Chapter 2
1D Analysis: Summarization and Visualization of a Single Feature

2.1 Quantitative Feature: Distribution and Histogram

1D data is a set of entities represented by one feature, categorical or quantitative. There is no simple criterion to tell a quantitative feature or categorical one. For practical purposes a good criterion is this: a feature is quantitative if averaging it makes sense. Let us first consider the quantitative case.

P2.1.1 Presentation

A most comprehensive, and quite impressive for the eye, way of summarization is the distribution. On the plane, one draws an x axis and the feature range boundaries, that is, its minimum and maximum. The range interval is divided then into a number of non-overlapping equal-sized sub-intervals, *bins*. Then the number of entities that fall in each bin is counted, and the counts are reflected in the heights of the bars over the bins, forming a histogram. Histograms of Population resident in Market town dataset and Petal width in Iris dataset are presented on Fig. 2.1.

Q.2.1. Why the bins are not to overlap? **A.** Each entity falls in only one bin if bins do not overlap, and the total of all bin counts equals the total number of entities in this case. If bins do overlap, the principle "one entity – one vote" will be broken.

Q.2.2. Why the bar heights on the left are greater than those on the right in Fig. 2.1? **A.** Because bins on the right are as twice shorter than those on the left; therefore, the numbers of entities falling within them must be smaller.

Q.2.3. Is it true that when there are only two bins, the divider between them must be the midrange point? **A.** Yes, because the bin sizes are equal to each other (see Fig. 2.2).

On Figs. 2.3 and 2.4, two most popular types of histograms are presented. The former corresponds to the so-called power law, sometimes referred to as Pareto distribution. This type is frequent in social systems. According to numerous empirical studies, such features as wealth, group size, productivity and the like are all distributed according to a power law so that very few individuals or entities have

B. Mirkin, *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, Undergraduate Topics in Computer Science,
DOI 10.1007/978-0-85729-287-2_2, © Springer-Verlag London Limited 2011

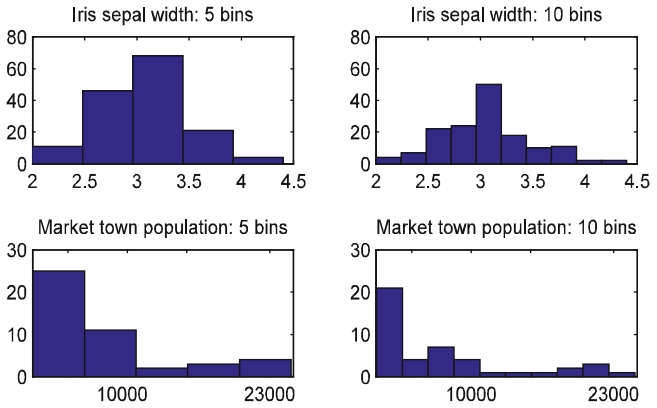

Fig. 2.1 Histograms of quantitative features in Iris and Market town data: the feature represented on *x*-axis and the counts on *y*-axis. The histogram shapes depend on the number of bins

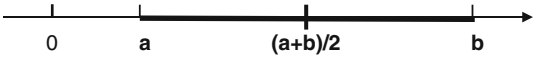

Fig. 2.2 With just two bins on the range, the divider is mid-range

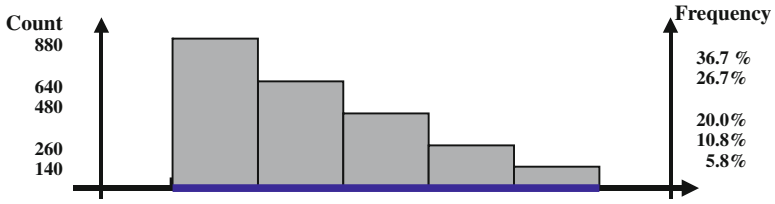

Fig. 2.3 A power type distribution

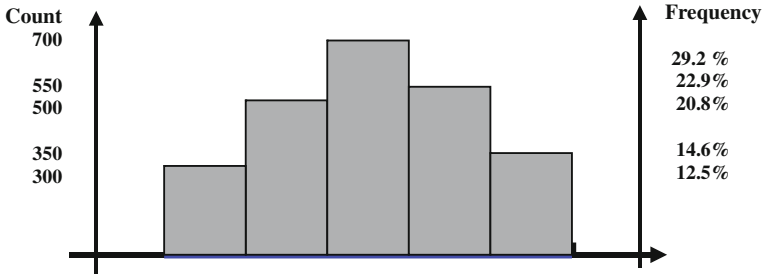

Fig. 2.4 Gaussian type distribution (bell curve)

huge amounts of wealth or members, whereas very many individuals are left with virtually nothing. However, they all are important parts of the same system with the have-nots creating the environment in which the lucky few can strive.

Another type, which is frequent at physical systems, is presented on Fig. 2.4.

This type of histograms approximates the so-called normal, or Gaussian, law. Distributions of measurement errors and, in general, features being results of small random effects are thought to be Gaussian, which can be formally proven within a mathematical framework of the probability theory.

Q.2.4. Take a look at the distributions on Fig. 2.1. Can you see which of the two types they are similar to? **A.** The Population's distribution is of power law type, and the Petal width is of Gaussian law type, as one would expect.

Another popular visualization of distributions is known as a pie-chart, in which the bin counts are expressed by the sizes of sectored slices of a round pie (see in the middle of Fig. 2.5).

As one can see, histograms and pie-charts cater for perception of two different aspects of the distribution; the former for the actual envelopment of the distribution along the axis x, whereas the latter caters for the relative sizes of distribution chunks falling into different bins. There are a dozen more formats of visualization of distributions, such as bubble, doughnut and radar charts, easily available in Microsoft Excel spreadsheet.

F2.1.2 Formulation

With N entities numbered from $i = 1, 2, \ldots, N$, data is a set of numbers $x_1, \ldots, x_N$. This set will be usually denoted by $X = \{x_1, \ldots, x_N\}$.

To produce n bins, one needs $n-1$ dividers at points $a + k(b-a)/n$ ($k = 1, 2, \ldots, n-1$). In fact, the same formula works for $k = 0$ and $k = n+1$ leading to the boundaries a as x_0 and b as x_{n+1}, which is useful for the operation of counting the number of entities N_k falling in each of the bins $k = 1, 2, \ldots, n$. Note that bin k has $a + (k-1)(b-a)/n$ and $a + k(b-a)/n$ as, respectively, its left and right boundary. One of them should be excluded from the bin so that the bins are not

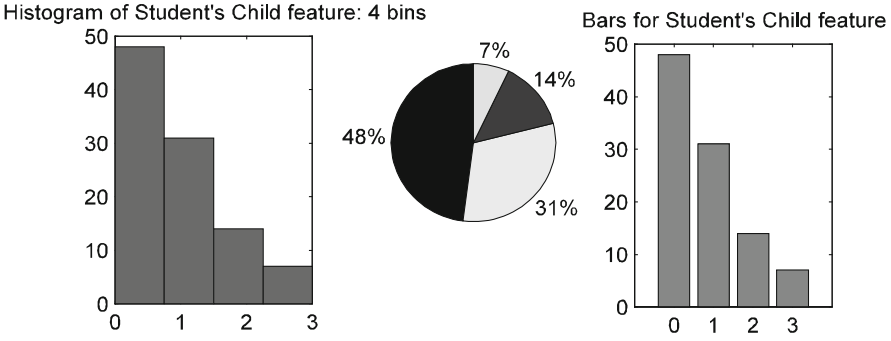

Fig. 2.5 Distribution of the number of children at Student data Child feature visualized as a 4-bin histogram on the *left*, pie-chart in the *middle*, and a bar set on the *right* – this seems the most appropriate of the three at the case

overlapping even on boundaries. These counts, N_k, $k = 1, 2, \ldots, n$ constitute the *distribution* of the feature. A *histogram* is a visual representation of the distribution by drawing a bar of the height N_k over each bin k, $k = 1, 2, \ldots, n$ (see Figs. 2.1, 2.3, 2.4 and 2.5). Note that the distribution is subject to the choice of the number of bins.

The histograms can be thought of as empirical expressions of theoretical probability distributions, the so-called density functions. A density function *p(x)* expresses the concept of probability, not straightforwardly with *p(x)* values, but in terms of their integrals, that is, the areas between the *p(x)* curve and *x*-axis, over intervals [*a,b*]: such an integral equals the probability that a random variable, distributed according to *p(x)*, falls within [*a,b*]. This implies that the total area between the curve an *x*-axis must be equal to 1, which is achieved with the corresponding scaling the curve with a constant factor.

The power law density function is $p(x) \approx a/x^{\lambda}$.where λ reflects the steepness of the frequency's fall (see Fig. 2.6 on the left). Such a law expresses what is called the Matthew's effect referring to the saying "He who has much, will get more; and he who has nothing, will lose even that little that he has," according to Matthew's gospel. The Matthew's effect is expressed, for example, in "the mechanism of preferential attachment": the probability that a new web surfer hits a web-site is proportional to the site's popularity, according to this mechanism.

The normal, or Gaussian, law is $p(x) = C\exp[-(x-a)^2/2\sigma^2]$, where C is a constant, which is sometimes denoted as $N(a, \sigma)$ (see Fig. 2.6 on the right). Distributions of measurement errors and, in general, features being results of small random effects are thought to be Gaussian, which can be formally proven within a mathematical framework of the probability theory. The parameters of this distribution, a and σ, have natural meaning: a expresses the expectation, or mean, and σ^2 – the variance, which naturally translates in data terms in Section 2.2. It should be pointed out that the probability of a value x falling in the interval $a \pm \sigma$ according to the normal distribution is about 88%, and falling in the interval $a \pm 3\sigma$ about 99.7%, virtually unity, so that at modest sample sizes it is highly unlikely that a value x can fall out of this interval, which is referred sometimes as "three sigma rule". The Gaussian distribution can be rescaled to the standard $N(0,1)$ form, with 0 expectation and 1 the variance, by shifting the variable x to the mean, a, and

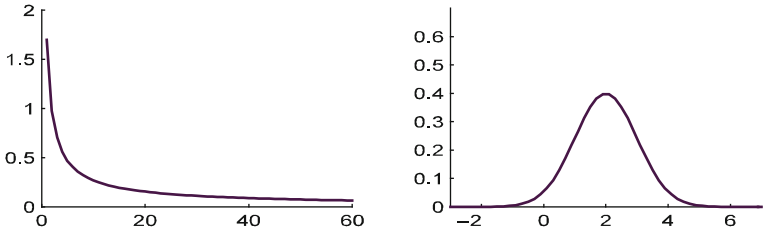

Fig. 2.6 Density functions of the power law with $\lambda = -0.8$, on the *left*, and normal distribution *N(2,1)*, on the *right*

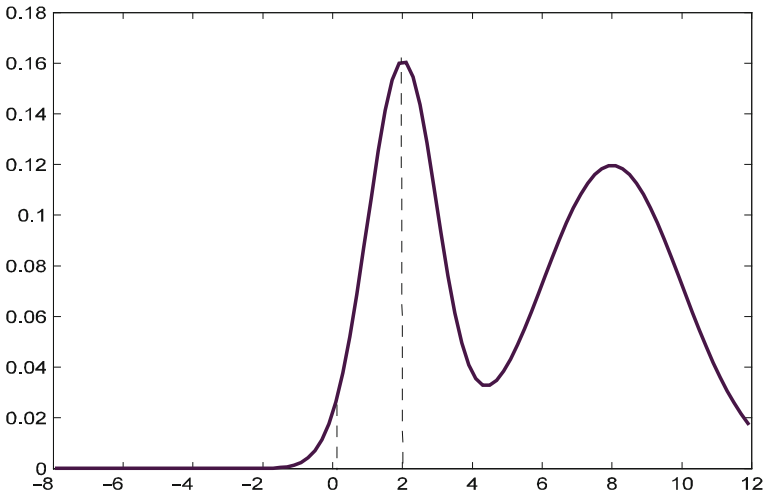

Fig. 2.7 A density function *p(x)*, which is a mixture of two normal distributions, *N(2,1)* weighted 0.4, and *N(8,2)* weighted 0.6. The area between the two *dashed lines* is the probability for value x to fall in the interval between 0 and 2 – not too high at this *p(x)*!

normalizing it afterwards by the square root of σ^2. This transformation, sometimes referred to as z-scoring, is expressed with formula $y = (x-a)/\sigma$, where y is the transformed feature. Sometimes a distribution can be approximated by a mixture of normal distributions (see Fig. 2.7).

