})$=.1, to evaluate the conditional probability formula in (1) one needs the probabilities $\mathbb{P}(A)$ and $\mathbb{P}(\overline{A})$:

$\mathbb{P}(A)$: the probability of a person randomly selected
from the particular population has the disease or not

Note that $\mathbb{P}(A)$ is often referred to as 'prevalence' or 'base rate'.

Let us also assume that a person randomly selected from a particular population, the town of Blacksburg, will have the disease with probability of .03, $\mathbb{P}(A)$=.03, $\mathbb{P}(\overline{A})$ = .97. Using these probabilities we can evaluate the remaining ones via:

$$\mathbb{P}(A|B)=\frac{\mathbb{P}(A)\cdot\mathbb{P}(B|A)}{\mathbb{P}(A)\cdot\mathbb{P}(B|A)+\mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A})}=\frac{(.03)(.95)}{(.03)(.95)+(.97)(.1)}=.227$$

where $\mathbb{P}(B)$=.1255. Indeed, one can relate the two probabilities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ by noting that:

$$\mathbb{P}(A|B)=\mathbb{P}(B|A)\left(\frac{\mathbb{P}(A)}{\mathbb{P}(B)}\right)=(.95)\left(\frac{.03}{.1255}\right)=.277. \tag{2}$$

Confusing the two probabilities is often referred to as the *base-rate fallacy*, i.e. the ratio $[\mathbb{P}(A)/\mathbb{P}(B)]$, referred to as base rates, is ignored. In psychology the *base-rate fallacy* refers to the error people make by ignoring the relative sizes of population subgroups when assessing the likelihood of contingent events involving the subgroups; see Tversky and Kahneman (1982).

In the context of Bayesian reasoning, the base-rate fallacy has been formalized as the error involved when the conditional probability of a hypothesis H given some evidence E, $P(H|E)$ — known as the *posterior* probability — is assessed on the basis of $P(E|H)$ without taking account of the *prior* probability (*base rate*) of H, $P(H)$. This fallacy stems from the fact that since probability calculus gives the relationship:

$$P(H|E)=P(E|H)\left(\frac{P(E)}{P(H)}\right), \tag{3}$$

the base rates $[P(H), P(E)]$ can be ignored when evaluating $P(H|E)$ at one's peril.

REMARK: it is very important to note that in this Bayesian context hypothesis H **is not observable,** in contrast to the above numerical example.

What is not so apparent is whether frequentist testing, which attaches *no* probabilities to hypotheses [$P(H)$ is meaningless], is also vulnerable to such a fallacy. Howson (1997, 2000) and Achinstein (2001, 2009), inter alia, contend that frequentist testing, in general, and the severity assessment (Mayo, 1996), in particular, are not just susceptible to this fallacy, they are totally undermined by it.

Howson (2000), (p. 54):

"The central methodological claims of a recent book on methodology (Mayo 1996) are based on committing it. In this book a version of the argument very similar to Fisher's is proposed. As in the tea-tasting, there is a notion of outcomes agreeing better or worse with some hypothesis H ... Similarly, spelling out explicitly what is implicit in Fisher's

discussion, if E 'fits' H, while there is a very small chance that the test procedure 'would yield so good a fit if H is false', then, 'E should be taken as good grounds for H to the extent that H has passed a severe test with E' (Mayo 1996:177). In the Harvard Medical School test we have Mayo's formal criteria for H 'passing a severe test with E' satisfied." Really!!

A similar view is expressed in Howson (1997), Howson and Urbach (2005) and Achinstein (2001, 2009).

2.1 Unraveling the Base-Rate Fallacy Argument

Achinstein (2009) example. In order to simplify the untangling of the various confusions vitiating the base-rate fallacy argument, consider the Harvard Medical School test example, specified in terms of the following probabilities:

$$\mathsf{P}(B|A)=.8, \quad \mathsf{P}(B|\overline{A})=.00002, \quad \mathsf{P}(A)=.0000001 \tag{4}$$

Recall that A - has the disease, B - tests positive. $(1-\mathsf{P}(B|A))$-false positive, $\mathsf{P}(B|\overline{A})$-false negative.

