Statistics 240 Lecture Notes

P.B. Stark `www.stat.berkeley.edu/~stark/index.html`

Revised 23 April 2008

1 Part 8: Resampling Methods

References:

- Beran, R., 1995. Stein confidence sets and the bootstrap, *Stat. Sinica, 5*, 109–127;
- Beran, R., 1990. Calibrating predictions regions, *J. Amer. Stat. Assoc., 85*, 715–723;
- Beran, R., 1990. Refining bootstrap simultaneous confidence sets, *J. Amer. Stat. Assoc., 85*, 417–426;
- Beran, R., 1987. Prepivoting to reduce level error of confidence sets, *Biometrika, 74*, 457–468;
- Feller, W., 1971. *An introduction to probability theory and its applications, V. II*, 2nd edition, John Wiley and Sons, Inc., New York;
- Freedman, D.A., 2005. *Statistical Models: Theory and Practice*, Cambridge University Press;
- Lehmann, E.L., 1975. *Nonparametrics*, Holden-Day, Oakland;
- Efron, B., 1982. *The Jackknife, the bootstrap, and other resampling plans,* SIAM, Philadelphia;
- Romano, J.P., 1988. A bootstrap revival of some nonparametric distance tests, *J. Amer. Stat. Assoc., 83*, 698–708;
- Romano, J.P., 1989. Bootstrap and randomization tests of some nonparametric hypotheses, *Ann. Stat., 17*, 141–159.

1.1 The Bootstrap

The setting for the next few lectures is that we observe an iid sample of size n, $\{X_j\}_{j=1}^n$ iid F. Each observation is real-valued. We wish to estimate some parameter of the distribution of F that can be written as a functional of F, $T(F)$. Examples include the mean, $T(F) = \int x dF(x)$, other moments, *etc.*

The (unpenalized) nonparametric maximum likelihood estimator of F from the data $\{X_j\}$ is just the empirical distribution $\hat{F}_n$, which assigns mass $1/n$ to each observation:

$$\arg \max_{\text{probability distributions } G} \mathbb{P}_G\{X_j = x_j,\ j = 1, \ldots, n\} = \hat{F}_n. \tag{1}$$

(Note, however, that the MLE of F is not generally consistent in problems with an infinite number of parameters, such as estimating a density or a distribution function.)

Using the general principle that the maximum likelihood estimator of a function of a parameter is that function of the maximum likelihood estimator of the parameter, we might be led to consider $T(\hat{F}_n)$ as an estimator of $T(F)$.

That is exactly what the sample mean does, as an estimator of the mean:

$$T(\hat{F}_n) = \int x d\hat{F}_n(x) = \sum_{j=1}^{n} \frac{1}{n} X_j = \frac{1}{n} \sum_j X_j. \tag{2}$$

Similarly, the maximum likelihood estimator of

$$\mathbf{Var}(X) = T(F) = \int \left(x - \int x dF\right)^2 dF \tag{3}$$

is

$$T(\hat{F}_n) = \int \left(x - \int x d\hat{F}_n\right)^2 d\hat{F}_n = \frac{1}{n} \sum_j \left(X_j - \frac{1}{n} \sum_k X_k\right)^2. \tag{4}$$

In these cases, we get analytically tractable expressions for $T(\hat{F}_n)$.

What is often more interesting is to estimate a property of the sampling distribution of the estimator $T(\hat{F}_n)$, for example the variance of the estimator $T(\hat{F}_n)$. The bootstrap approximates the sampling distribution of $T(\hat{F}_n)$ by the sampling distribution of $T(\hat{F}_n^*)$, where $\hat{F}_n^*$ is a size-n iid random sample drawn from $\hat{F}_n$. That is, the bootstrap approximates the sampling distribution of an estimator applied to the empirical distribution $\hat{F}_n$ of a random sample of size n from a distribution F by the sampling distribution of that estimator applied to a random sample $\hat{F}_n^*$ of size n from a particular realization $\hat{F}_n$ of the empirical distribution of a sample of size n from F.

When T is the mean $\int x dF$, so $T(\hat{F}_n)$ is the sample mean, we could obtain the variance of the distribution of $T(\hat{F}_n^*)$ analytically: Let $\{X_j^*\}_{j=1}^n$ be an iid sample of size n from $\hat{F}_n$. Then

$$\mathbf{Var}_{\hat{F}_n} \frac{1}{n} \sum_{j=1}^n X_j^* = \frac{1}{n^2} \sum_{j=1}^n (X_j - \bar{X})^2, \tag{5}$$

where $\{X_j\}$ are the original data and $\bar{X}$ is their mean. When we do not get a tractable espression for the variance of an estimator under resampling from the empirical distribution, we could still approximate the distribution of $T(\hat{F}_n)$ by generating a large number of size-n iid $\hat{F}_n$ data sets (drawing samples of size n with replacement from $\{x_j\}_{j=1}^n$), and applying T to each of those sets.

The idea of the bootstrap is to approximate the distribution (under F) of an estimator $T(\hat{F}_n)$ by the distribution of the estimator under $\hat{F}_n$, and to approximate *that* distribution by using a computer to take a large number of pseudo-random samples of size n from $\hat{F}_n$.

This basic idea is quite flexible, and can be applied to a wide variety of testing and estimation problems, including finding confidence sets for functional parameters. (It is not a panacea, though: we will see later how delicate it can be.) It is related to some other "resampling" schemes in which one re-weights the data to form other distributions. Before doing more theory with the bootstrap, let's examine the jackknife.

1.2 The Jackknife

The idea behind the jackknife, which is originally due to Tukey and Quenouille, is to form from the data $\{X_j\}_{j=1}^n$, n sets of $n-1$ data, leaving each datum out in turn. The "distribution" of T applied to these n sets is used to approximate the distribution of $T(\hat{F}_n)$. Let $\hat{F}_{(i)}$ denote the empirical distribution of the data set with the ith value deleted; $T_{(i)} = T(\hat{F}_{(i)})$ is the corresponding estimate of $T(F)$. An estimate of the expected value of $T(\hat{F}_n)$ is

$$\hat{T}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n T(\hat{F}_{(i)}). \tag{6}$$

Consider the bias of $T(\hat{F}_n)$:

$$E_F T(\hat{F}_n) - T(F). \tag{7}$$

Quenouille's jackknife estimate of the bias is

$$\widehat{\mathrm{BIAS}} = (n-1)(\hat{T}_{(\cdot)} - T(\hat{F}_n)). \tag{8}$$

It can be shown that if the bias of T has a homogeneous polynomial expansion in n^{-1} whose coefficients do not depend on n, then the bias of the bias-corrected estimate

$$\tilde{T} = nT(\hat{F}_n) - (n-1)T_{(\cdot)} \tag{9}$$

is $O(n^{-2})$ instead of $O(n^{-1})$.

