

Chapter 2

The Sample and Its Properties

When you're dealing with data, you have to look past the numbers.

– Nathan Yau

WHAT IS COVERED IN THIS CHAPTER

- MATLAB Session with Basic Univariate Statistics
- Numerical Characteristics of a Sample
- Multivariate Numerical and Graphical Sample Summaries
- Time Series
- Typology of Data



2.1 Introduction

The famous American statistician John Tukey once said, “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step.” The term *exploratory data analysis* is self-defining. Its simplest branch, *descriptive statistics*, is the methodology behind approaching and summarizing experimental data. No formal statistical training is needed for its use. Basic data manipulations such as calculating averages of experimental responses, translating data to pie charts or histograms, or assessing the variability and inspection for unusual measurements are all

examples of descriptive statistics. Rather than focusing on the population using information from a sample, which is a staple of statistics, descriptive statistics is concerned with the description, summary, and presentation of the sample itself. For example, numerical summaries of a sample could be measures of location (mean, median, percentiles, mode, extrema), measures of variability (sample standard deviation/variance, robust versions of the variance, range of data, interquartile range, etc.), higher-order statistics (k th moments, k th central moments, skewness, kurtosis), and functions of descriptors (coefficient of variation). Graphical summaries of samples involve various visual presentations such as box-and-whisker plots, pie charts, histograms, empirical cumulative distribution functions, etc. Many basic data descriptors are used in everyday data manipulation.

Ultimately, exploratory data analysis and descriptive statistics contribute to the principal goal of statistics – inference about population descriptors – by guiding how the statistical models should be set.

It is important to note that descriptive statistics and exploratory data analysis have recently regained importance due to ever increasing sizes of data sets. Some complex data structures require several terrabytes of memory just to be stored. Thus, preprocessing, summarizing, and dimension-reduction steps are needed to prepare such data for inferential tasks such as classification, estimation, and testing. Consequently, the inference is placed on data summaries (descriptors, features) rather than the raw data themselves.

Many data managing software programs have elaborate numerical and graphical capabilities. MATLAB provides an excellent environment for data manipulation and presentation with superb handling of data structures and graphics. In this chapter we intertwine some basic descriptive statistics with MATLAB programming using data obtained from real-life research laboratories. Most of the statistics are already built-in; for some we will make a custom code in the form of m-functions or m-scripts.

This chapter establishes two goals: (i) to help you gently relearn and refresh your MATLAB programming skills through annotated sessions while, at the same time, (ii) introducing some basic statistical measures, many of which should already be familiar to you. Many of the statistical summaries will be revisited later in the book in the context of inference. You are encouraged to continuously consult MATLAB's online help pages for support since many programming details and command options are omitted in this text.

2.2 A MATLAB Session on Univariate Descriptive Statistics

In this section we will analyze data derived from an experiment, step by step with a brief explanation of the MATLAB commands used. The whole session

can be found in a single annotated file  `carea.m` available at the book's Web page.

The data can be found in the file  `cellarea.dat`, which features measurements from the lab of Todd McDevitt at Georgia Tech: <http://www.bme.gatech.edu/groups/mcdevitt/>.

This experiment on cell growth involved several time durations and two motion conditions. Here is a brief description:

Embryonic stem cells (ESCs) have the ability to differentiate into all somatic cell types, making ESCs useful for studying developmental biology, in vitro drug screening, and as a cell source for regenerative medicine and cell-based therapies. A common method to induce differentiation of ESCs is through the formation of multicellular spheroids termed embryoid bodies (EBs). ESCs spontaneously aggregate into EBs when cultured on a nonadherent substrate; however, under static conditions, this aggregation is uncontrolled and EBs form in various sizes and shapes, which may lead to variability in cell differentiation patterns. When rotary motion is applied during EB formation, the resulting population of EBs appears more uniform in size and shape.

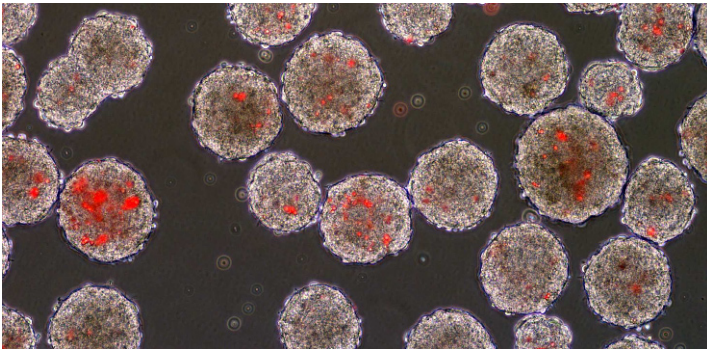



Fig. 2.1 Fluorescence microscopy image of cells overlaid with phase image to display incorporation of microspheres (*red stain*) in embryoid bodies (*gray clusters*) (courtesy of Todd McDevitt).

After 2, 4, and 7 days of culture, images of EBs were acquired using phase-contrast microscopy. Image analysis software was used to determine the area of each EB imaged (Fig. 2.1). At least 100 EBs were analyzed from three separate plates for both static and rotary cultures at the three time points studied.

