Aris Spanos [Dept. of Economics, Virginia Tech]

Outline:
1. Statistical Inference and Reliability
   how statistical misspecification leads to untrustworthy evidence
2. Theory testing in Economics: revisiting the CAPM
   Endemic misspecification bringing about a dead end
3. Theory appraisal in the context of Error Statistics
   Disentangling the statistical and substantive premises
4. Questions of ontology: statistical vs. substantive models
   Empirical examples: discovery of Argon, Kepler’s first law
5. Summary and conclusions

1Presentation at the Ontology and Methodology conference, Virginia Tech, May 4-5, 2013.
1 Statistical Inference and Reliability

1.1 The Frequentist recasting of statistical induction: model-based

R. A. Fisher’s (1922) most remarkable achievement was to recast statistical induction by replacing generalizing observed ‘events’ relating to data $z_0$, to modeling the underlying ‘process’ that gave rise to $z_0$, via the notion of a statistical model:

$$M_\theta(z) = \{ f(z; \theta), \theta \in \Theta \}, \quad z \in \mathbb{R}^n_Z, \quad \Theta \subset \mathbb{R}^m, \quad m < n,$$

(1.1.1)

$\Theta$-parameter space, $\mathbb{R}^n_Z$-sample space, $f(z; \theta)$-joint distribution of the sample $Z$. Fisher’s devised a general way to ‘operationalize’ inferential errors by:

(a) embedding the material experiment into a statistical model $M_\theta(z)$,
(b) calibrating the capacity of inference procedures in terms of the relevant error probabilities in the context of $M_\theta(z)$. The recasting also rendered the inductive premises testable vis-a-vis $z_0$. Neyman and Pearson (1933) supplemented Fisher’s ‘optimal’ estimation theory by proposing a theory of ‘optimal’ testing.

The technical apparatus of frequentist statistical inference was largely in place by the late 1930s, but the nature of the underlying inductive reasoning was clouded in disagreements: ‘inductive inference’ vs. ‘inductive behavior’, etc. Moreover, neither reasoning could adequate answer the question (Mayo, 1996): ‘when do data $z_0$ provide evidence for or against a hypothesis or a claim?’
1.2 Statistical misspecification: how serious is the problem?
What is often unappreciated is that behind every substantive model $M_\varphi(z)$ there is (often implicit) a statistical model $M_\theta(z)$ whose validity vis-a-vis data $z_0$ underwrites the reliability of any statistical inferences based on $M_\varphi(z)$.

\[ \text{\textbullet\ All statistical approaches are vulnerable since they invoke statistical models.} \]

Invalid $M_\theta(z) \Rightarrow$ wrong likelihood $L(\theta; z_0) \propto f(z_0; \theta) \Rightarrow \text{Unreliable Inferences}$

**Bayesian inference:** erroneous posterior $\pi(\theta|z_0) = \pi(\theta)L(\theta; z_0), \ \theta \in \Theta$.
Kadane (2011): “... likelihoods are just as subjective as priors.” Nonsense! The assumptions of $L(\theta; z_0) [M_\theta(z)]$ are testable with data $z_0$, those of $\pi(\theta)$ aren’t.

**Nonparametric:** invalid heterogeneity/dependence/distributional assumptions.
**Frequentist:** erroneous error probabilities and fit/prediction measures.
A frequentist inference (estimation, testing, prediction) is rendered unreliable when the actual error probabilities differ from the nominal (assumed) ones!

\[ \text{\textbullet\ Applying a .05 $\alpha$-level test when the actual type I error probability is closer to .90, is very likely to lead to erroneous inferences. But lame excuses abound.} \]

All models are wrong, but some are useful! Useful for what?
Example. Consider a simulation experiment using the Linear Regression model.

