Meeting in honor of Evarist's 70th birthday

Cambridge, 24 June 2014

Some reminiscences relating to Evarist

and

some explorations about the bootstrap

Evarist was my 4th Ph. D. student and third at MIT. Originally from Catalonia, he arrived in the Fall of 1970 from Venezuela, where he had been teaching. His thesis, finished in 1973, was on "Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms." It was published in Ann. Stat. in 1975.

After finishing at MIT Evarist went back to Venezuela for a year. He spent the academic year 1974–75 in Berkeley on a short-term faculty appointment. Then he again returned to Venezuela to fulfill the obligation of his 3-year graduate fellowship to MIT. He was at "IVIC," the Venezuelan institute of scientific research.

INFINITE-DIMENSIONAL LIMIT THEOREMS

The one previous meeting I've been to in the UK was in late July 1974, in Durham, a London Mathematical Society Symposium on Functional Analysis and Stochastic Processes. I was struck by the remarkably cool midsummer weather in Durham, as I was also by remarkably warm February weather here in Cambridge on my one previous visit, as a tourist for a day in 1976.

In 1955–56 R. Fortet and E. Mourier had proved that the CLT (central limit theorem) held in L^p spaces for $2 \leq p < \infty$ for i.i.d. random variables X_j with $EX_1 = 0$ and $E\|X_1\|^2 < \infty$. In 1968 Volker Strassen and I proved that in $C[0,1]$ the CLT could fail even for bounded random variables but held under a metric entropy condition.

In 1974 as far as I knew, in L^p for $1 < p < 2$ the question was still open, but I thought it was due to be solved and said we should solve it there in Durham. At the end of my talk Gilles Pisier came up to me and told me it had been solved in the negative. The paper by Jørgen Hoffmann-Jørgensen and Gilles appeared in Ann. Prob. in 1976, characterizing the Banach spaces in which the CLT held under the classical conditions as those of type 2, which L^p spaces for $1 \leq p < 2$ are not.

In 1976, Evarist began to publish on central limit theorems in Banach spaces. I think since then, he (and others) have known more than I do about limit theorems in separable Banach spaces. In 1980 Aloisio Araujo and Evarist published the book *The Central Limit Theorem for Real and Banach Space Valued Random Variables*. The book is by far Evarist's most cited work (Google Scholar).

I was lucky that my work on empirical processes was synergistic with that of people working at first in separable Banach spaces. In some fields of statistics I entered later I seemed to be more of a contrarian.

Related to my other topic today are Evarist's second and third-most cited works, both with Joel Zinn, in Ann. Prob.: "Some limit theorems for empirical processes," 1984, and "Bootstrapping general empirical measures," 1990.

When first reading the "Some limit theorems" paper, I noticed a lot of technical facts on symmetrization, desymmetrization, and randomization such as Poissonization. In the bootstrap paper, I saw how the technical facts are useful.

In January of some year, about 1989 or 1990, I visited the statistics departments in Seattle and Berkeley, lecturing on Evarist and Joel's work on the bootstrap, as I thought that was more interesting than any research I was doing around that time. My handout for those lectures was a mixture of pages from a preprint of Evarist and Joel's 1990 paper and copies of proofs by others it relied on, aiming to give self-contained proofs of Evarist and Joel's theorems.

I've taught about the bootstrap from a more applied point of view, a week or two each time, in graduate nonparametric statistics courses since 2007. Preparing this talk has helped me get ready for teaching that next spring.

Let S be a sample space. Let P be a probability measure on S and suppose we have $X_1, X_2, ...,$ S-valued, i.i.d. P. Let $P_n := \frac{1}{n} \Sigma_{j=1}^{n} \delta_{X_j}$. Suppose we have a functional $T(\cdot)$ defined on probability measures on S. A point estimate of $T(P)$ is $T(P_n)$. To estimate the amount of variability in T that could occur with different samples, there is the bootstrap, invented by B. Efron (1979 Ann. Stat.). Let X_{ni}^{B}, $i = 1, ..., m$, be i.i.d. (P_n). The case I'm familiar with is to take $m = n$, as I think statisticians generally do, although Evarist and others have treated $m = o(n)$.

Then let $P_n^B := \frac{1}{n}\Sigma_{i=1}^n \delta_{X_{ni}^B}$, the *bootstrap empirical measure.* Under some conditions, the conditional distribution of P_n^B given P_n may be such that for some functional(s) T, $T(P_n^B) - T(P_n)$ given P_n may behave as $T(P_n') - T(P)$, where P_n' is a random empirical measure for P as contrasted with the fixed given P_n. More precisely, for some sequence of constants a_n such as $\sqrt{n}$,

$$(a_n(T(P_n^B) - T(P_n))|P_n)$$
$$\sim a_n(T(P_n') - T(P)),$$

with $\sim$ indicating closeness in distribution, where $a_n(T(P_n') - T(P))$ is bounded away from 0 in probability.

