Resampling Methods
Math 6070, Spring 2013

Davar Khoshnevisan
University of Utah

April 4, 2013

Contents

1 Introduction **1**

2 Integration **3**
2.1 Stieldjes Integration 3
2.2 Some Examples 3

3 The Bootstrap **5**
3.1 A Bootstrap Confidence Interval 6
3.2 Subsampling 7
3.3 The Bootstrap for Variance 8
3.4 The Bootstrap for Quantiles 9
3.5 The Bootstrap for Skewness 11
3.6 The Bootstrap for Kurtosis 12
3.7 The Bootstrap Can Fail 14

4 The Jackknife **15**
4.1 The Jackknife, and Bias Reduction 15
4.2 The Jackknife Estimator of Variance 17
4.3 The Jackknife Can Fail 17

1 Introduction

One of the classical questions in asymptotic statistics is the problem of finding a reasonable asymptotic confidence interval for p for a population that is

distributed as Bernoulli(p). Let us recall the usual method of accomplishing this.

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from F, where F denotes the cumulative distribution function for Bernoulli(p) and $p \in (0\,,1)$ is unknown. That is,

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ q & \text{if } 0 \leqslant x < 1, \\ 1 & \text{if } x \geqslant 1, \end{cases}$$

where I am writing $q := 1 - p$ to keep with the traditional notation of this subject. In other words, $\mathrm{P}\{X_i = 1\} = p$ and $\mathrm{P}\{X_i = 0\} = q$. Note that $\bar{X} := n^{-1} \sum_{i=1}^n X_i$ is the sample proportion of ones. Since

$$\mathrm{E}(\bar{X}) = p \quad \text{and} \quad \mathrm{Var}(\bar{X}) = \frac{pq}{n},$$

the central limit theorem assures us that

$$\sqrt{n}\left(\bar{X} - p\right) \xrightarrow{d} \mathrm{N}(0\,, pq) \quad \text{as } n \to \infty.$$

In particular, when n is large, the following has probability very nearly equal to $(1 - \alpha) \times 100\%$:

$$p \in \left(\bar{X} - z_\alpha \sqrt{\frac{pq}{n}}\ ,\ \bar{X} + z_\alpha \sqrt{\frac{pq}{n}} \right), \tag{1}$$

where z_α is a non-random number that is chosen so that

$$\mathrm{P}\left\{ |\mathrm{N}(0\,,1)| \leqslant z_\alpha \right\} = 1 - \alpha.$$

However, despite first appearances, (1) is *not* an asymptotic level $(1 - \alpha) \times 100\%$ confidence interval for p, since the interval in question depends itself on p, which is sadly unknown. The most common remedy is to estimate further the unknown term pq—in (1)—by $\bar{X}(1 - \bar{X})$—which is an asymptotically-consistent estimator of pq thanks to the law of large numbers. This leads us to the following, which *is* asymptotically a level $(1 - \alpha) \times 100\%$ confidence interval for p, thanks to Slutsky's theorem:

$$p \in \left(\bar{X} - z_\alpha \sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}\ ,\ \bar{X} + z_\alpha \sqrt{\frac{\bar{X}(1 - \bar{X})}{n}} \right), \tag{2}$$

This is a typical exam of the [parametric] bootstrap: At the second stage of constructing our confidence interval we use the sample proportion $\bar{X}$ of ones in place of the population proportion p of ones. Magically, everything works more or less equally well.

2 Integration

2.1 Stieldjes Integration

Let F denote a cumulative distribution function, and let us write $f :=$ the mass function of F in the discrete case and $f :=$ the pdf of F in the absolutely continuous case. That is,

$$f(x) = \begin{cases} F(x) - F(x-) & \text{if discrete,} \\ F'(x) & \text{if absolutely continuous.} \end{cases}$$

Then we can define the "integral" $\int g\,\mathrm{d}F$ as

$$\int g\,\mathrm{d}F := \int_{-\infty}^{\infty} g(x)\,\mathrm{d}F(x) := \begin{cases} \sum_x g(x)f(x) & \text{if discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x & \text{if absolutely continuous,} \end{cases}$$

for every function g where the preceding definition makes sense and/or is finite. The integral $\int g\,\mathrm{d}F$ has all of the usual properties of an integral: Whenever $a_1, \ldots, a_n$ are real numbers and $g_1, \ldots, g_n$ are "integrable functions,"

$$\int \sum_{i=1}^{n} a_i g_i\,\mathrm{d}F = \sum_{i=1}^{n} a_i \int g_i\,\mathrm{d}F.$$

If $h(x) \geqslant 0$ for all x then $\int h\,\mathrm{d}F \geqslant 0$, we have the triangle inequality,

$$\left| \int g\,\mathrm{d}F \right| \leqslant \int |g|\,\mathrm{d}F,$$

and so forth. $\int g\,\mathrm{d}F$ is an example of what is called the *Stieldjes integral*.

2.2 Some Examples

Throughout, suppose $X, X_1, \cdots, X_n \sim F$ is an i.i.d. sample. As before, $\hat{F}_n$ denotes the empirical distribution function of $X_1, \ldots, X_n$.