Table 1: Adequate model

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean</th>
<th>Std</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>([\beta_0=1.5] \hat{\beta}_0)</td>
<td>1.502</td>
<td>.122</td>
<td>1.500</td>
<td>.087</td>
</tr>
<tr>
<td>([\beta_1=.5] \hat{\beta}_1)</td>
<td>0.499</td>
<td>.015</td>
<td>0.500</td>
<td>.008</td>
</tr>
<tr>
<td>([\sigma^2=.75] \hat{\sigma}^2)</td>
<td>0.751</td>
<td>.021</td>
<td>0.750</td>
<td>.010</td>
</tr>
<tr>
<td>([R^2=.25]\hat{R}^2)</td>
<td>0.253</td>
<td>.090</td>
<td>0.251</td>
<td>.065</td>
</tr>
</tbody>
</table>

Table 2: Misspecified model

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean</th>
<th>Std</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_t=1.5+0.13t+0.5x_t+u_t)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\hat{y}_t=\beta_0 + \beta_1x_t + u_t)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Replications</th>
<th>True/Estim: (y_t=1.5+0.5x_t+u_t)</th>
<th>Table 2: Misspecified model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N=10000)</td>
<td>(n=50) (n=100)</td>
<td>True: (y_t=1.5+0.13t+0.5x_t+u_t) Estim.: (\hat{y}_t=\beta_0 + \beta_1x_t + u_t)</td>
</tr>
</tbody>
</table>

Table 1: (i) estimates \(\hat{\theta}(z_0)\) are highly accurate and the empirical \(\alpha \approx\) nominal (.05), even for \(n=50\), and (ii) their accuracy improves as \(n\) increases to \(n=100\).

Table 2: (iii) point estimates \(\hat{\theta}(z_0)\) are highly inaccurate and the empirical \(\alpha \gg .05\), and (iv) the inaccuracies get worse as \(n\) increases!
The intuitive explanation of what goes wrong in table 4 is that one is using an inconsistent estimator of the means of \((x_t, y_t)\), which yields inconsistent (spurious) estimators of the variances, covariances and correlation coefficient.

**Inconsistent (statistically spurious)**

\[
\frac{1}{n} \sum_{t=1}^{n} (y_t - \bar{y})^2 = 34.21, \quad [\bar{y} = 12.1]
\]
\[
\frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2 = 18.04, \quad [\bar{x} = 8.07]
\]
\[
\frac{1}{n} \sum_{t=1}^{n} (y_t - \bar{y})(x_t - \bar{x}) = 24.29
\]
\[
\sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \bar{y})^2 \frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2} = .978
\]

**Consistent (statistically meaningful)**

\[
\frac{1}{n} \sum_{t=1}^{n} (y_t - 2 - .2t)^2 = 1
\]
\[
\frac{1}{n} \sum_{t=1}^{n} (x_t - 1 - .14t)^2 = 1,
\]
\[
\frac{1}{n} \sum_{t=1}^{n} (y_t-2-.2t)(x_t-1-.14t) = .5
\]
\[
\sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t-2-.2t)^2 \frac{1}{n} \sum_{t=1}^{n} (x_t-1-.14t)^2} = .5
\]

**Crucial distinction:** Statistical vs. substantive meaningfulness
These spurious (artificially inflated) estimators will give rise to spurious test statistics and goodness-of-fit measures, rendering them highly unreliable:

\[
\tau(\beta_1) = \sqrt{\frac{\sum_{t=1}^{n}(x_t - \bar{x})(\hat{\beta}_1)}{s^2}}, \quad R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{\sum_{t=1}^{n} \hat{u}_t^2}, \quad s^2 = \frac{1}{n-2} \sum_{t=1}^{n} \hat{u}_t^2.
\]

∴ A statistically misspecified model (table 2) is useless for inference purposes.

2 Theory testing in Economics: a bird’s eye view

2.1 A key source of unreliability: foisting the theory on the data

Theory has generally held the pre-eminent role in economics, with data being given the subordinate role of: ‘quantifying theories’ presumed true. Cairnes (1888) articulated an extreme version of the Pre-Eminence of Theory (PET) perspective arguing that data is irrelevant for appraising the ‘truth’ of economic theories because they are infallible deductions from self-evident truths – established introspectively. In contrast to physics whose theories stem from mere inductive generalizations based on experimentation and inductive inferences, which are known to be fallible. The current PET perspective is almost as extreme: “The model economy which better fits the data is not the one used. Rather currently established theory dictates which one is used.” (Kydland and Prescott, 1991)

From the PET perspective data does not so much test as allow instantiation of
Theories: econometric methods offer elaborate (but often misleading) ways ‘to bring data into line’ with an assumed theory. Since the theory has little or no chance to be falsified, such instantiations provide no genuine tests of the theory.

