CHAPTER 4

# Diagnostic Plots for One-Dimensional Data

$Header: /u/math/j40/cvsroot/lectures/src/dataanalysis/onedim/chonedim.tex,v 1.15 2013/11/30 15:59:05 j40 Exp $

label ch:onedim

**Abstract**[1]

How do we draw a distribution on the line? We give a survey of some well known and some recent proposals to present such a distribution, based on sample data. We claim: a diagnostic plot is only as good as the hard statistical theory that is supporting it. To make this precise, one has to ask for the underlying functionals, study their stochastic behaviour and ask for the natural metrics associated to a plot. We try to illustrate this point of view for some examples.

## 4.1. Introduction

Though the general contribution of diagnostic plots to statistics is accepted, sometimes diagnostic plots seem more of a fashion than a tool. There are uncountable possibilities to design diagnostic plots, not all being of equal use. Diagnostic plots can and should be judged the same way as any other statistical method. We have to ask: What is their power? What is their reliability? While we may have to stay with accidental notes or examples for some time now, in the end a diagnostic plot is only as good as the hard statistical theory that is supporting it.

For many diagnostic plots, we are still far from having this theory. For some plots, we have to ask: What precisely are they trying to diagnose? How do we judge their reliability or confidence? For other plots, we know at least the statistical methods they are related to. In this paper, we consider plots as views on probability measures: We relate plots to functionals, operating on probability measures. If we have a functional defining a plot, we can proceed in three steps. We can ask which features are exhibited by the functional, and which are collapsed. As a second step, we can analyze what is retained by the

[1]Based on a paper presented at the $24^{th}$ meeting of the Arbeitsgruppe "Computational Statistics", Internationale Biometrische Gesellschaft (DR), Reisensburg 1992 (21.-24. Juni 1992). Published in P. Dirschedl, R. Ostermann (eds.): Computational Statistics. Heidelberg: Physica-Verlag 1994, ISBN 3-7908-0813-X. pp. 237-258.

empirical version and which stochastic fluctuation is to be expected. Third, we can optimize the functional and its empirical version to gain maximal power.

Where possible, we try to indicate classical tests related to the plot. If these tests meet the core of the plot, the power of the plot may be identified with and judged by the power of these tests. The associated functional may even indicate a notion of distance, or a metric, associated with the plot. We can use this to find natural neighbourhoods of a given empirical plot, leading to confidence sets of compatible models.

We restrict ourselves to a very modest case: Assuming a continuous distribution on the real line, we look at diagnostic plots based on a sample from this (unknown) distribution. We exclude some of the more difficult questions: we assume independent sample points with identical distribution. So we do not look at plots for the diagnosis of dependency, trend, heteroscedasticity or other inhomogeneities. As is to be expected, the opportunity is taken to advocate some new plots: the silhouette, the densitogram, and the shorth plot.

## 4.2. Diagnostic plots, what do we need them for?

We use diagnostic plots to investigate a data set by itself (a descriptive problem), or in comparison to a model distribution or model family (the one-sample-problem), or to compare two data sets (the two-sample-problem).

In a classical framework, we may want to apply a specific method, like regression or analysis of variance. The validity of this method will depend on conditions, often on conditions which in principle cannot be verified. The rôle of a diagnostic plot is that of a detector: Since we are unable to verify the preconditions, we use diagnostic plots to have at least a warning instrument.

In other situations, we may still be exploring. We have not settled on a specific model or method, but are looking at what the data are telling us. In the next step we select a certain model or decide to apply a certain method. Here diagnostic plots are a means to navigate through the models or methods at our disposition, and should be considered a model selection tool.

In any case diagnostic plots could be considered in a decision framework, either as filtering out bad situations after applying a model/method, or as selecting a model/method to be applied afterwards. It would be most appropriate to judge diagnostic plots as one step in an analysis process. But still too little is known about the interplay between use of diagnostic plots and application of formal models/methods.

The data we feed into diagnostic plots are rarely raw data. Often we use diagnostic plots on residuals. Of course the conditions we have to check refer to the error terms. The residuals are only some (model dependent) estimators of the errors. So the true story will be more complex than the i.i.d. simplification

told here. Where no model dependence is included initially, we still have had some choice how to measure the data: what we consider to be the data is a result of our choice of a measurement process. This can be a practical choice, or this may be culture dependent. Even in simple examples it may be more than just a linear change of scale (for example, energy consumption in a car is measured in miles per gallon in the U.S.A., and as liter per 100 km in Europe). Sometimes detection of the "proper" scale is the major achievement. For example the Weber-Fechner law in psychology tells us that the amount of energy which must be added to a stimulus to produce a detectable difference is proportional to the energy level of the stimulus. Hence using a logarithmic scale may be more appropriate for perception experiments than a linear energy scale. Choice of (nonlinear) scale may be a major application of diagnostic plots. Identifying the shape of a distribution is equivalent to finding a way to transform it to some model distribution.

## 4.3. What do we look for in diagnostic plots?

We use diagnostic plots to check for special features revealed by or inherent in the data. Of course, these checks are useful only if we know how presence or absence of these features affects the statistical methods we are going to apply (but then, if we do not know this, it would be wise not to apply these methods at all). Usually, for ordinary statistical applications, there are only few features we have to check. Here is a short check list:

**ToDo:** fix layou in item 1

**Missing or censored data.:** Contrary to what classical statistics would like to see, real data sets usually contain registrations meaning "below detection level", "not recorded", "too large". In survival analysis, respecting missing or censored data is a mark of the trade. Although missing or censored data are a pending challenge in practical statistics, we will not deal with this problem here.

**Discretization.:** Usually, all data we record are discretized (truncated or rounded to some finite precision, for example). For methods based on ranks, this may lead to ties, with appropriate corrections being well known. For other methods, these effects are often grossly ignored, although it would be easy to take them into account in tests of t or F type.

**Multi-modality.:** Sometimes, multi-modality is a hint to a factor which separates the modes and should be included in the analysis. In other cases, as for instance in psychological preferences and choices, multi-modality may be an inherent feature. Classical methods have notorious pitfalls if multi-modal distributions are involved.

**Symmetry and skewness.:** In best cases, skewness is an indicator for power transformations which might bring the data to a simpler model.

**Tail behaviour.:** Many classical methods are strongly affected by the tail behaviour of the distribution. Sometimes, tail problems may be avoided by going to more robust methods.

In real applications, we cannot assume an i.i.d situation. Hence we have to check for of dependency, trend, heteroscedasticity or other inhomogeneities as well. But for now we restrict ourselves to the i.i.d assumption.

## 4.4. Notation and conventions

We assume a distribution $F$ on the real line with density $f$ and look at diagnostic plots based on a sample $X_1, \ldots, X_n$ from the unknown distribution $F$. We assume that $X_1, \ldots, X_n$ are independent sample points. By $X_{(i:n)}$, or $X_{(i)}$ for short, we denote the i.th order statistics. $F_n$ is the empirical distribution function with $F_n(X_{(i:n)}) = i/n$. For any plot expressed in terms of $F$, the empirical version is the corresponding plot with $F_n$ replacing $F$. For simplicity, we identify distribution function and measure, allowing us to write $F((a, b]) = F(b) - F(a)$. For any plot we try to follow this sequence: We give a rough recipe how to generate the plot. Then we try to give a functional definition of the plot. Which features a preserved and which are lost by the functional? After that we study the fine points: what needs to be corrected in the rough plot? Then we turn to related tests and discuss optional choices. After each plot, we ask: how far have we got? What is the information we can gain so far, and what is still missing?

## 4.5. Histogram

4.5.0.1. *Recipe:* Chose histogram bins. For any bin, mark the hit count of data points hitting this bin.

4.5.0.2. *The underlying functional.* A histogram is "the" classical way to present a distribution. Its historical advantage is the ease of calculation - it can be reduced to putting registration notes into bins. The functional corresponding to a histogram is a discretization of the density: Given a decomposition of the real line into disjoint intervals $A_j, 0 \leq j \leq k$, we can define a histogram as the distribution, discretized to these intervals. The discretization gives a probability $p_j = F(A_j)$ for bin $A_j$. The vector of observed bin counts $n_j := \#\{i : X_i \in A_j\}$ has a multinomial distribution. Using Pearson's approximation (**Pearson, 1900**)

$$P(n_1, \ldots, n_k) \approx (2\pi n)^{-1/2} (\prod p_j)^{-1/2} \exp[-1/2 \sum (n_j - np_j)^2 / np_j + \ldots]$$

we see that the $\chi^2$-test statistics, the leading term in the exponential, is controlling the bin hit frequency for sufficiently large expected bin counts $np_j$.

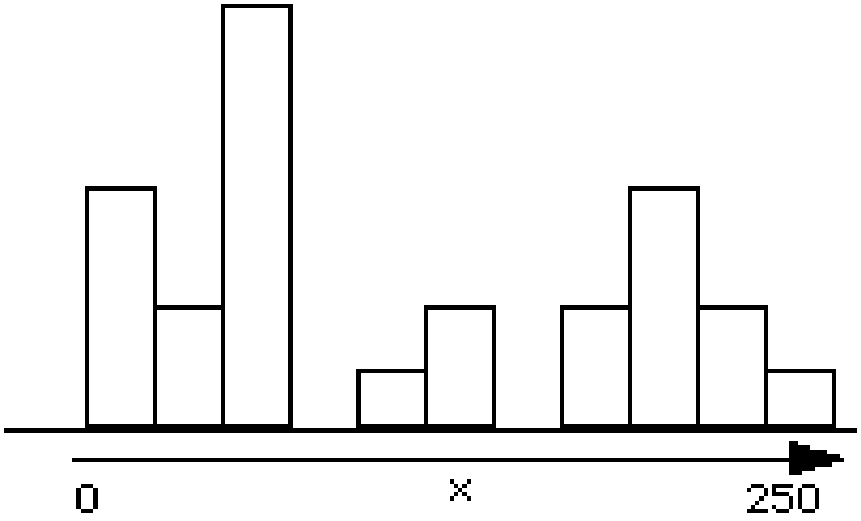


Figure 4.1: Histogram

Knowing the associated functional and its stochastic behaviour, we can tell what is to be expected from a histogram. A histogram can only show features which are preserved by the underlying functional, the discretization of the distribution. We loose all smoothness properties and local details of the distribution. The metric associated most naturally to the histogram is a $\chi^2$-metric. So if we want have an impression about the distributions which are compatible with our data, we should consider $\chi^2$ confidence bands. To obtain information about the power of the histogram, we can look at the $\chi^2$-test as a corresponding goodness of fit test.

