Revisiting Random Walk Questions with Newer and Better Tools

Richard H. Serlin, University of Arizona

Available at: https://works.bepress.com/richard_serlin/6/
Let’s jump right in with an instructive example. Suppose that given our prior information we feel that a standard linear parametric model is close enough to the true DGP that we are comfortable assuming that it is the true DGP. Then, we would start by assuming that the true model is $y_t = x_t \beta + \epsilon_t, \quad \epsilon \sim iid \ N(0, \sigma^2)$.

We then might want to have a null hypothesis; $H_0 : \beta = 0$ and an alternative hypothesis, $H_A : \beta \neq 0$. Now, it is important to note that, unlike in Lo and MacKinlay (1988), here whether the null hypothesis is true or the alternative hypothesis is true, the parametric model is exactly the same. It is $y_t = x_t \beta + \epsilon_t, \quad \epsilon \sim iid \ N(0, \sigma^2)$. The parametric model does not differ under the null and alternative hypotheses, only the value of the parameter vector does. So, in this situation the parametric model is identified, and so now it is possible to generate posterior densities for the parameters, and for functions of the parameters.

My goal in this paper is to provide posteriors (and confidence intervals) because they make economic significance much more clear than point hypothesis tests, but I can’t provide posteriors without assuming a parametric model. So, I propose assuming various interesting parametric time series models and then deriving the posteriors of their parameter vectors. What this means is that all my paper is doing so far is Bayesian/Monte Carlo estimating models appropriate for asset returns. Maybe all the interesting models have already been estimated in a Bayesian Monte Carlo way. Perhaps, that’s one of the things I’m hoping you could tell me. But, nonetheless, the parts of the paper which I will propose later involve significantly newer techniques, and I think they are much less likely to have already been done.

Also, perhaps I could make the paper more interesting by doing a Bayesian estimation of a time series model that’s been estimated before, but for significantly different assets. Per-
haps I could do a Bayesian estimation for some interesting asset that hasn’t been Bayesian
estimated before, perhaps REITs; I don’t know. I’m just posing the question to you guys
who really know the literature. Let’s now look at an example that I think will be instructive.

Suppose we assume that stock market data was generated by an AR(q) model. That
is:

\[ r_t = \phi_1 r_{t-1} + \phi_2 r_{t-2} + \ldots + \phi_q r_{t-q} + \epsilon_t \]  

(1)

where the \( \epsilon_t \) are iid, and we could assume that they are normal, or perhaps some fatter tailed
density like a t. When using a noninformative prior, the posterior of a parameter vector is
proportional to the likelihood function. Here, the likelihood function is multivariate normal,
so it will not be difficult to obtain the joint posterior of \( \phi_1, \ldots, \phi_q \). If the stock or index
in question does follow a random walk, or is close enough that for all practical purposes
we would consider this to be the case, then we would expect that for a "large" sample, the
posteriors of the \( \phi \)'s would be tightly wrapped around 0, meaning that, given the data, the
probability of the true \( \phi \)'s being much different from 0 is very small. This is in contrast to
a point hypothesis test like those which utilize VR and VD statistics. There, even if the
\( \phi \)'s are not much different from 0, the probability of generating a t-value greater than 8 can
easily be over 90% just by gathering a "very large" sample. However, if we gathered just
as large of a sample, but posted graphs of the posteriors of the \( \phi \)'s, instead of the t-values
greater than 8, then the reader would see graphs of the posteriors of the \( \phi \)'s that were
wrapped very tightly around values not economically significantly different from 0, which
would lead her to correctly think that the process is not economically significantly different
from a random walk. On the other hand, if she just saw t-stats of 8 and above, she might
think that the data very strongly rejects that the process is a random walk, and she might
think that the process is economically significantly different from a random walk, or at
the very least, the discerning reader would be stumped to determine the level of economic
significance, as economic significance cannot be determined from a point hypothesis test
without certain prior information or the ability to construct a confidence interval from the
data provided in the paper. Sometimes such data is there, but in the case of the VD and
VR tests it is not. And, I will discuss this in detail later.

By the way, the two sided p-value for a t-stat of 8 is \( 3.1 \times 10^{-16} \) ! And, we’ve all seen
papers that use very large time series that post t-values in the 30s, which, given that the
probabilities of a standard normal decrease exponentially, is completely off the chart. Even
a t-value of 4 has a two sided p-value of \( 1.5 \times 10^{-5} \).

So, in this paper, we could post graphs of the posteriors of all the \( \phi \)'s so the reader could
then look at these graphs to get a clear idea of economically how close the stock or index is to following a random walk. We also could post tables of confidence intervals, which are easy to generate once we have the posteriors. Such a table might look like the one below.

<table>
<thead>
<tr>
<th>95% Confidence Interval</th>
<th>99% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phi 1</td>
<td>-0.00000014</td>
</tr>
<tr>
<td>Phi 2</td>
<td>-0.00000085</td>
</tr>
<tr>
<td>Phi 3</td>
<td>-0.00000017</td>
</tr>
</tbody>
</table>

We could generate posteriors and confidence intervals for several models that we hope would be considered interesting. Of course, as I noted earlier, what we are doing is just estimating models of stock and index returns via Bayesian methods, which may have already been done for all of the interesting models, but if there were one or more interesting models where this hasn’t been done, posteriors for it’s parameters, and confidence intervals would be nice information to get out to the field, and perhaps would be publishable in a good journal. But, I would like to add more to the paper to increase its publishability. The next section includes material that is less likely to have been done already, because it involves fairly new or very new advances in econometrics and/or statistics.

