Contents

1 Types of survey samples ........................................... 2
  1.1 Random samples ................................................. 2
  1.2 Using weights ................................................... 3
  1.3 Stratified samples .............................................. 3
  1.4 Grouped data .................................................... 4

2 Natural estimators and resampling methods ............... 4
  2.1 The use of order statistics .................................... 5
  2.2 Jacknife and bootstraping ..................................... 6

3 Non parametric estimation of densities ..................... 8
  3.1 Histograms ........................................................ 8
  3.2 Kernel estimation .............................................. 9

4 Sampling properties of kernel estimates ..................... 10
  4.1 Assumptions and notations .................................... 12
  4.2 Bias and variance of a kernel estimate ....................... 12
  4.3 Approximating the bias and the variance ..................... 12
  4.4 What are the ideal kernel and window size? .................. 13

5 Choosing the window size ......................................... 14
  5.1 Subjective choices .............................................. 15
  5.2 Reference to a known distribution ............................ 15
  5.3 Likelihood with cross validation .............................. 16
  5.4 Least squares cross validation ................................ 16
  5.5 Density estimation with weighted samples ................... 17
  5.6 Using R .......................................................... 17
1 Types of survey samples

Please read the first chapter of Deaton (1997) which contains much more material than this section.

The data we are interested in are survey data concerning households. Many type of information can be asked to household such as unemployment, wages, education, health status. Here we are mainly concerned with income and sometime consumption. We have a finite population of size $N$, like the French of the UK population. We want to draw a sample of a smaller size $n$ from that population. How can we proceed? The design of a survey has to follow precise rules. We want to get information on a population and it is too costly to ask the entire population every year. A census occurs at most every five years and gives information on the whole population. The coverage of the population is usually not complete: homeless people, armed forces,...

1.1 Random samples

A survey has to be framed, which means that we have to know the size and composition of the true population. A census is useful to frame a survey, other administrative data can be used too. The census for instance provide a list of households to sample. Or social security numbers.

Then we have to decide about the size $n$ of the survey. The sample survey is then drawn at random. The sample mean

$$\bar{x} = \frac{1}{n} \sum_{i} x_i$$
is a good estimator for the population mean. As we can obtain different samples for the same population, this estimator has a variance estimated by

\[
\text{Var}(\bar{x}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

Remember the classical result about the sample mean \( \bar{x} \sim N(\mu, \sigma^2/n) \).

### 1.2 Using weights

Let us now suppose that we want to get more information on a particular group. That group will be more sampled than the other groups on purpose. It will be over represented: for instance to study the economic impact of AIDS, it is useful to sample in regions where AIDS is more present. If we compute the mean using the simple above formula, the mean will be biased. In this case the sample has to be reweighted to make it representative of the population.

Suppose that we have a population of \( N \) households and a sample of \( n \) observations. Each household has a probability \( \pi_i \) of being drawn in a sampling scheme with replacement (simplification assumption). For each household, we define a weight

\[
w_i = \frac{1}{n \pi_i}.
\]

In the usual random case, \( \pi_i = 1/N \), so that all the weights are the same and equal to \( N/n \) and the sum of the weights is equal to \( N \). We can now compute the weighted mean:

\[
\hat{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}.
\]

This is an unbiased estimator of the population mean. The variance of this estimator is

\[
\text{Var}(\hat{x}_w) = \frac{n}{n-1} \left( \sum_{i=1}^{n} \nu_i^2 (x_i - \bar{x}_w)^2 \right)
\]

where \( \nu_i = w_i / \sum w_i \) are the normalised weights. This variance is minimum when the sampling probabilities are chosen proportional to \( x_i \).

### 1.3 Stratified samples

The effect of stratification is to break up a single survey into multiple independent surveys. This is interesting to do when sub-populations vary considerably. Members of the population are grouped into relatively homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Then random sampling is applied within each stratum.
Suppose that we have $S$ strata, that the population size is $N$ while the population in each strata is $N_s$. The mean of the population is now estimated by

$$\bar{x} = \frac{1}{S} \sum_{s=1}^{S} N_s \bar{x}_s,$$

where $\bar{x}_s$ is the estimated mean for each strata. In each strata, we can of course have a particular weighting scheme which is superimposed to the stratification. Stratification often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population. In fact

$$\text{Var}(\bar{x}) = \sum_{s=1}^{S} \left( \frac{N_s}{N}\right)^2 \text{Var}(\bar{x}_s),$$

because the strata are independent. It can be shown that this variance is lower than the variance of

$$\bar{x}_{srs} = \frac{1}{S} \sum_{s=1}^{S} \frac{n_s}{n} \bar{x}_s,$$

where the weights are formed not using the population size, but the sample size and is finally just the sample mean of the unstratified sample.

### 1.4 Grouped data

Survey data report private information on households. These data are politically sensitive depending on their content. For instance, there are in France questionings about the use of racial information to study discrimination. In Belgium, it is forbidden to ask question on the language used at home (French or Flemish). So for a long time, these data were simply not available. Researcher had access to data that were so aggregated, that they were presented in groups. The treatment of these grouped data needed special tools and estimation techniques. For instance, Singh and Maddala or McDonald use grouped data for the US income. The remaining columns represent the class frequency. We reproduce here these data in Table 1 as given in McDonald (1984). We have percentages summing 100% in all the columns with dates. The first column represent the end of class for each group. It is presumably in thousands dollars per year per household. This lead to an histogram that has to be drawn by hand.

### 2 Natural estimators and resampling methods

In this section, we give indications on how to estimate usual quantities such as cumulative distributions, Lorenz curves, Gini indices using order statistics. The method can be extended so as to consider FGT poverty indices, poverty deficit curves and dominance curves. Most of the time, standard errors or small sample distributions are difficult to obtain so that resampling techniques such as the bootstrap are very useful.
Table 1: US Data on income

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>6.6</td>
<td>3.5</td>
<td>2.1</td>
</tr>
<tr>
<td>5.0</td>
<td>12.5</td>
<td>8.5</td>
<td>4.1</td>
</tr>
<tr>
<td>7.5</td>
<td>15.2</td>
<td>10.6</td>
<td>6.2</td>
</tr>
<tr>
<td>10.0</td>
<td>16.6</td>
<td>10.6</td>
<td>6.5</td>
</tr>
<tr>
<td>12.5</td>
<td>15.8</td>
<td>11.4</td>
<td>7.3</td>
</tr>
<tr>
<td>15.0</td>
<td>11.0</td>
<td>10.9</td>
<td>6.9</td>
</tr>
<tr>
<td>20.0</td>
<td>13.1</td>
<td>18.8</td>
<td>14.0</td>
</tr>
<tr>
<td>25.0</td>
<td>4.6</td>
<td>11.6</td>
<td>13.7</td>
</tr>
<tr>
<td>35.0</td>
<td>3.0</td>
<td>9.5</td>
<td>19.8</td>
</tr>
<tr>
<td>50.0</td>
<td>1.1</td>
<td>3.2</td>
<td>12.8</td>
</tr>
<tr>
<td>∞</td>
<td>0.5</td>
<td>1.4</td>
<td>6.7</td>
</tr>
</tbody>
</table>


2.1 The use of order statistics

The first estimation techniques that we shall present now are relatively simple. They use order statistics which come from the ordering of the observations. Suppose that the observations from $X$ are ordered by increasing value and let us note this ordering as

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}.$$  \hfill (1)

$x^{(1)}$ represents the smallest observation and $x^{(n)}$ the largest one. In this case, it becomes easy to estimate a cumulative distribution and its quantiles. As a matter of fact, a distribution is defined as $F(x) = \text{Prob}(X < x)$. It can be approximated by

$$\text{Prob}(X \leq x^{(i)}) \approx i/n$$ \hfill (2)

when we have enough observations.