Practical situations are slightly more complex. One complication may arise from the sampling scheme. A common case to sample for a certain time, instead of taking a fixed sized sample. This makes the total sample size a random variable. Under independence assumptions the vector of bin counts has a multivariate poisson distribution instead of a multinomial. A similar limit applies, but we gain one degree of freedom in the limit. Another complication may arise if we define our bins in a data dependent way. If the number of bins is small compared to the number of observations, the approximation still holds with good quality even if we use the data first to estimate location and scale, and use bins based on these estimators. We still have to correct for the degrees of freedom in the goodness of fit test.

If we have fixed reference distributions, we can head for optimal choice of bins. Common strategies are to take bins of equal probability with respect to the reference distribution, or to take bins of constant width with cut points $a_j = a_0 + j \cdot h,\ 0 > j > k$, for some bin width $h$, defining bins $A_0 = (-\infty, a_0]$, $A_j = (a_{j-1}, a_j]$. This still leaves us with the problem to decide about the number of bins. The classical recommendation for equal probability bins is to let the number of bins increase as $n^{2/5}$ (**Mann and Wald, 1942**). A more

detailed study of optimal choices of the bins with respect to the $\chi^2$-statistics for various types of reference distributions is presented in **Kallenberg et al. (1985**).

Walking through our short check list, it is an easy exercise to see how histograms can be modified to compensate for known censoring or discretization. If the data recording is discretized (for example by cutting down to a small number of reported digits) and the histogram is not adapted, this discretization may interfere with the histogram discretization and lead to patterns - an effect known as "aliasing". Tuning a histogram to look for multi-modality is a problem. Tuning it to identify asymmetry or tail behaviour again is comparably simple.

Perception of the histogram is a different matter. Prominent features, such as local maxima or general skewness, are perceived first: the eye does not perform the rescaling which would be necessary from a statistical point of view. Choice of bin width and bin locations are most critical to histograms (see 4.5.0.2). The rule is: if you use a histogram, don't use one. Use several.

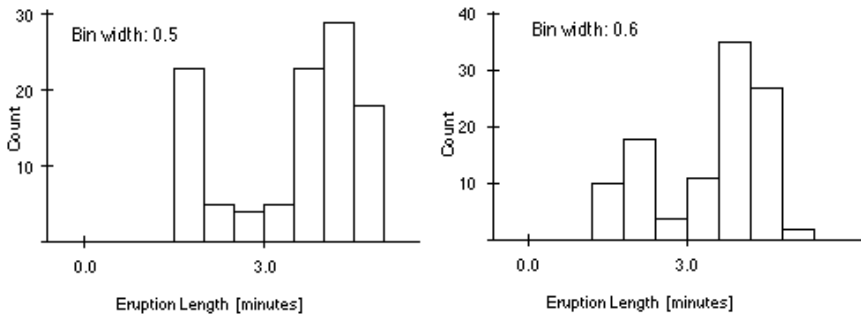


Figure 4.2: Histogram of eruption lengths of Old Faithful geyser. Same data set, but different bin widths. After **Silverman (1986**).

The quality of displays can be judged visually, and the usual concepts of statistics apply. The results may be surprising. Consider for example using a histogram as a goodness-of-fit-test for uniformity. What is the minimum sample size to achieve a (modest) confidence level of 95%

To qualify as a goodness of fit test for the hypotheses of uniformity with a confidence level of 95% for example, at most 1 histogram out of 20 should deviate from uniformity when taken from uniform data. So you can experiment to find the sample size needed to achieve this level.

With a sample size of 100 for example, unimodal structures, bimodal images, skewed histograms are common (see Figure figure 4.3 on page 4-7). To start with, you need about 500 data points to start working with a histogram on this confidence level. You can train yourself to work with smaller sample sizes. But then there are more powerful alternatives, and histograms should be put in their historical place.

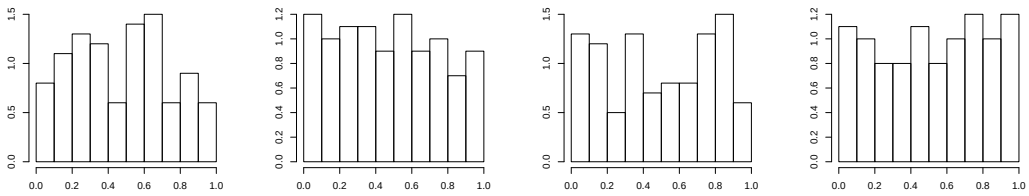

Figure 4.3: Random histograms of uniform samples, $n = 100$. Unimodal structures, bimodal images, skewed histograms are common.

Using histograms to analyze for discretization effects or multi-modality is inherently unsafe. A fairly safe use of histograms is for restricted purposes. Using a small number of cells to check for symmetry or tail behaviour is rather stable. But these purposes can be followed more effectively using Tukey's Box & Whisker plot, as discussed below.

4.5.0.3. *Histogram density estimators.* For the empirical version with bin counts $n_j := \#\{i : a_{j-1} < X_i \leq a_j\}$, the empirical histogram takes value $\widehat{p_j} = n_j/n$ on $A_j$. We can view a histogram as an estimator for the density, the histogram density estimator being defined by $\hat{f}(x) = \widehat{p_j}$ for $x \in A_j$. Judging the quality of this estimator by the integrated mean square error, $IMSE = \int E\{\hat{f}(x) - f(x)\}^2 dx$, at least for the case of constant bin width $h$ asymptotic results are available (**Scott, 1979**). For the histogram density estimator,

$$IMSE = \frac{1}{nh} + \frac{1}{12}h^2 \int f'(x)^2 dx + O(\frac{1}{n} + h^3),$$

that is the error goes to zero in order $o(n^{-2/3})$. The integrated mean square error is minimized for $h^* = n^{-1/3}\{6/\int f'(x)^2 dx\}^{1/3}$.

In particular, for the Gaussian distribution, $h^* \approx 3.49 \cdot \sigma \cdot n^{-1/3}$. Choosing a suboptimal bin width $h = c \cdot h^*$ gives an error $IMSE_h \approx IMSE_{h^*}(c^3 + 2)/30$. For example: choosing a bin width as 50% of the optimal bin width increases the error by 42%. Under regularity assumption, the optimal bin width requires approximately at least $(2n)^{1/3}$ bins. An upper bound for the bin width is $3.55 \cdot \sigma \cdot n^{-1/3}$. The regularity assumptions are: $\int f'(x)^2 dx > 0, \int f''(x)^2 dx < \infty$.

But judging histograms by the integrated mean square error of the histogram density estimator is not fair. First, in constructing the histogram we deliberately restricted possible estimators to stepwise constant functions. Using the same information as contained in histograms, we can generate better density estimators by allowing piecewise linear estimators. For example we can join the bin centres of a histogram. The frequency polygon, the resulting density estimator, can achieve better error rates (**Terell and Scott, 1985**). For their optimal bin width $h^* = 2[15/49n^{-1} \int \{f''(x)\}^2 dx]^{1/5}$ the error decreases as $n^{-4/5}$ under regularity conditions: $f''$ absolutely continuous, $f''(x)^2 dx > 0$, $f'''(x)^2 dx < \infty$.

Under these regularity conditions, we need at least $(147n/2)^{1/5}$ bins to achieve optimal bin width. An upper bound for the optimal bin width is $2.24 \cdot \sigma \cdot n^{-1/5}$. In particular, for the Gaussian distribution, $h^* \approx 2.15 \cdot \sigma \cdot n^{-1/5}$.

Thus the quality of the histogram, interpreted as a density estimator, can be greatly improved by allowing a slightly modified class of derived density estimators. But IMSE still is not an adequate criterion to judge histograms. Histograms are discretized versions of the distribution. The empirical histogram should be judged in comparison to this discretization, and competitors should be ranked on how they perform at (possible discretized) data.

## 4.6. Scatterplots

4.6.0.4. *Recipe:* Mark the data points.

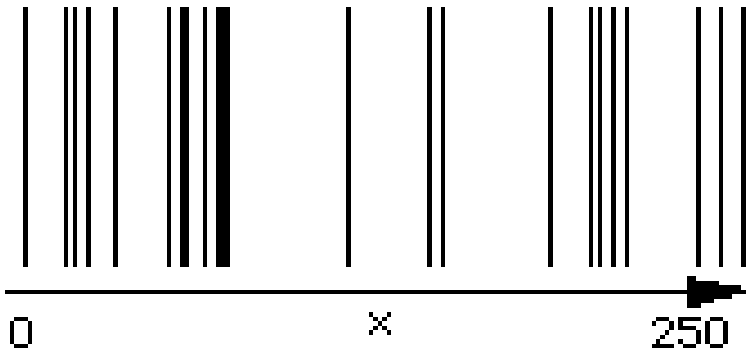


Figure 4.4: Scatterplot. Same data as in figure 2.

