Revisiting the Two Cultures in Statistical Modeling and Inference: the Statistics Wars and Their Potential Casualties

Aris Spanos [Virginia Tech, USA]

1. **Introduction**
   Paradigm shifts in statistics during the 20th century
2. **Karl Pearson's descriptive statistics** (1894-1920s)
   The original curve-fitting
3. **Fisher's model-based statistical induction** (1922)
   Securing statistical adequacy and the trustworthiness of evidence
4. ***Graphical Causal modeling** (1990s)
   Curve-fitting substantive models
5. ***The nonparametric turn for model-based statistics** (1970s)
   Replacing 'distribution' assumptions with non-testable assumptions
6. **Data Science (Machine Learning and all that!)** (1990s)
   Curve-fitting using algorithmic searches
7. **Summary and Conclusions**
   Potential casualties of the statistics wars

1 Introduction

Breiman (2001): "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown."

During the 20th century **statistical modeling and inference** experienced several **paradigm shifts**, the most notable being:

Karl Pearson's descriptive statistics (– 1920s), **Fisher's model-based statistics** (1920s), **Nonparametric statistics** (1970s), **Graphical Causal modeling** (1990s), and **Data Science (**Machine Learning, Statistical Learning Theory, etc.**)** (1990s).

Key points argued in the discussion that follows

- The discussions on **non-replication** overlook a key contributor to **untrustworthy evidence**, **statistical misspecification**: invalid probabilistic assumptions imposed (explicitly or implicitly) on one's data $\mathbf{x}_0{:=}(x_1, ..., x_n)$.
- **There is a direct connection** between Karl Pearson's descriptive statistics, Nonparametric statistics and the Data Science curve-fitting.

- All three approaches **rely on** (i) **curve-fitting**, (ii) **goodness-of-fit/prediction** measures, and (iii) **asymptotic inference results** (as $n \to \infty$) based on **non-testable** probabilistic/mathematical assumptions.
- **The Curve-Fitting Curse:** when empirical modeling relies on curve-fitting of mathematical functions with a sufficiently **large number of parameters** to fine-tune (e.g. neural networks, orthogonal polynomials), one will **always find a 'best'** model on goodness-of-fit/prediction grounds, even if that model is **totally false**. Worse, one will be **oblivious to the fact** that such a 'best' model will commonly yield untrustworthy evidence!
- '**Best' goodness-of-fit/prediction**, i.e. 'small' residuals/prediction errors relative to a particular **loss function**, is neither necessary nor sufficient for trustworthy evidence! What ensures the latter is the statistical adequacy (approximate validity) of the the invoked statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ comprising the probabilistic assumptions imposed one one's data Spanos (2007).
- **Trustworthy evidence** stems from procedures whose actual error probabilities **approximate 'closely'** the nominal ones – derived by presuming the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. That is, the trustworthiness of evidence originates in the relevant **error probabilities** as they relate to the **severity principle**.

All approaches to statistics require three basic elements:

(i) **substantive questions** of interest–however vague or highly specific,

(ii) **appropriate data $\mathbf{x}_0$** to shed light on these questions (learn from $\mathbf{x}_0$),

(iii) **probabilistic assumptions** imposed (implicitly or explicitly) on the observable process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$. These are the assumptions that **matter for statistical inference purposes**, and NOT those of any **error terms**.

Key differences of alternative approaches to statistics

[a] **Inductive premises**: their framing of the inductive premises (probabilistic assumptions imposed on the data), and the interpretation of the selected model.

[b] **Model choice**: the selection of the 'best' model for the particular data.

[c] **Inductive inference**: the underlying inductive reasoning and the nature and interpretation of their inferential claims.

[d] **Substantive vs. statistical** information/model: how they conciliate the substantive (theory-based) and statistical (data-based) information.

2 Karl Pearson's descriptive statistics

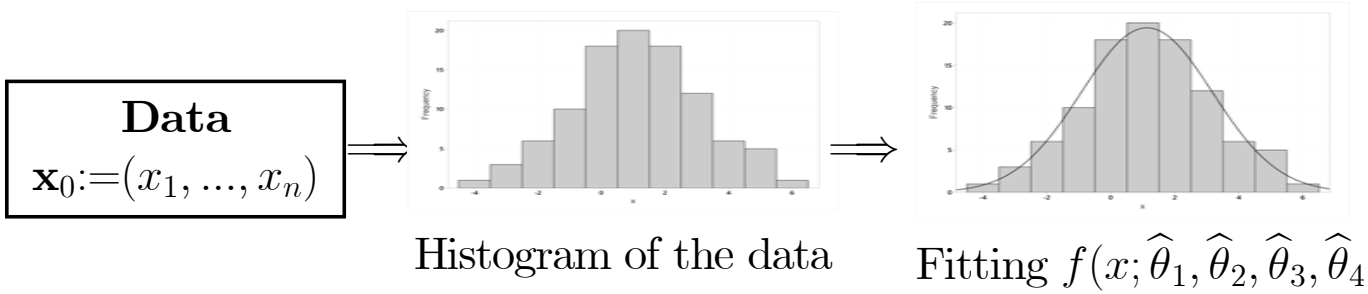


Diagram 1: Karl Pearson's approach to statistics

One begins with the raw data $\mathbf{x}_0:=(x_1, ..., x_n)$, whose initial 'rough summary' takes the form of a **histogram** with $m \geq 10$ bins. To provide a more succinct descriptive summary of the histogram Pearson would use the first four raw moments of $\mathbf{x}_0$ to select a frequency curve within a particular family known today as *the Pearson family* ($\mathcal{F}_P$). Members of this family are generated by:

$$\mathcal{F}_P: \ \frac{d\ln f(x;\boldsymbol{\psi})}{dx}=[(x-\theta_1)/\left(\theta_2+\theta_3 x+\theta_4 x^2\right)], \ \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^4, \ x\in\mathbb{R}:=(-\infty,\infty), \qquad (2.0.1)$$

that includes several well-known distributions. $\mathcal{F}_P$ is characterized by the four unknown parameters $\theta:=(\theta_1,\theta_2,\theta_3,\theta_4)$ that are estimated using $\widehat{\mu}_k(\mathbf{x}_0)=\frac{1}{n}\sum_{t=1}^{n} x_t^k$, $k=1,2,3,4$, yielding $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0):=(\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3,\widehat{\theta}_4)$. $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)$ is used to select $f_0(x)\in\mathcal{F}_P$

based on the estimated curve $\widehat{f}(x;\widehat{\boldsymbol{\theta}}(\mathbf{x}_0))$ that 'best' fits the histogram using Pearson's (1900) **goodness-of-fit test**:

$$\eta(\mathbf{X})=\sum_{i=1}^{m}[(\widehat{f}_i-f_i)^2/f_i] \underset{n\to\infty}{\backsim} \chi^2(m). \tag{2.0.2}$$

