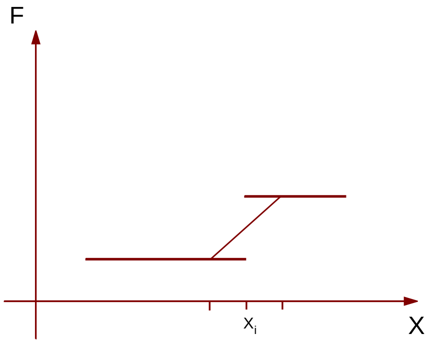

Figure 4.1: Smoothing the empirical distribution function.

4.5 The smoothed bootstrap

In the simple nonparametric bootstrap we have assumed that the empirical distribution assigning equal mass to each observation, $\hat{F}$, is a suitable estimate of F. However, $\hat{F}$ is discrete and it is natural to ask if a smooth estimate of F might be better, particularly when we expect F to be continuous.

$\hat{F}$ is the c.d.f. which places an atom of probability with mass $\frac{1}{n}$ to each data point. Smoothing consists of replacing each data point with a continuous distribution of total mass $\frac{1}{n}$ centered at the point. The most common smoothing distribution is uniform on interval $[-h, h]$. The uniformly smoothed empirical c.d.f. $\hat{F}^U$ is similar to $\hat{F}$ except that the jumps of size $\frac{1}{n}$ at each data point are replaced by straight lines with slope $\frac{1}{2nh}$ which pass through the midpoint of the jump. See Figure (4.1).

Another common smoothing distribution is $N(x_i, h^2)$.

In any case, there is the question of the choice of h. If it is too small, then the resulting distribution will not be very smooth, if it is too large, then the smoothed portions overlap and we loose information given in the original data.

Simulation from $\hat{F}^U$ proceeds in two stages:

1. generate a bootstrap sample in the usual way,
2. add a simulated r.v. from $Uniform[-h, h]$ to each member of the bootstrap sample.

Then calculate bias and variance of and estimator θ as in the simple bootstrap.

4.6 The balanced bootstrap

Since bootstrap samples are chosen randomly and independently, 'unrepresentative' collections of samples may occur, that is some values may occur many more times than other. In a balanced bootstrap, each of the n observations is constrained to occur exactly N times in the N samples. Hence, each bootstrap sample is a random sample from $\hat{F}$ but the samples are no longer independent (for example, knowing the first $N-1$ samples tell us what the Nth sample must be).

To implement a balanced bootstrap we need a random permutation of the vector

$$(1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots, n, n, \ldots, n)'$$

In the randomized vector we use the first n entries to index the first bootstrap sample, the next n entries to index the second bootstrap sample, and so on. For example, let $n = 10$ and $N = 2$ and let the sample from a population be (see Practical 5) :

9.6 10.4 13.0 15.0 16.6 17.2 17.3 21.8 24.0 33.8

The randomized vector might be

$$(2, 2, 3, 8, 4, 6, 1, 9, 7, 4, 5, 6, 1, 5, 8, 9, 3, 10, 10, 7)'$$

which gives the following two bootstrap samples:

10.4 10.4 13.0 21.8 16.6 17.2 9.6 24.0 17.3 15.0

16.6 17.2 9.6 16.6 21.8 24.0 13.0 33.8 33.8 17.3

Using the frequencies of occurrence we may put the bootstrap samples in the following table:

data	9.6	10.4	13.0	15.0	16.6	17.2	17.3	21.8	24.0	33.8
frequencies	1	2	1	2	0	1	1	1	1	0
frequencies	1	0	1	0	2	1	1	1	1	2

Now, consider $\hat{\theta} = \bar{X}$. The value of the estimator obtained from the original sample is $\bar{x} = 17.87$, the bootstrap replications of the estimate are: $\bar{x}_1^* = 15.37$ and $\bar{x}_2^* = 20.37$ which give the mean of $\bar{x}^* = 17.87$. The balanced bootstrap forces the average value of $\theta_i^* = \bar{x}_i^*$ to be the same as the value of $\hat{\theta} = \bar{X}$.

4.7 Bootstrapping bivariate data

Suppose we have bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, a sample of bivariate i.i.d. random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ each with the same c.d.f. $F_{X,Y}$. The way of bootstrapping this kind of data depends on what we know about the relationship between X and Y.

4.7.1 Non-parametric bootstrap

This method is appropriate when the pairs (x_i, y_i) are the random sample, we have no prior control over the values of r.vs X and Y and the model of Y in terms of X is either unknown or theoretically untractable.

Bootstrapping:

- For each bootstrap sample randomly choose n numbers $j_i, j_2, \ldots, j_n$ from $\{1, 2, \ldots, n\}$ with replacement.
- Then, the bootstrap sample consists on the pairs $(x_{j_1}, y_{j_1}), (x_{j_2}, y_{j_2}), \ldots, (x_{j_n}, y_{j_n})$ (x and y get bootstrapped together).
- Calculate the replications of the parameter estimate for each bootstrap sample and average the replications.

4.7.2 Fully parametric bootstrap

Suppose we know that
$$Y_i = f(X_i, \psi) + \varepsilon_i,$$
where the r.v. ε follows a distribution which belongs to a known parametric family of distributions. For example
$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

However, we neither know the values of the parameters ψ nor the error distribution parameters (α, β, σ^2 in the example). If we can control the values of r.v. X then we obtain a bootstrap sample by fixing each value x_i and simulating y_i from the fitted model $\hat{Y}_i$. For example from

$$N(\hat{\alpha} + \hat{\beta} x_i, \hat{\sigma}^2),$$

where $\hat{\alpha}$, $\hat{\beta}$and $\hat{\sigma}^2$ are preliminary estimates of the parameters.