Applying the jackknife estimate of bias to correct the plug-in estimate of variance reproduces the formula for the sample variance (with $1/(n-1)$) from the formula with $1/n$: Define

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j, \tag{10}$$

$$\bar{X}_{(i)} = \frac{1}{n-1}\sum_{j\neq i} X_j, \tag{11}$$

$$T(\hat{F}_n) = \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2, \tag{12}$$

$$T(\hat{F}_{(i)}) = \frac{1}{n-1}\sum_{j\neq i}(X_j - \bar{X}_{(i)})^2, \tag{13}$$

$$T(\hat{F}_{(\cdot)}) = \frac{1}{n}\sum_{i=1}^{n} T(\hat{F}_{(i)}). \tag{14}$$

Now

$$\bar{X}_{(i)} = \frac{n\bar{X} - X_i}{n-1} = \bar{X} + \frac{1}{n-1}(\bar{X} - X_i), \tag{15}$$

so

$$\begin{aligned}(X_j - \bar{X}_{(i)})^2 &= \Big(X_j - \bar{X} - \frac{1}{n-1}(\bar{X} - X_i)\Big)^2 \\ &= (X_j - \bar{X})^2 + \frac{2}{n-1}(X_j - \bar{X})(X_i - \bar{X}) + \frac{1}{(n-1)^2}(X_i - \bar{X})^2. \end{aligned} \tag{16}$$

Note also that

$$\sum_{j\neq i}(X_j - \bar{X}_{(i)})^2 = \sum_{j=1}^{n}(X_j - \bar{X}_{(i)})^2 - (X_i - \bar{X}_{(i)})^2. \tag{17}$$

Thus

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j\neq i}(X_j - \bar{X}_{(i)})^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{n}\Big[(X_j - \bar{X})^2 + \frac{2}{n-1}(X_j - \bar{X})(X_i - \bar{X}) + \\ &\quad + \frac{1}{(n-1)^2}(X_i - \bar{X})^2\Big] - (X_i - \bar{X})^2 - \\ &\quad - \frac{2}{n-1}(X_i - \bar{X})^2 - \frac{1}{(n-1)^2}(X_i - \bar{X})^2\Bigg]. \end{aligned} \tag{18}$$

The last three terms all are multiples of $(X_i - \bar{X})^2$; the sum of the coefficients is

$$1 + 2/(n-1) + 1/(n-1)^2 = n^2/(n-1)^2. \tag{19}$$

The middle term of the inner sum is a constant times $(X_j - \bar{X})$, which sums to zero over j. Simplifying the previous displayed equation yields

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j \neq i} (X_j - \bar{X}_{(i)})^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(n\hat{\sigma}^2 + \frac{n}{(n-1)^2}(X_i - \bar{X})^2 - \frac{n^2}{(n-1)^2}(X_i - \bar{X})^2 \right) \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (n\hat{\sigma}^2 - \frac{n}{n-1}(X_i - \bar{X})^2) \\
&= \frac{1}{n-1} \left[n^2\hat{\sigma}^2 - \frac{n^2}{n-1}\hat{\sigma}^2 \right] \\
&= \frac{n(n-2)}{(n-1)^2}\hat{\sigma}^2. \qquad (20)
\end{aligned}$$

The jackknife bias estimate is thus

$$\widehat{\mathrm{BIAS}} = (n-1)\left(T(\hat{F}_{(\cdot)}) - T(\hat{F}_n) \right) = \hat{\sigma}^2 \frac{n(n-2) - (n-1)^2}{n-1} = \frac{-\hat{\sigma}^2}{n-1}. \tag{21}$$

The bias-corrected MLE variance estimate is therefore

$$\hat{\sigma}^2 \left(1 - \frac{1}{n-1} \right) = \hat{\sigma}^2 \frac{n}{n-1} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2 = S^2, \tag{22}$$

the usual sample variance.

The jackknife also can be used to estimate other properties of an estimator, such as its variance. The jackknife estimate of the variance of $T(\hat{F}_n)$ is

$$\hat{\mathbf{Var}}(T) = \frac{n-1}{n} \sum_{j=1}^{n} (T_{(j)} - T_{(\cdot)})^2. \tag{23}$$

It is convenient to think of distributions on data sets to compare the jackknife and the bootstrap. We shall follow the notation in Efron (1982). We condition on $(X_i = x_i)$ and treat the data as fixed in what follows. Let $\mathcal{S}_n$ be the n-dimensional simplex

$$\mathcal{S}_n \equiv \{\mathbb{P}^* = (P_i^*)_{i=1}^n \in \mathbb{R}^n : P_i^* \geq 0 \text{ and } \sum_{i=1}^{n} P_i^* = 1\}. \tag{24}$$

A *resampling vector* $\mathbb{P}^* = (P_k^*)_{k=1}^n$ is any element of $\mathcal{S}_n$; *i.e.*, an n-dimensional discrete probability vector. To each $\mathbb{P}^* = (P_k^*) \in \mathcal{S}_n$ there corresponds a re-weighted empirical measure $\hat{F}(\mathbb{P}^*)$ which puts mass P_k^* on x_k, and a value of the estimator $T^* = T(\hat{F}(\mathbb{P}^*)) = T(\mathbb{P}^*)$. The resampling vector

$\mathbb{P}^0 = (1/n)_{j=1}^n$ corresponds to the empirical distribution $\hat{F}_n$ (each datum x_j has the same mass). The resampling vector

$$\mathbb{P}_{(i)} = \frac{1}{n-1}(1, 1, \ldots, 0, 1, \ldots, 1), \tag{25}$$

which has the zero in the ith place, is one of the n resampling vectors the jackknife visits; denote the corresponding value of the estimator T by $T_{(i)}$. The bootstrap visits all resampling vectors whose components are multiples of $1/n$.

The bootstrap estimate of variance tends to be better than the jackknife estimate of variance for nonlinear estimators because of the distance between the empirical measure and the resampled measures:

$$\|\mathbb{P}^* - \mathbb{P}^0\| = O_P(n^{-1/2}), \tag{26}$$

while

$$\|\mathbb{P}_{(k)} - \mathbb{P}^0\| = O(n^{-1}). \tag{27}$$

To see the former, recall that the difference between the empirical distribution and the true distribution is $O_P(n^{-1/2})$: For any two probability distributions $\mathbb{P}_1$, $\mathbb{P}_2$, on $\mathbb{R}$, define the Kolmogorov-Smirnov distance

$$d_{KS}(\mathbb{P}_1, \mathbb{P}_2) \equiv \|\mathbb{P}_1 - \mathbb{P}_2\|_{KS} \equiv \sup_{x \in \mathbb{R}} |\mathbb{P}_1\{(-\infty, x]\} - \mathbb{P}_2\{(-\infty, x]\}|. \tag{28}$$

There exist universal constants $\chi_n(\alpha)$ so that for every continuous (w.r.t. Lebesgue measure) distribution F,

$$\mathbb{P}_F \left\{ \|F - \hat{F}_n\|_{KS} \geq \chi_n(\alpha) \right\} = \alpha. \tag{29}$$

This is the Dvoretzky-Kiefer-Wolfowitz inequality. Massart (*Ann. Prob., 18*, 1269–1283, 1990) showed that the constant

$$\chi_n(\alpha) \leq \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}} \tag{30}$$

is *tight.* Thinking of the bootstrap distribution (the empirical distribution $\hat{F}_n$) as the true cdf and the resamples from it as the data gives the result that the distance between the cdf of the bootstrap resample and the empirical cdf of the original data is $O_P(n^{-1/2})$.

