Monte Carlo Tests

A random variable $\tau = \tau(\boldsymbol{y}, \theta)$ is **pivotal** if the distribution of $\tau(\boldsymbol{y}, \theta_0)$ is the same for every DGP in $\mathbb{M}$ with $\theta = \theta_0$.

In particular, the CDF $F(\tau)$ does not depend on any nuisance parameters. It may only depend on things we observe, like N and $\boldsymbol{X}$.

If $F(\tau)$ did vary with the DGP in finite samples but not asymptotically, then τ would be **asymptotically pivotal**.

In the classical normal linear model, t and F statistics are pivotal.

- If a test statistic is pivotal, we can perform an exact test, or construct an exact confidence interval, by simulation.
- We simply need to generate B simulated test statistics τ_b^* from some DGP in $\mathbb{M}$ with $\theta = \theta_0$.
- It is essential to choose B so that $\alpha(B + 1)$ is an integer, where α is the level of the test; see below.

The τ_b^* are used to calculate a **Monte Carlo *P* value** for τ.

A **bootstrap *P* value** (below) is computed just like a Monte Carlo P value, but since τ is not pivotal the test is not exact.

The EDF of the τ_b^* is given by

$$\hat{F}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq x). \tag{1}$$

If a test rejects in the upper tail, the **Monte Carlo *P* value**, or **simulated bootstrap *P* value**, is

$$\hat{p}^*(\tau) = 1 - \hat{F}^*(\tau) = 1 - \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq \tau) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \tau). \tag{2}$$

In principle, we could let $B \to \infty$, so that $\hat{p}^*(\tau) \to p^*(\tau)$, the **ideal bootstrap *P* value**.

Like every P value, $\hat{p}^*(\tau)$ must lie between 0 and 1.

For example, if $B = 999$, and 36 of the τ_b^* were greater than τ, we would have $\hat{p}^*(\tau) = 36/999 \cong .036$.

This procedure yields an exact test for pivotal test statistics even for finite values of B, provided B is chosen so that $\alpha(B+1)$ is an integer.

- If $\alpha = .05$, values of B that satisfy this condition are 19, 39, 59, and so on. If $\alpha = .01$, they are 99, 199, 299, and so on.
- That is why $B = 999$ in the above example. 999 works for all interesting values of α, including 0.001, 0.01, 0.025, 0.05, and 010.

Suppose we sort the original test statistic τ and the B bootstrap statistics τ_b^*, $b = 1, \ldots, B$, from largest to smallest. Since τ is pivotal, these are independent draws from the same distribution.

There are exactly R simulations for which $\tau_b^* > \tau$. Thus, if $R = 0$, τ is the largest value in the set, and if $R = B$, it is the smallest.

- The estimated P value $\hat{p}^*(\tau)$ is just R/B.
- The bootstrap test rejects if $R/B < \alpha$, that is, if $R < \alpha B$.

Let $[\alpha B]$ be the largest integer smaller than αB.

There are $[\alpha B] + 1$ such values of R, namely, $0, 1, \ldots, [\alpha B]$. Thus the probability of rejection is $\big([\alpha B] + 1\big)/(B + 1)$.

