Standing on the Shoulders of Machine Learning:

Can We Improve Hypothesis Testing?

Gary Cornwall
Bureau of Economic Analysis

Jeff Chen
Bennett Institute for Public Policy
University of Cambridge

Beau Sauley
University of Cincinnati

April 28, 2021

- We can think of hypothesis testing as a classification problem: either you reject the null or you don't.
- Classification algorithms built upon entropy (Shannon, 1948) are foundational pieces of modern machine learning.
- Despite improvements in computational power and classification algorithms we have not revisited the basic framework of hypothesis testing in any meaningful fashion.

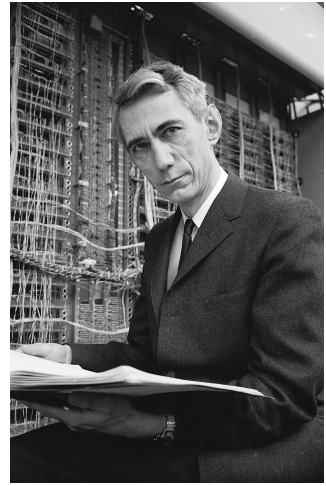


Photo of Claude Shannon / Alfred Eisenstaedt / The LIFE Picture Collection / Getty

Research Question(s)

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests?

Research Question(s)

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests?
2. If hypothesis tests and weak learners are equivalent, can we use modern machine learning algorithms to make testing more accurate?

Research Question(s)

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests?
2. If hypothesis tests and weak learners are equivalent, can we use modern machine learning algorithms to make testing more accurate?
3. **Can we use this approach to improve upon what has been a hard hypothesis testing problem in time series econometrics, that of the unit root?**

Quick Answers

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests? Yes, in fact these are equivalent in both single and two-tailed tests for some $\alpha = \alpha'$.

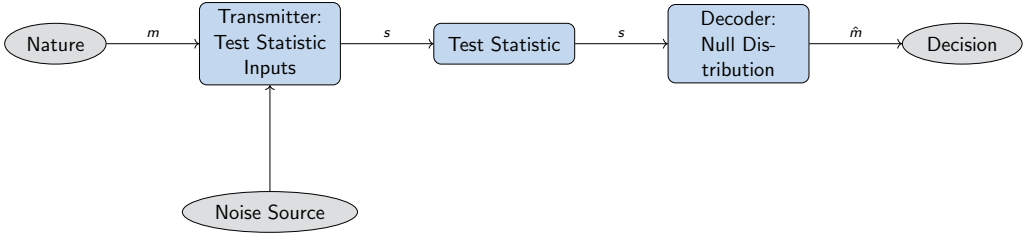
Quick Answers

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests? Yes, in fact these are equivalent in both single and two-tailed tests for some $\alpha = \alpha'$.
2. If hypothesis tests and weak learners are equivalent, can we use modern machine learning algorithms to make testing more accurate? Yes, since we know how to aggregate weak base learners and create more powerful ensemble prediction methods we can use tools such as random forests and gradient boosting to improve hypothesis test accuracy.

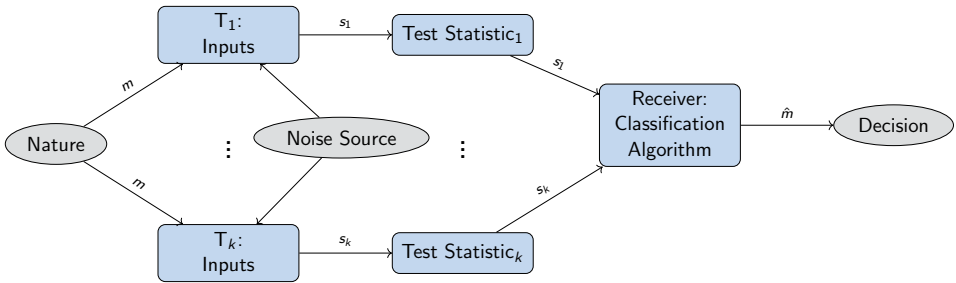
Quick Answers

1. If we think about hypothesis testing as a simple classification problem, can we draw an equivalence between "weak learners" and hypothesis tests? Yes, in fact these are equivalent in both single and two-tailed tests for some $\alpha = \alpha'$.
2. If hypothesis tests and weak learners are equivalent, can we use modern machine learning algorithms to make testing more accurate? Yes, since we know how to aggregate weak base learners and create more powerful ensemble prediction methods we can use tools such as random forests and gradient boosting to improve hypothesis test accuracy.
3. **Can we use this approach to improve upon what has been a hard hypothesis testing problem in time series econometrics, that of the unit root? Yes, we find that our approach can improve upon the traditional unit root test(s) by seventeen percentage points over the best single-test alternative.**

A Signal Representation of Hypothesis Tests



The Process: A 30,000 Foot View



Why Unit Roots?