To see that the cdfs of the jackknife samples are $O(n^{-1})$ from the empirical cdf $\hat{F}_n$, note that for univariate real-valued data, the difference between $\hat{F}_n$ and the cdf of the jackknife data set that leaves out the jth ranked observation $X_{(j)}$ is largest either at $X_{(j-1)}$ or at $X_{(j)}$. For $j = 1$ or $j = n$, the jackknife samples that omit the smallest or largest observation, the L_1 distance between the jackknife measure and the empirical distribution is exactly $1/n$. Consider the jackknife cdf $\hat{F}_{n,(j)}$, the cdf of the sample without $X_{(j)}$, $1 < j < n$.

$$\hat{F}_{n,(j)}(X_{(j)}) = (j-1)/(n-1), \tag{31}$$

while $\hat{F}_n((X_{(j)}) = j/n$; the difference is

$$\frac{j}{n} - \frac{j-1}{n-1} = \frac{j(n-1) - n(j-1)}{n(n-1)} = \frac{n-j}{n(n-1)} = \frac{1}{n-1} - \frac{j}{n(n-1)}. \tag{32}$$

On the other hand,

$$\hat{F}_{n,(j)}(X_{(j-1)}) = (j-1)/(n-1), \tag{33}$$

while $\hat{F}_n((X_{(j-1)}) = (j-1)/n$; the difference is

$$\frac{j-1}{n-1} - \frac{j-1}{n} = \frac{n(j-1) - (n-1)(j-1)}{n(n-1)} = \frac{j-1}{n(n-1)}. \tag{34}$$

Thus

$$\|\hat{F}_{n,(j)} - \hat{F}_n\| = \frac{1}{n(n-1)} \max\{n-j, j-1\}. \tag{35}$$

But $n/2 \le \max\{n-j, j-1\} \le n-1$, so

$$\|\hat{F}_{n,(j)} - \hat{F}_n\| = O(n^{-1}). \tag{36}$$

The neighborhood that the bootstrap samples is larger, and is probabilistically of the right size to correspond to the uncertainty of the empirical distribution function as an estimator of the underlying distribution function F (recall the Kiefer-Dvoretzky-Wolfowitz inequality—a K-S ball of radius $O(n^{-1/2})$ has fixed coverage probability). For linear functionals, this does not matter, but for strongly nonlinear functionals, the bootstrap estimate of the variability tends to be more accurate than the jackknife estimate of the variability.

Let us have a quick look at the distribution of the K-S distance between a continuous distribution and the empirical distribution of a sample $\{X_j\}_{j=1}^n$ iid F. The discussion follows *Feller* (1971, pp. 36ff). First we show that for continuous distributions F, the distribution of $\|\hat{F}_n - F\|_{KS}$ does not depend on F. To see this, note that $F(X_j) \sim U[0,1]$: Let $x_t \equiv \inf\{x \in \mathbb{R} : F(x_t) = t\}$. Continuity of F ensures that x_t exists for all $t \in [0,1]$. Now the event $\{X_j \le x_t\}$ is equivalent to the event $\{F(X_j) \le F(x_t)\}$ up to a set of F-measure zero. Thus

$$t = \mathbb{P}_F\{X_j \le x_t\} = \mathbb{P}_F\{F(X_j) \le F(x_t)\} = \mathbb{P}_F\{F(X_j) \le t\}, \ t \in [0,1]; \tag{37}$$

i.e., $\{F(X_j)\}_{j=1}^n$ are iid $U[0,1]$. Let

$$\hat{G}_n(t) \equiv \#\{F(X_j) \le t\}/n = \#\{X_j \le x_t\}/n = \hat{F}_n(x_t) \tag{38}$$

be the empirical cdf of $\{F(X_j)\}_{j=1}^n$. Note that

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{t \in [0,1]} |\hat{F}_n(x_t) - F(x_t)| = \sup_{t \in [0,1]} |\hat{G}_n(t) - t|. \tag{39}$$

The probability distribution of $\hat{G}_n$ is that of the cdf of n iid $U[0,1]$ random variables (it does not depend on F), so the distribution of the K-S distance between the empirical cdf and the true cdf is the same for every continuous distribution. It turns out that for distributions with atoms, the K-S distance between the empirical and the true distribution functions is stochastically smaller than it is for continuous distributions.

1.3 Bootstrap and Randomization Tests

This section is about Romano's papers. The set-up is as follows: We observe $\{X_j\}_{j=1}^n$ iid P, where P is a distribution on an abstract sample space $\mathcal{X}$. The distribution $P \in \Omega$, where Ω is a known collection of distributions on $\mathcal{X}$. The null hypothesis is that $P \in \Omega_0 \subset \Omega$. We assume that Ω_0 can be characterized as a set of distributions that are invariant under a transformation on Ω: let $\tau : \Omega \to \Omega_0$; we assume that $\tau(P) = P$ for all $P \in \Omega_0$.

Let $\mathcal{V}$ be a collection of subsets of a set $\mathcal{X}$. For a finite set $D \subset \mathcal{X}$, let $\Delta^{\mathcal{V}}(D)$ be the number of distinct sets $\{V \cap D : V \in \mathcal{V}\}$. For positive integers n, let

$$m^{\mathcal{V}}(n) = \max_{D \subset \mathcal{X} : \#D = n} \Delta^{\mathcal{V}}(D). \tag{40}$$

Let

$$c(\mathcal{V}) \equiv \inf\{n : m^{\mathcal{V}}(n) < 2^n\}. \tag{41}$$

If $c(\mathcal{V}) < \infty$, $\mathcal{V}$ is a Vapnik-Cervonenkis (V-C) class. That is, $\mathcal{V}$ is a V-C class if the maximum number of distinct intersections of sets in $\mathcal{V}$ with sets containing n points grows sub-exponentially with n. Intersections, finite unions, and Cartesian products of V-C classes are V-C classes. In $\mathbb{R}^n$, the set of all ellipsoids, the set of all half-spaces, the set of all lower-left quadrants, and the set of all convex sets with at most p extreme points are all V-C classes.

An alternative, equivalent definition of a V-C class is based on the following definition:

Definition 1 *Suppose $\mathcal{V}$ is a collection of subsets of a set $\mathcal{X}$, and that D is a finite subset of $\mathcal{X}$. We say D is* shattered *by $\mathcal{V}$ if every subset $d \subset D$ can be written $d = V \cap D$ for some $V \in \mathcal{V}$.*

Suppose D has n elements. Because there are 2^n subsets of a set with n elements, this is equivalent to saying that there are 2^n different subsets of the form $D \cap V$ as V ranges over $\mathcal{V}$.

A collection $\mathcal{V}$ is a V-C class if for some finite integer n, there exists a set $D \subset \mathcal{X}$ with n elements that is not shattered by $\mathcal{V}$.

Example. Half lines on $\mathbb{R}$. Consider a set $D = \{x_j\}_{j=1}^n$ of points on the real line. Let $\mathcal{V} = \{(-\infty, y] : y \in \mathbb{R}\}$. How many sets are there of the form $V \cap D$, for $V \in \mathcal{V}$? Just $n + 1$. Suppose the points are in increasing order, so that $x_1 < x_2 < \cdots < x_n$. Then the possibilities for $V \cap D$ are $\{\}$, $\{x_1\}$, $\{x_1, x_2\}$, …, $\{x_j\}_{j=1}^n$. Thus $m^{\mathcal{V}}(n) = n + 1$, and $c(\mathcal{V}) \equiv \inf\{n : m^{\mathcal{V}}(n) < 2^n\} = 2$ (for $n = 0$, we have $0 + 1 = 2^0$, and for $n = 1$, we have $1 + 1 = 2^1$, but for $n = 2$, we have $2 + 1 < 2^2$).