Here we focus only on the measurements of visible surface areas of cells (in μm^2) after growth time of 2 days, $t = 2$, under the static condition. The data are recorded as an ASCII file  `cellarea.dat`. Importing the data set into MATLAB is done using the command

 `load('cellarea.dat');`

given that the data set is on the MATLAB path. If this is not the case, use `addpath('foldername')` to add to the search path `foldername` in which the file resides. A glimpse at the data is provided by histogram command, `hist`:

```
hist(cellarea, 100)
```

After inspecting the histogram (Fig. 2.2) we find that there is one quite unusual observation, inconsistent with the remaining experimental measurements.

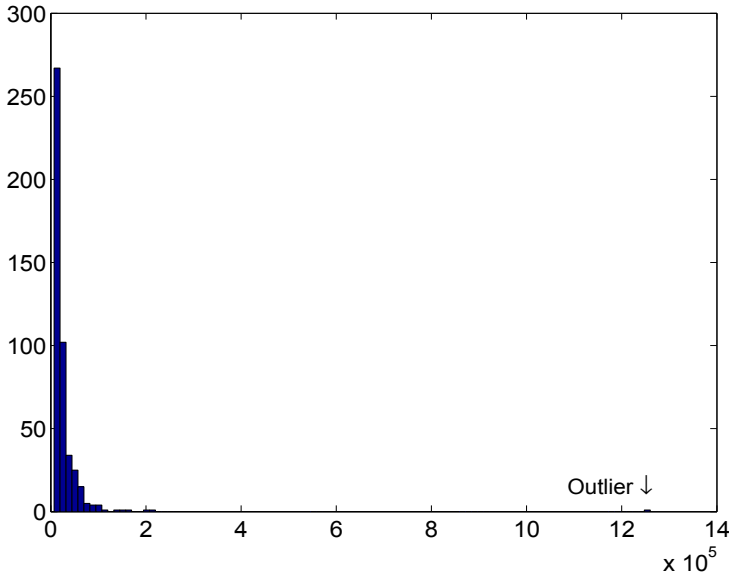


Fig. 2.2 Histogram of the raw data. Notice the unusual measurement beyond the point 12×10^5 .

We assume that the unusual observation is an outlier and omit it from the data set:

```
car = cellarea(cellarea ~= max(cellarea));
```

(Some formal diagnostic tests for outliers will be discussed later in the text.)

Next, the data are rescaled to more moderate values, so that the area is expressed in thousands of μm^2 and the measurements have a convenient order of magnitude.



```
car = car/1000;
n = length(car); %n is sample size
%n=462
```

Thus, we obtain a sample of size $n = 462$ to further explore by descriptive statistics. The histogram we have plotted has already given us a sense of the distribution within the sample, and we have an idea of the shape, location, spread, symmetry, etc. of observations.

Next, we find numerical characteristics of the sample and first discuss its location measures, which, as the name indicates, evaluate the relative location of the sample.

2.3 Location Measures

Means. The three averages – arithmetic, geometric, and harmonic – are known as Pythagorean means.

The arithmetic mean ([mean](#)),

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i,$$

is a fundamental summary statistic. The geometric mean ([geomean](#)) is

$$\sqrt[n]{X_1 \times X_2 \times \cdots \times X_n} = \left(\prod_{i=1}^n X_i \right)^{1/n},$$

and the harmonic mean ([harmmean](#)) is

$$\frac{n}{1/X_1 + 1/X_2 + \cdots + 1/X_n} = \frac{n}{\sum_{i=1}^n 1/X_i}.$$

For the data set $\{1, 2, 3\}$ the mean is 2, the geometric mean is $\sqrt[3]{6} = 1.8171$, and the harmonic mean is $3/(1/1 + 1/2 + 1/3) = 1.6364$. In standard statistical practice geometric and harmonic means are not used as often as arithmetic means. To illustrate the contexts in which they should be used, consider several simple examples.

Example 2.1. You visit the bank to deposit a long-term monetary investment in hopes that it will accumulate interest over a 3-year span. Suppose that the investment earns 10% the first year, 50% the second year, and 30% the third year. What is its average rate of return? In this instance it is not the arithmetic mean, because in the first year the investment was multiplied by 1.10, in the second year it was multiplied by 1.50, and in the third year it was multiplied by 1.30. The correct measure is the geometric mean of these three numbers, which is about 1.29, or 29% of the annual interest. If, for example, the ratios are averaged (i.e., ratio = new method/old method) over many experiments, the geometric mean should be used. This is evident by considering an example. If one experiment yields a ratio of 10 and the next yields a ratio of 0.1, an

arithmetic mean would misleadingly report that the average ratio was near 5.0. Taking a geometric mean will report a more meaningful average ratio of 1. (1985). This data set can be found on the book's Web page as well, as `fat.dat`.

Example 2.2. Consider now two scenarios in which the harmonic mean should be used.

(i) If for half the distance of a trip one travels at 40 miles per hour and for the other half of the distance one travels at 60 miles per hour, then the average speed of the trip is given by the harmonic mean of 40 and 60, which is 48; that is, the total amount of time for the trip is the same as if one traveled the entire trip at 48 miles per hour. Note, however, that if one had traveled for half the time at one speed and the other half at another, the arithmetic mean, in this case 50 miles per hour, would provide the correct interpretation of average.

(ii) In financial calculations, the harmonic mean is used to express the average cost of shares purchased over a period of time. For example, an investor purchases \$1000 worth of stock every month for 3 months. If the three spot prices at execution time are \$8, \$9, and \$10, then the average price the investor paid is \$8.926 per share. However, if the investor purchased 1000 shares per month, then the arithmetic mean should be used.