What Pearson and his contemporaries did not appreciate sufficiently is that, irrespective of whether one is summarizing the data for **descriptive** or **inferential purposes**, one implicitly imposes probabilistic assumptions on the data. For instance, the move from the raw data $\mathbf{x}_0$ to a histogram invokes a 'random' (IID) sample $\mathbf{X}:=(X_1,...,X_n)$ underlying data $\mathbf{x}_0$, and so do the formulae:

$$\overline{x}_n=\tfrac{1}{n}\sum_{t=1}^{n}x_t,\ \widehat{\sigma}_x^2=\tfrac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x}_n)^2,\ \overline{y}_n=\tfrac{1}{n}\sum_{t=1}^{n}y_t,\ \widehat{\sigma}_x^2=\tfrac{1}{n}\sum_{t=1}^{n}(y_t-\overline{y}_n)^2,$$
$$\widehat{\rho}_{xy}=\left[\left(\sum_{t=1}^{n}(x_t-\overline{x}_n)(y_t-\overline{y}_n)\right)/\sqrt{\left[\sum_{t=1}^{n}(x_t-\overline{x}_n)^2\right]\left[\sum_{t=1}^{n}(y_t-\overline{y}_n)^2\right]}\right],$$

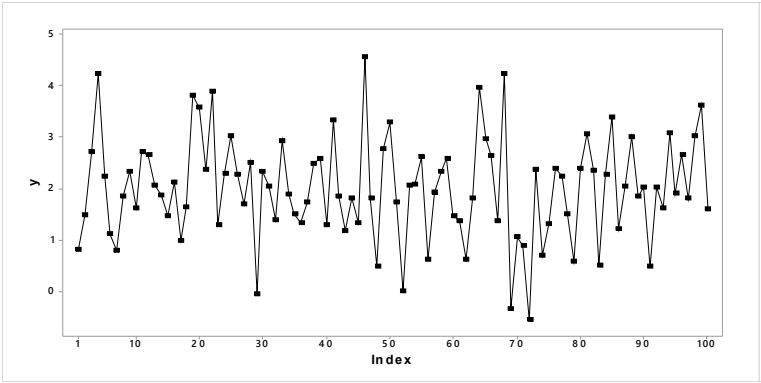
when estimating $E(X_t)$, $Var(X_t)$, $Corr(X_t,Y_t)$, etc.; see Yule (1926).

▶ Charging Karl Pearson with ignorance will be **anachronistic** since the theory of **stochastic processes** needed to understand the concept of **non-IID** samples was framed in the late 1920s early 1930s; Khitchin and Kolmogorov!

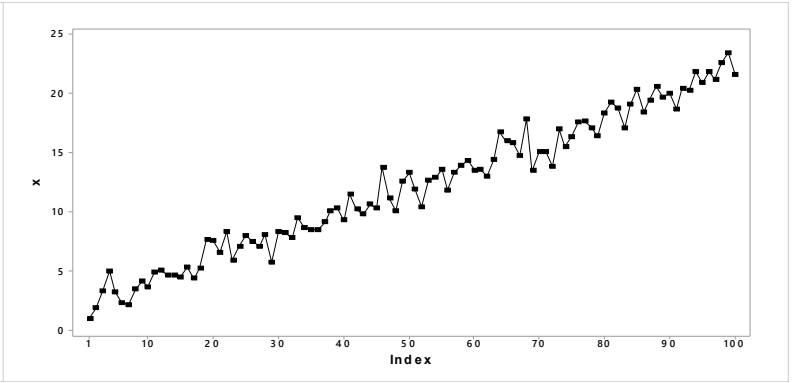
What about the **current discussions on the replication crisis?**

Amrhein, Trafimow, Greenland (2019) "Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication", **is ill-conceived**.

▶ The validity of the same probabilistic assumptions that **underwrite** the reliability of inferences also ensure the 'pertinence' of descriptive statistics.



Case 1: t-plot of IID data $\mathbf{y}_0$



Case 2 (ID false): t-plot of data $\mathbf{x}_0$

Case 1: Consistent (*valid*)

$$\overline{y}=\tfrac{1}{n}\sum_{t=1}^{n} y_t=2.03 \overbrace{[E(Y_t)=2]}^{\text{true}},$$

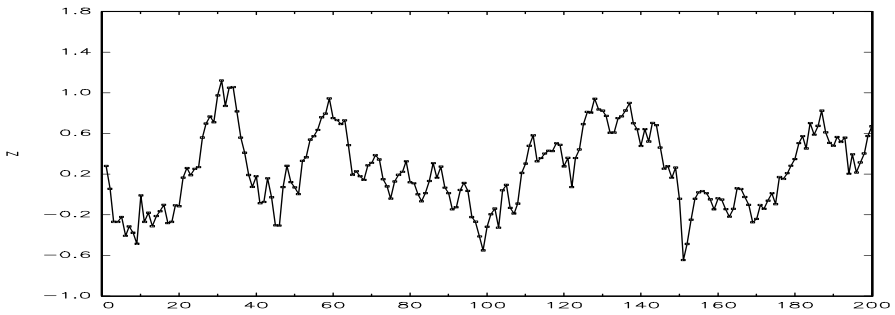
$$s_y^2=\tfrac{1}{n}\sum_{t=1}^{n}(y_t-\overline{y})^2=1.01 \overbrace{[Var(Y_t)=1]}^{\text{true}}$$

Case 2: Inconsistent (*spurious*)

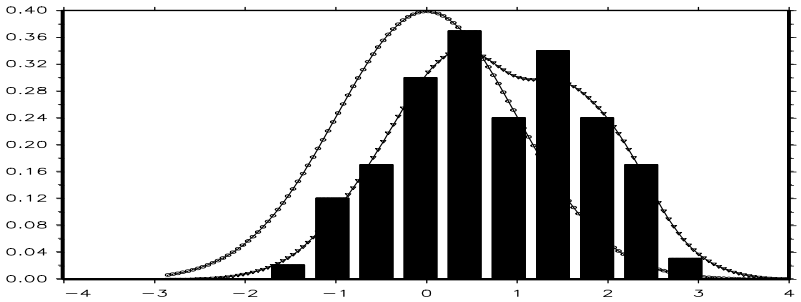
$$\overline{x}=\tfrac{1}{n}\sum_{t=1}^{n} x_t=12.1 \overbrace{[E(X_t)=2-.2t]}^{\text{true}}$$

$$s_x^2=\tfrac{1}{n}\sum_{t=1}^{n}(x_t-\overline{x})^2=34.21 \overbrace{[Var(X_t)=1]}^{\text{true}}$$

Consider **case 3** where the **Independence** assumption is invalid.