The first decile of this distribution corresponds to the value $x_{0.10}$ such that $\text{Prob}(X \leq x_{0.10}) = 0.10$. It will be enough to find the observation which rank $i$ corresponds roughly to $i/n = 0.10$ in the ordered sequence $X$. In the general case, let us note $Q(p)$ the quantile of order $p$; it can be estimated as

$$Q(p) = x^{(s)} \quad s - 1 \leq np \leq s.$$ \hfill (3)

This means that the quantile of order $p$ is the observation having rank $s^{th}$ so that the above inequality is verified. This solution is valid in large samples. In smaller samples, an interpolation can be needed.

The estimated quantiles can lead to the computation of the dispersion measure known as the interquartile range $(x_{0.75} - x_{0.25})/x_{0.50}$. 

5
Using the same order statistics, we can define an estimator for the **generalized Lorenz curve**. The generalized Lorenz curve is defined by the partial sum of the ordered quantiles. Thus

$$L(p = i/n) = \frac{1}{n} \sum_{j=1}^{i} x(j). \quad (4)$$

We have used here partial sums of order statistics. The usual Lorenz curve obtains by normalizing this curve by the sample mean.

Finally, the **Gini coefficient** can be estimated as seen in the previous chapter using a simple weighted sum of order statistics. Which is simpler than just evaluation the double sum of the original definition based on the mean of the absolute difference between each possible pair of observations:

$$\hat{I}_G = \frac{2}{n(n-1)\bar{z}} \sum_{i = 1}^{n-1} i x(i) - \frac{n + 1}{n - 1}. \quad (5)$$

This type of computation can also be used to for Sen-Schorocks-Thon poverty index:

$$\hat{I}_{SST} = \frac{1}{n^2} \sum_{i=1}^{q} (2n - 2i + 1) \frac{z - x(i)}{z}.$$  

where $q$ corresponds to the rank of the poverty line $z$ in the distribution of $X$.

### 2.2 Jacknife and bootstraping

Thus we have simple estimators, but we do not know all the time how to compute standard deviations. For instance it was rather easy to compute the variance of the mean. But the variance of the mode is much more difficult to establish, especially when the sampling design is more complex. The bootstrap and eventually the jackknife are two methods for assessing sampling variability of an estimator.

Two sources of randomness

1. We have samples from a finite population. We must know the sample design, which can be quite complicated in order to appreciate the source of randomness. Not always easy. For instance $N$ might not be known precisely.

2. There are errors of observations, or simply the nature of the variable which is observed is random as it results from decision making under uncertainty.

Two types of methods were designed in the literature, which are resampling techniques

1. The jacknife provides $n - 1$ samples by eliminating one observation each time of the original sample.

2. The bootstrap resamples with replacement $n$ data from the original sample.
With each technique, the statistics for which we want to compute a variance is evaluated for each bootstrap or jackknife sample. The resampling technique can be quite complicated, because it has to mimic the data generating process.

The bootstrap is available in R with the package `boot`. We must first call the library `boot`. Then define a function with two arguments: the first argument represents the original data, the second argument indicates the weights of the bootstrapping generated by the package. Here we have given an example with the Gini coefficient, asking for 1000 replications.

```r
library(boot, Gini)

r = boot(y79, function(d, i) {a = Gini(d[i])}, R = 1000)

hist(r$t, probability = T, col = 'light blue',
     main = "Distribution of the Gini")

lines(density(r$t), col = "red")

boot.ci(r, type = "norm")
```

![Distribution of the Gini](image)

Figure 1: Bootstrapping the Gini
The `boot.ci` function generates 5 different types of equi-tailed two-sided nonparametric confidence intervals. These are the first order normal approximation, the basic bootstrap interval, the studentized bootstrap interval, the bootstrap percentile interval, and the adjusted bootstrap percentile interval. The type of interval is selected in the calling list. In the example, type = "norm" is selected.

The bootstrap gives us a standard deviation and a 95% confidence interval in Table 2. In Table 2: Bootstrap results for the Gini coefficient using the 1979 FES

<table>
<thead>
<tr>
<th>Gini</th>
<th>Bias</th>
<th>std. error</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.256</td>
<td>-7.55e-05</td>
<td>0.00233</td>
<td>[0.252, 0.261]</td>
</tr>
</tbody>
</table>

Figure we give a graphical representation of the small sample dispersion of the Gini coefficient. We do not claim that this is the right way to bootstrap the Gini coefficient. This is just an illustration.

3 Non parametric estimation of densities

Densities are much complex to estimate than distributions, just because the above natural estimate of a distribution is not differentiable. Some smoothing has to be used, so this section is devoted to nonparametric estimation using kernels. Most of the material presented in this section and the next ones comes from the book by Pagan and Ullah (1999) which is a valuable reference.

3.1 Histograms

If \( X \) is a continuous random variable, we define a neighbourhood of \( x \) by \( x \pm h/2 \) and we count the number of observations \( x_i \) that belong to this neighbourhood. Let us define the transformation \( \psi_i = (x - x_i)/h \), then

\[
\hat{f}_1(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \mathbf{1}(-1/2 \leq \psi_i \leq 1/2).
\]  

We notice that \( x \) is the center of the class and that \( h \) implicitly defines the number of classes. The indicator function integrates up to 1 as well as \( \hat{f}_1(x) \). Intuitively, we understand that the number of classes can grow with the number of observations, so that \( h \to 0 \) when \( n \to \infty \).

This is a rather crude way of estimating a density. But this is the only way when using group data as the one given above for the US income.

In R, this can be programmed directly using the function `hist`:

\[
\text{hist(y79,breaks=50)}
\]

where \( y79 \) is the FES data for 1979. The corresponding graph is given in Figure 2.
3.2 Kernel estimation

The histogram has the bad property of being a step function: it is discontinuous and not differentiable. We would like to get a smooth representation, and we feel that this is possible when we have a full sample and not grouped data. Rosenblatt (1956) had the idea of replacing the indicator function by a kernel $K$ which integrates to one like the indicator function. We thus have the new estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K(\psi).$$

We can deduce some of the properties of a Kernel estimator from those of the indicator function associated with the histogram.

- $\int K(\psi) d\psi = 1$,
- $h \to 0$ when $n \to \infty$,
- $K(\pm \infty) = 0$,
- A common choice for $K$ is the standardised normal density. Then $K(|\psi| \geq 3) \simeq 0$.
- The value chosen for $h$ is capital for defining the neighbourhood $|x - x_i|/h \leq 3$.

It is very important to understand the role played by $h$ in determining the shape of the obtained density. We have simulated 500 observations drawn from a mixture of normals $N(\mu_i, 1)$

![Histogram of y79](image)

Figure 2: Histogram with 50 cell of FES 1979
with $\mu_1 = 1$, $\mu_2 = 5$ and $p = 0.75$.

$$f(x) = 0.75f(x|1, 1) + 0.25f(x|5, 1).$$

We then have estimated the density of these random draws using the kernel approach and three values form the window size $h$. We give the resulting graphs in Figure 8. For the while, we accept the fact that the optimal value of $h$ is given by

$$h = c\hat{\sigma} \times n^{-1/5}.$$

We have selected three values for $c$ in the following graphs. The bimodal nature of the density is well captured in the central graph; it disappears in the first graph where we have oversmoothing while sampling errors are well visible in the last graph where we have undersmoothing.