- The unit root problem is a difficult time series econometrics problem which has produced nearly five decades of research and many different test statistics.
- The test for unit roots is important because failing to identify a unit root can invalidate all subsequent inferences (Granger & Newbold, 1974).
- Co-integrated relationships between series means you can't just assume everything has a unit root (Granger, 1981; Engle and Granger, 1987).
- The difficulty comes from differentiating unit roots from *near unit roots*, as a result these test statistics have low power (Ng & Perron, 2001)

Let y_t be an autoregressive time series generated such that,

$$y_t = \phi y_{t-1} + \epsilon_t, \quad t = (1, \dots T)$$

- We assume $\epsilon_t \sim N(0, \sigma^2)\ \forall\ t$ and that $\sigma_1^2 = \dots = \sigma_T^2$.
- We can write this as $(1 - \phi L)y_t = \epsilon_t$ such that $Ly_t = y_{t-1}$.
- $(1 - \phi L)$ has a root of $1/\phi$ and if $|\phi| < 1$ then y_t is considered stationary.
- Tests are often using an $H_0 : \phi = 1$ and $H_1 : |\phi| < 1$ structure (*e.g.* Augmented Dickey Fuller test).

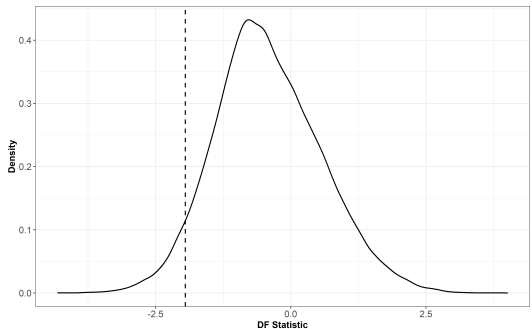
- The Dickey-Fuller test statistic:

$$\hat{\tau} = (\hat{\phi} - 1) S_e^{-1} (\sum_{t=2}^{N} Y_{t-1}^2)^{1/2},$$

$$S_e^{-1} = (n-2)^{-1} \sum_{t=2}^{N} (Y_t - \hat{\phi} Y_{t-1})^2,$$

with limiting distribution outlined in Dickey & Fuller (1979).

- Calculated on first difference of Y with $H_0 : \phi = 1$ and $H_1 : |\phi| < 1$.



- Different assumed DGPs result in different null distributions and decision thresholds:

$$y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 | \alpha = 0.05$$
$$y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 | \alpha = 0.05$$
$$y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 | \alpha = 0.05$$

How do we test for a Unit Root?

- Different assumed DGPs result in different null distributions and decision thresholds:

$$y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 | \alpha = 0.05$$
$$y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 | \alpha = 0.05$$
$$y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 | \alpha = 0.05$$

- Choice of DGP opens up an additional error path beyond Type I and Type II errors.

- Different assumed DGPs result in different null distributions and decision thresholds:

$$y_t = \lambda + \phi y_{t-1} + \delta t + \epsilon_t \rightarrow x_\alpha = -3.45 | \alpha = 0.05$$
$$y_t = \lambda + \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -2.89 | \alpha = 0.05$$
$$y_t = \phi y_{t-1} + \epsilon_t \rightarrow x_\alpha = -1.95 | \alpha = 0.05$$

- Choice of DGP opens up an additional error path beyond Type I and Type II errors.
- There are many tests that are similarly structured, *e.g.,* ADF (Dickey and Fuller, 1981), PP (Phillips and Perron, 1988), KPSS (Kwiatkowski et al., 1992), PGFF (Pantula et al., 1994), Breit (Breitung, 2002; Breitung and Taylor, 2003), ERS (Elliot et al., 1996), URSP (Schmidt and Phillips, 1992), and URZA (Zivot and Andrews, 2002).

A Simple Procedure for Composite Test Construction

1. Simulate a balanced training, validation, and test set containing representative cases of the null and alternative hypotheses
2. Derive transmitters from one or multiple test statistics and attributes of the time series
3. Train a set of supervised classifiers, then select the model that fairs the best in cross-validation; then
4. Given a cost ratio, $c(e_2)/c(e_1)$, that indicates the relative importance of Type I and Type II errors, calculate the optimal probability threshold from the validation set for classifying individual instances.

Simulate a balanced, representative data set

For any hypothesis test we can write down a DGP which will satisfy the null, *e.g.* unit roots.

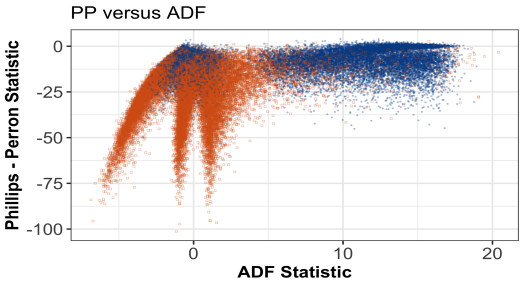
1. Generate $500,000$ time series with $350,000$ for training, $75,000$ for validation, and $75,000$ for testing.
2. A series will contain a unit root, that is $\phi = 1$ with probability 0.50 and $\phi \in \{0.9000, 0.9999\}$ otherwise.
3. Series will be uniformly distributed over the three unit root DGPs mentioned earlier.
4. All noise is Gaussian white noise.