4.6.0.5. *The underlying functional.* The functional corresponding to the scatterplot is the density; the empirical version is $\sum \delta_{X_i}$. Problems may arise from discretisations or drawing resolutions. These effects may lead to ties in the empirical version. A simple technical solution is to use a gray pen, instead of a simple black pen. If the pen has gray level $K(y)$ at distance $y$ from its centre, we get gray level $\sum K(x-X_i)$ at a point $x$: the plot corresponds to a kernel density estimator with kernel $K$, or to the smoothed density as a functional. Any pen (or kernel) gives rise to a whole scaled family $K_h$, with $K_h(y) = 1/hK(y/h)$, where $h$ is the pen size, or bandwidth. For the functional, this means replacing the density $f$ by the smoothed density $f * K_h$, the convolution of $f$ and $K_h$. Of course you gain smoothness but loose details by using a larger bandwidth.

Scatterplots as such can be barely considered diagnostic tools, but they may provide helpful orientation when combined with other plots: we continue to use them, but it is hard to pin down what we gain. To use them for diagnostic purposes, they must be enhanced. For example to check the symmetry behaviour,

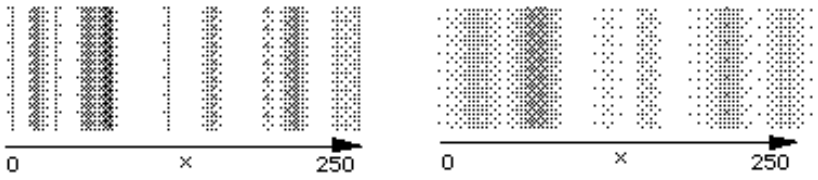


Figure 4.5: Scatterplot, using two different gray pens. Same data as in figure 4

we can look at symmetry centres. An appropriate plot suggested by J.Tukey (after (**Wilk and Gnanadeskian, 1968**)) is to show $(X_{(n-i+1)}+X_{(i)})$, plotted against the distance $X_{(n-i+1)}-X_{(i)}$. To check for discretization effects, we can look at the plot of the differences $X_{(i+1)}-X_{(i)}$ against $X_{(i)}$.

**4.6.1. Smoothed scatterplots and kernel density estimators.** In principle, the complete information of a sample is represented in a scatterplot. Perception however is easily trapped by sample size effects: small sample sizes will give the impression of pattern and inhomogeneities even for uniform samples; large sample sizes will hide non-uniformities for any distribution. This problem is even more complicated for kernel density estimators: we have the choice of a pen (or a kernel, if you like). Conventionally, this problem is split into two: choosing the pen shape (or kernel type) and the pen size (or bandwidth). We meet the same problem we have encountered with histograms: what we see depends critically on these choices. But we do not know how to judge these choices. There is a mathematical hideaway. If we accept that the density is our target functional, any distance measure between the (normalized) kernel density estimate and the true density can be used as a measure of fit, and of course $L_2$ distance is the easiest to deal with. Call $\hat{f}_h(x) = n^{-1}h^{-1}\sum K((x-X_i)/h)$ the kernel density estimator for kernel $K$ and bandwidth $h$. Let $\hat{h}_0 = \hat{h}_0(f,X)$ be the smallest minimiser of the integrated square error $\Delta(h) = \int(\hat{f}_h - f)^2dx$ and $h_0$ the smallest minimiser of the mean integrated square error $M(h) = \int E(\hat{f}_h - f)^2dx$. Under regularity conditions, for any (empirical) bandwidth $\hat{h}$ we have.

$$\Delta(\hat{h}) - \Delta(\hat{h}_0) = 1/2(1+o_p(1))(\hat{h}-\hat{h}_0)^2M''(h_0)$$

(**Hall and Marron, 1987**). While this does not help to estimate the error, it says that minimizing the integrated square error is essentially equivalent to optimizing the bandwidth for the data at hand. But $\hat{h}_0$ can be represented as $\hat{h}_0 = A_1 + n^{1/5}A_2\int f'^2dx + o_p(n^{-3/10}$, where $A_1$ and $A_2$ are functions of the data, not depending on $f$ (**Hall and Johnstone, 1992**). $A_2$ does not vanish asymptotically. So determining an optimal bandwidth is related to estimating $\int f'^2dx$. An optimal rate of $n^{-1/2}$ for the estimation of $\int f'^2dx$ makes the relative error of approximating $\hat{h}_0$ at best of order $n^{-1/10}$ (**Hall and Johnstone, 1992**). These results tell us why optimal bandwidth selection is a hard problem

even for very large sample sizes and continuous distributions, let alone for real data, that is for finite sample sizes and data truncated or rounded to some finite precision.

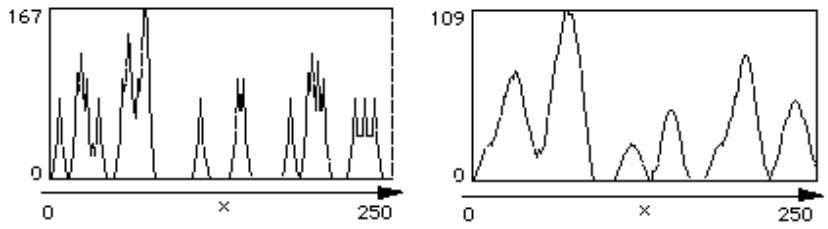


Figure 4.6: Pixel intensity for gray level plots of figure 5

It is possible to base goodness of fit tests on kernel density estimators (**Mammen, 1992**, Ch. 3). But the stochastic behaviour of kernel density estimators is difficult. There is no clear notion of distance or variation associated to kernel density estimators. There are candidates, among them distances based on (penalized) square errors. These are a treatable mathematical concept, but $L_2$ confidence bands are not too helpful from a data analytical point of view. The information gained from scatterplots, including kernel density estimators, is doubtful. Checking the list of critical features given above, it is hard to spot a feature that is reliably detected and reported by a scatterplot. **Silverman (1981**) made an attempt to exploit kernel density estimators as a diagnostic tool to analyze for multi-modality. A simpler approach, the densitogram (related to the excess mass test), is given below.

## 4.7. Distribution function and related plots

4.7.0.1. *Recipe:* Sort the data points. For any point, mark the proportion covered (the frequency of data points not exceeding this point).

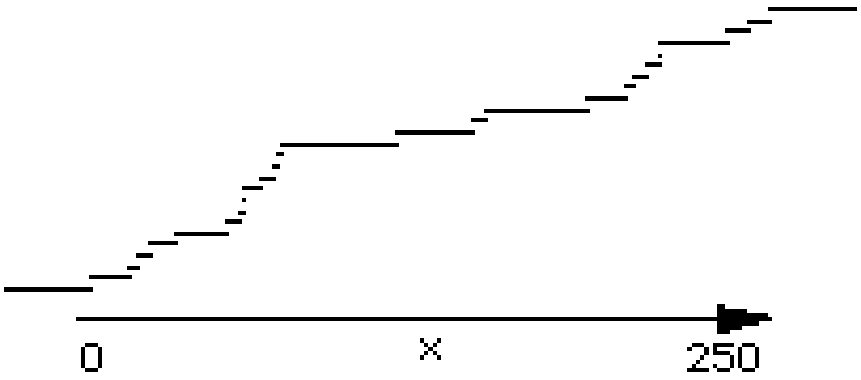


Figure 4.7: Distribution function

4.7.0.2. *The underlying functional.* The distribution function gives the probability of half-lines $F : x \mapsto F(x) = P[X \leq x]$. It can be estimated by its empirical version, $F_n$ where $F_n(x_{(i:n)}) = i/n$. The stochastic behaviour is described by the Glivenko-Cantelli Lemma: we have $\sup_x |F_n(x) - F(x)| \to^P 0$. The error has a Brownian bridge asymptotics: $\sqrt{n}(F_n - F) \to Z$. Viewed as an estimator for $F$, $F_n$ has a certain general optimality: For any loss function of supremum type, the empirical distribution function is asymptotically a minimax estimator (Dvoretzky-Kiefer-Wolfowitz-Theorem). For continuous distributions $F$, the distance $D_n = \sup_x |F_n(x) - F(x)|$ has a distribution which does not depend on $F$. This allows for simultaneous confidence bands: if $c$ denotes the $\alpha$-quantile of the Kolmogorov-Smirnov statistics, we have $F(x) \in [F_n(x) - c, F_n(x) + c]$ for all $x$ with probability of at least $1 - \alpha$. The Kolmogorov distance is the metric associated most naturally to the distribution function; the distribution function is easy to reconstruct, and its statistics is well understood. Interpreting it needs some education.

**4.7.1. Plots related to the distribution function.** Comparing two distribution functions visually is quite difficult. We have to compare two graphs, both piecewise constant and monotonous. Most interesting features are hidden in details. We can help perception by using a transformation which gives a near-to linear graph for corresponding distributions. If we have a given reference distribution, our choices are to align quantiles by transforming the probability scale (the quantile-quantile-plot), or to align probabilities by transforming the data scale (the percentage-percentage plot).

**4.7.2. Quantile-quantile-plot (Q-Q-plot).**

4.7.2.1. *Recipe:* Choose a reference distribution. Sort the data points. For any data point, find the proportion of observations not exceeding this data point. Plot the data point against the corresponding quantile of the reference distribution.

4.7.2.2. *Q-Q-plot details.* To transform the probability scale, we transform a probability to the corresponding quantile. The Q-Q-plot compares two distributions by plotting quantile against quantile. If $F$ and $G$ are the distributions to be compared, $X \sim F$, $Y \sim G$, the Q-Q-plot shows the curve $\alpha \mapsto (x_\alpha, y_\alpha)$. In terms of the probability distributions, this is the graph of $x \mapsto G^{-1} \circ F(x)$. Again, orientation has been chosen to give an easy empirical version $x \mapsto G^{-1} \circ F_n(x)$.