Case 3 (**I** false): t-plot of data $\mathbf{z}_0$



Case 3: Histogram of data $\mathbf{z}_0$

▶ When the IID assumptions are invalid for $\mathbf{x}_0$, not only the descriptive statistics, but also the estimated frequency curve chosen $f(x; \widehat{\boldsymbol{\theta}}(\mathbf{x}_0))$, will be highly misleading.

3 Fisher's Model-based frequentist approach

Fisher (1922) recast Pearson's curve-fitting into modern **model-based statistical induction** by viewing the data $\mathbf{x}_0$ as a 'typical realization' of a **parametric statistical model**, generically defined by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for}\ \Theta\subset\mathbb{R}^m,\ m<n. \tag{3.0.3}$$

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

$$X_t\backsim\mathsf{NIID}(\mu,\sigma^2),\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\Theta:=(\mathbb{R}\times\mathbb{R}_+),\ x_t\in\mathbb{R},\ t\in\mathbb{N}\}, \tag{3.0.4}$$

$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is framed in terms of probabilistic assumptions from 3 broad categories:

(D) Distribution	**(M) Dependence**	**(H) Heterogeneity**
Normal	Independence	Identically Distributed
Beta	Correlation	Strict Stationarity
Gamma	Markov	Weak Stationarity
Bernoulli	Martingale	Separable heterogeneity
⋮	⋮	⋮

assigned to the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$.

These assumptions determine the joint distribution $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, of the sample $\mathbf{X}:=(X_1,...,X_n)$, including its parametrization $\boldsymbol{\theta}\in\Theta$, as well as the likelihood function $L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\boldsymbol{\Theta}$; see Spanos (1986).

Fisher proposed a complete reformulation of statistical induction by **modeling the statistical Generating Mechanism (GM)** $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})]$ framed in terms of the observable stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$.

Fisher (1922) asserts that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is chosen by responding to the question: "Of what population is this a random sample?" (p. 313), and adding that "and the adequacy of our choice may be tested posteriori." (314). The 'adequacy' can be evaluated using Mis-Specification (M-S) testing; see Spanos (2006).

That is, $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is selected to account for **the chance regularities** in data $\mathbf{x}_0$, but its appropriateness is **evaluated** by M-S testing.

The **primary objective** of frequentist inference is to use the sample information, as summarized by $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$, in conjunction with data $\mathbf{x}_0$, to learn from data about $\boldsymbol{\theta}^*$- true value of $\boldsymbol{\theta}\in\Theta$; shorthand for saying that $\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\}$, $\mathbf{x}\in\mathbb{R}^n_X$, could have generated data $\mathbf{x}_0$.

Learning from $\mathbf{x}_0$ takes the form of 'statistical approximations' around $\boldsymbol{\theta}^*$, framed in terms of the sampling distribution, $f(y_n;\boldsymbol{\theta})$, $\forall y_n\in\mathbb{R}$, of a statistic (estimator,

test, predictor) $Y_n{=}g(X_1,...,X_n)$, derived using two different forms of reasoning via:

$$F_n(y){=}\mathbb{P}(Y_n\leq y){=}\underbrace{\int\int\cdots\int}_{\{\mathbf{x}:\ g(\mathbf{x})\leq y\}} f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x},\ \ \forall y{\in}\mathbb{R}, \tag{3.0.5}$$

(i) **Factual** (estimation and prediction): presuming that $\boldsymbol{\theta}{=}\boldsymbol{\theta}^*{\in}\Theta$, and
(ii) **Hypothetical** (hypothesis testing): various hypothetical scenarios based on different prespecified values of $\boldsymbol{\theta}$, under H_0: $\boldsymbol{\theta}{\in}\boldsymbol{\Theta}_0$ (presuming that $\boldsymbol{\theta}{\in}\boldsymbol{\Theta}_0$) and H_1: $\boldsymbol{\theta}{\in}\boldsymbol{\Theta}_1$ (presuming that $\boldsymbol{\theta}{\in}\boldsymbol{\Theta}_1$), where $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ *partition* $\boldsymbol{\Theta}$.

▶ **Crucially important**: (i) the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ ensures that $\boldsymbol{\theta}^*$ lies within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, and thus learning from data $\mathbf{x}_0$ is attainable.
(ii) **Neither form of frequentist reasoning** (factual or hypothetical) involves **conditioning on $\boldsymbol{\theta}$**, an unknown constant.
(iii) The decision-theoretic reasoning, **for all values of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$** ($\forall\boldsymbol{\theta}{\in}\boldsymbol{\Theta}$), undermines learning from data about $\boldsymbol{\theta}^*$, and gives rise to Stein-type paradoxes and admissibility fallacies; Spanos (2017).

▶ **Misspecification.** When $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is **misspecified,** $f(\mathbf{x};\boldsymbol{\theta})$ is incorrect, and this distorts $f(y_n;\boldsymbol{\theta})$, and often induces **inconsistency** in estimators and **size-**

able discrepancies between the actual and nominal error probabilities in Confidence Intervals (CIs), testing and prediction. This is why **Akaike-type model** selection procedures often go astray, since all goodness-of-fit/prediction measures presume the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; Spanos (2010).

How can one apply Fisher's **model-based statistics** when the empirical modeling begins with a **substantive model** $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$?

[i] Bring out the **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ implicit in $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$; there is always one that comprises solely the probabilistic assumptions imposed on data $\mathbf{x}_0$! It is defined as an unrestricted parametrization that follows from the probabilistic assumptions imposed on the process $\{X_t,\ t\in\mathbb{N}\}$ underlying $\mathbf{x}_0$ which includes $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ as a special case.

[ii] Relate the **substantive parameters** $\boldsymbol{\varphi}$ to $\boldsymbol{\theta}$ via restrictions, say $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$, ensuring that the restrictions $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ define $\boldsymbol{\varphi}$ **uniquely** in terms of $\boldsymbol{\theta}$.