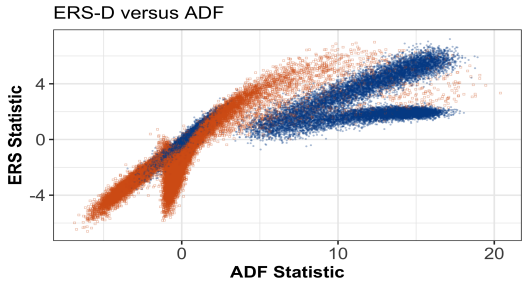
What are the transmitters?

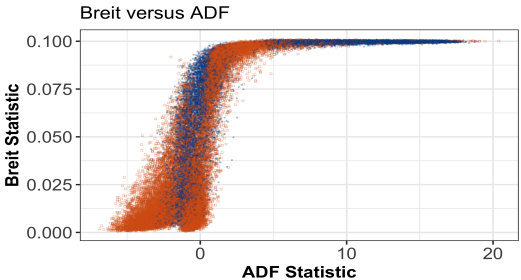
UR Tests	Level and First Difference	STL Decomposed Series	Miscellaneous
ADF	Skewness	TNN Test	Length
PP	Kurtosis	Skewness	Frequency
PGFF	Box Statistic	Kurtosis	$\text{var}(\Delta y)/\text{var}(y)$
KPSS	Lyapunov Exponent	Box Statistic	
ERS (d & p)	TNN Test		
URSP	Hurst Exponent		
URZA	Strength of Trend		
Breit	Strength of Seasonality		

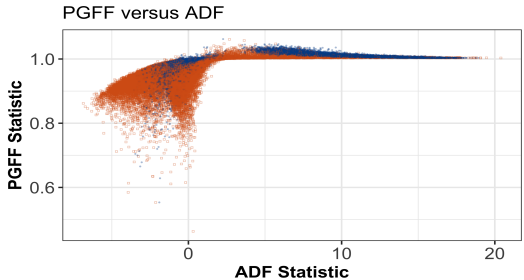
While we generate the data from one of the three possible "cases" outlined in the literature all test statistics are calculated on the most parsimonious DGP assumption possible, *e.g.* no drift or trend for the ADF.

Is there variation in our transmitters?

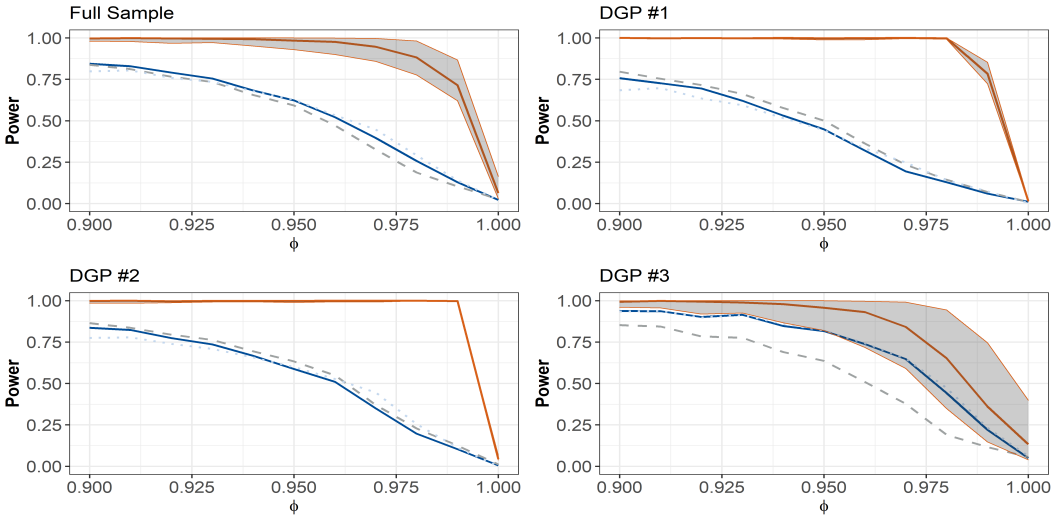








Power Curves



Conclusion

1. The framework for modern hypothesis testing comes from a time with limited computational power.
2. We have updated the framework to incorporate modern thoughts on signal processing and computational power.
3. In a single test statistic environment we have shown that this is equivalent to the current paradigm, but...
4. the proposed method allows for the exploitation of multiple transmitters (statistics) and other information to improve accuracy.
5. Practitioner can be explicit about cost of error types in our framework rather than fixing a specific error rate.
6. An R package for unit roots, "hypML", has been constructed and will be available for use.

Thank you!