Example 1 (Mean)**.** The mean of X can always be thought of as a Stieldjes integral. For instance, in the continuous case,

$$\int_{-\infty}^{\infty} x\,\mathrm{d}F(x) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = \mathrm{E}(X),$$

and

$$\int_{-\infty}^{\infty} x\,\mathrm{d}F(x) = \sum_x xf(x) = \mathrm{E}(X),$$

in the discrete case. Similarly, the mean of the random cdf $\hat{F}_n$ is

$$\int_{-\infty}^{\infty} x \, \mathrm{d}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

In somewhat loose words, the population mean is the mean of F and the sample mean is the mean of $\hat{F}_n$. □

Example 2 (Variance)**.** For our second elementary example, let us note that if $\mu := \mathrm{E}(X)$, then

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \, \mathrm{d}F(x).$$

The proof is similar to that of the previous example. Likewise, the variance of $\hat{F}_n$ is

$$\int_{-\infty}^{\infty} (x - \bar{X})^2 \, \mathrm{d}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

which is the [biased] sample variance. □

Example 3 (Median)**.** Recall that a *median* of X is any number m such that

$$\mathrm{P}\{X \leqslant m\} = F(m) \geqslant \frac{1}{2} \quad \text{and} \quad \mathrm{P}\{X \geqslant m\} = 1 - F(m-) \geqslant \frac{1}{2}.$$

Define the *left-continuous inverse* F^{-1} *of* F as follows:

$$F^{-1}(z) := \min\{x : \, F(x) \geqslant z\}.$$

Then, the function F^{-1} is left-continuous, monotone increasing, and has the property that

$$F(x) \geqslant z \quad \text{if and only if} \quad x \geqslant F^{-1}(z).$$

Let us note that

$$\mathrm{P}\left\{X \leqslant F^{-1}(1/2)\right\} = \mathrm{P}\left\{F(X) \leqslant 1/2\right\}.$$

If X is absolutely continuous, then $F(X) \sim \mathrm{unif}(0\,,1)$ and hence

$$\mathrm{P}\left\{X \leqslant F^{-1}(1/2)\right\} = \mathrm{P}\left\{\mathrm{unif}(0\,,1) \leqslant 1/2\right\} = \frac{1}{2},$$

and

$$\mathrm{P}\left\{X \geqslant F^{-1}(1/2)\right\} = \mathrm{P}\left\{\mathrm{unif}(0\,,1) \geqslant 1/2\right\} = \frac{1}{2}$$

And if X is discrete, then

$$\mathrm{P}\left\{X \leqslant F^{-1}(1/2)\right\} = \sum_{x \leqslant F^{-1}(1/2)} f(x) = \sum_{x:\, F(x) \leqslant 1/2} f(x) \geqslant \frac{1}{2},$$

and $\mathrm{P}\{X \geqslant F^{-1}(1/2)\} \geqslant 1/2$, similarly. In other words,

$$F^{-1}(1/2) := \text{ the population median.}$$

Similarly,

$$\hat{F}_n^{-1}(1/2) := \text{ the sample median;}$$

that is the smallest point such that at least half of the sample lies to its left. In practice, any point such that at least half of the sample lies to its left and at least half lies to its right can be used. For our particular choice, we have

$$\hat{F}_n(1/2) = \begin{cases} X_{(n+1)/2:n} & \text{if } n \text{ is odd,} \\ X_{n/2:n} & \text{if } n \text{ is even,} \end{cases}$$

written in terms of the order statistics. [Usually, people use $\hat{F}_n(x) := \frac{1}{2}\{X_{n/2:n} + X_{(n/2)+1:n}\}$ when n is even, in place of $\hat{F}_n^{-1}(1/2)$, as we have done.] □

Example 4 (Quantiles)**.** Given a number $p \in (0\,,1)$, a pth *quantile* Q_p of X [and/or F] is any number that satisfies

$$\mathrm{P}\{X \leqslant Q_p\} \geqslant p \quad \text{and} \quad \mathrm{P}\{X \geqslant Q_p\} \geqslant 1 - p.$$

One argues, as we did in the previous example, in order to see that $Q_p = F^{-1}(p)$ is always a pth population quantile, and $\hat{F}_n^{-1}(p)$ is the pth sample quantile. □

3 The Bootstrap

Let F denote the [typically unknown] cumulative distribution function of a given population of interest to us. A *statistical functional* is a real-valued function $T(F)$ of F. You should think of F as the variable of the functional T. Thus, for instance,

$$T(F) := \mu_F := \int_{-\infty}^{\infty} x \, \mathrm{d}F(x)$$

is the mean functional, and in particular, $\mu_{\hat{F}_n} = \bar{X}$ is the sample mean. The mean is an example of a *linear statistical functional*. A general linear statistical function has the form

$$T(F) := \mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)\,\mathrm{d}F(x). \tag{3}$$

Even though g need not be a linear function [e.g., as is the case with the variance functional $\int_{-\infty}^{\infty}(x-\mu_F)^2\,\mathrm{d}F(x)$], such functionals are linear in the following sense: If $\lambda \in (0\,,1)$ and F and G are two cdf's, then $\lambda F + (1-\lambda)G$ is also a cdf,[1] and

$$T(\lambda F + (1-\lambda)G) = \lambda T(F) + (1-\lambda)T(G).$$

3.1 A Bootstrap Confidence Interval

Let T be a statistical functional, and suppose that we wish to estimate $T(F)$, where F denotes the unknown cdf for our population. The *bootstrap* estimate of $T(F)$ is simply $T(\hat{F}_n)$. We have seen already that $\hat{F}_n \to F$ uniformly in probability as $n \to \infty$ [Glivenko–Cantelli theorem]. Therefore, it stands to reason that many statistical functional T have the property that $T(\hat{F}_n) \to T(F)$. [Such are called *continuous functionals*.] Rather than study the functional analysis aspect of such problems, let us concentrate on a simple though telling special case.