The first question that naturally arises is that, if this is supposed to be an example of statistical inference:

2.1.1 What is the underlying Statistical Model?

Since frequentist statistical inference is **model-based**:

$$\boxed{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ m<n,} \tag{5}$$

the first thing one needs to uncover is the *implicit statistical model* which can then be used to delineate what constitute legitimate events, hypotheses, data, test statistics, error probabilities, etc. Why?

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ plays a **pivotal role** in frequentist inference because:

(i) it specifies the *inductive premises of inference*,

(ii) it determines what constitutes *a legitimate* event (any function $h(\mathbf{x})$)

(iii) it assigns probabilities to all *legitimate events* via $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$,

(iv) it defines what are *legitimate hypotheses* and/or inferential claims,

(vi) it designates what constitute *legitimate data* $\mathbf{x}_0$ for inference purposes,

(v) it determines the relevant error probabilities in terms of which the optimality and reliability of inference methods is assessed. **To wit**: no $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and $\mathbf{x}_0$, no error probabilities or frequentist inferences!

Harvard Medical School test example. Given that there are only *two outcomes* positive or negative, it is easy to recognize that the events A - has the disease, and B - tests positive correspond to two Bernoulli distributed random variables:

X denotes having the disease (X=1) or not (X=0), and

Y denotes the medical test result, positive (Y=1) or negative (Y=0).
Thus, the probabilities in (4) can be written more transparently in terms of the joint probabilities $\{p_{ij}:=\mathbb{P}(X=i,Y=j),\ i,j=0,1\}$:

Table 1: Joint distribution of (X,Y)			
$X\backslash Y$	0	1	$f(x)$
0	p_{00}=.99997998	p_{01}=.00001992	$(1-\theta_1)$=.9999999
1	p_{10}=.00000002	p_{11}=.00000008	θ_1=.0000001
$f(y)$	$(1-\theta_2)$=.99998	θ_2=.00002	1

$$\begin{aligned}&\mathbb{P}(Y=1|X=1)=\tfrac{p_{11}}{p_{10}+p_{11}}=.8, \qquad \mathbb{P}(Y=1|X=0)=\tfrac{p_{01}}{p_{00}+p_{01}}=.00002,\\ &\mathbb{P}(X=1)=p_{10}+p_{11}=.0000001, \quad p_{00}+p_{01}+p_{10}+p_{11}=1.\end{aligned} \tag{6}$$

Solving (6) for $\{p_{ij}\ i,j=0,1\}$ gives rise to the joint distribution in table 1.

This suggests that the underlying *statistical model* — assumed to describe the incidence of a disease in a particular population as it relates to the result of a medical test for that disease — is a **simple (bivariate) Bernoulli model:**

$$\boxed{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}):\quad \mathbf{Z}_k \backsim \mathsf{BerIID}(E(\mathbf{Z}_k),\ Cov(\mathbf{Z}_k)),\ k=1,2,...,n,...} \tag{7}$$

where $\mathbf{Z}_k:=\begin{pmatrix} X_k \\ Y_k \end{pmatrix}$, $E(\mathbf{Z}_k)=\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$, $Cov(\mathbf{Z}_k)=\begin{pmatrix} \theta_1(1-\theta_1) & p_{11}-\theta_1\theta_2 \\ p_{11}-\theta_1\theta_2 & \theta_2(1-\theta_2) \end{pmatrix}$.

$\boldsymbol{\theta}$**:**$=(p_{11},\theta_1,\theta_1)\in[0,1]^3$, $\theta_1=p_{10}+p_{11}$, $\theta_2=p_{01}+p_{11}$, denotes the *unknown* parameters.

The additional 'ugliness' notwithstanding, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is identical to the *simple Bernoulli*:

$$\mathcal{M}_{\theta}(\mathbf{x}):\quad X_k \backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\ 0\leq\theta\leq 1,\ x_k=0,1,\ k=1,2,...., \tag{8a}$$

apart from the underlying process $\{\mathbf{Z}_k,\ k\in\mathbb{N}\}$ being bivariate!

The first thing one notices about $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in (7) and table 1 is that the latter is an instantiation of the former that involves *no unknown parameters*. If frequentist inductive inference is about learning from data about the data-generating mechanism, the question that naturally arises is 'what could *learning* based on table 1 mean?'