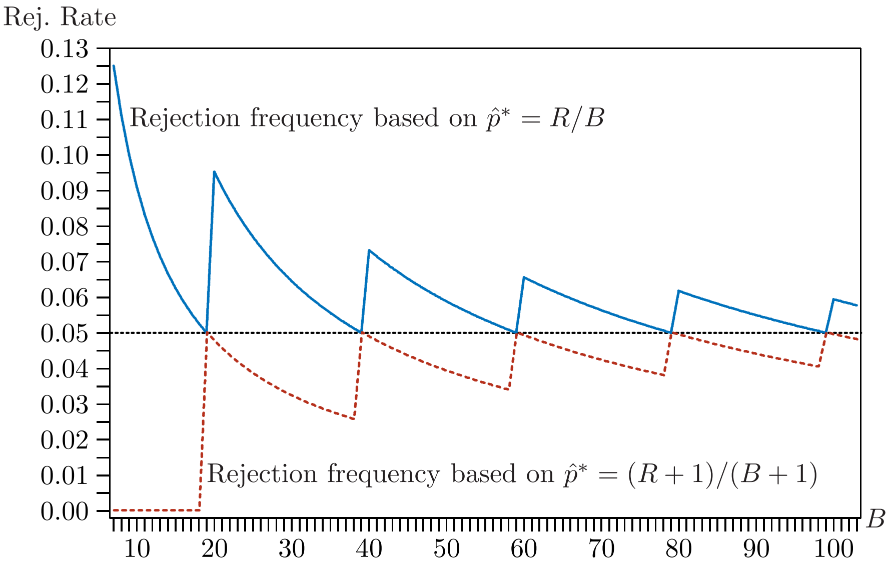
If we equate this probability to α and multiply by $B + 1$, we find that

$$\alpha(B + 1) = [\alpha B] + 1. \tag{3}$$

Equation (3) holds if and only if $\alpha(B + 1)$ is an integer. Therefore, the Type I error is precisely α if and only if that is the case.

- Let $B = 99$ and $\alpha = .05$. Then $\hat{p}^*(\tau) < .05$ whenever τ is in positions 1, 2, 3, 4, or 5. This occurs with probability $5/100 = .05$.
- When (3) does not hold, Monte Carlo tests will over-reject or under-reject in a manner that is $O(1/B)$ and depends on B.

The figure shows rejection frequencies for two types of Monte Carlo test. One rejects when $R < \alpha B$, and one rejects when $R + 1 \leq \alpha(B + 1)$. The latter is more conservative unless $\alpha(B + 1)$ is an integer.


Rej. Rate
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Rejection frequency based on $\hat{p}^* = R/B$
Rejection frequency based on $\hat{p}^* = (R+1)/(B+1)$
B
10
20
30
40
50
60
70
80
90
100

Monte Carlo Tests for Skewness and Kurtosis

For the normal distribution, the third moment of the disturbances is 0, and the fourth moment is $3\sigma^4$.

Define the normalized residuals e_i as $\hat{u}_i/\hat{\sigma}$, where $\hat{\sigma} = \sqrt{\text{SSR}/N}$. The sum of the e_i^2 is precisely N.

We can test for skewness using the test statistic

$$\tau_{\text{sk}} = \frac{1}{\sqrt{6N}} \sum_{i=1}^{N} e_i^3. \tag{4}$$

We can test for excess kurtosis using the test statistic

$$\tau_{\text{ku}} = \frac{1}{\sqrt{24N}} \sum_{i=1}^{N} (e_i^4 - 3). \tag{5}$$

Both τ_{sk} and τ_{ku} are asymptotically distributed as $\text{N}(0,1)$. But skewed!

Because τ_{sk} and τ_{ku} are asymptotically independent, we can test both hypotheses jointly using the test statistic

$$\tau_{\text{skku}} = \tau_{\text{sk}}^2 + \tau_{\text{ku}}^2, \tag{6}$$

which is asymptotically distributed as $\chi^2(2)$.

All these test statistics are pivotal. If $\boldsymbol{\epsilon} \equiv \boldsymbol{u}/\sigma$, they depend on $\boldsymbol{y}$ solely through the vector

$$\boldsymbol{e} \equiv (\boldsymbol{u}^\top \boldsymbol{M_X u}/N)^{-1/2} \boldsymbol{M_X u} = (\boldsymbol{\epsilon}^\top \boldsymbol{M_X \epsilon}/N)^{-1/2} \boldsymbol{M_X \epsilon}. \tag{7}$$

Under classical assumptions, $\boldsymbol{\epsilon}$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

For Monte Carlo tests, generate BN standard normal random variates and form them into N-vectors $\boldsymbol{\epsilon}^b$ for $b = 1, \ldots, B$.