First, however, I would like to talk more about confidence intervals. It is interesting that I cannot recall ever seeing a confidence interval table like table 1 above in a paper. I really can’t think of a good reason why right now. Instead, papers ubiquitously produce tables with t-statistics and/or p-values from point hypothesis tests. Now, in most cases, a (asymptotic estimate of a) confidence interval can be derived from the information in these tables, and such a confidence interval can be derived roughly in ones head. A specific common example is with OLS on the linear model. There, the asymptotic density of the estimator of \( \beta_k \) is \( N \left[ \beta_k, \text{var} \left( \hat{\beta}_k \right) \right] \), thus:

\[
\left( \frac{1}{\sqrt{\text{var} \left( \hat{\beta}_k \right)}} \right) \left( \hat{\beta}_k - \beta_k \right) \rightarrow N(0,1) \quad \text{as } T \rightarrow \infty
\]  

for any density of the error terms, with mild regularity conditions. This is a result of the central limit theorem, as \( \hat{\beta}^{OLS} \), like many estimators and statistics, is, in fact, a linear
function of a sum of independent random variables\textsuperscript{1}. This implies that as $T \to \infty$:

$$\text{Prob} \left[ -1.96 < \left( \frac{1}{\sqrt{\text{var} (\hat{\beta}_k)}} \right) (\hat{\beta}_k - \beta) < 1.96 \right] \to .05$$

$$\Rightarrow \text{Prob} \left[ -1.96 \sqrt{\text{var} (\hat{\beta}_k)} < \hat{\beta}_k - \beta_k < 1.96 \sqrt{\text{var} (\hat{\beta}_k)} \right] \to .05$$

$$\Rightarrow \text{Prob} \left[ \hat{\beta}_k - 1.96 \sqrt{\text{var} (\hat{\beta}_k)} < \beta_k < \hat{\beta}_k + 1.96 \sqrt{\text{var} (\hat{\beta}_k)} \right] \to .05 \quad (3)$$

Thus, line 3 gives us an \textit{asymptotic approximation} of a confidence interval, not the exact confidence interval given our actual finite sample size. Nonetheless, except for the case where we can feel comfortable assuming that the errors are distributed normally, the p-values are asymptotic approximations too. This is a drawback of much of frequentist statistics relative to Bayesian statistics. Often, because small sample densities are so hard to analytically derive in a nice closed form, asymptotic densities are substituted, while Bayesians are more precise in this regard. Their empirical posteriors are exactly the finite sample posteriors, not asymptotic approximations of them\textsuperscript{2}. As Kennedy (2003) notes:

If the errors are not distributed normally, the distribution of $\hat{\beta}_{\text{OLS}}$ is difficult to describe and to utilize for hypothesis testing. Instead of trying to derive the exact distribution of $\hat{\beta}_{\text{OLS}}$ in this circumstance, what is usually done is to approximate this exact distribution with what is called the asymptotic distribution of $\hat{\beta}_{\text{OLS}}$. (pg. 432)

And, I present again one of my favorite quotes, Wei and Tanner (1990):

The EM and MCEM algorithms provide a minimal amount of information to the data analyst. These algorithms yield the location of the normal approximation to the posterior distribution but so does standard MLE, so this is not an argument for standard MLE in preference to EM type algorithms...At the upper end are the data augmentation algorithm and the Gibbs sampler [Regarding the Gibbs sampler, see Geman and Geman (1984), Li (1988), and Gelfand and Smith (1990).] The latter two algorithms are iterative and can be shown to converge to the true posterior distribution under mild regularity conditions.\textsuperscript{3} (pg. 703).

\textsuperscript{1}An interesting side point is that any continuous function of an asymptotically normal random variable is also asymptotically normal. This is the basis of the Delta Method. See page 913, Greene, 5th ed.

\textsuperscript{2}This is true for all practical purposes if an "ample" number of draws is taken.
So, in a typical frequentist empirical paper we can’t derive exact finite sample confidence intervals, but we can derive asymptotic approximations of them. That may be the best we can do, and perhaps we might have good reason to believe the true DGP is close enough to normal and the sample is large enough that the asymptotic confidence intervals, and the p-values for that matter, are reasonably close to correct. A frequentist might also provide simulation based evidence that this is the case.

In a typical paper with tables of linear regression estimates and t-values, the author assumes that the central limit theorem approximation of normality is reasonably close. Let’s assume he is right. We can then construct confidence intervals from the information in these tables. First note the formula for the t-statistic under the usual null hypothesis; $H_0 : \beta_k = 0$:

$$t\text{-stat} = \frac{\hat{\beta}_k - 0}{\sqrt{\text{var}(\hat{\beta}_k)}}$$

This implies:

$$\sqrt{\text{var}(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{t\text{-stat}} \tag{4}$$

Now, if we substitute this expression for $\sqrt{\text{var}(\hat{\beta}_k)}$ into our confidence interval formula (3) we get:

$$\text{Prob} \left[ \beta_k - 1.96 \frac{\hat{\beta}_k}{t\text{-stat}} < \beta_k < \beta_k + 1.96 \frac{\hat{\beta}_k}{t\text{-stat}} \right] \rightarrow .05$$

which implies

$$\text{Prob} \left[ \hat{\beta}_k - \left( \frac{2}{t\text{-stat}} \right) \hat{\beta}_k < \beta_k < \hat{\beta}_k + \left( \frac{2}{t\text{-stat}} \right) \hat{\beta}_k \right] \rightarrow .0455$$

Thus a quick, in your head, estimate of the 95% confidence interval for $\beta_k$ is:

$$\beta_k = \hat{\beta}_k \pm \left( \frac{2}{t\text{-stat}} \right) \hat{\beta}_k \tag{5}$$

So if the t-stat is around positive 2, a very quick, in your head, approximation of the 95% confidence interval is just $\beta_k$ is between 0 and $2\hat{\beta}_k$. And, if the t-stat is around negative 2, a very quick, in your head, approximation of the 95% confidence interval is just $\beta_k$ is between 0 and $-2\hat{\beta}_k$.

Thus, perhaps one reason why I have never seen a confidence interval in a paper is because the t-stat tables give you everything you need to calculate the confidence intervals.
yourself, and you can just use a quick formula like formula (5) to approximate them in your head. But, this still does not seem like a good reason to not include confidence interval tables in any paper I have ever seen.

First, it would be nice to have the precise confidence Intervals, rather than in your head approximations, and it would be nice to just have it all right there without the reader having to do a bunch of calculations, and the CIs do seem like important info that in many cases the reader should know to help him determine economic and decision making significance.

Second, in any case, from my reading and asking around, it seems like few if any people, besides myself, try to…figure out these CIs in their head using quick formulas or approximations. The quick CI formula was something I derived myself my first year here when I first thought of this statistical significance problem. On the other hand, instead of approximating CIs, one could just basically look at the size of $\hat{\beta}_k$, see if it’s an economically significant size, think about the ramifications of its size, and then look at its t-stat to get an idea of the standard error around that estimate of the true $\beta_k$. Perhaps, but having CI tables would make the analysis more precise and less of the kind of back of the envelope ball park approximating in your head that should really be reserved for relatively unimportant things and not financial and economic research. And, some papers don’t report t-stats just p-values, so there’s an extra level of in your head approximation and calculation that must be done.