Example. For the substantive model known as the **Capital Asset Pricing**:

$$\begin{aligned}
&\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z}): && (Y_t-x_{2t})=\alpha_1(x_{1t}-x_{2t})+\varepsilon_t,\ (\varepsilon_t|\mathbf{X}_t=\mathbf{x}_t)\backsim\mathsf{NIID}(0,\ \sigma^2_\varepsilon),\ t\in\mathbb{N},\\
&\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z}): && Y_t=\beta_0+\beta_1x_{1t}+\beta_2x_{2t}+u_t,\ (u_t|\mathbf{X}_t=\mathbf{x}_t)\backsim\mathsf{NIID}(0,\ \sigma^2_u),\ t\in\mathbb{N},\\
&\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}: && \beta_0=0,\ \beta_1+\beta_2-1=0,\ \text{where } \boldsymbol{\varphi}=(\alpha_1,\sigma^2_\varepsilon),\ \boldsymbol{\theta}=(\beta_0,\beta_1,\beta_2,\sigma^2_u).
\end{aligned}$$

[iii] **Test the validity** of H_0: $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})=\mathbf{0}$ vs. H_1: $\mathbf{g}(\boldsymbol{\varphi},\boldsymbol{\theta})\neq\mathbf{0}$ to establish whether the **substantive model** $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ belies data $\mathbf{z}_0$.

Main features of the Fisher model-based approach

[a] **Inductive premises**: $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ comprises a set of complete, internally consistent, and testable probabilistic assumptions, relating to the **observable** process $\{X_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{x}_0$, from the Distribution, Dependence and Heterogeneity categories. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is viewed as a statistical stochastic mechanism aiming to account for all the **chance regularity patterns** in data $\mathbf{x}_0$.

[b] **Model choice**: the appropriate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is chosen on **statistical adequacy grounds** using comprehensive **Mis-Specification (M-S) testing** to ensure that inferences are *reliable*: the **actual** $\simeq$ **nominal error probabilities**.

[c] **Inductive inference**: the interpretation of probability is frequentist and the underlying **inductive reasoning** is either **factual** (estimation, prediction) or **hypothetical** (testing) and relates to learning from data about $\boldsymbol{\theta}^*$.
The **effectiveness** (optimality) of inference procedures **is calibrated** using **error probabilities** based on a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$!

Regrettably, the replication crisis literature often **confuses** hypothetical reasoning with conditioning on H_0!

Diaconis and Skyrms (2018) claim (tongue-in-cheek) that p-value testers conflate $P(H_0|\mathbf{x}_0)$ with $P(\mathbf{x}_0|H_0)$: "The untutored think they are getting the probability of effectiveness [of a drug] given the data, while they are being given conditional probabilities going in the opposite direction." (p. 67)

▶ The 'untutored' **know** from basic probability theory that conditioning on H_0: $\theta{=}\theta_0$ is formally illicit since θ is neither an event nor a random variable!

[d] Substantive vs. statistical: the substantive model, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, is embedded into a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ via restrictions $\mathbf{g}(\boldsymbol{\theta},\boldsymbol{\varphi}){=}\mathbf{0}$, $\boldsymbol{\theta}{\in}\Theta$, $\boldsymbol{\varphi}{\in}\Phi$, whose rejection indicates that the substantive information in $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ **belies** $\mathbf{x}_0$!

▶ The above modeling strategy **[i]-[iv]** can be used to provide **sound statistical foundations** for **Graphical Causal Modeling** that revolves around substantive causal models $[\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})]$. It will enable a harmonious blending of the **statistical** with the **substantive information** without undermining the credibility of either and allow for probing the validity of causal information.

4 Graphical Causal (GC) Modeling

Quantifying a **Graphical Causal (GC) model** based on directed acyclic graphs (DAG) (Pearl, 2009; Spirtes et. al 2000) constitutes another form of **curve-fitting** a **priori postulated** substantive model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$.

An crucial **weakness** of the GC modeling is that the **causal information** (substantive) is usually treated as **established knowledge** instead of best-daresay conjectures whose **soundness** needs to be tested against data $\mathbf{Z}_0$.

► Foisting a DAG substantive model, $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$, on data $\mathbf{Z}_0$, will usually yield a **statistically** and **substantively misspecified model**!

This is because the **estimation** of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ invokes a set of probabilistic assumptions relating to the observable process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ underlying data $\mathbf{Z}_0$, the **implicit** statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ whose **adequacy** is unknown!

► Can one **guard against** statistical and **substantive misspecification**?

Embed the DAG model into the Fisher model-based framework

Step 1. Unveil the **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ **implicit** in the GC model.

Step 2. Establish the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ using comprehensive **M-S testing**, and **respecification** when needed.

Substantive (GC) model [a] Z - **confounder**	Substantive (GC) model [b] Z - **mediator**	Substantive (GC) model [c] Z - **collider**
$Y_k=\beta_0+\beta_1X_k+\beta_2Z_k+\varepsilon_{1k}$, $X_k=\alpha_0+\alpha_1Z_k+\varepsilon_{2k},\ k\in\mathbb{N}$	$Y_k=\beta_0+\beta_1X_k+\beta_2Z_k+\varepsilon_{1k}$, $Z_k=\gamma_0+\gamma_1X_k+\varepsilon_{3k},\ k\in\mathbb{N}$,	$Y_k=\delta_0+\delta_1X_k+\varepsilon_{4k},\ k\in\mathbb{N}$, $Z_k=c_0+c_1Y_k+c_2X_k+\varepsilon_{5k}$,
Statistical model for [a] $Y_k=a_{01}+a_{11}Z_k+u_{1k}$, $X_k=a_{02}+a_{12}Z_k+u_{2k}$,	Statistical model for [b] $Y_k=b_{01}+b_{11}X_k+u_{3k}$, $Z_k=b_{02}+b_{12}X_k+u_{4k}$,	Statistical model for [c] $Y_k=b_{01}+b_{11}X_k+u_{3k}$, $Z_k=b_{02}+b_{12}X_k+u_{4k}$,

Diagram 2: Functional Graphical Causal models

Step 3. Use a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ to address the **identification and estimation** of the structural parameters $\boldsymbol{\varphi}\in\boldsymbol{\Phi}$.

Step 4. Test the validity of the **overidentifying restrictions** stemming from $\mathbf{g}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0},\ \boldsymbol{\theta}\in\Theta,\ \boldsymbol{\varphi}\in\Phi$.

Excellent **goodness-of-fit/prediction** is relevant for **substantive adequacy**, which can only be probed after:

(i) establishing the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, and

(ii) evaluating the **validity** of the restrictions: H_0: $\mathbf{g}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$ vs. H_1: $\mathbf{g}(\boldsymbol{\theta},\boldsymbol{\varphi})\neq\mathbf{0}$.

Rejecting H_0 indicates that the substantive information in $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$ belies $\mathbf{x}_0$!

5 Nonparametric statistics & curve-fitting

Nonparametric statistics began in the 1970s extending Kolmogorov (1933a):

> when the sample $\mathbf{X}:=(X_1, X_2, ..., X_n)$ is IID, the empirical cdf $\widehat{F}_n(x)$ is a good estimator of the cdf $F(x)$, for n large enough.