Order Statistic. If the sample X_1, \dots, X_n is ordered as $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ so that $X_{(1)}$ is the minimum and $X_{(n)}$ is the maximum, then $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ is called the *order statistic*. For example, if $X_1 = 2$, $X_2 = -1$, $X_3 = 10$, $X_4 = 0$, and $X_5 = 4$, then the order statistic is $X_{(1)} = -1$, $X_{(2)} = 0$, $X_{(3)} = 2$, $X_{(4)} = 4$, and $X_{(5)} = 10$.

Median. The median¹ is the middle of the sample sorted in numerical order. In terms of order statistic, the median is defined as

$$Me = \begin{cases} X_{((n+1)/2)}, & \text{if } n \text{ is odd,} \\ (X_{(n/2)} + X_{(n/2+1)})/2, & \text{if } n \text{ is even.} \end{cases}$$

If the sample size is odd, then there is a single observation in the middle of the ordered sample at the position $(n+1)/2$, while for the even sample sizes, the ordered sample has two elements in the middle at positions $n/2$ and $n/2+1$ and the median is their average. The median is an estimator of location robust to extremes and outliers. For instance, in both data sets, $\{-1, 0, 4, 7, 20\}$ and $\{-1, 0, 4, 7, 200\}$, the median is 4. The means are 6 and 42, respectively.

Mode. The most frequent (fashionable²) observation in the sample (if such exists) is the mode of the sample. If the sample is composite, the observation x_i corresponding to the largest frequency f_i is the mode. Composite samples consist of realizations x_i and their frequencies f_i , as in $\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ f_1 & f_2 & \dots & f_k \end{pmatrix}$.

¹ Latin: *medianus* = middle

² *Mode* (fr) = fashion

Mode may not be unique. If there are two modes, the sample is bimodal, three modes make it trimodal, etc.

Trimmed Mean. As mentioned earlier, the mean is a location measure sensitive to extreme observations and possible outliers. To make this measure more robust, one may trim $\alpha \cdot 100\%$ of the data symmetrically from both sides of the ordered sample (trim $\alpha/2 \cdot 100\%$ smallest and $\alpha/2 \cdot 100\%$ largest observations, Fig. 2.3b).

If your sample, for instance, is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, then a 20% trimmed mean is a mean of $\{2, 3, 4, 5, 6, 7, 8, 9\}$.

Here is the command in MATLAB that determines the discussed locations for the cell data.



```
location = [geomean(car) harmmean(car) mean(car) ...
            median(car) mode(car) trimmean(car,20)]
%location = 18.8485 15.4211 24.8701 17 10 20.0892
```

By applying $\alpha 100\%$ trimming, we end up with a sample of reduced size $[(1 - \alpha)100\%]$. Sometimes the sample size is important to preserve.

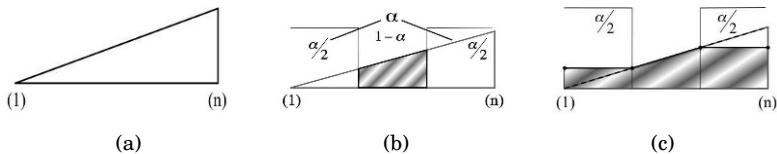


Fig. 2.3 (a) Schematic graph of an ordered sample; (b) Part of the sample from which α -trimmed mean is calculated; (c) Modified sample for the winsorized mean.

Winsorized mean. A robust location measure that preserves sample size is the winsorized mean. Similarly to a trimmed mean, a winsorized mean identifies outlying observations, but instead of trimming them the observations are replaced by either the minimum or maximum of the trimmed sample, depending on if the trimming is done from below or above (Fig. 2.3c).

The winsorized mean is not a built-in MATLAB function. However, it can be calculated easily by the following code:



```
alpha=20;
sa = sort(car);
sa(1:floor( n*alpha/200 )) = sa(floor( n*alpha/200 ) + 1);
sa(end-floor( n*alpha/200 ):end) = ...
    sa(end-floor( n*alpha/200 ) - 1);
winsmean = mean(sa) % winsmean = 21.9632
```

Figure 2.3 shows schematic graphs of of a sample

2.4 Variability Measures

Location measures are intuitive but give a minimal glimpse at the nature of a sample. An important set of sample descriptors are variability measures, or measures of spread. There are many measures of variability in a sample. Gauss (1816) already used several of them on a set of 48 astronomical measurements concerning relative positions of Jupiter and its satellite Pallas.

Sample Variance and Sample Standard Deviation. The variance of a sample, or sample variance, is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Note that we use $\frac{1}{n-1}$ instead of the “expected” $\frac{1}{n}$. The reasons for this will be discussed later. An alternative expression for s^2 that is more suitable for calculation (by hand) is

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n (X_i^2) - n(\bar{X})^2 \right),$$

see Exercises 2.6 and 2.7.

