Electronic Journal of Statistics
Vol. 10 (2016) 1526–1550
ISSN: 1935-7524
DOI: 10.1214/16-EJS1146

# Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox

**Michael Naaman**

*Christensen Associates*
*800 University Bay Drive, Suite 400*
*Madison, Wi 53705-2299*
*e-mail:* mtnaaman@lrca.com

**Abstract:** A new method of hypothesis testing is proposed ensuring that as the sample size grows, the probability of a type I error will become arbitrarily small by allowing the significance level to decrease with the number of observations in the study. Furthermore, the corresponding sequence of hypothesis tests will make only a finite number of errors with probability one, under mild regulatory conditions, including both i.i.d. and strongly mixing processes. It can be used as an alternative to arbitrary fixed significance levels such as 0.05 or 0.01.

This approach resolves the Jeffreys-Lindley paradox. It is also robust to multiple comparisons, and to a lesser extent, optional stopping making it somewhat robust to data fishing or p-hacking.

As an example of the practical applications of this technique, it is used as a lag order selection mechanism in simulations. It performs well relative to other model selection criteria. In another simulation, hypothesis tests and confidence intervals for the mean are investigated, demonstrating the improved performance even in small samples. We also show that under mild regularity conditions, any sequence of two-sided hypothesis tests with fixed significance level will make an infinite number of mistakes with positive probability.

**MSC 2010 subject classifications:** Primary 60K35, 60K35; secondary 60K35.
**Keywords and phrases:** Hypothesis testing, Edgeworth expansions, model selection, p-hacking.

Received August 2015.

## 1. Introduction

For the frequentist, the basic framework of hypothesis testing has changed little since the pioneering work of Pearson and Neyman nearly 80 years ago. One chooses both a null hypothesis and an alternative. If the probability of observing the test statistic under the null hypothesis is smaller than some significance level, which is usually 0.05 or 0.01, then the null is rejected in favor of the alternative.

Furthermore, one may invert the test statistic to create a confidence interval that will contain the parameter of interest with a large probability. Under suitable conditions, the sample mean, $\overline{x}$, can be used to construct a 95% confidence

interval for the mean, $\mu$, and as the sample size grows

$$\Pr\left(|\overline{x}-\mu|<\hat{\sigma}n^{-.5}1.96\right)\to 0.95 \tag{1.1}$$

where $\hat{\sigma}^2$ is a consistent estimator of the variance. In this case, the probability that the 95% confidence interval contains $\mu$ approaches 0.95. Many results in statistics focus on issues relating to the rejection of the null when it is false. However, Fisher, who introduced the term null hypothesis, did not even specify an alternative, instead focusing on a well-defined null, see [10]. In this paper, we will be primarily focused on hypothesis testing that performs well, regardless of the validity of the null.

In general, the sequence of confidence intervals under a fixed significance level, like in Eq. (1.1), will collapse around the true population parameter, but there will be an infinite number of confidence intervals that do not contain the population parameter. This concept is not entirely new. In [9], this issue is discussed, and conditions for the existence of a sequence of tests that address this issue is proved.

The choice of significance level is essentially arbitrary, but it does not have to be. Recall the consistency of the sample mean: for any $\epsilon>0$,

$$\Pr\left(|\overline{x}-\mu|\geq\epsilon\right)\to 0$$

and consider the interval given by $I_\epsilon=(\overline{x}-\epsilon,\overline{x}+\epsilon)$, which has the following property.

$$\Pr\left(\mu\in I_\epsilon\right)\to 1$$

In some sense, $I_\epsilon$ is a better predictor of the mean because the probability that the mean is in the interval converges to one. The downside is: the interval never gets any smaller as the sample size increases.

However, it is possible to have the best of both worlds. If one allows the significance level to decrease as the sample size gets larger by choosing a bandwidth, $h$, depending on the sample size and a function, $f$, decreasing in $h$ such that

$$\Pr\left(|\overline{x}-\mu|<\hat{\sigma}n^{-.5}f\left(h\right)\right)\approx 1-h$$

with $h\to 0$ and $n^{-.5}f\left(h\right)\to 0$, then the sequence collapses around the mean, and the intervals contain the mean with probability approaching unity. Even with this improvement, the sequence could still have an infinite number of confidence intervals that do not contain the mean. However, if the bandwidth falls off rapidly enough, we will show that there will be a finite number of errors made with probability one.

The problem with fixed significance level confidence intervals is that the sequence of intervals is collapsing too quickly, so it does not catch an infinite number of errors. By allowing the critical values to diverge slowly, one may catch almost all the errors.

In statistical inference, asymptotic theory plays a central role. It will often be said that some property holds as the sample size approaches infinity. But hypothesis testing is stuck in a static setting where the size of the test is fixed, and everything else is allowed to change as the sample size grows. Keeping the significance level fixed leads to what is known as the Jeffreys-Lindley paradox.

The paradox is that under certain conditions, for any fixed significance level, $\alpha$, one can find a sample size such that there is a statistically significant effect, but the posterior probability of the null hypothesis is approximately $1-\alpha$. This is in direct conflict with the frequentist approach. However, [21] recognized the source of the paradox was keeping the significance level fixed saying, "the theory does not justify the practice of keeping the significance level fixed" and even "some computations by Prof. Pearson in the discussion to that paper emphasized how the significance level would have to change with the sample size, if the losses and prior probabilities were kept fixed." While the paradox is of secondary importance, we will show that this approach allows the significance levels to change with the sample size in a way that resolves the paradox.

This paradox is not just of interest to theoretical statistics, but has tangible empirical implications. In particle physics, a result is not considered a discovery unless the estimate is 5 standard deviations away from the null hypothesis. This is an arbitrary choice, but researchers noticed that even 3 standard deviations may result in spurious results, see [4]. In fact, [23] states that the significance level should "decrease with increasing amount of data. However, there appears to be no obvious way of implementing this, and Particle Physics tends to use fixed levels of cuts, independent of the data size." The role of the paradox in physics is also discussed in [7]. This approach solves that problem by allowing the significance level to decrease as the number of observations grows.

We will also provide a methodology that allows for the construction of a sequence of confidence intervals such that only a finite number of the confidence intervals in the sequence fail to contain the population parameter with probability one. A similar result will hold for hypothesis testing, and the probability of a type I or type II error will converge to zero. This results in an arbitrarily small probability of making a mistake in inference as the sample size grows. In addition, any sequence of hypothesis tests with fixed significance level will make an infinite number of mistakes with positive probability.

For example, a 95% confidence interval has the interpretation that if the interval was recomputed on a large number of samples, then roughly 95% of the calculated confidence intervals would contain the true population parameter. Instead of thinking about what happens for a large number of repeated samples of a confidence interval with fixed sample size, consider a sequence of confidence intervals for a single data set as the sample size grows. In application, it is often the case that data comes as a stream in daily, monthly or yearly rates. More generally, as our sample size increases, there is nothing to stop us from performing statistical inference each time the data set grows. If we have yearly data, confidence intervals can be calculated every year, and there will be a sequence of confidence intervals corresponding to each year we recomputed a confidence interval.

In practice, the example above is not usually how inference is done. There is not a sequence of tests, but rather a single sample for which inference is made. This example is given solely to fix ideas.

However, even though we may only observe a fixed sample size for any given study, it is still possible to discuss notions of almost sure convergence of a confidence interval, or a hypothesis test, just as we do with the almost sure convergence of other statistics like the sample mean.

As another example of the power of this approach, if an academic journal only accepts papers with p-values less than 0.05, then roughly 1 in 20 independent studies of the same effect would find a significant result when there was none. However, if the journal required a minimum sample size of 100, and results would only be accepted using this paper's methodology, then one would expect roughly 1 in 250 studies would find an effect when there was none (if the minimum sample size was $1,000$, it would be 1 in $4,000$).[1]

It will also be shown that unlike fixed significance level testing, this methodology is robust to multiple comparisons. It is also somewhat robust to optional stopping, which occurs when one increases the sample size, and retests the hypothesis repeatedly, until a statistically significant result is found. This is important because, in empirical work, researchers sometimes unwittingly perform multiple comparisons or optional stopping without proper corrections. The approach is robust to a certain level of data fishing or p-hacking.