One more popular distribution is the uniform distribution, over a range *[l,r]*. Its density is a constant function equal to $p(x) = 1/(r-l)$, so that the probability of an interval *(a,b)* within the range is just $p = (b-a)/(r-l)$, proportional to the length of the interval.

C2.1.3 Computation

To compute the distributions on Fig. 2.1, one should first load the Iris and Market town data sets with a MatLab command such as

≫ st=load('Data\town.dat');
% the Market data is stored at subfolder "Data" under the name "town.dat"
after which the Population feature can be put in a different variable
≫ pop=st(:,1);
% meaning all the rows of column 1 corresponding to the Population feature.
Then command
≫ h=hist(pop,5);

will produce a 5×1 array h containing counts of entities within each of 5 bins, and command

≫ hist(pop, 5);

will produce a figure of the histogram.

To create a figure with four windows such as on Fig. 2.1, one should use subplot commands, along with corresponding rearrangements of the axes:

≫ subplot(2,2,1);hist(sw,5);axis([2 4.5 0 80]);
≫ subplot(2,2,2);hist(sw,10);axis([2 4.5 0 80]);
% assuming sw denotes Sepal width, column 2 of the Iris data set
≫ subplot(2,2,3);hist(pop,5);axis([2000 24000 0 30]);
≫ subplot(2,2,4);hist(pop,10);axis([2000 24000 0 30]);

The command axis([a b c d]) puts the image coordinate box so that it has the interval [a,b] over x-axis and interval [c,d] over y-axis.

Bar- and pie-charts are produced with `pie` and `bar` commands, respectively.

2.2 Further Summarization: Centers and Spreads

P2.2.1 Centers and Spreads: Presentation

Further summarization of the data leads to presenting a feature with just two numbers, one expressing the distribution's location, its "central" or other important point, and the other representing the distribution's dispersion, the spread. We review some most popular characteristics for both, the center, Table 2.1, and the spread, Table 2.2, see also Lohninger (1999).

Worked example 2.1. Mean

For set $X = \{1, 1, 5, 3, 4, 1, 2\}$, mean is $c = (1 + 1 + 5 + 3 + 4 + 1 + 2)/7 = 17/7 = 2.42857\ldots$, or rounded up to two decimals, $c = 2.43$.

This is as close an approximation to the numbers as one can get, which is good. A less satisfactory property is that the mean is not stable against outliers. For example, if X in Worked example 2.1 is supplemented with value 23, the mean becomes $c = (17+23)/8 = 5$, a much greater number. This is why it is a good idea to remove some observations on both extremes of the data range, both the minimum and maximum, before computing the mean, which is utilized in the concept of trimmed mean in statistics.

Worked example 2.2. Median

To compute the median of the set from the previous example, $X = \{1, 1, 5, 3, 4, 1, 2\}$, it must be sorted first: 1, 1, 1, 2, 3, 4, 5. The median is defined as the element in the middle, which is 2. This is rather far away from the mean, 2.43, which witnesses that the distribution is biased towards the left end, the smaller entities. With the outlier 23 added, the sorted set becomes 1, 1, 1, 2, 3, 4, 5, 23, thus leading to two

Table 2.1 A review of location or central point concepts

#	Name	Explanation	Comments
1	Mean	The feature's arithmetic average	0. Minimizes the summary error squared 1. Estimates the distribution's expected value 2. Sensitive to outliers and distribution's shape
2	Median	The middle of the sorted list of feature values	1. Minimizes the summary absolute error 2. Estimates the distribution's expected value 3. Not-sensitive to outliers 4. Sensitive to distribution's shape
3	Mid-range	Middle of the range	1. Minimizes the maximum absolute error 2. Estimates the distribution's expected value 3. Very sensitive to outliers 4. Not sensitive to distribution's shape
4	P-quantile	A value dividing the entire entity set in proportion P or (1−P) of feature values so that those with higher values constitute P proportion (upper P-quantile) or 1−P proportion (bottom P-quantile)	1. Not-sensitive to outliers 2. Sensitive to distribution's shape
5	Mode	A maximum of the histogram	1. Depends on the bin size 2. Can be multiple

Table 2.2 A review of spread concepts

#	Name	Explanation	Comments
1	Standard deviation	The quadratic average deviation from the mean	1. Minimized by the mean 2. Estimates the square root of the variance
2	Absolute deviation	The average absolute deviation from the median	Minimized by the median
3	Half-range	The maximum deviation from the midrange	Minimized by the mid-range

elements in the middle, 2 and 3. The median in this case is the average of the two, (2+3)/2 = 2.5, which is by far lesser change than the mean of the extended set, 5.

The more symmetric a distribution, the closer its mean and median to each other. Sepal width of Iris data set (Table 1.3) has mean = 3.05 and median = 3, quite close values. In contrast, in Market town data (Table 1.4), Population resident's median, 5,258, is predictably much less than the mean, 7,351.4. The mean of a power law distribution is always biased towards the great values achieved by the few outliers;

this is why it is a good idea to use the median as its central value. The median is very stable against outliers: the values on the extremes just do not affect the middle of the sorted set if added uniformly to both sides.

The midrange corresponds to the mean of a flat distribution, in which all bins are equally likely. In contrast to the mean and median, the midrange depends only on the range, not on the distribution. It is obviously highly sensitive to outliers, that is, changes of the maximum and/or minimum values of the sample.

The concept of p-quantile is an extension of the concept of median, which is a 50% quantile.

Worked example 2.3. P-quantile (percentile)

Take p = 10% and determine the upper 10% quantile of Population resident feature. This should be 5th value in its descending order, that is, 18,966. Why is the 5th value? Because 10% of the total number of entities, 45, is 4.5; therefore, the 5-th value leaves out p = 10% of the largest towns in the sample. Similarly, the lower 10% quantile of the feature is 5th value in its ascending order, 2,230.

Worked example 2.4. Mode

According to the histograms in the bottom of Fig. 2.1, it is the very first bin which is modal in the Population resident distribution. In the 5-bin setting, it takes one fifth of the feature range, 23,801–2,040 = 21,761, that is, 4,352. In the 10-bin setting, it is one tenth of the feature range, that is, 2,176. In the latter case, the modal bin is interval [2,040, 4,216], and the modal bin is as twice wider, [2,040, 6,392], in the former case.

Each of the characteristics of spread in Table 2.2 parallels, to an extent, a location characteristic under the same number.

These measures intend to give an estimate of the extent of error in the corresponding centrality index. The standard deviation is the average quadratic error of the mean. Its use is related to the least-squares approach that currently prevails in data analysis and can be justified by good properties of the solutions, within the data analysis perspective, and properties of the normal distribution, within the probabilistic perspective. These paradigms are explained later in Section 2.2.2.

The absolute deviation expresses the average absolute deviation from the median. Usually, it is calculated regarding the mean, as the average error in representing the feature values by the mean. However, it is more related to the median, because it is the median that minimizes it.

The half-range expresses the maximum deviation from the mid-range; so they should be used on par, as it is done customarily by the research community involved in building classifying rules.

F2.2.2 Centers and Spreads: Formulation

There are two perspectives on data summarization and correlation that very much differ from each other. One, of the classical mathematical statistics, views the data as generated by a probabilistic mechanism and uses the data to recover the mechanism or, at least, some properties of it. The other, of data analysis, does not much care of the mechanism and tries to look for patterns in the data instead.

F2.2.2.1 Data Analysis Perspective

Given a series $X = \{x_1, \ldots, x_N\}$, one defines the centre of X as a minimizing the average distance

$$D(X,a) = [d(x_1,a) + d(x_2,a) + \ldots + d(x_N,a)]\,/N \tag{2.1}$$

Depending on the definition of the distance, the optimal a can be expressed as follows.

Consider first the least-squares formulation. According to this approach the distance is measured as the squared difference, $d(x,a) = |x-a|^2$. The minimum distance (2.1) then is reached at a equal to the mean c defined by expression

$$c = \sum_{i=1}^{N} x_i/N \tag{2.2}$$

and distance *D(X,c)* itself is equal to the variance s^2 defined by expression

$$s^2 = \sum_{i=1}^{N} (x_i-c)^2/N \tag{2.3}$$

At the more traditional distance measure $d(x,a) = |x-a|$ in (2.1), the optimal a (center) is but the median, m, and *D(X,a)* the absolute deviation from the median,

$$ms = \sum_{i=1}^{N} |x_i-m|/N \tag{2.4}$$

To be more precise, the optimal a in this problem is median, that is the value $x_{(N+1)/2}$ in the sorted order of X, when N is odd. When N is even, any value between $x_{N/2}$ and $x_{N/2+1}$ in the sorted order of X is a solution, including the median.

If *D(X,a)* is defined not by the sum, but by the maximum of the distances, $D(X,a) = \max(d(x_1,a), d(x_2,a), \ldots, d(x_N,a))$, then the midrange mr is the solution, for *d(x,a)* specified as both $|x-a|^2$ and $|x-a|$.

These statements explain the parallels between the centers and corresponding spread evaluations reflected in Tables 2.1 and 2.2, with each of the centers minimizing its corresponding measure of spread.

The distance minimization problem can be reformulated in the data recovery perspective. In the data recovery perspective, the observed values are assumed to be but noisy realizations of an unknown value a. This is reflected in the form of an equation expressing x_i through a:

$$x_i = a + e_i, \text{ for all } i = 1, 2, \ldots, N, \tag{2.5}$$

in which e_i are additive errors, or residuals, that are to be minimized.

