THE ROCKEFELLER UNIVERSITY

Science for the benefit of humanity

The Parametric Bootstrap and Particle Physics

Luc Demortier

Progress on Statistical Issues in Searches

SLAC National Accelerator Laboratory, June 4–6, 2012

Outline

1. Motivation and Core Ideas

Motivation I

Particle Physics deals with data-generating processes that

1. often have a complicated dependence on both interest and nuisance parameters, but
2. can be modeled by Monte Carlo simulation.

Hence inferences are difficult to obtain by exact methods. The bootstrap provides a tractable numerical approach that approximates exact methods.

The bootstrap is a frequentist methodology. In terms of coverage accuracy, it can be thought of as a bridge between exact frequentist methods, which are typically not applicable to our complex data analysis problems, and asymptotic methods, which lack coverage in small data samples.

All that is required for applying bootstrap methods is the ability to generate "toy" experiments. The vast majority of bootstrap methods assume that the data are independent and identically distributed, which is the case in particle physics.

Motivation II

"An important point is that, subject to mild conditions, [a parameter of interest] can be the output of an algorithm of almost arbitrary complexity, shattering the naive notion that a parameter is a Greek letter appearing in a probability distribution and showing the possibilities for uncertainty analysis for the complex procedures now in daily use, but at the frontiers of the imagination a quarter of a century ago."

[Davison, A.C., Hinkley, D.V., and Young, G.A. (2003)]

Basic Ideas Underlying the Bootstrap I

There are two ideas at the core of the bootstrap:

1. The plug-in (or substitution) principle
   Bootstrap inferences apply to parameters that can be expressed as functionals of the data distribution: if F is the true distribution of the data, $F^\star$ an estimate of F, and θ a parameter of interest, then the true value of θ can be written as $\theta(F)$ and can be estimated by $\theta(F^\star)$.

   This is very general:

   - The functional can be as simple as the mean of a distribution:

$$\theta(F) = \int x\, dF(x) \quad \text{and} \quad \theta(F^\star) = \int x\, dF^\star(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

   - . . . or as complicated as the sampling distribution itself: inferring f from a random sample of x, $\{x_1, \ldots, x_n\}$, without assuming that f belongs to a known, finite-dimensional family of distributions.

So how does the bootstrap use $F^\star$ to make inferences about F?

Basic Ideas Underlying the Bootstrap II

2. **Resampling**

   Generate toy data samples that are statistically equivalent to the observed data sample. Two possibilities:

   (a) **Parametric resampling**

   The underlying distribution of the data is known up to one or more parameters; sampling is done from that distribution after substituting estimates for the parameter(s).

   (b) **Non-parametric resampling**

   The underlying distribution of the data is unknown; sampling is done directly from the data, with replacement. Given a dataset $\{x_1, \ldots, x_n\}$, this is done as follows: for each resample, n data points are successively drawn at random from the original dataset, and each selected data point is "replaced" in the set before selecting the next one. Thus, some of the original data points will appear more than once in the resampled dataset, and some will not appear at all.

This talk focuses mostly on parametric situations, because they are the most common in HEP. However, bootstrap methods can often be formulated either way.

Main Uses of the Bootstrap

- Bias reduction;
- Variance estimation;
- Confidence interval construction;
- Hypothesis testing.

Roger Barlow gave lectures at SLAC on the first two uses of the bootstrap in 2000, see `http://www-group.slac.stanford.edu/sluo/lectures/stat_lecture_files/sluolec6.pdf`. Here I'll concentrate on the last two.

Consistency of the Bootstrap

Under what conditions is the bootstrap consistent? When can we be sure that a bootstrap estimator converges to the true value as the sample size increases? There are theorems describing such conditions - and known examples where the conditions are not satisfied:

1. The bootstrap does not consistently estimate the distribution of a parameter estimator when the true value of the parameter lies on the boundary of parameter space.
2. The bootstrap does not consistently estimate the distribution of the maximum of a sample.

For such problems there exist alternative sampling schemes that are consistent, e.g. replacement subsampling and non-replacement subsampling (see Horowitz 2001).

2. Interval Estimation

"Since the early 1980s, a bewildering array of methods for constructing bootstrap confidence intervals have been proposed[. . .]"

[Carpenter, J., and Bithell, J., 2000]

A Confidence Interval Example from Particle Physics

To illustrate various bootstrap techniques for constructing confidence intervals, we consider a typical background subtraction problem in high energy physics.

Assume that we observe

$$N \sim \text{Poisson}(\theta + \nu),$$

where θ is an unknown signal intensity (parameter of interest) and ν an unknown background contamination (nuisance parameter). To determine ν we have an additional measurement

$$K \sim \text{Poisson}(\tau\nu),$$

with τ a known constant.

We wish to construct a confidence interval for θ.

This is a simple problem, with three complications: there is a nuisance parameter ν, there are physical boundaries at $\theta = 0$ and $\nu = 0$, and the sample space is discrete.

Some Notation for the Confidence Interval Example

The joint probability mass function of the primary measurement n and the auxiliary measurement k is:

$$f(n, k \mid \theta, \nu) = \frac{(\theta + \nu)^n e^{-\theta-\nu}}{n!} \frac{(\tau\nu)^k e^{-\tau\nu}}{k!},$$

and the likelihood ratio for testing a given value of θ is:

$$\lambda(n, k \mid \theta) = \frac{f(n, k \mid \theta, \hat{\hat{\nu}}(\theta))}{f(n, k \mid \hat{\theta}, \hat{\nu})},$$

where $(\hat{\theta}, \hat{\nu})$ is the maximum likelihood estimate (MLE) of (θ, ν) and $\hat{\hat{\nu}}(\theta)$ is the profile MLE of ν. All these MLEs are constrained to be positive for physical reasons.

From observed values n_{obs} and k_{obs} we can derive estimates $\hat{\theta}_{\text{obs}}$, $\hat{\nu}_{\text{obs}}$, and $\hat{\hat{\nu}}_{\text{obs}}(\theta)$.

A Non-Exhaustive List of Confidence Interval Techniques

Four non-bootstrap techniques:

1. Exact likelihood ratio test inversion ("exact" in the sense that the method never undercovers, but it does overcover);
2. Asymptotic likelihood ratio test inversion;
3. Naïve method, i.e. using a naïve estimator for the standard deviation of the estimate $\hat{\theta}$;
4. Bayesian elimination of the nuisance parameter;

and four bootstrap techniques:

5. Simple percentile;
6. Automatic percentile;
7. Bootstrap likelihood ratio test inversion;
8. Profile bootstrap likelihood ratio test inversion.