2.2 Probabilistic vs. Statistical Inference

It is well-known that **statistical hypotheses** in frequentist testing are always framed in terms of the unknown parameters $\boldsymbol{\theta}$, and they invariably pertain to the stochastic mechanism — $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in (7) — that gave rise to data:

$$\mathbf{z}_0:=([x_1,y_1],[x_2,y_2],...,[x_n,y_n]).$$

An example of such N-P hypotheses in the context of the bivariate Bernoulli model (7) might be:

$$H_0\textbf{:}\ \phi\leq\phi_0 \text{ vs. } H_1\text{: } \phi>\phi_0,\ \text{for } \phi_0=.05, \tag{9}$$

where ϕ=$\mathbb{P}(X$=1$|Y$=1)=(p_{11}/θ_2) (*unknown*), relates to the the probabilitity that a proportion of the population that has the disease, given that they tested positive.

The N-P hypotheses in (9) pose the sharp question whether or not $\phi \leq .05$, with a view to *learn from data* by narrowing the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ down to a subset:

$$\phi^* \in \mathcal{M}_0(\mathbf{z})=\{f(\mathbf{z};\phi),\ \phi \leq \phi_0\} \text{ or } \phi^* \in \mathcal{M}_1(\mathbf{z})=\{f(\mathbf{z};\phi),\ \phi > \phi_0\}.$$

2.2.1 Statistical hypotheses vs. events

A frequentist test for properly defined hypotheses such as (9) will involve constructing a test similar to T_α in (**??**), say:

$$\tau(\mathbf{Z})=\frac{(\widehat{\phi}-\phi_0)}{Var(\widehat{\phi})^{\frac{1}{2}}},\ C_1(\alpha)=\{\mathbf{z}\colon\ \tau(\mathbf{z}) > c_\alpha\}, \tag{10}$$

for an appropriate estimator $\widehat{\phi}$ of ϕ. Using the sampling distribution of $\tau(\mathbf{Z})$ under both the null and alternative hypotheses, one can evaluate the *pre-data* and *post-data* error probabilities as shown in Lecture Notes 3-4. One feature of the hypotheses in (9) worth bringing out is that because ϕ_0=.05 is rather small, one can guesstimate that the sample size n needed to give adequate power to the test to detect small discrepancies (third decimal) from ϕ_0 is likely to be huge.

How does this frequentist set up relate to the 'hypothesis':

$$h-\text{ Joan has the disease,} \tag{11}$$

around which the base-rate fallacy example revolves? The assignment of probability to h is a giveaway that it's **not a legitimate frequentist hypothesis**, since in that context hypotheses are framed in terms of $\boldsymbol{\theta}$, which is assumed to be an unknown constant, *not* a random variable! The difficulty, however, is to disentangle the connections between h, the data $\mathbf{z}_0$ and $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

In contrast to (9), h in (11) assumes that ϕ is *known* [ϕ=.004 (table 1)] and poses the question whether, Joan — a particular member of the target population — has the disease or not. Due to the ambiguity as to the status of h vis-a-vis $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, one needs to distinguish between two different cases.

Case 1: Joan is *randomly selected* from the target population. In this case h concerns the *event,* say (X_{13}=1) and represents an element of the random sample $\mathbf{Z}$:= $[X_{13}, Y_{13}]$. In view of the fact that h does not pertain to the data-generating mechanism, it is *not* a legitimate frequentist hypothesis; see Mayo (1997a). Indeed, in this case $[x_{13}, y_{13}]$=$[1, 1]$ constitutes a single observation from a size n sample $\mathbf{Z}$.

Case 2: Joan is *not* randomly selected from the target population, e.g. she requested the test. In this case the single observation $[x_{13}, y_{13}]$=$[1, 1]$ is no longer legitimate for $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. Indeed, the event h (x_{13}=1) lies outside the intended scope of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, which is to provide an idealized description of the disease's incidence in the target *population,* and not the affliction of a particular individual; for the latter one needs to specify a different statistical model (see Spanos,

2009). h is illegitimate as an event in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ because $[x_{13}, y_{13}]$=$[1, 1]$ does not constitute a 'typical' realization of $[X_{13}, Y_{13}]$; purposeful selection invalidates the IID assumptions underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. Any attempt to interpret $[x_{13}, y_{13}]$=$[1, 1]$ as an instantiation of the *generic event* $(X$=$1, Y$=$1)$ will introduce statistical misspecifications that often lead to unreliable inferences.