One cannot minimize all the residuals in (2.5) simultaneously. An integral criterion is needed to embrace them all. A general family of such criteria is known as Minkowski's criterion or L_p norm. It is specified by using a positive number p as

$$L_p = \left(|e_1|^p + |e_2|^p + \ldots + |e_N|^p\right)^{1/p}$$

At a given p, minimizing L_p or, equivalently, its p-th power $L_p{}^p$, would lead to a specific solution. Most popular are values $p = 1$, 2, and ∞ (infinity) leading to:

(1) Least-squares criterion $L_2^2 = e_1^2 + e_2^2 + \ldots + e_N^2$ at $p = 2$.
Its minimization over unknown a is equivalent to the task of minimizing the average squared distance, thus leading to the mean as the optimal a.
(2) Least-modules criterion $L_1 = |e_1| + |e_2| + \ldots + |e_N|$ at $p = 1$.
Its minimization over unknown a is equivalent to the task of minimizing the average absolute deviation, thus leading to the median, optimal $a = m$.
(3) Least-maximum criterion $L_\infty = \max(|e_1|, |e_2|, \ldots |e_N|)$ at p tending to ∞.
Minimization of L_∞ with respect to a is equivalent to the task of minimizing the maximum deviation leading to the midrange, optimal $a = mr$.

The Minkowski's criteria (1)–(3) may look just as trivial reformulations of the distance approximation criterion (2.1). This, however, is not exactly so. The Equation (2.5) adds to the solution one more equation. It allows for a decomposition of the data scatter involving the corresponding data recovery criterion.

This is rather straightforward for the least-squares criterion L_2^2 whose minimal value, at a equal to the mean c (2.1) is $L_2^2 = (x_1 - c)^2 + (x_2 - c)^2 + \ldots + (x_N - c)^2$. With little algebra, this becomes $L_2^2 = x_1^2 + x_2^2 + \ldots + x_N^2 - 2c(x_1 + x_2 + \ldots + x_N) + Nc^2 = x_1^2 + x_2^2 + \ldots + x_N^2 - Nc^2 = T(X) - Nc^2$.where $T(X)$ is the quadratic data scatter defined as $T(X) = x_1^2 + x_2^2 + \ldots + x_N^2$.

This leads to equation $T(X) = Nc^2 + L_2^2$ decomposing the data scatter in two parts: that explained by the model (2.5), Nc^2, and that unexplained, L_2^2. Since the data scatter is constant, minimizing L_2^2 is equivalent to maximizing Nc^2. The decomposition of the data scatter allows measuring the adequacy of model (2.5) not by just the averaged square criterion, the variance, by the relative value of the explained

part $L_2^2/T(X)$. A similar decomposition can be derived for the least modules L_1 (see Mirkin 1996).

Q.2.5. Consider a multiplicative model for the error, $x_i = a(1+e_i)$, assuming that errors are proportional to the values. What center a fits the data with the least-squares criterion? **A.** According to the least squares approach, the fit should minimize the summary errors squared. Every error can be expressed, from the model, as $e_i = x_i/a-1 = (x_i-a)/a$. Thus the criterion can be expressed as $L_2^2 = e_1^2+e_i^2+\ldots+e_N^2 = (x_1/a-1)^2+(x_2/a-1)^2+\ldots.(x_N/a-1)^2$. Applying the first order optimality condition, let us take the derivative of ${L_2}^2$ over a and equate it to zero. The derivative is $L_2^{2'} = -(2/a^3)\Sigma_i(x_i-a)x_i$. Assuming the optimal value of a is not zero, the first order condition can be expressed as $\Sigma_i(x_i-a)x_i = 0$, so that $a = \Sigma_i x_i^2/\Sigma_i x_i = (\Sigma_i x_i^2/N)/(\Sigma_i\ x_i/N)$. The denominator here is but the mean, c, whereas the numerator can be expressed through the variance s^2 because of equation $s^2 = \Sigma_i x_i^2/N - \Sigma_i x_i/N$ which is not difficult to prove. With little algebraic manipulation, the least-squares fit can be expressed as $a = s^2/c+1$. The variance to mean ratio s^2/c, equal to $a-1$ according to the model, emerges also in statistics as a good relative estimate of the spread.

It seems rather natural that both, the standard deviation and absolute deviation, are not greater than half the range, which can be proven mathematically (see Section 2.3.2).

Q.2.6. Prove that Minkowski's center is not sensitive with respect to changing the scale factor.

Q.2.7. Prove that Minkovski's center grows whenever power p grows.

Q.2.8. For the Population resident feature in Market town data compute Minkowski center at p = 0.5, 1, 2, 3, 4, 5. **A.** See solutions found using the cm.m code developed in Project 1.1 (and confirmed, at p > 1, with the steepest descent AG-MC method) in Table 2.3.

Table 2.3 Minkowski's metric centers of the Population resident in Market town dataset for different power values p

p	Minkowski's center	Data scatter unexplained
0.5	2,611	0.7143
1	5,258.0 (median)	0.6173
2	7,351.4 (mean)	0.4097
3	8,894.9	0.2318
4	9,758.8	0.1186
5	10,294.5	0.0584

F2.2.2.2 Probabilistic Statistics Perspective

In classical mathematical statistics, a set of numbers $X = \{x_1, x_2, \ldots, x_N\}$ is usually considered a random sample from a population defined by probabilistic distribution with density *f(x)*, in which each element x_i is sampled independently from the others. This involves an assumption that each observation x_i is modeled by the distribution $f(x_i)$ so that the mean's model is the average of distributions $f(x_i)$. The population analogues to the mean and variance are defined over function *f(x)* so that the mean, median and the midrange are unbiased estimates of the population mean. Moreover, the variance of the mean is N times less than the population variance, so that the standard deviation tends to decrease by N when N grows.

Let us further assume that the population's probabilistic distribution is Gaussian $N(\mu, \sigma)$ with density function

$$f(u) = Ce^{-(u-\mu)^2/2\sigma^2}, \tag{2.6}$$

where C stands for a constant term equal to $C = (2\pi\sigma^2)^{-1/2}$. Then c in (2.2) is an estimate of μ, and s in (2.3), of σ in (2.6). These parameters amount to the population analogues of the mean and variance defined, for any density function *f(u)*, as $\mu = \int uf(u)du$ and $\sigma^2 = \int (u-\mu)^2 f(u)du$ where the integral is taken over the entire u axis.

Consider now that the set X is a random independent sample from a population with a Gaussian, for the sake of simplicity, probabilistic density function $f(x) = C\exp\left\{-(x-\mu)^2/2\sigma^2\right\}$ (2.6) where μ and σ^2 are unknown parameters. The likelihood of randomly obtaining x_i then will be $C\exp\left\{-(x_i-\mu)^2/2\sigma^2\right\}$. The likelihood of the entire sample X will be the product of these values, because of the independence assumption. Therefore, the likelihood of the sample is $L(X) = \Pi_{i\in I} C\exp\left\{-(x_i-\mu)^2/2\sigma^2\right\} = C^N \exp\left\{-\Sigma_{i\in I}(x_i-\mu)^2/2\sigma^2\right\}$. One may even go further and express $L(X)$ as $L(X) = \exp\left\{N\ln(C) - \Sigma_{i\in I}(x_i-\mu)^2/2\sigma^2\right\}$ where *ln* is the natural logarithm (over base *e*). A well established approach in mathematical statistics, the principle of maximum likelihood, claims that the values of μ and σ^2 best fitting the data X are those at which the likelihood $L(X)$ or, equivalently its logarithm, $\ln(L(X))$, reaches its maximum. The maximum $\ln(L) = N\ln(C) - \Sigma_{i\in I}(x_i-\mu)^2/2\sigma^2$ is reached at μ minimizing the expression in the exponent, $E = \Sigma_{i\in I}(x_i-\mu)^2$, which is in fact the summary quadratic distance (2.1), that is, the least- squares criterion, which thus can be derived from the assumption that the sample is randomly drawn from a Gaussian population. This, however, does not mean that the least-squares criterion is only meaningful under the normality assumption: the least-squares criterion has a meaning of its own within the data analysis paradigm.

Likewise, the optimal σ^2 minimizes part of $ln(L)$ depending on it, $g(\sigma^2) = -N\ln(\sigma^2)/2 - \Sigma_{i\in I}(xi-\mu)^2/2\sigma^2$. It is not difficult to find the optimal σ^2 from the first-order optimality condition for $g(\sigma^2)$. Let us take the derivative of the function over σ^2 and equate it to 0: $dg/d(\sigma^2) = -N/(2\sigma^2) + \Sigma_{i\in I}(xi-\mu)^2/2(\sigma^2)^2 = 0$. This equation leads to $\sigma^2 = \Sigma_{i\in I}(xi-\mu)^2/N$, which means that the variance s^2 is the maximum likelihood estimate of the parameter in the Gaussian distribution.

However, when μ is not known beforehand but rather found from the sample according to formula (2.2) for the mean, s^2 in (2.3) is a slightly biased estimate of σ^2 and must be corrected by taking the denominator equal to $N-1$ rather than N which is the case in many statistical packages. The intuition behind the correction is that Equation (2.2) is a relation imposed by us on the N observed values, which effectively decreases the "degree of freedom" in the observations from N to $N-1$.

In situations in which the data entities can be plausibly assumed to have been randomly and independently drawn from a Gaussian distribution, the derivation above justifies the use of the mean and variance as the only theoretically valid estimates of the data center and spread. The Gaussian distribution has been proven to approximate well situations in which there are many small independent random effects adding to each other. However, in many cases the assumption of normality is highly unrealistic, which does not necessarily lead to rejection of the concepts of the mean and variance – they still may be utilized within the general data analysis perspective.

In some real life situations, the assumption that X is an independent random sample from the same distribution seems rather adequate. However, in most real-world databases and multivariate samplings this assumption is far from being realistic.

C2.2.3 Centers and Spreads: Computation

In MatLab, there are commands to compute mean(X) and median(X), which can be done over X being a matrix, not just a vector. The result will be a row of within-column means or medians, respectively. To compute the row of mid-ranges, one can use a combined command mr = (max(X)+min(X))/2. To compute an upper p-quantile of a feature vector x, one should first sort it, in the descending order, with command sx = sort(x, 'descent'), after which the quantile is determined as sx(k) where k = ceil(p*length(x)).

The standard deviation is computed with command std(x), with N−1 in the denominator of (2.3), or std(x,1), with N in the denominator.

A stable version of the range that can be used at large N values or when outliers are expected, can be defined by utilizing the concept of quantile. Initially, a value of the proportion p, say 1 or 2% is specified. The upper (lower) p-quantile is a value xp of X such that the proportion of entities with larger (smaller) than xp values is p. The $2p$-quantile range is defined as the interval between these p-quantiles, stretched up according to the proportion of entities taken out, $(xp-px)/(1-2p)$, where xp and px are the upper and lower p-quantiles, respectively. For example, at $p = 0.05\%$ and $N = 100,000$, xp cuts off 50 largest and px, 50 smallest, values of X.

2.3 Binary and Categorical Features

P2.3.1 Presentation

A categorical feature differs from a quantitative one not just because its values are strings, not numbers – they are coded by numbers anyway to be processed. The

average of a quantitative feature is always meaningful, whereas the averaging of categories, such as Occupations – BA, IT or AN – in Student data or Sector of Economy – Retail, Utility or Industry – in Company data, makes no sense even after they are coded by numbers. The applicability of the operation of averaging is indeed a defining property of being quantitative. For example, one may claim that a feature like the number of children in Student data (see Fig. 2.5) is not quantitative because its values can only be whole numbers. Still, a statement like "the average number of children per woman is 1.85" does make sense because it can be easily made legitimate by moving to counting by hundreds: there are 185 children per every hundred women.