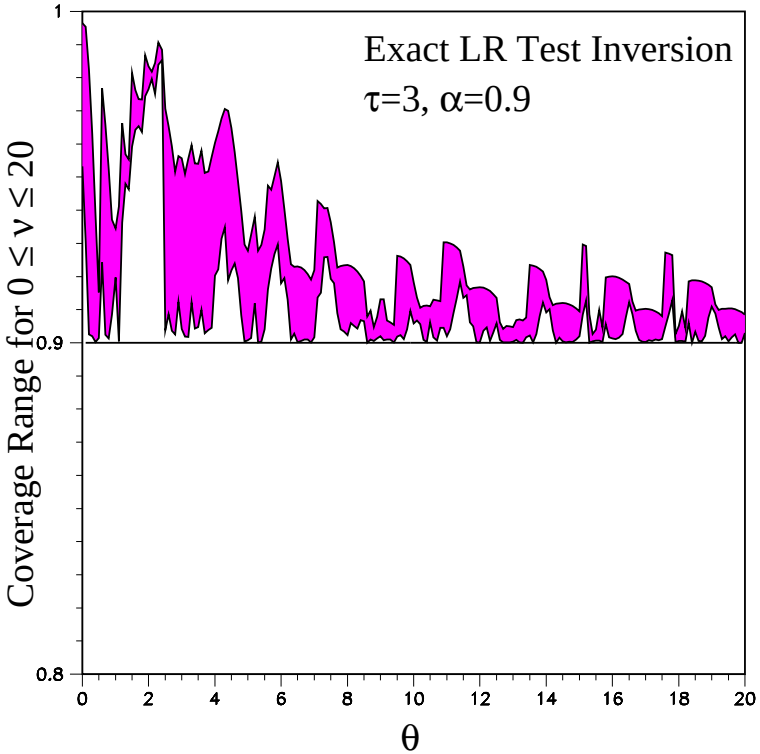
Interval Method I: Exact LR Test Inversion (Non-Bootstrap)

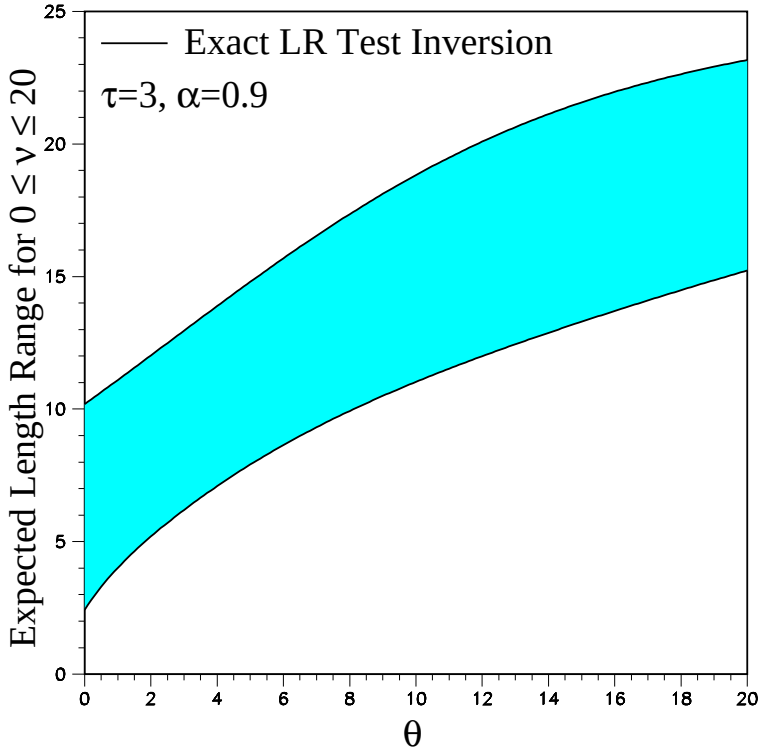
An α C.L. region is defined as the set of θ values for which

$$\min_{\nu}\left\{\mathbb{P}\left[-2\ln\lambda(N,K\,|\,\theta)\leq -2\ln\lambda(n_{\rm obs},k_{\rm obs}\,|\,\theta)\,\middle|\,\theta,\nu\right]\right\}\;\leq\;\alpha.$$

If we write $q_\alpha(\theta,\nu)$ for the α-quantile of the distribution of $-2\ln\lambda(N,K\mid\theta)$, this can be rewritten as:

$$-2\ln\lambda(n_{\rm obs},k_{\rm obs}\,|\,\theta)\;\leq\;q_\alpha(\theta)\;\equiv\;\max_{\nu}\,q_\alpha(\theta,\nu).$$



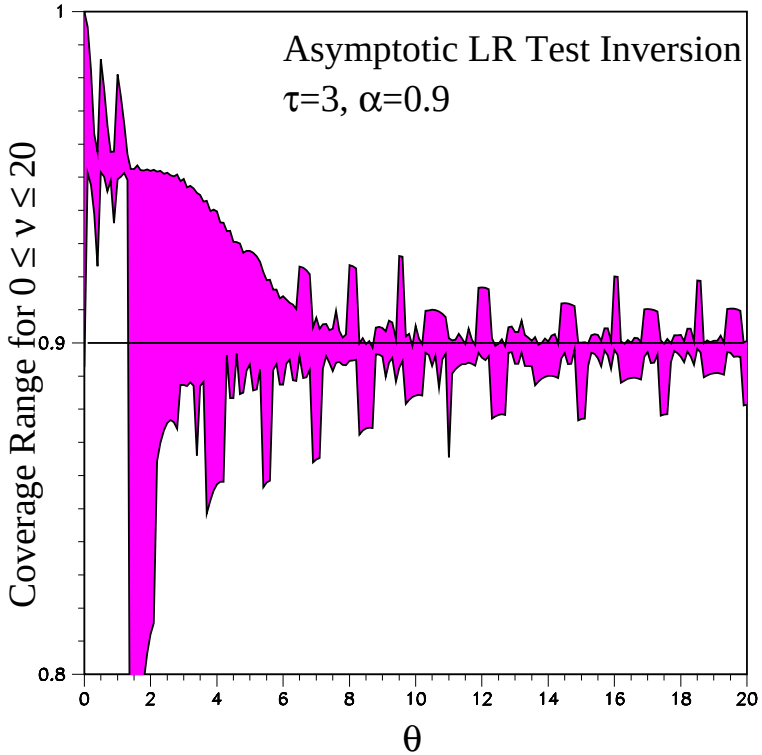


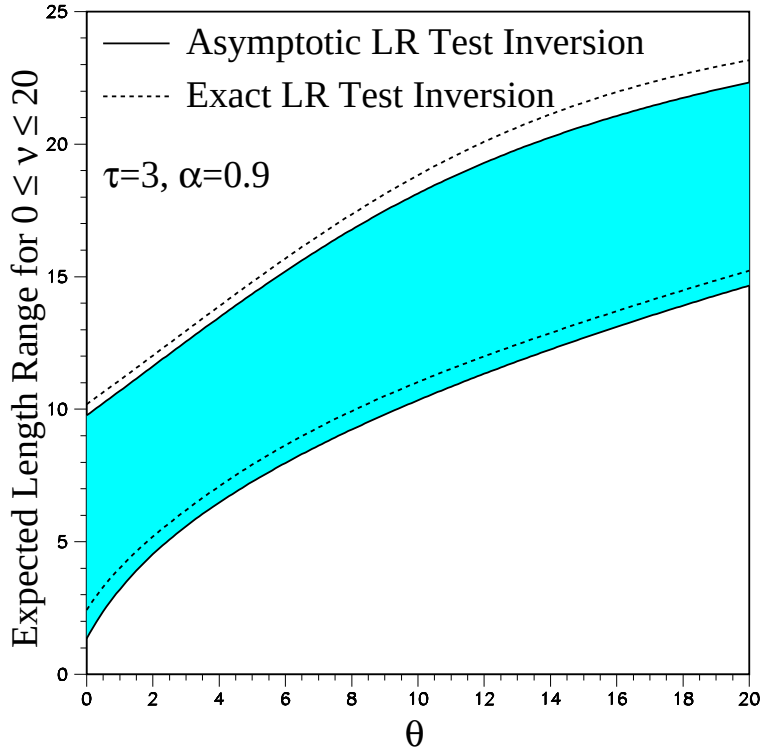
Interval Method II: Asymptotic LR Test Inversion (Non-Bootstrap)