In MATLAB, the sample variance of a data vector \mathbf{x} is `var(x)` or `var(x,0)`. Flag 0 in the argument list indicates that the ratio $1/(n-1)$ is used to calculate the sample variance. If the flag is 1, then `var(x,1)` stands for

$$s_*^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

which is sometimes used instead of s^2 . We will see later that both estimators have good properties: s^2 is an unbiased estimator of the population variance while s_*^2 is the maximum likelihood estimator. The square root of the sample variance is the *sample standard deviation*:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

In MATLAB the standard deviation can be calculated by `std(x)=std(x,0)` or `std(x,1)`, depending on whether the sum of squares is divided by $n-1$ or by n .



```
%Variability Measures
var(car) % standard sample variance, also var(car,0)
%ans = 588.9592
var(car,1) % sample variance with sum of squares
% divided by n
%ans = 587.6844
std(car) % sample standard deviation, sum of squares
% divided by (n-1), also std(car,0)
%ans = 24.2685
std(car,1) % sample standard deviation, sum of squares
% divided by n
%ans = 24.2422
sqrt(var(car)) %should be equal to std(car)
%ans = 24.2685
sqrt(var(car,1)) %should be equal to std(car,1)
%ans = 24.2422
```

Remark. When a new observation is obtained, one can update the sample variance without having to recalculate it. If \bar{x}_n and s_n^2 are the sample mean and variance based on x_1, x_2, \dots, x_n and a new observation x_{n+1} is obtained, then

$$s_{n+1}^2 = \frac{(n-1)s_n^2 + (x_{n+1} - \bar{x}_n)(x_{n+1} - \bar{x}_{n+1})}{n},$$

where $\bar{x}_{n+1} = (n\bar{x}_n + x_{n+1})/(n+1)$.

MAD-Type Estimators. Another group of estimators of variability involves absolute values of deviations from the center of a sample and are known as MAD estimators. These estimators are less sensitive to extreme observations and outliers compared to the sample standard deviation. They belong to the class of so-called robust estimators. The acronym MAD stands for either *mean absolute difference from the mean* or, more commonly, *median absolute difference from the median*. According to statistics historians (David, 1998), both MADs were already used by Gauss at the beginning of the nineteenth century.

MATLAB uses `mad(car)` or `mad(a,0)` for the first and `mad(car,1)` for the second definition:

$$\text{MAD}_0 = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|, \quad \text{MAD}_1 = \text{median}\{|X_i - \text{median}\{X_i\}|\}.$$

A typical convention is to multiply the MAD_1 estimator `mad(car,1)` by 1.4826, to make it comparable to the sample standard deviation.



```
mad(car) % mean absolute deviation from the mean;
          % MAD is usually referring to
          % median absolute deviation from the median
ans = 15.3328
realmad = 1.4826 * median( abs(car - median(car)))
          %real mad in MATLAB is 1.4826 * mad(car,1)
realmad = 10.3781
```

Sample Range and IQR. Two simple measures of variability, or rather the spread of a sample, are the range R and interquartile range (IQR), in MATLAB `range` and `iqr`. They are defined by the order statistic of the sample. The range is the maximum minus the minimum of the sample, $R = X_{(n)} - X_{(1)}$, while IQR is defined by sample quantiles.



```
range(car) %Range, span of data, Max - Min
ans = 212
iqr(car)    %inter-quartile range, Q3-Q1
ans = 19
```

If the sample is bell-shape distributed, a robust estimator of variance is $\hat{\sigma}^2 = (\text{IQR}/1.349)^2$, and this summary was known to Quetelet in the first part of the nineteenth century. It is a simple estimator, not affected by outliers (it ignores 25% of observations in each tail), but its variability is large.

Sample Quantiles/Percentiles. Sample quantiles (in units between 0 and 1) or sample percentiles (in units between 0 and 100) are very important summaries that reveal both the location and the spread of a sample. For example, we may be interested in a point x_p that partitions the ordered sample into two parts, one with $p \cdot 100\%$ of observations smaller than x_p and another with $(1-p)100\%$ observations greater than x_p . In MATLAB, we use the commands `quantile` or `prctile`, depending on how we express the proportion of the sample. For example, for the 5, 10, 25, 50, 75, 90, and 95 percentiles we have



```
%5%, 10%, 25%, 50%, 75%, 90%, 95% percentiles are:
prctile(car, 100*[0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95] )
ans = 7      8      11      17      30      51      67
```

The same results can be obtained using the command

```
qts = quantile(car,[0.05 0.1 0.25 0.5 0.75 0.9 0.95])
%qts = 7      8      11      17      30      51      67
```

In our dataset, 5% of observations are less than 7, and 90% of observations are less than 51.

Some percentiles/quantiles are special, such as the median of the sample, which is the 50th percentile. Quartiles divide an ordered sample into four parts; the 25th percentile is known as the first quartile, Q_1 , and the 75th percentile is known as the third quartile, Q_3 . The median is Q_2 , of course.³

³ The range is equipartitioned by a single median, two terciles, three quartiles, four quintiles, five sextiles, six septiles, seven octiles, eight naniles, or nine deciles.