Statistical adequacy — the validity (vis-a-vis data $\mathbf{z}_0$) of the IID assumptions underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ — ensures that the *actual* error probabilities are approximately equal to the *nominal* ones, rendering the reliability of inference ascertainable; see Mayo and Spanos (2004). Applying a .025 significance level test when its actual type I error probability is closer to .99 will lead the inference astray with certainty!

In summary, $h-$ Joan has the disease, does not constitute a legitimate frequentist hypothesis because, at best, it pertains to an event [not to the data-generating mechanism], and at worst it lies outside the intended scope of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

2.2.2 Error Probabilities vs. Conditional Probabilities of Events

Focusing on the best case scenario, it is clear that when Joan is randomly selected, the event $[x_{13}, y_{13}]$=$[1, 1]$ constitutes a single observation from $\mathbf{Z}$. It is well-known, however, that with n=1 no *consistent* (minimally reliable) estimator or test concerning $\boldsymbol{\theta}$ is possible; see Cox and Hinkley (1974).

In light of this, how do the proponents of the base-rate argument make their case that frequentist testing is vulnerable to the fallacy?

Their argument is that the hypothesis h: $(X$=$1)$ has 'passed a severe test' on the basis of the following two conditional probabilities (table 1):

$$\mathbb{P}(Y\text{=}1|X\text{=}1)\text{=}.8, \qquad \mathbb{P}(Y\text{=}1|X\text{=}0)\text{=}.00002, \tag{12}$$

the *false positive* and *false negative probabilities*, respectively. As the critics of severity argue, since $\mathbb{P}(X$=$1)$=.0000001 (the *prior* probability of h) is low, the conditional (*posterior* or *epistemic*) probability:

$$\mathbb{P}(X\text{=}1|Y\text{=}1)\text{=}\tfrac{.00000008}{.00002}\text{=}0.004, \tag{13}$$

is also low, and thus, on the basis of (13) Joan's positive result gives very little reason to believe h:=$(X$=$1)$, despite h's passing a severe test; see Achinstein (2009), p. 182.

Mayo (1997a-b, 2005) has called repeatedly into question the basic claim that 'h has passed a severe test' on the basis of the probabilities in (12)-(13) on several grounds. Howson (2000), however, maintains:

> "Mayo discusses the example in Mayo 1997, but I cannot see that she mitigates
>
> in any way its force." (p. 54); see also Howson and Urbach (2005), p. 25.

Focusing on the most telling of such grounds: the conditional probabilities (12)-(13) have **nothing to do** with any proper pre-data or post-data error probabilities.

To be more specific, the post-data severity functions have three arguments, a test T, data $\mathbf{z}_0$ and a frequentist hypothesis or a claim H, and *none* of them is present in the base-rate argument. This argument replaces a frequentist hypothesis with an event, the test and its error probabilities with conditional probabilities among events, and the data, at best, amount to a single observation (n=1). That is, the critics use a purely probabilistic argument to call into question an argument based on error probabilities which are statistical in nature.

One can go further and call into question the base-rate argument on other grounds. Even when h is the legitimate event (X=1) in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ – Joan has been selected randomly from the particular population – the claim that $\mathbb{P}(X{=}1)$ and $\mathbb{P}(X{=}1|Y{=}1)$ represent the prior and posterior probabilities of h is unwarranted on Bayesian statistics grounds as well, since these assignments have nothing to do with unknown parameters, their priors or any likelihoods.

▶ $\mathbb{P}(X{=}1|Y{=}1)$ in (13) is nothing more than a **deductive calculation** within the context of a **known statistical model** (table 1).

Worse, any attempt to associate $\mathbb{P}(X{=}1)$ and $\mathbb{P}(X{=}1|Y{=}1)$ with sub-groups of the original target population as a basis of any form of inductive inference is misplaced. One can test frequentist hypotheses about $\theta_1{=}\mathbb{P}(X{=}1)$ and $\mathbb{P}(X{=}1|Y{=}1){=}\phi$ in the context of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in (7), but that requires a proper data set $\mathbf{z}_0$ associated with $n > 1$ randomly selected individuals from the target population.

Mayo (2009) elaborates further on how misleading is the claim that h passes a severe test with data $[x_{13}, y_{13}]{=}[1, 1]$ on the basis of the conditional probabilities in (6).