### 4 Sampling properties of kernel estimates

We have investigated many factors that influenced the final aspect of a non-parametric density estimate. The two basic ingredients are the choice of the kernel and the choice of the smoothing window size. How could we measure exactly their influence on the precision of the final result? The basic question is to find a way to measure the distance between the true density and the estimated density. A natural distance measure between an estimator and a true value is the Mean Squared Error:

$$\text{MSE}_x(\hat{\theta}) = E[\hat{\theta} - \theta]^2$$

that can be easily decomposed into

$$\text{MSE}_x(\hat{\theta}) = \text{Biais}[\hat{\theta}]^2 + \text{Var}[\hat{\theta}].$$

But this indicator concern a ponctual estimator and not a complete density. We are thus looking for a global measure valid for the whole range of $x$. We are thus going to integrate over $x$ in order to get the MISE, or Mean Integrated Squared Error:

$$\text{MISE}_x(\hat{f}) = E\int[\hat{f}(x) - f(x)]^2dx.$$  \hspace{1cm} (10)

This corresponds to a notion of risk. If we want to minimise the loss, we simply have to consider the ISE (Integrated Squared Error):

$$\text{ISE}_x(\hat{f}) = \int[\hat{f}(x) - f(x)]^2dx.$$  \hspace{1cm} (11)

The MISE is the most commonly used indicator, but it might be difficult to compute. So that most of the time we rely on approximations that are found by noting that the MISE can be decomposed into:

$$\text{MISE}_x(\hat{f}) = \int[E(\hat{f}(x)) - f(x)]^2dx + \int\text{Var}[\hat{f}(x)]dx.$$  \hspace{1cm} (12)

It is then sufficient to find approximations for the bias and the variance and report those values in this expression.
Figure 3: Oversmoothing and undersmoothing in density estimation
4.1 Assumptions and notations

We already made some assumptions concerning the Kernel and the window size. We recall them and introduce some useful notations:

- \( \int K(t) \, dt = 1 \)
- \( \int K^2(t) \, dt = c_K < \infty \)
- \( \int tK(t) \, dt = 0 \)
- \( \int t^2K(t) \, dt = \mu_2 \)

The quantity \( \mu_2 \) is going to play an important role in the sequel. Finally, concerning the window size, we have the following assumptions:

- \( h \to 0 \) when \( n \to \infty \)
- \( nh \to \infty \) when \( n \to \infty \)

The window size has to go to zero as the sample size grows, but at a speed which is not too high.

4.2 Bias and variance of a kernel estimate

The bias and the variance of an estimator can be computed as expectations with respect to the true and unknown distribution \( f(\cdot) \). Let us start from the usual kernel density estimator

\[
E(\hat{f}(x)) = \int \frac{1}{h} K \left( \frac{x-y}{h} \right) f(y) \, dy
\]  

(13)

in order to compute the bias. For the variance we have:

\[
\begin{align*}
\text{Var}(\hat{f}(x)) &= \int \frac{1}{h^2} K \left( \frac{x-y}{h} \right)^2 f(y) \, dy \\
&- \left\{ \int \frac{1}{h} K \left( \frac{x-y}{h} \right) f(y) \, dy \right\}^2.
\end{align*}
\]  

(14)

4.3 Approximating the bias and the variance

The exact formulae that we have just given include integrals that cannot readily be evaluated and thus are of a direct practical interest. We have to find approximations, using a first order Taylor expansion, reduced to the first order.

Let us first propose the change of variable \( y = x - ht \) with Jacobian \( h \). With this change of variable, the bias becomes:

\[
\text{bias} = \int K(t)[f(x-ht) - f(x)] \, dt.
\]  

(15)
Let us develop \( f(x - ht) \) around \( h = 0 \):

\[
f(x - ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + \ldots.
\]  

Using the fact that a kernel is of zero expectation and of variance \( \mu^2 \),

\[
\text{bias} \simeq \frac{1}{2}h^2f''(x)\mu_2 + \ldots.
\]

Similar computations for the variance show that

\[
\text{Var}(\hat{f}(x)) \simeq \frac{1}{nh}f(x)c_K,
\]

supposing that \( n \) is big and \( h \) small. The approximation for the MISE is thus:

\[
\text{AMISE} \simeq \frac{1}{4}h^4\mu_2^2 \int f''(x)^2 dx + \frac{1}{nh}c_K
\]

The bias depends on the window size and not on the sample size. On the contrary, the variance is a function of the sample size. Moreover, we can minimize the bias by decreasing the window size \( h \), but at the same time we increase the variance. Choosing a value for \( h \) implies a trade-off between systematic error and random errors, between bias and variance. If we want to minimize the MISE (or the AMISE here), we see that the first term is of the same order as \( h^4 \), when the second term is of the same order as \( 1/(nh) \). Bias and variance are of the same order for

\[
h \propto n^{-1/5}.
\]

This rate of convergence for the window size is quite general for the whole non-parametric inference.

### 4.4 What are the ideal kernel and window size?

We are going to differentiate the approximate MISE with respect to \( h \) in order to find the ideal \( h \) by setting this expression to zero. We have:

\[
h_{\text{opt}} = \mu_2^{-2/5}c_K^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5}
\]

\[
= \left[ \frac{c_K}{n \mu_2^2 \int f''(x)^2 dx} \right]^{1/5}
\]

The ideal window size is a function of quite different things:

- It tends to zero at a very low speed

- It depends on the fluctuations of \( f \). If \( f \) fluctuates a lot beaucoup, a small \( h \) will be needed. Some methods will determine \( h \) with respect to a known density like the Normal (Silverman’s rule of thumb).
- Finally, \( h \) depends on the kernel. The latter can always be normalised so that \( \mu_2 = 1 \). So that the kernel takes part to the final result only with \( c_K = \int K^2(t) dt \). Silverman’s rule will again take advantage of this result.

Let us plug the optimal \( h \) into the expression of the MISE. We get:

\[
\text{MISE} \approx \frac{5}{4} \mu_2^{2/5} c_K^{1/5} \left( \int f'(x)^2 dx \right)^{1/5} \cdot n^{-4/5}
\]  

(22)

The ideal kernel is the one that minimises the MISE for a given \( f \). In order to find it, we have to minimise \( c_K \) under the provision that this kernel is a density, that is to say integrates to one and is normalised, which means that \( \mu_2 = 1 \). One can show that this ideal kernel is the Epanechnikov kernel that has a very simple expression:

\[
K(t) = \begin{cases} 
\frac{3}{4\sqrt{5}}(1 - \frac{t^2}{5}) & \text{if } |t| \leq \sqrt{5} \\
0 & \text{otherwise}
\end{cases}
\]  

(23)

We can compare the efficiency of the other kernels with respect to the Epanechnikov kernel by defining the ratio:

\[
Ef = \frac{\sqrt{\int t^2 K(t) dt \int K(t)^2 dt}}{\sqrt{\int t^2 K_e(t) dt \int K_e(t)^2 dt}}.
\]  

(24)

And using the properties of the Epanechnikov kernel, this ratio is simplified into:

\[
Ef = \frac{2/(5\sqrt{5})}{\sqrt{\int t^2 K(t) dt \int K(t)^2 dt}}.
\]  

(25)

Let us now compute the efficiency of the usual kernels. The most inefficient kernel is the rectangular kernel which leads to the histogram. With this kernel, we have an efficiency which is very near from one. It is thus not very useful to spend much time finding an efficient kernel. To justify the search for an efficient kernel, we have to take into account other criteria than efficiency. For instance, the Epanechnikov kernel is not differentiable at an order greater than one, when the biweight kernel is differentiable at the order two and when the Gaussian kernel is infinitely differentiable. Some kernels have a finite support, while others have an infinite support. This makes a difference in term of numerical efficiency. With the Gaussian, a lot of time can be spend computing very small weights.

