Modeling vs. Inference in frequentist statistics: securing the trustworthiness of empirical evidence

Aris Spanos

Preamble. The discussion that follows was prompted by the presentation "Testing in models that are not true" by Christian Hennig who posed two key questions: do the probabilistic assumptions imposed on one's data have to be fulfilled? Can this be checked? The answer to both questions is an emphatic YES!

1 Modeling vs. inference

1.1 Introduction

Model-based frequentist inference revolves around the concept of a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \Theta\subset\mathbb{R}^m,\ m<n, \tag{1}$$

where $f(\mathbf{x};\boldsymbol{\theta})$ is the distribution of the sample $\mathbf{X}:=(X_1,...,X_n)$, $\mathbb{R}^n_X$-sample space, and Θ-parameter space. Inference is framed in terms of mappings: $h(.)$: $\mathbb{R}^n_X\to\Theta$.

The *main objective* of model-based inference is to learn from data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$ by narrowing down Θ as much as possible, ideally to a single point $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ denotes the 'true' value of $\boldsymbol{\theta}$ in Θ. 'True' value $\boldsymbol{\theta}^*$ in this context is shorthand for saying that $\mathcal{M}^*(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, could have generated data $\mathbf{x}_0$; see Spanos and Mayo (2015). In practice, the ideal situation $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ is unlikely to be reached, except by happenstance, but that does not preclude learning from $\mathbf{x}_0$.

Example. The *simple Normal model* is specified by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_t\backsim \mathsf{NIID}(\mu,\sigma^2),\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ t\in\mathbb{N}:=(1,2,...,n,..) \tag{2}$$

where $\mathbb{R}:=(-\infty,\infty)$, $\mathbb{R}_+:=(0,\infty)$ and '$\mathsf{NIID}(\mu,\sigma^2)$' stands for 'Normal, Independent and Identically Distributed (IID), with mean μ and variance σ^2.

1.2 The modeling facet

The **modeling** facet comprises the process from selecting $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ to establishing its statistical adequacy: the validity of the probabilistic assumptions imposed on data $\mathbf{x}_0$, which includes the following stages.

(a) *Specification* [the initial choice of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$] based on the statistical systematic information (chance regularity patterns) exhibited by data $\mathbf{x}_0$ and the substantive questions of interest that influence the parameterization $\boldsymbol{\theta}$.

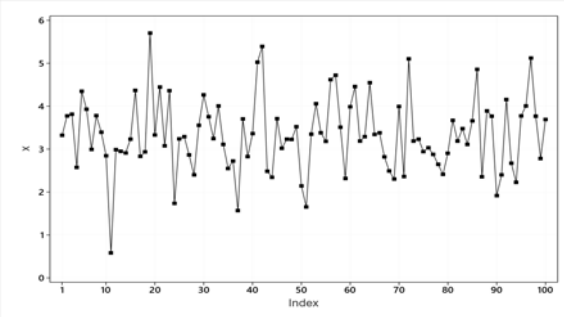

Fig. 1: t-plot of x_{1t}, $t=1,2,..,n$

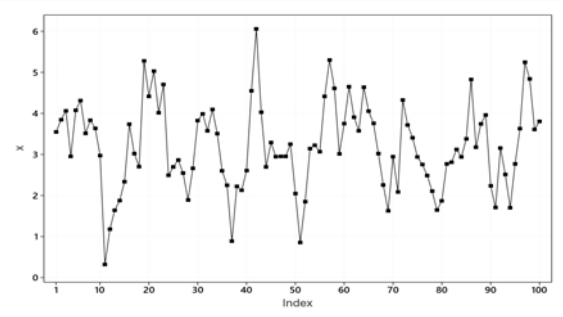

Fig. 2: t-plot of x_{2t}, $t=1,2,..,n$

Figures 1 and 2 show the t-plots of data sets $\mathbf{x}_1$ and $\mathbf{x}_2$, where data $\mathbf{x}_1$ exhibit NIID regularity patterns (appropriate for (2)), but data $\mathbf{x}_2$ exhibit irregular cycles, a form of dependence, which is a departure from [4]; see Spanos (2019).

(b) *Mis-Specification (M-S) testing* evaluates the (approximate) validity of the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ by probing for statistical systematic information in $\mathbf{x}_0$ unaccounted for by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Using the universal set $\mathcal{P}(\mathbf{x})$ of all possible models that could have given rise to $\mathbf{x}_0$, the M-S testing hypotheses of interest are generically framed as:

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \text{ vs. } H_1\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in [\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]\,. \tag{3}$$

In contrast, Neyman-Pearson (N-P) testing is based on the generic hypotheses:

$$H_0\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_0\} \text{ vs. } H_1\text{: } f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_0(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta_1\}, \tag{4}$$

where Θ_0 and Θ_1 constitute a partition of Θ; see figures (3-4). The obvious difference between N-P testing and M-S testing is that, as it stands, however, $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ is non-operational and needs to be particularized for viable M-S testing; see Spanos (2019).