One can even use this approach for lag order selection in autoregressive models. In simulations, almost sure (A.S.) hypothesis testing picks the correct lag more often than other popular lag selection mechanisms. In other simulations, the effectiveness of A.S. hypothesis tests is compared with standard approaches in terms of size and power.

A.S. hypothesis testing is applicable not only to sample means, but any estimators that fall under the smooth function model including correlations, maximum likelihood, and regressions. It is a much more powerful tool for inference.

## 2. The normal case with known variance

### *2.1. One-sided hypothesis tests*

First, we consider the simplest of cases to make the methodology more clear. Suppose $x_i = \mu + \epsilon_i$ where $\{\epsilon_i\}_{i\in\mathbb{N}}$ is an i.i.d. sequence of standard normal random variates. It is of interest to test

$$\begin{aligned} H_0 &: \quad \mu \leq t \\ H_1 &: \quad \mu > t \end{aligned} \tag{2.1}$$

[1]This assumes the suggested bandwidth in Eq. (3.5) is used, and all of the studies are at the minimum sample size. If the bandwidth in Eq. (3.8) is used (which will have better small sample performance with regard to type I error when multiple comparisons are a concern) one would expect roughly 1 in $10,000$ studies would find an effect when there was none (if the minimum sample size was $1,000$, it would be 1 in $1,000,000$).

which is a one-sided test that will reject $H_0$ whenever

$$(\overline{x} - t)\sqrt{n} \ > \Phi^{-1}(1-\alpha)$$

for a size $\alpha$ test and $\Phi^{-1}$ is the inverse of the normal distribution function. This generates a sequence of hypothesis tests based on the sample size, which is denoted in the following way.

$$\begin{aligned} H_{0n}: &\quad \mu \leq t \\ H_{1n}: &\quad \mu > t \end{aligned} \tag{2.2}$$

The probability of a type I error, which is rejecting the null hypothesis when it is true, is given below.

$$\begin{aligned} &\Pr\left((\overline{x} - \mu)\sqrt{n} \ > \Phi^{-1}(1-\alpha) - (\mu - t)\sqrt{n}\right) \\ &\leq \Pr\left((\overline{x} - \mu)\sqrt{n} \ > \Phi^{-1}(1-\alpha)\right) = \alpha \end{aligned}$$

So the null hypothesis is rejected incorrectly up to $\alpha$ percent of the time. Instead of choosing a fixed type I error rate, $\alpha$, consider choosing a smoothing parameter, $h$, that is chosen to converge to zero.

Unfortunately, this means the critical values will diverge to infinity, so if a smoothing parameter is chosen to converge to zero, then the corresponding confidence interval might blow up. Fortunately, this is not necessarily the case.

**Lemma 2.1.** *Let $h(n)$ satisfy $0 \leq h(n) \leq 1$ for all $n$ and $z_h = \Phi(1-h)$. Suppose*

$$\lim_{n \to +\infty} \frac{\ln(h)}{n} = 0 \tag{2.3}$$

*then*

$$\lim_{n \to +\infty} \frac{z_h}{\sqrt{n}} = 0 \tag{2.4}$$

*Proof.*

$$\begin{aligned} \mathrm{z_h} &= \Phi^{-1}(1-h) = \sqrt{2}\mathrm{erf}^{-1}(2(1-h)-1) \\ &= \sqrt{2}\mathrm{erfc}^{-1}(2h) \ \leq \sqrt{-2\ln(2h)} \end{aligned}$$

for all $n \geq N$ such that $h(N) \leq \frac{1}{2}$. The inequality follows from Chernoff's bound, which is derived in [6].

$$\lim_{n \to \infty} \frac{\mathrm{z_h}}{\sqrt{n}} \ \leq \lim_{n \to \infty} \sqrt{-\frac{2\ln(2h)}{n}} = 0 \tag{2.5}$$

□

This implies that even though the critical values of the test are diverging, the sequence of confidence intervals will still collapse around $\mu$, albeit with a slower rate of convergence. A sufficient but not necessary condition for Eq. (2.3) to hold is $h \propto n^{-p}$ with $p > 0$.

Of course, any choice of $h$ that converges to zero results in the probability of a type I error vanishing asymptotically. But this has not completely solved the

problem, as we could still have an infinite number of type I errors with some positive probability.

Let's examine the sequence of hypothesis tests that could be performed as the sample size grows. Consider the sequence of not necessarily independent events defined by the realization of a type I error. We will use the indicator function

$$A_n = \mathbb{1}\left((\overline{x} - t)\sqrt{n} > \Phi^{-1}(1-h)\right) \tag{2.6}$$

where a type I error occurs when $A_n = 1$. If $h$ satisfies $\sum_{i=1}^{+\infty} h(i) < +\infty$, then the Borel-Cantelli lemma can be applied.

$$\Pr\left(\lim_{n\to+\infty} \sup A_n = 1\right) = 0 \tag{2.7}$$

The implication is that probability of an infinite number of type I errors is zero, so we may sometimes say that the sequence makes a finite number of errors without reference to probabilities. It is important to realize that no assumptions about independence have been made. The Borel-Cantelli lemma allows for arbitrary dependence of the events; indeed, the sequence of tests will be highly dependent. As more data becomes available, and new hypothesis tests are made, not only will the inference become arbitrarily accurate, but the sequence of tests as a whole will be very accurate. A sufficient but not necessary condition for this limit to hold is $h \propto n^{-p}$ with $p > 1$.

Of course, we have yet to consider the number of type II errors that occur. It will be useful to write the power function in terms of the Q-function, which is defined by $Q(x) \equiv 1 - \Phi(x)$. The power of the test is given by

$$\beta_{\mathrm{n}} = \mathrm{Q}\left(z_h - (\mu - t)\sqrt{n}\right) \geq Q(z_h) = h$$

whenever the null hypothesis is false. Notice that when the null hypothesis is false, our previous bound will be flipped; however, the following holds

$$\lim_{n\to+\infty} z_h - (\mu - t)\sqrt{n} = \begin{cases} +\infty & \mu < t \\ +\infty & \mu = t \\ -\infty & \mu > t \end{cases}$$

which is the result of Lemma 2.1 and convergence of the power function is given below.

$$\beta_{\mathrm{n}} \longrightarrow \begin{cases} 0 & \mu < t \\ 0 & \mu = t \\ 1 & \mu > t \end{cases}$$

Using this scheme as the sample size grows, the probability of a type I or type II error goes to zero asymptotically. Again we would like to show that the sum of the probabilities of a type II errors is finite.

**Theorem 2.1.** *Let $x_i = \mu + \epsilon_i$ where $\{\epsilon_i\}_{i\in\mathbb{N}}$ is an i.i.d. sequence of standard normal random variates and the sequence of hypothesis tests in (2.2) have critical values given by $z_h = \Phi^{-1}(1-h)$. Suppose $h \propto n^{-p}$ with $p > 1$, then with probability 1 the sequence of hypothesis test will make a finite number of errors.*

*Proof.* It has already been shown that there are a finite number of type I errors. As with the type I error, define a sequence of events that determines the probability of a type II error using the indicator function.

$$B_n = \mathbb{1}\big((\overline{x}-\mu)\sqrt{n} < \Phi^{-1}(1-h) - (\mu - t)\sqrt{n}\big)$$

where $B_n = 1$ means that a type II error has occurred. It must be shown that the sum of the probabilities of a type II error is finite. The probability of a type II error can be bounded for all $n > N$ where $N$ is chosen large enough for the following to hold.

$$\begin{aligned}\Pr(B_n = 1) = 1 - \beta_n = \; & Q\left((\mu - t)\sqrt{n} - z_h\right) \\ \leq \tfrac{1}{2} e^{-\frac{\left((\mu-t)\sqrt{n}-z_h\right)^2}{2}} \; & \leq e^{-\frac{(\mu-t)^2\sqrt{n}}{4}}\end{aligned}$$

The first inequality is the Chernoff bound for the Q-function, and it will be valid whenever $(\mu - t)\sqrt{n} - z_h > 0$. The second inequality follows because $(\mu - t)\sqrt{n} - z_h > (\mu - t)\sqrt{n/2}$ for suitably large $n$ which means

$$\sum_{i=N}^{+\infty} 1 - \beta_i \; \leq \sum_{i=N}^{+\infty} e^{-\frac{(\mu-t)^2\sqrt{n}}{4}} \; \leq M + \int_N^{+\infty} e^{-\frac{(\mu-t)^2\sqrt{x}}{4}} dx < +\infty$$

for some $M > 0$ and $N$ suitably large. Since $\sum_{i=N}^{+\infty} 1 - \beta_i \; \leq +\infty$, then the Borel-Cantelli lemma can be applied again to conclude

$$\Pr\left(\lim_{n\to+\infty} \sup B_n = 1 \;\right) = 0$$

which shows almost sure convergence and it follows that the number of type I and II errors must be finite. □

This demonstrates that by choosing the smoothing parameter appropriately, inference will not only become arbitrarily accurate as the sample size grows, but the entire sequence of tests will also perform very well in terms of type I error.