An α C.L. region is defined as the set of θ values for which

$$-2 \ln \lambda(n_{\text{obs}}, k_{\text{obs}} \,|\, \theta) \;\leq\; \chi^2_{1,\alpha},$$

where $\chi^2_{1,\alpha}$ is the α-quantile of a chisquare distribution for one degree of freedom



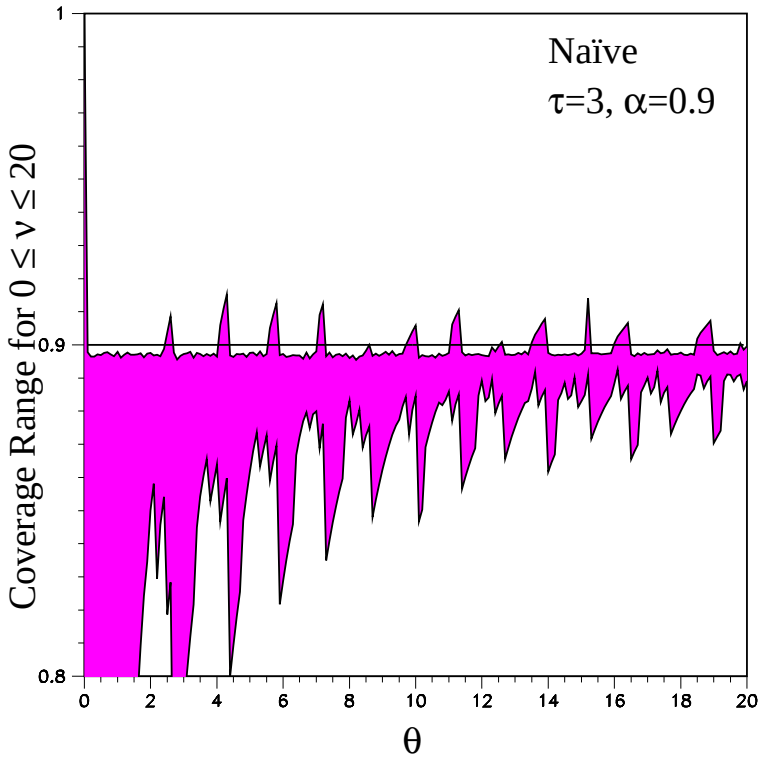


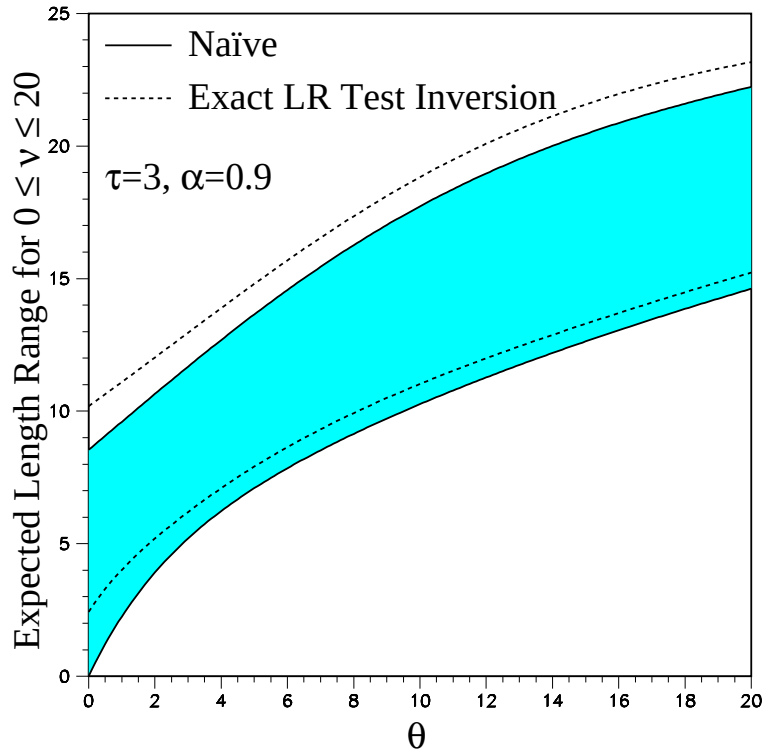
Interval Method III: Naïve (Non-Bootstrap)

Under our Poisson signal+background model, the MLE of θ is $\hat{\theta} = n - k/\tau$. Ignoring the physical constraint $\hat{\theta} \geq 0$, the variance of this MLE is $\theta + \nu + \nu/\tau$, which can be estimated by $n + k/\tau^2$. Thus an approximate α C.L. interval for θ is given by the intersection of

$$\left[\hat{\theta}_{\rm obs} - z_{\frac{1+\alpha}{2}}\sqrt{n_{\rm obs} + \frac{k_{\rm obs}}{\tau^2}},\ \hat{\theta}_{\rm obs} + z_{\frac{1+\alpha}{2}}\sqrt{n_{\rm obs} + \frac{k_{\rm obs}}{\tau^2}}\right]$$

with the physical region $\theta \geq 0$ (z_γ is a standard normal quantile).





Interval Method IV: Bayesian Elimination Plus LR Test Inversion (Non-Bootstrap)

In the original model:

$$f(n,k \mid \theta,\nu) \;=\; \frac{(\theta+\nu)^n e^{-\theta-\nu}}{n!}\,\frac{(\tau\nu)^k e^{-\tau\nu}}{k!},$$

the k-dependent component is reinterpreted as a prior for ν:

$$\pi(\nu) \;=\; \frac{\tau(\tau\nu)^k e^{-\tau\nu}}{\Gamma(k+1)},$$

and the pdf of n is integrated over $\pi(\nu)$ to obtain a distribution that depends on θ only:

$$f^\dagger(n \mid \theta) \;=\; \int \frac{(\theta+\nu)^n e^{-\theta-\nu}}{n!}\,\pi(\nu)\,d\nu.$$