A feature admitting only two, either "Yes" or "No", values is conventionally considered Boolean in Computer Sciences, thus relating it to Boolean algebra with its "True" and "False" statement evaluations. We do not adhere to this strict logic approach but rather engage the numbers and arithmetic. The values not only can be coded by numerals 1, for "Yes", and 0, for "No", but also arithmetic operations on them can be meaningful too. Two-valued features will be referred to as binary ones.

The mean of a 1/0 coded binary feature is the proportion of its "Yes" values, which is rather meaningful. The other above defined central values bear much less information. The median is 1 only if the proportion of ones is 0.5 or greater; otherwise, it is 0. In a rare event when the number of entities is even and the proportion of ones is exactly one half, the median is one half too. The mode is ether 1 or 0, the same as the median.

For categorical features, there is no need to define bins: the categories themselves play the role of bins. However, their histograms typically are visualized with bars or stems, like on Fig. 2.8 that represents the distribution of categories IT, BA and AN of Occupation feature in Student data.

The distribution of the feature can be expressed in absolute numbers of entities falling in each of the categories, that is, D = (35, 34, 31), or on the relative scale, by using proportions found by dividing frequencies over their total, 35+34+31 = 100, which leads to the relative frequency distribution d = (0.35, 0.34, 0.31).

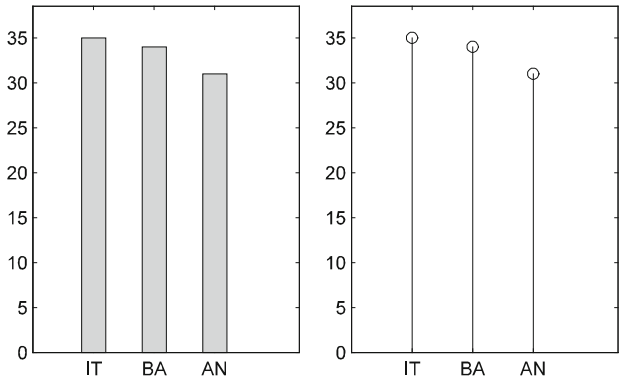

Fig. 2.8 The distribution of categories IT, BA and AN of occupation feature in Student data shown with *bars on the left* and *stems on the right*

Table 2.4 Race distribution of stop-and-search cases in England and Wales in 2005/6

Race	Number of "stop-and-searches"	Relative frequency, (%)
Black (B)	131, 723	15
Asian (A)	70, 250	8
White (W)	676, 180	77
Total	878, 153	100

This distribution is close to the uniform one in which all frequencies are equal to each other. In real life, many distributions are far from that. For example the distribution by race of the 878,153 stop-and-search cases performed by police in England and Wales was widely discussed in the media (see Table 2.4. and BBC's website http://news.bbc.co.uk/1/hi/uk/7069791.stm of 29/10/07.) This is far from uniform indeed: the proportion of W category is thrice greater than of the other two taken together. Yet it was a claim of racial bias because the proportion of W category in the population is even higher than that (for further analysis, see Section 3.4).

Q. 1.9. What is the modal category in the distribution of Table 2.4? In Occupation on Student data? **A.** These are most likely categories, W in Table 2.4 and IT in Student data.

A number of coefficients have been proposed to evaluate how much a distribution differs from the uniform distribution. The most popular are the entropy and Gini index. The latter also is referred to as the categorical variance.

The entropy measures the amount of information in signals being transferred over a communication channel. Intuitively, a rare signal bears more information than a more frequent one. Additionally, the levels of information in independent signals are to be summed up to estimate the total information. These two requirements lead to the choice of logarithm of $1/p$, that is, $-log(p)$, for scoring the level of information in a signal which appears with the probability (frequency) p. The logarithm's base is taken to be 2, because all the digital coding uses the binary number system. The entropy is defined as the averaged level of information in categories of a categorical feature. The unit of entropy has been chosen to be the bit, which is the entropy of a uniformly distributed binary feature, also referred to as a binary digit with two equally likely states. Intuitively, one bit is the level of information given in an answer to a Yes-or-No question in which no prior knowledge of the possible answer is assumed. The maximum entropy of a feature with m categories, $H = log(m)$, is reached when their distribution is uniform. The maximum Gini index, $(m-1)/m$ is reached at the uniform distribution too. Gini index measures the average level of error of the method of proportional classifier. Given a sequence of entities with unknown values of a categorical feature, the proportional classifier assigns entities with values chosen randomly, each with a probability proportional to its frequency. The average error of a category whose frequency is p is equal to $p(1-p) = p-p^2$. If, for example, $p = 20\%$, then the average error is $0.2-0.2^*0.2 = 16\%$.

Worked example 2.5. Entropy and Gini index of a distribution

Table 2.5 presents all the steps to compute the value of entropy, the summary $-plog(p)$ value, and Gini index, the summary $p(1-p)$ value where p are probabilities (relative frequencies) of categories.

Entropy is the averaged amount of information in the three categories, $H = -p_1 \log(p_1) - p_2 \log(p_2) - p_3 \log(p_3)$. The entropy in Table 2.5 relative to the maximum is $0.99/1.585 = 0.625$ because at $m = 3$ the maximum entropy is $H = \log(3) = 1.585$. Gini index is defined as the average error of the proportional prediction. The proportional prediction mechanism is defined over a stream of entities of which nothing is known beforehand except for the distribution of categories $\{p_l\}$. This mechanism predicts category l at an entity in p_l proportion of all instances. In our case, $G = p_1(1 - p_1) + p_2(1 - p_2) + p_3(1 - p_3) = 0.378$. The maximum Gini index value, $(m-1)/m$, is reached at the uniform distribution, that is, $G = 2/3$. The relative Gini index, thus, is $0.378/(2/3) = 0.567$, which is not that different from the relative entropy.

Table 2.5 Entropy and Gini index for race distribution in Table 2.4

Distribution		Entropy		Qualitative variance	
Category	Relative frequency p	Information $-log(p)$	Weighted $-plog(p)$	Error $1-p$	Variance $p(1-p)$
B	0.15	2.74	0.41	0.85	0.128
A	0.08	3.64	0.29	0.92	0.074
W	0.77	0.38	0.29	0.23	0.177
Total	1.00		0.99		0.378

F2.3.2 Formulation

A categorical feature such as Occupation in Students data or Protocol in Intrusion data, partitions the entity set in such a way that each entity falls in one and only one category. Categorical features of this type are referred to as *nominal* ones.

If a nominal feature has L categories $l = 1, \ldots, L$, its distribution is characterized by amounts $N_1, N_2, \ldots, N_L$ of entities that fall in each of the categories. Because of the partitioning property these numbers sum to the total number of entities, $N_1 + N_2 \ldots, N_L = N$. The relative frequencies, defined as $p_l = N_l/N$ sum to the unity ($l = 1, 2, \ldots, L$).

Since categories of a nominal feature are not ordered, their distributions are better visualized by pie-charts than by histograms.

The concepts of centrality, except for the mode, are not applicable to categorical feature distributions. Spread here is also not quite applicable. However, the variation – or diversity – of the distribution $(p_1, p_2, \ldots, p_L)$ can be measured. There are two rather popular indexes that evaluate dispersion of the distribution, Gini index, or qualitative variance, and entropy.

Gini index G is the average error of the proportional prediction rule. According to the proportional prediction rule, each category l, $l = 1, 2, \ldots, L$ is predicted randomly with the distribution (p_l), so that l is predicted at Np_l cases out of N. The average error of predictions of l in this case is equal to $1 - p_l$, which makes the average arror to be equal to:

$$G = \sum_{l=1}^{L} p_l(1 - p_l) = 1 - \sum_{l=1}^{L} {p_l}^2$$

Entropy averages the quantity of information in category l as measured by $\log(1/p_l) = -\log(p_l)$ over all l. The entropy is defined as

$$H = -\sum_{l=1}^{L} p_l \log p_l$$

This is not too far away from the Gini index, the qualitative variance, because at small p, $-\log(1-p)$ and $1-p$ coincide, up to a very minor difference, as is well known from calculus (see Fig. 2.9).

A very important class of nominal features consists of features with only two categories – binary features. They may emerge independently as some attributes or divisions. They also can be produced by converting categories of categorical features into binary attributes. For example, IT occupation in Student data can be converted into a question "Is it true that the student's occupation is IT?", that is, a binary feature with answers Yes and No.

These combine properties of both categorical and quantitative features. Indeed, an important difference between categorical and quantitative features is in their admissible coding sets. An admissible numerical recoding of values of a feature changes them consistently, in such a way that the relations between entities according to the feature remain intact. For example, the human heights in centimeters can be recoded in millimeters, by multiplying them by 10, or temperatures at various locations expressed in Fahrenheit can be recoded in Celsius, by subtracting 32 and dividing the result by 1.8. Such a recoding would not change the relations between locations that have been put in effect when Fahrenheit temperatures had been recorded. If, however, we assign arbitrary values to the temperatures, the new set will be inconsistent with the previous one and give a very different information. This is the borderline between quantitative and nominal features: the nominal feature can only compare if the categories are the same or not, thus admitting any

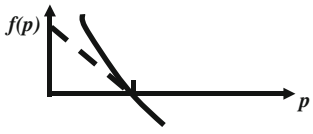

Fig. 2.9 Graphs of functions of the error $f(p) = 1-p$ involved in Gini index (*dashed line*) and the information $f(p) = -\log(p)$

one-to-one recoding as admissible, whereas the quantitative feature can only admit shifts of the origin of the scale and change of the scale factor. This borderline however is not quite hard. The binary features, as nominal ones, admit any numerical recoding. But the recoding, in this case, can always be expressed as a shift of the origin and change of the scale factor. Indeed, for any two numbers, α and β, a conversion of the feature values from 0 to α and from 1 to β can be achieved in a conventional quantitative fashion by using two rescaling parameters: the shift of the origin (α) and scaling factor (β–α).

Thus, a binary feature can be always considered as coded into a quantitative 1/0 format, 1 for Yes and 0 for No. Thus coded, a binary feature sometimes is referred to as a dummy variable.

To compute the variance of a binary feature with mean $c = p$, sum Np items $(1-p)^2$ and $N(1-p)$ items p^2, which altogether leads to $s^2 = p(1-p) = 1-p^2$. Accordingly, the standard deviation is the square root of the variance, $s = \sqrt{p(1-p)}$. Obviously, this is maximum when $p = 0.5$, that is, both binary values are equally likely. The range is always 1. The absolute deviation, in the case when $p < 0.5$ so that median $m = 0$, comprises Np items that are 1 and $N(1-p)$ items that are 0, so that $sm = p$. When $p < 0.5$, $m = 1$ and the number of unity distances is $N(1-p)$ leading to $sm = 1-p$ That means that, in general, $sm = \min(p, 1-p)$, which is less than or equal to the standard deviation. Indeed, if $p \leq 0.5$, then $p \leq 1-p$ and, thus, $p^2 \leq p(1-p)$, so that $ms \leq s$. Analogously, if $p > 0.5$ then $p > 1-p$ and, thus, $p(1-p) > (1-p)^2$, so that again $sm < s$, which proves the statement.