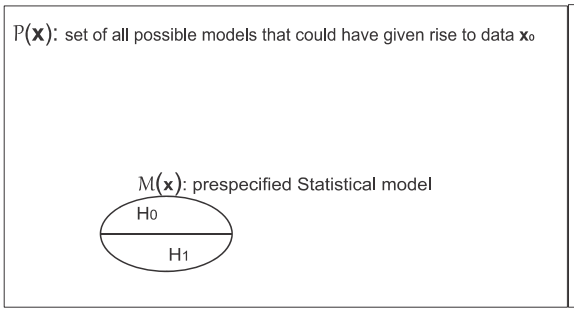

Fig. 3: Testing within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: N-P

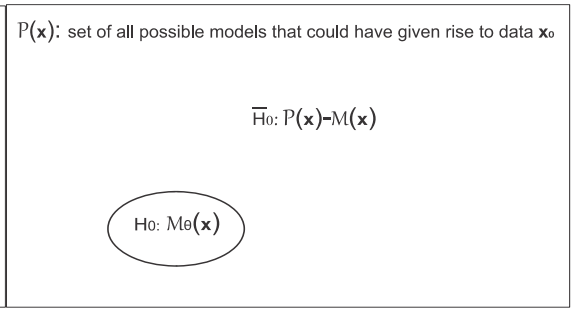

Fig. 4: Testing outside $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: M-S

The success of M-S testing depends crucially on the capacity of the designated departures to probe as much of the set $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ as possible. The traditional way is of limited scope since it encompasses $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ into a nesting model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$ which is often marginally broader (see fig. 5 and (21) for an example):

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\subset\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})\subset [\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]\,. \tag{5}$$

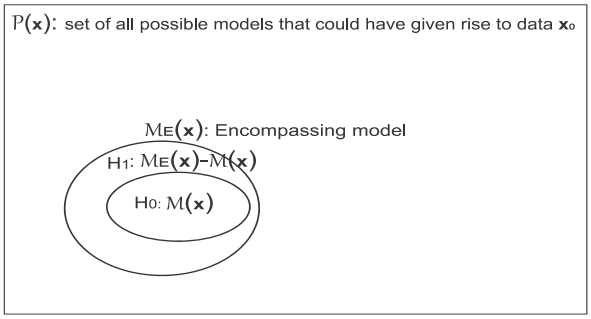

Fig. 5: M-S testing by encompassing

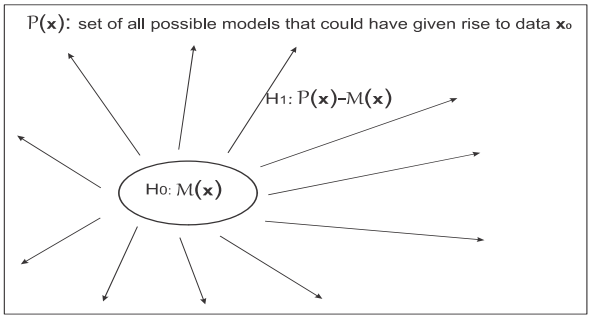

Fig. 6: M-S testing: directions of departures

The framing in (3)-(5) implies that rejecting a particular assumption as invalid, does not entail adopting the particular form of departure considered for two reasons.

(i) The generic null in M-S testing is H_0: all the model assumptions are valid, not just the one being considered.

(ii) The default generic alternative $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ is always much broader than the encompassing model $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{x})$.

A more effective way to operationalize $[\mathcal{P}(\mathbf{x})-\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ is in terms of joint M-S tests relying on auxiliary regressions based on the residuals of the estimated $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, with potential departures indicated by generic directions of departure; see fig. 6. For testing the validity of the assumptions of (2), [1] Normality, [2] Constant mean, [3] Constant variance, and [4] independence, the first auxiliary regression based on the residuals $\widehat{u}_t=(X_t-\overline{X}_n)$ specifies how departures from [2] and [4] might affect the mean (μ) by adding generic terms, such as trends and lags, to the original specification $X_t=\mu+u_t$:

$$\widehat{u}_t=\delta_0+\overbrace{\delta_1 t+\delta_2 t^2}^{\neg[2]}+\overbrace{\delta_3 x_{t-1}}^{\neg[4]}+\varepsilon_{1t}, \tag{6}$$

$$H_0\text{: } \delta_1=\delta_2=\delta_3=0 \text{ vs. } H_1\text{: } \delta_1\neq 0 \text{ or } \delta_2\neq 0 \text{ or } \delta_3\neq 0.$$

where $\neg[2]$ denotes the 'negation' of [2]. The second auxiliary regression uses similar generic terms to probe for departures from [3]-[4] that could potentially affect $E(u_t^2)=\sigma^2$:

$$\widehat{u}_t^2=\gamma_0+\overbrace{\gamma_1 t+\gamma_2 t^2}^{\neg[3]}+\overbrace{\gamma_3 x_{t-1}^2}^{\neg[4]}+\varepsilon_{2t}, \tag{7}$$

$$H_0\text{: } \gamma_1=\gamma_2=\gamma_3=0 \text{ vs. } H_1\text{: } \gamma_1\neq 0 \text{ or } \gamma_2\neq 0 \text{ or } \gamma_3\neq 0,$$