### *2.2. Two-sided hypothesis tests*

The two-sided case is nearly a trivial extension, but there are slight differences. As in the previous section, it is of interest to construct the following sequence of two-sided tests.

$$\begin{aligned} H_{0n}: \;& \mu = t \\ H_{1n}: \;& \mu \neq t \end{aligned} \tag{2.8}$$

The incorrect rejection of the null can be written with the indicator function

$$C_n = \mathbb{1}\big(|\overline{x} - \mu|\sqrt{n} > z_{h/2}\big) \tag{2.9}$$

where $1-\Phi\left(z_{h/2}\right)=h/2$. Of course, this means we will reject the null incorrectly $h$ percent of the time when the null hypothesis is true, i.e. the probability of a type I error will be $\Pr\left(C_n\right)=h$.

As with the one-sided test, if $h\propto n^{-p}$ with $p>1$, there can be only a finite number of type I errors. The probability of a type II error for the two-sided test is described below.

$$\begin{aligned}1-\beta_n &= \Pr\left(|\overline{x}-t|\sqrt{n}\ \leq z_{h/2}\right)\\ &= \Phi\left(z_{h/2}+(\mu-t)\sqrt{n}\right)+\Phi\left(z_{h/2}-(\mu-t)\sqrt{n}\right)-1\end{aligned}$$

The power of the two-sided test will converge to one for all alternatives. The sum of the probabilities of type II error will also converge, by a similar argument, as in the one-sided case, so the conclusion of Theorem 2.1 is valid for the two-sided case.

With this result in hand, we are able to resolve the Jeffreys-Lindley paradox, see [21]. Suppose the assumptions of Theorem 2.1 hold, except one is interested in the two-sided test given by Eq. (2.8). If one assumes a uniform prior density for $\mu$ over some interval $(t+I/2,t-I/2)$ centered at $t$ with a prior probability of $H_0$ given by $\pi_0$. Under these conditions, if it happens to be the case that $|\overline{x}-t|=z_{h/2}/\sqrt{n}$, then the posterior probability of the null hypothesis, $H_0$, is

$$\Pr\left(H_0\mid x_1,...,x_n\right)=\frac{I\pi_0\phi(z_{h/2})\sqrt{n}}{I\pi_0\phi(z_{h/2})\sqrt{n}+1-\pi_0}\to 1 \tag{2.10}$$

as $n\to+\infty$ whenever $\pi_0>0$ and $h=\alpha$ is fixed. Since the posterior probability converges to one, there will be some $n$ suitably large such that $H_0$ is rejected at the $\alpha$ significance level. But the posterior probability that the null is true is approximately $1-\alpha$. This puts the Bayesian and frequentist approach in direct conflict, which is the paradox.

But this paradox will not occur for A.S. hypothesis testing that sets $h=n^{-p}$ with $p>1$. First, note the Q-function satisfies the following well-known inequality,[2]

$$\frac{z_{h/2}^2}{1+z_{h/2}^2}\phi(z_{h/2})\leq z_{h/2}Q(z_{h/2})\leq\phi(z_{h/2})$$

whenever $z_{h/2}>0$. Since $z_{h/2}$ is increasing in $n$, it follows

$$\frac{1}{k_n\left(z_{h/2}\right)}\equiv\frac{z_{h/2}Q(z_{h/2})}{\phi(z_{h/2})}\to 1 \tag{2.11}$$

as $n\to+\infty$ and $Q(z_{h/2})=h/2$, so Eq. (2.10) can be rewritten as

$$\Pr\left(H_0\mid x_1,...,x_n\right)=\frac{I\pi_0 z_{h/2}hk_n\sqrt{n}}{I\pi_0 z_{h/2}hk_n\sqrt{n}+2\left(1-\pi_0\right)}\to 0$$

as $n\to+\infty$ whenever $\pi_0>0$ and $h=n^{-p}$ with $p>1/2$, which follows from Lemma 2.1. The posterior probability that the null hypothesis is true goes to

[2] See [14] for a similar discussion of the Q-function.

zero, which agrees with the frequentist rejection of the null. Thus, there is no paradox using A.S. hypothesis testing.

Now suppose the sequence of two-sided hypothesis tests are pairwise independent, such as the case would be if there was a sequence of hypothesis tests based on independent samples, but the size of each sample was growing. In this scenario, under any fixed significance level, we have $\sum_{i=1}^{+\infty} h\,(i) = +\infty$. This allows the second Borel-Cantelli lemma to be applied.

$$\Pr\left(\lim_{n\to+\infty} \sup C_n = 1\right) = 1 \tag{2.12}$$

If the significance level is fixed for such a sequence of tests, the sequence of tests will make an infinite number of errors, under the null, with probability one. Furthermore, one can always increase the sample size and retest until a statistically significant result is found, which is known as optional stopping.

However, in the present context, one cannot assume the samples are independent, but as Lindley points out in [21], the law of the iterated logarithm can be used when the independence assumption is dropped. To see why this is the case: assume $\mu = t = 0$, then we have the following modification of the law of the iterated logarithm

$$\Pr\left(\limsup_{n\to+\infty} \frac{\sqrt{n}\,|\overline{x}| - z_{h/2}}{\sqrt{\ln\ln n}} = \sqrt{2}\right) = 1$$

whenever $h = \alpha$ is fixed. This implies $\sqrt{n}\,|\overline{x}| - z_{h/2} > 0$ infinitely often with probability one, so one can increase the sample size until statistical significance is found with probability one. However, once the significance level is allowed to decrease with the sample size

$$\Pr\left(\limsup_{n\to+\infty} \frac{\sqrt{n}\,|\overline{x}| - z_{h/2}}{\sqrt{\ln\ln n}} = -\infty\right) = 1$$

whenever $h = n^{-p}$ with $p > 0$. This is because $z_{h/2} \approx \sqrt{2p\ln(n)}$ for large $n$, see [14], and $2p\ln(n)/\ln\ln n \to +\infty$, so A.S. hypothesis testing will be more robust to optional stopping. Of course, this is not too surprising because we have already shown that when $p > 1$, $\sqrt{n}\,|\overline{x}| - z_{h/2} > 0$ at most a finite number of times with probability one.

For example, suppose a researcher performed an experiment with a sample size of 10 and found no statistically significant result. Then suppose she decided to add one more observation, and retest continuing this process until a significant result was found. Under this scenario,[3] given the initial batch of 10 observations resulted in an insignificant result, the probability that the experiment will be stopped at some finite sample size, $N_s$, can be bounded using Boole's inequality

$$\Pr\left(N_s < +\infty\right) < \sum_{n=11}^{\infty} h < 0.0952$$

[3] A similar process is considered by [11] for a simulation in the context of animal testing.

where $h = n^{-2}$. This compares favorably with fixed significance level testing, which has a finite stopping time with probability one; however, this bound will not be meaningful for all bandwidths, as the above sum can be greater than one (the bandwidth in Eq. (3.5) would be one example). But even using that bandwidth, if the testing was done in batches of 10, then

$$\Pr\left(N_s < +\infty\right) < \sum_{i=2}^{\infty} (10i)^{-1.2} < 0.3$$

which results in a relatively large probability that the process will never end. Of course, these bounds rely on the assumption of normality, which can be untenable in many cases.

## 3. The general case with i.i.d. data

The analysis up to this point has assumed normality with unit variance for the special case of the sample mean, which is not very helpful for real world applications, which brings us to the main results.

### *3.1. Hypothesis testing*

If the data is not normal, then the test statistics are only valid asymptotically, so another approach must be taken. The Edgeworth expansion is a natural choice because it allows the finite sample distribution of the test statistic to be approximated by a normal distribution function. Furthermore, the smooth function model of the Edgeworth expansion is quite general and includes statistics such as means, variances, and even M-estimators with appropriate modifications.