If T is a *linear* statistical functional of the form (3), then

$$T(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

The asymptotic theory of linear statistical functionals is elementary: Since $T(F) = \int g\,\mathrm{d}F = \mathrm{E}g(X)$, the law of large numbers ensures that

$$T(\hat{F}_n) \xrightarrow{\mathrm{P}} T(F) \qquad \text{as } n \to \infty,$$

and in accord with the central limit theorem,

$$\sqrt{n}\left[T(\hat{F}_n) - T(F)\right] \xrightarrow{d} \mathrm{N}\left(0\,,\sigma^2(F)\right),$$

[1] The easiest way to understand the meaning of $\lambda F + (1-\lambda)G$ is as follows: Suppose $X \sim F$ and $Y \sim G$ are independent. Set $Z := X$ with probability λ and $Z := Y$ with probability $1-\lambda$. Then, you should check that $Z \sim \lambda F + (1-\lambda)G$.

where

$$\sigma^2(F) := \mathrm{Var}\left(g(X)\right) = \int_{-\infty}^{\infty} \left(g(x) - T(F)\right)^2 \mathrm{d}F(x)$$

denotes the asymptotic variance. Therefore, if n is large then the following has probability $\approx 1 - \alpha$:

$$T(F) \in \left(T(\hat{F}_n) - \frac{\sigma(F)}{\sqrt{n}} \ , \ T(\hat{F}_n) + \frac{\sigma(F)}{\sqrt{n}} \right).$$

But this is not a proper confidence interval [with asymptotic level $1 - \alpha$], because the "limiting standard deviation" $\sigma(F)$ is unknown. Therefore, we use the following *bootstrap* estimator for $\sigma^2(F)$:

$$\begin{aligned} \sigma^2(\hat{F}_n) &= \int_{-\infty}^{\infty} \left(g(x) - T(\hat{F}_n)\right)^2 \mathrm{d}\hat{F}_n(x) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(g(X_i) - \frac{1}{n} \sum_{j=1}^{n} g(X_j) \right)^2. \end{aligned}$$

The quantity $\sigma^2(\hat{F}_n)$ is the *bootstrap estimator for the standard error* of $T(\hat{F}_n)$. And our *bootstrap confidence interval for* $T(F)$ is

$$\left(T(\hat{F}_n) - \frac{\sigma(\hat{F}_n)}{\sqrt{n}} \ , \ T(\hat{F}_n) + \frac{\sigma(\hat{F}_n)}{\sqrt{n}} \right). \tag{4}$$

By the law of large numbers, $\sigma^2(\hat{F}_n)$ converges in probability to $\sigma^2(F)$ as $n \to \infty$, and therefore Slutsky's theorem justifies the remainder of the following.

Theorem 5. *If* $\mathrm{E}(|g(X)|^2) < \infty$*, then the preceding bootstrap confidence interval for* $T(F)$*—see* (4)*—has asymptotic level* $(1-\alpha) \times 100\%$ *as* $n \to \infty$*.*

3.2 Subsampling

One can view bootstrapping quite abstractly as a resampling and/or subsampling scheme. Suppose F is the unknown cdf and we wish to know about some functional $T(F)$ of F. In order to learn about $T(F)$ we select a random sample $X_1, \ldots, X_n$ from F and estimate $T(F)$ with $T_n := T(\hat{F}_n)$. We can also estimate statistical functionals such as

$$S(F) := \mathrm{P}\{T_n \leqslant t\}.$$

Let us write this quantity, somewhat more precisely, as

$$S(F) = \mathrm{P}_F\{T_n \leqslant t\},$$

in order to emphasize that we are assuming that F is indeed the true cdf. Then we can estimate $S(F)$ by $S(\hat{F}_n)$. This looks easy [and is, in some sense] but its meaning is quite subtle, Indeed, $S(\hat{F}_n)$ is the [conditional] probability that we have sampled [and conditioned on] $X_1, \ldots, X_n$ and obtain $\hat{F}_n$, then subsample from $\hat{F}_n$ a sample $X_1^*, \ldots, X_n^*$ of size n, compute the corresponding T_n^* from the subsample $X_1^*, \ldots, X_n^*$, and then find the odds that T_n^* is at most t. A question that arises is: *What does it mean to sample from* $\hat{F}_n$?