NOTE that the above choices of the additional terms in (6)-(7) are *only indicative* of the direction of departure from the model assumptions!

For the data in figures 1 [$\mathbf{x}_1$] and 2 [$\mathbf{x}_2$], the auxiliary regression (6) yields:

$$\begin{aligned}&\text{(i)}\ \ \widehat{u}_{1t}=\underset{(.363)}{-0.247}+\underset{(.153)}{.019}t+\underset{(.297)}{.075}t^2+\underset{(.102)}{.067}x_{1t-1},\ \ s=.882,\\ &\text{(ii)}\ \widehat{u}_{2t}=\underset{(.294)}{-1.888}+\underset{(.154)}{.020}t+\underset{(.298)}{.079}t^2+\underset{(.084)}{.583}^{*}x_{2t-1},\ \ s=.881,\end{aligned} \tag{8}$$

indicating no departures from assumption [2], [4] for data $\mathbf{x}_1$, but the statistical significance of the coefficient of x_{2t-1} indicates that assumption [4] is invalid for data $\mathbf{x}_2$; note that the numbers in brackets denote the standard errors. When assumptions [2]-[4] are valid, one can proceed to test [1] Normality. For the $\mathbf{x}_1$ data the Anderson and Darling test yields: A-D($\mathbf{x}_1$)=.471[.240], whose p-value, in square brackets, indicates no departure.

(c) *Respecification,* when any generic departures from the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ are detected, one needs to select a different statistical model $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$ that *does* account for all the statistical systematic information in $\mathbf{x}_0$. This is not just a 'patch up' of the original model using various forms of error-fixing strategies (Spanos, 2018). The detected departures from the original assumptions indicate that

there is systematic information (lingering chance regularities) in the residuals that has not been accounted for, and it calls for returning to the specification stage armed with the added information of the detected departures. The new respecified model $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$ is chosen to account for all the systematic statistical information, including the part the original model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ did not.

For instance, in the case of (ii) in (8), the only inference one can drawn on the basis of t-test for the coefficient of x_{2t-1}, $\tau(\mathbf{x}_2)=\frac{.583}{.084}=6.941[.00001]$, is that the p-value, in square brackets, indicates a departure from [4] independence, e.g. $Cov(X_t, X_s)\neq 0$, for $t\neq s$. Whether the generic form, based on $\delta_3 x_{2t-1}$, is appropriate when selecting the respecified model $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$ will be decided on the basis of its statistical adequacy, not on the basis of the form of the alternative specification contrived for the M-S testing relating to $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. That is, M-S testing informs one about generic departures from [1]-[4]: $\neg$[1] $X_t \backsim \neg\mathsf{N}(.,.)$, $\neg$[2] $E(X_t)=\mu(t)\neq\mu$, $\neg$[3] $Var(X_t)=\sigma^2(t)\neq\sigma^2$, $\neg$[4] $Cov(X_t, X_s)\neq 0$, for $t\neq s$, where $\mu(t)$ and $\sigma^2(t)$ are unknown functions of the index t. The particular form of departure used to operationalize the M-S test can only provide hints for selecting the respecified model $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$. That is, M-S testing does not entail the validity of the particular framing of the departure one happened to select. For this reason, claiming that the specification of $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$ is 'conditional on' the results of the M-S testing, as often argued by advocates of the decision-theoretic perspective, is fallacious. In this context, 'conditioning' is not just a throw away clause, but a formal probabilistic claim that amounts to institutionalizing the fallacies of acceptance and rejection; see section 4.