In addition, there is the issue of what percentage of readers understand well the issue of statistical versus economic significance, and what percentage of readers know how to take that information in the t-stat table and from it come up with good approximations of the level of economic significance and make good appropriate conclusions and advice. If the reader does not have a good grasp of the concept of economic significance versus statistical significance, and is not that clear on what is the economic significance, I think a confidence interval is a very good device for making this clear.

Consider, for example, a hypothetical paper which runs a regression of autoworkers salaries on sex, and assume the specification is perfect and all significant controls have been used. If the paper reports a t-stat of 5.2 for the sex coefficient and a value of $4.00 in annual salary for the estimate of the sex coefficient, then I suppose one would think that most readers would focus on the $4.00 and not just the t-stat of 5.2, and they’d in their head think that the density for the true sex coefficient is tightly wrapped around an insignificant amount of money. You’d think most readers would be at least that sophisticated, but
still if you were to post that the 99% confidence interval for the difference between male and female pay is between $1.50 in annual salary and $6.50 in annual salary, then it’s completely obvious to any reader, without a second thought, that as long as the study was done correctly, there is, for all practical purposes, no wage discrimination, in spite of the highly statistically significant t-stat. Plus, sometimes it may not be this clear cut; what if the t-stat is 1.98 and the sex coefficient estimate is $327. It seems like a confidence interval table would at least sometimes be worthwhile for making economic magnitude and significance more quickly and precisely clear.

And again, there is the question of what percentage of the readers of academic articles understand well the issue of statistical versus economic significance, and know how to take the information in the t-stat tables and process it into good approximations of economic magnitude and significance. When I first thought of this issue and asked Professor Oaxaca about it my first year, he told me that there was an economist by the name of Diedre McCloskey (formerly Donald) who came out (no pun intended) strongly against what she perceived as a substantial problem in economics of mistaking statistical significance for economic significance. So now, 4 1/2 years later, I looked up her research to see if she had any evidence as to what percentage of those in economics and finance aren’t good at evaluating economic significance. Her two major articles on this are, American Economic Review (1985) and Journal of Economic Literature (1996) with Stephen Ziliak.

The evidence she presents is at least somewhat vague and difficult to interpret, at least without time to evaluate it carefully, but in 1985 she claimed, "Roughly three quarters of the contributors to American Economic Review misuse the test of significance" (pg. 201) and "The most important question is whether the economists in the sample mix up statistical and substantive significance. Even on purely statistical grounds the news is not good: none of the papers mention the word ‘power’ though all mention ‘significance’.

By 1996, in her JEL paper, she (and her coauthor) still claims, "We would not assert that every economist misunderstands statistical significance, only that most do, and these some of the best economic scientists".

My big question is do papers present t-stat tables but not confidence intervals because they correctly assume that the vast majority of their readers will be savvy enough to construct their own confidence intervals from the t-table information, and otherwise to properly assess economic significance from the information in those tables without the author having to lengthen the paper by providing the reader explicitly with confidence intervals and the authors own analysis of economic significance. McKloskey and Ziliak address this issue on
page 99 when they criticize how most graduate economic texts say little or nothing about statistical versus economic significance:

One might defend contemporary usage by arguing that the advanced texts assume their readers already grasp the difference between economic and statistical significance. Economy of style would dictate the unqualified word ‘significance’, its exact meaning, economic or statistical, to be supplied by the sophisticated reader. Under such a hypothesis the contemporary usage would be no more than a shorthand way to refer to an estimated coefficient. The implied reader would be educated enough to supply the appropriate caveats about economic significance.

The hypothesis is not borne out by the evidence. The evidence, at least that which is provided in the two papers, as I said, is hard to evaluate, but doesn’t seem that strong. However, I think she also may be relying on experience and other more formal evidence that she has accumulated as an economist since the 60s, including as a professor at the University of Iowa between 1980 and 1996.

In any case, the bottom line for me is all considered I don’t currently see good justification for the fact that I cannot recall seeing a paper with a confidence interval table, or any notable presentation of confidence intervals. If you know of a good argument for why this is, I would be grateful if you would let me know. In the meantime, I am basically proposing a paper that clearly and explicitly analyzes the random walk hypothesis in terms of economic significance, that is, how economically significantly different is the DGP of popular assets and indices from a random walk. And, the tools I plan on using to make this hopefully clearer, more explicit, and more precise than most previous papers are confidence intervals and posteriors.

To help make the paper original enough for publication, I hope to generate posteriors for models which haven’t been estimated in a Bayesian way before (or perhaps for models which have been estimated before in a Bayesian way, but with different data), and in estimating confidence intervals I plan on using some original statistics, and some parametric and non-parametric techniques which have only recently been developed. An example is the interesting robust and efficient nonparametric confidence interval proposed by Romano and Wolf in their 2000 Annals of Statistics article, which I will discuss later.

I think probably the best way to determine the economic significance of how different the DGP of assets and indices is from a random walk is to test how much excess risk
adjusted return can be made from trading strategies based on past pricing information, both including and not including transactions and other costs. And, there have been a lot of studies like this, but I still think analyses with statistics, posteriors, and confidence intervals can be useful and complementary. They give a view of what’s going on from a different angle, and so can contribute to a richer and more valuable understanding. And, of course, it will be nice to work on a paper that allows me to learn a lot more econometrics, statistics, math, and mathematical computing, which hopefully I will then be able to apply fruitfully in future projects.

**How Does One Generate a Confidence Interval for VD(q) or VR(q)?**

Earlier I showed how the central limit theorem could be invoked to generate an asymptotic estimate of the confidence interval for $\beta_i$ in an iid linear regression model. However, in the setting of the VD(q) and VR(q) it is not possible to do a standard parametric estimation of a confidence interval. The information provided in the Lo and MacKinlay (1988) VD(q) and VR(q) test paper is not sufficient to construct any confidence intervals. Even for the savvy reader, it is not possible to ascertain a precise idea of the level of economic significance associated with their t-values. The best that can be expected is that those readers who are very savvy, and who have a great deal of prior knowledge regarding the random walk question can perhaps in a very loose or creative way ascertain some moderate idea of the economic significance behind Lo and MacKinlay’s tests, if that. This isn’t saying much. Test statistics that come with clear confidence intervals (or if for some reason we want to perhaps gratuitously make the reader do extra work, not the confidence intervals, but at least the ingredients so that readers can construct the confidence intervals themselves) give a much more precise understanding of the economic significance. And, the understanding is even more precise if a posterior can also be provided.