Since $\hat{F}_n$ is a proper cdf on n points [albeit a random one], we sample $\hat{F}_n$ as follows: Independently from [and conditionally on] the original sample $X_1, \ldots, X_n$, we sample with replacement $X_1^*, \ldots, X_n^*$ from the set $\{X_1, \ldots, X_n\}$, all points being equally likely. Then, the probability of interest is the conditional probability

$$S(\hat{F}_n) = \mathrm{P}\left(T(F_n^*) \leqslant t \mid X_1, \ldots, X_n\right),$$

where F_n^* denotes the empirical cdf of the subsamples; that is,

$$F_n^*(x) := \frac{1}{n}\sum_{j=1}^{n} I\left\{X_j^* \leqslant x\right\}.$$

In general, one cannot compute the random probability $S(F_n^*)$. Therefore, one simulates it by sampling N conditionally independent groups of n random variables $X_{1,1}^*, \ldots, X_{1,n}^*, \ldots, X_{N,1}^*, \cdots, X_{N,n}^*$, and then applying the law of large numbers [valid as $N \to \infty$]:

$$S(\hat{F}_n) \approx \frac{1}{N}\sum_{j=1}^{N} I\left\{F_{j,n}^* \leqslant t\right\},$$

where

$$F_{j,n}^*(x) := \frac{1}{n}\sum_{k=1}^{n} I\left\{X_{j,k}^* \leqslant x\right\}$$

denotes the empirical cdf of the jth group $X_{j,1}^*, \ldots, X_{j,n}^*$ of subsamples.

3.3 The Bootstrap for Variance

In order to gel the ideas further, let us now consider a statistical functional T, and apply the bootstrap statistic $T_n := T(\hat{F}_n)$ to estimate $T(F)$. A

question that arises is, *how do we estimate the variance of* T_n? The standard methods for doing this sort of thing, at this level generality, typically involve resampling; one uses either the bootstrap [which we do now] and the jackknife [which we cover later]. Both methods work well under some regularity assumptions on the functional T, though I will not discuss this sort of optimality in great detail here because the discussion requires a great deal of functional analysis that is beyond the level of this course.

In order to estimate $\theta = \text{Var}(T_n)$ via the bootstrap, we perform the following algorithmically:

1. Sample with replacement and at random $X_1^*, \ldots, X_n^*$ from the data, all data points being equally likely;
2. Compute $T_n^* := T(X_1^*, \ldots, X_n^*)$, using the same functional T as the one that led us here;
3. Repeat steps 1 and 2 N times—for a large N—in order to obtain N conditionally independent replicas $T_{1,1}^*, \ldots, T_{N,n}^*$;
4. The bootstrap estimate for $T(F) := \text{Var}(T_n)$ is

$$V_{\text{bootstrap}} := \frac{1}{N} \sum_{j=1}^{N} \left(T_{j,n}^* - \frac{1}{n} \sum_{k=1}^{n} T_{j,k}^* \right)^2 .$$

3.4 The Bootstrap for Quantiles

Let us say a few things about the bootstrap estimation of the pth quantile $T(F) := F^{-1}(p)$ of the cdf F, where $0 < p < 1$ is known. Of course, the bootstrap estimate is the pth sample quantile $T(\hat{F}_n) = \hat{F}_n^{-1}(p)$, as we have seen earlier. The issue is to decide on the asymptotic behavior of $\hat{F}_n^{-1}(p)$.

Theorem 6 (Efron, 1979)**.** *Suppose that $f = F'$ exists, is continuous, and $f(F^{-1}(p)) > 0$. Then,*

$$\sqrt{n} \left(\hat{F}_n^{-1}(p) - F^{-1}(p) \right) \xrightarrow{d} \text{N} \left(0 \, , \, \frac{pq}{|f(F^{-1}(p))|^2} \right),$$

as $n \to \infty$, where $q := 1 - p$ as before.

Sketch of the Proof. In order to understand this result, consider first the special case where the data comes from unif$(0\,,1)$. In that case, $f(x) = 1$, $F(x) = x$, and $F^{-1}(p) = p$. The theorem states that, in this case,

$$\sqrt{n} \left(p - \hat{F}_n^{-1}(p) \right) \xrightarrow{d} \text{N}(0\,, pq).$$

But this is not hard to prove, given our existing knowledge of empirical process theory. Indeed, we know that uniformly for all $0 \leqslant x \leqslant 1$,

$$\sqrt{n}\left(\hat{F}_n(x) - x\right) \approx B^\circ(x),$$

and hence the preceding is true even at the random point $x = \hat{F}_n^{-1}(p)$. That is,

$$\sqrt{n}\left(p - \hat{F}_n^{-1}(p)\right) \approx B^\circ\left(\hat{F}_n^{-1}(p)\right).$$

One ready consequence of this is that $p - \hat{F}_n^{-1}(p)$ is roughly of order $n^{-1/2}$ and hence $\hat{F}_n^{-1}(p) \approx p$ with high probability when $n \gg 1$. Since the Brownian bridge is continuous, it follows that

$$\sqrt{n}\left(p - \hat{F}_n(p)\right) \approx B^\circ(p) \sim \mathrm{N}\left(0\,, pq\right).$$