Then regress the $\boldsymbol{\epsilon}^b$ on $\boldsymbol{X}$, compute normalized residuals $\boldsymbol{e}^b$, and calculate test statistics using (4), (5), or both of them plus (6).

These tests are asymptotically valid if regressors are not exogenous.

Bootstrap Tests

We have seen how to perform a bootstrap test for $\theta = \theta_0$ based on bootstrap standard error $\text{se}^*(\hat{\theta})$ and assumption that $\hat{\theta} \overset{a}{\sim} \text{N}\big(0, \text{Var}(\hat{\theta})\big)$.

Another (often better) approach is like Monte Carlo testing. Compare test statistic τ with the distribution of B bootstrap test stats τ_b^*.

This sort of bootstrap test differs somewhat from Monte Carlo tests.

- Monte Carlo test statistics are pivotal.
- Bootstrap test statistics may or may not be asymptotically pivotal.
- Monte Carlo tests are exact, provided $\alpha(B + 1)$ is an integer.
- Bootstrap tests are almost never exact in finite samples.
- Bootstrap tests based on asymptotically pivotal test statistics may provide **asymptotic refinements**.

Both simulation results and higher-order theory suggest that this sort of bootstrap test should work well in certain circumstances.

We may hope that bootstrap tests will work well whenever:

1. The test statistic τ is close to being pivotal.
2. The bootstrap DGP does a good job of mimicking the true DGP under the null hypothesis. This matters more if #1 does not hold.
3. The parameters of the bootstrap DGP are estimated under the null hypothesis. This helps make #2 hold.
4. The distribution of the bootstrap statistics τ_b^* is (almost) independent of τ. This is critical and often overlooked.

There are two ways to perform a bootstrap test:

- Compute a bootstrap P value.
- Compute a bootstrap critical value, say c_α^*, and check whether τ is more extreme than c_α^*.

When $\alpha(B+1)$ is an integer, both methods yield identical inferences.

Bootstrap P values are more informative unless τ is enormous.

There are three main ways to compute bootstrap P values:

1. One-sided (upper tail) P value:

$$\hat{p}^*(\tau) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \tau). \tag{8}$$

Use this for test statistics that are asymptotically χ^2 or F.

We also want to use (8) for one-sided t tests against an alternative in the upper tail.

2. Symmetric P value:

$$\hat{p}^*(\tau) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\tau_b^*| > |\tau|). \tag{9}$$

Use this for two-sided t tests when we believe that $F(\tau)$ is roughly symmetric around zero.

3. Equal-tail P value:

$$\hat{p}^*(\tau) = \frac{2}{B} \min\left(\sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq \tau),\ \sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \tau) \right) \tag{10}$$

Use this for two-sided t tests when we believe that $F(\tau)$ is not symmetric around zero. Note the factor of 2!

Equal-tail and symmetric P values can differ greatly when τ is a t stat based on a biased parameter estimate.

Perhaps use **bootstrap bias correction** (MacKinnon and Smith, 1998).

4. Bootstrap critical values:

If we sort the τ_b^* from smallest to largest, the bootstrap critical value c_α^* is simply number $(1-\alpha)(B+1)$.

For example, when $\alpha = .05$ and $B = 999$, c_α^* is number 950.

Rejecting when $\tau > c_\alpha^*$ is equivalent to rejecting when the one-sided P value (8) is less than α.

Be careful if the bootstrap DGP does not impose the null hypothesis!

Consider the bootstrap t statistic for testing $\theta = \theta_0$:

$$t_b^* = \frac{\hat{\theta}_b^* - \theta_0}{\text{s.e.}(\hat{\theta}_b^*)}. \tag{11}$$

When the bootstrap DGP imposes the null, we would expect $F(\hat{\theta}_b^*)$ to be centered near $\mathrm{E}(\hat{\theta} \,|\, \theta = \theta_0)$.

But if the bootstrap DGP does not impose the null, it is going to be centered near $\mathrm{E}(\hat{\theta} \,|\, \theta = \hat{\theta})$.