Example. Closed intervals $\{[y, z] \colon\ y < z\}$ on $\mathbb{R}$. For finite sets D as discussed above, the possibilities for $V \cap D$ include all sets of adjacent values, such as $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_1, x_2\}$, $\{x_2, x_3\}$, and $\{x_1, x_2, x_3\}$, but not, for example, $\{x_1, x_3\}$. Clearly, $m^{\mathcal{V}}(2) = 4$ but $m^{\mathcal{V}}(3) = 7$, so $c(\mathcal{V}) = 3$. (The general rule is $\binom{m^{\mathcal{V}}(n)=1+n+n}{2}$. Why?)

Suppose that $\mathcal{V}$ and $\mathcal{W}$ are V-C classes on a common set $\mathcal{X}$. Then $\mathcal{V} \cup \mathcal{W}$ is also a V-C class, as is $\mathcal{V} \cap \mathcal{W}$.

Exercise. Show that intersections and finite unions of V-C classes are V-C classes. Show by example that a countable union of V-C classes need not be a V-C class.

We return now to the approach Romano advocates for testing hypotheses. Let $\mathcal{V}$ be a VC class of subsets of $\mathcal{S}$. Define the pseudo-metric

$$\begin{aligned} \delta : \Omega \times \Omega &\rightarrow \mathbb{R}^+ \\ (P, Q) &\rightarrow \sup_{V \in \mathcal{V}} |P(V) - Q(V)|. \end{aligned} \tag{42}$$

This is a generalization of the Kolmogorov-Smirnov distance for distributions on the line. In that case, the sets in $\mathcal{V}$ are the half-lines $\{(-\infty, y] : y \in \mathbb{R}\}$ (which comprise a V-C class).

Assume that $\mathcal{V}$ and τ have been selected such that $\delta(P, \tau P) = 0$ iff $P \in \Omega_0$. Romano proposes using the test statistic

$$T_n = n^{1/2} \delta(\hat{P}_n, \tau \hat{P}_n), \tag{43}$$

where $\hat{P}_n$ is the empirical measure of $\{X_j\}_{j=1}^n$. One rejects the hypothesis when $\tau \hat{P}_n$ is far from $\hat{P}_n$; i.e., when T_n is sufficiently large.

But how large? One way to obtain a critical value for the test is with the bootstrap: resample from $\tau(\hat{P}_n)$, tabulate the distribution of the distance between the empirical distribution of the bootstrap samples and τ applied to them, use the $1-\alpha$ quantile of that distribution as the critical value for an approximate level α test. (We have to resample from $\tau(\hat{P}_n)$ rather than $\hat{P}_n$ because the significance level is computed under the assumption that the null hypothesis is *true*. The null hypothesis is true for $\tau(\hat{P}_n)$ but not necessarily for $\hat{P}_n$.)

Suppose that there is a (known) group $\mathcal{G}_n$ of transformations of the sample space $\mathcal{S}_n$ such that under the null hypothesis, P is invariant under $\mathcal{G}_n$. Then we can also construct a *randomization test* of the hypothesis H_0. For simplicity, suppose that $\mathcal{G}_n$ is finite, with M_n elements $\{g_{nj}\}_{j=1}^{M_n}$. Under the null hypothesis, conditional on $X = x$, the values $\{g_{nj}x\}_{j=1}^{M_n}$ are equally likely.[1] Compute the test statistic for each $g_{nj}x$ in the orbit of x. Reject the null hypothesis if the statistic for x exceeds the $1-\alpha$ quantile of the test statistic for the set of values obtained from the orbit; do not reject if it is less; reject with a given probability if the statistic equals the $1-\alpha$ quantile, in such a way as to get a level α test. This is a randomization test. Because the level of the randomization test is α, conditional on the data, integrating over the distribution of the data shows that it is α unconditionally.

1.3.1 Examples of hypotheses and functions τ

Examples Romano gives include testing for independence of the components of each X_j, testing for exchangeability of the components of each X_j, testing for spherical symmetry of the distribution of X_j, testing for homogeneity among the X_j, and testing for a change point.

In the example of testing for independence, the mapping τ takes the marginal distributions of the joint distribution, then constructs a joint distribution that is the product of the marginals. For distributions with independent components, this is the identity; otherwise, it maps a distribution into one with the same marginals, but whose components are independent. For testing for spherical symmetry, τ maps a distribution into one with the same mass at every distance from the origin, but that is uniform on spherical shells. For testing for exchangability, Romano proposes looking at the largest difference between P and a permutation of the coordinates of P, over all permutations of the coordinates. See his paper for more details.

[1] The *orbit* of an point x in a space S acted on by a group $\mathcal{G}$ is the set of all elements of S that can be obtained by applying elements of $\mathcal{G}$ to x. That is, it is the set $\{g(x) : g \in \mathcal{G}\}$. For example, consider points in the plane and the group of rotations about the origin. Then the orbit of a point x is the circle with radius $\|x\|$.

Romano shows that these tests are consistent against all alternatives, and that the critical values given by the bootstrap and by randomization are asymptotically equal with probability one. Because the randomization tests are exact level α tests, they might be preferred. Romano also briefly discusses how to implement the tests computationally.

Let's consider the implementation in more detail, for two hypotheses: independence of the components of a k-variate distribution, and rotational invariance of a bivariate distribution.

1.3.2 Independence

We observe $\{X_j\}_{j=1}^n$ iid P, where each $X_j = (X_{ij})_{i=1}^k$ takes values in $\mathbb{R}^k$. Under the null hypothesis, P is invariant under the mapping τ that takes the k marginal distributions of P and multiplies them together to give a probability on $\mathbb{R}^k$ with independent components. Let $\hat{P}_n$ be the empirical measure; let the V-C class $\mathcal{V}$ be the set of lower left quadrants $\{Q(x) : x \in \mathbb{R}^k\}$ where

$$Q(x) \equiv \{y \in \mathbb{R}^k : y_i \le x_i, \;\; i = 1, \ldots, k\}. \tag{44}$$

Then

$$\hat{P}_n(Q(x)) = \frac{1}{n}\#\{X_j : X_{ij} \le x_i, \;\; i = 1, \ldots, k\}, \tag{45}$$

and

$$\tau\hat{P}_n(Q(x)) = \prod_{i=1}^{k} \frac{1}{n}\#\{X_j : X_{ij} \le x_i\}. \tag{46}$$

The maximum difference in the probability of a lower left quadrant $Q(x)$ occurs when x is one of the points of support of $\tau\hat{P}_n$:

$$\begin{aligned}\sup_{V\in\mathcal{V}} |\hat{P}_n(V) - \tau\hat{P}_n(V)| &= \sup_{x\in\mathbb{R}^k} |\hat{P}_n(Q(x)) - \tau\hat{P}_n(Q(x))| \\ &= \max_{x\in\mathbb{R}^k : x_i \in \{X_{ij}\}_{j=1}^n, \; i=1,\ldots,k} |\hat{P}_n(Q(x)) - \tau\hat{P}_n(Q(x))|. \end{aligned} \tag{47}$$

The probability of a lower left quadrant is straightforward to compute for $\hat{P}_n$ and for $\tau\hat{P}_n$; here is Matlab code. Let X be an n by k matrix whose rows are the observations $\{X_j\}_{j=1}^n$, and let $x = (x_i)_{i=1}^k$ be a row vector.

```
    temp = X <= (ones(n,1)*x);

    phatn = sum(prod(temp, 2))/n;

    tauphatn = prod(sum(temp))/(n^k);
```

We can simulate a sample of size n iid $\tau\hat{P}_n$ in Matlab as follows:

```
[n k] = size(X);
tausam = zeros(size(X));
for i=1:k,
    tausam(:,i) = X(ceil(n*rand(n,1)),i);
end;
```

To test the null hypothesis of independence, we would compute

$$T(X) = \max_{x \in \mathbb{R}^k : x_i \in \{X_{ij}\}_{j=1}^n,\ \ i=1,\ldots,k} |\hat{P}_n(Q(x)) - \tau\hat{P}_n(Q(x))| \tag{48}$$

from the data X, then repeatedly draw iid samples X^* of size n from $\tau\hat{P}_n$, computing

$$T(X^*) = \max_{x \in \mathbb{R}^k : x_i \in \{X^*_{ij}\}_{j=1}^n,\ \ i=1,\ldots,k} |\hat{P}^*_n(Q(x)) - \tau\hat{P}^*_n(Q(x))| \tag{49}$$

for each. We would reject the null hypothesis that the components of P are independent (at approximate significance level α) if $T(X)$ exceeds the $1-\alpha$ quantile of the simulated distribution of $T(X^*)$.