This proves the result in the unif$(0\,,1)$ case. The general case can be reduced to this one, after a little more work. I will skip almost all of the remaining details, except for the fact that the key ingredient involves *variance stabilization*. What this means, in rough terms is this: Suppose $\sqrt{n}(Z_n - \mu) \approx \mathrm{N}(0\,, pq)$ in distribution, say thanks to some sort of central limit theorem argument. Then, $Z_n \approx \mu$ with high probability and so a Taylor expansion yields the following for every nice function g:

$$\sqrt{n}\left(g(Z_n) - g(\mu)\right) \approx g'(\mu) \times \sqrt{n}\left(Z_n - \mu\right) \approx \mathrm{N}\left(0\,, pq\left|g'(\mu)\right|^2\right),$$

in distribution. The connection to the Efron's theorem is through the application $g(x) := F^{-1}(x)$. If F^{-1} were the classical [continuous] inverse function to F, then it follows that $g'(x) = 1/f(F^{-1}(x))$, thanks to elementary calculus. But, as you have been warned, the remaining details are skipped. □

Let us return to Theorem 6 and build an estimator for $F^{-1}(p)$. According to Theorem 6, if $f = F'$ exists and is continuous with $f(F^{-1}(p)) > 0$, then $\hat{F}_n^{-1}(p)$ is a consistent estimator for $F^{-1}(p)$. Next we might wish to develop a confidence interval. But that task presents us with a serious challenge: We need to estimate the asymptotic variance

$$\tau^2 = \frac{pq}{|f(F^{-1}(p))|^2}.$$

Once we do this, we know how to proceed, using standard theory.

The quantites p and $q = 1 - p$ are, of course, known. Therefore, one possible approach is to use the estimator

$$\hat{\tau}^2 := \frac{pq}{\left|\hat{f}\left(\hat{F}_n^{-1}(p)\right)\right|^2},$$

where $\hat{f}$ is a reliable kernel density estimator of f. Theorem 6 ensures that $\hat{F}_n^{-1}(p) \xrightarrow{\mathrm{P}} F^{-1}(p)$. Therefore, under the conditions of Parzen's uniform consistency theorem, $\hat{\tau}^2 \xrightarrow{\mathrm{P}} \tau^2$ as $n \to \infty$ [Slutsky's theorem]. A more classical approach is to use a *subsampling method* [Hall, DiCiccio, and Roman, 1989; Falk and Kaufman ,1991]. Recall that the real issue is to find z_α such that

$$\mathrm{P}_F\left\{\sqrt{n}\left(\hat{F}_n^{-1}(p) - F^{-1}(p)\right) \leqslant z_\alpha\right\} \approx 1 - \alpha.$$

This can be done by resampling methods. That is, we can seek to find z_α such that

$$\mathrm{P}_{\hat{F}_n}\left\{\sqrt{n}\left((F_n^*)^{-1}(p) - \hat{F}_n^{-1}(p)\right) \leqslant z_\alpha\right\} \approx 1 - \alpha,$$

where $(F_n^*)^{-1}$ denotes the left-continuous inverse to the resampled empirical cdf F_n^*.

3.5 The Bootstrap for Skewness

The *skewness* of a random variable X is defined as

$$\gamma := \frac{\mathrm{E}[(X - \mathrm{E}X)^3]}{[\mathrm{Var}(X)]^{3/2}} = \mathrm{E}\left[\left(\frac{X - \mathrm{E}X}{\mathrm{SD}(X)}\right)^3\right].$$

If the cdf of X is F, then we can always write κ in terms of F [the "skewness of F"] as follows:

$$\gamma = \frac{\int_{-\infty}^{\infty}(x - \mu_F)^3\,\mathrm{d}F(x)}{\left[\int_{-\infty}^{\infty}(x - \mu_F)^2\,\mathrm{d}F(x)\right]^{3/2}}, \quad \text{where} \quad \mu_F := \int_{-\infty}^{\infty} x\,\mathrm{d}F(x).$$

The skewness γ of F is a "unit-free quantity" and a fairly good measure of how asymmetric the distribution of X is. Stated in terms of F, X is symmetric if and only if $F(x) = 1 - F((-x)-)$ for all $x \in \mathbf{R}$, where $F(y-) := \lim_{z \uparrow y} F(z)$. Whenever X is symmetric $\gamma = 0$. If $\gamma \gg 0$ then X is mostly "skewed to the right," whereas $\gamma \ll 0$ implies that X is mostly "skewed

to the left." We may think of γ as $\gamma(F)$ and apply our boostrap [plugin] estimator

$$\begin{aligned} G_{\text{bootstrap}} &:= \frac{\int_{-\infty}^{\infty}(x-\mu_{\hat{F}_n})^3\,\mathrm{d}\hat{F}_n(x)}{\left[\int_{-\infty}^{\infty}(x-\mu_{\hat{F}_n})^2\,\mathrm{d}\hat{F}_n(x)\right]^{3/2}} \\ &= \frac{n^{-1}\sum_{j=1}^{n}(X_j-\bar{X})^3}{\left[n^{-1}\sum_{j=1}^{n}(X_j-\bar{X})^2\right]^{3/2}} \\ &= \frac{\sqrt{n}\,\sum_{j=1}^{n}(X_j-\bar{X})^3}{\left[\sum_{j=1}^{n}(X_j-\bar{X})^2\right]^{3/2}}. \end{aligned}$$

3.6 The Bootstrap for Kurtosis

The *kurtosis* of a random variable X is defined as

$$\kappa := \frac{\mathrm{E}\left(|X-\mathrm{E}X|^4\right)}{[\mathrm{Var}(X)]^2} - 3 = \mathrm{E}\left(\left[\frac{X-\mathrm{E}X}{\mathrm{SD}(X)}\right]^4\right) - 3.$$

This is a unit-free parameter.