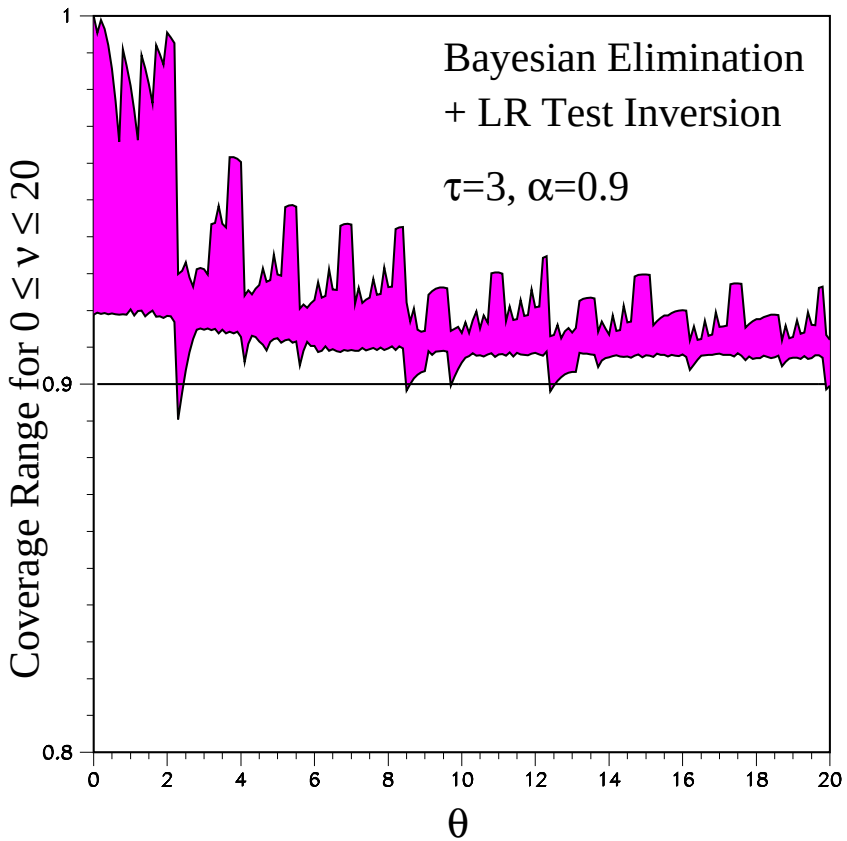
The likelihood ratio is

$$\lambda^\dagger(n_{\text{obs}} \mid \theta) \;=\; \frac{f^\dagger(n_{\text{obs}} \mid \theta)}{f^\dagger(n_{\text{obs}} \mid \hat\theta^\dagger)},$$

where $\hat\theta^\dagger$ maximizes $f^\dagger$ at the observed value n_{obs} of N.

Interval Method IV Continued

Next, obtain the α-quantile $q_\alpha^\dagger(\theta)$ of the distribution of $-2\ln\lambda^\dagger(N \mid \theta)$ under $f^\dagger(n \mid \theta)$, and θ values for which $-2\ln\lambda^\dagger(n_{\text{obs}} \mid \theta) \leq q_\alpha^\dagger(\theta)$ form the desired interval.

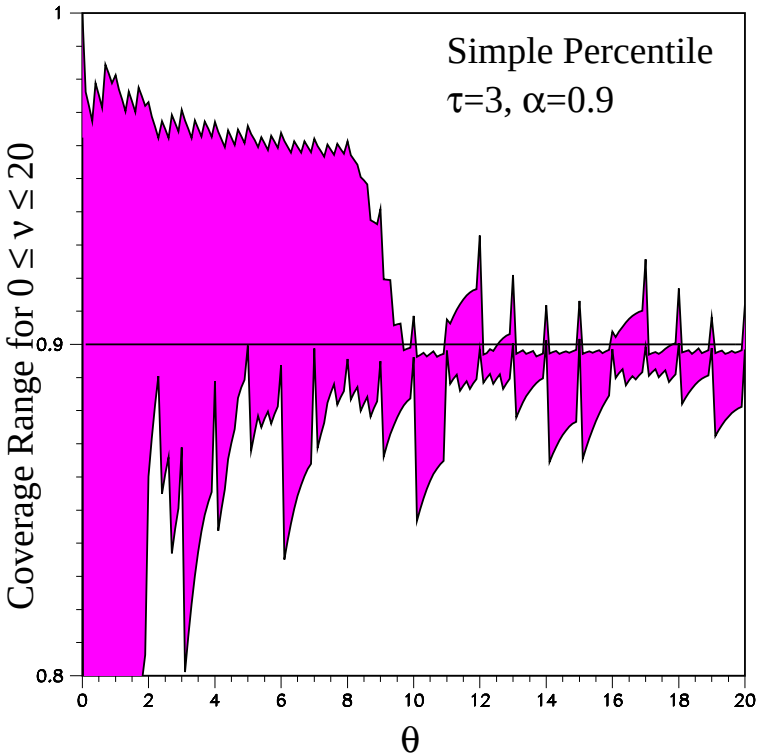


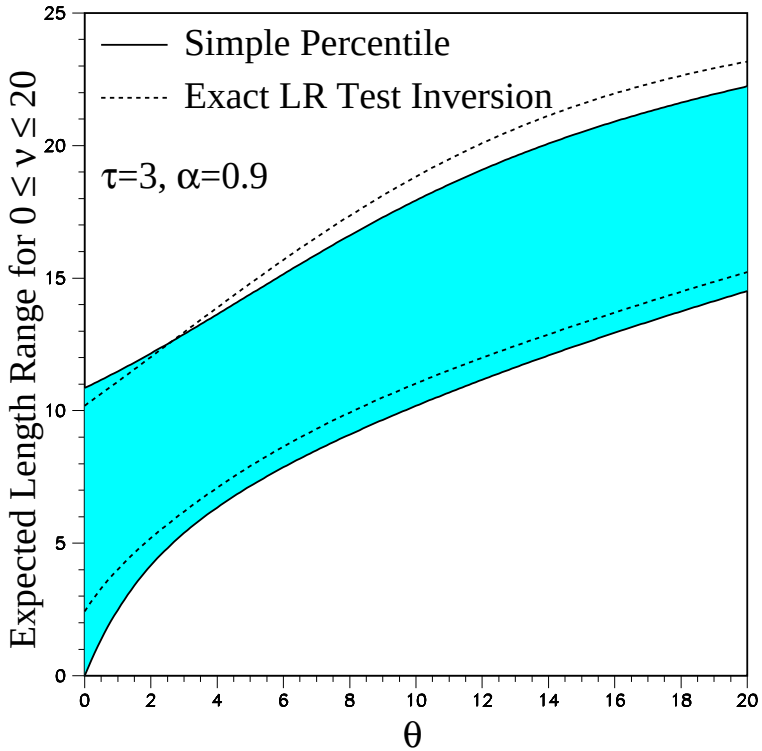
Interval Method V: Simple Percentile Bootstrap

Interval endpoints are set to the $\frac{1-\alpha}{2}$- and $\frac{1+\alpha}{2}$-quantiles of the distribution of the estimator

$$\hat{\theta} \equiv \max(N - K/\tau, 0).$$

The distribution of $\hat{\theta}$ is derived from $f(n, k \mid \hat{\theta}_{\text{obs}}, \hat{\nu}_{\text{obs}})$, where $\hat{\theta}_{\text{obs}}$ and $\hat{\nu}_{\text{obs}}$ are the MLEs determined from n_{obs} and k_{obs}.