5 Choosing the window size

The choice of the window size is crucial for the final aspect of the graph of the density. This choice can be driven by the final aim of the study. If we want to present the empirical content
of a data set, a subjective choice is convenient. If we want to derive statistical conclusions, some under-smoothing could be necessary, as the reader is able to smooth visually when he cannot rebuild details that would have been smoothed out by using a too large $h$. When many results have to be presented, an automatic method can be useful. If we want to compare results, a standardised method will be preferable. We must note that automatic methods cannot be qualified of being objective as they all rely on particular assumptions.

### 5.1 Subjective choices

We consider several graphs of the density, each one corresponding to a given choice for the window size. We chose the window size which produces the more aesthetics graph. Just have a look at previous Figures where under or over smoothing are easily detected.

### 5.2 Reference to a known distribution

We have seen that the optimal $h$ was given by:

$$h_{opt} = \mu_2^{-2/5} c_K^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5}$$

Some of the elements of this expression are known as $n$ and $K(,)$. But $f$ is of course unknown, as we want to estimate it. We have to compute $\int f''(x)^2 dx$. If we suppose that the true distribution $f$ is Normal of zero mean and of variance $\sigma^2$, then

$$\int f''_{N(0,\sigma^2)}(x)^2 dx = \sigma^{-5} \frac{0.375 \sqrt{\pi}}{\sqrt{\pi}} \simeq 0.212 \sigma^{-5}$$

Let us now choose a normal, we can verify that $\mu_2 = 1$ and $c_K = 0.5/\sqrt{\pi}$. Gathering all these small bits, we have an expression for the optimal $h$:

$$h \simeq 1.06 \sigma n^{-1/5}.$$
The only remaining question is to find a consistent estimate for the variance of the sample to get an estimate for the optimal $h$. This is the rule of Silverman which is the most popular way of finding easily a window size.

This procedure is very efficient as soon as we are not far from the Normal case, but lacks efficiency when we are far from it. In particular, if the true distribution $f$ is a mixture, the rule of Silverman will tend to over smooth the density as soon as the modes of the mixture get apart. Different articles have also shown that we have over smoothing when $f$ is asymmetric, but no over smoothing in the case of kurtosis. In particular if $f$ is Student, the rule of Silverman is rather efficient.

5.3 Likelihood with cross validation

We are going to develop the idea of a likelihood function in order to apply it for choosing a value for $h$. If a likelihood function is given by $\sum \log f(x_i)$, a pseudo likelihood function is defined as

$$\log L = \sum \log \hat{f}(x_i, h).$$

(29)

Here is however a serious problem as the optimum value of this function is obtained for $h = 0$. In order to circumvent this problem, we are going to apply the cross-validation principle. Instead of evaluating $\hat{f}(x_i, h)$, we are going to compute $\hat{f}_{-i}(x_i, h)$ defined as

$$\hat{f}_{-i}(x_i, h) = \frac{1}{h(n - 1)} \sum_{\substack{j = 1 \\text{i} \neq j}}^{n} K\left(\frac{x_j - x_i}{h}\right).$$

(30)

This means dropping one observation each time. This is a general principle in non-parametric statistics that will be used in the sequel.

This likelihood method is equivalent to choosing the $h$ that minimizes Kulback-Leibler distance between $f$ et $\hat{f}$, or in other terms

$$\int f(x) \log \left(\frac{f(x)}{\hat{f}(x)}\right) dx.$$ 

(31)

However, the $h$ which is selected by this method is extremely sensitive to the behaviour of the tails of the true density $f$. So it is not very much used. It might be useful to know it, just because it is a good way to introduce the non-parametric regression.

5.4 Least squares cross validation

Instead of considering a pseudo likelihood function as a criterion to optimize, we shall consider this time the Integrated Squared Error:

$$ISE(h) = \int (\hat{f}(x, h) - f(x))^2 dx.$$ 

(32)
Let us develop the square. This resulting expression can be simplified as one of its terms does not depend on $h$:

$$ISE(h) \propto \int \hat{f}(x, h)^2 \, dx - 2 \int \hat{f}(x, h) f(x) \, dx$$  \hspace{1cm} (33)$$

We have to find the value of $h$ that minimises as estimation of the $ISE(h)$. Here again, the cross-validation method is the right solution for evaluating this criterion. We have

$$\hat{f}_{-i}(x, h) = \frac{1}{h(n-1)} \sum_{j \neq i} K \left( \frac{x - x_j}{h} \right)$$  \hspace{1cm} (34)$$

The notation $-i$ means that we drop observation $i$ for evaluating $f(x_i)$. We can now notice that $\int \hat{f}(x, h) f(x) \, dx$ is the expectation of $\hat{f}(x, h)$. An unbiased estimator of this expectation is given by the empirical mean of $\hat{f}_{-i}(x, h)$, or in other terms

$$E(\hat{f}(x, h)) \approx \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i, h).$$  \hspace{1cm} (35)$$

We have now to compute the first element of the $ISE$ by means of

$$\int \hat{f}^2 \, dx = \frac{1}{n^2 h^2} \sum_i \sum_j \int x K \left( \frac{x_i - x}{h} \right) K \left( \frac{x_j - x}{h} \right) \, dx,$$  \hspace{1cm} (36)$$

with a solution given by

$$\int \hat{f}^2 \, dx = \frac{1}{n^2 h^2} \sum_i \sum_j \tilde{K} \left( \frac{x_i - x_j}{h} \right)$$  \hspace{1cm} (37)$$

$\tilde{K} = K \circ K$. If the kernel $(0,1)$, then $\tilde{K} = N(0, 2)$.

The method is rather intensive in term of computer time. For every value of $h$, we have to evaluate $ISE(h)$ which contains a double sum. Moreover, the function can have several local minima. Pagan and Ullah mention the “binning” technique which is used for instance in the software Xplore for reducing computer time.

### 5.5 Density estimation with weighted samples

When there are weights $w_i$, we must first impose that the weights sum to unity. The usual formula is simply modified into

$$f(x) = \frac{1}{n} \sum w_i K \left( \frac{x - x_i}{h} \right)$$

### 5.6 Using R

The standard stats package includes a routine for estimating densities. The density object is created by simply calling density(x) where x represents the data set, assuming that the data are presented in a column. By default a Gaussian kernel is used and the classical rule of
Table 4: R options for density estimation

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Kernel</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>(bw = \text{nrd0}(x))</td>
<td>(\text{kernel = ”gaussian”})</td>
<td>(\text{weights = rep(1/nx, nx)})</td>
</tr>
<tr>
<td>(bw = \text{bw.ucv}(x))</td>
<td>(\text{kernel = ”epanechnikov”})</td>
<td></td>
</tr>
<tr>
<td>(bw = \text{bw.SJ}(x))</td>
<td>(\text{kernel = ”triangular”})</td>
<td></td>
</tr>
</tbody>
</table>

Silverman for the bandwidth. Of course many options are possible which can be found on the help. We present these options in Table 4. To obtain a graph, it suffices to use the routine plot together with the output object of density. For instance, \(\text{plot(density}(x))\). If we want to change the default method for determining the bandwidth, using for instance the cross validation method, we can use

\[
\text{plot(density}(y79, bw=\text{bw.ucv}(y79)))
\]

We are not obliged to use the same sample for estimating the density and for computing the bandwidth. In particular, we can use a subsample for computing the bandwidth. We can draw a subsample at random for instance.