1.3.3 Rotational invariance in $\mathbb{R}^2$

We observe $\{X_j\}_{j=1}^n$ iid P, where each $X_j = (X_{1j}, X_{2j})$ takes values in $\mathbb{R}^2$. For $y \in \mathbb{R}^2$, define $|y| \equiv \sqrt{y_1^2 + y_2^2}$ to be the distance from y to the origin. Except at the origin, the mapping from Cartesian coordinates (x_1, x_2) to polar coordinates (r, θ) is one-to-one; identify the origin with the polar coordinates $(0, 0)$. Under the null hypothesis, P is invariant under the mapping τ that produces a distribution with the same marginal distribution of $|X|$ but that is uniform on θ for each possible value of $|X|$.

As before, let $\hat{P}_n$ be the empirical measure; let the V-C class $\mathcal{V}$ be the set of lower left quadrants $\{Q(x) : x \in \mathbb{R}^2\}$ where

$$Q(x) \equiv \{y \in \mathbb{R}^k : y_i \le x_i,\ \ i = 1, 2\}. \tag{50}$$

Then

$$\hat{P}_n(Q(x)) = \frac{1}{n}\#\{X_j : X_{ij} \le x_i,\ \ i = 1, 2\}. \tag{51}$$

To proceed, we need to find the probability of lower left quadrants $Q(x)$ for the distribution $\tau\hat{P}_n$. Consider the contribution from each X_j separately. Let $R_j = |X_j| = \sqrt{X_{1j}^2 + X_{2j}^2}$. The contribution of X_j to $\tau \hat{P}_n(Q(x))$ is $1/n$ times the fraction of the circle $\{y \in \mathbb{R}^2 : |y| = R_j\}$ that is in the quadrant $Q(x)$. There eight cases to consider:

1. $x_1^2 + x_2^2 > R_j^2$, $x_1,\ x_2 < 0$ or $x_1 < -R_j$ or $x_2 < -R_j$. The contribution is 0: the quadrant does not intersect the circle $|y| = R_j$.

2. $x_1, x_2 > R_j$. The contribution is $1/n$: the quadrant contains the entire circle $|y| = R_j$.

3. $x_1^2 + x_2^2 \leq R_j^2$. The quadrant includes an arc that is at most half the circle. Let the points at which the quadrant boundary intersects the circle be (x_1', x_2) and (x_1, x_2'). Then x_1' is the negative root of $x_1'^2 = R_j^2 - x_2^2$ and x_2' is the negative root of $x_2'^2 = R_j^2 - x_1^2$. The fraction of the circle included in $Q(x)$ is
$$\frac{1}{\pi} \sin^{-1} \frac{1}{\sqrt{2}} \left(1 + \frac{x_1}{R_j} \sqrt{1 - \frac{x_2^2}{R_j^2}} + \frac{x_2}{R_j} \sqrt{1 - \frac{x_1^2}{R_j^2}} \right)^{1/2}. \tag{52}$$

4. $x_1^2 + x_2^2 > R_j^2$, $-R_j < x_1 \leq 0$, $x_2 \geq 0$. The fraction of the circle within $Q(x)$ is
$$q(x_1) \equiv \frac{1}{\pi} \sin^{-1} \frac{1}{\sqrt{2}} \left(1 - \frac{x_1^2}{R_j^2} \right)^{1/2}. \tag{53}$$

5. $x_1^2 + x_2^2 > R_j^2$, $0 \leq x_1 < R_j$, $x_2 \geq R_j$. The fraction of the circle within $Q(x)$ is $1 - q(x_1)$.

6. $x_1^2 + x_2^2 > R_j^2$, $x_1 \geq 0$, $-R_j < x_2 < 0$. The fraction of the circle within $Q(x)$ is $q(x_2)$.

7. $x_1^2 + x_2^2 > R_j^2$, $x_1 \geq R_j$, $0 \leq x_2 < R_j$. The fraction of the circle within $Q(x)$ is $1 - q(x_2)$.

8. $x_1^2 + x_2^2 > R_j^2$, $0 \leq x_1 < R_j$, $0 \leq x_2 < R_j$. The fraction of the circle within $Q(x)$ is $1 - q(x_1) - q(x_2)$.

At which points x should we evaluate the discrepancy $D(x) = |\hat{P}_n(Q(x)) - \tau \hat{P}_n(Q(x))|$? Let $R = \max_j R_j$. Then for $x_1, x_2 > R$, $D(x) = 0$. Similarly, for $x_1, x_2 < -R$, $D(x) = 0$. We might take x on a fine grid in the square $[-R, R] \times [-R, R]$, but this is wasteful. Some thought shows that the maximum discrepancy occurs when some datum is just included in $Q(x)$, which makes $\hat{P}_n$ relatively large compared with $\tau \hat{P}_n$, or when some datum is just excluded from $Q(x)$, which makes $\tau \hat{P}_n$ relatively large compared with $\hat{P}_n$. The possible points of maximum discrepancy are x of the form $(X_{1j} - s\epsilon, X_{2k} - s\epsilon)$ with $1 \leq j, k \leq n$, $s \in \{0, 1\}$, and ϵ small, together with the points $(X_{1j} - s\epsilon, R)$ and $(R, X_{2j} - s\epsilon)$. This is a large $(2n^2 + 4n)$ but finite number of points. Denote this set by $\mathcal{X}(\{X_j\}, \epsilon)$.

To draw an iid sample of size n from $\tau \hat{P}_n$, we draw n values iid uniform on $\{r_j\}_{j=1}^n$ and draw n iid $U[0, 2\pi]$ random variables, and treat these as the polar coordinates (r, θ) of n points in $\mathbb{R}^2$.

To test the null hypothesis of rotational invariance, we would compute

$$T(X) = \max_{x \in \mathbb{R}^k : x \in \mathcal{X}(\{X_j\}, \epsilon)} |\hat{P}_n(Q(x)) - \tau \hat{P}_n(Q(x))| \tag{54}$$

from the data X, then repeatedly draw iid samples $\{X_j^*\}$ of size n from $\tau \hat{P}_n$, computing

$$T(X^*) = \max_{x \in \mathbb{R}^k : x \in \mathcal{X}(\{X_j^*\}, \epsilon)} |\hat{P}_n^*(Q(x)) - \tau \hat{P}_n^*(Q(x))| \tag{55}$$

for each. We would reject the null hypothesis that P is rotationally invariant (at approximate significance level α) if $T(X)$ exceeds the $1 - \alpha$ quantile of the simulated distribution of $T(X^*)$.