### Traditional Econometric Modeling Procedure

<table>
<thead>
<tr>
<th>Theory model</th>
<th>Statistical model</th>
<th>Data</th>
<th>Statistical Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_t^e = \alpha + \beta r_{Mt}^e )</td>
<td>( r_t^e = \alpha + \beta r_{Mt}^e + \varepsilon_t )</td>
<td>( (r_t, r_{ft}, r_{Mt}), t=1, \ldots, n )</td>
<td>estimation ((\alpha, \beta, \sigma^2))</td>
</tr>
<tr>
<td>( \alpha = 0, \beta \neq 0 )</td>
<td>[i] ( \varepsilon_t \sim \text{NID}(0, \sigma^2) ),</td>
<td>( \downarrow \rightarrow )</td>
<td>testing ((\alpha, \beta, \sigma^2))</td>
</tr>
<tr>
<td></td>
<td>[ii] ( E(r_{Mt}^e \varepsilon_t) = 0, t \in \mathbb{N} ).</td>
<td></td>
<td>prediction: ( r_{n+p}^e, p \geq 1 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>simulation</td>
</tr>
</tbody>
</table>

The Capital Asset Pricing Model (CAPM), relates excess returns of a single asset \( r_t^e = (r_t - r_{ft}) \) and market excess returns \( r_{Mt}^e = (r_{Mt} - r_{ft}) \), where \( r_t \) – returns on the particular asset (CITI shares), \( r_{Mt} \) – market returns (S&P500), \( r_{ft} \) – returns of risk free asset (treasury bills); the CAPM assumes \( m > 1 \) assets.

**Data:** monthly Aug. 2000 to October 2005 \((n=64)\), on log returns \( r_t := \ln(P_t/P_{t-1}) \):
$r_t^e = 0.0053 + 1.137 r_{Mt}^e + \hat{\varepsilon}_{3t}, \quad R^2 = .725, \quad s = .0188, \quad n = 64. \quad (2.1.2)$

Lai & Xing claim that (2.1.2) provides strong evidence for the CAPM since:
(a) the signs and magnitudes of the estimated $(\alpha, \beta)$ are as predicted:
(i) the CAPM restriction $\alpha = 0$ is accepted by the data: $\tau(z_0; \alpha) = \frac{0.0053}{0.0033} = 1.606[.110]$,  
(ii) the beta coefficient $\beta$ is statistically significant: $\tau(z_0; \beta) = \frac{1.137}{0.089} = 12.775[.000]$,  
(b) the goodness-of-fit $(R^2 = .725)$ provides additional support to the CAPM.

What potential errors could render unreliable an inferences based on (2.1.2)?

[a] **Substantively inadequate model**: improper ceteris paribus clauses, missing confounding factors, false causal claims, etc.

[b] **Inaccurate data**: systematic errors imbued by the compilation process.

[c] **Incongruous measurement**: data $Z_0$ do not adequately quantify the concepts envisioned by the theory, e.g. intentions vs. realizations.

[d] **Statistically misspecified inductive premises**: error term assumptions 

\[ i \] impose (indirectly) invalid probabilistic assumptions on data $Z_0$:

\[ \varepsilon_t \sim \text{NIID}(0, \sigma^2), \quad t \in \mathbb{N}, \]
\[ \quad E(r_{Mt}^e \varepsilon_t) = 0, \quad t \in \mathbb{N}. \]

\[ \Leftrightarrow \{ Z_t := (r_t, r_{ft}, r_{Mt}) \sim \text{NIID}(\mu, \Sigma), \quad t \in \mathbb{N} \} \]

The discussion that follows focuses primarily on [d]: statistical misspecification.
Step 1: Complete specification of inductive premises: $\mathcal{M}_\theta(z)$.

| $\{(\varepsilon_t|X_t=x_t), \ t \in \mathbb{N}\}$ | $\{(y_t|X_t=x_t), \ t \in \mathbb{N}\}$ |
|------------------------------------------------|----------------------------------|
| $\varepsilon_t|X_t=x_t \sim \mathcal{N}(.,.)$ | $[1] \ (Y_t|X_t=x_t) \sim \mathcal{N}(.,.)$ |
| $E(\varepsilon_t|X_t=x_t) = 0$ | $[2] \ E(\varepsilon^2_t|X_t=x_t) = \sigma^2$ |
| $E(\varepsilon_t|X_t=x_t) = \sigma^2$ | $[3] \ Var(Y_t|X_t=x_t) = \sigma^2$ |
| $\{(\varepsilon_t|X_t=x_t), \ t \in \mathbb{N}\}$ | $[4] \ \{(Y_t|X_t=x_t), \ t \in \mathbb{N}\}$ indep. process |
| $\? \quad [5] (\beta_0, \beta_1, \sigma^2)$ do not change with $t$ |

Recasting [i]-[ii] into a complete set of probabilistic assumptions pertaining to the observable process $\{(y_t|X_t=x_t), \ t \in \mathbb{N}\}$ yields [1]-[5] (table 3).