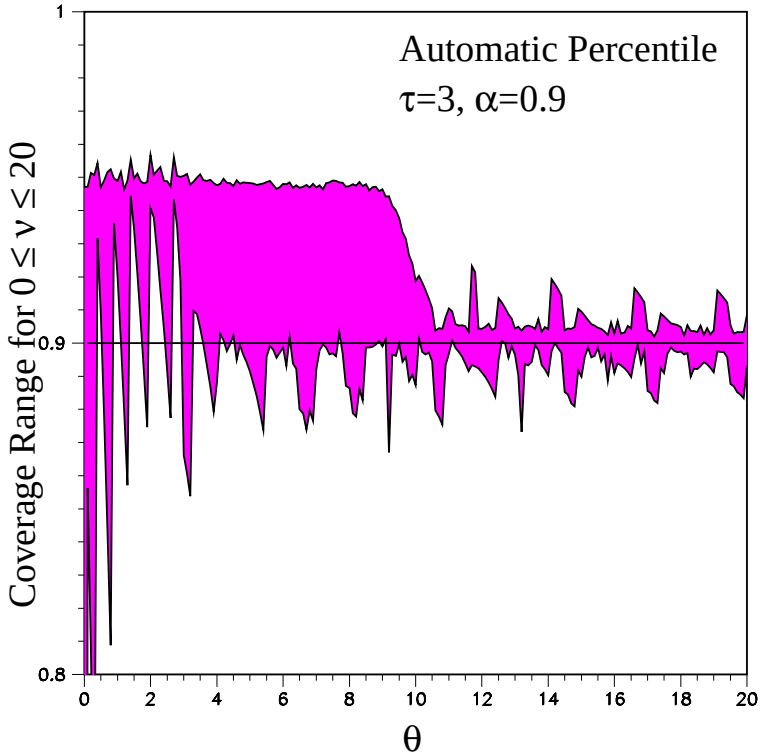


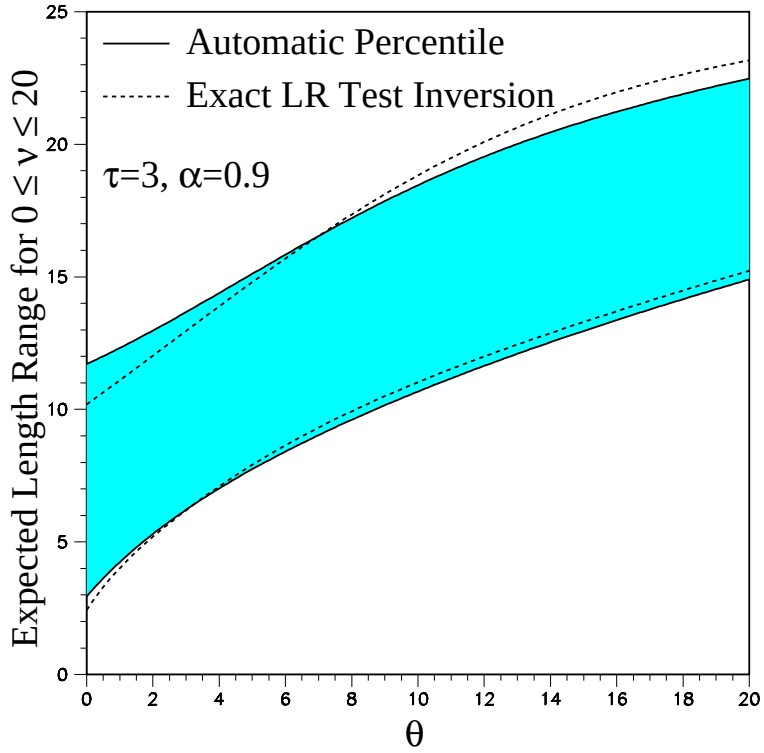
Interval Method VI: Automatic Percentile Bootstrap

Let $G(\hat{\theta} \mid \theta, \nu)$ be the cumulative distribution of the estimate $\hat{\theta}$. The interval endpoints θ_1 and θ_2 are the solutions of

$$G(\hat{\theta}_{\text{obs}} \mid \theta_1, \hat{\hat{\nu}}(\theta_1)) = \frac{1+\alpha}{2} \quad \text{and} \quad G(\hat{\theta}_{\text{obs}} \mid \theta_2, \hat{\hat{\nu}}(\theta_2)) = \frac{1-\alpha}{2},$$

where $\hat{\hat{\nu}}(\theta)$ is the profile MLE of ν. Note that in the absence of the nuisance parameter ν, an *exact* method is to solve $G(\hat{\theta}_{\text{obs}} \mid \theta_1) = 1 - G(\hat{\theta}_{\text{obs}} \mid \theta_2) = \frac{1+\alpha}{2}$, so the above method is a profile likelihood generalization of the exact one.





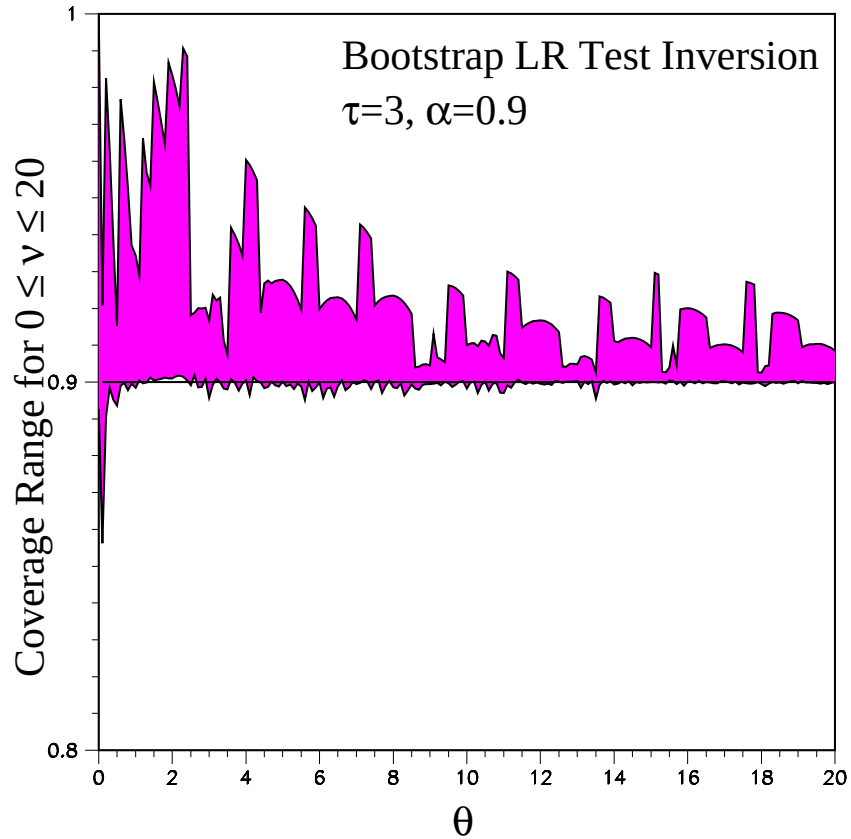
Interval Method VII: Bootstrap LR Test Inversion

This is a simplification of the exact LR test inversion method: instead of minimizing the LR tail probability w.r.t. ν, one substitutes an MLE. Thus, an α C.L. region is defined as the set of θ values for which

$$\mathbb{P}\Big[-2\ln\lambda(N,K\,|\,\theta) \le -2\ln\lambda(n_{\rm obs},k_{\rm obs}\,|\,\theta)\;\Big|\;\theta,\hat{\nu}_{\rm obs}\Big] \;\le\; \alpha.$$

If we write $q^{\star}_{\alpha}(\theta)$ for the α-quantile of the distribution of $-2\ln\lambda(N,K\mid\theta)$ under $f(n,k\mid\theta,\hat{\nu}_{\rm obs})$, the above equation can be rewritten as:

$$-2\ln\lambda(n_{\rm obs},k_{\rm obs}\,|\,\theta) \;\le\; q^{\star}_{\alpha}(\theta).$$