Under the null hypothesis, the distribution of the data is invariant under the action of the rotation group. This is not a finite group, so we cannot exhaust the set of transformations on a computer. However, we might consider the subgroup of rotations by multiples of $2\pi/M$ for some large integer M. We could get an alternative approximate level α test of the hypothesis of rotational invariance by comparing $T(X)$ with the $1 - \alpha$ quantile of T over all such rotations of the data—the orbit of the data under this finite subgroup.

1.4 Bootstrap Confidence Sets

Let $\mathcal{U}$ be an index set (not necessarily countable). Recall that a collection $\{\mathcal{I}_u\}_{u \in \mathcal{U}}$ of confidence intervals for parameters $\{\theta_u\}_{u \in \mathcal{U}}$ has simultaneous $1 - \alpha$ coverage probability if

$$\mathbb{P}_\theta \{\cap_{u \in \mathcal{U}} \{\mathcal{I}_u \ni \theta_u\}\} \geq 1 - \alpha. \tag{56}$$

If $\mathbb{P}\{\mathcal{I}_u \ni \theta_u\}$ does not depend on u, the confidence intervals are said to be *balanced.*

Many of the procedures for forming joint confidence sets we have seen depend on *pivots*, which are functions of the data and the parameter(s) whose distribution is known (even though the parameter and the parent distribution are not). For example, the Scheffé method relies on the fact that (for samples from a multivariate Gaussian with independent components) the sum of squared differences between the data and the corresponding parameters, divided by the variance estimate, has an F distribution, regardless of the parameter values. Similarly, Tukey's maximum modulus method relies on the fact that (again, for independent Gaussian data) the distribution of the maximum of the studentized absolute differences between the data and the corresponding parameters does not depend on the parameters. Both of those examples are parametric, but the idea is more general: the procedure we looked at for finding bounds on the density function subject to shape restrictions just relied on the fact that there are uniform bounds on the probability that the K-S distance between the empirical distribution and the true distribution exceeds some threshold.

Even in cases where there is no known exact pivot, one can sometimes show that some function of the data and parameters is asymptotically a pivot. Working out the distributions of the functions involved is not typically straightforward, and a general method of constructing (possibly simultaneous) confidence sets would be nice.

Efron gives several methods of basing confidence sets on the bootstrap. Those methods are substantially improved (in theory, and in my experience) by Beran's pre-pivoting approach, which leads to iterating the bootstrap.

Let X_n denote a sample of size n from F. Let $R_n(\theta) = R_n(X_n, \theta)$ have cdf H_n, and let $H_n^{-1}(\alpha)$ be the largest α quantile of the distribution of R_n. Then

$$\{\gamma \in \Theta : R_n(\gamma) \le H_n^{-1}(1-\alpha)\} \tag{57}$$

is a $1-\alpha$ confidence set for θ.

1.4.1 The Percentile Method

The idea of the percentile method is to use the empirical bootstrap percentiles of some quantity to approximate the true percentiles. Consider constructing a confidence interval for a single real parameter $\theta = T(F)$. We will estimate θ by $\hat{\theta} = T(\hat{F}_n)$. We would like to know the distribution function $H_n = H_n(\cdot, F)$ of $D_n(\theta) = T(\hat{F}_n) - \theta$. Suppose we did. Let $H_n^{-1}(\cdot) = H_n^{-1}(\cdot, F)$ be the inverse cdf of D_n. Then

$$\mathbb{P}_F\{H_n^{-1}(\alpha/2) \le T(\hat{F}_n) - \theta \le H_n^{-1}(1-\alpha/2)\} = 1-\alpha, \tag{58}$$

so

$$\mathbb{P}_F\{\theta \le T(\hat{F}_n) - H_n^{-1}(\alpha/2) \text{ and } \theta \ge T(\hat{F}_n) - H_n^{-1}(1-\alpha/2)\} = 1-\alpha, \tag{59}$$

or, equivalently,

$$\mathbb{P}_F\{[T(\hat{F}_n) - H_n^{-1}(1-\alpha/2), T(\hat{F}_n) - H_n^{-1}(\alpha/2)] \ni \theta\} = 1-\alpha, \tag{60}$$

so the interval $[T(\hat{F}_n) - H_n^{-1}(1-\alpha/2), T(\hat{F}_n) - H_n^{-1}(\alpha/2)]$ would be a $1-\alpha$ confidence interval for θ.

The idea behind the percentile method is to approximate $H_n(\cdot, F)$ by $\hat{H}_n = H_n(\cdot, \hat{F}_n)$, the distribution of D_n under resampling from $\hat{F}_n$ rather than F. An alternative approach would be to take $D_n(\theta) = |T(\hat{F}_n) - \theta|$; then

$$\mathbb{P}_F\{|T(\hat{F}_n) - \theta| \le H_n^{-1}(1-\alpha)\} = 1 - \alpha, \tag{61}$$

so

$$\mathbb{P}_F\{[T(\hat{F}_n - H_n^{-1}(1-\alpha), T(\hat{F}_n + H_n^{-1}(1-\alpha)] \ni \theta\} = 1 - \alpha. \tag{62}$$

In either case, the "raw" bootstrap approach is to approximate H_n by resampling under $\hat{F}_n$.

Beran proves a variety of results under the following condition:

Condition 1. (Beran, 1987) For any sequence $\{F_n\}$ that converges to F in a metric d on cdfs, $H_n(\cdot, F_n)$ converges weakly to a continuous cdf $H = H(\cdot, F)$ that depends only on F, and not the sequence $\{F_n\}$.

Suppose Condition 1 holds. Then because $\hat{F}_n$ is consistent for F, the estimate $\hat{H}_n$ converges in probability to H in sup norm; moreover, the distribution of $\hat{H}_n(R_n(\theta))$ converges to $U[0, 1]$.

Instead of D_n, consider $R_n(\theta) = |T(\hat{F}_n) - \theta|$ or some other (approximate) pivot. Let $\hat{H}_n(\cdot, \hat{F}_n)$ be the bootstrap estimate of the cdf of R_n; The set

$$\begin{aligned} B_n &= \{\gamma \in \Theta : \hat{H}_n(R_n(\gamma)) \le 1 - \alpha\} \\ &= \{\gamma \in \Theta : R_n(\gamma) \le \hat{H}_n^{-1}(1-\alpha)\} \end{aligned} \tag{63}$$

is (asymptotically) a $1 - \alpha$ confidence set for θ.

The level of this set for finite samples tends to be inaccurate. It can be improved in the following way, due to Beran.

The original root, $R_n(\theta)$, whose limiting distribution depends on F, was transformed into a new root $R_{n,1}(\theta) = \hat{H}_n(R_n(\theta))$, whose limiting distribution is $U[0, 1]$. The distribution of $R_{n,1}$ depends less strongly on F than does that of R_n; Beran calls mapping R_n into $R_{n,1}$ *prepivoting*. The confidence set 63 acts as if the distribution of $R_{n,1}$ really is uniform, which is not generally true. One could instead treat $R_{n,1}$ itself as a root, and pivot to reduce the dependence on F.