If $F$ and $G$ coincide, the Q-Q-plot is a diagonal line. If one is a linear transformation of the other, the Q-Q-plot is linear. The Q-Q-plot shows a high resolution in regions of low densities and vice versa. As a consequence, it emphasizes the tail behaviour for long-tailed distributions (**Wilk and Gnanadeskian, 1968**), and emphasis on the tails combines unluckily with high variation.

If $G$ is the true distribution, $G = F$, the Q-Q-plot of $F_n$ against $F$ is given by $(X_{(i:n)}, x_{i/n})$ where $x_{i/n}$ is the $i/n$ quantile. In particular, for $G = U[0,1]$ we have $x_{i/n} = i/n$; so in this case the Q-Q-plot coincides with the empirical distribution function.

If $U_i$ are iid $\sim U[0,1]$, $U_{(i:n)}$ is distributed as $\beta(i, n-i+1)$. Hence in the general case $E(F(X_{(i:n)})) = E(U_{(i:n)}) = i/(n+1)$: the empirical version of the Q-Q-plot is biased. We can take this into account to get an "unbiased" empirical plot by using plot positions $(X_{(i:n)}, G^{-1}(i/(n+1)))$ for an empirical Q-Q-plot to compare the observations $X_i$ with the model distribution $G$. This is the convention used by **Weibull (1939)**. But getting the mean behaviour right is only one part of the difficulty. Since you will not apply a diagnostic plot to a mean situation, but to a sample, you are prone to be affected by the notorious skewness of empirical quantile distributions. This is the origin for many fine points to be considered in the actual mapping (**Kimball, 1960**; **Harter, 1984**).

Direct relatives of the Q-Q-plot are goodness-of-fit tests based on the regression of order statistics on expected order statistics, like for example the Shapiro-Wilk test (**Shapiro and Wilk, 1965**).

### 4.7.3. Percentage-percentage plot (P-P-plot).

4.7.3.1. *Recipe:* Choose a reference distribution. Sort the data points. For any data point, find the proportion of observations not exceeding this data point. Plot the proportion against the corresponding proportion of the reference distribution.

4.7.3.2. *P-P-plot details.* To transform the data scale for linearity, we have to transform X to the corresponding probability under the reference distribution. If $F$ and $G$ are the distributions to be compared, $X \sim F$, $Y \sim G$, the P-P-Plot shows the curve $X \mapsto (G(X), F(X))$. In terms of the probability distributions, this is the graph of $\alpha \mapsto F \circ G^{-1}(\alpha)$. We apply this with $G$ in the role of a reference distribution. Orientation has been chosen here to avoid the discontinuities in $F_n$, that is to give an empirical version $\alpha \mapsto F_n \circ G^{-1}(\alpha)$.

If $F$ and $G$ are identical, the P-P-plot will be straight line. P-P-plots are not preserved under linear transformations: they are not equivariant. So usually P-P-plots will be applied only to distributions standardized for location and scale. For the empirical version, this preferably is done using robust estimators of location and scale. As for the Q-Q-plot, the skewness of the empirical quantile function should be considered in the actual mapping. But in contrast to the Q-Q-plot, for the P-P-plot, high variability is not combined with sensitivity in the tails. So choice of the proper plotting position is a fine point for the P-P-plot, whereas it is critical choice for the Q-Q-plot. Goodness-of-fit tests can be constructed based on the linearity of the P-P-plots (see **Gan and Koehler (1990)**).

**4.7.4. Other plots related to the distribution function plot.** The plots based on the distribution function suffer from the tail-orientation of the distribution function. It measures half infinite intervals, and local behaviour can be judged only by looking at differences. This is easy to compensate using a third dimension: you can localize the probability mass to intervals and define a probability mass plot $(a, b) \mapsto F(a, b]$, with the obvious empirical version. But readability and practical use are doubtful.

## 4.8. Box & Whisker-Plot

4.8.0.1. *Recipe:* Find the median and quartiles, and mark them. Connect the range of points which are not too far from the median (judged by the interquartile distance). Highlight all points which are out or far out.

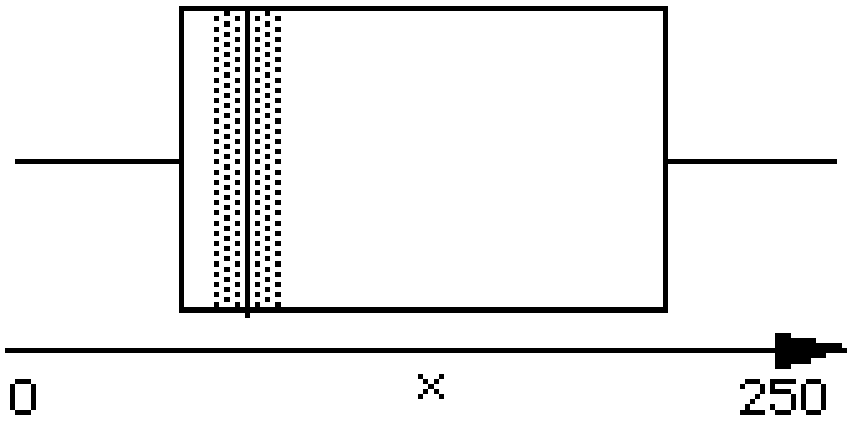


Figure 4.8: Box & Whisker plot, with a confidence interval for the median marked grey.

4.8.0.2. *Box & Whisker plot details.* In more detail, the construction is: Find the median of the data points, and mark it. Find the median of the subset above the general median, mark it, and call it the upper hinge. Find the median of the subset below the general median, mark it, and call it the lower hinge. Let $\Delta$ be the distance between the hinges. Draw a whisker from the box to the last data point not exceeding *upper hinge* $+ 1.5 \cdot \Delta$. Mark all data points in the out area between *hinge* $+ 1.5 \cdot \Delta$ and *hinge* $+ 2.5 \cdot \Delta$. Highlight all far out points exceeding *hinge* $+ 2.5 \cdot \Delta$. Do the corresponding for the lower hinge.

John Tukey's Box & Whisker plot (**Tukey, 1975**) is one of the gems of data analysis. Like the histogram, the Box & Whisker plot represents a discretization of the density. But where the histogram discretizes on the observation scale, the Box & Whisker plot discretizes on the probability scale. The discretization varies, from a rough 25%-discretization in the centre part, to a $1/n$ discretization for a sample size n in the tails.

The Box & Whisker plots achieve to present general information about the core of the data, with information hiding in this area. On the other hand, they highlight the exceptional. The exceptional data might be just tail effects, or it might be genuine outliers - they are worth a second look anyway.

The Box & Whisker plot is best understood by following its construction. Roughly, the Box & Whisker plot marks median and quartiles, and exceptional points. We will try to look at the ideas of Box & Whisker plots more carefully here. For the Box & Whisker plot, first we try to get an estimator for the location. The data median is used as the obvious (robust) candidate. The centre line of the Box & Whisker plot marks the median. Now we estimate the scale. Since we have already estimated the location, we can use this information. Given an estimator for the location, estimating the scale would be useful in exceptional cases: it would be meaningful only for symmetric distributions. Given the location estimator, we construct two scale estimators, a lower and an upper scale estimator. In the absence of ties, we could use the differences between median and lower/upper quartile as estimators. Since we must be prepared for discretization effects, we must be more careful. We use the median of the lower or upper half instead - Tukey's hinges. Finally, using these scale estimators, we estimate "central" areas, and mark all points outside.

The construction gives us some idea when Box & Whisker plots start to work. Although it is not made for this purpose, we can look at the box as some confidence set for the median. For a continuous distribution $F$, $F(X) \sim U$ Hence $F(X_{(i:n)}) \sim \beta(i, n-i+1)$. So if the lower hinge is at order statistics $i$, we can use the beta integral to find $P(X_{(i:n)} \leq X_{0.5})$. Instead of using the beta distribution, you can use the binomial: $X_{(i:n)} \leq X_{0.5}$ iff at most $i$ observations do not exceed $X_{0.5}$, and this count is a binomial event. In the presence of ties, calculations become more intricate, but for our purposes, a rough calculation is sufficient. If we take $F(X_{0.5}) \approx 0.5$ and $i \approx [n/4]$, we need $n \geq 13$ to have at least 90% probability for the box to cover the true median. To have a coverage of at least 99%, we need $n \geq 29$. But, has been said: be aware of ties!! As a rough rule, $n = 40$ may be a good starting point.

Tukey's Box & Whisker plot takes into account many possibilities and pitfalls of real data sets. It is very easy to miss these fine points, as can be seen from popular software packages. The Box & Whisker plot is particularly powerful in analyzing the overall structure of a distribution, like location, scale and outliers. But it still leaves the needs to diagnose other features. Discretisations are in no way reflected in the Box & Whisker plot. The tail behaviour is made a caricature: if there are tails, outliers are identified. But if the tails are too steep, heavier tails are invented: even a uniform distribution is shown with tails. Multi-modality is ruled out: the Box & Whisker plot knows about central location, but has no space for modes. It must be accompanied by other plots.

Higher dimensional variations are in (**Rousseuw et al., 1999**) and more in (**McGill et al., 1978**).

