10 Lecture 10: Introduction to the Bootstrap

10.1 Motivation

In Statistical Inference we are "learning from experience": we observe a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and wish to infer properties of the complete population that yielded the sample. A complete knowledge about the population is obtained from the *population distribution function* $F(.)$. This (unknown (!)) function is usually estimated in a non-parametric framework (i.e. when no additional information about the population is available except the sample itself) by the *empirical distribution function* (EDF)

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i).$$

Basically, when calculating the EDF, we look at the empirical proportion of the realizations in the sample that happen to be $\leq x$ and use this expirical proportion to *estimate* the true unknown probability $F(x)$ for a realisation in the interval $(-\infty, x]$. The remarkable fact about the EDF is that despite its simplicity, it has been defended over the years by prominent statisticians as *the* asymptotically optimal estimator of the population distribution function. A statement first proved in a paper by Dvoretsky, Kiefer and Wolfowitz in 1956, it has been shown that the EDF is *asymptotically minimax* among the collection of all continuous distributions. As Millar (1979) notes, "This paper has stood for over 20 years as one of the pivotal achievements of nonparametric decision theory".

We will not discuss the precise meaning of the above mentioned asymptotic optimality in this course. Note that it goes about estimating a function (not a finite-dimensional parameter) and justifying the optimality requires defining it in a suitably chosen function space, with a suitable loss over the functions in this space etc. This requires some preparation in itself. The convergence of

$$\sqrt{n}(\hat{F}(x) - F(x))$$

towards a limiting Gaussian process (See Lecture 8) helps and is utilised but you probably can appreciate that the details are subtle and are skipped here.

However, encouraged by the existence of such results in the literature, we can be tempted to estimate some interesting aspect of $F(.)$ (such as its mean or median, higher order moment etc.) by using the corresponding aspect of $\hat{F}(.)$. For example, we would be tempted to estimate the population mean $\mu = \int x dF(x)$ by the sample mean $\bar{X} = \int x d\hat{F}(x)$. This approach to estimation is called *plug-in principle* (we plug-in the good estimator $\hat{F}$ instead of F in the formula for the theoretical mean and hope to get a good estimator of the theoretical mean itself). In this example, we *were* successful- we got the sample mean which is known to be a good estimator of the theoretical mean. More generally, any parameter of interest θ can usually be written as a statistical functional $\theta = t(F)$ and the obvious plug-in point estimator for θ would be $\hat{\theta} = t(\hat{F})$ then.

The (nonparametric) *bootstrap method* is an *application of the above discussed plug-in principle*. Originally, the bootstrap was suggested by B. Efron as a method to derive an estimate for the the *standard error* of arbitrary estimator. Finding the standard error of

an estimator is a significant activity for every statistician since, not being satisfied with a point estimator only, he/she is always looking for the *variablity* of the estimator. He/she would be interested in the bias, standard error or *even in the complete distribution* of the estimator itself. If available, these can be used to construct confidence intervals, to test hypotheses etc. for the parameter of interest.

Since theoretical investigation of the standard error is possible only in a limited number of "textbook cases" and even in these "textbook cases" the treatment is only asymptotic, Efron's idea was to use the bootstrap as a means to get standard error estimates of $\hat{\theta} = t(\hat{F})$ in an "*automatic way*", no matter how complicated the functional mapping $\theta = t(F)$ may be. We describe the idea below.

10.2 Nonparametric bootstrap.

In the original settings of the (nonparametric) bootstrap we assume that we need to estimate the *standard error* of the estimator $\hat{\theta} = t(\hat{F})$. *If we had several estimators* of size n of $t(F)$ then of course we could use their empirical standard error as an estimator for the unknown standard error! Unfortunately, we only have got *one sample* of size n that allows us to construct *one* plug-in estimator $\hat{\theta} = t(\hat{F})$. How could we the evaluate the sampling accuracy of this estimator when we only have one only realisation? The bootsrap helps us here in a very simple manner by helping us to generate **many** samples that look like the original sample hence to get many versions of the estimator by evaluating it on each of these samples.