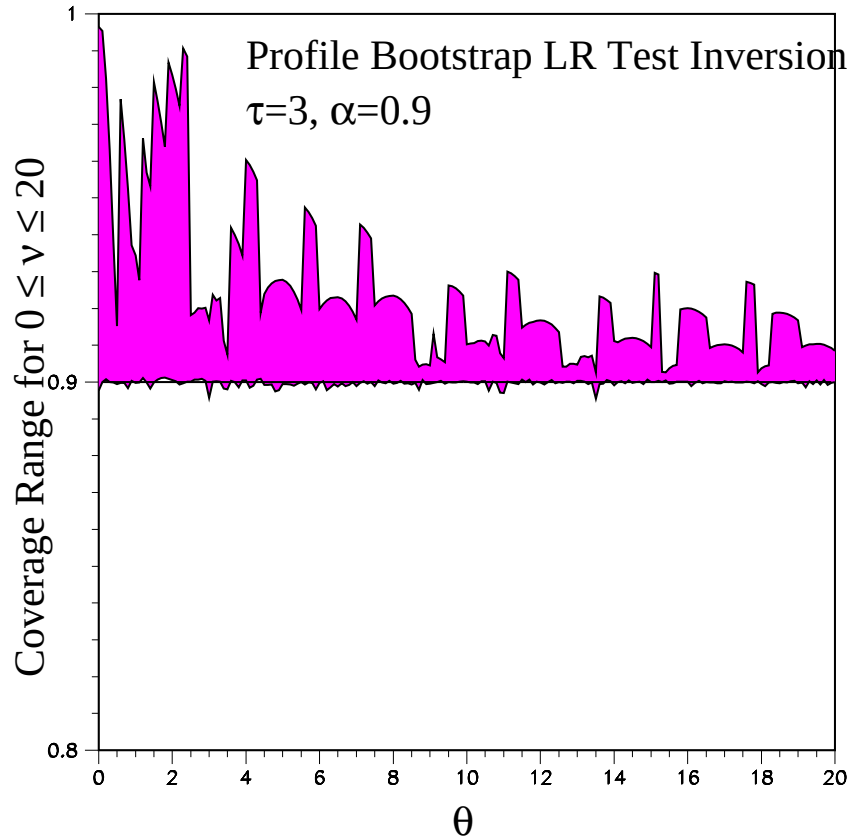
Interval Method VIII: Profile Bootstrap LR Test Inversion

This is a variation on the previous method: instead of substituting the MLE $\hat{\nu}_{\text{obs}}$ of ν into the LR tail probability, one substitutes the profile MLE $\hat{\hat{\nu}}_{\text{obs}}(\theta)$. An α C.L. region is defined as the set of θ values for which

$$\mathbb{P}\left[-2\ln\lambda(N,K\,|\,\theta) \le -2\ln\lambda(n_{\text{obs}},k_{\text{obs}}\,|\,\theta)\;\middle|\;\theta,\hat{\hat{\nu}}_{\text{obs}}(\theta)\right] \;\le\; \alpha.$$

If we write $q^{\star\prime}_{\alpha}(\theta)$ for the α-quantile of the distribution of $-2\ln\lambda(N,K\mid\theta)$ under $f(n,k\mid\theta,\hat{\hat{\nu}}_{\text{obs}}(\theta))$, the above equation can be rewritten as:

$$-2\ln\lambda(n_{\text{obs}},k_{\text{obs}}\,|\,\theta) \;\le\; q^{\star\prime}_{\alpha}(\theta).$$



Lesson 1 of 3 for Bootstrap Confidence Intervals

Use a pivot or an approximate pivot whenever possible. A pivot is a function of both data and parameters, whose distribution does not depend on any unknowns. Examples include:

- The studentized pivot for a location parameter μ: $(\hat{\mu} - \mu)/\hat{\sigma}$;
- The likelihood ratio;
- The cumulative distribution function of an estimator of the quantity of interest.

Suppose that $U(X, \theta)$ is an exact pivot, and that it is a monotone function of the quantity of interest θ. Then if u_γ denotes the γ-quantile of U, the set of θ values such that

$$u_\alpha \leq u(x_{\text{obs}}, \theta) \leq u_\beta$$

forms a $\beta - \alpha$ confidence interval for θ. If $U(X, \theta)$ is only an approximate or asymptotic pivot, a bootstrap confidence interval can be obtained by computing the u_γ from a bootstrap sample.

For example, if μ is a location parameter and one can compute a decent estimate of the variance σ^2 of its estimator $\hat{\mu}(X)$, then it is better to bootstrap the distribution of $U(X, \mu) = (\hat{\mu} - \mu)/\hat{\sigma}$ than that of $\hat{\mu}$.

Lesson 2 of 3 for Bootstrap Confidence Intervals

Test inversion improves the performance of the boostrap. This can be seen from the difference between the simple and automatic percentile methods. Test inversion can be computationally demanding. However, there exist approximate solutions. Suppose for example that we are interested in an α C.L. automatic percentile limit $\hat{\theta}_{\mathrm{AP},\alpha}$. This limit satisfies

$$\mathbb{P}\big[\hat{\theta} \leq \hat{\theta}_{\mathrm{obs}} \mid \hat{\theta}_{\mathrm{AP},\alpha}, \hat{\hat{\nu}}(\hat{\theta}_{\mathrm{AP},\alpha})\big] \;=\; 1-\alpha.$$

An approximate algorithm proceeds as follows:

1. Start from an initial approximation $\hat{\theta}_{0,\alpha}$, for example $\hat{\theta}_{0,\alpha} = \hat{\theta}_{\mathrm{SP},\alpha}$.
2. Find the $(1-\alpha)$-quantile $\hat{\theta}'_{0,\alpha}$ of $\hat{\theta}$ under the parameter value $(\hat{\theta}_{0,\alpha}, \hat{\hat{\nu}}(\hat{\theta}_{0,\alpha}))$: $\quad \mathbb{P}\big[\hat{\theta} \leq \hat{\theta}'_{0,\alpha} \mid \hat{\theta}_{0,\alpha}, \hat{\hat{\nu}}(\hat{\theta}_{0,\alpha})\big] \;=\; 1-\alpha.$
3. Find $\hat{\theta}_{1,\alpha}$ such that
$$\mathbb{P}\big[\hat{\theta} \leq \hat{\theta}_{1,\alpha} \mid \hat{\theta}_{\mathrm{obs}}, \hat{\nu}_{\mathrm{obs}}\big] \;=\; \mathbb{P}\big[\hat{\theta} \leq \hat{\theta}_{0,\alpha} \mid \hat{\theta}'_{0,\alpha}, \hat{\hat{\nu}}(\hat{\theta}'_{0,\alpha})\big].$$

Then $\hat{\theta}_{1,\alpha}$ differs from $\hat{\theta}_{\mathrm{AP},\alpha}$ by $\mathcal{O}_P(n^{-3/2})$, and this only requires calculating the distribution of $\hat{\theta}$ under $(\hat{\theta}_{0,\alpha}, \hat{\hat{\nu}}(\hat{\theta}_{0,\alpha}))$, $(\hat{\theta}'_{0,\alpha}, \hat{\hat{\nu}}(\hat{\theta}'_{0,\alpha}))$, and $(\hat{\theta}_{\mathrm{obs}}, \hat{\nu}_{\mathrm{obs}})$.

Lesson 3 of 3 for Bootstrap Confidence Intervals

Handle nuisance parameters by reduction to a least favorable family. Suppose that $X \sim f(x \mid \eta)$, where η is a vector of unknown parameters, and we are interested in the one-dimensional quantity $\theta = \theta(\eta)$. The first step is to reduce $f(x \mid \eta)$ to a one-parameter least favorable family through $\hat{\eta}$,

$$\hat{f}(x \mid \tau) \equiv f(x \mid \hat{\eta} + \tau\vec{\delta}),$$

where

$$\vec{\delta} \equiv I(\hat{\eta})^{-1}\vec{\nabla}\theta; \qquad I(\hat{\eta}) \equiv \left[-\frac{\partial^2 \ln L(x \mid \eta)}{\partial\eta\partial\eta^T}\right]_{\eta=\hat{\eta}}; \qquad \vec{\nabla}\theta \equiv \left.\frac{\partial\theta(\eta)}{\partial\eta}\right|_{\eta=\hat{\eta}}.$$

The Fisher information bound for an unbiased estimate of θ in this one-parameter family is the same as in the full multiparameter family: inferences about θ have not been made easier by restricting attention to a one-parameter family.