When a categorical feature is converted into a set of binary features corresponding to its categories, the total variance of the L binary variables is equal to the Gini index, or qualitative variance, of the original feature.

There are some probabilistic underpinnings to binary features. Two models are popular, one by Bernoulli and another by Poisson. Given p, $0 \leq p \leq 1$, Bernoulli model assumes that every x_i is either 1, with probability p, or 0, with probability $1 - p$. Poisson model suggests that, among the N binary numerals, random pN are unities, and $(1-p)N$ zeros. Both models yield the same mathematical expectation, p. However, their variances differ: the Bernoulli distribution's variance is $p(1-p)$, whereas the Poisson distribution's variance is p, which is obviously greater for all positive p, because the factor at Bernoulli standard deviation, $1 - p$, is less than 1 under this condition. Similar models can be considered for nominal features with more than two categories.

There is a rather natural, though somewhat less recognized, relation between quantitative and binary features: the variance of a quantitative feature is always smaller than that of the corresponding binary feature. To explicate this according to Mirkin (2005), assume the interval [0,1] to be the range of data $X = \{x_1, \ldots, x_N\}$. Assume that the mean c divides the interval in such a way that a proportion p of the data is greater than or equal to c, whereas proportion of those smaller than c is $1 - p$. The question then is this: given p, at what distribution of X the variance is maximized. To address the question, assume that X be any given distribution within interval [0,1] with its mean at some interior point c. According to the assumption,

there are Np observations between 0 and c. Obviously, the variance can only increase if we move each of these points to the boundary, *0*. Similarly, the variance will only increase if we push each of $N(1-p)$ points between c and 1, into the opposite boundary 1. That means that the variance $p(1-p)$ of a binary variable with Np zero and $N(1-p)$ unity values is the maximum, at any p. The following is proven. A binary variable, whose distribution is $(p, 1-p)$, has the maximum variance, and the standard deviation, among all quantitative variables of the same range and p entries below its average.

This implies that no variable over the range [0,1] has its variance greater than the maximum ¼ reached by a binary variable at $p = 0.5$. The standard deviation of this binary variable is ½, which is just half of the range. Therefore, the standard deviation of any variable cannot be greater than its half-range.

The binary variables also have the maximum absolute deviation among the variables of the same range, which can be proven similarly.

C2.3.3 Computation

If the distribution of a feature is in vector df, then a command like

≫ bar(df, .4);h=axis;axis(1.1*h);
will produce its bar drawing. The parameters here are: 0.4 the width of bars, 1.1 the rescaling to allow some air between the histogram and the border in the drawing frame (see Fig. 2.8).

Computation of the entropy and Gini index for the distribution presented in vector df can be done with commands:
≫ df=df/sum(df); h=−sum(df .*log2(df)); % h is entropy
≫ df=df/sum(df); g=−sum(df .*(1−df)); % h is Gini

Q.2.10 Take nominal features from the Intrusion data set and generate category-based binary features, after which compute their individual means and variances Compare the variances with Gini index for the original features.

2.4 Modeling Uncertainty: Intervals and Fuzzy Sets

2.4.1 Individual Membership Functions

In those cases when the probability distributions are unknown or inapplicable, intervals and fuzzy sets are used to reflect uncertainty in data. When dealing with complex systems, feature values cannot be determined precisely, even for such a relatively stable and homogeneous dimension as the population resident in a country. The so-called "linguistic variables" (Zadeh 1970) express imprecise categories and concepts in terms of appropriate quantitative measures, such as the concept of "normal temperature" of an individual – a body temperature from about 36.0 to 36.9

Celsius or "normal weight" – the Body Mass Index BMI (the ratio of the weight, in kg, to the height, in meters, squared) somewhat between 20 and 25. (Those with BMI > 25 are considered overweight or even obese if BMI > 30; and those with BMI < 20, underweight). In these examples, the natural boundaries of a category are expressed as an interval.

A more flexible description can be achieved using the concept of fuzzy set A expressed by the membership function $\mu_A(x)$ defined, on the example of Fig. 2.10, as:

$$\mu_A(x) = \begin{cases} 0 \quad if \quad x \leq 18 \quad or \quad x \geq 27 \\ 0.25x - 4.5 \quad if \quad 18 \leq x \leq 22 \\ 1 \quad if \quad 22 \leq x \leq 24 \\ -x/3+9 \quad if \quad 24 \leq x \leq 27 \end{cases}$$

This function says that the normal weight does not occur outside of the BMI interval [18, 27]. Moreover, the concept applies in full, with the membership 1, only within BMI interval [22, 24]. There are "grey" areas expressed with the slopes on the left and the right so that, say, a person with BMI = 20 will have the membership value $\mu_A(20) = 0.25^*20 - 4.5 = 0.5$ and the membership of that with BMI = 26.1, will be $\mu_A(26.1) = -26.1/3 + 9 = -8.7 + 9 = 0.3$.

In fact, a membership function may have any shape; the only requirement perhaps is that there should exist at least one point or sub-interval at which the function reaches the maximum value 1. A fuzzy set formed with straight lines, such as that on Fig. 2.10, is referred to as a trapezoidal fuzzy set. Such a set can be represented by four points on the axis x :(a, b, c, d) such that $\mu_A(x) = 0$ outside the outer interval *[a,d]* and $\mu_A(x) = 1$ inside the inner interval *[b,c]* (with the straight lines connecting points (a,0) and *(b,1)* as well as *(c,1)* and *(d,0)* (see Fig. 2.10).

Both the precise and interval values can be considered special cases of trapezoidal fuzzy sets. An interval *(a, b)* can be equivalently represented by a trapezoidal fuzzy set *(a, a, b, b)* in which all points of *(a, b)* have their membership value equal to 1, and a point a can be represented by trapezoidal fuzzy set *(a, a, a, a)*.

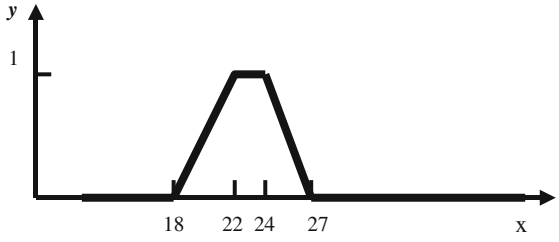

Fig. 2.10 A trapezoidal membership function expressing the concept of normal body mass index; a positive degree of membership is assigned to each point within interval [18, 27] and, moreover, those between 22 and 24 certainly belong to the set

The so-called triangular fuzzy sets are also popular. A triangular fuzzy set A is represented by an ordered triplet *(a,b,c)* so that $\mu_A(x) = 0$ outside the interval *[a,c]* and $\mu_A(x) = 1$ only at $x = b$, with values of $\mu_A(x)$ in between are represented by the straight lines between points *(a,0)* and *(b,1)* and between *(c,0)* and *(b,1)* on the Cartesian plane, see Fig. 2.11.

Fuzzy sets presented on Figs. 2.10 and 2.11 are not equal to each other: only those fuzzy sets A and B are equal at which $\mu_A(x) = \mu_B(x)$ for every x, not just outside of the base interval

A fuzzy set should not be confused with a probabilistic distribution such as a histogram: there may be no probabilistic mechanism nor frequencies behind a membership function, just an expression of the extent at which a concept is applicable. A conventional, crisp set S, can be specified as a fuzzy set whose membership function μ admits only values 0 or 1 and never those between; thus, $\mu(x) = 1$ if $x \in S$ and $\mu(x) = 0$, otherwise.

Q.2.11. Prove that triangular fuzzy sets are but a special case of trapezoidal fuzzy sets. **A.** Indeed a triangular fuzzy set *(a,b,c)* can be represented by a trapezoidal fuzzy set *(a,b,b,c)*.

There are a number of specific operations with fuzzy sets imitating those with the "crisp" sets, first of all, the set-theoretic complement, union and intersection.

The complement of a fuzzy set A is fuzzy set B such that $\mu_B(x) = 1 - \mu_A(x)$. The union of two fuzzy sets, A and B, is a fuzzy set denoted by $A \cup B$ whose membership function is defined as $\mu_{A\cup B}(x) = \max(\mu_A(x),\ \mu_B(x))$. Similarly, the intersection of two fuzzy sets, A and B, is a fuzzy set denoted by $A \cap B$ whose membership function is defined as $\mu_{A\cap B}(x) = \min(\mu_A(x),\ \mu_B(x))$.

It is easy to prove that these operations indeed are equivalent to the corresponding set theoretic operations when performed over crisp membership functions. It should be noted, though, that of all these operations only the union is always correct; the others can bring forward a fuzzy set whose maximum is less than 1.

Q.2.12. Draw the membership function of fuzzy set A on Fig. 2.11.

Q.2.13. What is the union of the fuzzy sets presented in Figs. 2.10 and 2.11.

Q.2.14. What is the intersection of the fuzzy sets presented in Figs. 2.10 and 2.11.

Q.2.15. Draw the membership function of the union of two triangular fuzzy sets represented by triplets (2,4,6), for A, and (3,5,7), for B. What is the membership function of their intersection?

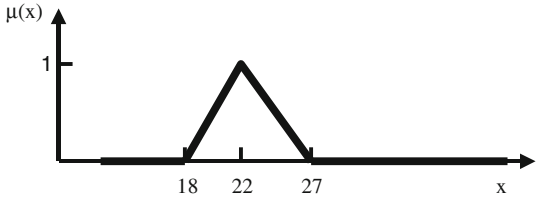

Fig. 2.11 A triangular fuzzy set for the normal weight BMI

Q.2.16. What type of a function is the membership function of the intersection of two triangular fuzzy sets? Of two trapezoidal fuzzy sets? Does it always represent a fuzzy set?

2.4.2 Central Fuzzy Set

The conventional center and spread concepts can be extended to intervals and fuzzy sets. Let us consider an extension of the concept of average to the triangular fuzzy sets using the least-squares data recovery approach.

Given a set of triangular fuzzy sets A_1, A_2, ..., A_N, the central triangular set A can be defined by such a triplet *(a, b, c)* that approximates the triplets (a_i, b_i, c_i), $i = 1, 2, \ldots, N$. The central triplet can be defined by the condition that the average difference squared,

$$L(a, b, c) = (\Sigma_i (a_i - a)^2 + \Sigma_i (b_i - b)^2 + \Sigma_i (c_i - c)^2)/(3N)$$

is minimized by it. Since the criterion L is additive over the triplet's elements, the optimal solution is analogous to that obtained in the conventional case: the optimal a is the mean of $a_1, a_2, \ldots, a_N$; and the optimal b and c are the means of b_i and c_i, respectively.

Q.2.17. Prove that the average a_i indeed minimizes L. **A.** Let us take the derivative of L over a: $\partial L/\partial a = -2\Sigma_i (a_i - a)/(3N)$. The first-order optimality condition, $\partial L/\partial a = 0$, has the average as its solution described.