Following the formulation in [14], suppose $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. distributed d-dimensional vectors satisfying $EX_1 = \mu$ with sample mean given by $\bar{X}_n = \frac{1}{n}\sum_i X_i$. Let $g$ and $h$ be smooth functions satisfying $g, h : \mathbb{R}^d \to \mathbb{R}$. Suppose $\hat{\theta} = g\left(\bar{X}_n\right)$ is an estimator of some parameter of interest, $\theta_0 = g\left(\mu\right)$. Similarly, $\hat{\sigma} = h(\bar{X}_n)$ is some consistent estimator of the asymptotic variance of $\hat{\theta}\sqrt{n}$ and define the relevant studentized test statistic by $S_n\sqrt{n} = \frac{\left(\hat{\theta}-\theta_0\right)\sqrt{n}}{\hat{\sigma}}$, then under the following conditions:

**Assumption 1.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. distributed d-dimensional vectors satisfying $E\|X_1\|^4 < +\infty$, $\limsup_{\|t\|\to+\infty} \left|E\left(e^{it'X_1}\right)\right| < 1$, $\hat{\sigma} \to_p \sigma > 0$ and $S_n$ has 4 continuous derivatives in a neighborhood of $\mu$ with $\nabla g\left(\mu\right) \neq 0$*

there exists an Edgeworth expansion of the studentized statistic given by

$$\sup_{|y|<+\infty} \left|\Pr\left(S_n\sqrt{n} \leq y\right) - \Phi\left(y\right) - \frac{\pi_1\left(y\right)\phi(y)}{\sqrt{n}} - \frac{\pi_2\left(y\right)\phi(y)}{n}\right| = O(n^{-\frac{3}{2}}) \quad (3.1)$$

where $\pi_j$ is a polynomial of degree $3j-1$. This kind of expansion is quite general, and the assumptions can be modified so that Eq. (3.1) may also be

applied to studentized regression estimates, see [14, 25]. Continuing with the smooth function model, the main results are discussed below.

**Theorem 3.1.** *Suppose Assumption 1 is satisfied and define the sequence of hypothesis tests by*

$$\begin{aligned} H_{0n}: &\quad \theta_0 \leq t \\ H_{1n}: &\quad \theta_0 > t \end{aligned}$$

*using the test statistic* $S_n\sqrt{n}$ *with critical values given by* $z_h = \Phi^{-1}(1-h)$ *and* $h \propto n^{-p}$ *with* $p > 1$*, then with probability* 1 *the sequence of hypothesis test will make a finite number of errors.*

*Proof.* The probability of a type I error can be bounded

$$\Pr\left(S_n\sqrt{n} > z_h + \frac{\sqrt{n}\,(t-\theta_0)}{\hat{\sigma}}\right) \leq P\left(S_n\sqrt{n} > z_h\right)$$

because $\sqrt{n}\,(t-\theta_0) \geq 0$ under the null hypothesis. Using Eq. (3.1) and Eq. (2.11), we have the following

$$\Pr\left(S_n\sqrt{n} > z_h\right) \leq Cn^{-\frac{3}{2}} + h\left[1 + \frac{|\pi_1(z_h)|z_h k_n(z_h)}{\sqrt{n}} + \frac{|\pi_2(z_h)|z_h k_n(z_h)}{n}\right]$$

where $\pi_1$ and $\pi_2$ are polynomials of degree two and five respectively and $Q\left(z_h\right) = h$. Since $z_h$ is increasing and $k_n\left(z_h\right) \to 1$, the leading term of the polynomial dominates, which means a constant, $D > 0$, can be chosen such that,

$$\frac{|\pi_1|z_h k_n(z_h)}{\sqrt{n}} < \frac{Dz_h^3}{\sqrt{n}} \leq \frac{D(-2\ln(2h))^{\frac{3}{2}}}{\sqrt{n}} = o(1)$$

for any $h \propto n^{-p}$ and $n$ sufficiently large. A similar argument will hold for $\pi_2$ which means

$$P\left(S_n\sqrt{n} > z_h\right) \leq Cn^{-\frac{3}{2}} + \mathrm{h} + \mathrm{o(h)} \tag{3.2}$$

for sufficiently large $n$. Thus almost sure convergence will follow from the summability of $h$. The probability of a type II error can be bounded similarly

$$1 - \beta_n \leq Cn^{-\frac{3}{2}} + \mathrm{Q}\left(w_n\right) + \frac{|\pi_1\left(w_n\right)|\phi\left(w_n\right)}{\sqrt{n}} + \frac{|\pi_2\left(w_n\right)|\phi\left(w_n\right)}{n}$$

where $w_n = \frac{(\theta_0-t)\sqrt{n}}{\hat{\sigma}} - z_h = \frac{(\theta_0-t)\sqrt{n}}{\sigma} + o\left(\sqrt{n}\right)$ and the order of convergence will now be $\sqrt{n}$, so for suitably large N, q, and $M > 0$.

$$\sum_{n=N}^{\infty} \frac{|\pi_1\left(w_n\right)|\phi\left(w_n\right)}{\sqrt{n}} \leq |\pi\left(w_N\right)|\,\phi\left(w_N\right)N^{-.5} + \int_N^{+\infty} M(\theta_0 - t)^{2q} x^{q-1/2} e^{-\frac{(\theta_0-t)^2\sqrt{x}}{2\sigma^2}} dx < \infty$$

Similar arguments can be made for the other two terms of the expansion, so the probabilities of both a type I and type II error are summable, allowing the Borel-Cantelli lemma can be applied justifying the claim. □

The two-sided test is a trivial extension of the previous result. Theorem 3.1 relies not only on the moments of the random variables, but also on the smoothness of the distribution, which manifests itself in the following condition.

$$\limsup_{\|t\|\to+\infty} \left| E\left(e^{it'X_1}\right)\right| < 1 \tag{3.3}$$

This condition requires that at least one of the $d$ random variables is absolutely continuous. In many practical applications such as randomized experiments, one cannot expect such an assumption to hold.

Nevertheless, as demonstrated in [14] and [3], an Edgeworth expansion can still be constructed for lattice random variables with some modifications, so Theorem 3.1 remains true by replacing Condition 3.3 with the following

$$\Pr\left(X_{1i} = a_i + jb_i\right) = 1 \tag{3.4}$$

where $j = \pm 1, \pm 2, ...$ and $i = 1, ..., d$. To see why this is the case, the one term Edgeworth expansion for a lattice random variable in one dimension is

$$\Pr\left(S_n\sqrt{n} \leq y\right) = \Phi\left(y\right) + \frac{\pi_1\left(y\right)\phi(y)}{\sqrt{n}} + \frac{bR\left(b^{-1}y\sqrt{n} - b^{-1}an\right)\phi(y)}{\sqrt{n}} + O(n^{-1})$$

where $R(y) = [y] - y - 1/2$ and $[y]$ is the floor function. Since $|R| < 3/2$, the summability in Theorem 3.1 will not be affected; however, the small sample performance for lattice random variates may not be as good.

The proof of Theorem 3.1 provides some guidance as to the choice of $p$. As one can see from Eq. (3.2), under the null the bandwidth will be a tight approximation of the type I error in the following sense

$$\frac{P\left(S_n\sqrt{n} > z_h\right)}{h} = O\left(1\right)$$

as long as $p \leq 3/2$. If $p > 3/2$, then one may require additional assumptions in order to be sure the bandwidth accurately reflects the probability of rejection under the null hypothesis.

There is also a trade-off between type I and type II errors. For any fixed sample size, as $p$ approaches one from above, the probability of a type I error increases, but so does the power. It would seem a choice of $p$ closer to one would strike a balance between type I and type II errors. The following bandwidth seems to perform well in simulations.

$$h_1 = n^{-\frac{6}{5}} \tag{3.5}$$

Table 1 shows various critical values for the one-sided and two-sided A.S. hypothesis tests with smoothing parameter defined by Eq. (3.5). It may seem unsettling that the critical values these hypothesis tests are based upon are diverging to infinity, but the rate of growth is very slow.

Even with a million observations, the critical value for the two-sided test is only about three times as large as the normal critical value associated with 0.05 level test. The hurdle that researchers must overcome is bigger, but not excessive.