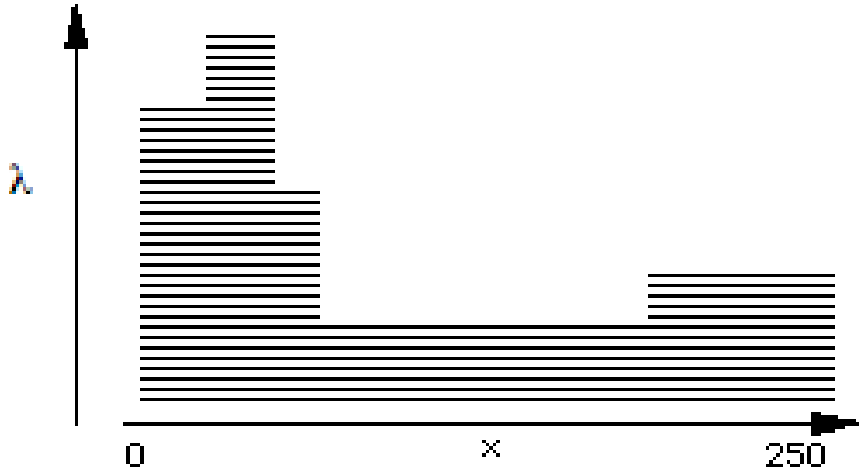


Figure 4.9: Silhouette. Locations of maximal excess mass $E(\lambda) := \int (f-\lambda)^+ dx$ for varying levels $\lambda$.

## 4.9. Silhouette and Densitogram

4.9.0.3. *Recipe:* Choose a family of sets serving as a model (e.g. sets composed of one or two intervals, if you are looking for bi-modality). Choose a level $\lambda$ . Mark the maximal set with average hit density exceeding level $\lambda$. Do this for a choice of levels $\lambda$.

4.9.0.4. *The underlying functional.* If you are looking for specific features in your data, it is possible to design diagnostic plots for these features. ***Silhouette*** and its accompanying plot, the ***densitogram***, are plots tuned to inspect multimodality (**Müller and Sawitzki, 1991**). Both are based on the idea that a mode of a distribution is a location where the probability mass is concentrated. A corresponding functional is the excess mass,

$$E(\lambda) := \int (f-\lambda)^+ dx,$$

giving the probability mass exceeding $\lambda$. Restricting the allowed sets in an appropriate way to a family $\mathscr{C}$, we define

$$E_{\mathscr{C}}(\lambda) := \sup_{C\in\mathscr{C}} \int_C (f-\lambda)dx = \sup_{C\in\mathscr{C}} (F - \lambda\mathrm{Leb})(C).$$

The silhouette marks the maximizing sets, for any level $\lambda$. The densitogram shows the excess mass, as a function of lambda. The cue lies in the freedom to choose $\mathscr{C}$. For unimodal distributions, $\mathscr{C} = \mathscr{C}_1$ should be the family of intervals and $E(\lambda) = E_{\mathscr{C}_1}(\lambda)$. For bimodal distributions, $\mathscr{C} = \mathscr{C}_2$ is made of the disjoint unions of two intervals and $E(\lambda) = E_{\mathscr{C}_2}(\lambda)$. Given a hypothesis on the modality, silhouette and densitogram can be estimated using the empirical excess mass

$$\widehat{E_{\mathscr{C}_m}(\lambda)} = \sup_{C\in\mathscr{C}_m} (F_n - \lambda\mathrm{Leb})(C), \quad m = 1, 2, \ldots.$$

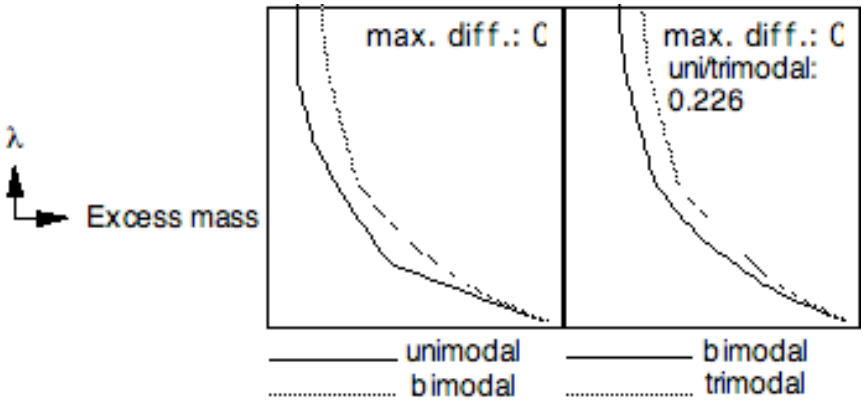


Figure 4.10: Densitogram, the excess mass concentration curve. Same data as figure 2. Excess mass estimated under assumption of uni- bi- and tri-modality. By assuming bi-modality, an additional excess mass of 16.4% of the data is covered.

4.9.0.5. *Silhouette and densitogram details.* As an estimator for the location of the mode, the silhouette shares a rate of order $n^{-1/3}$ with density estimation based methods. But the number of modes in the silhouette is more reliable even for small sample size. For the densitogram, the associated test is the excess mass test for multi-modality (**Müller and Sawitzki, 1991**): $\sup_\lambda \widehat{E_{\mathscr{C}_2}(\lambda)} - \widehat{E_{\mathscr{C}_1}(\lambda)}$, the maximal difference between excess mass, estimated on the assumption of bi-modalitity, and excess mass, estimated on the assumption of uni-modality, can be used as a test statistic for multi-modality. For a bimodal distribution, the maximal excess mass difference $\sup_\lambda \widehat{E_{\mathscr{C}_2}(\lambda)} - \widehat{E_{\mathscr{C}_1}(\lambda)}$ is half the total variation distance, between F and the closest unimodal distribution. This points to the Kuiper metric as a distance measure related to excess mass.

On the unimodal distributions, the error rate of these excess mass estimates is of order $n^{-1/2}$. In more practical terms: the difference between both excess mass curves starts providing a reliable indicator for multi-modality for a sample size n in the range 20 to 50.

## 4.10. Shorth-Plot

4.10.0.6. *Recipe:* Choose a coverage $\alpha$. For any point, get the length of the shortest interval containing this point and covering at least an $\alpha$-fraction of the data (at least $\alpha \cdot n$ data points). Do this for a selection of coverages $\alpha$.

4.10.0.7. *The underlying functional.* The shorth is the smallest interval containing at least 50% the distribution:

$$S = \arg\min\{|I| : I = [a, b], P(X \in I) \geq 0.5\}.$$

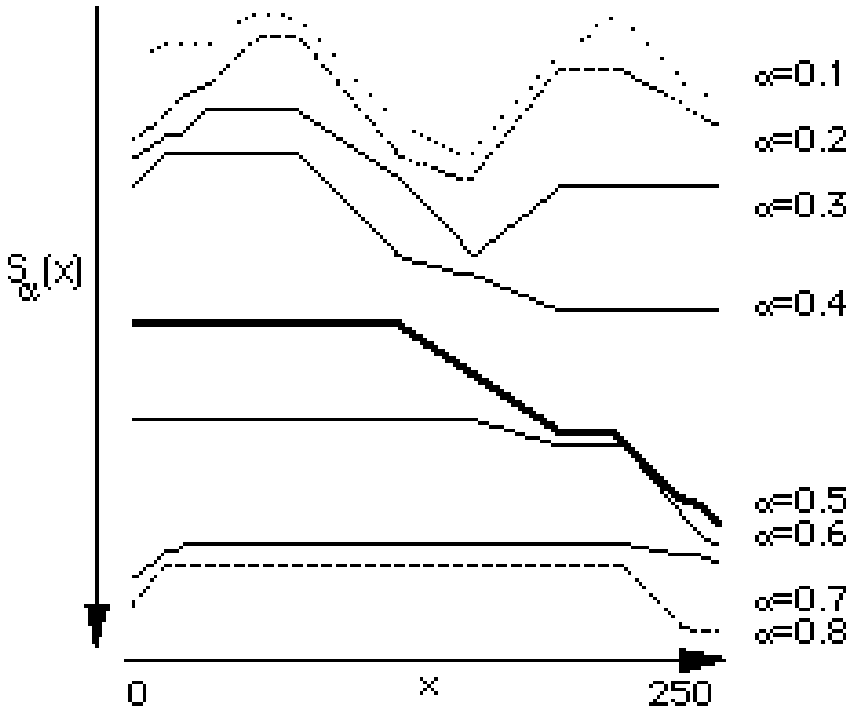


Figure 4.11: Shorth Plot. The shorth length axis points downwards.

Here $|I|$ is the length of he interval $I$. More generally the $\alpha$-shorth is the smallest interval containing at least an $\alpha$ fraction of the distribution:

$$S_\alpha = \arg\min\{|I| : I = [a, b], P(X \in I) \geq \alpha\}.$$

For data analysis, we can localize the shorth. We define the $\alpha$-shorth at $x$ as the smallest interval at $x$ containing at least a proportion $\alpha$ of the distribution

$$S_\alpha(x) = \arg\min\{|I| : I = [a, b], x \in I, P(X \in I) \geq \alpha\}.$$

In particular, the shorth at $x$ is defined as $S(x) := S_{0.5}(x)$. The shorth plot is the graph of $x \mapsto |S_\alpha(x)|$ for a selection of coverages $\alpha$. **FixMe:** ref

4.10.0.8. *More about shorth plots.* Andrews et al. (1972) use the (not localized) shorth to construct a robust estimator of location. The shorth procedure takes the centre of the empirical shorth as location estimator. Unfortunately this estimator of location has an asymptotic rate of only $n^{-1/3}$, with non-trivial limiting distribution. However **Grübel (1988)** shows that the length of the empirical shorth is a reasonable estimator of scale, converging with a rate of $n^{-1/2}$ to a Gaussian limit. Mass concentration now can be represented by the graph of $x \mapsto |S_\alpha(x)|$. A small length of the shorth signals a large mass concentration. To facilitate optical interpretation, we plot the negative of the lengths. The shorth plot was proposed as a means to investigate mass concentration by **Sawitzki (1992)**. It is easy to compute, avoids the bandwidth selection problems, and allows scanning for local as well as for global features of the distribution. This asymptotics can be extended to cover the asymptotics as a process (**Einmahl et al., 2010March**; **2010November**). The good rate

of convergence of the shorth estimator makes it useful already for moderate sample size.