Attempts to find good estimators for the **density function** $f(x),\ x\in\mathbb{R}$, led to:
(a) **kernel smoothing** and related techniques, including regression-type models,
(b) **series estimators** of $\widehat{f}(x)=\sum_{i=0}^{m}\beta_i\phi_i(x_k)$, where $\{\phi_i(x_k),\ i=1, 2, ..., m\}$ are **polynomials**, usually **orthogonal**; see Wasserman (2006).
A **nonparametric statistical model** is specified in terms of a broader family $\mathcal{F}$ of distributions (Wasserman, 2006):

$$\mathcal{M}_{\mathcal{F}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\psi}_n),\ f\in\mathcal{F}\},\ \boldsymbol{\psi}_n\in\boldsymbol{\Psi},\ \mathbf{x}\in\mathbb{R}^n_X,$$

where $\mathcal{F}$ is defined in terms of indirect & non-testable **Distribution** assumptions such as: (a) the **existence of moments** up to order $p \geq 1$, (see Bahadur and Savage, 1956, on such assumptions),
(b) **smoothness restrictions** on the **unknown** density function $f(x)$, $x\in\mathbb{R}_X$ (symmetry, differentiability, unimodality of $f(x)$, etc.).

Dickhaus (2018), p. 13: "Of course, the advantage of considering $\mathcal{F}$ is that the issue of model misspecification, which is often problematic in parametric models, is avoided." Really?

Nonparametric models always impose **Dependence** and **Heterogeneity** assumptions (**often** Independent and Identically Distributed (IID))!

What are the consequences of replacing $f(\mathbf{x};\boldsymbol{\theta})$ with $f(\mathbf{x};\boldsymbol{\psi}_n),\ f\in\mathcal{F}$?

The likelihood-based inference procedures are replaced by **loss function-based procedures** driven by mathematical approximations and goodness-of-fit measures, relying on asymptotic inference results at a high price in reliability and precision of inference since (i) the adequacy of $f(\mathbf{x};\boldsymbol{\psi}_n),\ f\in\mathcal{F}$ is impossible to establish, and (ii) the 'indirect' and non-testable distribution assumptions invariably contribute substantially to the imprecision/unreliability of inference.

▶ As argued by Le Cam (1986), p. xiv:

"... limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. Unfortunately, the approximation bounds we could get are too often too crude and cumbersome to be of any practical use."

6 Data Science (Machine Learning and all that!)

Big Data and Data Science includes Machine Learning (ML), Statistical Learning Theory (SLT), pattern recognition, data mining, etc.

As claimed by **Vapnik (2000)**: "Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, ... was replaced by a new one. This paradigm reflects a new answer to the fundamental question: What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations? In Fisher's paradigm the answer was very restrictive – one must know almost everything. Namely, one must know the desired dependency up to the values of a finite number of parameters. ... In the new paradigm ... it is sufficient to know some general properties of the set of functions to which the unknown dependency belongs." (ix).

In **Fisher's model-based** approach one selects $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ to account for the chance regularities in data $\mathbf{Z}_0$, and evaluates its validity before any inferences are drawn, respecifying it when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is misspecified. The form of dependence follows from the probabilistic assumptions of a statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$.

Example. Assuming that $\{Y_t,\ t\in\mathbb{N}\}$ is Normal, Markov and stationary, the dependence is an AR(1) model $Y_t=\alpha_0+\alpha_1 Y_{t-1}+u_t,\ (u_t|\sigma(Y_{t-1}))\backsim\mathsf{NIID}\left(0,\sigma^2\right),\ t\in\mathbb{N}$.

Machine Learning views **statistical modeling** as an **optimization problem** relating to how a machine can **'learn from data'**:
(a) **learner's input**: a domain set $\mathcal{X}$, a label set $\mathcal{Y}$,
(b) **training data** $\mathcal{X} \times \mathcal{Y}$: $\mathbf{z}_i{:=}(\mathbf{x}_i, y_i),\ i{=}1, 2, ..., n$,
(c) with an **unknown distribution** $f^*(\mathbf{z})$, and
(d) **learner's output**: $h(.)$: $\mathcal{X} \to \mathcal{Y}$.
The **learning algorithm** is all about **choosing** $h(.)$ **to approximate 'closely'** the true relationship $y_i{=}g(\mathbf{x}_i),\ i{\in}\mathbb{N}$, by minimizing the distance $\|h(\mathbf{x}) - g(\mathbf{x})\|$.

Barriers to entry? The **underlying probabilistic and mathematical framework** comes in the form of **functional analysis:** the study of infinite-dimensional **vector (linear) spaces** endowed with a **topology** (metric, norm, inner product) and a **probability measure**.
Example. The normed linear space $(C[a, b], \|\ .\ \|_p)$ of all real-valued continuous functions $g(x)$ defined on $[a, b]{\subset}\mathbb{R}$, with the p-norm (Murphy, 2022):

$$\|g\|_p = \left(\int_a^b |g(x)|^p\, dx\right)^{\frac{1}{p}},\ \text{or}\ \|g\|_p = (\textstyle\sum_{i=0}^n |g(x_i)|^p)^{\frac{1}{p}},\ p{=}1, 2, \infty. \qquad (6.0.6)$$

The mathematical approximation problem is transformed into an optimization in the context of a **vector space** employing powerful theorems such as the open

mapping, the Banach-Steinhaus, the Hahn-Banach theorems; see Carlier (2022). To ensure the **existence** and **uniqueness** of the optimization solution, the **approximation problem** is often embedded in a **complete inner product vector (linear) space** $(C[a,b], \| \ . \ \|_2)$ of real or complex-valued functions $h(\mathbf{X})$, defined on $[a,b] \subset \mathbb{R}$, with the 2-norm, also known as a **Hilbert space** of **square-integrable functions** $(E(|\mathbf{X}_t|^2))$, where $\{\mathbf{X}_t, \ t \in \mathbb{N}\}$ is a stochastic process. A Hilbert space generalizes the n-dimensional Euclidean geometry to an infinite dimensional **inner product space** that allows **lengths** and **angles** to be defined to render optimization possible.

Supervised learning (Regression). A typical example is a regression model:

$$y_t = g(\mathbf{x}_t; \boldsymbol{\psi}_n) + v_t,$$

where $g(\mathbf{x}_t; \boldsymbol{\psi}_n) \in \mathcal{G}$, where $\mathcal{G}$ is a family of smooth enough mathematical functions, is approximated using data $\mathbf{Z}_0 := \{(\mathbf{x}_t, y_t), t = 1, 2, ..., n\}$.