For data $\mathbf{x}_2$ in fig. 2, (2) is misspecified ([4] is invalid), but the respecified model:

$$\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})\text{: } X_t=\alpha_0+\alpha_1 X_{t-1}+u_t,\ (u_t|X_{t-1})\backsim\mathsf{NIID}(0,\sigma^2),$$

an AutoRegressive model with one lag [AR(1)], yields:

$$\underset{}{\widehat{x}_{2t}}=\underset{(.278)}{1.331}+\underset{(.084)}{.584}x_{2t-1},\ s=.874,$$

which turns out to be statistically adequate; its assumptions, [1] $(X_t|X_{t-1})\backsim\mathsf{N}(.,.)$, [2] Linearity $E(X_t|X_{t-1})=\alpha_0+\alpha_1 X_{t-1}$, [3] Homoskedasticity $Var(X_t|X_{t-1})=\sigma_0^2$, [4] Markov dependence, [5] $\boldsymbol{\phi}:=(\alpha_0,\alpha_1,\sigma_0^2)$ are t-invariant, are valid for $\mathbf{x}_2$ in fig. 2. That is, the presence of x_{2t-1} was not decided on the basis of the auxiliary regression (i) in (8), but on the basis of the statistical adequacy of $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{x})$. Indeed, one would reach the same conclusion, $Cov(X_t, X_s)\neq 0$, for $t\neq s$, using a form of the Durbin-Watson (D-W) test [D-W($\mathbf{x}_2$)=.834] based on a particular alternative, $u_t=\rho u_{t-1}+\varepsilon_t$, $|\rho|<1$, or even a nonparametric runs test $d_R(\mathbf{x}_2)=-5.266[.000]$; see Spanos (2019).

1.3 The inference facet

The crucial importance of the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ stems from the fact that it ensures that . $\boldsymbol{\theta}^*$ lies within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, i.e. $f(\mathbf{x};\boldsymbol{\theta}^*)\in\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and thus, optimal inference procedures could give rise to learning from data about $\boldsymbol{\theta}^*$. The modeling facet precedes the inference facet since the latter takes the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ as given,

and proceeds to establish the effectiveness (optimality) of the inference procedures. For instance, the probabilistic assumptions [1]-[4] comprising (2) predetermine the relevant *distribution of the sample* $f(\mathbf{x};\boldsymbol{\theta})$:

$$f(\mathbf{x};\boldsymbol{\theta})\overset{[4]}{=}\prod_{k=1}^{n}f_k(x_k;\boldsymbol{\theta}_k)\overset{[2]\text{-}[4]}{=}\prod_{k=1}^{n}f(x_t;\boldsymbol{\theta})\overset{[1]\text{-}[4]}{=}\prod_{k=1}^{n}\tfrac{1}{\sqrt{2\pi\sigma^2}}\exp(-\tfrac{(x_k-\mu)^2}{2\sigma^2}).$$

Using $f(\mathbf{x};\boldsymbol{\theta})$ one can define the likelihood function, $L(\boldsymbol{\theta};\mathbf{x}_0)=c(\mathbf{x}_0)\cdot f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, which provides the cornerstone of model-based $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ frequentist inference, since both Maximum Likelihood (ML) estimators and Likelihood Ratio (LR) tests play a crucial role.

Statistical inference is framed in terms of a statistic (estimator – point and interval – test, predictor), say $Y_n=g(X_1,X_2,...,X_n)$, and its the *sampling distribution* $f(y_n;\boldsymbol{\theta})$. For instance, the sampling distributions of the estimators:

$$\overline{X}_n=\tfrac{1}{n}\sum_{k=1}^{n}X_k\ \text{and}\ s^2=\tfrac{1}{(n-1)}\textstyle\sum_{k=1}^{n}(X_k-\overline{X}_n)^2,$$

of (μ,σ^2) in (2) depend crucially on assumptions [1]-[4]:

$$\text{(a)}\ \overline{X}_n\overset{[1]\text{-}[4]}{\backsim}\mathsf{N}\left(\mu,\tfrac{\sigma^2}{n}\right),\quad\text{(b)}\ \tfrac{(n-1)s^2}{\sigma^2}\overset{[1]\text{-}[4]}{\backsim}\chi^2(n-1). \tag{9}$$

(a) indicates that $\overline{X}_n$ is an excellent (unbiased, fully efficient, sufficient and consistent) estimator of μ; analogous properties hold for s^2; see Spanos (2019). (a)-(b) provide the basis for all inferences relating to $\boldsymbol{\theta}:=(\mu,\sigma^2)$.

For instance, hypothesis testing relating to μ is based on the sampling distributions of a test statistic, say $\tau(\mathbf{X})$, under the null and alternative hypotheses, based on a hypothesized value μ_0, say:

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

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s}\overset{\mu=\mu_0}{\backsim}\mathsf{St}(n-1), \tag{10}$$

$$\tau(\mathbf{X})=\tfrac{\sqrt{n}\left(\overline{X}_n-\mu_0\right)}{\sigma}\overset{\mu=\mu_1}{\backsim}\mathsf{St}(\delta_1;n-1),\ \delta_1=\tfrac{\sqrt{n}(\mu_1-\mu_0)}{\sigma},\ \text{for all}\ \mu_1>\mu_0, \tag{11}$$

where $\mathsf{St}(.)$ denotes a Student's t distribution, and δ_1 is the noncentrality parameter. The sampling distribution in (10) is used to evaluate the type I error probability and the p-value:

$$\text{(i)}\ \mathbb{P}(\tau(\mathbf{X})>c_\alpha;\ \mu=\mu_0)=\alpha,\ \text{(ii)}\ \mathbb{P}(\tau(\mathbf{X})>\tau(\mathbf{x}_0);\ \mu=\mu_0)=p(\mathbf{x}_0), \tag{12}$$

and that in (11) is used to evaluate the power of T_α:

$$\mathcal{P}(\mu_1)=\mathbb{P}(\tau(\mathbf{X})>c_\alpha;\ \mu=\mu_1),\ \text{for all}\ \mu_1>\mu_0, \tag{13}$$

which, in conjunction with α defines the optimality of test:

$$T_\alpha:=\{\tau(\mathbf{X})=\tfrac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha)=\{\mathbf{x}\text{: }\tau(\mathbf{x})>c_\alpha\}\}. \tag{14}$$

The above pre-data error probabilities (type I and power) are used to calibrate the effectiveness of the Neyman-Pearson (N-P) test T_α, which is Uniformly Most Powerful (UMP); see Lehmann and Romano (2005).

The form of the test statistic $\tau(\mathbf{X})$ in (14) poses to data $\mathbf{x}_0$ the question whether 'μ_0 is 'close enough' to μ^*' in the sense that the difference $(\mu^*-\mu_0)$ [reflected in $(\overline{X}_n-\mu_0)$] is 'statistically negligible', which is operationalized using the above error probabilities. The hypotheses H_0: $\mu\in\Theta_0$ vs. H_1: $\mu\in\Theta_1$ are framed as a partitioning of the parameter space Θ, since all values of μ in $\Theta:=\mathbb{R}$ are relevant for statistical inference purposes. This eliminates the possibility that μ^* belongs to a subset of Θ excluded by $\Theta_0\cup\Theta_1$. What ensures that μ^* is within the boundaries of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ in (2)? The statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. That is, the reliability of inference and the trustworthiness of the ensuing empirical evidence depends crucially on the statistical adequacy of the prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ ensures that the actual error probabilities approximate closely the nominal ones – the ones assuming the validity of [1]-[4]. For instance, departures from assumptions [1]-[4] undermine the reliability of inference by inducing substantial discrepancies between the actual and nominal error probabilities by distorting the sampling distributions in (10) and/or (11); see Spanos and McGuirk (2001). That is, one might be applying the t-test in (14) assuming α=.05, but the actual type I error probability might be closer to .7 due to certain departures from assumptions [1]-[4].

2 Misspecification and unreliable inferences

The data in figure 2 exhibit a distinct departure from assumption [4], in the form of irregular cycles, indicating Markov dependence:

$$Cov(X_i,X_j)=\sigma^2\rho^{|i-j|}\neq 0,\quad i\neq j,\ i,j=1,...n,... \tag{15}$$

If one were to use the data in figure 2 to estimate (2), all inferences will be unreliable because (15) will distort all sampling distributions used in section 1.

In particular, it can be shown (Spanos, 2009) that under (15) the sampling distributions of $\overline{X}$ and s^2 are no longer (a) and (b) above, but:

$$\text{(a)}^*\ \overline{X}\backsim \mathsf{N}\left(\mu,\tfrac{\sigma^2}{n}[c_n(\rho)]\right),\ \text{(b)}^*\tfrac{(n-1)s^2}{\sigma^2}\backsim [d_n(\rho)]\chi^2(n-1). \tag{16}$$

were the distortions in relation to (a) and (b) in square brackets are:

$$c_n(\rho)=1+\tfrac{2\rho(n(1-\rho)-1+\rho^n)}{n(1-\rho)^2}\quad d_n(\rho)=\tfrac{n(n-1)(1-\rho)^2-2\rho(n(1-\rho)-1+\rho^n)}{n^2(1-\rho)^2}$$

Naturally, (a)*-(b)* distort the distributions of $\tau(\mathbf{X})$ both under H_0 and H_1:

$$\begin{aligned}\tau(\mathbf{X})&=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s}\overset{\mu=\mu_0}{\backsim}[\sqrt{c_n(\rho)/d_n(\rho)}]\cdot\mathsf{St}(n-1),\\ \tau(\mathbf{X})&=\tfrac{\sqrt{n}(\overline{X}-\mu_0)}{s}\overset{\mu=\mu_1}{\backsim}[\sqrt{c_n(\rho)/d_n(\rho)}]\mathsf{St}\left(\delta_1;n-1\right),\ \text{for all } \mu_1>\mu_0.\end{aligned} \tag{17}$$

Depending on the 'true' value and sign of ρ, both the type I error probability and the power will be distorted, inducing a discrepancies between the actual and nominal error probabilities.