2.3 False Positive and False Negative rates?

The medical test is conflated with a *proper frequentist test* using a beguiling analogical argument concerning the former's false positive and false negative conditional probabilities as being equivalent to the type I and II error probabilities (Howson, 2000). The analogy misrepresents frequentist testing because there is no such a thing as a *generic* type I and II error probability for a frequentist test. This is because the false positive/negative rates for medical devices and procedures are usually established mainly by the manufacturers of medical equipment before they make them available. These devices are *calibrated* by running numerous tests with specimens of blood, urine, etc. that are known to be positive or negative and evaluate their false positive/negative rates. No such calibrations are possible for frequentist tests for several reasons.

First, frequentist testing inference is *local* and *data specific* in the sense that it depend crucially on the particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the relevant test $(d(\mathbf{X}),\ C_1)$, which includes the framing of the hypotheses and the particular data $(\mathbf{x}_0)$. Hence, error probabilities are also local in this sense. When applying a test $(d(\mathbf{X}),\ C_1)$ using data $\mathbf{x}_0$, the primary aim is to learn about θ^* that could have generated $\mathbf{x}_0$. In contrast. a medical device or procedure is not local to the specific medical facility,

but calibrated for any such facility, irrespective of whether it will be used for 1 or 10^{10} tests. Frequentist error probabilities depend crucially on the sample size. Any attempt to identify the false positive/negative with *asymptotic* error probabilities, as being more generic, will not work because the power of any half-decent (consistent) N-P test goes to 1 as the sample size n goes to infinity; see Lehmann (1986).

Second, a medical testing devices or procedures aim to *prognosticate* the occurrence of two events [Joan has the disease or not], but a frequentist test aims to guide the learning about the true data-generating mechanism $\mathcal{M}^*(\mathbf{x})$ by probing the data for different values $\theta \in \Theta$; where Θ is usually uncountably infinite! Hence, the use of *conditional probabilities* associated with particular *events* by the base-rate argument is a far cry from proper error probabilities pertaining to inference procedures. Why? the type I and II error probabilities and the power of the test are *never* conditional on a hypothesis — they are *evaluated under different hypothetical values* of θ and they invariably involve an infinity of events (tail areas). For example, the type I error probability involves all outcomes $\mathbf{x}\in\mathbb{R}^n_X$ such that $d(\mathbf{x}) > c_\alpha$.

Third, the *ampliative* dimension of frequentist induction, in the sense of 'learning from data' about the underlying data-generating mechanism, is completely absent. When viewed in the context of frequentist induction, $\mathbb{P}(X{=}1|Y{=}1)$ amounts to a deductive calculation of conditional probabilities associated with particular events within the context of a *known* model (table 1). This is the reason why from that perspective the base-rate fallacy arguments appears to rely on $n{=}1$, and there are no unknown parameters in table 2; totally at odds with the implicit bivariate Bernoulli model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ in (7).

3 Positive Predictive Value (PPV)

The claim by Ioannidis (2005) that 'most published research findings are false' has been particularly influential in blaming the abuse of significance testing as the main culprit.

His argument revolves around a posterior probability measure, the Positive Predictive Value (PPV), adapted from medical screening that aims to evaluate the reliability of medical diagnostic tests for detecting a disease in patients that revolves around the notions of 'false positive' and 'false negative' for the screening devices; see Fletcher and Fletcher (2005). The PPV is defined in terms of H_0: no disease, $F{=}H_0$ is false, $R{=}$test rejects H_0, and takes the Bayesian probability formulation:

$$\text{PPV}=\Pr(F|R)=\frac{\text{number of true positive detections}}{\text{number of positive detections}}=\frac{\Pr(R|F)\Pr(F)}{\Pr(R|F)P(F)+\Pr(R|\overline{F})P(\overline{F})}, \quad (14)$$

where $\Pr(R|F)$ and $\Pr(\overline{R}|\overline{F})$ are referred to as 'sensitivity' and 'specificity', respectively. Sensitivity aims to measure the proportion of correct rejections of H_0 (when false), specificity aims to measure the proportion correct acceptances of H_0 when true.