Let $\mathcal{F}$ be a class of real-valued measurable functions on S. It is called a Donsker class for P if the empirical process $\sqrt{n}(P_n - P)$ converges in distribution to a limiting Gaussian process G_P indexed by $\mathcal{F}$ with respect to uniform convergence over $\mathcal{F}$. In 1978 I gave a definition of Donsker class (of sets), then later twice changed the definition, extended to classes of functions. The final definition, by Hoffmann-Jørgensen, defines convergence *in* distribution of a sequence of possibly nonmeasurable random elements to a measurable random variable limit such as G_P.

Details are in my book *Uniform Central Limit Theorems* (Cambridge University Press; a second edition appeared early this year) or in van der Vaart and Wellner, *Weak Convergence and Empirical Processes.* $\mathcal{F}$ is called a *bootstrap P-Donsker class* in probability if the bounded Lipschitz distance between the distribution of $\sqrt{n}(P_n^B - P_n)$ given P_n and that of G_P converges to 0 in outer probability as $n \to \infty$.

Disclaimer: Although some people kindly put my name on it, I did not define the bounded Lipschitz metric in separable metric spaces, Fortet and Mourier did in the 1950's. Also, I did not prove that weak convergence implies convergence in the bounded Lipschitz norm, R. Ranga Rao did in a paper I was lucky to see while in graduate school because my adviser assigned me the paper to referee. It seems I did first prove that bounded Lipschitz convergence implies weak convergence in any separable metric space, which moreover need not be complete, and I was involved in extending the definition to nonseparable metric spaces, which we need for empirical processes.

Evarist and Joel proved in their 1990 Ann. Prob. paper that P-Donsker implies bootstrap P-Donsker in probability. I include a proof of that in my book UCLT (both editions), although not the almost sure version which they also proved. I think statisticians generally apply the bootstrap for a fixed sample, as opposed to sequentially using it for increasing samples. $\mathcal{F}$ is called a *uniform Donsker class* if the convergence is also uniform in all P on the same σ-algebra in S. Evarist and Joel in 1991 characterized uniform Donsker classes as those which are uniformly pregaussian or finitely uniformly pregaussian.

I put a proof of the theorem into the second edition (2014) of my book. I also put in a proof of the Bousquet, Koltchinskii, and Panchenko (2002) theorem that the convex hull of a uniformly bounded uniform Donsker class is still uniform Donsker. It seems to me that in bootstrapping, little may be known about P except for the observed P_n, so the uniform Donsker property of $\mathcal{F}$ may be desirable, although it is highly restrictive, requiring $\mathcal{F}$ to be uniformly bounded up to additive constants. Pollard's entropy condition provides a useful sufficient condition for uniform Donsker.

Slow convergence in the central limit theorem for empirical processes. József Beck in 1985 (ZW) proved that for the uniform distribution P on the unit cube in $\mathbf{R}^d$, for $d \geq 2$, for the class $B(d)$ of all balls, which as a nicely measurable VC class of sets is a uniform Donsker class of functions, we have for the limiting Gaussian process G_P, $\sup_{B \in B(d)} |[\sqrt{n}(P_n - P) - G_P](B)|$ is no smaller than of order $n^{-1/(2d)}$, which seems disappointingly slow.

In a seminar in the Fall of 2006, which alternated between MIT and University of Connecticut, Storrs, we verified that Beck's proof was essentially correct although some details needed fixing. I and Richard Nickl, then a postdoc with Evarist, gave most or perhaps all of the talks, but Dmitry Panchenko and a then graduate student, Wen Dong, contributed useful ideas.

Richard was familiar with the very relevant series of papers by different authors called "Irregularities of distribution." There is a 1987 book with that title by Beck and W. W. L. Chen.

"Fast" convergence — $O((\log n)/\sqrt{n})$. For the 1-dimensional empirical process $\sqrt{n}(F_n - F)(x)$, the rate of convergence in sup norm to $b_{F(x)}$ for a Brownian bridge process b_t is $O((\log n)/\sqrt{n})$ as stated by Komlós–Major–Tusnády and proved with not too large constants by Bretagnolle and Massart. Just after the first edition of my book UCLT appeared (1999) I gave an exposition of the Bretagnolle–Massart theorem and proof with more details, which is now in the second edition. I also put in a proof of Massart's form of the Dvoretzky–Kiefer–Wolfowitz inequality with sharp constant, detailed except that in one step, a proof by calculus is replaced by numerical verification on a grid.