Now why is it not possible to construct confidence intervals with the information provided in Lo and MacKinlay’s (1988) VD(q) and VR(q) test paper? Let’s start with a brief review of the VD(2) test in their paper$^3$.

Lo and MacKinlay test the null, $H_0$: $r_t = \mu + \epsilon_t$, $\epsilon_t \sim iid \ N(0, \sigma^2)$, where $r_t$ is the continuously compounded return for some asset from time $t-1$ to time $t$. The alternative

---

$^3$A quick side comment: It would be so much nicer to be writing an idea for a paper that criticises and/or extends on Lo and MacKinlay 1988 in 1988, or 1989, rather than in 2004, when much or all of this may have already been done. That’s why I so want to specialize in an area or sub-area so I can become an expert and get to the fronteir, so then when I think of ideas they won’t already have been done. Still, some of the statistics I’m going to propose later is pretty new or very new, so perhaps it will allow me the opportunity to make some new and valued contributions.
hypothesis is the negation of the null. Under the null:

\[ \text{var} \{ r_t(2) \} = \text{var} \{ r_t + r_{t-1} \} = \text{var} \{ r_t \} + \text{var} \{ r_{t-1} \} = 2 \text{var} \{ r_t \} \]

\[ \Rightarrow V D (2) = \text{var} \{ r_t(2) \} - 2 \text{var} \{ r_t \} = 0 \quad (6) \]

So, under the null, \( V D (2) = 0 \). Of course, \( V D (2) \) cannot be observed, so we use an unbiased and consistent estimator for it:

\[ \widehat{V D} (2) \equiv \text{var} \{ r_t(2) \} - 2 \widehat{\text{var}} \{ r_t \}, \quad \text{or} \quad \widehat{V D} (2) \equiv 1/2 \widehat{\text{var}} \{ r_t(2) \} - \widehat{\text{var}} \{ r_t \} \quad (7) \]

where \( \text{var} \{ r_t(2) \} \) and \( \text{var} \{ r_t \} \) are the asymptotically efficient MLE estimates of \( \text{var} \{ r_t(2) \} \) and \( \text{var} \{ r_t \} \). They use an interesting insight from Hauseman (1978) to obtain the asymptotic density of \( \widehat{V D} (2) \) under the null:

\[ \sqrt{T} \widehat{V D} (2) \sim N \left[ 0, (\sigma^2)^2 \right] \quad (8) \]

So, what do we have here: we have a DGP, a parametric model, under the null hypothesis, and we have a set of DGPs that are compatible with the alternative hypothesis. As noted in the first section of this write up, what we don’t have is a DGP, a parametric model, that we’re assuming is the true parametric model that generated the data. The true parametric model is not identified. Therefore, we cannot find the true density, finite or asymptotic, for \( \widehat{V D} (2) \). (8) is not the asymptotic density for \( \widehat{V D} (2) \). It is the asymptotic density for \( \widehat{V D} (2) \) if the null hypothesis, the random walk hypothesis, is true. So, we can’t find a confidence interval for the true \( V D (2) \) with what is in the Lo and MacKinlay (1988) paper; we can only find the confidence interval under the null, but even that’s not really a confidence interval, because under the null it doesn’t matter what value we get for \( \widehat{V D} (2) \); the confidence interval for \( V D (2) \) (note that there’s no hat here) – given the null – is 100% that it’s between 0 and 0!

So, unlike the case with our \( \beta \)’s in the standard linear model example earlier with \( H_0 : \beta_k = 0, H_A : \beta_k \neq 0 \), we cannot construct confidence intervals to help make clear to us whether we are observing economic significance or just statistical significance, and if we are observing economic significance, how much. A key difference between the situation in this standard linear case and the situation in Lo and MacKinlay (1988) described above is that in the standard linear case, there is a parametric data generating model that is assumed

\[ ^4 \text{nor can we find the true density, finite or asymptotic for } \left( \widehat{V D} (2) - V D (2) \right) \text{ which is what we really need if we want to find confidence intervals for } V D (2). \]
to be true, period. It is true under the null, and it is true under the alternative. It is true in any state. You may not know the value of its parameters, but you do know it’s the true parametric form. Unlike in Lo and Mackinlay, the null and alternative only partition the state space with respect to the values of the parameters, not with respect to what the parametric form of the DGP is.

Because in Lo and MacKinlay the true parametric form is unidentified, we cannot use parametric statistics to find a confidence interval for the true $VD(2)$. Therefore, we cannot estimate the economic magnitude of $VD(2)$. We can ascertain no confidence interval for it. We can ascertain no density for it, finite sample or asymptotic. Thus, I’m not even sure how one could possibly gauge economic significance without doing simulations which aren’t in the paper. I suppose one way one could get some idea of economic significance is if one had essentially done their own simulations ex-ante, in other words if prior they had worked with data enough in this area and read and seen enough in this area that they could have some kind of idea of the magnitudes likely to go with the t-values. Many or most readers will not have this kind of prior knowledge and experience, and for those who do this is very rough, vague, and imprecise. I propose providing better information, which much more clearly and precisely indicates economic magnitude and significance. And, I propose a way to do this without even changing Lo and MacKinlay’s null and alternative hypotheses.

So, how will we generate a confidence interval, given the state space implied by Lo and MacKinlay’s null and alternative hypotheses, when this state space does not identify a parametric form for the data generating process? We will use nonparametric statistics.

**Nonparametric Statistics**

Some say that nonparametric statistics and econometrics makes no assumptions about the form of the density that generated the data, but that is not actually true. The term non-parametric is used in statistics and econometrics in a relative way. Techniques that are "very little parametric", that is techniques where a parametric form is specified, but that form is very unconstraining, very flexible, or has many, but not an infinite number of, parameters, are still classified as "nonparametric". An example is neural nets, where a very, or extremely, flexible functional form with a high number of parameters is used to fit the data.