**Table 3: Normal, Linear Regression Model**

<table>
<thead>
<tr>
<th>Statistical GM: $Y_t = \beta_0 + \beta_1 x_t + u_t, \ t \in \mathbb{N}$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] Normality: $(Y_t</td>
</tr>
<tr>
<td>[2] Linearity: $E(Y_t</td>
</tr>
<tr>
<td>[3] Homoskedasticity: $Var(Y_t</td>
</tr>
<tr>
<td>[4] Independence: ${(Y_t</td>
</tr>
<tr>
<td>[5] t-invariance: $(\beta_0, \beta_1, \sigma^2)$ do not change with $t$,</td>
</tr>
</tbody>
</table>

$$
\beta_0 = E(Y_t) - \beta_1 E(X_t), \quad \beta_1 = \frac{Cov(Y_t,X_t)}{Var(X_t)}, \quad \sigma^2 = Var(Y_t) - \frac{[Cov(Y_t,X_t)]^2}{Var(X_t)}
$$

$\{\mathcal{M}_\theta(z)\}$
Step 2: Thoroughly test the inductive premises: assumptions [1]-[5] (table 3). Mis-Specification (M-S) testing using pertinent auxiliary regressions:

\[ \hat{v}_t = \gamma_{10} + \gamma_{11} X_t + \gamma_{12} t + \gamma_{13} t^2 + \gamma_{14} X_t^2 + \gamma_{15} X_{t-1} + \gamma_{16} Y_{t-1} + \varepsilon_{1t}, \]

\[ H_0: \gamma_{1i} = 0, \ i = 2, \ldots, 5 \]

\[ \hat{v}_t^2 = \gamma_{20} + \gamma_{22} t + \gamma_{23} t^2 + \gamma_{21} X_t + \gamma_{24} X_t^2 + \gamma_{25} X_{t-1}^2 + \gamma_{26} Y_{t-1}^2 + \varepsilon_{2t}, \]

\[ H_0: \gamma_{2i} = 0, \ i = 1, \ldots, 6 \]

\[ \hat{v}_t = \frac{\sqrt{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t)}{s} \] are the studentized residuals. Applying these M-S tests to (2.1.2) indicate that [3]-[5] are clearly invalid; p-values in square brackets.

<table>
<thead>
<tr>
<th>Mis-Specification (M-S) tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normality:</strong> $D'AP = 1.169[.557]^2$</td>
</tr>
<tr>
<td><strong>Homosk/city:</strong> $F(2, 59) = 4.950[.010]^*$</td>
</tr>
<tr>
<td><strong>t-invariance:</strong> $F_\beta(2, 60) = 4.611[.014]^*$</td>
</tr>
</tbody>
</table>

These M-S test results undermine the above authors’ inferences (a)-(b)!

No evidence for or against a theory can be drawn on the basis of a statistically misspecified model! Why? The actual type I error probabilities are likely to
be very close to 1.0, invaliding inferences (a)-(b)! Much worse than the simulation example in table 4; M-S results can be guessed from the t-plots of the data!

Fig. 1: t-plot of CITI excess returns

Fig. 2: t-plot of market excess returns
Step 3: When $M_\theta(z)$ is misspecified, pose no substantive questions. Why? It elicits dubious answers. Consider posing the classic *confounding variable* question: is last period’s excess returns of General Motors, $r_{1t-1}^e$ an omitted but relevant variable? Re-estimating the model with $r_{1t-1}^e$ added yields:

$$
r_t^e = 0.0055 + 1.131 r_{Mt}^e - 0.144 r_{1t-1}^e + \hat{\nu}_{1t}, \quad R^2 = 0.747, \quad s = 0.0182,
$$