## 4.11. Summary

How far have we got? The general purpose plots (histogram, scatterplot, distribution function) can be applied, but provide doubtful information per se. They can be sufficiently restricted to provide reliable information on questions as rough symmetry or tail behaviour. But the rough information seems to be read off more readily from constructions as the Box & Whisker plot. The general purpose plots may have an advantage if we move to the one-sample problem or the two sample problem, where no immediate generalization of the Box & Whisker is available. Multi-modality stays a critical feature. The classical general purpose plots tend to be misleading: random fluctuation may appear as modes, and no controlled measure of significance is available. The general purpose plots are not likely to oversee modes, but are prone to show more than should be shown. The Box & Whisker plot does not address the problem of modes at all. We can construct special plots for the detection of modes, such as silhouette and densitogram. We loose information on density and tails in these plot. The shorth plot tries to make a compromise, allowing for information about modality as well as on local density, but avoiding the fluctuation affecting (smoothed) scatterplots and other classical plots. It may be a candidate for a general purpose plot. But practical evaluation and analysis is still necessary.

The true distribution is usually hidden from our eyes. Since we were using simulated data here, we are able to look at the true distribution. In the examples shown here, we used a bimodal distribution with two strong modes. Sample size for the illustrations was 25 data points.

## 4.12. Higher Dimensions

While we want to restrict to dimension one here, it may be worth to have an outlook to higher dimensions.

The histogram immediately generalizes to arbitrary dimension. The critical statistics stays the $\chi^2$ statistics. The only change is that we have more freedom define the cells for the histogram. Rectangles defined by marginal intervals are immediate, but not the only candidates.

Kernel density estimators generalise to higher dimensions without any change. But the convergence rates decrease as the dimension increases.

The distribution function can be generalised using

$$F(x) = P(X|X_i \leq x_i \text{ for } i = 1 \dots n).$$

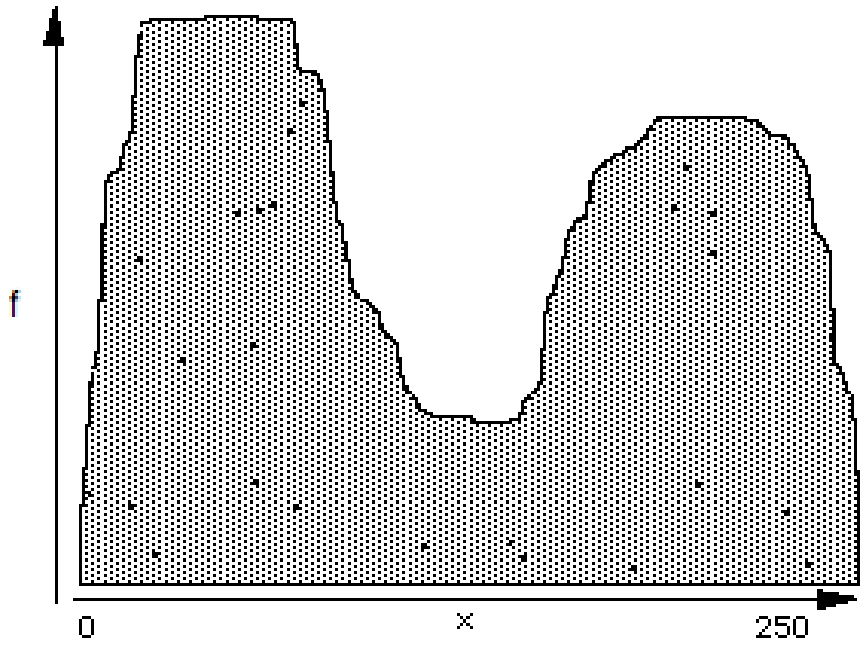


Figure 4.12: True distribution used for the examples; data points sampled by acceptance/rejection.

But we loose universal bounds as provided by Kolmogorov-Smirnov in dimension 1.

There have be various attempts to generalise box&whisker plots to higher dimensions (**Rousseeuw and Ruts, 1996**; **Rousseuw et al., 1999**). But so far, these are proposals under discussion.

The excess mass approach, leading to silhouette and densitogram, extends to higher dimensions as a concept (**Polonik, 1995**; **Nolan, 1991**). But efficient algorithms become a problem already in dimension 2.

Generalisations of the shorth length to higher dimensions are not know.

## 4.13. Some Sample Displays

The use of diagnostic plot can be trained using simulated data, and you should spend some time to generate simulated data for frequent challenges in your area, and try diagnostic tools. Here is a small collection. All data in this section are simulated for a sample size $n = 100$, and the same simulated data is shown in several displays.