TABLE 1
*Critical values*

| N | One-sided | Two-sided | Digit rule |
|---|---|---|---|
| 5 | 1.06 | 1.46 | 2 |
| 10 | 1.53 | 1.86 | 3 |
| 15 | 1.76 | 2.07 | 3 |
| 25 | 2.03 | 2.31 | 3 |
| 50 | 2.36 | 2.61 | 3 |
| 100 | 2.65 | 2.88 | 4 |
| 1,000 | 3.48 | 3.66 | 5 |
| 10,000 | 4.16 | 4.32 | 6 |
| 100,000 | 4.75 | 4.89 | 7 |
| 1,000,000 | 5.28 | 5.41 | 8 |

Another interesting thing to notice is that the growth of the critical values is roughly proportional to the number of digits in the sample size. For example, when the sample size is $9,999$, there are 4 digits in the number representing the sample size, while the critical value for this sample size is 4.32 for the two-sided test. If, for simplicity, we took the number of digits plus one as our critical value, then almost sure convergence would be guaranteed.

When researchers are reading papers, they might be interested to know if some statistical test is significant in the A.S. hypothesis setting, which we call A.S. significance. If the test statistic is available, then a simple test can be used to verify A.S. significance. We say that the "digit-rule"is the critical value given by the number of digits in the sample size plus one.

Since the digit-rule only increases at a new order of magnitude, it is sufficient to compare the critical values right before the next order of magnitude is reached. So the digit-rule for a sample size of 99 would be 3, and one can infer from Table 1 that the critical value for the digit-rule is greater than the critical value for the bandwidth in Eq. (3.5), and a similar argument can be made for the other orders of magnitude, which implies A.S. significance. As the sample size grows, the digit-rule becomes increasingly conservative, and relatively less powerful, so it should be used with caution in large samples.

Previously, we used the law of the iterated logarithm to argue that using a fixed significance level results in an infinite number of type I errors with probability one. However, the law of the iterated logarithm is not directly applicable to the studentized statistic. Nevertheless, a similar result can be proven: recall the Kochen-Stone lemma, see [1] or [17].

**Lemma 3.1.** *Suppose $E_1, E_2, ...$ are events satisfying*

$$\sum_{n=1}^{+\infty} \Pr(E_n) = +\infty$$

$$\liminf_{k\to+\infty} \frac{\sum_{n=1}^{k}\sum_{m=1}^{k} \Pr\left(E_n \bigcap E_m\right)}{\left(\sum_{n=1}^{k} \Pr(E_n)\right)^2} < +\infty$$

*then infinitely many of the events take place with positive probability.*

Now we may demonstrate any sequence of tests with fixed significance level will make an infinite number of mistakes, at least with positive probability.

**Lemma 3.2.** *Suppose* $C_n = \mathbb{1}\left(|S_n|\sqrt{n} > z_{h/2}\right)$ *with* $h = \alpha$ *fixed under the null and the assumptions of Theorem 3.1 are satisfied, then an infinite number of errors are made with positive probability.*

*Proof.* From equation (3.1), it is clear that $\Pr(C_n) = \alpha - o\left(\frac{1}{\sqrt{n}}\right)$, so $\Pr(C_n)$ converges to $\alpha$ which means there exists an $N$ such that $\Pr(C_n) > \frac{\alpha}{2}$ for all $n > N$ giving $\sum_{n=1}^{+\infty} \Pr(C_n) = +\infty$, so the first part of the Lemma 3.1 is satisfied, furthermore, it follows that there exists some $c > 0$ such that

$$\frac{1}{k}\sum_{n=1}^{k} \Pr(C_n) > c$$

for all $k \leq +\infty$. For the second part, using Boole's inequality

$$\frac{\sum_{n=1}^{k}\sum_{m=1}^{k} \Pr\left(C_n \bigcap C_m\right)}{\left(\sum_{n=1}^{k} \Pr(C_n)\right)^2} < 2k\frac{\sum_{n=1}^{k} \Pr\left(C_n\right)}{\left(\sum_{n=1}^{k} \Pr(C_n)\right)^2} \leq \frac{2}{\frac{1}{k}\sum_{n=1}^{k} \Pr(C_n)} < \frac{2}{c}$$

upon taking limits the result follows. □

### *3.2. Confidence intervals*

Confidence intervals play a central part in most problems of statistical inference, so it is useful to demonstrate that the approach can be used to construct a sequence of confidence intervals that cover the true parameter all but a finite number of times, which we call an A.S. confidence interval. Consider the following one-sided and two-sided confidence intervals.

$$I_{1n}(h) = \left(-\infty, \hat{\theta} + n^{-.5}\hat{\sigma}z_h\right) \tag{3.6}$$

$$I_{2n}(h) = \left(\hat{\theta} - n^{-.5}\hat{\sigma}z_{h/2}, \hat{\theta} + n^{-.5}\hat{\sigma}z_{h/2}\right) \tag{3.7}$$

Again, an Edgeworth expansion can be used to demonstrate these confidence intervals are, in fact, A.S. confidence intervals.

**Theorem 3.2.** *Suppose the conditions of Theorem 3.1 are satisfied, then*

$$\Pr\left(\theta_0 \in I_{1n}(h)\right) \to 1$$
$$\Pr\left(\theta_0 \in I_{2n}(h)\right) \to 1$$

*and with probability* 1 *there are only a finite number of confidence intervals which do not contain* $\theta_0$.

*Proof.* It follows from Theorem 3.1

$$\Pr(\theta_0 \in I_{1n}) = \Pr\left(\theta_0 \leq \hat{\theta} + n^{-1/2}\hat{\sigma} z_h\right) = \Pr(S_n > -z_h)$$
$$= 1 - \Phi(-z_h) - n^{-1/2}\pi_1(-z_h)\phi(-z_h) - n^{-1}\pi_2(-z_h)\phi(-z_h) + O(n^{-3/2})$$
$$= 1 - h + O(n^{-3/2})$$

and the $\Pr(\theta_0 \notin I_{1n})$ will be summable demonstrating the result. As for the two-sided confidence interval,

$$\Pr(\theta_0 \in I_{2n}) = \Pr\left(S_n > -z_{h/2}\right) + \Pr\left(S_n > z_{h/2}\right)$$
$$= 1 - h + 2n^{-1}\pi_2\left(z_{h/2}\right)\phi\left(z_{h/2}\right) + O(n^{-2}) = 1 - h + O(n^{-2})$$

because $\pi_1$ and $\pi_3$ are odd functions resulting in a cancellation, so $\Pr(\theta_0 \notin I_{2n})$ is summable. □

The interpretation is the same as the previous results. As the sample size gets large, the probability that the confidence interval will not cover $\theta_0$ goes to zero; furthermore, there will be only a finite number of times that this confidence interval will fail to contain $\theta_0$.

### *3.3. Multiple comparisons*

In standard hypothesis testing, an issue of multiple comparisons arises when a large number of hypothesis tests are conducted. Since the type I error is fixed, a large number of hypothesis tests results in the high likelihood of finding a statistically significant effect which is spurious, sometimes called a false discovery.

For example, if one performs 100 two-sided hypothesis tests with a significance level of 0.05, then one would expect that at least one of the tests would find a significant result when there was none; however, we will show A.S. hypothesis testing is robust to such multiple comparisons.

Of course, one can directly control for multiple comparisons with a bandwidth of the form

$$h \propto \frac{n^{-p}}{m}$$

where $p > 1$ and $m$ is the number of comparisons, which would be similar to a Bonferroni correction. For the above example, if we also assumed a sample size of 150 with the bandwidth in Eq. (3.5), this would result in a critical value of 4.22 compared to 3.03 without the correction. Meanwhile, using a Bonferroni correction with a significance level of 0.05 results in a critical value of 3.48. So even in this simple example, one can see that A.S. critical values change less than under a fixed significance level when a correction for multiple comparisons is used. In this section, we will mainly focus on the role of multiple comparisons for confidence intervals, but the results hold for hypothesis testing as well.

If, for some reason, a researcher was unaware of exactly how many comparisons were made, this approach still works without any modifications. Instead of a single sequence of confidence intervals, we allow for a family of $m$ sequences of confidence intervals. If the same bandwidth is used for each of the $m$ sequences of confidence intervals, then *each* of the $m$ sequences will make a finite number of errors with probability one, so one would expect that *all* of the $m$ sequences will make a finite number of errors with probability one. This intuition is verified below.