Numerical example. For μ_0=0, s=1, n=100, a nominal α=.05 and ρ=.6, the actual type I error probability is .207, and for a discrepancy γ=.3 from μ_0, the nominal power is .911 but the actual is .745; see Spanos (2009).

It is important to emphasize that the above discussion of a particular departure from a statistical model assumption, is only indicative of what can happen in practice with real data because: (i) it concerns only a particular form of departure from a single assumption, with (ii) all the relevant parameters assumed known. In practice, there is an infinite number of potential departures from each model assumption, and more than assumptions are often invalid.

3 Separating modeling from inference formally

In addition to the above arguments, the separation of the modeling and inference facets can be formally justified in the case of statistical models whose underlying distribution belongs to the *Exponential family*, which includes the Normal, exponential, gamma, chi-square, beta, Bernoulli, Poisson, Dirichlet, Wishart, inverse Wishart, geometric, binomial, multinomial, negative Binomial, etc.

As shown in Spanos (2010), in that case $f(\mathbf{x};\boldsymbol{\theta})$, in terms of which $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is specified, simplifies to:

$$f(\mathbf{x};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s},\mathbf{r};\boldsymbol{\theta})=|J|\cdot f(\mathbf{s};\boldsymbol{\theta})\cdot f(\mathbf{r}),\ \forall\,(\mathbf{s},\mathbf{r})\in\mathbb{R}_S^m\times\mathbb{R}_R^{n-m}, \tag{18}$$

where $|J|$ denotes the Jacobian of the transformation $\mathbf{X}\rightarrow(\mathbf{S}(\mathbf{X}),\ \mathbf{R}(\mathbf{X}))$,

(a) $\mathbf{R}(\mathbf{X})$:=$(R_1,...,R_{n-m})$, is a *complete sufficient* statistic,

(b) $\mathbf{S}(\mathbf{X})$:=$(S_1,...,S_m)$ a *maximal ancillary* statistic, and

(c) $\mathbf{S}(\mathbf{X})$ and $\mathbf{R}(\mathbf{X})$ are *independent*.

The clear separation of $f(\mathbf{s};\boldsymbol{\theta})$ and $f(\mathbf{r})$ in (18) stemming from (c) implies that inference can be based exclusively on $f(\mathbf{s};\boldsymbol{\theta})$, since the likelihood function reduces to $L(\boldsymbol{\theta};\mathbf{x}_0)$=$c(\mathbf{x}_0)\cdot f(\mathbf{s};\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$, where $f(\mathbf{r})$ becomes a component of the proportionality constant $c(\mathbf{x}_0)$. On the other hand, since $f(\mathbf{r})$ is free of $\boldsymbol{\theta}$, it can be used to validate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ since it has nothing to do with the substantive questions of interest framed in terms of $\boldsymbol{\theta}$. For example, in the case of the simple Normal model in (2):

$$\mathbf{S}(\mathbf{X})=(\overline{X}_n,s^2),\ \mathbf{R}(\mathbf{X})=(\widehat{v}_3,..,\widehat{v}_n),\widehat{v}_k=(\sqrt{n}(X_k-\overline{X}_n)/s),\ k=3,4,..,n,$$

where $\widehat{v}_k,\ k=1,2,..,n$, denotes the studentized residuals; see Spanos (2010). The above results also extend to statistical models whose inference procedures are based on asymptotic Normality.

These results make a very strong case that carrying on with the inference facet when the (approximate) validity of the model assumptions has not been established is a bad strategy that often would give rise to untrustworthy evidence.

4 Combined procedures revisited

The case for validating the premises of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ before any inference is drawn is based on separating the modeling and inference facets in 'learning from data $\mathbf{x}_0$ about phenomena of interest'. Treating model validation and inference as a single combined

inference problem is akin to conflating the construction of a boat to given specifications (modeling) with sailing it in a competitive race (inference). The two are related since the better the construction the more competitive the boat, but imagine trying to build a boat from a pile of plywood in the middle of the ocean while racing it.

To illustrate this claim consider the substantive hypotheses of interest:

$$H_0\text{: } \beta_1{=}0, \text{ vs. } H_1\text{: } \beta_1{\neq}0. \tag{19}$$

in the context of the traditional Linear Regression (LR) model:

$$\begin{gathered} y_t = \beta_0 + \beta_1 x_t + u_t, \ t = 1, 2, ..., n, \\ \text{[i] } (u_t|X_t{=}x_t) \backsim \mathsf{N}(.,.) \text{ [ii] } E(u_t|X_t{=}x_t){=}0, \\ \text{[iii] } E(u_t^2|X_t{=}x_t){=}\sigma^2, \text{ [iv] } E(u_t u_s|X_t{=}x_t){=}0, t > s. \end{gathered} \tag{20}$$