In this case, we have to replace (11) by

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\text{s.e.}(\hat{\theta}_b^*)}. \tag{12}$$

If not, the bootstrap test will have no useful power.

Bootstrap Confidence Intervals

Inverting a bootstrap test yields a **bootstrap confidence interval**, or **bootstrap CI**.

Ideally, we invert a bootstrap test based on a restricted bootstrap DGP to obtain a **restricted bootstrap confidence interval**.

- Doing this requires an iterative procedure. We need to find two values of θ, say θ_l^* and θ_u^*.
- The equal-tail bootstrap P value for each of them must equal α, or the appropriate one-tail P value must equal $\alpha/2$.
- For each candidate value of, say, θ_u, we generate B bootstrap samples under the null hypothesis that $\theta = \theta_u$ and compute (10).
- If $P^*(\theta_u) < \alpha$, then θ_u is too large. If $P^*(\theta_u) > \alpha$, it is too small.

We need to use a root-finding algorithm such as **bisection** that does not use derivatives to find approximate value of θ_u.

Bisection Algorithm:

Define $f(\theta)$ as $\hat{p}^*(\theta) - \alpha$, where $\hat{p}^*(\theta)$ denotes the equal-tail P value evaluated at θ. We want to find a value θ_u^* for which $f(\theta_u^*) = 0$.

- To start the process, we need two values of θ, say θ_a and θ_b, with the properties that

$$f(\theta_a) > 0 \quad \text{and} \quad f(\theta_b) < 0. \tag{13}$$

  Since $f(\cdot)$ is non-increasing, it must be the case that $\theta_a < \theta_b$.
- At each step, the bisection method finds a new value $\theta_c = (\theta_a + \theta_b)/2$ and computes $f(\theta_c)$.
- Then θ_c replaces whichever of the previous values has $f(\theta)$ with the same sign as $f(\theta_c)$. New interval is half the length of old one.
- Eventually, when θ_a and θ_b are sufficiently close, the algorithm terminates, and the final value of θ_c becomes θ_u^*.

The **grid bootstrap** of Hansen (1999) is another way to obtain restricted bootstrap confidence intervals.

Because $\hat{p}^*(\theta)$ is based on a finite value of B, such as 999, it cannot be a smooth function of θ_u. It is a step function.

- There will typically exist no value θ_u^* for which $\hat{p}^*(\theta_u^*) = \alpha$.
- Instead, θ_u^* will be the value where $\hat{p}^*(\theta) < \alpha$ for $\theta > \theta_u^*$ and $\hat{p}^*(\theta) > \alpha$ for $\theta < \theta_u^*$.

It is essential to use the same seed (and thus the same sequence of random numbers) every time we calculate a bootstrap P value.

This applies to many simulation-based estimators.

Otherwise, $\hat{p}^*(\theta)$ would take on different values each time it was computed for the same value of θ, and the root-finding algorithm would never converge.

Procedure for finding θ_l^* is very similar to procedure for finding θ_u^*.

- Now define $f(\theta)$ as $\alpha - \hat{p}^*(\theta)$.
- If $\hat{p}^*(\theta) < \alpha$, then θ is too small. If $\hat{p}^*(\theta) > \alpha$, then θ is too large.
- Use bisection to find θ_l^*, exactly as before.

Studentized Bootstrap Confidence Intervals

When a test statistic is pivotal, we can calculate just one set of τ_b^*, for $b = 1, \ldots, B$ and use them to compute every bootstrap P value.

This will yield an exact confidence interval.

When τ is approximately pivotal, we can do the same thing, and with luck the interval will be reasonably accurate.

- For a **studentized bootstrap confidence interval**, the test statistic $\tau(\boldsymbol{y}, \theta)$ is the t statistic $(\hat{\theta} - \theta)/s_\theta$.
- Dividing an estimate by its standard error, in this case s_θ, to form a t statistic is often called **studentization**.
- For a linear regression model, s_θ could be a classical standard error, a heteroskedasticity-robust standard error, or a cluster-robust standard error.