In order to understand this definition better, consider the case that $X \sim \mathrm{N}(\mu\,,\sigma^2)$. In that case, we can let $Z := (X-\mu)/\sqrt{\mathrm{Var}(X)}$ in order to see that $\kappa = \mathrm{E}(Z^4) - 3$, where $Z \sim \mathrm{N}(0\,,1)$. Therefore,

$$\begin{aligned} \kappa &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4\mathrm{e}^{-x^2/2}\,\mathrm{d}x - 3 = \sqrt{\frac{2}{\pi}}\int_0^{\infty} x^4\mathrm{e}^{-x^2/2}\,\mathrm{d}x - 3 \\ &= \frac{4}{\sqrt{\pi}}\int_0^{\infty} y^{3/2}\mathrm{e}^{-y}\,\mathrm{d}y \qquad\qquad (y := x^2/2) \\ &= \frac{4}{\sqrt{\pi}}\Gamma(5/2). \end{aligned}$$

But $\Gamma(5/2) = (3/2)\Gamma(3/2) = (3/2)(1/2)\Gamma(1/2) = (3/4)\sqrt{\pi}$. Therefore, $\kappa = 0$ for any normal distribution!

It turns out that κ is usually very sensitive to the choice of normality in the preceding computation. That is, $\kappa \neq 0$ for most "natural" non-normal distributions. Therefore, we can fairly reliably know whether or not a population is normal by just checking [parametrically!] to see if $\kappa = 0$ or not. Because

$$\kappa = \frac{\int_{-\infty}^{\infty}(x-\mu_F)^4\,\mathrm{d}F(x)}{\left[\int_{-\infty}^{\infty}(x-\mu_F)^2\,\mathrm{d}F(x)\right]^2},$$

we have the plugin [bootstrap] estimate K of kurtosis,

$$K_{\text{bootstrap}} = \frac{n^{-1}\sum_{i=1}^n (X_i - \bar{X})^4}{\left[n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} - 3 = \frac{n\sum_{i=1}^n (X_i - \bar{X})^4}{\left[\sum_{i=1}^n (X_i - \bar{X})^2\right]^2} - 3.$$

We can understand κ better after we inspect a few more examples.

Example 7 (unif$(-1\,,1)$)**.** If $X \sim \text{unif}(-1\,,1)$, then the pdf of X is $f(x) = \frac{1}{2}\mathbf{I}\{-1 < x < 1\}$, and hence

$$\begin{aligned} \mathrm{E}X &= 0, \\ \mathrm{Var}(X) = \mathrm{E}(X^2) &= \frac{1}{2}\int_{-1}^1 x^2\,\mathrm{d}x = \frac{1}{3}, \\ \mathrm{E}\left(|X - \mathrm{E}X|^4\right) = \mathrm{E}(X^4) &= \frac{1}{2}\int_{-1}^1 x^4\,\mathrm{d}x = \frac{1}{5}. \end{aligned}$$

Therefore,

$$\kappa \approx \frac{1/5}{(1/3)^2} - 3 = \frac{9}{5} - 3 = -\frac{6}{5}.$$

This is strictly negative, which means that the uniform distribution on $(-1\,,1)$ has lighter tails [i.e., has *greater* concentration near the mean] than any normal distribution does. □

Example 8 (DE)**.** Consider the case that X has the double exponential distribution with pdf $f(x) = \frac{1}{2}\exp(-|x|)$. In that case,

$$\begin{aligned} \mathrm{E}X &= 0, \\ \mathrm{Var}(X) = \mathrm{E}(X^2) &= \frac{1}{2}\int_{-\infty}^{\infty} x^2 \mathrm{e}^{-|x|}\,\mathrm{d}x = \int_0^{\infty} x^2\mathrm{e}^{-x}\,\mathrm{d}x = \Gamma(3) = 2, \\ \mathrm{E}\left(|X - \mathrm{E}X|^4\right) = \mathrm{E}(X^4) &= \frac{1}{2}\int_{-\infty}^{\infty} x^4 \mathrm{e}^{-|x|}\,\mathrm{d}x = \int_0^{\infty} x^4\mathrm{e}^{-x}\,\mathrm{d}x = \Gamma(5) = 4! = 24. \end{aligned}$$

Therefore,

$$\kappa = \frac{24}{4} - 3 = 3.$$

The interpretation of this formula is that, since $\kappa > 0$, the double exponential distribution has heavier tails [i.e., has *smaller* concentration near the mean] than any normal does. □

3.7 The Bootstrap Can Fail

The bootstrap is a powerful method of broad utility. However, there are occasions where it fails demonstrably. Here is an example.