The Durbin-Watson (D-W) is an M-S test for appraising the validity of the no autocorrelation assumption [iv]. It frames the alternative hypothesis by particularizing $[\mathcal{P}(\mathbf{x}){-}\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ down to an encompassing LR model, $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, by assuming a particular form of departure, $u_t{=}\rho u_{t-1}{+}\varepsilon_t$, $|\rho| < 1$:

$$\begin{aligned} \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})\text{:} \quad & y_t{=}\beta_0 + \beta_1 x_t + u_t, \\ \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})\text{:} \quad & y_t{=}\beta_0 + \beta_1 x_t + u_t, \ u_t{=}\rho u_{t-1}{+}\varepsilon_t, \end{aligned} \tag{21}$$

where $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}) \subset \mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$. Hence, the M-S testing hypotheses are:

$$H_0\text{: } \rho = 0, \text{ vs. } H_1\text{: } \rho \neq 0. \tag{22}$$

The combined procedure re-frames the above M-S testing as a decision-theoretic estimation problem based on a hybrid model.

Step 1. View the M-S test based on (22) as a choice between $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ and $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, being particularized as a choice between $\widehat{\beta}_1$, the OLS estimator of β_1 in $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and $\widetilde{\beta}_1$, the GLS estimator of β_1 in $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$; see Spanos (1986).

Step 2. $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ are admixed by defining the *combined estimator*:

$$\widetilde{\widehat{\beta}}_1{=}\lambda\widehat{\beta}_1{+}(1{-}\lambda)\widetilde{\beta}_1, \ \lambda{=}\begin{cases} 1, & \text{if } H_0 \text{ is accepted,} \\ 0, & \text{if } H_0 \text{ is rejected.} \end{cases} \tag{23}$$

Step 3. The above decision-theoretic framing asserts that the relevant error probabilities are those based on the sampling distribution of $\widetilde{\widehat{\beta}}_1$, which is often non-Normal, biased and $Var(\widetilde{\widehat{\beta}}_1)$ is highly complicated; see Leeb and Pötscher (2005). As argued in Spanos (2017), the decision-theoretic setup misrepresents frequentist inference, and the above re-formulation of M-S testing as decision-theoretic estimation problem based on $\widetilde{\widehat{\beta}}_1$is problematic for several reasons.

First, the above combined procedure framing institutionalizes the fallacies of rejection and acceptance.

Reject H_0. Accepting H_1: $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ is a valid model constitutes a classic example of the fallacy of rejection. That is, rejecting H_0 on the basis of any M-S test entitles one to infer that the no-autocorrelation assumption is false, i.e.

$$E(u_t u_s|X_t{=}x_t) \neq 0 \text{ for } t{>}s, \ t{=}1, 2, ..., n.$$

On the other hand, accepting H_1 goes much further than that by assuming that the dependence is of a very specific form of autocorrelation, $u_t=\rho u_{t-1}+\varepsilon_t$:

$$E(u_t u_s|X_t=x_t)=(\rho^{|t-s|}/(1-\rho^2))\sigma_\varepsilon^2,\ t,s=1,2,...,n. \tag{24}$$

The validity of all the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ (not just 24) needs to be established separately by thoroughly testing all its assumptions. Hence, in a M-S testing one should *never* accept the alternative without further testing; see Spanos (2018).

Accept H_0. This procedure is also equally vulnerable to *the fallacy of acceptance*. It is possible that the particular M-S test did not reject the particular assumption or assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ because it had very low power to detect an existing departure; the D-W test is not a good M-S test for several reasoning; see Spanos and McGuirk (2001). In practice this can be remedied using a combination of joint tests together with a variety of additional M-S tests to cross-check the results.

Second, the best case scenario for the combined procedure is that one of the two model, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ or $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$, is statistically adequate for data $\mathbf{z}_0$:=$\{(x_t,y_t),\ t=1,2,..,n\}$. In such a case the error probabilities based on the sampling distribution of $\widetilde{\widehat{\beta}}_1$ will also be misleading. The situation is akin to mixing drinking with bath water; the result will always be bath water. Hence, any attempt to render $\widetilde{\widehat{\beta}}_1=\lambda\widehat{\beta}_1+(1-\lambda)\widetilde{\beta}_1$ a solution of the unreliability of inference problem is ill-fated.

The worse case scenario for the combined procedure is when neither $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ nor $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ is statistically adequate for data $\mathbf{z}_0$. This implies that the 'true' parameter of interest β_1^* will not lie within either of these models, contravening the sufficient condition for rendering $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ 'good' estimators of β_1, and ensuring the reliability of inference. For instance, in the case where the dynamic LR model:

$$\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z}):\quad y_t=\alpha_0+\alpha_1 x_t+\alpha_2 x_{t-1}+\alpha_3 y_{t-1}+u_t, \tag{25}$$

is statistically adequate for data $\mathbf{z}_0$ is $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ are practically useless (biased and inconsistent) estimators of α_1 in (25), rendering any inferences relating to β_1, including (19), unreliable; see Spanos (1986), Spanos and McGuirk (2001). Note that the inappropriateness of $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ arises despite $\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})$ in (25) being only marginally more general than $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})$ in (21) since $\mathcal{M}_{\boldsymbol{\psi}}(\mathbf{z})\rightleftarrows\{\mathcal{M}_{\boldsymbol{\phi}}(\mathbf{z})\ \&\ \alpha_2=-\alpha_1\alpha_3\}$.