These intervals are also called **percentile-*t*** confidence intervals or **bootstrap-*t*** confidence intervals.

Studentized bootstrap confidence intervals are widely used. They should work well if two assumptions hold:

1. The distribution of $\tau(\theta, \boldsymbol{y})$ does not depend very strongly on how $\boldsymbol{y}$ is generated.
2. The standard error of $\hat{\theta}$, s_θ, is reasonably accurate and not very correlated with $\hat{\theta}$.

Assumption #1 says that the t statistic is pivotal to a reasonably good approximation.

Assumption #2 is very important, because s_θ plays the same role in a studentized bootstrap CI as it does in a conventional CI based on the $t(N-k)$ distribution.

If either part of #2 fails, the interval may have poor coverage.

The procedure for constructing a studentized bootstrap confidence interval is quite easy.

Use any bootstrap DGP that does not impose a null hypothesis.

1. Calculate $\hat{\theta}$ and its standard error s_θ, along with anything needed for an unrestricted bootstrap DGP.
2. Generate B bootstrap samples $\boldsymbol{y}_b^*$, $b = 1, \ldots, B$, based on unrestricted estimates. Choose B so that $(\alpha/2)(B+1)$ is an integer.
3. For each bootstrap sample, compute $\hat{\theta}_b^*$ and its standard error s_b^*. Then use these to compute the bootstrap t statistic

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{s_b^*}. \tag{14}$$

4. Sort the t_b^* from smallest to largest. Let $c_{\alpha/2}^*$ denote number $(\alpha/2)(B+1)$, and let $c_{1-\alpha/2}^*$ denote number $(1-\alpha/2)(B+1)$.
5. Construct the studentized bootstrap confidence interval

$$\left[\hat{\theta} - s_\theta c_{1-\alpha/2}^*,\ \ \hat{\theta} - s_\theta c_{\alpha/2}^*\right]. \tag{15}$$

Notice that the upper-tail (lower-tail) quantile determines the lower (upper) limit of the interval.

The studentized bootstrap CI (15) looks very much like a conventional CI based on the $t(N-k)$ distribution.

- Bootstrap critical values are used instead of critical values from $t(N-k)$, which causes the interval to be asymmetric.
- When $\hat{\theta}$ is biased, the interval will generally not be centered at $\hat{\theta}$. In effect, it performs a sort of **bias correction**.

When we are interested in $\gamma = g(\theta)$, there are two obvious ways to obtain studentized bootstrap confidence intervals.

1. Construct a studentized bootstrap interval for γ, using the delta method to obtain s_γ. The result would be

$$\left[\hat{\gamma} - s_\gamma c^{\gamma *}_{1-\alpha/2},\ \hat{\gamma} - s_\gamma c^{\gamma *}_{\alpha/2}\right], \tag{16}$$

where $c^{\gamma *}_{\alpha/2}$ and $c^{\gamma *}_{1-\alpha/2}$ are the entries indexed by $(\alpha/2)(B+1)$ and $(1-\alpha/2)(B+1)$ in the sorted list of bootstrap t statistics for the hypothesis that $\gamma = g(\hat{\theta})$.

2. Transform both limits of the studentized bootstrap interval (15). If we did that, we would obtain the confidence interval

$$\big[g(\hat\theta - s_\theta c^*_{1-\alpha/2}),\ \ g(\hat\theta - s_\theta c^*_{\alpha/2})\big], \tag{17}$$

where $c^*_{\alpha/2}$ and $c^*_{1-\alpha/2}$ are the appropriate entries in the sorted list of bootstrap t statistics for the hypothesis that $\theta = \hat\theta$.

The intervals (16) and (17) will be different, perhaps quite different if the function $g(\cdot)$ is highly nonlinear in the neighborhood of $\hat\theta$.

There are many other ways to construct bootstrap confidence intervals.