Risk functions. The problem is framed in terms of a **loss function:**

$$L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)), \ \forall \boldsymbol{\psi}_n \in \boldsymbol{\Psi}, \ \forall \mathbf{z} \in \mathbb{R}_Z^{nm},$$

(a) $\|g\|_2$: $L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)) = (y - g(\mathbf{x}; \boldsymbol{\psi}_n))^2$, (b) $\|g\|_1$: $L(y, g(\mathbf{x}; \boldsymbol{\psi}_n)) = |y - g(\mathbf{x}; \boldsymbol{\psi}_n)|$.

To render $L(y, g(\mathbf{x};\boldsymbol{\psi}_n))$ only a function of $\forall\boldsymbol{\psi}_n\in\boldsymbol{\Psi}$, $\forall\mathbf{z}\in\mathbb{R}_Z^{nm}$ is eliminated by taking expectations wrt the distribution of the sample $\mathbf{Z}$, $f^*(\mathbf{z})$-presumed **unknown,** to define a **risk function**:

$$R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))=E_{\mathbf{z}}(L(y, g(\mathbf{x};\boldsymbol{\psi}_n)))=\int_{\mathbf{z}} L(y, g(\mathbf{x};\boldsymbol{\psi}_n))f^*(\mathbf{z})d\mathbf{z},\ \forall\boldsymbol{\psi}_n\in\boldsymbol{\Psi}.$$

The statistical model implicit in **Data Science** is: $\mathcal{M}_{\mathcal{F}}(\mathbf{z})=\{f^*(\mathbf{z}),\ f\in\mathcal{F}(\boldsymbol{\psi})\}$, $\boldsymbol{\psi}_n\in\boldsymbol{\Psi}$, $\mathbf{z}\in\mathbb{R}_Z^{mn}$, and the ensuing **inference** revolves around the risk function using the **decision theoretic reasoning** based on $\forall\boldsymbol{\psi}_n\in\boldsymbol{\Psi}$.

Hence, Data Science tosses away all forms of **frequentist inference** apart from the **point estimation**. Since f^* is unobservable $R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))$ is estimated using basic **descriptive statistics:** $\widehat{E}(X_i)=\frac{1}{n}\sum_{i=1}^n X_i$!

Assuming that $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ is **IID** (often not stated explicitly!) one can use the arithmetic average

$\widehat{R}(f^*, g(\mathbf{z};\boldsymbol{\psi}_n))=\frac{1}{n}\sum_{i=1}^n L(y_i, g(\mathbf{x}_i;\boldsymbol{\psi}_n(\mathbf{x}_i))$, and then minimize it to yield a **consistent** estimator of $g(\mathbf{x};\boldsymbol{\psi}_n)$:

$$\widehat{g}_m(\mathbf{z};\widehat{\boldsymbol{\psi}}_n(\mathbf{x}))=\arg\min_{g\in\mathcal{G}}[\tfrac{1}{n}\textstyle\sum_{i=1}^n L(y_i, g(\mathbf{x}_i;\boldsymbol{\psi}_n(\mathbf{x}_i))],\qquad(6.0.7)$$

where $\widehat{y}_i=\widehat{g}_m(\mathbf{z}_i;\widehat{\boldsymbol{\psi}}_n(\mathbf{x}_i))$ minimizing (6.0.7) is inherently overparametrized!

Regularization. Depending on the dimension (effective number of parameters) of the class of **functions** $\mathcal{G}$, (6.0.7) will usually give rise to serious **overfitting** – near-interpolation! To reduce the inherent overparametrization problem ML methods impose ad hoc restrictions on the parameters, euphemistically known as **regularization,** to derive to $\widehat{y}_i{=}\widehat{g}_m(\mathbf{z}_i;\widehat{\boldsymbol{\phi}}_n(\mathbf{x}_i))$ via minimizing:

$$R_r(f^*, g(\mathbf{z};\boldsymbol{\psi}_n)) = R(f^*, g(\mathbf{z};\boldsymbol{\psi}_n)) + \lambda C(g(\mathbf{z};\boldsymbol{\psi}_n)), \tag{6.0.8}$$

where $C(g(\mathbf{z};\boldsymbol{\psi}_n))$ is often related to the algorithmic complexity of the class $\mathcal{G}$.
Example. For the LR model $Y_t{=}\boldsymbol{\beta}^\top\mathbf{x}_t + u_t$, $\boldsymbol{\beta}$ is estimated by minimizing:

$$\overbrace{\sum_{t=1}^{n}\left(Y_t{-}\boldsymbol{\beta}^\top\mathbf{x}_t\right)^2}^{\text{least-squares}}{+}\overbrace{\lambda_1\sum\nolimits_{i=1}^{p}|\beta_i| + \lambda_2\sum\nolimits_{i=1}^{p}\beta_i^2}^{\text{regularization term}}.$$

The idea is that regularization reduces the variance of $\widehat{\boldsymbol{\beta}}$ with only small increases in its bias, which improves prediction MSE $\sum_{t=n+1}^{n+N}(Y_t{-}\widehat{\boldsymbol{\beta}}^\top\mathbf{x}_t)^2$ (artificially!), at the expense of learning from data; Spanos (2017).

Probably Approximately Correct (PAC) **learnability** refers to 'learning' (computationally) about $f^*(\mathbf{z})$ using a polynomial-time algorithm (N^k) to chose

$g(\mathbf{z};\boldsymbol{\psi}_n)$ in $\mathcal{G}$, in the form of an upper bound (Murphy, 2022):

$$\mathbb{P}(\max_{g\in\mathcal{G}}|\widehat{R}(f^*,g(\mathbf{z};\boldsymbol{\psi}_n))-R(f^*,g(\mathbf{z};\boldsymbol{\psi}_n))|>\epsilon)\leq 2\dim(\mathcal{G})\exp(-2N\epsilon^2).$$

Statistical inference in Data Science is invariably based on asymptotic results ('as $n\to\infty$') invoking IID, such as the **Uniform** Law of Large Number (ULLN) for the whole family of functions $\mathcal{G}$, as well as more general asymptotic results derived by invoking **non-testable** mathematical and probabilistic assumptions ('as $n\to\infty$'), such as mixing conditions (Dependence) and asymptotic homogeneity (Heterogeneity)!

Main features of the Data Science approach:

[a] **Inductive premises**: normed linear spaces $L_p\left(S,\Im,\mathbb{P}(.)\right)$ endowed with a probability measure relating to the stochastic process $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$. The probabilistic assumptions underlying $\{\mathbf{Z}_t,\ t\in\mathbb{N}\}$ are *often IID*, with an indirect distribution assumption relating to a family of mathematical functions $\mathcal{G}$ and chosen on mathematical approximation grounds.