Quantiles and sample quantiles. If F is a continuous distribution function, strictly increasing on $F^{-1}((0,1))$, and $0 < q < 1$, the qth quantile of F is $F^{-1}(q)$. If one observes $X_1, ..., X_n$ i.i.d. for such an F, one can give confidence intervals for $F^{-1}(q)$ with endpoints order statistics of X_j, without bootstrapping, although the bootstrap gives a relatively easy, off-the-shelf method.

For an empirical distribution function F_n, eight books, two on the bootstrap and six beginning statistics textbooks, give 9 different definitions of sample qth quantile for $0 < q < 1$ as order statistics, e.g. $X_{(\lceil nq \rceil)}$, or as convex combinations of adjoining ones. The difference in definitions is not crucial here.

If instead of n we have R, the number of bootstrap replications, sometimes called B, two books on the bootstrap give definitions whereby the sample quantiles used are order statistics: B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, 1993, and A. C. Davison and D. V. Hinkley *Bootstrap Methods and Their Application*, 1997, who give

$$X_{(k)} \text{ for } k = (R + 1)q$$

if that is an integer, as they arrange for usual q's for example by setting

$$R = 999$$

so that $R + 1$ is a round number. Also, having R odd gives a nice sample median ($q = 1/2$).

Bootstrap confidence intervals

The percentile interval. Let T be a real-valued functional, defined on some set $\mathcal{P}$ of probability measures P on a set S, containing all possible empirical measures P_n. Let $X_1, ..., X_n$ be i.i.d. P for some $P \in \mathcal{P}$, and take the empirical measure P_n. To get an approximate confidence interval for $T(P)$, form i.i.d. bootstrap empirical measures P_{ni}^B for $i = 1, ..., R$ and let $T_i := T(P_{ni}^B)$, $i = 1, ..., R$. Form the order statistics $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(R)}$. Given $0 < \alpha < 1/2$, let T_L and T_U be the $\alpha/2$ and $1 - (\alpha/2)$ sample quantiles of the T_i.

Then Efron had defined $[T_L, T_U]$ as the $1 - \alpha$ two-sided *percentile* confidence interval for $T(P)$, as also in Efron and R. Tibshirani, 1993. I have not seen any precise, general statements as to under what conditions and with what accuracy the percentile interval, or for that matter any bootstrap-based interval, works.

The BC_a *interval.* Efron and Tibshirani defined, along with the percentile interval, the "Bias-corrected, accelerated" $\mathrm{BC_a}$ intervals. I won't give here the complicated-looking definitions, in which there are transformations to and from normal scale. The endpoints of the $\mathrm{BC_a}$ $1-\alpha$ intervals are some q_{lo}, q_{hi} sample quantiles of the T_i, but instead of $\alpha/2$ and $1-\alpha/2$, the q's are adjusted. The bias correction is based on the deviation from 1/2 of the fraction of the bootstrap sampled statistics T_i that are $\leq T(P_n)$. The "acceleration" is a correction for skewness.

The "basic" interval. If the bootstrap works, i.e. $a_n(T(P_n^B)-T(P_n))$ given P_n is close to $a_n(T(P_n') - T(P))$ in distribution, then since $\Pr(T_L < T(P_n^B) < T_U|P_n)$ has been estimated as $1-\alpha$, we would also approximate

$$\Pr(T_L - T(P_n) < T(P_n') - T(P)$$
$$< T_U - T(P_n)) \ \sim \ 1 - \alpha.$$

Davison and Hinkley (1997) proposed, it seems to me, to plug P_n in place of P_n' in the above approximation. That may be plausible, as in some ways, we are treating P_n as a typical value of P_n', but it seems to me it is not clearly justified. One might say that P_n is fixed and P_n' is independent of it?

The plug-in yields

$$\Pr\left(2T(P_n) - T_U < T(P) < 2T(P_n) - T_L\right) \sim 1 - \alpha.$$

Davison and Hinkley call the interval $[2T(P_n) - T_U, 2T(P_n) - T_L]$ the *basic* bootstrap confidence interval for $T(P)$. It provides an interesting alternative to the percentile and other intervals, to be tried out to see how the intervals compare. The basic and percentile intervals are reflections of one another in $T(P_n)$. The percentile, basic, BC_{a}, and a normal interval are all implemented in the R system library "boot" and treated by W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th Ed. 2002. Venables and Ripley, and others, prefer the basic interval.