And, in fact, we would rarely, if ever, want to make no assumptions about the data generating process, because that would be wasting our prior information. Either formally
or informally it is more efficient to use that prior information. For example, consider the empirical distribution I presented in my literature review, which I reprint below. This empirical distribution comes from a sample of 100 draws taken from a standard normal:

![Empirical Distribution Function](image_url)

*Figure 21.6 Empirical distribution function based on 100 observations*

Note that this empirical distribution (and this is true of any empirical distribution) places 0 probability mass in the tails beyond the most extreme observations. Certainly this would not be sensible, at least in most cases, where our sample was moderate to small, and it may also not be sensible to have a distribution function comprised of flat plateaus, so that, for example, the probability of a number less than 2.2 equals the probability of a number less than 2.0 equals the probability of number less than 1.8. This means that there is 0 probability of a number between 1.8 and 2.2. Certainly we can use our prior information to get a more efficient distribution (although for some problems it may not be worth spending the extra time, or it may be preferable to incorporate our prior information in an informal way).

The field of nonparametrics offers many techniques for incorporating assumptions based on prior information, for example the local histogram approach eliminates the flat spots
and the truncation of the tails in the empirical distribution., and, as I have noted, there are various "nonparametric" techniques that do fit functional forms to the data, albeit highly non-constraining functional forms. In addition to neural nets, there is a popular technique called kernel density estimation.

The specific nonparametric techniques I would now like to propose are nonparametric bootstraps. They involve taking draws from the empirical distribution or from the "smoothed", "fitted", or "slightly adjusted" empirical distribution. A smoothed empirical distribution is one where, as discussed above, the plateaus have been smoothed out and/or the tails have been un-truncated. A "fitted" empirical distribution is one where a highly flexible, little constraining, functional form has been fitted to the data, and a slightly adjusted empirical distribution is one where slight adjustments have been made to improve the performance of the bootstrap. Specifically what these slight adjustments are and why they make sense is quite involved and for the most part beyond the scope of this paper, although I'll touch on some of this later.

There are many advanced new bootstrapping techniques in the statistics and econometrics literature that we may be able to use to make the paper more unique, valuable, and publishable. And, I will describe some of these later, but first let's discuss the basic technique I am proposing and some of the intuition behind it.

Consider Lo and MacKinlay's (1988) $\bar{V}D(2)$ statistic:

$$\bar{V}D(2) \equiv \text{var} [r_t] - 1/2 \text{var} [r_t(2)]$$

This is equivalent to:

$$\frac{1}{T} \left\{ (r_1 - \bar{r})^2 + (r_2 - \bar{r})^2 + \ldots + (r_T - \bar{r})^2 \right\}$$

$$- \left( \frac{1}{1/2} \frac{1}{T} \right) \left\{ (r_1 + r_2 - 2\bar{r})^2 + (r_3 + r_4 - 2\bar{r})^2 + \ldots + (r_{T-1} + r_T - 2\bar{r})^2 \right\}$$

but because we have a time series which might have autocovariation, $r_t$ is really $(r_t \mid r_{-t})$

5Some interesting points here: First, $\text{var} [r_t(2)]$ can be estimated more efficiently by utilizing the overlapping intervals. Thus, a more efficient estimate is:

$$\bar{\text{var}} [r_t(2)] = \frac{1}{(1/2) T} \left\{ (r_1 + r_2 - 2\bar{r})^2 + (r_2 + r_3 - 2\bar{r})^2 + \ldots + (r_{T-1} + r_T - 2\bar{r})^2 \right\}$$

Second, maximum likelihood estimates, like the one above, can be biased, sometimes severely in small samples. And, in small samples, there may be alternative estimators that not only are unbiased, but have
even though we didn’t explicitly write it that way in (9) above. To make the exposition clearer, let’s restate (9) with the potential autocovariation made explicit:

\[
\hat{V} D (2) = \frac{1}{T} \left\{ (r_1 - \overline{r})^2 + (r_2 - \overline{r})^2 + \ldots + (r_T - \overline{r})^2 \right\}
\]

\[- \left( \frac{1}{2} \right) \frac{1}{(1/2) T} \left\{ [r_1 + (r_2 | r_1) - 2\overline{r}]^2 + [(r_3 | r_2, r_1) + (r_4 | r_3, r_2, r_1) - 2\overline{r}]^2 + \ldots + [(r_{T-1} | r_{-(T-1)}) + (r_T | r_{-T}) - 2\overline{r}]^2 \right\} \]

Next, let’s assume that the autocovariation goes for \(k\) lags and after that it has died out completely. We can make this assumption without loss of generality because we can make \(k\) any size we want, even infinity. Then, \(\hat{V} D (2)\) can be expressed as:

\[
\hat{V} D (2) = \frac{1}{T} \left\{ (r_1 - \overline{r})^2 + (r_2 - \overline{r})^2 + \ldots + (r_T - \overline{r})^2 \right\}
\]

\[- \left( \frac{1}{2} \right) \frac{1}{(1/2) T} \left\{ [r_1 + (r_2 | r_1) - 2\overline{r}]^2 + [(r_3 | r_2, r_1) + (r_4 | r_3, r_2, r_1) - 2\overline{r}]^2 + \ldots + [(r_{T-1} | r_{-(T-1)}) + (r_T | r_{-T}) - 2\overline{r}]^2 \right\} \] (10)

Now, given that \(\hat{V} D (2)\) equals expression (10) above, let’s think about what would happen if we did an ordinary bootstrap to empirically estimate the density of this statistic. In an ordinary bootstrap we would say that the sample \(r_1, \ldots, r_T\) comprises an empirical density for \(r\). To draw from this empirical density, we would just draw from \(r_1, \ldots, r_T\) with replacement. In this way we could generate \(B\) bootstrap samples, each of size \(T\). Let’s note these bootstrap samples \(\left( r_1^{(1)}, \ldots, r_T^{(1)} \right), \ldots, \left( r_1^{(B)}, \ldots, r_T^{(B)} \right)\). Next, we generate an empirical density for \(\hat{V} D (2)\) by calculating a \(\hat{V} D (2)\) statistic for each of the \(B\) bootstrap samples.