Taking the t-test ($\tau(z_0) = \frac{144}{0.56} = 2.571$) at face value, the answer is YES! A dubious answer. Any variable that proxies the trends, shifts or cycles in fig. 1-2, is likely to appear statistically significant, including generic trends and lags:

$$
r_t^e = 0.296 + 1.134 r_{Mt}^e - 0.134 t + 0.168 t^2 + \hat{\nu}_{2t}, \quad R^2 = 0.745, \quad s = 0.0184,
$$

$$
r_t^e = 0.0034 + 1.24 r_{Mt}^e - 0.175 r_{t-1}^e + \hat{\nu}_{3t}, \quad R^2 = 0.75, \quad s = 0.0181.
$$

**Important:** *Statistical misspecification* also undermines *goodness-of-fit/prediction*. Intuitively, the latter calls for *small* but the statistical adequacy calls for *nonsystematic (white-noise) residuals*. It also undermines all *goodness-of-fit/prediction* driven procedures, including the Akaike-type model selection:

$$
\text{min.} \quad \text{AIC} = -2 \ln L(\theta; z_0) + 2K, \quad \text{when} \quad L(\theta; z_0) \text{ is erroneous?}
$$
What is one supposed to do next?

**Strategy (?) 1:** Respecify the CAPM \( \mathcal{M}_\varphi(z) \) to improve it. **Problem:** how does one assess *improvement*? Goodness-of-fit/prediction measures are unreliable.

**Strategy 2:** Respecify the statistical model (2.1.2) to secure statistical adequacy and then test the CAPM and other modifications/extensions. **Problem:** How does one respecify (2.1.2) without changing the underlying theory?

Unreliable ‘error-fixing’ strategies for ‘upholding’ a theory. The PET perspective encourages modelers to uphold the original model \( \mathcal{M}_\varphi(z) \) and apply ad hoc modifications to [i]-[ii] in an attempt to account for an apparent ‘anomaly’.

**Example 1.** When the *independence* ([4]) assumption is rejected, one is supposed to ‘fix’ the problem by replacing \( H_0: E(\varepsilon_t \varepsilon_s) = 0 \) for \( t \neq s \) with \( \varepsilon_t = \rho \varepsilon_{t-1} + v_t \), i.e. adopting \( H_1 \) of an M-S test – the *fallacy of rejection*; Mayo & Spanos (2004).

**Example 2.** When the *homoskedasticity* ([3]) assumption is invalid a twofold recommendation is often prescribed. One ‘fixes’ the problem by: (i) an ad hoc modeling of heteroskedasticity using weighted least squares with \( \tilde{u}_t^2 = c_0 + c_1^\top z_t + v_t \), or (ii) retaining the original OLS estimator \( \hat{\beta} \) but replacing its *Standard Errors* (SEs) with *Heteroskedasticity-Consistent SEs* (HCSE).

*None of these ‘fixes’ brings the actual error probabilities closer to the nominal ones!*
The problem with the traditional approach. Foisting a theory on the data by attaching stochastic error terms to the theory model, gives rise to structural model which is an amalgam of substantive and statistical assumptions. Hence, the estimated $M_{\varphi}(z)$ is often both statistically and substantively misspecified, but one has no principled way to distinguish between:

- the theory is false or the inductive premises are invalid.

The key to circumventing this *Duhemian ambiguity* is to find a way to disentangle the statistical from the substantive premises, without compromising the credibility of either source of information.

### 3.1 Disentangling the substantive and statistical premises

What is often not adequately appreciated in empirical research is that:

- when confronting theory with data, **behind every substantive model** $M_{\varphi}(z)$, there is a statistical model $M_{\theta}(z)$; the two models are ontologically distinct!

- $M_{\varphi}(z)$ comes from the theory in light of data $Z_0$, but **where does the statistical model $M_{\theta}(z)$ come from**, if not from $M_{\varphi}(z)$? From the probabilistic structure of the stochastic process $\{Z_t, \; t \in \mathbb{N}\}$ underlying data $Z_0$. 

14
An alternative perspective on the relationship between $\mathcal{M}_\theta(z)$ and $\mathcal{M}_\varphi(z)$:

- The construction of the statistical model (premises) $\mathcal{M}_\theta(z)$ begins with a given data $Z_0$, irrespective of the theory or theories that led to the choice of $Z_0$. Once selected, data $Z_0$ take on ‘a life of its own’ as a particular realization of a generic stochastic process $\{Z_t, \ t \in \mathbb{N}\}$. The link between data $Z_0$ and the process $\{Z_t, \ t \in \mathbb{N}\}$ is provided by a pertinent answer to the key question:

  what probabilistic structure, when imposed on the process $\{Z_t, \ t \in \mathbb{N}\}$ would render data $Z_0$ a typical realization thereof?’