If s_θ is not available, we could use the bootstrap to estimate it and then construct a studentized bootstrap CI. But this would involve a **double bootstrap**, with $B \times B_2$ bootstrap samples.

Theory suggests that methods based directly on $\hat\theta$ and the $\hat\theta^*_b$, i.e. not based on asymptotically pivotal test statistics, should be avoided.

However, this advice may be wrong if $\text{s.e.}(\hat\theta)$ is a poor estimator.

Power Loss from Bootstrapping

A bootstrap test may reject more or less often than the corresponding asymptotic test; see Davidson and MacKinnon (2006).

Generally, bootstrap tests appear to have less power than the corresponding asymptotic test, but only because the latter over-rejects.

- If an asymptotic test under-rejects, the corresponding bootstrap test will probably have more power.
- A bootstrap test based on finite B must reject less often than one based on $B = \infty$, athough the power loss is often negligible.
- When B is finite, $\hat{p}^*$ differs from p^* because of random variation in the bootstrap samples.
- Adding randomness to p^* is equivalent to adding randomness to τ. In both cases, this reduces test power.

The power loss due to B being finite is $O(1/B)$; see Davidson and MacKinnon (2000).

Consider z_{β_2} and t_{β_2} for the classical normal linear model.

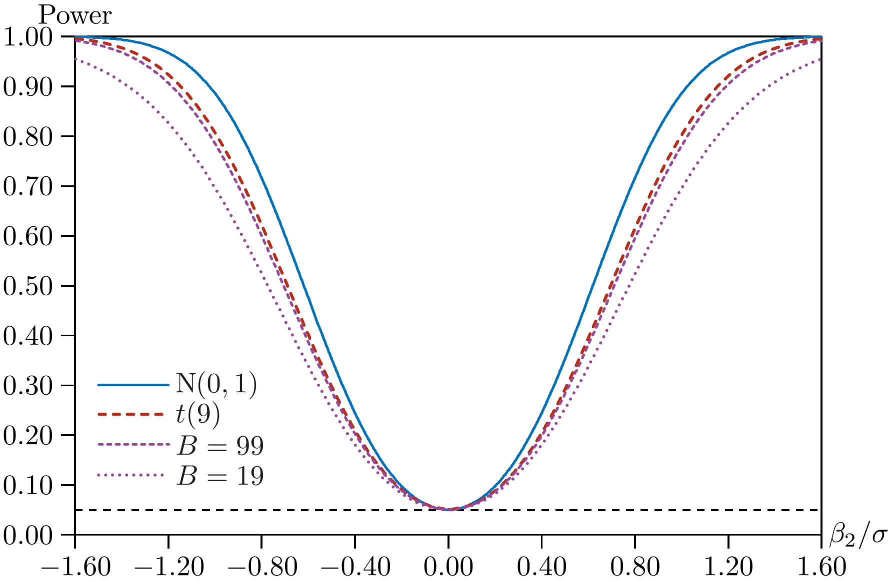
z_{β_2} follows the $\mathrm{N}(0,1)$ distribution, because σ is known. In contrast, t_{β_2} follows the $t(N-k)$ distribution, because σ is estimated.

t_{β_2} is equal to z_{β_2} times the random variable σ/s, which is independent of z_{β_2} and the same for both H_0 and H_1.

- Multiplying z_{β_2} by σ/s adds independent random noise.
- This requires us to use a larger critical value, which in turn causes the test based on t_{β_2} to be less powerful than the test based on z_{β_2}.

The figure illustrates power loss in going from z_{β_2} to t_{β_2}, plus the additional power loss from bootstrapping with finite B.

- Power loss is very rarely a problem when $B = 999$, and it is never a problem when $B = 9{,}999$.
- For confidence intervals, randomness due to finite B shows up as intervals that are longer than necessary.


Power
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
N(0, 1)
t(9)
B = 99
B = 19
β2/σ
−1.60
−1.20
−0.80
−0.40
0.00
0.40
0.80
1.20
1.60