To this end we consider the so called *bootstrap sample* of size n drawn *from* $\hat{F}$ instead of F! It is defined as $\mathbf{x}^* = (x_1^*, x_2^*, .., x_n^*)$, the star indicating that this is not the original data set but a randomized, or *resampled* version of **x.** Hence the bootstrap data points are a random sample *with replacement* from the population of n objects $(x_1, x_2, .., x_n)$ rather than from the original population (but we use them as if they were a new sample from the entire population). Correspondingly to the bootstrap sample, we can get a *bootstrap replication* $\hat{\theta}^*$ via plug-in. The process of getting new bootstrap samples and then plug-in, can be repeated as many times as we like! As a consequence, we get B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, .., \mathbf{x}^{*B}$, each consisting of n data values drawn with replacement from $\mathbf{x}$. To each such sample, we calculate the corresponding $\hat{\theta}^*(b), b = 1, 2, .., B$ by the plug- in method and finally, we can calculate an estimator of the standard error $s_F(\hat{\theta})$ by applying the classical formula $\hat{s}_B = \{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \bar{\hat{\theta}^*}]^2/(B-1)\}^{1/2}, \bar{\hat{\theta}^*} = \sum_{b=1}^{B}\hat{\theta}^*(b)/B$.

Note that for a given sample size n, letting $B \to \infty$ means that we get in a limit the ideal bootstrap estimate $s_{\hat{F}}(\hat{\theta}^*)$ of $s_F(\hat{\theta})$, i.e. $\lim_{B\to\infty}\hat{s}_B = s_{\hat{F}}(\hat{\theta}^*)$. This ideal bootstrap estimate $s_{\hat{F}}(\hat{\theta}^*)$ is called a *nonparametric bootstrap estimate of the standard error*. In computer simulations, using the cheap computer power, the value B can be taken very large so that indeed $\hat{s}_B \approx s_{\hat{F}}(\hat{\theta}^*)$ holds and because of this, sometimes the value $\hat{s}_B$ itself is called a *nonparametric bootstrap estimate of the standard error*. Our hope is that when n is large and B is large, $\hat{s}_B$ will be close to $s_{\hat{F}}(\hat{\theta}^*)$ which will be close to $s_F(\hat{\theta})$. Much of the theory in bootstrap has been developed to justify such type of statements. From the discussion so far we see that the bootstrap can be considered as a *data- based simulation method for statistical inference*.

The use of the term bootstrap derives from the phrase *to pull oneself up by one's*

bootstrap from the eighteenth century adventures of Baron Münhausen. The method is extremely powerful and has made Efron an instant celebrity. Efron mentions that he thought about calling the method the *shotgun* since it " .. can blow the head off any problem if the statistician can stand the resulting mess". This quotation relates to bootstrap's wide applicability in combination to the, generally speaking, large volume of numerical work associated with its application.

It should be noted that when enough bootstrap resamples have been generated, *not only the standard error* but *any aspect of the distribution of the estimator* $\hat{\theta} = t(\hat{F})$ *could be estimated!* One can, for example, have a *histogram* of the distribution of $\hat{\theta} = t(\hat{F})$ by constructing a histogram based on the observed $\hat{\theta}^*(b), b = 1, 2, .., B$ values! This histogram can be used as an estimator of the density of the es timator $t(\hat{F})$. Specifically, the histogram would be very useful since it will indicate, for example, if normal approximation to the distribution of $t(\hat{F})$ is justified for the given fixed sample size n.

10.3 Parametric bootstrap

Of course, it is to be expected that in situations where there **is** more information about the population's distribution function F other than that provided by the sample, the plug-in principle would not be very good and the standard bootstrap method will need some modifications in order to be applied. For example, if there existed a *parametric model* for F, exact (or approximate asymptotic) analytic formulae may exist for the standard errors of some estimators and we could use them instead of the nonparametric bootstrap. But even in parametric situations the bootstrap's idea *still* can be used *parametrically*. We can apply the so called *parametric bootstrap* and the results are similar (and sometimes even better) than the textbook formulae. The parametric bootstrap estimate of standard error is defined as $s_{\hat{F}_{\text{par}}}(\hat{\theta}^*)$ where $\hat{F}_{\text{par}}$ is an estimate of F derived from a *parametric model of the data*. That is, in this case, instead of drawing independent samples with replacement from the data, we draw B samples of size n from the parametric estimate of the function F. After generating the B samples, we proceed in the usual way outlined in *1.* to calculate the estimator of the standard error. When used in the above parametric mode, bootstrap can provide more accurate results than approximate textbook formulae results based on asymptotic normality approximations and it also provides answers in problems for which no textbook formulae exist.

10.4 Numerical illustration

(B. Efron and R. Tibshirani give the example below to illustrate a non-trivial application of the bootstrap and to compare the accuracy of the method with the accuracy of alternative methods that exist to deal with this example). The goal is to **evaluate standard error of the estimator of the correlation coefficient**. A sample of size $n = 15$ is available from a set of American law schools participating in a large study of admission practices.