In the column Bandwidth of Table 4, \(bw = \text{nrd0}(x)\) is a slight modification of the rule of Silverman as it uses an improved estimator of the sample variance. \(bw = \text{bw.ucv}(x)\) was already explained as being the unbiased cross validation. \(bw = \text{bw.SJ}(x)\) is the implementation of the Sheather and Jones (1991) plug-in rule. It estimates non-parametrically the integral of the squared second order derivative of the true density. This method is very popular, as it is a robust plug-in rule which in general gives better results than the simple Silverman rule. But is requires the fourth order derivative of the Kernel. So it cannot be used with the Epanechnikov kernel. But it is safe with a Gaussian kernel.

6 General estimation methods

Inference, including Bayesian inference for Pareto and other simple densities. R and inference for SM using the FES data.

6.1 Inference for grouped data

Grouped data used to be very common because they solve the question of anonymity when individual data are involved. Considering group data can also be a way to solve difficult estimation problems. For instance, it is quite impossible to use the maximum likelihood principle to make inference with the Generalized Gamma density due to its awkward parameterization (see Johnson, Kotz, and Balakrishnan (1995)).

When data are grouped into clusters, inference is based on the comparison of two quantities
- $p_i(\theta)$ is the theoretical probability to belong to cluster $i$th among the $g$ possible clusters of the population:

$$p_i(\theta) = \int_{I_i} f(x; \theta) \, dx.$$  

This probability is given by integrating the density to be estimated over the range of cluster $i$. The cluster corresponds to the interval $[x_{i-1}, x_i]$, the integral is computed over this range.

- $n_i/n$ are the observed frequencies, they are given by the data. For instance, the cluster frequencies in an histogram. $n$ is the total sample size, while $n_i$ is the number of observations in cluster $i$.

McDonald and Ranson (1979) give different ways two confront these two quantities.

In a likelihood framework, we have to represent the multinomial process generating the histogram. The likelihood function is thus:

$$L(\theta) = n! \prod_{i=1}^{g} \frac{p_i(\theta)^{n_i}}{n_i!}.$$  

They call this approach a scoring method because we have to compute the first derivative of the likelihood function in order to find its maximum.

The Pearson minimum chi-squared estimator minimizes a chi-squared distance between the theoretical probability and its empirical counterpart

$$n \sum_{i=1}^{g} \left( \frac{n_i}{n} - \frac{p_i(\theta)}{p_i(\theta)} \right)^2.$$  

This quantity is distributed as a $\chi^2$ with $g - k - 1$ degrees of freedom which give a direct way for testing the adequation between the data and the model. This a goodness-of-fit test. This method of estimation is asymptotically equivalent to the maximum likelihood.

The least squares estimator minimizes a simpler distance between theoretical and empirical probabilities with

$$\sum_{i=1}^{g} \left( \frac{n_i}{n} - p_i(\theta) \right)^2.$$  

This last method gives often different results than the previous ones and is not recommended. The Pearson method corresponds to a weighted least-squares.

On US grouped data for 1970, 1972, 1974, 1975, McDonald and Ranson (1979) found that in general the Singh-Maddala distribution gave the better fit, much better than the logNormal. Scoring and Pearson methods gave very similar results either for the parameters or the implied Gini coefficient. Least squares gave sometimes rather different results.
6.2 A regression based on the empirical distribution

When the data are not grouped, it is possible to use other methods to fit a density. The method we examine here is used for instance in Singh and Maddala (1976). It is still based on the comparison between a statistic and its theoretical counterpart. But here, Singh and Maddala (1976) take advantage of the fact that the distribution has an analytical form. They confront it to the natural nonparametric estimator of the distribution. For the SM distribution, we have

\[ F(x) = 1 - \frac{1}{(1 + a_1 x^{a_2})^{a_3}}. \]

The estimation procedure consists in minimising the least squares distance between \( F(x, a) \) and \( \hat{F}(x) \) computed either for each sample value or for a given grid. Only \( \hat{F} \) has to make use of the whole sample. The minimisation problem is:

\[ \hat{a} = \arg\min \sum [\log(1 - \hat{F}) + a_3 \log(1 + a_1 x_i^{a_2})]^2. \]

This is a nonlinear regression problem which has to be solved by numerical optimisation in a quite simple way.

We can make two comments concerning this method:

- it uses a least squares distance and not a \( \chi^2 \) distance. We can have a first source of errors by not using weighted least squares as underlined in the previous subsection.

- We have a problem at the right infinite boundary as we cannot compute \( \log(1 - F) \) because \( F(x_{\text{max}}) = 1 \). This problem does not exist when probabilities are confronted to their empirical counterparts.

The same regression method can be used for making inference on the Pareto parameter because we have then a linear regression. For the Pareto density, this was in fact the original method. We have

\[ (1 - F(x_i)) = (x_i/x_m)^{-\alpha}. \]

Taking the logs each side and using a natural estimate for \( F \) leads to the regression

\[ \log(1 - \hat{F}(x_i)) = \text{cste} - \alpha \log(x_i) + \epsilon_i. \]

If we do not get a straight line when plotting the two logs, it is a test that the sample does not come from a Pareto distribution. We can also estimate \( \alpha \) in a similar way using the empirical Lorenz curve. These estimators are consistent.

Finally, let us consider the Weibull case. The cumulative distribution is

\[ F(x) = 1 - \exp(-(kx)^\alpha). \]

Taking log twice and paying attention to the signs, we have the following regression

\[ \log(- \log(1 - \hat{F}(x_i))) = \alpha \log k + \alpha \log x_i + \epsilon_i. \]

This regression is similar to that obtained for the Pareto case, except that we have to take twice the logs for the left hand side. A graphical device is also a good test for the adequacy of the Weibull model to the data.
7 Using the likelihood function for making inference

When individual data are available, it is possible to write the likelihood function of the model and use it for making inference. In this section, we shall apply this principle of inference for two standard processes the Pareto density and the lognormal density.

7.1 Maximum likelihood for Pareto samples

Inference is quite easy for the usual Pareto I model. It is detailed for instance in Arnold (2008). Let us suppose that we have an IID sample of \( X \) which is drawn from a Pareto I model. The likelihood function is

\[
L(x; x_m, \alpha) = \alpha^n x_m^\alpha (\prod x_i)^{-(\alpha+1)} \mathbb{I}(x_i \geq x_m).
\]

It is easy to see that we have two sufficient statistics which give immediately the MLE

\[
\hat{x}_m = x_{(1)} \\
\hat{\alpha} = \left[ \frac{1}{n} \sum \log(x_i/x_{(1)}) \right]^{-1}.
\]

As underlined by Arnold (2008), these estimators are positively biased in a small sample as

\[
E(\hat{x}_m) = x_m (1 - 1/(n\alpha))^{-1} \\
\text{Var}(\hat{x}_m) = x_m^2 n\alpha(n\alpha - 1)^{-2} (n\alpha - 2)^{-1} \\
E(\hat{\alpha}) = \alpha n / (n - 2) \\
\text{Var}(\hat{\alpha}) = \alpha^2 (n - 2)^{-2} (n - 3)^{-1}.
\]

Knowing the bias, it is easy to propose unbiased estimators by simply correcting the initial maximum likelihood estimators. Once we know the estimates of \( x_m \) and of \( \alpha \), it is easy to produce an estimate for the needed transformations of these parameters such as for instance the Gini coefficient and to find their standard deviation using the delta method (which is not very precise, however).