Let $H_{n,1} = H_{n,1}(\cdot, F)$ be the cdf of the new root $R_{n,1}(\theta)$, estimate $H_{n,1}$ by $\hat{H}_{n,1} = H_{n,1}(\cdot, \hat{F}_n)$, and define

$$\begin{aligned} B_{n,1} &= \{\gamma \in \Theta : \hat{H}_{n,1}(R_{n,1}(\gamma)) \le 1 - \alpha\} \\ &= \{\gamma \in \Theta : \hat{H}_{n,1}(\hat{H}_n(R_n(\gamma))) \le 1 - \alpha\} \\ &= \{\gamma \in \Theta : R_n(\gamma) \le \hat{H}_n^{-1}(\hat{H}_{n,1}^{-1}(1-\alpha))\}. \end{aligned} \tag{64}$$

Beran shows that this confidence set tends to have smaller error in its level than does B_n. The transformation can be iterated further, typically resulting in additional reductions in the level error.

1.5 Approximating $B_{n,1}$ by Monte Carlo

I'll follow Beran's (1987) notation (mostly).

Let x_n denote the "real" sample of size n. Let x_n^* be a bootstrap sample of size n drawn from the empirical cdf $\hat{F}_n$. The components of x_n^* are conditionally iid given x_n. Let $\hat{F}_n^*$ denote the "empirical" cdf of the bootstrap sample x_n^*. Let x_n^{**} denote a sample of size n drawn from $\hat{F}_n^*$; the components of x_n^{**} are conditionally iid given x_n and x_n^*. Let $\hat{\theta}_n = T(\hat{F}_n)$, and $\hat{\theta}_n^* = T(\hat{F}_n^*)$. Then

$$H_n(s, F) = \mathbb{P}_F\{R_n(x_n, \theta) \le s\}, \tag{65}$$

and

$$H_{n,1}(s, F) = \mathbb{P}_F\left\{\mathbb{P}_{\hat{F}_n}\{R_n(x_n^*, \hat{\theta}_n) < R_n(x_n, \theta)\} \le s\right\}. \tag{66}$$

The bootstrap estimates of these cdfs are

$$\hat{H}_n(s) = H_n(s, \hat{F}_n) = \mathbb{P}_{\hat{F}_n}\{R_n(x_n^*, \hat{\theta}_n) \le s\}, \tag{67}$$

and

$$\hat{H}_{n,1}(s) = H_{n,1}(s, \hat{F}_n) = \mathbb{P}_{\hat{F}_n}\left\{\mathbb{P}_{\hat{F}_n^*}\{R_n(x_n^{**}, \hat{\theta}_n^*) < R_n(x_n^*, \hat{\theta}_n)\} \le s\right\}. \tag{68}$$

The Monte Carlo approach is as follows:

1. Draw $\{y_k^*\}_{k=1}^M$ bootstrap samples of size n from $\hat{F}_n$. The ecdf of $\{R_n(y_k^*, \hat{\theta}_n)\}_{k=1}^M$ is an approximation to $\hat{H}_n$.

2. For $k = 1, \cdots, M$, let $\{y_{k\ell}^{**}\}_{\ell=1}^N$ be N size n bootstrap samples from the ecdf of y_k^*. Let $\hat{\theta}_{n,k}^* = T(\hat{F}_{n,k}^*)$. Let Z_k be the fraction of the values

$$\{R_n(y_{k,\ell}^{**}, \hat{\theta}_{n,k}^*)\}_{\ell=1}^N \tag{69}$$

that are less than or equal to $R_n(y_k^*, \hat{\theta}_n)$. The ecdf of $\{Z_k\}$ is an approximation to $\hat{H}_{n,1}$ that improves (in probability) as M and N grow.

Note that this approach is extremely general. Beran gives examples for confidence sets for directions, *etc.* The pivot can in principle be a function of any number of parameters, which can yield simultaneous confidence sets for parameters of any dimension.

1.6 Other approaches to improving coverage probability

There are other ways of iterating the bootstrap to improve the level accuracy of bootstrap confidence sets. Efron suggests trying to attain a different coverage probability so that the coverage attained in the second generation samples is the nominal coverage probability. That is, if one wants a 95% confidence set, one tries different percentiles so that in resampling from the sample, the attained coverage probability is 95%. Typically, the percentile one uses in the second generation will be higher than 95%. Here is a sketch of the Monte-Carlo approach:

- Set a value of α^* (initially taking $\alpha^* = \alpha$ is reasonable)
- From the sample, draw M size-n samples that are each iid $\hat{F}_n$. Denote the ecdfs of the samples by $\{\hat{F}^*_{n,j}\}$.
- For each $j = 1, \ldots, M$, apply the percentile method to make a (nominal) level $1-\alpha^*$ confidence interval for $T(\hat{F}_n)$. This gives M confidence intervals; a fraction $1 - \alpha'$ will cover $T(\hat{F}_n)$. Typically, $1 - \alpha' \neq 1 - \alpha$.
- If $1 - \alpha' < 1 - \alpha$, decrease α^* and return to the previous step. If $1 - \alpha' > 1 - \alpha$, increase α^* and return to the previous step. If $1 - \alpha' \approx 1 - \alpha$ to the desired level of precision, go to the next step.
- Report as a $1 - \alpha$ confidence interval for $T(F)$ the (first generation) bootstrap quantile confidence interval that has nominal $1 - \alpha^*$ coverage probability.

An alternative approach to increasing coverage probability by iterating the bootstrap is to use the same root, but to use a quantile (among second-generation bootstrap samples) of its $1 - \alpha$ quantile rather than the quantile observed in the first generation. The heuristic justification is that we would ideally like to know the $1 - \alpha$ quantile of the pivot under sampling from the true distribution F. We don't. The percentile method estimates the $1 - \alpha$ quantile of the pivot under F by the $1 - \alpha$ quantile of the pivot under $\hat{F}_n$, but this is subject to sampling variability. To try to be conservative, we could use the bootstrap a second time find an (approximate) upper $1 - \alpha^*$ confidence interval for the $1 - \alpha$ quantile of the pivot.

Here is a sketch of the Monte-Carlo approach:

- Pick a value $\alpha^* \in (0, 1/2)$ (*e.g.*, $\alpha^* = \alpha$). This is a tuning parameter.

- From the sample, draw M size-n samples that are each iid $\hat{F}_n$. Denote the ecdfs of the samples by $\{\hat{F}^*_{n,j}\}$.

- For each $j = 1, \ldots, M$, draw N size-n samples, each iid $\hat{F}_{n,j}$. Find the $1 - \alpha$ quantile of the pivot. This gives M values of the $1 - \alpha$ quantile. Let c be the $1 - \alpha^*$ quantile of the M $1 - \alpha$ quantiles.

- Report as a $1 - \alpha$ confidence interval for $T(F)$ the interval one gets by taking c to be the estimate of the $1 - \alpha$ quantile of the pivot.

In a variety of simulations, this tends to be more conservative than Beran's method, and more often attains at least the nominal coverage probability.

Exercise. Consider forming a two-sided 95% confidence interval for the mean θ of a distribution F based on the sample mean, using $|\bar{X} - \theta|$ as a pivot. Implement the three "double-bootstrap" approaches to finding a confidence interval (Beran's pre-pivoting, Efron's calibrated target percentile, and the percentile-of-percentile). Generate 100 synthetic samples of size 100 from the following distributions: normal, lognormal, Cauchy, mixtures of normals with the same mean but quite different variances (try different mixture coefficients), and mixtures of normals with different means and different variances (the means should differ enough that the result is bimodal). Apply the three double bootstrap methods to each, resampling 1000 times from each of 1000 first-generation bootstrap samples. Which method on the average has the lowest level error? Which method tends to be most conservative? Try to provide some intuition about the circumstances under which each method fails, and the circumstances under which each method would be expected to perform well. How do you interpret coverage for the Cauchy? **Warning:** You might need to be clever in how you implement this to make it a feasible calculation in S or Matlab. If you try to store all the intermediate results, the memory requirement is huge. On the other hand, if you use too many loops, the execution time will be long.