School	LSAT (X)	GPA (Y)
1	576	3.39
2	635	3.3
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

Two measurements are available on the entering classes of each school : LSAT and GPA. It is believed that these scores are highly correlated. One is interested in estimating the correlation coefficient **and** in obtaining an estimate of the standard error of the estimator. **Precise** theoretical formula for the standard error of the estimator is **unavailable** but an **asymptotic approximation formula exists**. As an alternative to it, and especially when the sample size is **not very large** ($n = 15$ only), the bootstrap estimate is considered. The numerical values are compared and show very close values.

A bivariate scatterplot indicates relatively strong linear relationship. Methods for testing bivariate normality suggest the data is likely to have been generated from a bivariate normal distribution. The parameter of interest is the correlation coefficient $\rho = \frac{E[(X-EX)(Y-EY)]}{\{E[X-EX]^2.E[Y-EY]^2\}^{1/2}}$. Its typical estimator is $\hat{\rho} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{[\Sigma x_i^2 - n\bar{x}^2]^{1/2}.[\Sigma y_i^2 - n\bar{y}^2]^{1/2}} = .776$ (the sample correlation coefficient).To estimate its standard error, B bootstrap samples of $n = 15$ points selected at random with replacement from the actual sample are performed and the coefficients $\hat{\rho}^*(1), \hat{\rho}^*(2), .., \hat{\rho}^*(B)$ are obtained. Efron and Tibshirani experiment with different values of B. They notice a stabilization of the empirical standard errors of these B bootstrap replications towards 0.132. Closeness is observed already at values of B around 1000. This illustrates the **nonparametric** bootstrap approach.

To perform a **parametric bootstrap** in this example, we assume that the the LSAT and the GPA results do follow a bivariate normal distribution. We estimate this distribution by $\hat{F}_{\text{norm}}$ via substitution of the empirical estimators of the mean vector and of the covariance matrix in the formula for the bivariate normal density. Then B samples of size 15 each from $\hat{F}_{\text{norm}}$ are simulated and the correlation coefficient is computed for each sample. The parametric bootstrap estimate for $B = 3200$ repetitions has been .124, close to the value obtained by the nonparametric bootstrap.

Finally, for this example, there exists a celebrated theoretical result which states that asymptotically, for a sample from a bivariate normal, the standard error of the empirical correlation coefficient is approximated by $(1 - \hat{\rho}^2)/\sqrt{n-3}$. Substituting $\hat{\rho} = .776$ gives .115 in our case. Furthermore, another celebrated result in this direction is **Fisher's** z

transformation (a variance stabilizing transformation (!)) which implies that

$$z = 0.5.\log(\frac{1+\hat{\rho}}{1-\hat{\rho}}) \sim N(0.5\log(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

holds asymptotically. This result is used to make inference (test hypotheses, construct confidence intervals etc.) for ρ by first making inference about $0.5\log(\frac{1+\rho}{1-\rho})$ and then transforming back to inference about ρ. The empirical standard deviation of the 3200 z values obtained from the transformation of the parametric bootstrap's $\hat{\rho}^*$ values has been equal to .290 which is very close to the value $\frac{1}{\sqrt{15-3}} = .289$.

These numerical values are quite convincing about the accuracy of the bootstrap values obtained (where **no textbook results are necessary to be used !**). Also, one more merit of bootstrap should be appreciated. Note that one of the main reasons for making parametric assumptions in traditional statistical inference is to facilitate the derivation of analytically tractable formulae for standard errors. This restricts the applicability (we could only apply the parametric method to cases where indeed its assumptions are believed to hold for the data available (!)) while still possibly leading to quite painful theoretical derivations. **But in bootstrap approach** we **do not need** these formulae, hence we can also **avoid making** restrictive parametric assumptions.

10.5 Bootstrap estimate of bias

The bootstrap was first introduced as a method for evaluating standard errors of estimators. It should be noted, however, that it can easily be adapted to estimate the **bias** of an estimator and therefore can be applied as a bias correction procedure.