What do we mean by a typical realization of a generic process $\{Z_t, \ t \in \mathbb{N}\}$?

Fig. 3 exhibits dependence (cycles), and fig. 4 also exhibits mean heterogeneity.
In practice $\mathcal{M}_\theta(z)$ is chosen with two objectives in mind.

**A. Typicality.** The *truly typical realization* answer provides the apropos probabilistic structure for $\{Z_t, \ t \in \mathbb{N}\}$; testable using thorough $M$-$S$ testing.

The typicality generalizes Fisher’s (1922) original specification question: “Of what population is this a random sample?” to accommodate *non-random samples*.

**B. Relevant parameterization.** $\mathcal{M}_\theta(z)$ is specified by choosing a parameterization $\theta \in \Theta$ for $\{Z_t, \ t \in \mathbb{N}\}$ to the structural model $\mathcal{M}_\psi(z)$ via $G(\theta, \psi)=0$.

This replaces the notion of a ‘population’ with that of a ‘generating mechanism’.

**Example 1.** For the data in fig. 4, an appropriate model is the Normal AR(1) with a trend $(\theta:=(\alpha_0, \alpha_1, \sigma^2) \in \mathbb{R} \times (-1, 1) \times \mathbb{R}_+)$:

$$\mathcal{M}_\theta(x): (Z_t|Z_{t-1}) \sim \mathcal{N}(\alpha_0+\delta_1 t+\delta_2 t^2+\alpha_1 Z_{t-1}, \sigma_0^2), \ t \in \mathbb{N}.$$  

The statistical model for the data in fig. 3: $\mathcal{M}_\theta(x)|_{\delta_1=\delta_2=0}$.

Disentangling $\mathcal{M}_\theta(z)$ from $\mathcal{M}_\psi(z)$ delineates two very different questions:

[a] **statistical adequacy:** does $\mathcal{M}_\theta(z)$ account for the chance regularities in $Z_0$?  
[b] **substantive adequacy:** does the model $\mathcal{M}_\psi(z)$ shed adequate light (describe, explain, predict) on the phenomenon of interest?

To establish [b] one needs to secure [a] and then probe for potential errors, like *ceteris paribus* clauses, confounding effects, false causal claims, etc.
Questions of ontology: statistical vs. substantive models

A. A statistical model $\mathcal{M}_\theta(z)$, when untangled from the substantive model $\mathcal{M}_\phi(z)$, is built exclusively on the statistical information contained in data $Z_0$, and acts as a mediator between $\mathcal{M}_\phi(z)$ and $Z_0$.

Ontological commitments in specifying $\mathcal{M}_\theta(z)$ concern the existence of:

- (a) a rich enough probabilistic structure to 'model' the chance regularities in $Z_0$.
- (b) $\exists \theta^* \in \Theta : \mathcal{M}_{\theta^*}(z) = \{f(z; \theta^*)\}, \ z \in \mathbb{R}^n$, could have generated data $Z_0$.

The 'trueness' of $\theta^*$ is limited to whether $\mathcal{M}_{\theta^*}(z)$ could have generated data $Z_0$.

In this sense, the ontological status of $\mathcal{M}_\theta(z)$ is inextricably bound up with a particular frequentist methodology of inference and modeling.

- M-S testing assesses whether the family $\mathcal{M}_\theta(z)$, $\theta \in \Theta$ could have generated $Z_0$, regardless of the particular value $\theta^*$.
- Statistical inference aims to narrow $\Theta$ down to $\theta^*$—whatever $\theta^*$ happens to be!

Inductive reasoning based on sampling distributions

Example. In the case of the simple Normal model:

$$\mathcal{M}_\theta(x) : X_k \sim \text{NIID}(\mu, \sigma^2), \ \sigma^2\text{-known}, \ k \in \mathbb{N},$$

Estimation and testing adopt different forms of inductive reasoning, but their
The primary aim is identical: learn from \( x_0 \) about \( \mu^* \) the ‘true’ (but unknown) value of \( \mu \in \mathbb{R} \).