1.7 Bootstrap confidence sets based on Stein (shrinkage) estimates

Beran (1995) discusses finding a confidence region for the mean vector $\theta \in \mathbb{R}^q$, $q \geq 3$, from data $X \sim N(\theta, I)$. This is an example illustrating that *what* one bootstraps is important, and that naive plug-in bootstrapping doesn't always work.

The sets are spheres centered at the shrinkage estimate

$$\hat{\theta}_S = \left(1 - \frac{q-2}{\|X\|^2}\right) X, \tag{70}$$

with random diameter $\hat{d}$. That is, the confidence sets C are of the form

$$C(\hat{\theta}_S, \hat{d}) = \left\{\gamma \in \mathbb{R}^q : \|\hat{\theta}_S - \gamma\| \leq \hat{d}\right\}. \tag{71}$$

The problem is how to find $\hat{d} = \hat{d}(X; \alpha)$ such that

$$\mathbb{P}_\gamma\{C(\hat{\theta}_S, \hat{d}) \ni \gamma\} \geq 1 - \alpha \tag{72}$$

whatever be $\gamma \in \mathbb{R}^q$.

This problem is parametric: F is known up to the q-dimensional mean vector θ. We can thus use a "parametric bootstrap" to generate data that are approximately from F, instead of drawing directly from $\hat{F}_n$: if we have an estimate $\hat{\theta}$ of θ, we can generate artificial data distributed as $N(\hat{\theta}, I)$. If $\hat{\theta}$ is a good estimator, the artificial data will be distributed nearly as F. The issue is in what sense $\hat{\theta}$ needs to be good.

Beran shows (somewhat surprisingly) that resampling from $N(\hat{\theta}_S, I)$ or from $N(X, I)$ do not tend to work well in calibrating $\hat{d}$. The crucial thing in using the bootstrap to calibrate the radius of the confidence sphere seems to be to estimate $\|\theta\|$ well.

Definition 2 *The* geometrical risk *of a confidence set C for the parameter $\theta \in \mathbb{R}^q$ is*

$$G_q(C, \theta) \equiv q^{-1/2} E_\theta \sup_{\gamma \in C} \|\gamma - \theta\|. \tag{73}$$

That is, the geometrical risk is the expected distance to the parameter from the most distant point in the confidence set.

For confidence spheres

$$C = C(\hat{\theta}, \hat{d}) = \{\gamma \in \mathbb{R}^q : \|\gamma - \hat{\theta}\| \leq \hat{d}\}, \tag{74}$$

the geometrical risk can be decomposed further: the distance from θ to the most distant point in the confidence set is the distance from θ to the center of the sphere, plus the radius of the sphere, so

$$\begin{aligned} G_q(C(\hat{\theta}, \hat{r}), \theta) &= q^{-1/2} E_\theta \left(\|\hat{\theta} - \theta\| + \hat{d}\right) \\ &= q^{-1/2} E_\theta \|\hat{\theta} - \theta\| + q^{-1/2} E_\theta \hat{d}. \end{aligned} \tag{75}$$

Lemma 1 *(Beran, 1995, Lemma 4.1). Define*

$$W_q(X, \gamma) \equiv (q^{-1/2}(\|X - \gamma\|^2 - q), q^{-1/2}\gamma'(X - \gamma). \tag{76}$$

Suppose $\{\gamma_q \in \mathbb{R}^q\}$ *is any sequence such that*

$$\frac{\|\gamma_q\|^2}{q} \to a < \infty \text{ as } q \to \infty. \tag{77}$$

Then

$$W_q(X, \gamma_q) \underset{W}{\rightarrow} (\sqrt{2}Z_1, \sqrt{a}Z_2) \tag{78}$$

under $\mathbb{P}_{\gamma_q}$*, where* Z_1 *and* Z_2 *are iid standard normal random variables. (The symbol* $\underset{W}{\rightarrow}$ *denotes weak convergence of distributions.)*

Proof. Under $\mathbb{P}_{\gamma_q}$, the distribution of $X - \gamma$ is rotationally invariant, so the distribution of the components of W_q depend on γ only through $\|\gamma\|$. Wlog, we may take each component of γ_q to be $q^{-1/2}\|\gamma_q\|$. The distribution of the first component of W_q is then that of the sum of squares of q iid standard normals (a chi-square rv with q df), minus the expected value of that sum, times $q^{-1/2}$. The standard deviation of a chi-square random variable with q df is $\sqrt{2q}$, so the first component of W_q is $\sqrt{2}$ times a standardized variable whose distribution is asymptotically (in q) normal. The second component of W_q is a linear combination of iid standard normals; by symmetry (as argued above), its distribution is that of

$$\begin{aligned} q^{-1/2}\sum_{j=1}^{q} q^{-1/2}\|\gamma_q\|Z_j &= \|\gamma_q\|\sum_{j=1}^{q} Z_j \\ &\to a^{1/2}Z_2. \end{aligned} \tag{79}$$

Recall that the squared-error risk (normalized by $q^{-1/2}$) of the James-Stein estimator is $1 - q^{-1}E_\theta\{(q-2)^2/\|X\|^2\} < 1$. The difference between the loss of $\hat{\theta}_S$ and an unbiased estimate of its risk is

$$D_q(X, \theta) = q^{-1/2}\{\|\hat{\theta}_S - \theta\|^2 - [q - (q-2)^2/\|X\|^2]\}. \tag{80}$$

By rotational invariance, the distribution of this quantity depends on θ only through $\|\theta\|$; Beran writes the distribution as $H_q(\|\theta\|^2/q)$. Beran shows that if $\{\gamma_q \in \mathbb{R}^q\}$ satisfies 77, then

$$H_q(\|\gamma_q\|^2/q) \xrightarrow[W]{} N(0, \sigma^2(a)), \tag{81}$$

where

$$\sigma^2(t) \equiv 2 - 4t/(1+t)^2 \geq 1. \tag{82}$$

Define

$$\hat{\theta}_{\mathrm{CL}} = [1 - (q-2)/\|X\|^2]_+^{1/2} X. \tag{83}$$

Theorem 1 *(Beran, 1995, Theorem 3.1) Suppose $\{\gamma_q \in \mathbb{R}^q\}$ satisfies 77. Then*

$$H_q(\|\hat{\theta}_{CL}\|^2/q) \xrightarrow[W]{} N(0, \sigma^2(a)), \tag{84}$$

$$H_q(\|X\|^2/q) \xrightarrow[W]{} N(0, \sigma^2(1+a)), \tag{85}$$

and

$$H_q(\|\hat{\theta}_S\|^2/q) \to N(0, \sigma^2(a^2/(1+a))), \tag{86}$$

all in P_{γ_q} probability.

It follows that to estimate H_q by the bootstrap consistently, one should use

$$\hat{H}_B = H_q(\|\hat{\theta}_{\mathrm{CL}}\|^2/q) \tag{87}$$

rather than estimating using either the norm of X or the norm of the James-Stein estimate $\hat{\theta}_S$ of θ.

Proof. Lemma 1 implies that under the conditions of the theorem, $\|\hat{\theta}_{\mathrm{CL}}\|^2/q \to a$, $\|X\|^2/q \to 1+a$, and $\|\hat{\theta}_S\|^2/q \to a^2/(1+a)$.