In MATLAB, `Q1=prctile(car,25); Q3=prctile(car,75)`. Now we can define the IQR as $Q_3 - Q_1$:



```
prctile(car, 75)- prctile(car, 25) %should be equal to iqr(car).
%ans = 19
```

The five-number summary for univariate data is defined as (Min, Q_1, Me, Q_3, Max) .

z-Scores. For a sample x_1, x_2, \dots, x_n the z -score is the standardized sample z_1, z_2, \dots, z_n , where $z_i = (x_i - \bar{x})/s$. In the standardized sample, the mean is 0 and the sample variance (and standard deviation) is 1. The basic reason why standardization may be needed is to assess extreme values, or compare samples taken at different scales. Some other reasons will be discussed in subsequent chapters.



```
zcar = zscore(car);
mean(zcar)
%ans = -5.8155e-017
var(zcar)
%ans = 1
```

Moments of Higher Order. The term *sample moments* is drawn from mechanics. If the observations are interpreted as unit masses at positions X_1, \dots, X_n , then the sample mean is the first moment in the mechanical sense – it represents the balance point for the system of all points. The moments of higher order have their corresponding mechanical interpretation. The formula for the k th moment is

$$m_k = \frac{1}{n}(X_1^k + \dots + X_n^k) = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

The moments m_k are sometimes called *raw* sample moments. The *power k mean* is $(m_k)^{1/k}$, that is,

$$\left(\frac{1}{n} \sum_{i=1}^n X_i^k \right)^{1/k}.$$

For example, the sample mean is the first moment and power 1 mean, $m_1 = \bar{X}$. The *central* moments of order k are defined as

$$\mu_k = \frac{1}{n} \sum_{i=1}^n (X_i - m_1)^k.$$

Notice that $\mu_1 = 0$ and that μ_2 is the sample variance (calculated by `var(.,1)` with the sum of squares divided by n). MATLAB has a built-in function `moment` for calculating the central moments.



```
%Moments of Higher Orders
%kth (row) moment: mean(car.^k)
mean(car.^3) %third
%ans = 1.1161e+005
%kth central moment mean((car-mean(car)).^k)
mean( (car-mean(car)).^3 ) %ans=5.2383e+004
%is the same as
moment(car,3) %ans=5.2383e+004
```

Skewness and Kurtosis. There are many uses of higher moments in describing a sample. Two important sample measures involving higher-order moments are *skewness* and *kurtosis*.

Skewness is defined as

$$\gamma_n = \mu_3/\mu_2^{3/2} = \mu_3/s_*^3$$

and measures the degree of asymmetry in a sample distribution. Positively skewed distributions have longer right tails and their sample mean is larger than the median. Negatively skewed sample distributions have longer left tails and their mean is smaller than the median.

Kurtosis is defined as

$$\kappa_n = \mu_4/\mu_2^2 = \mu_4/s_*^4.$$

It represents the measure of “peakedness” or flatness of a sample distribution. In fact, there is no consensus on the exact definition of kurtosis since flat but fat-tailed distributions would also have high kurtosis. Distributions that have a kurtosis of <3 are called *platykurtic* and those with a kurtosis of >3 are called *leptokurtic*.



```
%sample skewness mean(car.^3)/std(car,1)^3
mean( (car-mean(car)).^3 )/std(car,1)^3 %ans = 3.6769
skewness(car) %ans = 3.6769
%sample kurtosis
mean( (car-mean(car)).^4 )/std(car,1)^4 %ans = 22.8297
kurtosis(car)%ans = 22.8297
```

A robust version of the skewness measure was proposed by Bowley (1920) as

$$\gamma_n^* = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1},$$

and ranges between -1 and 1 . Moors (1988) proposed a robust measure of kurtosis based on sample octiles:

$$\kappa_n^* = \frac{(O_7 - O_5) + (O_3 - O_1)}{O_6 - O_2},$$

where O_i is the $i/8 \times 100$ percentile (i th octile) of the sample for $i = 1, 2, \dots, 7$. If the sample is large, one can take O_i as $X_{(\lfloor i/8 \times n \rfloor)}$. The constant 1.766 is sometimes added to κ_n^* as a calibration adjustment so that it is comparable with the traditional measure of kurtosis for samples from Gaussian populations.



```
%robust skewness
(prctile(car, 75)+prctile(car, 25) - ...
2 * median(car))/(prctile(car, 75) - prctile(car, 25))
%0.3684

%robust kurtosis
(prctile(car,7/8*100)-prctile(car,5/8*100)+prctile(car,3/8*100)- ...
prctile(car,1/8*100))/(prctile(car,6/8*100)-prctile(car,2/8*100))
%1.4211
```

Coefficient of Variation. The coefficient of variation, CV, is the ratio

$$CV = \frac{s}{\bar{X}}.$$

The CV expresses the variability of a sample in the units of its mean. In other words, a CV equal to 2 would mean that the variability is equal to $2\bar{X}$. The assumption is that the mean is positive. The CV is used when comparing the variability of data reported on different scales. For example, instructors A and B teach different sections of the same class, but design their own final exams individually. To compare the effectiveness of their respective exam designs at

creating a maximum variance in exam scores (a tacit goal of exam designs), they calculate the CVs. It is important to note that the CVs would not be related to the exam grading scale, to the relative performance of the students, or to the difficulty of the exam.



```
%sample CV [coefficient of variation]
std(car)/mean(car)
%ans = 0.9758
```

The reciprocal of CV, \bar{X}/s , is sometimes called the signal-to-noise ratio, and it is often used in engineering quality control.

Grouped Data. When a data set is large and many observations are repetitive, data are often recorded as grouped or composite. For example, the data set

4	5	6	3	4	3	6	4	5	4	3
7	3	5	2	5	6	4	2	4	3	4
7	7	4	2	2	5	4	2	5	3	8

is called a simple sample, or raw sample, as it lists explicitly all observations. It can be presented in a more compact form, as grouped data:

X_i	2	3	4	5	6	7	8
f_i	5	6	9	6	3	3	1

where X_i are distinctive values in the data set with frequencies f_i , and the number of groups is $k = 7$. Notice that $X_i = 5$ appears six times in the simple sample, so its frequency is $f_i = 6$.