7.2 Bayesian inference for the Pareto

Instead of using the frequentist estimation approaches discussed above, we may consider a Bayesian formulation of the problem. See for instance the summary available in Arnold (2008). If \( x_m \) is known, the problem is quite simple. In the case where \( x_m \) is also an unknown parameter, inference becomes more delicate and a Gibbs sampler is needed. We treat here only the case where \( x_m \) is known.

Let us recall that in a classical framework, the sample space is probabilised and that one looks for the value of the parameter \( \theta \) that gives the maximum probability to get the observed
In a Bayesian framework, the parameter space is also probabilised. It is endowed with a prior \( p(\theta) \) possibly non-informative and the product of inference is a posterior density obtained by applying Bayes’ theorem:

\[
p(\theta|y) = \frac{l(y; \theta)p(\theta)}{\int l(y; \theta)p(\theta)d\theta},
\]

where the denominator is the integrating constant of the posterior density. It is usually the case to work up to a constant of proportionality as the denominator does not depend on the parameters (they are integrated out). So that the posterior is defined as:

\[
p(\theta|y) \propto l(y; \theta)p(\theta).
\]

In the natural conjugate framework, the prior \( p(\theta) \) is chosen in such a way that it combines easily with the likelihood function \( l(y; \theta) \). The natural framework relies on the exponential family where sufficient statistics of two samples combine easily.

The Pareto distribution is related to the exponential distribution as follows. Suppose \( X \) is Pareto-distributed with minimum \( x_m \) and index \( \alpha \). Let us consider the following transformation:

\[
Y = \log \left( \frac{X}{x_m} \right).
\]

Then \( Y \) is exponentially distributed with intensity parameter \( \alpha \), or equivalently with expected value \( 1/\alpha \):

\[
\Pr(Y > y) = e^{-\alpha y}.
\]

The cumulative density function is thus \( 1 - e^{-\alpha y} \) and the pdf:

\[
f(y; \lambda) = \begin{cases} \alpha e^{-\alpha y}, & y \geq 0, \\ 0, & y < 0. \end{cases}
\]

The likelihood function for \( \alpha \), given an independent and identically distributed sample \( y = (y_1, \ldots, y_n) \) drawn from that variable, is

\[
L(\alpha; y) = \prod_{i=1}^{n} \alpha \exp(-\alpha y_i) = \alpha^n \exp\left(-\alpha \sum_{i=1}^{n} y_i\right) = \alpha^n \exp\left(-\alpha n\bar{y}\right),
\]

where

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

is the sample mean of \( y \). The conjugate prior for the exponential distribution is the gamma distribution (of which the exponential distribution is a special case). The following parametrisation of the gamma pdf is useful:

\[
\text{Gamma}(\alpha; \nu, s) = \frac{s^{\nu}}{\Gamma(\nu)} \alpha^{\nu-1} \exp(-\alpha s),
\]
with moments given by
\[ \text{E}(\alpha) = \nu/s \quad \text{Var}(\alpha) = \nu/s^2. \]

The posterior distribution \( p \) can then be expressed in terms of the likelihood function defined above and a gamma prior:

\[
p(\alpha|y) \propto L(\alpha; y) \times \text{Gamma}(\alpha; \nu, s)
= \alpha^n \exp(-\alpha n\overline{y}) \times \frac{s^\nu}{\Gamma(\nu)} \alpha^{\nu-1} \exp(-\alpha s)
\propto \alpha^{(\nu+n)-1} \exp(-\alpha (s + n\overline{y})).
\]

Now the posterior density \( p \) has been specified up to a missing normalizing constant. Since it has the form of a gamma pdf, this can easily be filled in, and one obtains
\[
p(\alpha|y) = \text{Gamma}(\alpha; \nu + n, s + n\overline{y}).
\]

Here the parameter \( \nu \) can be interpreted as the number of prior observations, and \( s \) as the sum of the prior observations.

Knowing the posterior parameters, we can compute easily the posterior moments by applying simply the above analytical formulae. We can draw the graph of the posterior density of \( \alpha \). More interestingly, we can generate random numbers from the posterior density in order to find the distribution of any inequality index such as the Gini coefficient or the Atkinson index or of any of the other transformation of \( \alpha \). We have in this way \( np \) draws of transformations of \( \alpha \) for which we can compute a mean, a standard deviation and estimate a density using a nonparametric kernel estimate.

### 7.3 Maximum likelihood for Lognormal samples

The probability density function of a log-normal distribution is:
\[
f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0
\]
where \( \mu \) and \( \sigma \) are the mean and standard deviation of the variable’s natural logarithm. This means for instance that \( \mu = \text{E}(\log(x)) \). The likelihood function is rather simple to write once we note that this pdf is just the normal pdf times the Jacobian of the transformation which is \( 1/x \).

We have
\[
f_L(x; \mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{x_i}\right) f_N(\ln x_i; \mu, \sigma)
\]
where by \( f_L \) we denote the probability density function of the log-normal distribution and by \( f_N \) that of the normal distribution. Therefore, using the same indices to denote distributions, we can write the log-likelihood function in the following way:
\[
\ell_L(\mu, \sigma|x_1, x_2, \ldots, x_n) = -\sum_i \ln x_i + \ell_N(\mu, \sigma|\ln x_1, \ln x_2, \ldots, \ln x_n)
= \text{constant} + \ell_N(\mu, \sigma|\ln x_1, \ln x_2, \ldots, \ln x_n).
\]
Since the first term is constant with regard to \( \mu \) and \( \sigma \), both logarithmic likelihood functions, \( \ell_L \) and \( \ell_N \), reach their maximum with the same \( \mu \) and \( \sigma \). Hence, using the formulas for the normal distribution maximum likelihood parameter estimators and the equality above, we deduce that for the log-normal distribution it holds that

\[
\widehat{\mu} = \frac{\sum_i \ln x_i}{n}, \quad \widehat{\sigma^2} = \frac{\sum_i (\ln x_i - \widehat{\mu})^2}{n}.
\]

This means that in a lognormal sample, the two parameters can be estimated by the sample mean of the logs and the variance of the logs.

### 7.4 Bayesian inference for the Lognormal

The likelihood function is the same as in the classical case, but some rewriting is convenient for combining with the prior:

\[
L(\mu, \sigma^2 | x) = \left( \prod_{i=1}^{n} (x_i)^{-1} \right) (2\pi)^{-n/2} \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log x_i - \mu)^2 \right)
\]

\[
\propto \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i} (\log x_i - \mu)^2 \right)
\]

\[
\propto \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \left( s^2 + n(\mu - \bar{x})^2 \right) \right), \quad (38)
\]

where

\[
\bar{x} = \frac{1}{n} \sum_i \log x_i, \quad s^2 = \frac{1}{n} \sum_i (\log x_i - \bar{x})^2.
\]

As we can neglect the Jacobian \( \left( \prod_{i=1}^{n} (x_i)^{-1} \right) \), Bayesian inference in the log normal process proceed in the same way as for the usual normal process. In particular, we have natural conjugate prior densities for \( \mu \) and \( \sigma^2 \). We select a conditional normal prior on \( \mu | \sigma^2 \) and an inverted gamma prior on \( \sigma^2 \):

\[
\pi(\mu | \sigma^2) = f_N(\mu | \mu_0, \sigma^2/n_0) \propto \sigma^{-1} \exp \left( -\frac{n_0}{2\sigma^2} (\mu - \mu_0)^2 \right), \quad (39)
\]

\[
\pi(\sigma^2) = f_{\gamma}(\sigma^2 | \nu_0, s_0) \propto \sigma^{-(\nu_0+2)} \exp \left( -\frac{s_0}{2\sigma^2} \right). \quad (40)
\]