A \((1-\alpha)\) Confidence Interval stems from the sampling distribution of a pivot:
\[
\sqrt{n} (\bar{x}_n - \mu) \overset{\mu=\mu^*}{\sim} \mathcal{N}(0, 1), \quad \text{where} \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]
evaluated under \( \mu=\mu^* \). The N-P test for \( H_0: \mu=\mu_0 \) vs. \( H_1: \mu \neq \mu_0 \), uses the sampling distribution of a test statistic \( \sqrt{n} (\bar{x}_n - \mu_0) \) evaluated under \( H_0 \) & \( H_1 \):

\[
\text{type I error prob.} \quad \sqrt{n} (\bar{x}_n - \mu_0) \overset{\mu=\mu_0}{\sim} \mathcal{N}(0, 1), \quad \text{type II error prob./power} \quad \sqrt{n} (\bar{x}_n - \mu_1) \overset{\mu=\mu_1}{\sim} \mathcal{N}(\frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}, 1), \quad \text{for} \quad \mu_1 \neq \mu_0.
\]

What is often overlooked is that \( \sqrt{n} (\bar{x}_n - \mu^* - \mu_0) \) is a scaled difference between \( \mu^* \) [that generated \( \bar{x}_n \)] and \( \mu_0 \).

- The pre-data error probabilities calibrate the generic capacity (power) of frequentist tests: how effective a test is in detecting different discrepancies \( \gamma \) from \( \mu_0 \). This capacity is often conflated with the empirical relative frequencies of errors associated with the long-run metaphor, which invokes the ‘repeatability in principle’ [not in time or space] of the GM defined by \( \mathcal{M}_\theta(x) \). Error probabilities are relevant for inferences with any \( x \in \mathbb{R}_X^n \), including data \( x_0 \).
Error Statistics (ER) can be viewed as a refinement/extension of the Fisher-Neyman-Pearson (F-N-P) motivated by the call to address several foundational problems. In particular, ER (a) refines the F-N-P approach by proposing a broader framework with a view to secure statistical adequacy, and reliability of inference, and (b) extends the F-N-P approach to provide an evidential interpretation for the accept/reject rules and p-values. This is achieved using a post-data severity assessment that takes into account the test’s generic capacity (power) to output the discrepancies \( \gamma \) from \( \mu = \mu_0 \) warranted or ruled out by data \( x_0 \); Mayo and Spanos (2006). Severity also bridges the inferential gap between statistical (\( \theta \)) and substantive (\( \varphi \)) parameters related via \( G(\theta, \varphi) = 0 \), e.g. statistical vs. substantive significance, etc.

B. A substantive model \( M_\varphi(z) \) is often viewed as aiming to approximate the actual Data Generating Mechanism by abstracting, simplifying and focusing on particular aspects (selecting the relevant observables \( Z_t \)) of the phenomenon of interest. In practice, however, the form of \( M_\varphi(z) \) can vary widely from a highly complex system of dynamic (integro-differential) equations to a simple question of interest. Consider two examples from the two ends of this spectrum.
Example 1. The discovery of Argon was partly based on probing the difference in weight between nitrogen produced by two different broad methods, referred to as Atmospheric (X) and Chemical (Y); see table 4. The data were generated over a period of 18 months using different experimental techniques and procedures to produce nitrogen and measure the weight of the end product.

<table>
<thead>
<tr>
<th>Table 4 - Rayleigh data (1894)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
</tr>
<tr>
<td>Y</td>
</tr>
</tbody>
</table>

Substantive model $M_\phi(z)$: is there a difference in weight between the nitrogen produced by the two procedures? The $M_\theta(z)$ is a simple Normal:

$$M_\theta(z) : X_k \sim \text{NIID}(\mu_1, \sigma^2), \ Y_k \sim \text{NIID}(\mu_2, \sigma^2), \ k \in \mathbb{N}. \quad (4.0.3)$$

The substantive question is framed in terms of the statistical hypotheses:

$$H_0 : (\mu_1 - \mu_2) = 0, \ \text{vs.} \ \ H_1 : (\mu_1 - \mu_2) > 0, \quad (4.0.4)$$

in the context of $M_\theta(z)$ in (4.0.3). The *material experiment* is embedded in $M_\theta(z)$, which is substantively fictitious. However, $M_\theta(z)$ is statistically adequate because the data $z_0$ in figure 5 could have been generated by (4.0.3).
Fig. 5: $t$-plots of $z_0 := (x_t, y_t), \ t=1, \ldots, n$.