To illustrate this, let us assume again that a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is available from an unknown distribution F and $\theta = t(F)$ is a parameter to be estimated by a statistic $\hat{\theta} = s(\mathbf{x})$. We would like to estimate the (unknown) bias $b_F(\hat{\theta}, \theta) = E_F[s(\mathbf{x})] - t(F)$. The bootstrap estimate of the theoretical bias is obtained naturally by plugging-in $\hat{F}$ instead of F in the above bias formula, i.e.: $b_{\hat{F}}(\hat{\theta}^*, \theta) = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})$. Hence, we can get the *bias-adjusted* estimate as:

$$t(\hat{F}) - \{E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})\} = 2t(\hat{F}) - E_{\hat{F}}[s(\mathbf{x}^*)].$$

It is an easy exercise for you to show that if the parameter θ is *the mean :* $\theta = t(F) = \int xdF(x)$ and $\hat{\theta} = \bar{\mathbf{x}}$ then $b_{\hat{F}}(\hat{\theta}, \theta) = 0$ and there is no bias correction necessary. This is but a rather exceptional case. In most other situations a bias will exist and bias correction via $b_{\hat{F}}$ makes sense. This is demonstrated with the next example.

Example: When estimating the variance σ^2 by $s(x) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ the bias is well known to be $-\frac{1}{n}\sigma^2$ and in this case $b_{\hat{F}}(s^*, \sigma^2) = -\frac{1}{n^2}\sum_{i=1}^n (x_i - \bar{x})^2$ (show it (!)). After the bias correction: $s(x) - b_{\hat{F}}(s, \sigma^2) = \frac{n+1}{n^2}\sum_{i=1}^n (x_i - \bar{x})^2$ and you can see that now the bias of the corrected estimator is $-\frac{1}{n^2}\sigma^2$! This is of much smaller order than the bias $-\frac{1}{n}\sigma^2$ before the corrective action thus illustrating the effect of the bias correction.

In the above example, because of its simplicity, we were able to give a **theoretical** bootstrap estimate of the bias. In more involved situations this will not be possible but again simulation can help us to avoid the difficulties! We can do the bias correction without any theoretical derivations by simply:

- generate B independent bootstrap samples of size n;
- evaluate the bootstrap replications $\hat{\theta}^*(b) = s(\mathbf{x}^{*b}), b = 1, 2, .., B$;
- approximate the bootstrap expectation $E_{\hat{F}}[s(\mathbf{x}^*)]$ by $\frac{1}{B}\sum_{b=1}^{B} s(\mathbf{x}^{*b})$;
- get the estimate (approximation) of the bootstrap estimate of bias based on B replications by $\hat{b}_B(\hat{\theta}, \theta) = \frac{1}{B}\sum_{b=1}^{B} s(\mathbf{x}^{*b}) - t(\hat{F})$

Of course, in the last step we could calculate **both** the estimated standard deviation $\hat{s}_B$ **and** $\hat{b}_B(\hat{\theta}, \theta)$ from the same set of bootstrap replications thus being in a position to give a bootstrap estimate of the mean squared error of the estimator,too.

10.6 The jackknife estimate of bias.

In fact , the original method proposed for bias reduction was not the bootstrap but the **jackknife** (M. Quenouille as early as in the mid $1950's$). We shall discuss it briefly here and also its relation to the bootstrap will be pointed out. Given the original sample $\mathbf{x}$, one defines n *jackknife samples* $\mathbf{x}_{(i)} = (x_1, x_2, .., x_{i-1}, x_{i+1}, .., x_n)$ and the ith *jackknife replication* $\hat{\theta}_{(i)}$ of the statistic $\hat{\theta} = s(\mathbf{x})$ is defined as $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$. Then the *jackknife estimate of bias* is defined as $\hat{b}_{\text{jack}} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta})$ where $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$. Note also that the jackknife method has been also extended to estimating the standard error by using the formula

$$\hat{s}_{\text{jack}} = [\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2]^{1/2}$$

Finally , one **warning** should be given: **sometimes** bias correction could be dangerous in practice. It may happen that it reduces the bias at too high a price,namely by increasing the variance quite significantly thus increasing the mean squared error. Fortunately, we have means to check if this unpleasant effect has occurred! Indeed, (again using the bootstrap), we can estimate the mean squared error, as well! Then, bias correction should be only applied when it does not blow up significantly the estimated mean squared error.

10.7 Relation of bootstrap and jackknife methods.

The bootstrap and the jackknife **standard error estimators** look also quite similar except that the factor $(n-1)/n$ in the jackknife estimator formula is much larger than the $1/(n-1)$ factor in the bootstrap estimator formula. This is so since the jackknife deviations $(\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2$ tend to be smaller than the bootstrap deviations $[\hat{\theta}^*(b) - \bar{\hat{\theta}}^*]^2$ (intuitively, the jackknife sample is more similar to the original sample than is the bootstrap sample). More refined arguments show that the jackknife can be considered as a **linear approximation to the bootstrap:** i.e. it agrees with bootstrap for a certain *linear statistic in the form* const$+\frac{1}{n}\sum_{i=1}^{n} \alpha(x_i)$ that approximates the possibly nonlinear statistic $\hat{\theta}$. In fact, the accuracy of the jackknife estimate depends essentially on how close is $\hat{\theta}$ to a linear statistic. In terms of our discussion in the robustness lecture, if the remainder in

the approximation of the functional $T(F)$ via $T(F_n)$ "behaves well" then the results when using bootsrap and jackknife for bias correction are very similar.