The prior moments are easily derived as:

\[
E(\mu | \sigma^2) = E(\mu) = \mu_0, \quad \text{Var}(\mu | \sigma^2) = \frac{1}{n_0} \sigma^2 \quad \text{Var}(\mu) = \frac{1}{n_0} \frac{s_0}{\nu_0 - 2} \quad (41)
\]

\[
E(\sigma^2) = \frac{s_0}{\nu_0 - 2}, \quad \text{Var}(\sigma^2) = \frac{s_0^2}{(\nu_0 - 2)(\nu_0 - 4)} \quad (42)
\]

Let us now combine the prior with the likelihood function to obtain the joint posterior probability density function of \( (\mu, \sigma^2) \) in such a way that isolates the conditional posterior densities of each parameter.

\[
\pi(\mu, \sigma^2 | x) \propto \sigma^{-(n+n_0+3)} \exp \left( -\frac{1}{2\sigma^2} \left( s_0 + s^2 + n(\mu - \bar{x})^2 + n_0(\mu - \mu_0)^2 \right) \right).
\]
As we are in the natural conjugate framework, we must identify the parameters of the product of an inverted gamma in $\sigma^2$ by a conditional normal density in $\mu | \sigma^2$. After some algebraic manipulations: the conditional normal posterior is
\[
\pi(\mu | \sigma^2, x) \propto \sigma^{-1} \exp \left( -\frac{1}{2\sigma^2} \left( (n_0 \mu_0 + n \bar{x}) / n_\ast \right) \right),
\]

\[
\propto f_N(\mu | \mu_\ast, \sigma^2 / n_\ast),
\]

with
\[
n_\ast = n_0 + n, \quad \mu_\ast = (n_0 \mu_0 + n \bar{x}) / n_\ast.
\]

Then the marginal posterior density of $\mu$ is Student with
\[
\pi(\mu | x) = f_t(\mu | \mu_\ast, s_\ast, n_\ast, \nu_\ast),
\]

\[
\propto [s_\ast + n_\ast (\mu - \mu_\ast)]^{-(\nu_\ast + 1)/2}
\]

where
\[
\nu_\ast = \nu_0 + n, \quad s_\ast = s_0 + s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{x})^2.
\]

The posterior density of $\sigma^2$ is given by
\[
\pi(\sigma^2 | x) \propto \sigma^{-(n+\nu_0+2)} \exp \left( -\frac{1}{2\sigma^2} \left( s_0 + s^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{x})^2 \right) \right),
\]

\[
\propto f_{i\gamma}(\sigma^2 | \nu_\ast, s_\ast).
\]

The posterior densities of $\mu$ and $\sigma^2$ belong to well known family. Their moments are obtained analytically and no numerical integration is necessary. We recover the classical results under a non-informative prior.

### 7.5 Using R for Pareto and lognormal fit

Using the same data set as before (UK family expenditure survey in real terms), we shall here compare the fit obtained by using a Pareto density and a lognormal density.

We first try to fit a Pareto density. There is a simple way to test the Pareto assumption. We just have to plot the graph of $\log(y)$ against $\log(1 - F)$. For this the following R routine is convenient. It assumes that the observations are ordered. The boudary problem is solved by dropping the last observation:

```r
pareto = function(y){
    n = length(y)
    F = (1:n)/n
    F = F[1:n-1]
    y = y[1:n-1]
    plot(log(y),log(1-F))
    lines(log(y),log(1-F))
}
```
Figure 4: Pareto tail for the income distribution
Figure 4 shows that the Pareto assumption might be valid only above a certain income level. The black line represents 1979, while the red line corresponds to 1988, blue to 1992 and green to 1996.

A Bayesian inference for $\alpha$ is easy to program, provided we take into account the way the Gamma distribution is parameterised in R. The shape parameter corresponds to the sample size and the scale parameter corresponds to $1/(n\bar{x})$ when using a non informative prior. This can be implemented in the following routine which include the computation of the Gini index together with its small sample properties.

\[
\text{Bayes} = \text{function}(x, \text{np}) \{
\text{# Bayesian inference for alpha when xm is known.}
\text{# Simulation of the Gini}
\text{yb} = \text{sum}(\log(x/\text{min}(x)))
\text{n} = \text{length}(x)
\text{alpha} = \text{rgamma}(\text{np, scale} = 1/\text{yb, shape} = \text{n})
\text{a} = \text{alpha[alpha>0.6]}
\text{g} = 1/(2*\text{a}-1)
\text{cat("Gini = ",mean(g)," S.D. =", sd(g),"\n")}
\text{plot(density(g))}
\}
\]

The answer given by the Bayesian inference using a Pareto model depends heavily on the truncation point. We have chosen 120, which leaves only 971 observations out of 6230 for 1979. But do not forget that Pareto is for high incomes. Bayesian inference produced an $\alpha$ with posterior mean of 4.365 and a standard deviation of 0.143. Fitting a Pareto model leads to a Gini coefficient which is slightly greater than that obtained when computing it directly using the sole sub-sample.

The bootstrap produces a density which is slightly more concentrated than its Bayesian counterpart as shown in Figure 5 where the Bayesian estimate is in black while the bootstrap is in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian with Pareto</td>
<td>0.129</td>
<td>0.00486</td>
</tr>
<tr>
<td>Bootstrap parameter free</td>
<td>0.118</td>
<td>0.00426</td>
</tr>
</tbody>
</table>

The Pareto model does not fit correctly the complete sample. Using the 1979 FES data, the MLE for $\alpha$ in the Pareto process is 1.974. If we now turn to the lognormal process, the MLE estimate for $\sigma$ is 0.459. We can now plug these two values into the expression of the Lorenz curves for the two models and compare the result to the natural estimate of the Lorenz curve. This is done in Figure 6 using the following R code.
Figure 5: Comparing Bayes and bootstrap estimates for the Gini

```r
plot(Lc(y79))
p = seq(0,1,0.05)
lines(p,Lc.pareto(p, parameter=2),col="red")
  text(0.9,0.6,"Pareto 2.0")
lines(p,Lc.lognorm(p, parameter=0.45),col="blue")
  text(0.45,0.4,"Lognormal 0.45")
```

The lognormal seems to fit the data quite well when of course the Pareto is not able to produce a good account of the whole sample. So, we could perform the same exercise as we did for the Gini coefficient with the Pareto process. The posterior density of $\sigma$ is an inverted gamma with hyperparameters $\nu_*$ and $s_*$ based on sample mean and variance of the log variable under a non-informative prior. We could then simulate $\sigma^2$ and compute the Gini as $2\Phi(\sigma/\sqrt{2}) - 1$ for each draw. This is done in the next chapter.
Figure 6: Lorenz for Pareto and Lognormal
8 Using mixtures for IID samples

We are presenting in this section an intermediate approach between a fully parametric model for the income distribution and a fully nonparametric density estimation. It is a semiparametric approach as it is based on the combination of parametric densities where the number of needed densities has to be determined by the sample.

8.1 Informal introduction

Let us go back to the FES data sets. Which kind of density can we fit to these data? We have illustrated several stylised facts

- The Pareto does not fit the data as shown by the Lorenz curve
- The lognormal seems to fit the data better as shown again by the Lorenz curve
- The high incomes, greater than 120, seem to behave like a Pareto

Does the lognormal fit really well the data as the Lorenz curve would suggest? In Figure 7, we compare the adjusted parametric lognormal density with a non-parametric estimate of the density using the following R code:

```r
plot(density(y79))
lines(dlnorm(seq(0,350,1), meanlog=mean(ly79),
                 sdlog=sd(ly79)),col="red")
```

We see clearly that if the overall fit of the lognormal could pass for being nice, the two modes are of course smoothed into something with is even not in between, while the right tail seems to be fitted quite well. So the lognormal model is not adequate to describe completely the sample.