The $t$-test $\tau(z_0) = 37.2678[.000]$ rejects $H_0$, and a severity evaluation:

$$\text{SEV}(T_\alpha; z_0; \mu > \gamma) = \mathbb{P}(\tau(Z) \leq \tau(z_0); \mu_1 = \gamma) = .9,$$

indicates that a substantive discrepancy $\gamma \geq .011$ is warranted by data $z_0$.

It is interesting to note that the **substantively true** value is $\gamma^* = .01186$.

**Lesson**: the statistical model $\mathcal{M}_\theta(z)$, acting as a *mediator* between the substantive question and $z_0$, enabled us to find out (learn from data) **true things** about the real difference in weight between atmospheric and chemical nitrogen, only to the extent that it does ‘account for the statistical regularities in $z_0$’.
Example 2. Kepler’s first law states that ‘a planet moves around the sun in an elliptical motion with one focus at the sun’. The loci of the elliptical motion in polar coordinates is: \[(1/r) = \alpha_0 + \alpha_1 \cos \vartheta,\]
where \(r\)-the distance of the planet from the sun, \(\vartheta\)-the angle between the line joining the sun and the planet and the principal axis of the ellipse.

For \(Y := (1/r), \ X := \cos \vartheta\) this yields the substantive model:
\[M_\varphi(z) : Y_k = \alpha_0 + \alpha_1 X_k + \epsilon_k, \ k=1, 2, \ldots, n, \ldots \quad (4.0.5)\]
The statistical model \(M_\theta(z)\) is the Linear Regression model (table 3), which when estimated Brahe’s original data (\(n=28\)) yields:
\[y_t = .662062 + .061333 x_t + \hat{u}_t, \ R^2 = .9999, \ s = .00001115. \quad (4.0.6)\]
(4.0.6) is statistically adequate, providing strong evidence for Kepler’s first law. In this case \(M_\varphi(z)\) and \(M_\theta(z)\) aim to approximate the actual DGM.
We learn from data about phenomena of interest by applying reliable [actual \( \simeq \) nominal error probabilities] and potent [optimal] inference procedures.

Foisting a theory on the data yields statistically and substantively inad-equate estimated models. A dead end! Addressing the Duhemian ambiguity:

**Step 1:** Separate, *ab initio*, the substantive \( \mathcal{M}_\varphi(z) \) from the statistical premises \( \mathcal{M}_\theta(z) \) by viewing the latter as a parameterization of the process \( \{Z_t, \ t \in \mathbb{N}\} \) underlying data \( Z_0 \), that nests \( \mathcal{M}_\varphi(z) \) via \( G(\theta, \varphi) = 0 \).

**Step 2:** Secure statistical adequacy using trenchant M-S testing.

**Step 3:** Respecify \( \mathcal{M}_\theta(z) \) with a view to find a statistically adequate model.

The ontological status of \( \mathcal{M}_\theta(z) \) is inextricably bound up with the error-statistical methodology of inference and modeling.

M-S testing assesses whether the family \( \mathcal{M}_\theta(z), \ \theta \in \Theta \) could have generated \( Z_0 \). Statistical inference aims to narrow \( \Theta \) down to \( \theta^* \) - whatever \( \theta^* \) happens to be!

**Step 4:** Test the validity of the substantive model in the context of a statistically adequate \( \mathcal{M}_\theta(z) \). When \( \mathcal{M}_\varphi(z) \) is found empirically invalid, one can use a statistically adequate \( \mathcal{M}_\theta(z) \) to guide the search for a better theory!
If $\mathcal{M}_\theta(z)$, then $Q(z; \theta)$, for $z \in \mathbb{R}^n$

**Statistical Model:**
$\mathcal{M}_\theta(z)$, $z \in \mathbb{R}^n$, $\Theta \subset \mathbb{R}^m$
(assumptions [1]-[5])

**Inferential Propositions:**
$Q(z; \theta)$ $G(\theta, \varphi) = 0$ $\Rightarrow$ $Q_1(z; \varphi)$

**Data:** $z_0 = (z_1, z_2, \ldots, z_n)$

**Structural Model** $\mathcal{M}_\varphi(z)$

**Inference results:**
$Q(z_0; \theta) \Rightarrow Q_1(z_0; \varphi)$

---

Model-based frequentist statistical induction

Key components of Error Statistics [A]-[C]