Both depend crucially on 'prevalence' $\Pr(F)$ that aims to measure the proportion of false H_0 in a certain population. To make sense of the PPV in hypothesis testing, as opposed to medical screening devices, one needs to imagine that there is a population of null hypotheses for a particular discipline, a proportion of which are false, say 20%! .Unfortunately, the analogical reasoning behind the adaptation from medical screening gives the impression that $\Pr(R|F)$ and $\Pr(R|\overline{F})$ relate directly to the frequentist concepts of the 'power' and the 'significance level' of a test, respectively. This semblance, however, is highly misleading and fundamentally false.

First, frequentist error probabilities are never defined as *conditional* on the 'H_0 being true or false', because the latter do not constitute legitimate events in the context of a prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, upon which one can condition. In the context of frequentist testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, '$H_0$ is true or false' represent *hypothetical scenarios* under which the sampling distribution of the test statistic $(d(\mathbf{X}))$ is evaluated.

Hence, the claim: "... for all practical purposes in my view, the p value, is indeed a probability conditional or conditioned on an assumption, the null hypothesis." (Schneider, 2018) bespeaks ignorance of basic probability theory; one can condition only on events and random variables, not assumptions. Moreover, pretending that it is a matter of notational ignores the fact that assigning probabilities to θ via 'H_0 is true or false' is illegitimate in the context of frequentist testing.

Second, in frequentist testing there is no such thing as **discipline wide false positive/negative** proportions that revolve around generic tests and generic null hypotheses analogous to medical screening devices. One can assert that the false positive of this screening device is 7% and the prevalence of this illness for this population is 20%, but the analogical reasoning used to transfer such notions to frequentist testing is completely misplaced for several reasons.

Frequentist testing is *local* in the sense that it depend crucially on the particular $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, the relevant test $(d(\mathbf{X}),\ C_1)$ and the particular data $(\mathbf{x}_0)$, including n; see Spanos (2013). Assuming that a certain proportion of the 'effects' tested in a particular field, say $\Pr(F)$=.2, are expected to be 'truly' non-null relates to what Bayesians call the 'base rate', which is meaningless in the context of frequentist testing; see Spanos (2010b). Similarly, the power of a test is never a *point probability* chosen by cherry-picking a value in Θ_1=$\Theta-\Theta_0$. It calibrates the capacity of the test to detect different discrepancies in Θ_1.

Third, ensuring that every practitioner in a particular discipline refrains from any form abuse (p-hacking, multiple testing, cherry-picking, etc.) or misinterpretation of p-values, will be a good starting point, but nowhere near enough to guarantee that the end result will be trustworthy empirical evidence. This becomes obvious when the main sources of untrustrworthiness are recalled: (i) statistical misspecification, (ii) poor implementation of inference procedures, and (iii) unwarranted evidential interpretations of their inferential results.

Finally, the bias inducing abuses of the p-value, finger pointed as the main culprit,

take place at the level of an **individual study**, based on a particular statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. In contrast, the PPV postulates implicitly an imaginary meta-model of discipline-wide null hypotheses and decisions, and assigns a posterior measure of *untrustworthiness by association* to the overall performance of diagnostic screening in that field that revolves around $\Pr(F)$, the proportion of false null hypotheses; a meaningless notion in the context of frequentist testing. Unfortunately, this meta-model is not just imaginary, it defines the inductive premises of inference in a way that renders its appropriateness a matter of speculation.

4 Summary and conclusions

The above discussion has demonstrated that the base-rate fallacy argument is riddled with obfuscations and false analogies which stem from inadequate understanding of frequentist testing anchored on the notion of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$. The same applies to the PPV with the additional problem that establishing the reliability of inference is not at level of the individual study, but there is a field-wide false positive and negative rates that render an individual study 'guilty of untrustworthy evidence by association'!

Despite the apparent similarities between the Harvard Medical School test example and a frequentist test, it was pointed out that the two are fundamentally different. The medical test and its false positive and negative probabilities have only superficial resemblance to a proper frequentist test and its error probabilities, because the latter are crucially dependent on the sample size n but the former are inbuilt.

In particular, learning from data about $\mathcal{M}^*(\mathbf{x})$ is absent from the medical diagnostic procedures because the relevant error probabilities — around which the ampliative dimension of frequentist inference revolves — are replaced by known conditional probabilities among events that are calibrated once and for all before such procedures are used for diagnostic checking. Their probabilities represent simple deductive calculations within the context of a *known* prognostication procedure.