When $\hat{\eta}$ is the MLE of η, another least favorable family is $f(x \mid \hat{\hat{\eta}}(\theta))$, i.e. the *profile* likelihood. This family has the property that it transforms correctly under reparametrization, and this property carries over to automatic percentile limits.

Other Bootstrap Confidence Intervals

There exist many other bootstrap confidence interval constructions, e.g.

- Bootstrap-t, based on bootstrapping a studentized pivot such as $(\hat{\mu} - \mu)/\hat{\sigma}$;
- Bias-corrected percentile, an improvement on the simple percentile;
- Bias-corrected, accelerated percentile (BCa), an improvement on the bias-corrected percentile;
- Double bootstrap: bootstrap iterations can improve the coverage accuracy of bootstrap intervals, but at considerable computational cost;
- Bootstrap calibration: one can use the bootstrap to estimate the actual coverage of a given confidence limit, and then to correct (recalibrate) that limit.

These constructions differ in their coverage accuracy and other properties (for example equivariance under reparametrization).

Table of Bootstrap Confidence Interval Methods

Coverage Accuracy	Methods		
	Pivoting	Percentile	Test Inversion
1st order	Non-studentized Bootstrap	Simple Percentile Bias Corrected Percentile	
2nd order	Bootstrap-t	BCa Percentile Automatic Percentile	Bootstrap LR Profile Bootstrap LR
higher order	Bootstrap Calibration or Iteration		

3. Hypothesis Testing

Correct Resampling of Parametric Bootstrap P-Values

Hypothesis testing in particle physics is typically done with p-values. When bootstrapping a p-value to test a hypothesis H_0, it is important that the resampling reflect H_0. For a non-parametric example, suppose that we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, and wish to test whether the underlying population means are equal, $H_0 : \mu_X = \mu_Y$. A good test statistic is the pooled t statistic:

$$t_p = \frac{\bar{X} - \bar{Y}}{\sqrt{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \sqrt{\frac{m+n-2}{1/m + 1/n}}.$$

1. Resampling should not be done separately from the X and Y samples (this tends to have very limited power). Instead, one should draw both samples with replacement from the pooled sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.
2. Furthermore, even though in principle one could bootstrap the distribution of $\bar{X} - \bar{Y}$, it is much better to bootstrap the distribution of the approximate pivot t_p.

Convergence of Parametric Bootstrap P-Values

An interesting result is the following:

If the p-value is based on a test statistic T, and the distribution of T is asymptotically normal with mean $a(\theta)$ and variance $b^2(\theta)/n$, where θ is the parameter of interest, then under some fairly general conditions the parametric bootstrap p-value is asymptotically uniform if a does not depend on θ, and asymptotically conservative otherwise.

[Boos, D.D. (2003)]

A Bootstrap P-Value Example

To illustrate *p*-value calculations, consider testing a signal with unknown magnitude θ in the presence of unknown background contamination ν:

$$f(n, x \mid \theta, \nu) \;=\; \frac{(\theta+\nu)^n\, e^{-\theta-\nu}}{n!}\;\frac{e^{-\frac{1}{2}\left(\frac{x-\nu}{\Delta\nu}\right)^2}}{\sqrt{2\pi}\,\Delta\nu}.$$

We wish to test $H_0 : \theta = 0$. The constrained maximum-likelihood estimate of ν under H_0 is obtained by setting $\theta = 0$ and solving $\partial \ln f/\partial\nu = 0$ for ν. This yields:

$$\hat{\hat{\nu}} \;=\; \frac{x_{\text{obs}} - \Delta\nu^2}{2} \;+\; \sqrt{\left(\frac{x_{\text{obs}} - \Delta\nu^2}{2}\right)^2 \;+\; n_{\text{obs}}\,\Delta\nu^2}.$$

The parametric bootstrap (or "plug-in") *p*-value is then:

$$p_{plug} \;\equiv\; \sum_{n=n_{\text{obs}}}^{+\infty} \frac{\hat{\hat{\nu}}^n\, e^{-\hat{\hat{\nu}}}}{n!}.$$

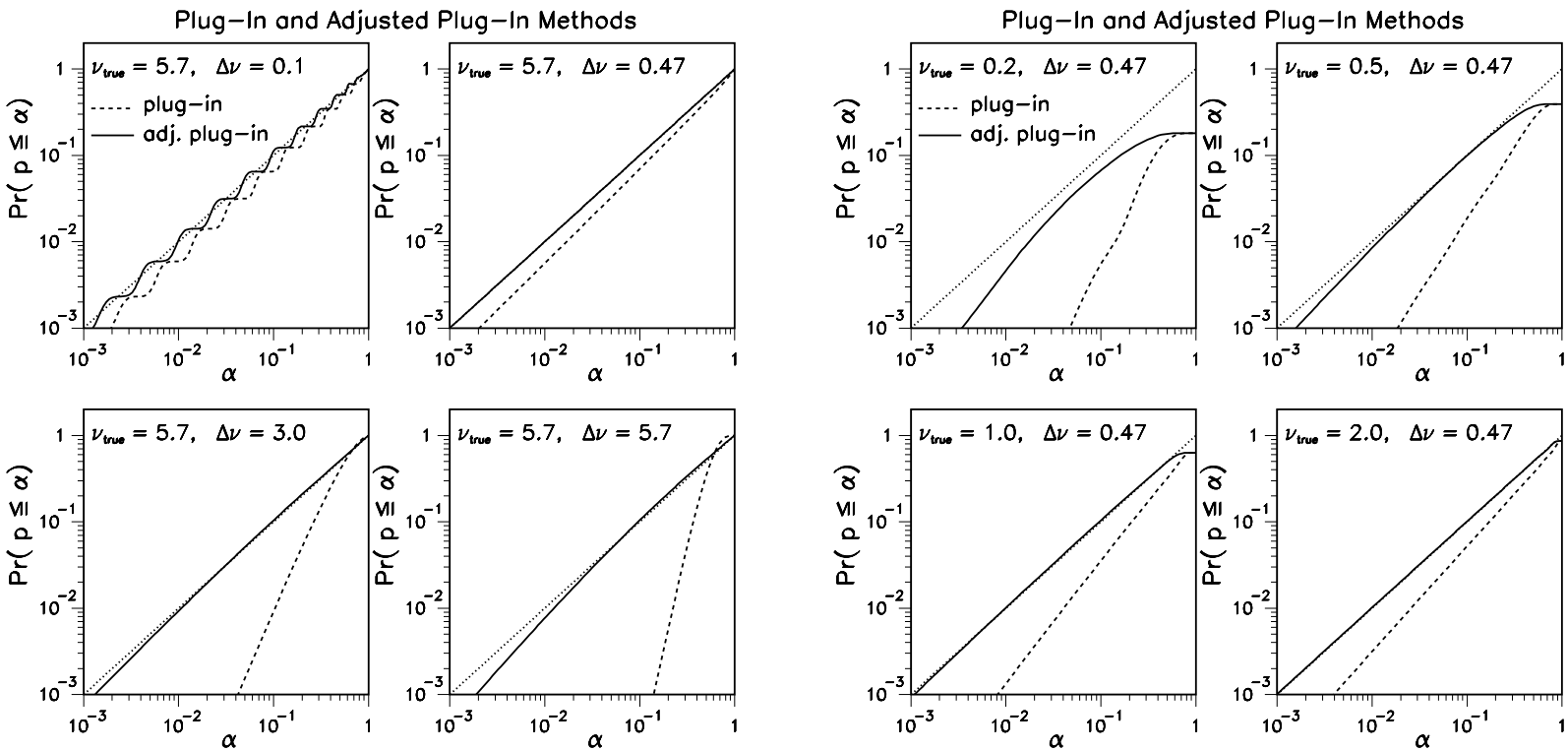
Note that p_{plug} makes double use of the data, first in the calculation of $\hat{\hat{\nu}}$, and second in the calculation of the tail probability. This tends to favor H_0. On the other hand, p_{plug} does not take into account the uncertainty on $\hat{\hat{\nu}}$, which works against H_0.