Q.2.18. Explore the concepts of central trapezoidal fuzzy set and central interval in an analogous way.

Project 2.1. Computing Minkowski metric's center

Consider a series x_i, $i = 1, 2, \ldots, N$ and given a positive $p > 0$, compute such an a that minimizes the summary Minkowski criterion, p-th power of the distance,

$$Lp = |x_1 - a|^p + |x_2 - a|^p + \ldots |x_N - a|^p \tag{2.7}$$

When $p \neq 2$, no generally applicable analytic expression can be derived for the minimizer. One way to proceed would be using the mechanisms of hill-climbing, a strategy of iteratively approaching a (local) minimum point by moving step-by-step in the anti-gradient direction which is frequently referred to as the steepest descent direction (see Polyak 1987). Another way is to use a nature-inspired strategy by letting a population of admissible solutions to interatively evolve and keeping track of the "best" points visited (see Engelbrecht 2002).

We take on both approaches to minimization of *Lp*:

(i) Steepest descent iterations, and
(ii) Nature inspired iterations.

(i) Steepest descent computation MC_SD

Before we proceed to computations, let us explore the criterion *Lp* in (2.7). For the sake of simplicity, assume $p \geq 1$. Consider that the N values in X are sorted in the ascending order so that $x_1 \leq x_2 \ldots \leq x_N$. Then it is easy to prove, first, that the criterion is a convex function shaped like that presented on Fig. 2.12, and, second, the optimal a-value is indeed between the minimum, x_1, and the maximum, x_N.

Assume the opposite: that the minimum is reached outside of the interval, say at $a > x_N$. Then, obviously, $Lp(x_N) < Lp(a)$ because $|x_i - x_N| < |x_i - a|$ for every $i = 1, 2, \ldots, N$, and the same holds for the p-th powers of those. As to the convexity, let us consider any a in the interval between x_1 and x_N. Criterion (2.7) then can be rewritten as:

$$Lp(a) = \sum_{i \in I_+} (a - x_i)^p + \sum_{i \in I_-} (x_i - a)^p \tag{2.8}$$

where I_+ is set of those indices i for which $a > x_i$, and I_- is set of such i's that $a \leq x_i$. The derivative of $Lp(a)$ in (2.8) is equal to:

$$Lp'(a) = p \left(\sum_{i \in I_+} (a - x_i)^{p-1} - \sum_{i \in I_-} (x_i - a)^{p-1} \right) \tag{2.9}$$

and the second derivative, to

$$Lp''(a) = p(p-1) \left(\sum_{i \in I_+} (a - x_i)^{p-2} + \sum_{i \in I_-} (x_i - a)^{p-2} \right).$$

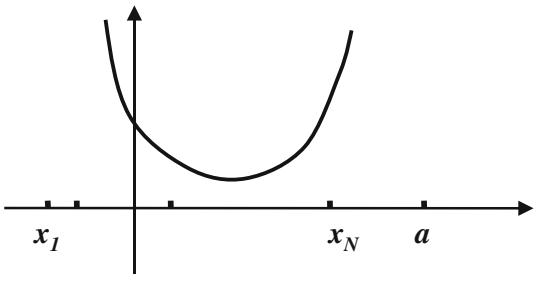

Fig. 2.12 A convex function of a

The latter expression is positive for each a value, provided that $p > 1$, which proves that $Lp(a)$ is convex. This leads to one more property: assume that $Lp(x_{i*})$ is minimum among all the $Lp(x_i)$ values ($i = 1, 2, \ldots, N$), then the minimum of $Lp(a)$ lies within interval $(x_{i'}, x_{i''})$ where $x_{i'}$ is that x_i-value, which is the nearest to x_{i*} among those on the left of it, at which $Lp(x_i) > Lp(x_{i*})$. And, similarly, $x_{i''}$ is that x_i-value, which is the nearest to x_{i*} among those to the right of it, at which $Lp(x_i) > Lp(x_{i*})$.

The properties above justify the following steepest descent algorithm applicable at $p > 1$:

MC_SD

1. Initialize with $a0 = x_{i*}$ and a positive learning rate λ.
2. Compute $a0 - \lambda Lp'(a0)$ according to formula (2.9) and take it as $a1$ if it falls within the interval $(x_{i'}, x_{i''})$. Otherwise, decrease λ a bit and repeat the step.
3. Test whether $a1$ and $a0$ coincide, up to a pre-specified precision threshold. If yes, halt the process and output $a1$ as the optimal value for a. If not, move on.
4. Test whether $Lp(a1) \leq Lp(a0)$. If yes, set $a0 = a1$ and $Lp(a0) \leq Lp(a1)$, and go to (2). If not, decrease λ a bit and go to 2 (without changing $a0$).

(ii) Nature-inspired computation MC_NI

According to the nature-inspired approach, a population of possible solutions rather than a single solution is maintained. In contrast to the classical approaches, the improvements here are a matter of a random evolution of the population from one generation to another, which is organized in such a way that improvements are likely to be acquired. Since this is a 1D search, it is likely that any random moves would approximate the optimal point soon enough. The simple algorithm MC_NI presented below works quite well in experiments:

1. *Determining the area of admissible solutions.* Determine an area A of admissible solutions – a set of points which should contain the optimum point(s).
   This is quite easy in this case: as proven above, the optimum lies between the minimum *lb* and maximum *rb* of the series $x_i, i = 1, 2, \ldots, N$. Thus, the area is interval *(lb,rb)*.
2. *Population setting.* Specify the size *pe* of the population to evolve, say, $pe = 15$, and randomly put points $s_1, s_2, \ldots, s_{pe}$ in the admissible area *(lb,rb)*.
3. *Elite initialization.* Evaluate values of the criterion, frequently referred to as the "fitting function", for each member of the population and store information of the best (elite), that is, the minimum, as the only record s_e to output when needed.
4. *Next generation.* Modify the population by, first, adding random Gaussian noise r:

$$s'_k = s_k + \lambda r$$

and, second, by moving all those of the resulting values that went out of the area of admissible solutions A back to the area.

5. *Elite maintenance.* Evaluate values of the criterion at the new generation, pick the best and worst of them, say s_b and s_w, and compare with the elite s_e. If s_b is better than s_e, change the elite for s_b. Else, that is, if s_b and, more so, s_w are worse than s_e, improve the current population by changing s_w in that for the record s_e.
6. *Stop condition.* If the number of iterations has not reached a pre-specified value, go to (4). Otherwise, output the elite solution.

Experiments show that the gradient based procedure of the steepest descent is faster than the nature-inspired one. But the latter works at any p, whereas the former only at $p > 1$.

Project 2.2. Analysis of a multimodal distribution

Let us take a look at the distributions of OOP and CI marks at the Student data. Assuming that the data file of Table 1.4 is stored as Data\stud.dat, the corresponding MatLab commands can be as follows:

```
≫ a=load('Data\stud.dat');
≫ oop=a(:,7); % column of OOP mark
≫ coi=a(:,8); % column of CI mark
≫ subplot(1,2,1); hist(oop);
≫ subplot(1,2,2); hist(coi);
```

With ten bins used in MatLab by default, the histograms are on Fig. 2.13.

The histogram on the left seems to have three humps, that is, three-modal. Typically, a homogeneous sample should have a unimodal distribution, to allow interpretation of the feature as its modal value with random deviations from it. The

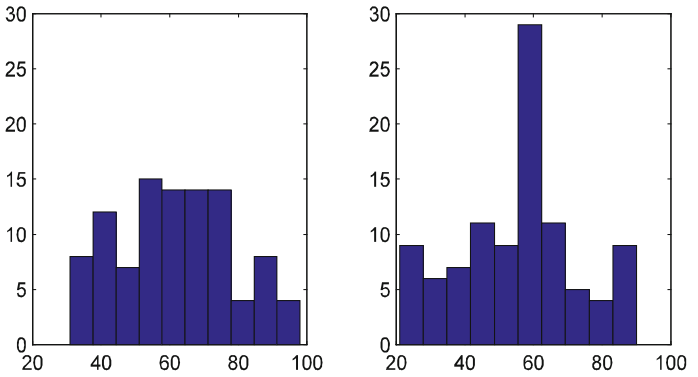

Fig. 2.13 Histograms of the distributions of marks for OOP (*on the left*) and CI (*on the right*) from students data

three modes on the OOP mark histogram require an explanation. For example, one may hypothesize that the modes can be explained by the presence of three different groups of students represented by their occupations so that IT group should have higher marks than BA group whose marks should still be higher than those at AN group.

To test this hypothesis, one needs to compare distributions of OOP marks at each of the occupations. To make the distributions comparable, we need to specify an array with boundaries between 10 bins that can be used for each of the samples. This array, *b*, can be computed as follows:

≫ r=max(oop)−min(oop);for i=1:11;b(i)=min(oop)+(i–1)*r/10;end;

Now we are ready to produce comparable distributions for each of the occupations with MatLab command histc:

≫ for ii=1:3;li=find(a(:,ii)==1);hp(:,ii)=histc(oop(li),b);end;

This generates a list, li, of student indices corresponding to each of the three occupations presented by the three binary columns. Matrix hp stores the three distributions in its three columns. Obviously, the total distribution of OOP, presented on the left of Fig. 2.13 is the sum of these three columns. To visualize the distributions, one may use "bar" command in MatLab:

≫ bar(hp);

which produces bar histograms for each of the three occupations (see Fig. 2.14). One can see that the histograms differ indeed and concur with the hypothesis, so that IT concentrates in top seven bins and, moreover, it shares the top three bins with no other occupation. The other two occupations overlap more, though AN takes over on the leftmost, worst marks, positions indeed.

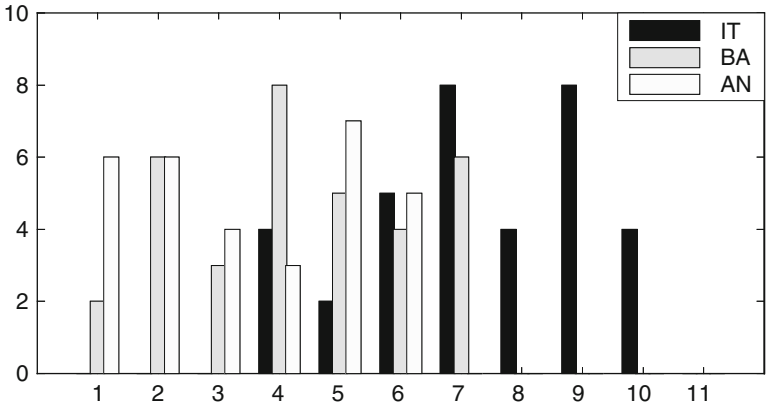

Fig. 2.14 Histograms of OOP marks for each of three occupations, IT, BA and AN, each presented with bars filled in according to the legend

Q.2.19. What would happen if array *b* is not specified once for all but the histogram is drawn by default for each of the sub-samples? **A.** The 10 default bins depend on the data range, which may be different at different sub-samples; if so, the histograms will be incomparable.