8.2 Mixture of distributions

When a single density is not enough to represent correctly the distribution of a sample, a simple explanation is that the observed sample is heterogenous and this result from the mixing of different populations, each being represented by a particular density indexed by a given parameter. The trouble is that we do not know first how many different sub-populations there are and second what is their proportion. This lack of knowledge makes the problem difficult. For a simplification, let us suppose that we have only two subpopulations, each one being described by a density indexed by \( \theta_i \) and in unknown proportion \( p \). The density of one observation is

\[
f(x|\theta) = p \times f_N(x|\mu_1, \sigma_1^2) + (1 - p) \times f_N(x|\mu_2, \sigma_2^2)
\]

if we suppose as a simplification that the two members of the mixture are normal densities. If we knew the sample separation, i.e. which observation belongs to group 1 or 2, the inference problem would be very simple. But of course, the allocation of the observations is unknown.
8.3 Estimation procedures

It is convenient to introduce a new random variable called $Z$ that will be associated to each observation $x_i$ and that will say if $x_i$ belongs to the first component of the mixture $z_i = 1$ or to the second component of the mixture $z_i = 2$. Suppose that we know the $n$ values of $z$. We can compute easily the following statistics:

\[
\begin{align*}
    n_1(z) &= \sum \mathbf{1}(z_i = 1) \\
    \bar{x}_1(z) &= \frac{1}{n_1} \sum x_i \times \mathbf{1}(z_i = 1) \\
    s_1(z) &= \frac{1}{n_1} \sum (x_i - \bar{x}_1(z))^2 \times \mathbf{1}(z_i = 1) \\
    n_2(z) &= \sum \mathbf{1}(z_i = 2) \\
    \bar{x}_2(z) &= \frac{1}{n_2} \sum x_i \times \mathbf{1}(z_i = 2) \\
    s_2(z) &= \frac{1}{n_2} \sum (x_i - \bar{x}_2(z))^2 \times \mathbf{1}(z_i = 2)
\end{align*}
\]

These statistics give direct estimates for the parameters of the two members that we shall call $\theta_1$ and $\theta_2$. Of course we do not know the $z_i$, but we can compute the following probabilities for
each observation

\[
\Pr(z_i = 1|x, \bar{\theta}) = \frac{\hat{p} \times f_N(x_i|\bar{\theta}_1)}{\hat{p} \times f_N(x_i|\bar{\theta}_1) + (1 - \hat{p}) \times f_N(x_i|\bar{\theta}_2)}
\]

provided we have estimated \( p \) as \( \hat{p} = n_1/n \). We have then two solutions for allocating the observations between the two regimes:

- We allocate observation \( i \) to the first member if \( \Pr(z_i = 1|x, \bar{\theta}) > 0.5 \).
- We randomly allocate observation \( i \) to one regime according to a binomial experience with probability \( \Pr(z_i = 1|x, \bar{\theta}) \).

Once we have chosen between the two possibilities, we iterate the process. A deterministic allocation corresponds to the EM algorithm of Dempster et al. (1977) while a random allocation corresponds to an algorithm which is not far from a Bayesian Gibbs sampler.

### 8.4 Difficulties of estimation

As we have already said, estimating a mixture of densities is not a simple task. In the above writing of the data density, all the parameters are free to move in their domain. The likelihood function

\[
L(x; \theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j \times f(x|\mu_j, \sigma_j^2)
\]

goes to infinity if one of the \( \sigma_j \) goes to zero which happens if there are less observations in one cluster than there are parameters to estimate. So only a local maximum can be found.

The EM algorithm or the Gibbs sampler have global convergence properties. The EM algorithm converges to the maximum likelihood estimator. But both algorithms are sensitive to starting values.

There is a fundamental identification problem which is called a labeling problem. The likelihood function does not change is we change the order of the parameters. So, a usual way of identifying the parameters consists in imposing an ordering, for instance on the means.

### 8.5 Estimating mixture in \( \mathbb{R} \)

The complexity of the estimation procedures is reflected in the packages proposed in R. One of the many different available packages is \texttt{mixdist}. We shall now detail its use. In order to simplify the problem, the program start by considering an histogram, which means grouped data. So we have first to select the number of cells in the histogram. Then we have to give starting values for the parameters, and first of all the number of components. It it is quite safe to start by estimating a two component mixture. Mixture of a higher order are difficult to manipulate and many references in the empirical literature indicate that they are rarely successful. Usually an equal weight is given as a starting value for the \( p_i \). A visual inspection of the histogram gives clues about plausible values for the mean. The prior variance is small when the prior mean
correspond to a sharp part of the histogram and much larger for the prior mean corresponding to the tail.

library(mixdist)
FES.mix = function(y){
  chist = hist(y,breaks=100)
  y.gd = mixgroup(y,breaks=chist$breaks)
  y.par = mixparam(mu = c(50,80), sigma = c(10,50))
  y.res = mix(y.gd,y.par,"lnorm")
  print(y.res)
  plot(y.res)
}
FES.mix(y79)

In this code, we first determine break points with the instruction hist. Then, mixgroup is used for grouping the observations using the previously computed break points. mixgroup creates a data frame containing grouped data, a data frame being a special type of object in R. mixparam creates a data frame containing starting values for the mean and the standard deviation. If no other argument is given, it is assumed that the starting p are all equal while summing to one. mix is the proper function for estimation. It has at least three arguments: two data frames for the observations and the parameters. The third arguments give the density which is used. The choices for continuous densities are "norm", "lnorm", "gamma" and "weibull". Note that the last case weibull needs special type of entry for its parameters. The function weibullpar takes as an entry the prior mean and the prior standard deviation and creates a data frame containing the shape, scale and location parameters of the Weibull.

For FES 1979, we could not estimate a mixture of more than two components. We fitted two lognormals. The estimated parameters were

<table>
<thead>
<tr>
<th>member</th>
<th>p</th>
<th>µ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1369</td>
<td>45.42</td>
<td>6.764</td>
</tr>
<tr>
<td>2</td>
<td>0.8631</td>
<td>89.14</td>
<td>40.811</td>
</tr>
</tbody>
</table>

The graph show that the fit is rather good. It is rather difficult to identify a particular to group to each of these members. The second group seems to correspond to the large segment of the population as $p_2 = 0.85$ and the corresponding mean is not too large with $\mu_2 = 90$. The first group correspond to poorer people. A poverty line of half the mean is equal to 41.54.
Figure 8: Mixture of two lognormal densities
9 Exercises

1. Imagine a method of estimation for the poverty deficit curves and the dominance curves using order statistics.

2. Using the data provided in Table 1, draw an histogram for each year of the US income distribution. Draw the corresponding dominance curves at the order 1.

3. Compare the bootstrap results for the Gini index with the Davidson and the Giles methods which were given in Chapter 4, using the FES data.

4. When estimating an histogram, the number of cells has to be given. Compute the implicit bandwidth which is implied by the number of cells.

5. The Weibull density has an analytical cumulative distribution. Use this property to propose a way to adjust a Weibull density to the grouped data given in Table 1 for the US income distribution. Run the program in R.

6. Propose a regression method for estimating the main parameter of a Pareto distribution using the empirical Lorenz curve.

7. Propose an unbiased estimator for the Pareto I model, starting from the maximum likelihood estimator as given above.

8. Using the FES data set, fit a mixture of normal densities after taking the logs of the observations. Compare your results with the results obtained by considering directly a mixture of two lognormal densities.

References