Example 9 (Efron, 1979)**.** Let $X_1, \ldots, X_n$ be i.i.d. unif$(0\,, \theta)$ where $\theta > 0$ is unknown. The usual estimator for θ is the largest order statistic $X_{n:n}$; this is, for example, the MLE. In order to develop confidence bounds, we may study

$$T_n := n\,(\theta - X_{n:n})\,.$$

The scaling factor is n and not $\sqrt{n}$ because in this way

$$T_n \xrightarrow{d} \text{Exp}(\theta).$$

The proof is easy, and many of you have seen it in your Math. 5080–5090 sequence: For all $t > 0$,

$$\begin{aligned}
\text{P}\left\{T_n < t\right\} &= 1 - \text{P}\left\{n\,(\theta - X_{n:n}) > t\right\} = 1 - \text{P}\left\{X_{n:n} < \theta - \frac{t}{n}\right\} \\
&= 1 - \left(\text{P}\left\{X_1 < \theta - \frac{t}{n}\right\}\right)^n \\
&= 1 - \left(1 - \frac{t}{n\theta}\right)^n \qquad \text{(if } n > t/\theta \text{ is large enough)} \\
&\to 1 - \exp\left(-\frac{t}{\theta}\right) \qquad \text{as } n \to \infty.
\end{aligned}$$

But the bootstrap estimate of this probability is

$$\text{P}_{\hat{F}_n}\left\{T_n^* < t\right\},$$

where T_n^* is obtained by taking n independent subsamples $X_1^*, \ldots, X_n^*$ from $X_1, \ldots, X_n$ and then computing the conditional probability

$$\text{P}_{\hat{F}_n}\left\{n\,(X_{n:n} - X_{n:n}^*) < t\right\}.$$

Since $X_{n:n} \geqslant X_{n:n}^*$, **the preceding probability is greater than or equal to the following:**

$$\begin{aligned}
\text{P}_{\hat{F}_n}\left\{n\,(X_{n:n} - X_{n:n}^*) = 0\right\} &= \text{P}_{\hat{F}_n}\left\{X_{n:n} = X_{n:n}^* \mid X_1, \ldots, X_n\right\} \\
&= 1 - \text{P}_{\hat{F}_n}\left\{X_{n:n}^* < X_{n:n} \mid X_1, \ldots, X_n\right\} \\
&= 1 - \left(\frac{n-1}{n}\right)^n \\
&\to 1 - \text{e}^{-1}.
\end{aligned}$$

That is, the bootstrap probability estimate satisfies[2]

$$\lim_{n\to\infty} \mathrm{P}_{\hat{F}_n}\{T_n^* < t\} \geqslant 1 - \mathrm{e}^{-1} \qquad \text{for all } t > 0.$$

Therefore, it is easy to see that the preceding does not have the same limit as $\mathrm{P}\{T_n < t\} \approx 1 - \exp(-t/\theta)$ when $t < \theta$. In other words, the bootstrap probability estimate [for the confidence values] fails to converge to the correct value for many values of t. □

4 The Jackknife

4.1 The Jackknife, and Bias Reduction

The jackknife refers to a loosely-connected family of resampling schemes that are, for example, used in improving the asymptotic behavior of biased estimators. Suppose T is a statistical functional, and we are interested in estimating $\theta := T(F)$ with $T(\hat{F}_n)$. Let us think of $T(\hat{F}_n)$ as a function of $X_1, \ldots, X_n$, instead of as a function of $\hat{F}_n$. In other words, our estimator for θ has the form

$$T_n := T_n(X_1, \ldots, X_n).$$

Let $T_{(-i)} := T_{(-i),n}$ denote the same estimator, but applied to the data after we excise X_i from the data. Then the *jackknife estimator* of the bias of T_n is

$$b_{\text{jackknife}} := (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} T_{(-i)} - T_n\right) := (n-1)\left(\bar{T}_n - T_n\right).$$

And the *bias-corrected jackknife estimate* for θ is

$$T_{\text{jackknife}} := T_n - b_{\text{jackknife}}.$$

The choice of these jackknife estimators has to do with some heuristics that are justified in various concrete settings. Indeed, it turns out that some times T_n has a bias which has the asymptotic form

$$\text{bias}(T_n) := \frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + \frac{\alpha_3}{n^3} + \cdots, \tag{5}$$

as $n \to \infty$, for constants $\alpha_1, \alpha_2, \ldots$ with $\alpha_1 \neq 0$. Since $T_{(-i)}$ has the same structure as T_n but with the ith sample point removed, it follows from (5) that

$$\text{bias}(T_{(-i)}) = \frac{\alpha_1}{n-1} + \frac{\alpha_2}{(n-1)^2} + \cdots.$$

[2] To be perfectly honest, the limit should be replaced by a $\liminf$.