[b] **Model choice**: the best fitted curve $\widehat{\mathbf{y}}_i=\mathbf{G}_m(\mathbf{x}_i;\widehat{\boldsymbol{\phi}}_n(\mathbf{x}_i))$ in $\mathcal{G}$ is chosen on **goodness-of-fit/prediction** grounds or/and Akaike-type information criteria.

[c] **Inductive inference**: the interpretation of probability can be both frequentist and Bayesian, but the underlying reasoning is **decision theoretic** $(\forall \boldsymbol{\psi}_n \in \boldsymbol{\Psi})$ which is at odds with frequentist inference. The optimality of inference procedures is based on the the risk function and framed in terms of asymptotic theorems that invoke non-testable mathematical and probabilistic assumptions. The 'best' **fitted-curve** $\widehat{\mathbf{y}}_i = \mathbf{G}_m(\mathbf{x}_i; \widehat{\boldsymbol{\phi}}_n(\mathbf{x}_i))$ in $\mathcal{G}$ is used for '**predictive learning**'.

[d] **Substantive vs. statistical**: the fitted curve $\widehat{\mathbf{y}}_t = \mathbf{G}_m(\mathbf{x}_t; \widehat{\boldsymbol{\phi}}_n)$ in $\mathcal{G}$ is rendered a **black box** free of any statistical/substantive interpretation since the curve-fitting and regularization imposes **arbitrary restrictions** on $\boldsymbol{\psi}_n$ to **fine-tune** the prediction error. This **obviates** any possibility for **interpreting** $\widehat{\boldsymbol{\phi}}_n(\mathbf{x}_i)$ or establishing any evidence for potential causal claims, etc.

Weaknesses of Data Science (ML, SLT, etc.) algorithmic methods

1. The Curve-Fitting Curse: viewing the **modeling facet** with data as an **optimization problem** in the context of a **Hilbert space** of overparametrized functions will **always guarantee a unique solution** on goodness-of-fit/prediction grounds, **trustworthiness be damned**.

▶ Minimizing $\sum_{i=n+1}^{N} (y_i - \widehat{g}(\mathbf{x}_i; \boldsymbol{\psi}_n))^2$, using additional data in the **testing and**

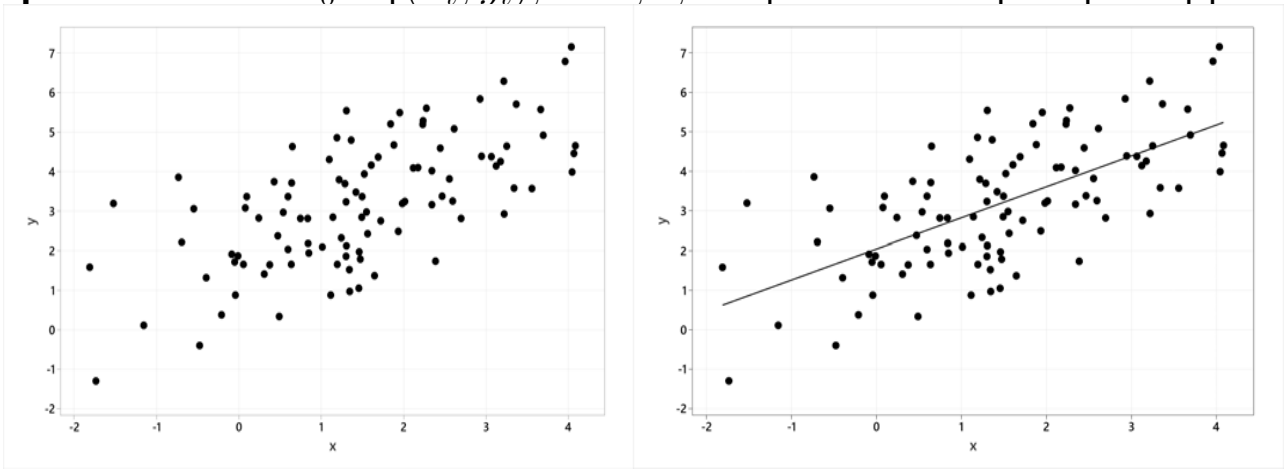
validation facets $\mathbf{z}_i{:=}(\mathbf{x}_i,y_i)$ $i{=}n{+}1,n{+}2,....,N$, has no added value in **learning from data** when training-based choice $\widehat{y}_t{=}\widehat{g}(\mathbf{x}_t;\boldsymbol{\psi}_n)$ is **statistically misspecified**. It just adds more scope for **tweaking!**

2. Is $L(\boldsymbol{\theta})= -\ln L(\boldsymbol{\theta};\mathbf{Z}_0)$, $\boldsymbol{\theta}{\in}\Theta$, just another loss function (Cherkassky and Mulier, 2007, p. 31)? No! $\ln L(\boldsymbol{\theta};\mathbf{Z}_0)$ is based on **testable probabilistic assumptions** comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$, as opposed to **arbitrary loss functions** based on information other than data $\mathbf{Z}_0$ (Spanos, 2017).

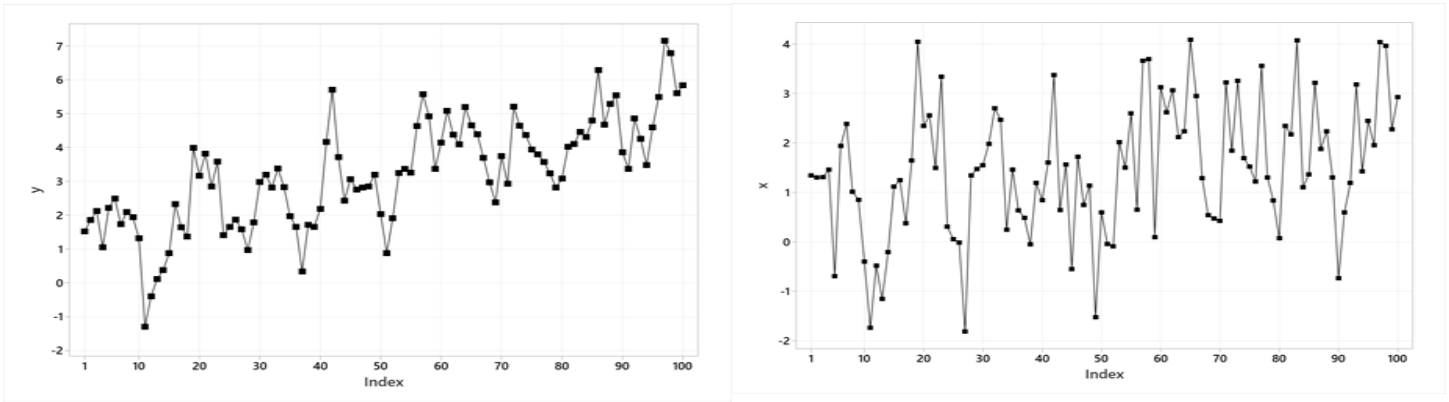
3. **Mathematical approximation error terms** are very different from **white-noise** statistical error terms**.** The former rely on Jackson-type upper bounds that are never statistically non-systematic. Hence, one can conflate the two errors at the **peril** of the **trusthworthiness of evidence**; Spanos (2010).