**Lemma 3.3.** *Let the conditions of Theorem 3.1 be satisfied for m sequences of two-sided confidence intervals*

$$I_{2ni}(h) = \left(\hat{\theta}_i - n^{-.5}\hat{\sigma}_i z_{h/2}, \hat{\theta}_i + n^{-.5}\hat{\sigma}_i z_{h/2}\right)$$

*and* $i = 1, .., m$*, then with probability* 1 *the family of confidence intervals makes only a finite number of errors.*

*Proof.* It follows from Theorem 3.1 that

$$\Pr\left(\theta_{0i} \in I_{2i_n}(h)\right) = 1 - h + o(h)$$

and the probability that at least one of the members of the family of confidence intervals makes an error can be bounded by Boole's inequality.

$$\Pr\left(\bigcup_i \left[\theta_{0i} \notin I_{2ni}(h)\right]\right) \leq mh + o(h)$$

which will be summable for fixed $m$ as long as $h \propto n^{-p}$ with $p > 1$. □

Even with a fixed number of confidence intervals, Lemma 3.3 may have poor small sample performance unless $m \ll n$. For instance, continuing with the previous example, if there are 100 independent hypothesis tests with a sample size of 150 for each test, the probability of finding a false discovery will be approximately 0.24 using Eq. (3.5), so one may find better small sample performance in terms of type I error by setting the bandwidth in the following manner:

$$h_2 = n^{-2} \tag{3.8}$$

resulting in a critical value of 4.08 for the example above. This is similar to controlling for the number of comparisons directly, which has a critical value of 4.22. The rationale of Eq. (3.8) is that in smaller samples whenever $m \approx n$, one would still have a small probability of false discovery, $mh \approx n^{-1}$.

### *3.4. Simulations*

Since all of the previous results are asymptotic, it is important to see how well the approach holds up in finite samples. The first simulation compares the size of A.S. hypothesis testing with a standard significance level and compares the

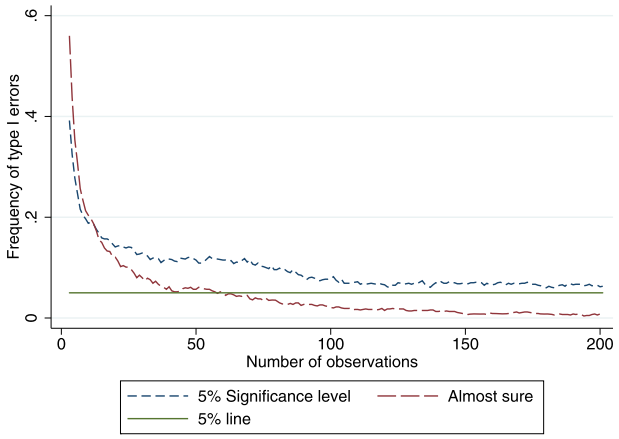


FIG 1. *Frequency of type I errors for two-sided hypothesis test.*

empirical coverage probabilities of the 95%, almost sure, and digit-rule two-sided confidence intervals. The second simulation will compare the power of the two approaches. The third simulation compares A.S. hypothesis testing for lag selection in an autoregressive model with other lag selection criteria.[4]

Consider the following sequence of two-sided hypothesis tests, where under the null the mean, $\theta_0$, will be 1 and under the alternative the mean will be 2.

$$\begin{aligned} H_{0n}: &\quad \theta_0 = 1 \\ H_{1n}: &\quad \theta_0 = 2 \end{aligned}$$

For the first simulation, 200 independent draws are made from a Chi-square distribution with one degree of freedom to simulate the case when the null is true. We proceed by performing a sequence of hypothesis tests for every sample size between 3 and 200 and then repeat the process for $1,000$ sequences of hypothesis tests and the frequency of type I errors are tabulated in Figure 1.

As one can see from the graph, the number of type I errors for the A.S. hypothesis test converges to zero while the hypothesis tests with a 5% significance level fall to around the 5% mark as expected. Except in very small samples, A.S. hypothesis testing performs better than the standard approach in terms of type I error.

In Figure 2 using the same simulated data, a sequence of two-sided confidence intervals are computed using the 95%, almost sure, and digit-rule. The frequency of errors for the 95% confidence interval approaches 5% from above. Meanwhile, both the digit-rule and A.S. confidence intervals have failures that are converging to zero quite quickly. The digit-rule covers the mean more often than the A.S. confidence interval because it has larger critical values.

The second simulation is performed in the same fashion as the first except the alternative hypothesis is imposed by taking draws from a Chi-square distribution

[4]Except for the digit-rule, all simulations will use the bandwidth in Eq. (3.5).

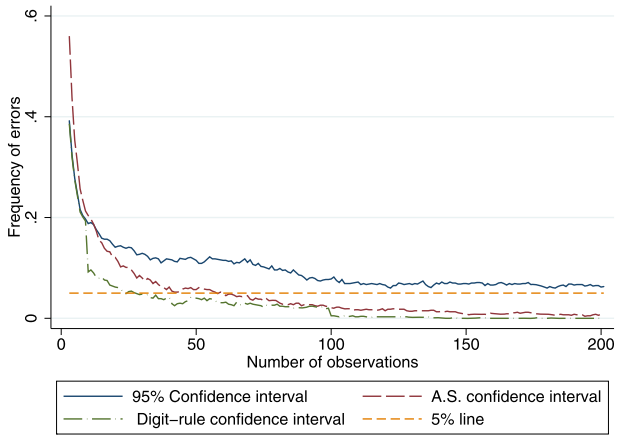


FIG 2. *Percentage of two-sided confidence intervals that failed to cover the mean.*

with 2 degrees of freedom. Generally, the power of the A.S. hypothesis test will be less than with a fixed significance level.[5] This is due to the divergence of the critical values with the A.S. approach making the tests generally less powerful relative to some fixed significance level. The rejection frequencies under the alternative are compared in Figure 3.

As expected, the power under the fixed significance level is generally greater than the A.S. test; however, the power of the A.S. test is still acceptable. The reason the A.S. test is more powerful in small samples is because the critical values happen to be smaller than 1.96, but as the sample grows the fixed significance level becomes more powerful. Regardless, depending on the problem, A.S. hypothesis testing could suffer from low power in small samples and one should proceed with caution in such cases.

For the final simulation, consider the problem of estimating an AR(p) process with unknown lag length.

$$y_i = \sum_{j=1}^{p} \gamma_j y_{i-j} + \epsilon_i \tag{3.9}$$

One way to determine the lag length is to chose some large fixed lag length, $L$, and then perform backward elimination to eliminate extraneous lags utilizing an appropriate hypothesis test, the likelihood ratio test (LR) is an example of this approach. In general, this will not produce a consistent estimate of the lag length, $p$.

Suppose $L = p+1$ and a significance level of 0.05 is used, then in large samples the probability of rejecting the null for the $p+1$ lag will be approximately 0.05

[5]In small samples, depending on the bandwidth, the A.S. critical value may be smaller than the critical value for some fixed significance level in which case the power may be greater for the A.S. hypothesis test.

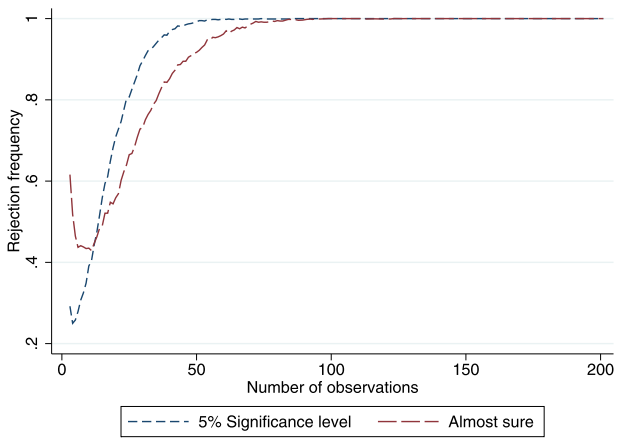


FIG 3. *Rejection frequency under the alternative.*

even with an infinite sample size, which is a type I error. Furthermore, in finite samples, there is a nonzero probability that the null for the pth lag will not be rejected, which is a type II error. In this case, both type I and type II errors need to be small to make for a reliable test of the lag length.