But Davison and Hinkley (1997) (its inventors?) do not find it best at all. They say "The studentized bootstrap and adjusted percentile [i.e. $\mathrm{BC_a}$] methods for calculating confidence limits are inherently more accurate than the basic bootstrap and percentile methods." They empirically compare 8 different bootstrap-based confidence intervals for the functional

$$T(P,Q) = \textstyle\int x\,dP(x)/\int y\,dQ(y)$$

for P and Q specified gamma distributions. In this case both the lower and upper endpoints of the basic interval tend to be substantially too small. The errors for the basic interval are the worst of those shown. "Studentized, log scale" intervals perform well.

The basic interval can give strange endpoints. Suppose for example that T takes values any positive number $0 < T < +\infty$. Then $0 < T_L < T_U$. But $2T(P_n) - T_U$ can be negative. Also, if $T(P_n) < T(P)/2$, then $T(P)$ will never be in the basic interval, for arbitrarily small α. Letting $\alpha \downarrow 0$, use of the basic interval seems to entail the usually untrue statement

$$\Pr(T(P_n) < T(P)/2) = 0,$$

perhaps a contradiction.

Neither of these possibilities can occur for the percentile nor BC_{a} intervals, with endpoints order statistics of the bootstrap sample, unless $T(P) < T_{(1)}$ or $T_{(R)} < T(P)$, which would rarely occur for the $R \geq 1000$ generally used.

In teaching about the bootstrap since 2007 I had recommended the basic interval, following Venables and Ripley. Now, I think it may not even be the second-best of the intervals mentioned.

For $p > 1$ the $Pareto(p)$ distribution has a density $(p-1)/x^p$ for $x \geq 1$ and 0 for $x < 1$. It has a finite mean if and only if $p > 2$ and finite variance if and only if $p > 3$. For sample means of 100 i.i.d. Pareto (7/2) variables I found by simulation that the basic interval has significantly worse coverage properties than the percentile interval.

In Efron and Tibshirani's Section 13.4, titled "Is the percentile interval backwards?", as one book called it, they say that sometimes, it's rather the basic interval that is backwards (skewed in the wrong direction), as I found for the mean in the Pareto (7/2) case.

Efron and Tibshirani, §13.4, say neither the percentile nor the basic interval works well in all cases. In §14.3 they say that the BC$_{\mathrm{a}}$ method has "two significant theoretical advantages." For one, it is *transformation respecting*, meaning preserved by (monotone) transformations of T [such as taking $\log(T)$ if $T > 0$], as the percentile intervals also are. The other stated advantage is *second-order accuracy*, meaning that coverage probabilities are approximated within $O(1/n)$ as opposed to $O(1/\sqrt{n})$ for first-order accuracy with the other methods.

There are bootstrap confidence intervals assuming approximate normal distribution of the T_i, using normal or t distributions, but for lack of time I won't say much more about these. Venables and Ripley, p. 136, say that if the distribution of the T_i is asymmetric, "intervals based on normality are not adequate."

Bootstrapping of the mean. Here one considers means $T(P) = \int f\,dP$ for a fixed function f, so that $T(P_n)$ are sample means. The bootstrap works in this case by the Giné–Zinn theorem if $f \in \mathcal{L}^2(P)$. They showed in a 1989 paper that the weaker condition that f be in the domain of attraction of a normal law suffices. Athreya (1987) had shown that if the distribution of f for P is in the domain of attraction of a stable law of index α with $1 < \alpha < 2$, then the bootstrap fails, as the sample means can be centered and normalized to have a limiting α-stable law, but corresponding bootstrapped sample means have a random limit law depending on the sequence $X_1, X_2, \ldots$.

Using Athreya's result, Giné and Zinn (1989) proved that for the bootstrap to work, being in the domain of attraction of a normal is necessary. For statistics, it seems that the main interest will be in $f \in \mathcal{L}^2$, or better.

In 200 experiments on 90% confidence intervals for the mean for Pareto(9/2), again with $n = 100$, the number of cases of non-coverage by the normal, basic, percentile, and BC$_{\text{a}}$ intervals respectively was 33, 37, 33, and 31 respectively. None of these is consistent with true average coverage probability of 0.9; BC$_{\text{a}}$ comes closest.

Conditioning on disagreement in coverage between the kinds of intervals, which occurred in just 16 of the 200 experiments, the apparent superiority of the BC$_{\text{a}}$ (relatively best, 11 coverages) over the basic (relatively worst, 5 coverages) interval was not significant in itself, even without a correction for multiple testing.