There can be other hypotheses as well, such as that the modes come from different age groups. To test that, one should define the age group boundaries first.

Project 2.3. Computational validation of the mean by bootstrapping

The data file short.dat in Appendix A5 is a 50 × 3 array whose columns are samples of three data types described in Table 2.6.

The normal data is in fact a sample from a Gaussian N(10,2), that has 10 as its mean and 2 as its standard deviation. The other two are Two-modal and Power law samples. Their histograms are on the left-hand sides of Figs. 2.15, 2.16, and 2.17. Even with the aggregate data in Table 2.6 one can see that the average of Power law does not make much sense, because its standard deviation is more than three times greater than the average.

Table 2.6 Aggregate characteristics of columns for short.dat array

Data type		Normal	Two-modal	Power law
Mean		10.27	16.92	289.74
Standard deviation	Real value	1.76	4.97	914.50
	Related to $\sqrt{N}$	0.25	0.70	129.33

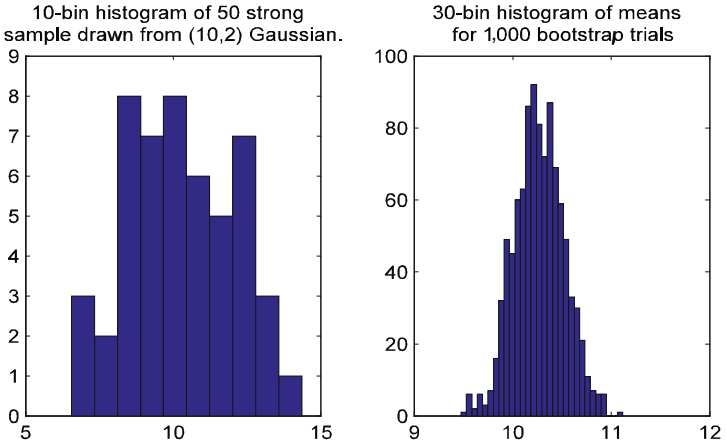

Fig. 2.15 The histograms of a 50 strong sample from a Gaussian distribution (*on the left*) and its mean's bootstrap values (*on the right*): all falling between 9.7 and 10.1

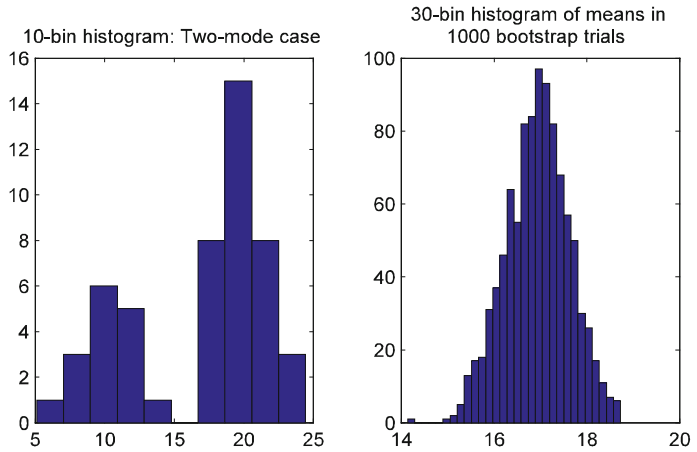

Fig. 2.16 The histograms of a 50 strong sample from a Two-mode distribution (*on the left*) and its mean's bootstrap values (*on the right*)

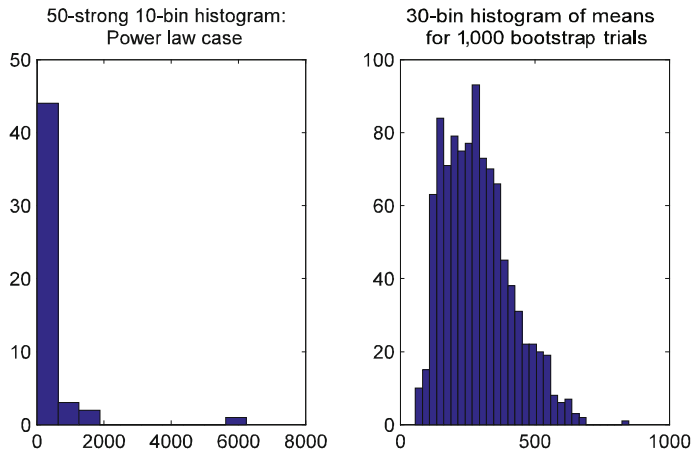

Fig. 2.17 The histograms of a 1,000 strong sample from a Power law distribution (*on the left*) and its mean's bootstrap values (*on the right*): all falling between 260 and 560

Many statisticians would argue the validity of characteristics in Table 2.6 not because of the distribution shapes – which would be a justifiable source of concern for at least two of the three distributions – but because of small sizes of the samples. Is the 50 entities available a good representation of the entire population indeed? To address these concerns, the Mathematical Statistics have worked out principles based on the assumption that the sampled entities come randomly and independently from a – possibly unknown but stationary – probabilistic distribution. The mathematical thinking would allow then, in reasonably well-defined situations, to arrive at a theoretical distribution of an aggregate index such as the mean, so that the distribution may lead to some confidence boundaries for the index.

Typically, one would obtain the boundaries of an interval at which 95% of the population falls, according to the derived distribution. For instance, when the distribution is normal, the 95% confidence interval is defined by its mean plus/minus 1.96 times the standard deviation related to the square root of the number observations, which is 7.07 at N = 50. Thus, for the first column data, the theoretically derived 95% confidence interval will be $10 \pm 1.96^*2/7.07 = 10 \pm 0.55$, that is, (9.45, 10.55) (if the true parameters of the distribution are known) or $10.27 \pm 1.96^*1.76/7.07 = 10.27 \pm 0.49$, that is, (9.78,10.76) (at the observed parameters in Table 2.6). The difference is rather minor, especially if one takes into account that the 95% confidence is a rather arbitrary notion. In probabilistic statistics, the so-called Student's distribution is used to make up for the fact that the sample-estimated standard deviation value is used instead of the exact one, but that distribution differs little from the Gaussian distribution when there are more than several hundred entities.

In many real life applications the shape of the underlying distribution is unknown and, moreover, the distribution is not necessarily stationary. The theoretically defined confidence boundaries are of little value then. This is why a question arises whether any confidence boundaries can be derived computationally by re-sampling the data at hand rather than by imposing some debatable assumptions. There have been developed several approaches to computational validation of sample based results. One of the most popular is bootstrapping which will be used here in its two basic, "pivotal" and "non-pivotal", formats as defined in Carpenter and Bithell (2000) (see also Efron and Tibshirani 1993).

Bootstrapping is based on a pre-specified number, say 1,000, of random trials. *A trial* involves randomly drawn N entities, with replacement, from the entity set. Note that N is the size of the entity set. Since the sampling goes with replacement, some entities may be drawn two or more times so that some others are bound to be left behind. Recalling that $e = 2.7182818...$ is the natural logarithm base, it is not difficult to see that, on average, only approximately $(e-1)/e = 63.2\%$ entities get selected into a trial sample. Indeed, at each random drawing from a set of N, the probability of an entity being not drawn is $1-1/N$, so that the approximate proportion of entities never selected in N draws is $(1-1/N)^N \approx 1/e = 1/2.71828 \approx 36.8\%$ of the total number of entities. For instance, in a bootstrap trial of 15 entities, the following numbers have been drawn: 8, 11, 7, 5, 3, 3, 11, 5, 9, 3, 11, 6, 13, 13, 9 so that seven entities have been left out of the trial while several multiple copies have got in.

A trial set of a thousand randomly drawn entity indices (some of them, as explained, would coincide) is assigned with the corresponding row data values from the original data table so that coinciding entities get identical rows. Then a method under consideration, currently "computing the mean", applies to this trial data to produce the trial result. After a number of trials, the user gets enough results to represent them with a histogram and derive confidence boundaries from that.

The bootstrap distributions for each of the three types of data generation mechanism, after 1,000 trials, are presented in Figs. 2.15, 2.16 and 2.17 on the right hand side.

The pivotal validation method is based on the assumption that the bootstrap distribution of means is Gaussian, so that having estimated its average m_b and standard deviation s_b, the 95% confidence interval is estimated as usual, with formula $m_b \pm 1.96^* s_b = 10.24 \pm 1.96^* 0.24 = 10.24 \pm 0.47$, which is the interval between 9.77 and 10.71 – which is very similar to that obtained under the hypothesis of Gaussian distribution – this is no wonder here because the hypothesis is true.

The non-pivotal method makes no assumption of the distribution of bootstrap means and uses the empirical bootstrap found distribution to cut it at its 2.5% upper and bottom quantiles. To do this, we can sort values of the vector of bootstrap means and find the values at its 26th and 975th components that cut out exactly 2.5% of the set each. This action produces interval between 9.78 and 10.70, which is very close to the previously found boundaries for the 95% confidence interval for the mean value of the first sample.

There is theoretical evidence, presented by E. Bradley (1993), to support the view that the bootstrap can produce somewhat tighter estimate of the marks deviation than the estimate based on the original sample. In our case, we can see in Table 2.7 that indeed, with the means almost unchanged, the standard deviations have been slightly reduced.

Unfortunately, the bootstrap results are not that helpful in analyzing the other two distributions: as can be seen in our example, both of the means, the Two-modal and Power law ones, are assigned rather decent boundaries while, in most applications, the mean of either of these two distributions may be considered meaningless. It is a matter of applying other data analysis methods such as clustering to produce more homogeneous sub-samples whose distributions would be more similar to that of a Gaussian.

The reader is requested to provide pivotal and not-pivotal estimates of 95% confidence interval for the other two samples in short.dat dataset (Two-modal and Power law).

Project 2.4. K-fold cross validation

Another set of validation techniques utilizes randomly splitting the entity set in two parts of pre-specified sizes, the so-called train and test sets, so that the method's results obtained for the train set are compared with the data on the test set. To

Table 2.7 Aggregate characteristics of the results of 1,000 bootstrap trials over short.dat array

Data type		Normal	Two-mode	Power law
Mean		10.27	16.94	287.54
Standard deviation	Original sample	0.25	0.70	129.33
	Bootsrap value	0.25	0.69	124.38
	Relative to mean, %	2.46	4.05	43.26

guarantee that each of the entities gets into a train/test set the same number of times, the so-called cross-validation methods have been developed.

The so-called K-fold cross validation works as follows. Randomly split entity set in K parts $Q(k)$, $k = 1, \ldots, K$, of equal sizes.[1] Typically, K is taken as 2 or 5 or 10. In a loop over k, each part $Q(k)$ is taken as test set while the rest forms the train set. A data analysis method under consideration is run over the train set ("training phase") with its result applied to the test set. The average score of all the test sets constitutes a K-fold cross-validation estimate of the method's quality.

The case when K is equal to the number of entities N is especially popular. It was introduced earlier under the term "jack-knife", but currently term "leave-all one-out" is more popular as better reflecting the idea of the method: N trials are run over the entire set except for just each one entity removed from the training.