The function  `[xi fi]=simple2comp(a)` provides frequencies `fi` for a list `xi` of distinctive values in `a`.



```
a=[ 4 5 6 3 4 3 6 4 5 4 3 ...
    7 3 5 2 5 6 4 2 4 3 4 ...
    7 7 4 2 2 5 4 2 5 3 8];
[xi fi] = simple2comp( a )
% xi =
%    2    3    4    5    6    7    8
% fi =
%    5    6    9    6    3    3    1
```

Here, $n = \sum_i f_i = 33$.

When a sample is composite, the sample mean and variance are

$$\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{n}, \quad s^2 = \frac{\sum_{i=1}^k f_i (X_i - \bar{X})^2}{n-1}$$

for $n = \sum_i f_i$. By defining the m th raw and central sample moments as

$$\overline{X^m} = \frac{\sum_{i=1}^k f_i X_i^m}{n} \quad \text{and} \quad \mu_m = \frac{\sum_{i=1}^k f_i (X_i - \bar{X})^m}{n-1},$$

one can express skewness, kurtosis, CV, and other sample statistics that are functions of moments.

Diversity Indices for Categorical Data. If the data are categorical and numerical characteristics such as moments and percentiles cannot be defined, but the frequencies f_i of classes/categories are given, one can define Shannon's diversity index:

$$H = \frac{n \log n - \sum_{i=1}^k f_i \log f_i}{n}. \quad (2.1)$$

If some frequency is 0, then $0 \log 0 = 0$. The maximum of H is $\log k$; it is achieved when all f_i are equal. The normalized diversity index, $E_H = H/\log k$, is called Shannon's homogeneity (equitability) index of the sample.

Neither H nor E_H depends on the sample size.

Example 2.3. Homogeneity of Blood Types. Suppose the samples from Brazilian, Indian, Norwegian, and US populations are taken and the frequencies of blood types (ABO/Rh) are obtained.

Population	O+	A+	B+	AB+	O-	A-	B-	AB-	total
Brazil	115	108	25	6	28	25	6	1	314
India	220	134	183	39	12	6	6	12	612
Norway	83	104	16	8	14	18	2	1	246
US	99	94	21	8	18	18	5	2	265

Which county's sample is most homogeneous with respect to blood type attribute?



```
br = [115 108 25 6 28 25 6 1];
in = [220 134 183 39 12 6 6 12];
no = [ 83 104 16 8 14 18 2 1];
us = [ 99 94 21 8 18 18 5 2];
```

```
Eh = @(f) (sum(f)*log(sum(f)) - ...
          sum( f.*log(f)))/(sum(f)*log(length(f)))
```

```
Eh(br) % 0.7324
Eh(in) % 0.7125
Eh(no) % 0.6904
Eh(us) % 0.7306
```

Among the four samples, the sample from Brazil is the most homogeneous with respect to the blood types of its population as it maximizes the statistic E_H . See also Exercise 2.13 for an alternative definition of diversity/homogeneity indices.



2.5 Displaying Data

In addition to their numerical descriptors, samples are often presented in a graphical manner. In this section, we discuss some basic graphical summaries.

Box-and-Whiskers Plot. The top and bottom of the “box” are the 25th and 75th percentile of the data, respectively, with the distances between them representing the IQR. The line inside the box represents the sample median. If the median is not centered in the box, it indicates sample skewness. Whiskers extend from the lower and upper sides of the box to the data’s most extreme values within 1.5 times the IQR. Potential outliers are displayed with red “+” beyond the endpoints of the whiskers.

The MATLAB command `boxplot(X)` produces a box-and-whisker plot for X . If X is a matrix, the boxes are calculated and plotted for each column. [Figure 2.4a](#) is produced by



```
%Some Graphical Summaries of the Sample
figure;
boxplot(car)
```

Histogram. As illustrated previously in this chapter, the histogram is a rough approximation of the population distribution based on a sample. It plots frequencies (or relative frequencies for normalized histograms) for interval-grouped data. Graphically, the histogram is a barplot over contiguous intervals or bins spanning the range of data ([Fig. 2.4b](#)). In MATLAB, the typical command for a histogram is `[fre,xout] = hist(data,nbins)`, where `nbins` is the number of bins and the outputs `fre` and `xout` are the frequency counts and the bin locations, respectively. Given the output, one can use `bar(xout,n)` to plot the histogram. When the output is not requested, MATLAB produces the plot by default.



```
figure;
hist(car, 80)
```

The histogram is only an approximation of the distribution of measurements in the population from which the sample is obtained.

There are numerous rules on how to automatically determine the number of bins or, equivalently, bin sizes, none of them superior to the others on all possible data sets. A commonly used proposal is Sturges’ rule (Sturges, 1926), where the number of bins k is suggested to be

$$k = 1 + \log_2 n,$$

where n is the size of the sample. Sturges’ rule was derived for bell-shaped distributions of data and may oversmooth data that are skewed, multimodal, or have some other features. Other suggestions specify the bin size as $h = 2 \cdot \text{IQR}/n^{1/3}$ (Diaconis–Freedman rule) or, alternatively, $h = (7s)/(2n^{1/3})$ (Scott’s rule; s is the sample standard deviation). By dividing the range of the data by h , one finds the number of bins.