\[\text{lower mean squared error, average absolute deviation, and other desirable properties. Suppose, for example, that a maximum likelihood estimator of } \theta, \hat{\theta}^{MLE}, \text{ is unbiased. Then, because it has been proven that MLE estimators are invariant, the MLE estimator of } f(\theta) \text{ is } f(\hat{\theta}^{MLE}). \text{ Now, if } \hat{\theta}^{MLE} \text{ is unbiased, as we are assuming, then } E[\hat{\theta}^{MLE}] = \theta, \text{ but if } E[\hat{\theta}^{MLE}] = \theta \text{ and } f(\cdot) \text{ is a non-linear function, then } E[f(\hat{\theta}^{MLE})] \neq f(\theta) \text{ because of Jensen’s inequality, so } f(\hat{\theta}^{MLE}) \text{ is biased. Now, suppose } f(\cdot) \text{ is an extremely non-linear function, extremely curved, like the log, and the sample is small so that the variance of } \hat{\theta}^{MLE} \text{ is large, then the Jensen’s inequality effect could create extreme bias, making typical estimates of } \log \theta \text{ far off. Instead, estimates of } \log \theta \text{ that come from estimating } \log \theta \text{ directly rather than estimating } \theta \text{ directly, and then taking the log of that estimate, can have less bias and lower mean squared error and average absolute deviation.}

\text{Lo and MacKinlay (1988) use overlapping intervals and an adjustment to the MLE to make it unbiased that in a 1987 paper they showed increased performance in small sample simulations.}

\text{14}
This will give us \( B \) \( \hat{V}D(2) \) statistics which we will notate: \( \hat{V}D(2)^{(1)}, \ldots, \hat{V}D(2)^{(B)} \), and that is our empirical density for \( \hat{V}D(2) \).

This is incorrect. Why? because the correct formula for \( \hat{V}D(2) \) is formula (10) above, and in formula (10) the inputs are:

\[
r_1, (r_2 \mid r_1), (r_3 \mid r_2, r_1), \ldots, (r_{k+1} \mid r_k, \ldots, r_1), \ldots, (r_T \mid r_{T-1}, \ldots, r_{T-k})
\]

not

\[
r_1, \ldots, r_T
\]

where it is not implied in (12) that the \( r_t \) are generated conditionally on their predecessors, because when we just take draws independently from an empirical approximation of the unconditional density of \( r \), as we did in the ordinary bootstrap above, we just have a series of \( r \)'s drawn independently of whatever their predecessors were, so we don't have series (11), which is what we really need as input for the formula for \( \hat{V}D(2) \); instead we just have series (12), a series of independently drawn \( r \)'s from \( r \)'s unconditional density.

So, I asked myself, how do I draw \( B \) bootstrap samples like (11), and the answer I came up with is to draw blocks of \( r \)'s, rather than individual \( r \)'s. So, in other words, suppose I have a sample of size 10,000, and suppose I have prior information that gives me good reason to believe that after 8 lags any covariation is essentially gone. Then, I can divide my 10,000 observations into 1,250 observations of 8 consecutive \( r \)'s: \( (r_1, r_2, \ldots, r_8), (r_9, r_{10}, \ldots, r_{16}), \ldots, (r_{9993}, r_{9994}, \ldots, r_{10000}) \). Let's notate these blocks: \( B_1, \ldots, B_{1250} \). Note that each block preserves, and just actually came from, whatever covariation exists in the true DGP. They're all still in the original order that actually came from the true DGP.

So, in accordance with the nonparametric bootstrap idea (as I exposited it in my literature review), the sample of 1,250 blocks; \( B_1, \ldots, B_{1250} \), constitutes an empirical density for blocks of 8 consecutive observations of \( r \). Each of these 1,250 blocks of 8 observations is a little sample of 8 taken from the true DGP, with precisely all of the covariations that truly exist, so for each we can calculate a \( \hat{V}D(2) \). So, it's like we observed \( \hat{V}D(2) \) 1,250 times in repeated samples of size 8. This will give us a density for \( \hat{V}D(2)'s \) drawn from samples of size 8.

Now, what do we do with these blocks? Generally, when I read academic material, I tend to ask myself first how would I solve a certain problem, or handle a certain situation; the question just pretty much pops into my mind as soon as I read about the situation, and I
usually spend (more than a safe amount of) time, trying to come up with an answer, before looking to see what's already been thought of. Here, the covariation problem and block idea came to mind quickly. They are pretty clear, and blocking, in fact, is always done when dependence is assumed, but, my initial idea for constructing statistic densities and confidence intervals with these blocks is different from what I saw when I started looking in the literature, which puzzled me for a while. Then, I found out my idea is a subsampling idea as opposed to a resampling idea (resampling is another word for bootstrapping because you resample from the original sample, with replacement). In any case, time is very short, but there are a few points worth making.

First, I thought of a subsampling idea because I was worried about the "join point" problem in block resampling (I'll explain this problem later). Subsampling avoids this problem, but under certain assumptions, where the problem is not that serious, it is less efficient. In fact, subsampling is for the most part less efficient than resampling, but more robust. So, it might make a nice conservative check to back up our bootstrap results. In the words of subsampling expert Dimitris Politis, "...the understanding has been that subsampling sacrifices some accuracy for its extremely general applicability; this was also the viewpoint adopted in Politis, Romano, and Wolf (1999), Chapter 10." (Politis 2003, pg. 221).

Second, resampling and subsampling are very interesting areas. It's hard to tear myself away from this material, although, of course, I must – and soon, and they're also areas with a lot of subtleties, so one must be careful to apply these techniques correctly. As Davison and Hinkley (1997) put it, "Despite its scope and usefulness, resampling must be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a solution to the wrong problem, or a bad solution to the right one." (pg. 4).

Third, one often hears the jackknife mentioned along with the bootstrap. The jackknife is very similar, but it is always statistically no better than the bootstrap, and often it is inferior, or far inferior. Its one advantage is that it requires less, or far less, computing. That is why it is so much older than the bootstrap, having been credited to Quenneville in 1949. At that time, the jackknife was feasible in some situations, but to have the computing power to do the bootstrap, and at a cost that any researcher could easily afford, was like something out of Jules Verne. As Efron and Tibshirani (1993) put it, "...However by looking at only at the n jackknife samples, the jackknife uses only limited information about the statistic \( \bar{\theta} \), and thus one might guess that the jackknife is less efficient than the bootstrap. In fact it turns out that the jackknife can be viewed as an approximation to
the bootstrap...For nonlinear statistics there is a loss of information. The jackknife makes a linear approximation to the bootstrap...For highly nonlinear functions the jackknife can be inefficient, sometimes dangerously so" (pg.s 145-146). Given the available computing power, it is unlikely that the jackknife will be used in this project.