Here we fit the model to the observed data, then simulate random samples from the fitted model, and refit the model with the simulated samples.

4.7.3 Semi-parametric bootstrap

Suppose that we know the expected model, i.e., $E(Y_i) = f(X_i, \psi)$ up to the model parameters ψ but we do not know the distribution of the errors $Y_i - E(Y_i)$. If we can control the values of X then we can bootstrap in the following way:

- Estimate the parameters ψ and obtain the residuals $r_i = y_i - \widehat{E(Y_i)}$.
- Fix the x_i values and bootstrap the residuals, i.e., randomly choose n numbers $j_i, j_2, \ldots, j_n$ from $\{1, 2, \ldots, n\}$ with replacement and put $(x_i, \widehat{E(Y_i)} + r_{j_i})$ for $i = 1, \ldots, n$ as a bootstrap bivariate sample.

Here we fit the expectation part of the model, then resample with replacement from the residuals, add the resampled residuals to the fitted expectation to get bootstrap samples, to which we refit the model.

Example of semi-parametric bootstrap
In recent years, physicians used the *dividing reflex* to reduce abnormally rapid heartbeats in humans by briefly submerging the patient's face in cold water. The reflex, triggered by cold water temperatures, is an involuntary neural response that shuts off circulation to the skin, muscles, and internal organs and diverts extra oxygen-carrying blood to the heart, lungs and brain. A research physician conducted an experiment to investigate the effects of various cold water temperatures on the pulse rate of small children. From his earlier experience, the physician knew that the expected pulse rate may be modeled as a linear function of water temperature, however he had no information about the measurement error distribution.

The relationship between X (water temperature) and Y (pulse rate) is assumed to be linear, so

$$E(Y) = \alpha + \beta X.$$

However, there is no information about the error distribution. We will use semi-parametric bootstrap.

GenStat program doing the calculations will be shown in the lectures.

4.7.4 Summary of the bivariate bootstrap

Relationship between X and Y	X controlled	X not controlled
known apart from the parameters	regression or GLM or parametric bootstrap	bootstrap x_i, then do parametric bootstrap
$E(Y_i)$ known apart from the parameters	fix the x_i and bootstrap the residuals	bootstrap x_i, then bootstrap the residuals
unknown	?	bootstrap the pairs (x_i, y_i)

So far we have used the bootstrap method to assess the properties of estimators based on a random sample. However, a major advantage of the bootstrap is that it can be used in an enormous range of statistical problems, including very complicated ones.

For a complicated statistical model involving many random variables, it is important to distinguish between:

1. non-parametric bootstrap in which we sample with replacement from all the data and estimate unknown parameters for each bootstrap sample,
2. fully parametric boostrap in which we fit the model to the observed data, then simulate random sample from the fitted model and finally refit the model to the simulated samples,
3. semi-parametric bootstrap in which we fit the expectation part of the model, then resample with replacement from the residuals, adding the resampled residuals to the fitted expectation to get new samples, to which we refit the model.

The distinction between the three cases is what we can assume. In the fully parametric bootstrap we must be prepared to assume that the statistical model is valid for some values of the parameters: the only problem is to find these values. In the semi-parametric bootstrap we must assume that the expectation part of the model is valid, but need not assume anything about the distribution of errors. In the non-parametric bootstrap we need not assume that any part of the model is valid.

If the whole model is valid then the parametric bootstrap will give better estimates.

There is no invariant formula for deciding which method to apply. It is a question of the statistician's judgement based on understanding the model and the underlying phenomenon.

4.8 Cross-validation

When a statistical model has been fitted to data, a good way to test the goodness-of-fit is to assess how well the model predicts any future data. However, often there is no additional data available. If future data become available, it is desirable to refit the model to all data.

Cross-validation provides methods of making use of all the data both to fit the model and to assess the goodness-of-fit. These are resampling methods, computer intensive, requiring refitting the model many times.

Method 1: leave-one-out samples
For each i in turn, omit the $i-th$ datum, fit the model to the $n-1$ remaining data and use the fitted model to predict the outcome at the $i-th$ point.

The cross-validation residual is

$r_i = predicted\ value\ at\ the\ i{-}th\ point\ -\ actual\ outcome\ at\ the\ the\ i{-}th\ point.$

If the model fits well, then

$$\sum_{i=1}^{n} r_i^2$$

will be small. To choose between two models, choose the one with the smaller value of $\sum_{i=1}^{n} r_i^2$.

Method 2: construction and test samples
Randomly divide the observed data into two sets: the construction sample and the test sample. Fit the model with the construction sample, then test the goodness-of-fit with the test sample. Use the explanatory variables of the test sample to predict the outcomes from the fitted model, then compare them with the observed outcomes. Repeat the procedure, with new random division into construction and test samples, so that every observed data point will occur several times in both construction and test samples.

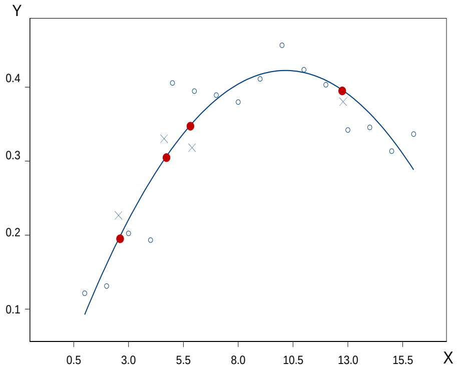

Figure 4.2: Method 2 of cross-validation: open blue circles - construction sample, cross - test sample, solid red points - prediction of the test sample