To summarize, instead of devising ways to circumvent the fallacies of rejection and acceptance and avoid erroneous entailments in M-S testing, the combined procedure institutionalizes these fallacies by recasting the original problem as an choice between two models (estimators) come what may; see Spanos (2010). That is, the above combined procedure, based on $\widetilde{\widehat{\beta}}_1$, is ill-conceived and highly questionable because it misrepresents M-S testing. Worse, it offers an alibi to practitioners who use the slogan "all models are wrong" since they do not constitute 'exact' pictures of reality, and hence it is perfectly legitimate to ignore model validation.

5 Models are not 'true'!

Any references to 'true models' disguises (intentionally or unintentionally) the important distinction between a **substantive model** $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, determined by some theory or theories, and its implicit **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, which comprises the set of probabilistic assumptions imposed (implicitly or explicitly) on data $\mathbf{x}_0$.

[a] statistical adequacy: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ account for the chance regularities in $\mathbf{x}_0$? or equivalently, are the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ valid for data $\mathbf{x}_0$? The underlying intuition is that if $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is statistically adequate for data $\mathbf{x}_0$, one can use it to generate (simulate) new data realizations, say $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, which will have the *same* probabilistic structure as the original data $\mathbf{x}_0$.

[b] substantive adequacy**:** does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ adequately accounts for (describes, explains, predicts) the phenomenon of interest? Substantive inadequacy arises from the narrowness of its scope, missing confounding factors, systematic approximation errors, etc.

The same distinction is also ignored by practitioners of statistics who invoke the slogan: "All models are wrong, but some are useful" attributed to Box (1979). These practitioners are seeking an alibi to side-step statistical model validation by misinterpreting the slogan as implying that 'statistical misspecification is inevitable, when in fact he was referring to substantive models: "Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model." Box (1979), p. 202. See also Mayo (2018).

In relation to statistical models, there is no such thing as a 'true' model since 'truth' for $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ refers to 'it could have generated $\mathbf{x}_0$'; see Spanos and Mayo (2015). The approximate validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be tested and established, without its probabilistic assumptions being "precisely true or valid". Its adequacy is established by probing for potential *generic departures* from its probabilistic assumptions of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ and eliminating them.

In contrast to $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, a substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ aims to approximate the actual mechanism underlying the phenomenon of interest that gave rise to data $\mathbf{x}_0$. Again, calling $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ 'not true' constitutes an unhelpful truism since, by definition, such models involve abstraction, simplification and idealization of the real-world phenomenon they aim to describe/explain. On the other hand, its 'adequacy' can be tested by evaluating its realisticness and broadness of scope in explaining such phenomena.

References

[1] Box, G.E.P. (1979) "Robustness in the strategy of scientific model building", pp. 201–236 in Launer, R.L. and G.N. Wilkinson, *Robustness in Statistics*, Academic Press, London.

[2] Leeb, H. and B.M. Pötscher (2005) "Model selection and inference: Facts and fiction", *Econometric Theory*, 18: 21-59.

[3] Lehmann, E.L. and J.P. Romano (2005) *Testing Statistical Hypotheses*, Springer, NY.

[4] Mayo, D.G. (2018) *Statistical Inference as Severe Testing: How to Get Beyond the Statistical Wars*, Cambridge University Press, Cambridge.

[5] Spanos, A. (2009) "Statistical Misspecification and the Reliability of Inference: the simple t-test in the presence of Markov dependence", *Korean Economic Review*, 25(2): 165-213.

[6] Spanos, A. (2010) "Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification", *Journal of Econometrics,* **158**: 204-220.

[7] Spanos, A. (2017) "Why the Decision-Theoretic Perspective Misrepresents Frequentist Inference", chapter 1, pp. 3-28, *Advances in Statistical Methodologies and Their Applications to Real Problems*, ISBN 978-953-51-4962-0.

[8] Spanos, A. (2018) "Mis-Specification Testing in Retrospect", *Journal of Economic Surveys*, **32**: 541–577.

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

[10] Spanos A. and D.G. Mayo (2015) "Error statistical modeling and inference: where methodology meets ontology", *Synthese*, 192(11): 3533-3555.

[11] Spanos, A. and A. McGuirk (2001) "The Model Specification Problem from a Probabilistic Reduction Perspective", *Journal of the American Agricultural Association*, **83**: 1168-1176.