As an upshot, the jackknife provides a simple approximation to the bootstrap for estimating bias and standard error. However, the jackknife can fail if the statistic is significantly non-linear.

10.8 Confidence intervals based on the bootstrap.

The ultimate goal in evaluating the standard deviation of an estimator is to utilize this standard deviation for constructing *confidence intervals*. Given $\alpha \in (0,1)$ the naive **asymptotic** $(1-\alpha).100\%$ confidence interval would be $[\hat{\theta} - z_{\alpha/2}.s_F(\hat{\theta}), \hat{\theta} + z_{\alpha/2}.s_F(\hat{\theta})]$. Despite its simplicity, this confidence interval does not have much merit. It has only approximate $(1-\alpha).100\%$ coverage since the $(1-\alpha).100\%$ coverage is **only** obtained in a limit. It is possible to derive **better** (still approximate but with better coverage accuracy) bootstrap confidence intervals based on the following observation. The distribution of $Z = \frac{\hat{\theta}-\theta}{s_F(\hat{\theta})}$ would not follow exactly the $N(0,1)$ law **but this very same distribution can be estimated directly from the data at hand.** A corresponding bootstrap table can be constructed by generating B bootstrap samples that give rise to B different Z values (the empirical percentiles can be obtained from the empirical distribution of these Z values). More specifically, we:

- generate B bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, .., \mathbf{x}^{*B}$
- for each sample, we calculate $Z^*(b) = (\hat{\theta}^*(b) - \hat{\theta})/\hat{s}^*(b), b = 1, 2, .., B$
- evaluate the αth percentile by the value $\hat{t}^{(\alpha)}$such that $\#(Z^*(b) \leq \hat{t}^{(\alpha)})/B = \alpha$.
- the **bootstrap t confidence interval** at level $(1-2\alpha)$ is then $(\hat{\theta} - \hat{t}^{(1-\alpha)}s_F(\hat{\theta}), \hat{\theta} - \hat{t}^{(\alpha)}s_F(\hat{\theta}))$(or with the empirical value $s_B(\hat{\theta})$ substituted if $s_F(\hat{\theta})$ or $s_{\hat{F}}(\hat{\theta}^*)$ unknown).

A subtle theoretical investigation of the procedure shows that in large samples the coverage of the bootstrap t based confidence interval is closer to the desired level than the coverage of the standard normality based interval or the interval based on the t-distribution table. The price of this accuracy is that generality is lost- the bootstrap t table can be applied **only for the data at hand.** The table will need to be generated from scratch for a new set of observations / new problem. Note also that the bootstrap t percentiles **are not necessarily symmetric around zero** and the resulting confidence intervals are not necessarily symmetric around $\hat{\theta}$!

The following problems are encountered when applying the bootstrap t and should be mentioned here:

i) for more complicated statistics than the arithmetic mean, we need to estimate $\hat{s}^*(b)$ and this needs in fact a **second nested level of bootstrapping.** This might increase the computational costs **dramatically.**

ii) for small samples, the bootstrap t interval may perform erratically. It may be preferable to work with the **bootstrap distribution of** $\hat{\theta}$ itself instead of working with

the bootstrap distribution of Z. This leads to the so called "BC_a" procedure. Its description is outside the scope of this course, however.

10.9 Extensions of bootstrap outside the i.i.d. setting

Many practically important models are **not** based on the assumption of availability of i.i.d. observations. These models are exactly the more complicated ones for which no textbook solutions for standard errors of estimators etc. exist. Hence extensions of the bootstrap method beyond the i.i.d. setting become of crucial importance. We do not have time to discuss these extensions here but will mention that many such extensions exist and cover situations such as:

- regression (model-based bootstrap)
- autoregressive type time series (block bootstrap)
- other weakly dependent series (sieve bootstrap)
- bootstrap in the frequency domain
- bootstrap methods tailored specifically for Markov Processes
- bootstrap for long range dependent data
- bootstrap for spatial data

These more involved extensions of the bootstrap may not yet exist in traditional packages such as SAS of SPLUS but have realisations in packages such as *R*.