In any case, in this paper while I might want to do subsampling in addition to bootstrapping, I at least want to do bootstrapping, and since we are dealing with a potentially dependent time series, that would be block bootstrapping. So, let’s move forward with that. Remember our goal is to find confidence intervals for statistics of interest, the specific statistic we are focused on now is $VD(2)$. So, how do we find a confidence interval for $VD(2)$ via nonparametric block bootstrapping? There are many ways, some better and newer than others, but I think it will be instructive and useful to first go through one of the most basic, the percentile method.

The Percentile Method

In the percentile method, whether using blocks or individual observations, the empirical distribution is treated as if it were the true distribution. The justification is that the empirical distribution – the sample – will tend to be like the distribution it was drawn from, and certainly this is true in the limit. With a large enough number of draws, the sample will almost surely, for all practical purposes, mimic the distribution it came from. For example, regarding the graph on page 12, with a large enough number of draws, the empirical distribution taken from standard normal drawing, will, for all practical purposes, almost surely, be indistinguishable from the actual standard normal distribution.

So now let’s look at our example of interest. We observe a time series of, say, daily returns from some stock or index, $r_1,...r_T$. That is our empirical distribution for $r$. We act as though that empirical distribution is the true distribution. Thus, for any statistics of the empirical distribution, we act as though they are the corresponding statistics from the true distribution. For example:

The empirical distribution = The sample $\equiv \hat{F}(r) \equiv (r_1,...r_T)$

and we act as though this equals The true distribution $= F(r)$
The empirical distribution’s mean = The sample’s mean = The mean of \( \hat{F}(r) = \bar{r} \)

\[
\text{and we act as though this equals} \quad \text{The mean of } F(r) = E(r)
\]

\[
\vdots
\]

The empirical distribution’s \( VD(2) \) = The sample’s \( VD(2) \) = The \( VD(2) \) of \( \hat{F}(r) \)

\[
= \hat{V}D(2) = 1/2\bar{r}^2 [r_1(2)] - \bar{r}^2 [r_1]
\]

\[
= \frac{1}{T} \left\{ (r_1 - \bar{r})^2 + (r_2 - \bar{r})^2 + \ldots + (r_T - \bar{r})^2 \right\}
\]

\[
- (1/2) \frac{1}{(1/2)T} \left\{ \left[ r_1 + (r_2 | r_1) - 2\bar{r} \right]^2 + \left[ (r_3 | r_2, r_1) + (r_4 | r_3, r_2, r_1) - 2\bar{r} \right]^2 + \ldots + \left[ (r_{T-1} | r_{T-2}, \ldots, r_{T-k}) + (r_T | r_{T-1}, \ldots, r_{T-k}) - 2\bar{r} \right]^2 \right\}
\]

\[
\equiv VD(2)^{EMP}
\]

\[
\text{and we act as though this equals} \quad \text{The } VD(2) \text{ of } F(r)
\]

I introduced the notation \( VD(2)^{EMP} \) in (13) because I think it may avoid confusion later. Given the explicitness of the superscript, it’s clear that this is the \( VD(2) \) of the empirical distribution.

Now, we want to utilize our empirical distribution (which again, remember, if the sample is large, will be a close and consistent approximation to the true distribution) to find a confidence interval for \( VD(2) \). This is how we will do it.

Assuming the empirical distribution is true, then the true \( VD(2) \) is \( VD(2)^{EMP} \). We can then ask, under the empirical distribution, what is the density of \( [\hat{V}D(2) - VD(2)^{EMP}] \) where \( \hat{V}D(2) \) is defined as in equation (10) on page 14. We generate this density by first drawing \( B \) samples of \( r \), each of size \( T \) from our empirical distribution: \( (r_1^{(1)}, \ldots, r_T^{(1)}), \ldots, (r_1^{(B)}, \ldots, r_T^{(B)}) \). Next, for each of the \( B \) samples of \( r \) we calculate \( \hat{V}D(2) \), and that will give us: \( [\hat{V}D(2)^{(1)} - VD(2)^{EMP}], \ldots, [\hat{V}D(2)^{(B)} - VD(2)^{EMP}] \), which is our empirical density for \( [\hat{V}D(2) - VD(2)^{EMP}] \).
Once we have this density, we can then generate confidence intervals because, for example, this density will tell us for what \( k^{lo} \) and \( k^{up} \):

\[
\text{Prob}\left[ k^{lo} \leq \left( \widehat{V D} (2) - V D (2)^{EMP} \right) \leq k^{up} \right] = .95
\]  

just by letting \( k^{lo} \) be the 5th percentile observation in our density: \( [\widehat{V D} (2)^{(1)} - V D (2)^{EMP}] \), \( ... [\widehat{V D} (2)^{(B)} - V D (2)^{EMP}] \), and letting \( k^{up} \) be the 95th percentile observation. That is why this method is called the percentile method. Now, let’s think about what equation (14) tells us. If our empirical density is accurate, it gives us a good approximation of how much the \( VD (2) \) estimator, \( \widehat{V D} (2) \), is likely differ from the true \( VD (2) \) when the sample is of size \( T \). It tells us that there is a 95% chance that \( \widehat{V D} (2) \) will be in error by more than \( k^{lo} \) units, but by less than \( k^{up} \) units. Because the empirical density is a consistent approximation to the true density, confidence interval (14) will be a consistent approximation to the true confidence interval, and in fact, there’s been a great deal of research in the statistics field to show that this method usually gives better approximations of confidence intervals than the standard asymptotic normal method.

From (14) we can do some algebra to get:

\[
\text{Prob}\left[ \widehat{V D} (2) - k^{up} \leq V D (2)^{EMP} \leq \widehat{V D} (2) - k^{lo} \right] = .95
\]

And this tells us that if we use the \( VD (2) \) estimator, \( \widehat{V D} (2) \), with a sample of size \( T \), then there is a 95% chance that the true \( VD (2) \) will be in the interval \( [\widehat{V D} (2) - k^{up}, \widehat{V D} (2) - k^{lo}] \).