4. **What patterns**? Curve-fitting using **mathematical approximation patterns** is very different from **recurring chance regularity patterns** in data $\mathbf{Z}_0$, that relate directly to probabilistic assumptions. Indeed, the former seek approximation patterns which often invoke the validity of certain probabilistic assumptions**.** For instance, supervised and unsupervised learning using scatterplots invokes IID for $\mathbf{Z}_0$; Wilmot (2019), p. 66.

Example. For data $\mathbf{Z}_0$:=$\{(x_t, y_t),\ t$=$1, 2, ...n\}$ the scatterplot presupposes IID!



Unfortunately, the IID assumptions are **false** for both data series, shown below, that exhibit trending means (non-ID) and irregular cycles (non-I).



4. **How reliable are Data Science inferences**? The training/testing/validation split of the data can improve the selected models on **prediction grounds**, but will nothing **not secure the reliability of inference**.

5. It contrast to **PAC learnability** that takes the fitted $\widehat{\mathbf{y}}_i=\mathbf{G}_m(\mathbf{x}_i;\widehat{\boldsymbol{\phi}}_n(\mathbf{x}_i))\in\mathcal{G}$ at face value to learn about $f^*(\mathbf{z})$, the learning in **Fisher's model-based statistics** stems from $f(\mathbf{z};\boldsymbol{\theta})$, $\mathbf{z}\in\mathbb{R}^n_Z$, to $g(\mathbf{x}_t;\boldsymbol{\psi}(\boldsymbol{\theta}))=E(h(Y_t)|\mathbf{X}_t)$, where the probabilistic structure of $f(\mathbf{z};\boldsymbol{\theta})$, determines $Y_i=g(\mathbf{x}_i;\boldsymbol{\psi}(\boldsymbol{\theta}))$ as well as $\boldsymbol{\psi}(\boldsymbol{\theta})$.

6. The impression in Data Science is that the combination of: (i) a very **large sample size** n for data $\mathbf{Z}_0$, (ii) the training/testing/validation split, (iii) the **asymptotic inference**, renders the **statistical adequacy** problem irrelevant, is an illusion! Departures from **IID** will render both the **reliability and precision worse & worse** as n increases (Spanos & McGuirk, 2001). Moreover, invoking limit theorems 'as $n\to\infty$' based on **non-testable** Dependence and Heterogeneity is another head game.

On a positive note, **ML can be useful** when: (i) the data $\mathbf{Z}_0$ is (luckily) IID, (ii) $\mathbf{Z}_t$ includes a large number of variables, (iii) one has meager substantive information, and (iv) the sole objective is a short horizon 'makeshift' prediction.

7 Summary and conclusions

7.1 'Learning from data' about phenomena of interest

Breiman's (2001) claim that in Fisher's paradigm "One assumes that the data are generated by a given stochastic data model" refers to a common **erroneous implementation** of **model-based statistics** where $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ is viewed as *a priori* postulated model – presumed to be valid no matter what; see Spanos (1986).

In fact, Fisher (1922), p. 314, emphasized the crucial importance of model validation: "For empirical as the specification of the hypothetical population $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ represents the whole of the available facts." i.e. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ accounts for all the chance regularities in $\mathbf{Z}_0$.

Fisher's parametric model-based $[\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})]$ statistics, relying on **strong** (not weak) probabilistic assumptions that are **validated** vis-a-vis data $\mathbf{Z}_0$, provide the best way **to learn from data** using 'statistical approximations' around $\boldsymbol{\theta}^*$, framed in terms of sampling distributions of 'statistics' because they secure the **effectiveness** (reliability and precision) of inference and the **trustworthiness** of the ensuing evidence.

The Data Science algorithmic and the Graphical Causal (GC) modeling approaches share an inbuilt proclivity to side-step the **statistical misspecification** problem. The obvious way to improve the trustworthiness of their evidence is to **integrate** them within a broad **Fisher model-based statistical framework**.
In turn, sophisticated algorithms can enhance the model-based approach in several ways, including more thorough M-S testing.
That, of course, would take a generation to be implemented mainly due to the pronounced differences in culture and terminology!

In the meantime the trustworthiness of evidence in Data Science, **can be ameliorated** using simple M-S testing to evaluate the non-systematicity of the **residuals** from the fitted curve $\widehat{y}_t=g_m(\mathbf{x}_t;\widehat{\boldsymbol{\psi}}_n)$, $\widehat{\epsilon}_t=y_t-\widehat{y}_t,\ t=1,2,...,n.$

It is important to emphasize that **statistical 'excellent' prediction** is NOT just small prediction errors relative to a loss function, but non-systematic and 'small' prediction errors relative to likelihood-based goodness-of-prediction measures; see Spanos (2007).

"All models are wrong, but some are useful!" NO statistically misspecified model is useful for '**learning from data**' about phenomena of interest!

7.2 Potential casualties of the STATISTICS WARS

(1) Frequentist inference, in general, and **hypothesis testing,** in particular, as well as the frequentist underlying reasoning: **factual** and **hypothetical**.

(2) Error probabilities and their key role in securing the **trustworthiness of evidence** by controlling & evaluating **how severely tested** claims **are**, including:

(a) **Statistical adequacy**: does $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{z})$ **account** for the chance regularities in data $\mathbf{Z}_0$?

(b) **Substantive adequacy:** does the model $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{z})$ shed **adequate light** on (describes, explains, predicts) the phenomenon of interest?

(3) Mis-Specification (M-S) testing and **respecification** to account for the chance regularity patterns exhibited by data $\mathbf{Z}_0$, and ensure that the **substantive information** does not **belie** the data.

(4) Learning from data about phenomena of interest. Minimizing a risk function to reduce the overall Mean Square Prediction Error (MSPE) '$\forall \boldsymbol{\psi}_n \in \boldsymbol{\Psi}$' undermines learning from $\mathbf{Z}_0$ about $\boldsymbol{\psi}^*$; Spanos (2017).

Thanks for listening!