The standard approach to overcome this issue is to minimize some sort of information criteria, which amounts to choosing the model that minimizes the sum of squared residuals subject to some penalty for the number of parameters in the model. Common criteria include final prediction error (FPE), Akaike's information criterion (AIC), Schwarz bayesian information criterion (SBIC), and Hannan-Quinn information criterion (HQIC). The different merits of these information criteria are discussed in [20, 22, 8]. Like the LR approach, both the FPE and AIC overestimate the true lag with positive probability while the SBIC and HQIC are consistent, see [22].

As discussed previously, A.S. hypothesis testing may suffer from low power; however, this certainly doesn't preclude its use even when power is important. If we assume the residuals in Eq. (3.9) are i.i.d., then OLS can be applied. If the lag length, $L$, does not depend on the sample size, then the Edgeworth expansion results of [14] can be applied to the OLS estimates. If an A.S. hypothesis test is used in the example above, then in large samples the probability of rejecting the null for the $p+1$ lag will be arbitrarily small; however, the power of the test still comes into play because one must reject the null for the pth lag. For the A.S. hypothesis approach, the bandwidth in Eq. (3.5) is used for a two-sided test with backward elimination, like the LR approach.

In order to compare these different lag selection mechanisms, a simulation is conducted similar in nature to the simulations in [20, 8]. First we generate an AR(2) series and estimate the lag length in the 6 ways outlined above using a maximum lag of 10 at various points in the sample. This is similar to the choice made in [20]. We also allow the second coefficient, $\gamma_2$, to be small, but

TABLE 2
*Comparison of different lag selection mechanisms*

| N | FPE | AIC | HQIC | SBIC | LR | A.S. |
|---|---|---|---|---|---|---|
| 25 | 15.7% | 15.7% | 11.3% | 14.5% | 9.0% | 34.8% |
| 50 | 39.5 % | 39.5 % | 43.0 % | 42.7 % | 28.2% | 44.0% |
| 100 | 57.1% | 57.1% | 65.0% | 64.0% | 35.6% | 58.6% |
| 500 | 69.7% | 69.7% | 88.2% | 91.9% | 38.3% | 87.5% |
| 1,000 | 71.4% | 71.4% | 94.0% | 97.6% | 38.6% | 94.5% |
| 5,000 | 70.9% | 70.9% | 94.5% | 99.5% | 38.1% | 100.0% |
| 10,000 | 72.8% | 72.8 % | 94.9% | 99.8% | 38.5% | 100.0% |

not arbitrarily close to zero.

1. Repeatedly draw $\gamma_1$ and $\gamma_2$ from the uniform distribution in the range $(-1, 1)$ until the following conditions are met: $|\gamma_1 + \gamma_2| < 1$ and $|\gamma_2| > 0.1$
2. Generate an AR(2) process according to $y_i = 1 + \gamma_1 y_{i-1} + \gamma_1 y_{i-2} + \epsilon_i$, where $\{\epsilon_i\}$ is an i.i.d. sequence of uniformly distributed random variables on $(-4, 4)$
3. Discard the first $20,000$ observations
   (a) Compute the 6 lag order selection mechanism and tabulate the results for the first $N$ observations using a maximum lag of 10
   (b) Repeat for the 6 different sample sizes
4. Repeat 1000 times and tabulate the results

The results of the simulation are reported in Table 2.[6] The LR approach, which uses the 5% significance level, is the only other sequential hypothesis test and it performs the worst; however, the sequential testing approach goes from the worst performing to arguably the best using the A.S. methodology.

In small samples, the A.S. approach selects the correct lag more often than any other lag selection criteria. Since there are 10 lags, there are as few as 15 observations in the regressions. It dominates the AIC, FPE, and LR for all of the tabulated sample sizes. This should be expected because the LR, AIC, and FPE model selection criteria are inconsistent.

The HQIC and SBIC are consistent estimators of the lag length. Both of these criteria are dominated by the A.S. approach in very small and very large samples. One could also use the A.S. critical values to determine the lag length without backward elimination and perform a single regression instead. In simulations this approach was less effective than using backwards elimination, but A.S. still dominated in very large samples.

## 4. Extension to the strongly mixing process

In many circumstances, the assumption of an i.i.d. process is far too strong, especially in a time series settings. In the previous section, we described how an

[6]In other variations of the simulation, the maximum lag was allowed to increase with the sample size. Alternative distributions were used for the error term. The assumption that the autoregressive coefficients being not arbitrarily close to zero was also dropped. In each of the variations, the effectiveness of the A.S. approach was similar to the results in Table 2.

Edgeworth expansion could be applied to the studentized statistic. In the case of the sample mean, the studentized statistic can be written in the following manner.

$$S_n = \frac{\sqrt{n}\overline{x}_n}{\sqrt{\frac{1}{n}\sum_i x_i^2 - \overline{x}_n}}$$

The studentized statistic is a smooth function of $\overline{x}_n$ and $\frac{1}{n}\sum_i x_i^2$, so the standard approach applies. However, even if the process is stationary, the estimation problem becomes more difficult.

For a covariance stationary process with summable autocovariances, the asymptotic variance will be

$$Var\left(\sqrt{n}\overline{x}_n\right) \to \gamma(0) + 2\sum_{i=1}^{\infty}\gamma(i) \tag{4.1}$$

where $\gamma(i) = Cov(x_1, x_{i+1})$ is the lag-covariance. The asymptotic estimate now depends on an infinite number of parameters. Obviously one cannot construct an infinite number of estimates in any finite sample, so the asymptotic variance in Eq. (4.1) must be approximated by a truncated estimate

$$Var\left(\sqrt{n}\overline{x}_n\right) \approx \hat{\gamma}(0) + 2\sum_{i=1}^{l}\hat{\gamma}(i) \tag{4.2}$$

where $\hat{\gamma}(i) = n^{-1}\sum_{j=1}^{n-i} x_j x_{j+i} - \overline{x}_n^2$ and $l$ is a bandwidth parameter that tends to infinity as the sample size grows. The approximation results in a biased estimator, but it can be shown the estimate is consistent. Even though there are only a finite number of estimates, the dimension of the estimation problem is unbounded, so alternative methods must be used.

The Edgeworth expansion for studentized statistics when the data is a strongly mixing process with an exponential decay rate has been derived in [19]; however, the results of the paper hold more generally under polynomial decay, see [18]. According to [19, 5], the assumptions for this Edgeworth expansion can be modified, so the following results are also applicable for a large class of M-estimators.

The assumptions used to derive the expansion in [19] are quite intricate. Since the Edgeworth expansion is of secondary interest, results will be demonstrated for the vector linear process, as the expansions are the same for the more general case.

**Assumption 2.** *Let $\{X_i\}_{i\in\mathbb{Z}}$ be a d-dimensional random vector generated by a sequence of nonrandom matrices, $\{A_i\}_{i\in\mathbb{Z}}$, and i.i.d. d-dimensional vectors, $\{\epsilon_i\}_{i\in\mathbb{Z}}$ in the following way*

$$X_i = \mu + \sum_{j\in\mathbb{Z}} A_j \epsilon_{i-j} \tag{4.3}$$

*Suppose* $\|A_i\| = O\left(e^{-\kappa|i|}\right)$ *for some* $\kappa \in (0,1)$ *as* $|i| \to 0$ *and* $\sum_{i\in\mathbb{Z}} A_i$ *is nonsingular. Further suppose that* $\epsilon_1$ *is absolutely continuous with* 0 *mean and* $E\|\epsilon_1\|^{12+\delta} < +\infty$. *Finally assume* $g_n$ *has has* 4 *continuous derivatives in a neighborhood of* $\mu$ *with* $\nabla g_n(\mu) \neq 0$.