A Bootstrap P-Value Example, Continued

If we knew the exact cumulative distribution function F_{plug} of p_{plug} under H_0, then $F_{plug}(p_{plug})$ would be an exact p-value. Since F_{plug} depends on unknown parameters, we can apply the parametric bootstrap by substituting estimates for these unknown parameters. This leads to the "adjusted plug-in" p-value:

$$p_{plug,adj} \equiv F_{plug}(p_{plug} \mid \theta = 0, \nu = \hat{\hat{\nu}}).$$

This is equivalent to a double bootstrap.



Hypothesis Testing: How Many Bootstrap Resamples?

It can be shown that using a finite number B of bootstrap replications often results in a test that is less powerful than the ideal test obtained in the limit $B \to \infty$. On the other hand, bootstrap replications can be CPU intensive. A compromise value of B must be found.

Suppose that we don't care about an exact p-value, we only wish to determine whether $p \leq \alpha$ or $p > \alpha$ for some predetermined α. Then the following pretest procedure can be used:

1. Compute test statistic T from data, set $B = B' = B_{\text{min}}$, and compute $T_i^\star$ for $B = B_{\text{min}}$ bootstrap samples.
2. Compute $\hat{p}^\star$ based on B bootstrap samples.
   - If $\hat{p}^\star < \alpha$, test the hypothesis that $p^\star \geq \alpha$ at level β, using binomial statistics. If the hypothesis is rejected, stop.
   - If $\hat{p}^\star > \alpha$, test the hypothesis that $p^\star \leq \alpha$ at level β. If the hypothesis is rejected, stop.
3. If you get to this step, set $B = 2B'$. If $B > B_{\text{max}}$, stop. Otherwise calculate $T_i^\star$ for a further B samples, set $B' = B$, and return to step 2.

For $\alpha = 0.05$, it is found that $\beta = 0.001$ and $B_{\text{max}} = 12800$ yield good performance (Davidson, R., and MacKinnon, J., 2000).

The Look-Elsewhere Effect (LEE)

When we search for a signal without knowing its precise location, the significance of any effect we observe must be degraded for the fact that the background could have fluctuated up *anywhere* in the search range. This is the LEE. In the statistics literature this effect shows up under various guises, e.g. "hypothesis testing when a nuisance parameter is present only under the alternative". Another guise is "inference for the number of components in a mixture distribution." A simple example of the latter is:

$$H_0 : f(x) = N(0,1) \quad \text{versus} \quad f(x) = (1-\pi)N(0,1) + \pi N(\mu,1),$$

where $N(\mu, \sigma^2)$ is a Gaussian density with mean μ and variance σ^2. The null hypothesis corresponds to $\pi = 0$ or $\mu = 0$. Hence the parameter is on the boundary of parameter space and H_0 corresponds to a non-identifiable subset of parameter space. The classic theorems about the asymptotic properties of the MLE and the likelihood ratio do not hold.

Bootstrapping the parameters will not work since they are not identifiable. However the likelihood is identifiable, and therefore also the likelihood ratio. Testing H_0 by bootstrapping the likelihood ratio will work, in contrast with parameter-based tests such as Wald's test (see Feng, Z.D. and McCullogh, C.E., 1996).

4. Bootstrap Diagnostics

Bootstrap Diagnostics

How can we assess the reliability of bootstrap calculations? There are several concerns (Canty, Davison, Hinkley, and Ventura (2006)):

1. *Data outliers*
   Their effect can be checked by removing data points one by one from the sample.
2. *Incorrect resampling*
   For the parametric bootstrap this can become a problem when there are many nuisance parameters.
3. *Nonpivotality*
   For the parametric bootstrap this is straightforward to check by varying the parameter value over a wide range that includes the estimated value, and examining how the mean, variance, and quantiles of the distribution of the bootstrapped quantity changes.
4. *Inconsistency*
   There are cases where it is possible to check whether or not the bootstrap is consistent, but there is no general recipe.

"Further theory and more diagnostics are necessary before bootstrap procedures become fully trustworthy additions to the statistical toolkit."

Summary

- The bootstrap is a frequentist tool that bridges exact calculations and asymptotic ones.
- Bootstrap inferences can be obtained for any quantity of interest that is the output of an algorithm, regardless of complexity.
- For confidence intervals, the coverage accuracy of bootstrap procedures can be improved by bootstrapping a pivot, by using test inversion, and by bootstrap iteration.
- For hypothesis testing, bootstrap resampling should always reflect the null hypothesis.
- The look-elsewhere effect can be handled by bootstrapping a likelihood ratio.
- There is a need for tools to help diagnose misbehavior of bootstrap procedures.

References

1. Boos, D.D. (2003), "Introduction to the bootstrap world," Statist. Sci. **18**, 168.
2. Canty, A.J., Davison, A.C., Hinkley, D.V., and Ventura, V. (2006), "Bootstrap diagnostics and remedies," Can. J. Statist. **34**, 5.
3. Carpenter, J., and Bithell, J. (2000), "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," Statist. Med. **19**, 1141.
4. Davidson, R., and MacKinnon, J. (2000), "Bootstrap tests: how many bootstraps?," Econometric Reviews **19**, 55.
5. Davison A.C. and Hinkley, D.V. (1997), "Bootstrap methods and their application," Cambridge University Press.
6. Davison, A.C., Hinkley, D.V., and Young, G.A. (2003), "Recent developments in bootstrap methodology," Statist. Sci. **18**, 141.
7. DiCiccio, T.J., and Romano, J.P. (1995), "On bootstrap procedures for second-order accurate confidence limits in parametric models," Statistica Sinica **5**, 141.
8. Efron, B. and Tibshirani, R.J. (1993), "An introduction to the bootstrap," Chapman & Hall.

Some references

9. Feng, Z.D. and McCullogh, C.E. (1996), "Using bootstrap likelihood ratios in finite mixture models," J. R. Statist. Soc. B **58**, 609.
10. Horowitz, J.L. (2001), "The bootstrap," in *Handbook of Econometrics, Vol. 5*, Heckman, J.J., and Leamer, E., eds., Elsevier.
11. Robins, J., Van Der Vaart, A., and Ventura, V. (2000),"Asymptotic distribution of *p* values in composite null models," J. Amer. Statist. Assoc. **95**, 1143.
12. Scholz, F.W. (2007), "The bootstrap small sample properties," Boeing Computer Services, Research and Technology, Report bcstech-93-051.