Let us apply the 10-fold cross-validation method to the problem of evaluation of the means of the three data sets. First, let us create a partition of our 1,000 strong entity set in 10 non-overlapping classes, a hundred entities each, with randomly assigning entities to the partition classes. This can be done by randomly putting entities one by one in each of the 10 initially empty buckets. Or, one can take a random permutation of the entity indices and divide then the permuted series in 10 chunks, 100 strong each. For each class $Q(k)$ of the 10 classes ($k = 1, 2, \ldots, 10$), we calculate the averages of the variables on the complementary 900 strong entity set, and use these averages for calculating the quadratic deviations from them – not from the averages of class $Q(k)$ – on the class $Q(k)$. In this way, we test the averages found on the complementary training set.

The results are presented in Table 2.8. The values found at the original distribution and with a ten fold cross validation are similar. Does this mean that there is no need in applying the method? Not at all, when more complex data analysis methods are used, the results may differ indeed. Also, whereas the ten quadratic deviations calculated on the ten test sets for the Gaussian and Two-modal data are very similar to each other, those at the Power law data set drastically differ, ranging from 391.60 to 2,471.03.

Table 2.8 Quadratic deviations from the means computed on the entity set as is and by using ten fold cross validation

Data type		Normal	Two-modal	Power law
Standard deviation	On set	1.94	5.27	1744.31
	ten fold cr.-val.	1.94	5.27	1649.98

Q.2.20. What is the bin size in the example of Fig. 2.18? **A.** 2.

[1] To do this, one may start from all sets $Q(k)$ being empty and repeatedly run a loop over $k = 1 : K$ in such a way that at each step, a random entity is drawn from the entity set (with no replacement!) and put into the current $Q(k)$; the process halts when no entities remain out of $Q(k)$.

Fig. 2.18 Range [2,12] divided in five bins

Q.2.21. Consider feature x whose range is between 1 and 10. When the range of x is divided in 9 bins (in this case, intervals of the lengths one: [1,2), [2,3), . . ., [9,10]), the x frequencies in the corresponding bins are: 10, 20, 10, 20, 30, 20, 40, 20, 30. Please answer these questions:

(i) How many observations of x are available?
(ii) What can be said about the value of the median of x?
(iii) Provide the minimum and maximum estimates of the average of x.
(iv) What can be said of 20% quantiles of x?
(v) What is the distribution of x when the number of bins is 3? What is the qualitative variance (Gini coefficient) for this distribution?

A.

(i) There are 200 observations.
(ii) The median lies between 100-th and 101-th values in a sorted order, that is, in the 6-th bin, that is, between 6 and 7.
(iii) The minimum estimate of the mean is computed with the minimal values in bins:
$(1*10+2*20+3*10+4*20+5*30+6*20+7*40+8*20+9*30)/200 = 5.7$
The maximum estimate is calculated using the same formula with all bin values increased by 1, which should lead to 5.7+1 = 6.7.
(iv) 20% of 200 is 40. That means that the 20% quantile on the left end of x is 4, while that on the right end must be in the 8-th bin, that is, between 8 and 9.
(v) The three-bin distribution will be 40, 70, 90 or, in the relative frequencies, 0.2, 0.35, 0.45, which leads to the Gini index equal to $1-0.2^2-0.35^2-0.45^2 = 0.635$.

Q.2.22. Occurrence and co-occurrence. Of 100 Christmas shoppers, 50 spent £60 each, 20 spent £100 each, and 30 spent £150 each. What are the (i) average, (ii) median and (iii) modal spending? Tip: How one can take into account in the calculation that there are, effectively, only three different types of customers?

A. Average: First, let us see that the proportions of shoppers who spent £60, £100 and £150 each are, respectively, 0.5, 0.2 and 0.3. The average can be calculated by weighting the expenditures by the proportions so that Average = 60*0.5 + 100*0.2 + 150*0.3 = 30.0 + 20.0 + 45.0 = 95.

Median: According to definition, the median of 100 numbers is the mid value between 50th and 51st entries in their sorted order, which are 60 and 100 in this case. Thus the median spending is £80.

Mode: The modal value is the most likely one, that is, 60.

Q.2.23. Consider two geological formations that are represented by 7 and 5 ore specimens, respectively. The mineral contents in formation A is described by vector a = (7.6, 11.1, 6.8, 9.8, 4.9, 6.1, 15.1), and in formation B, by vector b = (4.7, 6.4, 4.1, 3.7, 3.9). The average content in A is 8.77 and in B, 4.56. Test the hypothesis that the mineral contents in A is richer than in B (with 95% confidence) by using bootstrap. A. Because the sets are quite small, the number of trials should be taken rather moderate, not greater than 200. At 200 trials, 95% confidence interval will be found with boundaries at 6-th and 195-th values in the sorted series of bootstrap means. In our computation, this is interval (6.66, 11.09) for A and (3.82, 5.44) for B. Since all of the former interval is greater than all of the latter interval, the hypothesis can be considered confirmed. (There is a flaw in this solution, because of some imprecision in the notion that A is richer than B. If we define that A is richer than B with 95% confidence if a random sample from A is richer than a random sample from B in 95% of the cases, then the 95%-intervals are not enough – they cover only 0.95*0.95 = 90.25% of all possible pairs of bootstrap mean values, which means the hypothesis is proven with 90% confidence. Yet if we take a look at the minimum and maximum bootstrap mean values, we find that the entire range of means is (6.33, 11.94) for A and (3.82, 5.82) for B, which means that the hypothesis is proven now since 5.82 < 6.33 – within the limits of the method.)

Q.2.24. Central triangular fuzzy set. Given three triangular fuzzy sets defined by triples (0, 2, 3), (0, 3, 4), and (3, 4, 8), determine the corresponding central triangular fuzzy set. **A.** The central triangular fuzzy set is defined by the average values such as (0+0+3)/3 = 1, for the first component; so that it is (1, 3, 5).

Q.2.25. Iris feature distributions. Consider histograms of Iris dataset features and demonstrate that two of them are bimodal. **A:** With a MatLab command like

>> for k = 1:4;subplot(2,2,k);hist(iris(:,k),15);end; a figure like Fig. 2.19 will appear. Obviously the third and fourth features are bimodal.

Q.2.26. To run a computational experiment, a student is to randomly generate distributions of relative frequencies for a three-category nominal variable. The student decides first generate random numbers in interval (0,1) and then make them sum to unity by relating them to thir sum. Thus, for example, random numbers 0.7116, 0.1295, 0.6598 are first generated, and then divided by their sum 1.5009 to produce values 0.4741, 0.0863, 0.4396 totaling to 1. Is it a right way to go? **A.** Not exactly. A bias towards equal frequencies will be created. For example, take a look at Fig. 2.20a presenting the distribution of the first element of a pair of frequencies found by the described method: generate a pair of random numbers and then divide them by the sum. This distribution is far from that of a uniformly random value presented on Fig. 2.20b. (Can you explain the difference?) An appropriate way for generating uniformly random frequency triplets would be this.

First, generate just two random numbers, then sort them in ascending order and add 0 and 1 into the series: $r_0 = 0 < r_1 < r_2 < r_3 = 1$. Then define the frequencies as differences of neighboring values in the series, $p_k = r_k - r_{k-1}$ $(k = 1, 2, 3)$. For example, if 0.8775, 0.5658 were first generated, then the frequencies would be

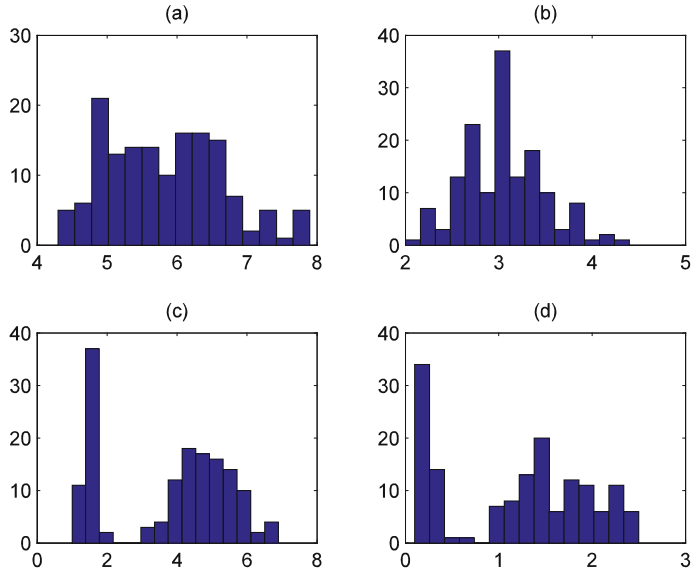

Fig. 2.19 Histograms of four Iris dataset featues; (**c**) and (**d**) are bimodal

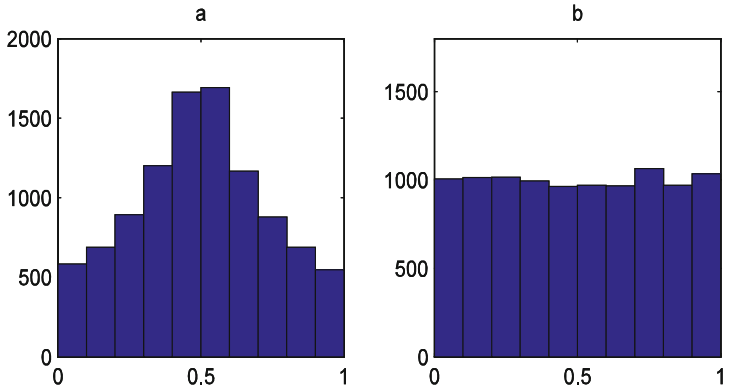

Fig. 2.20 Histograms of a 100,000 strong random sample of (**a**) the first element of a random pair after division by the pair summary value, and (**b**) just a random number

defined as $p_1 = 0.5658$, $p_2 = 0.8775 - 0.5658 = 0.3117$, and $p_3 = 1 - 0.8775 = 0.1225$. This method is easily extendable to any number of categories.

2.5 Summary

This chapter presents summaries of one-dimensional data, first of all, histograms, central points and spread evaluations. Two perspectives are outlined: one is the classical probabilistic and the other of approximation, naturally extending into the data

recovery approach to supply a decomposition of the data scatter in the explained and unexplained parts.

A difference between categorical and quantitative features is pointed out: the latter admit averaging whereas the former not. This difference is somewhat blurred at binary features especially the so-called dummies, 1/0 variables representing individual categories – they can be considered quantitative too.

Some attention is given to modeling uncertainty by using intervals and fuzzy sets, but not much. In fact, most of further methods can be extended to these more complex data types.

Several projects are presented to show how questions may arise and get computational answers. Computational intelligence and cross validation approaches are involved.

References

Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**, 1141–1164 (2000)

Efron, B., Tibshirani, R.: An Introduction to Bootstrap. Chapman & Hall, Boca Raton, FL (1993)

Engelbrecht, A.P.: Computational Intelligence. Wiley, New York, NY (2002)

Lohninger, H.: Teach Me Data Analysis. Springer, Berlin-New York-Tokyo (1999)

Polyak, B.: Introduction to Optimization. Optimization Software, Los Angeles, CA, ISBN: 0911575146 (1987)

Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning I–II. Inf. Sci., 8, 199-249, 301-375 (1975)