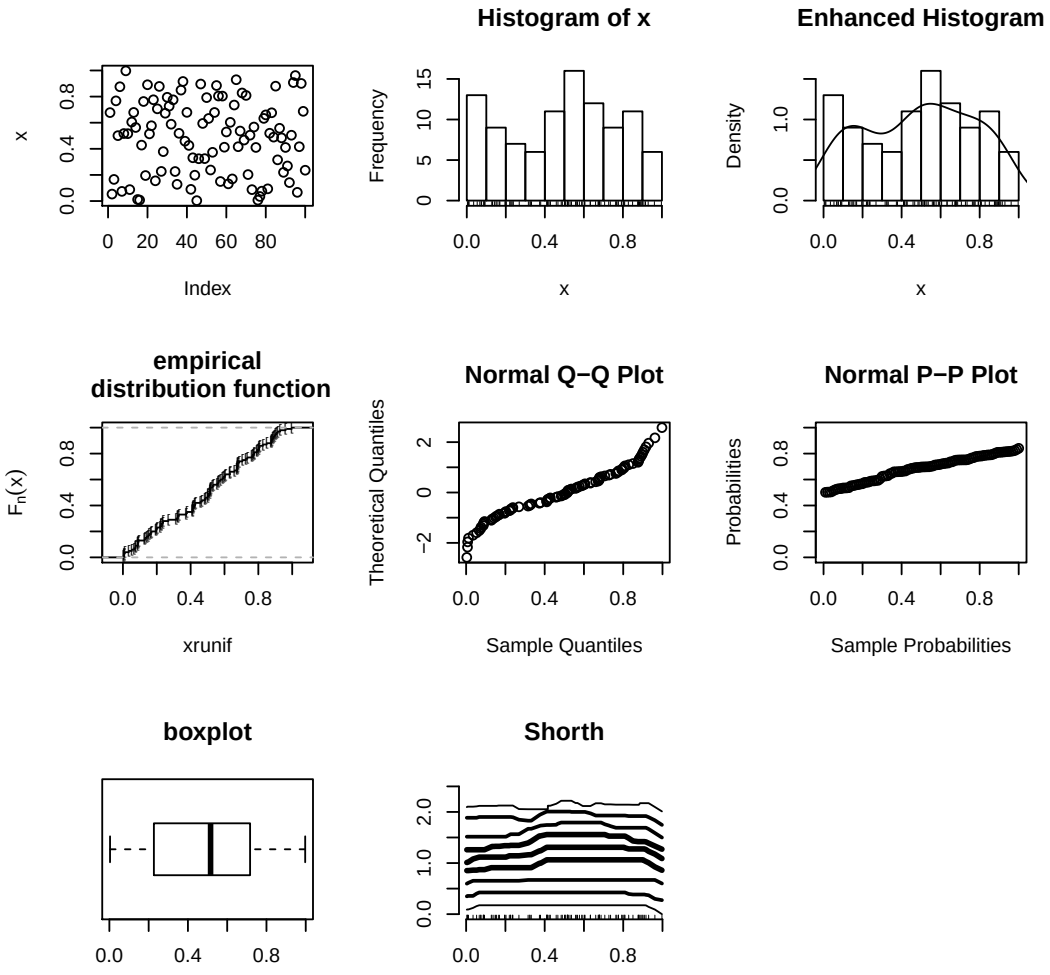


Figure 4.13: Uniform random numbers, $n = 100$

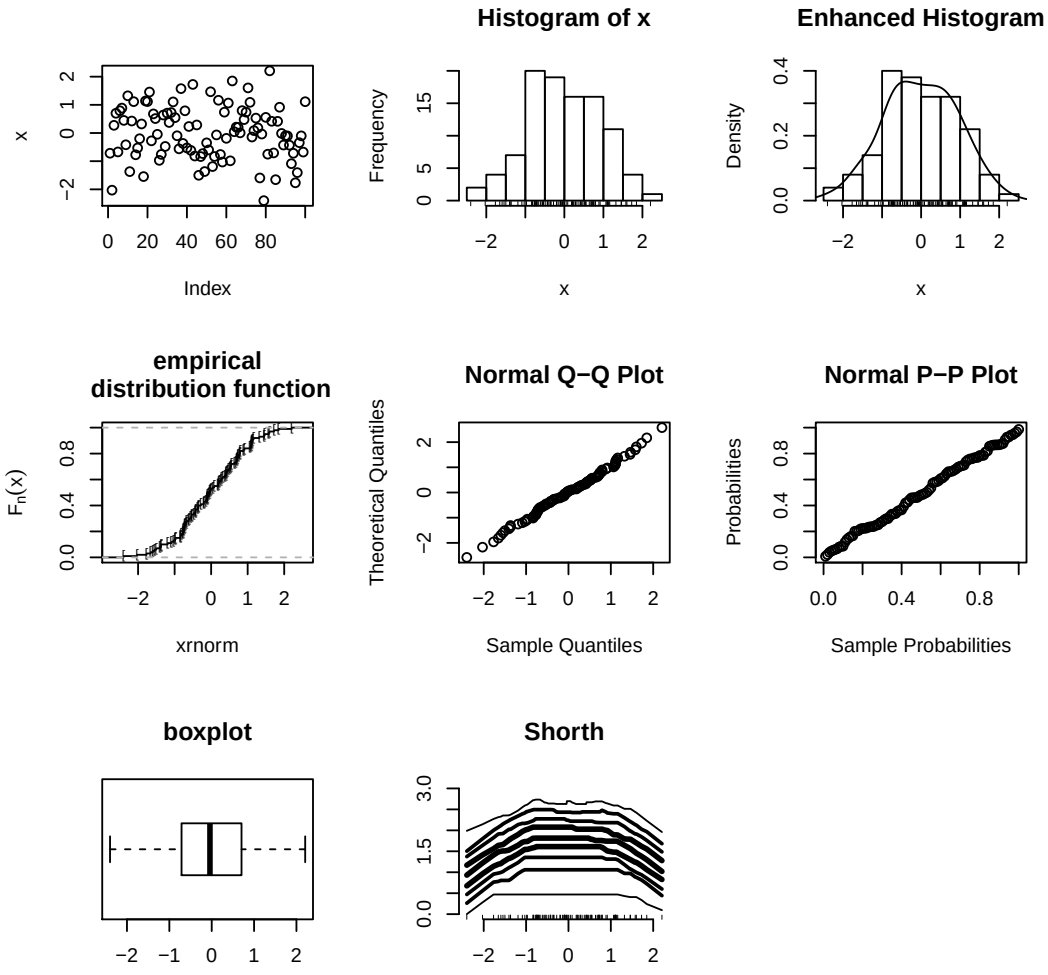


Figure 4.14: Normal random numbers, $n = 100$

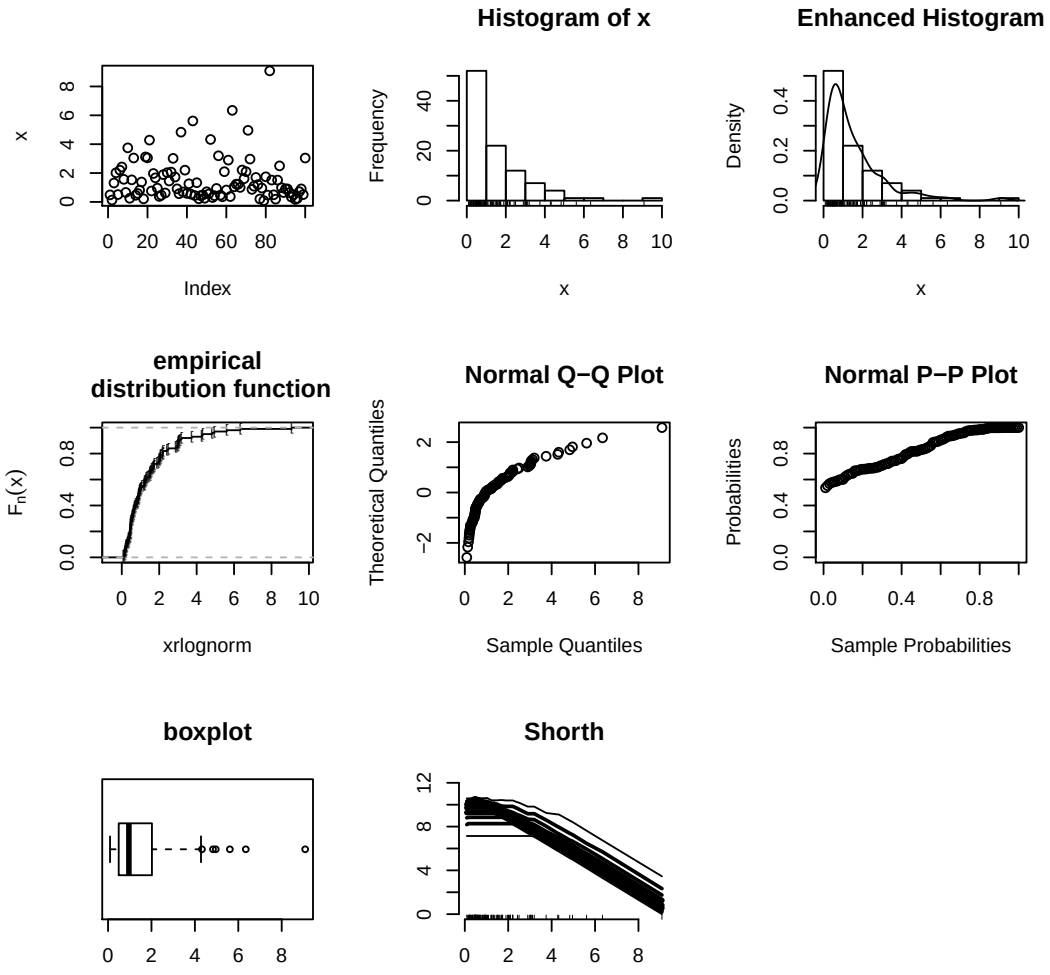


Figure 4.15: Log-Normal random numbers, $n = 100$

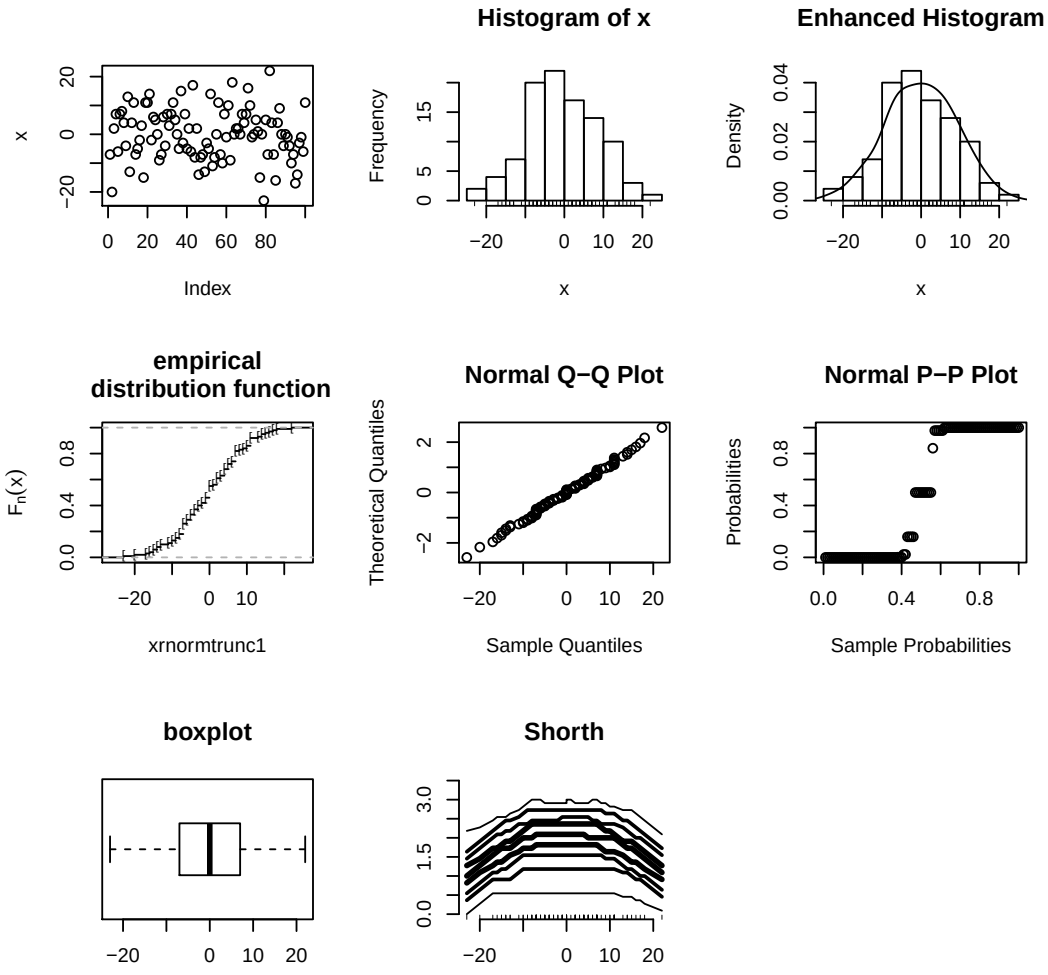


Figure 4.16: Normal random numbers with 1 decimal digit precision, $n = 100$

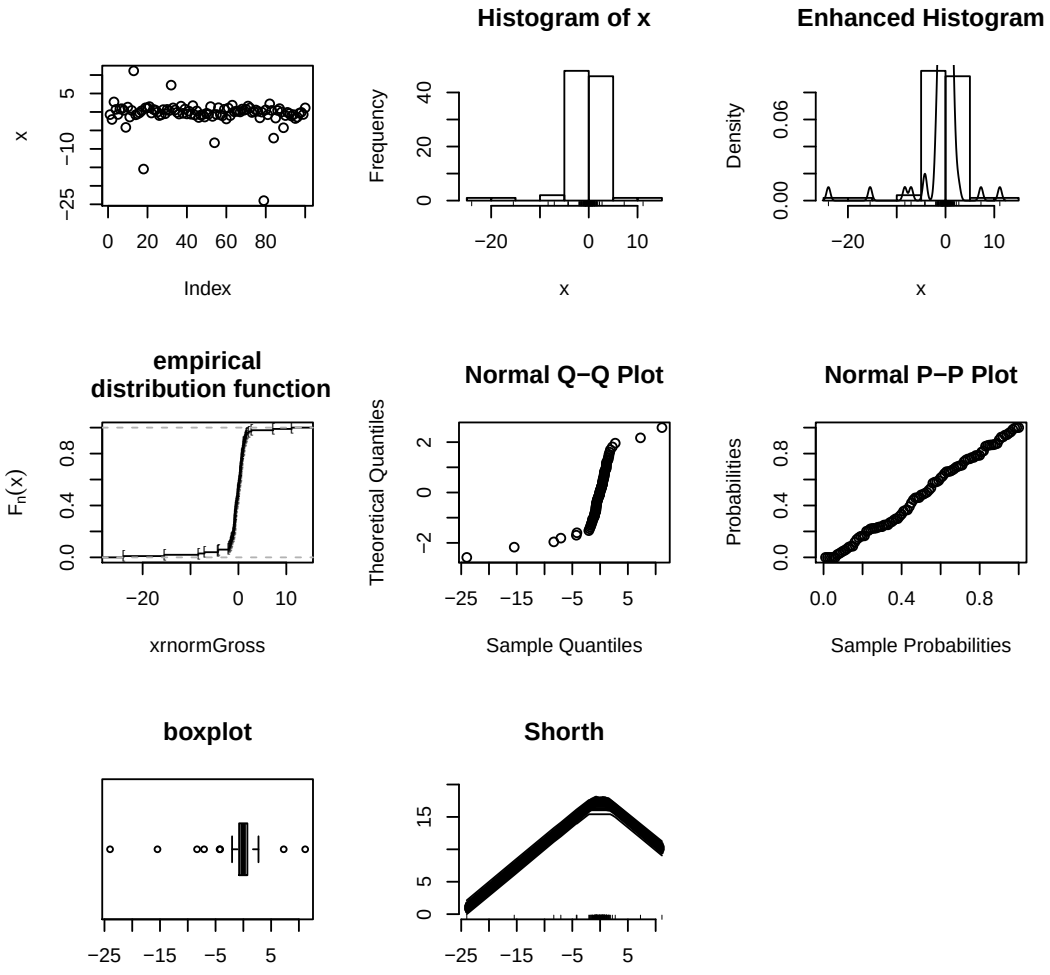


Figure 4.17: Normal random numbers with 10% gross errors, $n = 100$

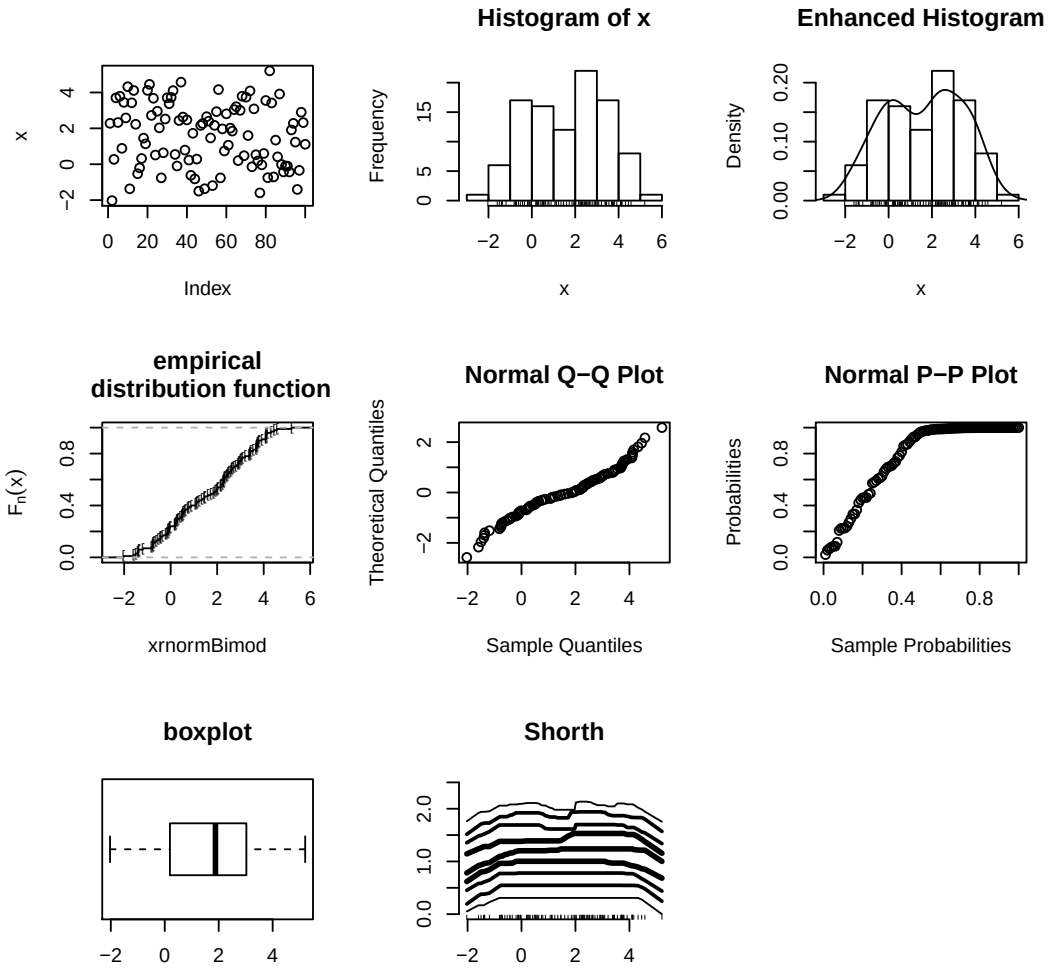


Figure 4.18: 50% mixture of two normal random variates, $n = 100$

## 4.14. Inspecting Independence and Identical Distribution

The standard assumptions in the theory of classical statistics commonly are "Let $X_1, \ldots, X_n$ be independent observations with identical distribution $F$. So far, we have concentrated on diagnostics of the distribution $F$. Let us turn to the other aspects.

Identical distribution seems to be the hard problem. The hypothetical situation becomes treatable if we have some idea about the distribution. But the alternatives are overwhelming. There are some common situations, like a trend or like periodic behaviour. For these, a remedy is to introduce a sequence indicator, are a time or some substitute of it an expand the one dimensional problem to a two-dimensional one. For sequential aspects inspecting successive differences should be a standard in data screening. Alternatively, you can check the autocorrelation of shifted data sequences. Periodic behaviour is difficult if we do not know the period, but in general it shows by runs or patterns in the difference. Spectral analysis (see 4.14.6) can be used.

A general purpose strategy is to divide the data into blocks, and to check for homogeneity between blocks. Taking disjoint blocks may help to keep block information independent, but it fragments data size. Using sliding blocks needs more intricate analysis, but the gain in information may pay.

Theoretically, independence is the easy question. Independence means that the joint distribution of $X_1, \ldots, X_n$ is the product of the marginal distributions of the $X_i$. This is actually used when inspecting pseudo random number generators, that is rules how to generate sequences to be used in simulations instead of random numbers. The fundamental case is the task to generate uniform random numbers. For inspection up to dimension $d$, you generate $d$ tuples and use the ***numeric discrepancy***, that is the supremum difference between the simulated frequency and the $d$ dimensional model distribution (uniform, in the fundamental task). Histogram and distribution function can be immediately generalized to be used here.

In a real world situation (as opposed to theory and simulation) this has its limits. Of course you can break down your sample to $d$ tuples. But then your original problem with $n$ data points in dimension 1 becomes a problem with $[n/d]$ tuples in dimension $d$. So an analysis for small $d$ can be done and should be done as a routine. But it breaks down if you cannot handle dimension $d$, or if $[n/d]$ becomes too small.