Under the smooth function model discussed in the previous section, we consider the studentized statistic given by $T_n\sqrt{n} = \frac{(\hat{\theta}-\theta_0)\sqrt{n}}{\hat{\tau}}$. The estimated variance, $\hat{\tau}$, is constructed in the following manner

$$\hat{\tau}^2 = max\left\{g\left(\overline{X}_n\right)'\left[\hat{\Gamma}(0) + \sum_{i=1}^{l} k\left(\frac{i}{l}\right)\left(\hat{\Gamma}(i) + \hat{\Gamma}(i)'\right)\right] g\left(\overline{X}_n\right), \frac{1}{n}\right\} \tag{4.4}$$

where $\hat{\Gamma}(i) = \frac{1}{n}\sum_{j=1}^{n-i}\left(X_j - \overline{X}_n\right)\left(X_{i+j} - \overline{X}_n\right)$ and $k$ is a weighting function. In the case of the sample mean, one could simply take the average of the autocovariances, but now we must estimate a matrix that grows arbitrarily large, so a weighting function must be used to obtain consistent estimation. We will consider the bounded kernel framework of [2]. This includes the most popular choices such as the Bartlet, Parzen and QS spectral kernel, but more general weighting schemes are also possible.

**Assumption 3.** *Let equation (4.4) be used as the studentizing factor where the kernel,* $k : [-1,1] \to \mathbb{R}$, *is an even function continuous at 0 and at all but a finite number of points. Suppose* $k(0) = 1$ *and* $\int x^2 k dx < +\infty$, *furthermore, assume the lag parameter,* $l$, *is chosen so that* $Bias\left(\hat{\tau}^2\right) = O\left(n^{-1/3}\right)$ *with* $\kappa \ln n < l \leq \kappa^{-1} n^{1/3}$ *for some* $\kappa \in (0,1)$ *and all suitably large* $n$.

This assumption is not necessary, but includes many of the common kernels used in application.

**Theorem 4.1.** *Under Assumptions 2 and 3, define the sequence of hypothesis tests*

$$\begin{aligned} H_{0n}: &\quad \theta_0 \leq t \\ H_{1n}: &\quad \theta_0 > t \end{aligned}$$

*using the test statistic* $T_n\sqrt{n}$ *with critical values given by* $z_h = \Phi^{-1}(1-h)$ *and* $h \propto n^{-p}$ *with* $p > 1$, *then the sequence of hypothesis test will make a finite number of errors with probability* 1.

*Proof.* Under Assumptions 2 and 3, [19] proves the following expansion

$$\sup_{|y|<+\infty}\left|P\left(T_n\sqrt{n} \leq y\right) - \Phi(y) - \Psi(y)\right| = O\left(\left(\frac{n}{l}\right)^{-3/2}\ln(n)^{-2}\right) \tag{4.5}$$

where

$$\Psi(y) = \phi(y)\left[y a_n^{-1} - \frac{y^3 a_n^{-2}}{2} + \frac{y^3\left(y^2-1\right)a_n^{-3}}{6} + \sum_{i=1}^{6} c_{in} p_{in}(y)\right]$$

where $Bias\left(\hat{\tau}^2\right) = a_n^{-1} = O\left(n^{-1/3}\right)$, $c_{in} = o\left(n^{-.5}\right)$ and $p_{in}$ is a polynomial with bounded coefficients. Notice $\Psi\left(z_h\right) = o(h)$, so we can focus on the summability of the right hand side of Eq. (4.5). By assumption, the lag length must be no greater than $l = O\left(n^{1/3}\right)$, which means

$$\sum_{n=2}^{+\infty}\left(\frac{n}{l}\right)^{-3/2}\ln\left(n\right)^{-2} \leq \frac{.5}{\ln(2)^2} + \int_2^{+\infty} x^{-1}\ln\left(x\right)^{-2}dx = \frac{.5+\ln(2)}{\ln(2)^2} \tag{4.6}$$

and we may conclude that a type I error occurs at most a finite number of times. The rest of the argument follows in the same fashion as Theorem 3.1. □

For example, if one were to use the Parzen kernel, then the optimal lag parameter would be $l = O\left(n^{1/5}\right)$ and the error of the Edgeworth expansion in Eq. (4.5) would be approximately $O\left(n^{-6/5}\ln\left(n\right)^{-2}\right)$ ensuring the bandwidth in Eq. (3.5) is appropriate. A notable exception is the Bartlet kernel which requires an optimal lag parameter to be $l = O\left(n^{1/3}\right)$, so in that case additional assumptions may be needed to justify the bandwidth in Eq. (3.5).

## 5. Conclusions

There is still more work to be done on this topic. It seems likely that this approach can be used in a bootstrap setting as well. The assumptions in this paper are far from minimal. The result could also be generalized to other test statistics such as an F-test.

A.S. hypothesis testing has similar critical values to the standard approach, but it is less reliant on arbitrary choices of significance level such as 0.05 or 0.01. Furthermore, using A.S. critical values resolves the Jeffreys-Lindley paradox and it is robust to multiple comparisons. A simple rule of thumb was also given that we call the digit-rule, which ensures almost sure convergence.

A.S. hypothesis testing was used as a lag order selection mechanism and in simulations the approach performed very well relative to other popular model selection criteria. Simulations were also conducted comparing the power and frequency of type I error with a fixed significance level.

The rates of convergence for the power and coverage probability may differ, but the procedure allows hypothesis testing and confidence intervals to be computed in such a way that the correct decision will be reached with a probability approaching one as the sample size increases. Despite the complexity of Edgeworth expansions, A.S. hypothesis testing can be conducted quite simply, so this approach has broad appeal.

## References

[1] Amghibech, S. (2006). On the Borel-Cantelli Lemma and moments. *Comment. Math. Univ. Carolin.* 669–679. MR2337421

[2] Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*. **59** 817–858. MR1106513
[3] Bhattacharya, R. and Rao, R. (1986). *Normal Approximation and Asymptotic Expansions*. Siam, Philadelphia.
[4] Bialik, C. (2012). How to be sure you've found a Higgs boson. *Wall Street Journal*. Retrieved from http://www.wsj.com.
[5] Bustos, O.H. (1982). General M-estimates for contaminated pth-order autoregressive process: consistency and asymptotic normality. *Z. Wahrsch. Verw. Gebiete*. **64** 211–239. MR0656512
[6] Chiani, M., Dardari, D. and Simon, M. K. (2003). New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*. **2** 840–845.
[7] Cousins, R. D. (2014). The Jeffreys-Lindley paradox and discovery criteria in high energy physics. *Synthese*. 1–38.
[8] Den Haan, W. and Levin, A. (1996). A practitioner's guide to robust covariance matrix estimation. Techincal Working Paper Series 197, NBER. MR1492717
[9] Dembo, A. and Peres, Y. (1994). A topological criterion for hypothesis testing. *The Annals of Statistics*. **22** 106–117. MR1272078
[10] Fisher, R.A. (1971)[1935]. *The Design of Experiments*. Macmillan, New York.
[11] Fitts, D. (2011). Ethics and animal numbers: informal analyses, uncertain sample sizes, inefficient replications, and type I errors. *Journal of the American Association for Laboratory Animal Science*. **50** 445–553.
[12] Hall, P. (1983). Inverting an Edgeworth expansion. *The Annals of Statistics*. **11** 569–576. MR0696068
[13] Hall, P (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. *The Annals of Probability*. **15** 920–931. MR0893906
[14] Hall, P (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237
[15] Hansen, B. (2000). Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions. *Manuscript, University of Wisconsin*. MR2235714
[16] Inoue, A. and Shintani, M. (2006). Bootstrapping GMM estimators for time series. *Journal of Econometrics*. **131** 531–555. MR2252908
[17] Kochen, S. and Charles, A. (1964). A note on the Borel-Cantelli Lemma. *Illinois J. Math*. **8** 248–251. MR0161355
[18] Lahiri, S.N. (1996). Asymptotic expansions for sums of random vectors under polynomial mixing rates. *Sankhya Ser. A*. **58** 206–224. MR1662519
[19] Lahiri, S.N. (2010). Edgeworth expansion for studentized statistics under weak dependence. *The Annals of Statistics*. **38** 388–434. MR2589326
[20] Liew, V.K. (2004). Which lag length selection criteria should we employ? *Economics Bulletin*. **33** 1–9.

[21] Lindley, D. V. (1957). A statistical paradox. *Biometrika*. **44** 187. MR0087273
[22] Lutkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. New York: Springer. MR2172368
[23] Lyons, L. (2013). Discovering the signifcance of 5 sigma. *arXiv:1310.1284*.
[24] Newey, W.K. and Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*. **72** 219–255. MR2031017
[25] Qumsiyeh, M.B. (1990). Edgeworth expansion in regression models. *Journal of Multivariate Analysis*. **35** 86–101. MR1084943