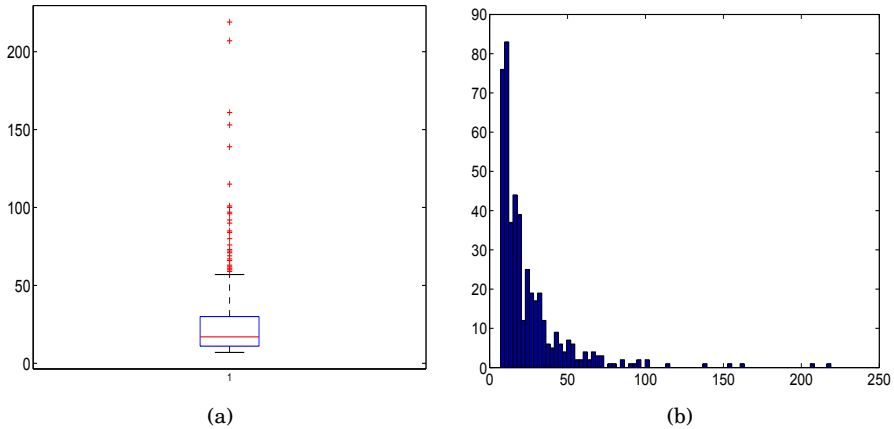


Fig. 2.4 (a) Box plot and (b) histogram of cell data `car`.

For example, for cell-area data `car`, Sturges' rule suggests 10 bins, Scott's 19 bins, and the Diaconis–Freedman rule 43 bins. The default `nbins` in MATLAB is 10 for any sample size.

The histogram is a crude estimator of a probability density that will be discussed in detail later on (Chap. 5). A more esthetic estimator of the population distribution is given by the *kernel smoother density* estimate, or `ksdensity`. We will not go into the details of kernel smoothing at this point in the text; however, note that the spread of a kernel function (such as a Gaussian kernel) regulates the degree of smoothing and in some sense is equivalent to the choice of bin size in histograms.

Command `[f,xi,u]=ksdensity(x)` computes a density estimate based on data `x`. Output `f` is the vector of density values evaluated at the points in `xi`. The estimate is based on a normal kernel function, using a window parameter `width` that depends on the number of points in `x`. The default width `u` is returned as an output and can be used to tune the smoothness of the estimate, as is done in the example below. The density is evaluated at 100 equally spaced points that cover the range of the data in `x`.



```
figure;
[f,x,u] = ksdensity(car);
plot(x,f)
hold on
[f,x] = ksdensity(car,'width',u/3);
plot(x,f,'r');
[f,x] = ksdensity(car,'width',u*3);
plot(x,f,'g');
legend('default width','default/3','3 * default')
hold off
```

Empirical Cumulative Distribution Function. The empirical cumulative distribution function (ECDF) $F_n(x)$ for a sample X_1, \dots, X_n is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq x) \quad (2.2)$$

and represents the proportion of sample values smaller than x . Here $\mathbf{1}(X_i \leq x)$ is either 0 or 1. It is equal to 1 if $\{X_i \leq x\}$ is true, 0 otherwise.

The function `empiricalcdf(x, sample)` will calculate the ECDF based on the observations in `sample` at a value `x`.



```
xx = min(car)-1:0.01:max(car)+1;
yy = empiricalcdf(xx, car);
plot(xx, yy, 'k-', 'linewidth', 2)
xlabel('x'); ylabel('F_n(x)')
```

In MATLAB, `[f xf]=ecdf(x)` is used to calculate the proportion `f` of the sample `x` that is smaller than `xf`. Figure 2.5b shows the ECDF for the cell area data, `car`.

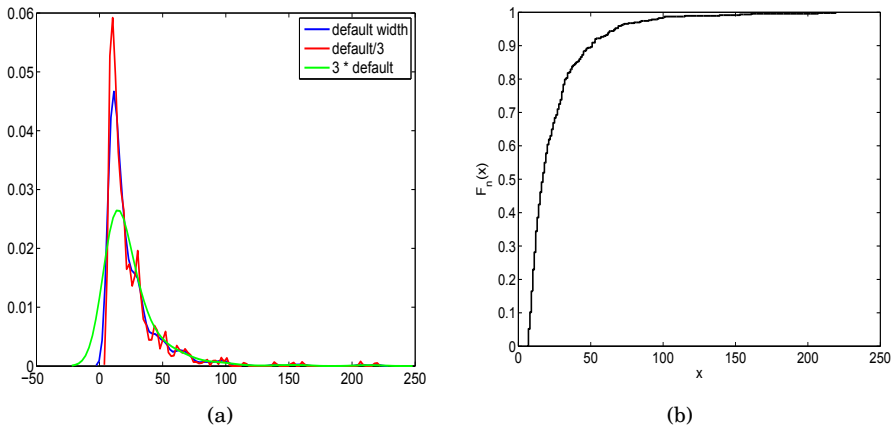


Fig. 2.5 (a) Smoothed histogram (density estimator) for different widths of smoothing kernel; (b) Empirical CDF.