A view to theory reveals a dilemma. Taking disjoint tuples puts us on the clean side, because for independent observations we have independent tuples. But it reduces the effective sample size. We can work with overlapping tuples, sliding the start point. This does not reduce the sample size drastically, but it leaves us with an inherent dependency structure in our data. In general, we prefer sliding windows, taking into account that we have to keep in mind the dependency.

We find more tools looking back to the beginning of this chapter. The statistics introduced there are made for inspecting the distribution for independent identically distributed samples. Although they are made to measure deviation from the distribution, they have information on the other aspects as well. Take the histogram with the associated $\chi^2$, for example. If only the distribution is of concern, we look at the $\chi^2$ distance between observed and expected frequency, and large distances are critical. We would perform a one sided test, rejecting by high values. However, a common mistake in fraud is to overfit the target: a dependency is introduced in the data to better fit the target distribution. However, under model assumptions we have asymptotically a $\chi^2$ distribution. Using the test in two-sided form we will be able to detect fraudal overfitting, and poor data fit. So in this form, the test is sensitive for both aspects as alternative. This applies to most goodness of fit tests.

The hypothesis in this area is very large, and the alternatives are overwhelming. Most proposals are motivated either by historic experiences/examples, or by alternative models in mind. As far as models are concerned, two standard models should be kept in mind gain beyond independence: the regression model, that is there is an explanatory variable that may "explain" a dependency. And the Markov model: in general form, you do not have independence as is, but conditional independence. The simple Markov model is illustrated by the idea that the distribution at time $t + 1$ is independent of the time paste, given (conditional) on the state at time $t$. That is, if you know $X_t$, you can forget all other information in the past if you are concerned with $X_{t+1}$.

In the context of testing pseudo random number generators, a battery of tests has been developed, culminating in George Marsaglia's series of DIEHARD tests (**Marsaglia, 1996**). The US National Institute of Standards and Technology (NIST) has developed a list of tests that largely overlap with DIEHARD (**Rukhin, 2000**).

Details for the following tests are in (**Marsaglia, 1996**) or (**Rukhin, 2000**).

### 4.14.1. digit and serial frequency test.

### 4.14.2. monobit tests.

### 4.14.3. poker tests.

### 4.14.4. runs tests. (Bradley and Laboratory, 1960)

### 4.14.5. gap and return tests.

**4.14.6. autocorrelation and spectral tests.** (**Rukhin, 2000**) Sec 2.6 Discrete Fourier Transform (Spectral) Test `http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf`.

$Source: /u/math/j40/cvsroot/lectures/src/dataanalysis/onedim/chonedim.tex,v $
$Revision: 1.15 $
$Date: 2013/11/30 15:59:05 $
$Name:  $
$Author: j40 $