Average over the first n values of i to see that

$$\text{bias}(\bar{T}_n) = \frac{\alpha_1}{n-1} + \frac{\alpha_2}{(n-1)^2} + \cdots .$$

Therefore,

$$\begin{aligned}
\text{bias}(b_{\text{jackknife}}) &= (n-1)\left[\left(\frac{\alpha_1}{n-1} + \frac{\alpha_2}{(n-1)^2} + \cdots\right) - \left(\frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + \cdots\right)\right] \\
&= (n-1)\left[\frac{\alpha_1}{n(n-1)} + \frac{\alpha_2\left(n^2 - (n-1)^2\right)}{n^2(n-1)^2} + \cdots\right] \\
&= \frac{\alpha_1}{n} + \frac{\alpha_2(2n-1)}{n^2(n-1)} + \cdots \\
&= \frac{\alpha_1}{n} + \frac{2\alpha_2}{n^2} + \cdots .
\end{aligned}$$

In particular, the bias of our jackknife estimator looks like

$$\begin{aligned}
\text{bias}(T_{\text{jackknife}}) &= \text{bias}(T_n) - \text{bias}(b_{\text{jackknife}}) \\
&= -\frac{\alpha_2}{n^2} + \cdots ,
\end{aligned}$$

and this quantity is much smaller than $(\alpha_1/n) + \cdots = \text{bias}(T_n)$ when n is large. In other words, if our estimator T_n has a bias that satisfies (5), then the jackknifed version $\theta_{\text{jackknife}}$ of T_n has a significantly smaller bias when the sample size is large. The following is a standard example of an estimator that satisfies (5).

Example 10. Suppose $\theta = \text{Var}(X_1) = T(F) > 0$ is unknown. The estimator that we will use is

$$T(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Clearly, this is biased. In fact, we know from Math. 5080 that $\text{E}(T_n) = (n-1)\theta/n$, whence

$$\text{bias}(T_n) = \frac{(n-1)\theta}{n} - \theta = -\frac{\theta}{n}.$$

In this case, $\alpha_1 = \theta > 0$ and $\alpha_2 = \alpha_3 = \cdots = 0$. Hence the jackknife estimator is, in fact, unbiased. □

4.2 The Jackknife Estimator of Variance

Suppose, as before, that $T_n = T(\hat{F}_n)$ is the estimator of some $\theta := T(F)$. The ith *pseudo-value* is the quantity

$$\widetilde{T}_i := nT_n - (n-1)T_{(-i)} \qquad (1 \leqslant i \leqslant n).$$

The jackknife estimator of $\mathrm{Var}(T_n)$ is defined as

$$V_{\text{jackknife}} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\widetilde{T}_i - \frac{1}{n} \sum_{j=1}^{n} \widetilde{T}_j \right)^2.$$

That is, $V_{\text{jackknife}}$ is a normalized version of the sample variance of all of the pseudo-values.

Example 11. Consider the linear case where $T_n = n^{-1} \sum_{i=1}^{n} g(X_i)$. In this case, $T_{(-i)} = (n-1)^{-1} \sum_{j \neq i} g(X_j)$, and hence the ith pseudo-value is $\widetilde{T}_i = g(X_i)$. Consequently, the jackknife estimator of the variance of T_n is

$$V_{\text{jackknife}} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(g(X_i) - \frac{1}{n} \sum_{j=1}^{n} g(X_j) \right)^2,$$

which is $1/(n-1)$ times the usual sample variance of the $g(X_i)$'s.[3] Thanks to the law of large numbers, $V_{\text{jackknife}}$ is a good estimator of $\mathrm{Var}(T_n) = \mathrm{Var}(g(X_1))/n$ in the following sense:

$$\frac{V_{\text{jackknife}}}{\mathrm{Var}(T_n)} = \frac{(n-1)^{-1} \sum_{i=1}^{n} \left(g(X_i) - n^{-1} \sum_{j=1}^{n} g(X_j) \right)^2}{\mathrm{Var}(g(X_1))} \xrightarrow{\text{P}} 1,$$

as $n \to \infty$. □

4.3 The Jackknife Can Fail

The preceding suggests [quite appropriately] that the jackknife variance estimate is typically a very good one, for instance when applied to linear statistical functionals. Striking Theorems of Bradley Efron [$p = 1/2$] and Michael Martin [$0 < p < 1$] show that there are natural settings in which the jackknife can fail completely.

[3] In particular, note that if $T_n = \bar{X}$ [that is, $g(x) = x$], then $V_{\text{jackknife}} = S^2/n$.

Theorem 12 (Efron, 1979; Martin, 1990)**.** *If $T(F) = F^{-1}(p)$ is the pth quantile of F, then*

$$\frac{V_{\text{jackknife}}}{\text{Var}(T_n)} \xrightarrow{d} \sqrt{\text{Exp}(1)},$$

as $n \to \infty$.

I will not prove this beautiful fact. The proof is neither particularly long, nor hard. But it rests on a series of specialized facts about the order statistics, which I will omit. Instead, let me point out that the preceding implies that

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \text{P} \left\{ \left| \frac{V_{\text{jackknife}}}{\text{Var}(T_n)} - 1 \right| > \epsilon \right\} = 1.$$

That is, the jackknife estimator of the variance of the pth sample quantile T_n is always inconsistent $[0 < p < 1]$![4]

[4]This is a much stronger statement than the one which says that the jackknife estimator is never consistent. I will let you ponder over the logical details.