Bootstrap of the variance

Googling “Bootstrap of the mean” (with the quotes) gave me 47,300 hits (although beyond some point, Google may begin to disrespect the literal quotes), but “bootstrap of the variance” gave just two (4 June 2014). The topic is mentioned in Davison and Hinkley. One of the two papers was on time series, leaving only one about i.i.d. data, a Report from Uppsala University by S. Amiri, D. von Rosen, and S. Zwanzig (2008) “On the comparison of parametric and nonparametric bootstrap,” concerned with estimates and not, as far as I saw, with confidence intervals.

If T is the variance functional $T(P) = \mathrm{Var}_P(x)$ for distributions P on the line,

$$T(P_n) = {s'_X}^2 := \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2.$$

It is well known that $ET(P_n) = \frac{n-1}{n}T(P)$. It follows, for P_n in place of P, that

$$E(T(P_n^B)|P_n) = \frac{n-1}{n}T(P_n).$$

The unbiased sample variance is a U-statistic,

$$\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2 =$$

$$\frac{1}{\binom{n}{2}} \sum_{1 \le i < k \le n} \frac{1}{2}(X_i - X_k)^2.$$

I have that in §11.9 on U-statistics of my book *Real Analysis and Probability*, 1989; 2002 Cambridge University Press.

My source on history of the bootstrap of U-statistics was Evarist's St.-Flour Lectures from 1996. The central limit theorem for nondegenerate U statistics of order 2, as this one is, holds under a hypothesis which in this case amounts to $E(X_1^4) < \infty$. It follows from results of von Mises (1947) on what are now called V-statistics. The corresponding bootstrap central limit theorem for order 2 also holds under the same hypothesis as shown by Bickel and Freedman, Ann. Stat. 1981.

For small n, the biases are not negligible. They are nonparametric, holding for all distributions with finite variance.

I did 100 experiments, in each of which I generated 20 i.i.d. $N(0,1)$ variables, and noted which of the normal, basic, percentile, and $\mathrm{BC_a}$ 90% intervals for the variance covered the target $\sigma^2 = 1$. There were respectively 19, 20, 19, and 11 cases of non-coverage. Seemingly because of the bias(es), the normal, basic, and percentile intervals for the variance are to the left of where they should be and have inferior coverage frequencies.

The probability of 81 or fewer successes in 100 independent trials with probability 0.9 of success on each is 0.00458. Multiplying by 3 to correct for multiple tests (Bonferroni) gives 0.0137. So there is evidence that the normal, basic, and percentile intervals with target 90% confidence do not attain the target in this case. The BC_a interval does much better. As its name implies, it does bias correction. A confidence interval for its coverage probability is $[0.828, 0.932]$, including 0.9.

I took $n = 20$ as suggested by the Wikipedia article *Bootstrapping (statistics)*, accessed May 26, 2014, which said that the coverage probability of the percentile 90% confidence interval in this case is 0.78. The 0.81 I found gives a 90% confidence interval *for that coverage probability* of $[0.738, 0.866]$, which indeed contains 0.78. The article favors the basic interval, but does not give a number for it to compare to 0.78. From my data, it seems no better.

Davison and Hinkley give a table of coverage probabilities of "confidence intervals for normal variance ... for 10 samples each of size two," in which the BC_a interval does very well, much better than the basic or percentile intervals.

One should consider non-normal distributions of X_j. I experimented with i.i.d. Pareto (11/2) variables, which do have fourth moments, $EX_1^4 = 9$. I took samples of $n = 200$ of them at a time, checking the coverage of the normal, basic, percentile and $\mathrm{BC_a}$ target 90% confidence intervals for the variance. In 100 experiments, non-coverage of σ^2 occurred respectively 29, 33, 29, and 22 times. There is a strong tendency for intervals to be to the left of where they should be. Of the total 113 non-coverages, in 111 the right endpoint of the interval was less than σ^2. The other two cases were both for the $\mathrm{BC_a}$ interval.

The target fraction of non-coverages on each side is .05, so one could say it's a merit of the BC_{a} interval that on the right it had two and the others none. The BC_{a} interval, intended to compensate for bias and skewness, has only partial success in this case; it is only relatively better than the other intervals.

In 5 of the 100 experiments, the sample variance $T(P_n)$ was less than half the true variance $T(P)$, meaning that the basic $1 - \alpha$ interval could not have covered $T(P)$ no matter for how small $\alpha > 0$.