And, we then just insert the actual \( \widehat{V D} (2) \) that we actually got from our sample (which is what we had been calling \( VD (2)^{EMP} \). It’s a little confusing. If this write up were for students, I’d put in a clear simple example with numbers that I think would make everything clearer) into this interval.

And along the same lines, we could generate a complete empirical density for \( VD (2) \), and display a nice graph, giving a good idea of how likely the true \( VD (2) \) is to be in the neighborhood of 0.

Now, there are still some issues. First, we have to take \( B \) bootstrap samples of \( r \), but we cannot just independently draw \( r \)’s with replacement from the sample because that will destroy any autocovariations that may have existed, so as we discussed earlier we must take our draws in blocks of size \( k \), and there are various techniques in the literature
for determining what $k$ should be. The blocks can be overlapping or non-overlapping depending on the technique. After we have drawn $T/k$ blocks, we then connect them all together to form a bootstrap sample of size $T$, and we repeat this process until we have $B$ bootstrap samples.

This brings us to the join point problem. The observations near, or at, the join will have unnaturally weakened dependence, thus any autocovariations will tend to be understated. There are two responses to this. First, in some sense, in the limit, it’s not a problem, because as $T \to \infty$, we can also let $k \to \infty$. So, there is, in some sense, still consistency. And, second, the field of statistics has developed techniques to alleviate the anti-dependence bias created by the join point problem. These techniques include observation weighting, post blackening, and blocks of blocks.

The most promising method I have seen, and one I propose using, is the Block-Block bootstrap, recently introduced by Yale econometrician Donald Andrews in the May 2004 issue of Econometrica. According to Andrews, "The asymptotic refinements of the block-block bootstrap are shown to be greater than those obtained with the block bootstrap and close to those obtained with the nonparametric iid and parametric bootstrap." (pg. 673), and with regard to small sample performance, "In sum, the Monte Carlo results show that all of the bootstrap CI’s considered outperform the delta method CI’s by a substantial margin. The results also show that the block-block bootstrap yields improved coverage probabilities in the cases considered compared to the standard block bootstrap." (pg. 697)

So, in summary, I propose using the Block-Block bootstrap to generate empirical densities and confidence intervals for statistics in the literature like $VD(q)$, and possibly for my own statistics, to give a clearer idea of how economically close (or far) asset return processes are to (or from) various definitions of a random walk. I propose to use this technique to make economic significance more clear and precise, especially in cases like Lo and MacKinlay’s 1988 $VD(q)$ paper, where it is impossible to construct even possibly highly inaccurate asymptotic confidence intervals with the information provided in the paper.


This is a very interesting and possibly important paper that essentially talks about constructing highly robust and efficient nonparametric confidence intervals. How is robust defined? There is some underlying true DGP, $F(X)$. If we knew $F(X)$, and if it was tractable, then we could just use $F(X)$ directly to construct confidence intervals for parameters and statistics, but in a setting where we decide to use nonparametrics, it is because,
given our prior information, we feel more comfortable not assuming a specific DGP form like a normal or a $t$ or gamma.

A robust 95% confidence interval, for example, is one where there is, at least, a 95% chance that the true value of the statistic lies somewhere in that interval, no matter whether the true DGP is a normal, a $t$, a gamma, or any other distribution within some large set of distributions $\mathbf{F}$. Of course, it’s easy to create a completely robust confidence interval, just select $(-\infty, \infty)$. There’s at least a 95% chance that the statistic is in that interval no matter what the true $F$ is, but that’s where efficiency comes in. Efficiency in this context means that the interval is not that wide, or not that much wider than a standard asymptotic normal confidence interval.

Romano and Wolf (2000) is a very advanced paper, and I have not spent much time on it, so I really can’t say much more about it now, except that as I currently understand it, for some statistics they are able to construct finite sample confidence intervals that are extremely robust and yet are only negligibly wider than the far less robust standard asymptotic normal confidence interval.

The paper looks like it may be a very important advance. First, highly robust and efficient confidence intervals can be very valuable in ascertaining economic significance and in economic and financial analysis and decision making in general. For example, consider the case of long period asset returns, say 6 months, 1, or 5 years, $r^5_t$, $r^1_t$, or $r^5_t$. Here we will have very few observations of $r^5_t$, $r^1_t$, or $r^5_t$, making the use of asymptotics, asymptotically normal p-values, and confidence intervals, terrible, even powerless to reject in a one sided test no matter what the underlying DGP is, as shown in an example in Campbell, Lo, and MacKinlay (1997, pg. 58).

In such cases, if we can still get robust confidence intervals around correlation coefficients that aren’t too wide, this would be important information regarding market efficiency or market behavior, and it would be extremely credibly relative to standard asymptotic normal approximation. There are potentially a lot of valuable uses for highly robust, yet not too wide confidence intervals, as in many cases standard asymptotic confidence intervals relying on the central limit theorem are terrible, or nonsense, in ordinary finite samples. This has been shown in simulations in Romano and Wolf (2000) and in many other papers. In finite samples, the actual density may be very different from it’s asymptotic normal approximation.
Why else might this paper be an important new advance? Again, it’s just impossible right now for me to spend the time to really understand this paper well. That would probably require learning some new statistics. But, in the authors’ words:

A variety of conservative methods are presented in Bickel (1992) including some justification for Stringer’s (1963) widely used proposal. However, none of the methods presented there are both conservative in level and efficient. We believe our proposal is the first one that is both nonparametric (or conservative) and efficient. (pg. 758)

Finally, the authors are very important in the fields of nonparametrics, resampling and subsampling. Romano is a full professor at Stanford.

The authors talked about expanding their work, and in 2002 came out with an extension, "Explicit Nonparametric Confidence Intervals for the Variance with Guaranteed Coverage". Again, I haven’t been able to go over this paper thoroughly, but just to give an idea, here is a quote:

In this paper we provide a method for constructing confidence intervals for the variance which exhibits guaranteed coverage probability for any sample size, uniformly over a wide class of probability distributions. In contrast, standard methods achieve guaranteed coverage only in the limit for a fixed distribution or for any sample size over a very restrictive (parametric) class of probability distributions. (pg. 1231)

References