Q–Q Plots. Q–Q plots, short for quantile–quantile plots, compare the distribution of a sample with some standard theoretical distribution, such as normal, or with a distribution of another sample. This is done by plotting the sample quantiles of one distribution against the corresponding quantiles of the other. If the plot is close to linear, then the distributions are close (up to a scale

and shift). If the plot is close to the 45° line, then the compared distributions are approximately equal. In MATLAB the command `qqplot(X,Y)` produces an empirical Q–Q plot of the quantiles of the data set X vs. the quantiles of the data set Y . If the data set Y is omitted, then `qqplot(X)` plots the quantiles of X against standard normal quantiles and essentially checks the normality of the sample.

Figure 2.6 gives us the Q–Q plot of the cell area data set against the normal distribution. Note the deviation from linearity suggesting that the distribution is skewed. A line joining the first and third sample quartiles is superimposed in the plot. This line is extrapolated out to the ends of the sample to help visually assess the linearity of the Q–Q display. Q–Q plots will be discussed in more detail in Chap. 13.

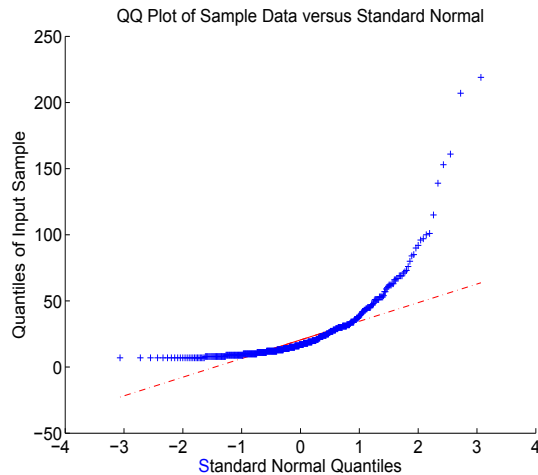


Fig. 2.6 Quantiles of data plotted against corresponding normal quantiles, via `qqplot`.

Pie Charts. If we are interested in visualizing proportions or frequencies, the pie chart is appropriate. A pie chart (`pie` in MATLAB) is a graphical display in the form of a circle where the proportions are assigned segments.

Suppose that in the cell area data set we are interested in comparing proportions of cells with areas in three regions: smaller than or equal to 15, between 15 and 30, and larger than 30. We would like to emphasize the proportion of cells with areas between 15 and 30. The following MATLAB code plots the pie charts (Fig. 2.7).



```
n1 = sum( car <= 15 ); %n1=213
n2 = sum( (car > 15 ) & (car <= 30) ); %n2=139
n3 = sum( car > 30 ); %n3=110
```

```
% n=n1+n2+n3 = 462
% proportions n1/n, n2/n, and n3/n are
%           0.4610, 0.3009 and 0.2381
explode = [0 1 0]
pie([n1, n2, n3], explode)
pie3([n1, n2, n3], explode)
```

Note that option `explode=[0 1 0]` separates the second segment from the circle. The command `pie3` plots a 3-D version of a pie chart (Fig. 2.7b).

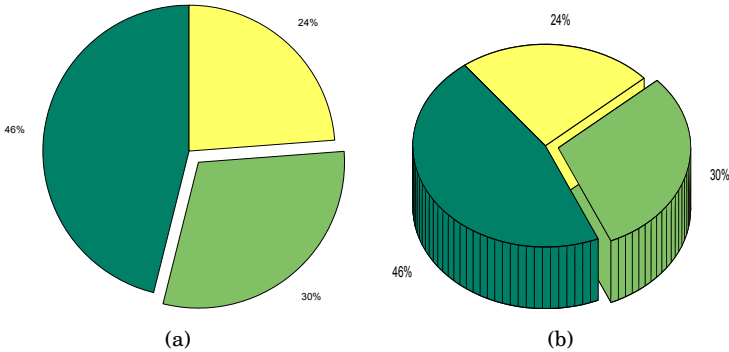


Fig. 2.7 Pie charts for frequencies 213, 139, and 110 of cell areas smaller than or equal to 15, between 15 and 30, and larger than 30. The proportion of cells with the area between 15 and 30 is emphasized.

2.6 Multidimensional Samples: Fisher's Iris Data and Body Fat Data

In the cell area example, the sample was univariate, that is, each measurement was a scalar. If a measurement is a vector of data, then descriptive statistics and graphical methods increase in importance, but they are much more complex than in the univariate case. The methods for understanding multivariate data range from the simple rearrangements of tables in which raw data are tabulated, to quite sophisticated computer-intensive methods in which exploration of the data is reminiscent of futuristic movies from space explorations.

Multivariate data from an experiment are first recorded in the form of tables, by either a researcher or a computer. In some cases, such tables may appear uninformative simply because of their format of presentation. By simple rules such tables can be rearranged in more useful formats. There are several guidelines for successful presentation of multivariate data in the form of tables. (i) Numbers should be maximally simplified by rounding as long as

it does not affect the analysis. For example, the vector (2.1314757, 4.9956301, 6.1912772) could probably be simplified to (2.14, 5, 6.19); (ii) Organize the numbers to compare columns rather than rows; and (iii) The user's cognitive load should be minimized by spacing and table lay-out so that the eye does not travel long in making comparisons.

Fisher's Iris Data. An example of multivariate data is provided by the celebrated Fisher's iris data. Plants of the family *Iridaceae* grow on every continent except Antarctica. With a wealth of species, identification is not simple. Even iris experts sometimes disagree about how some flowers should be classified. Fisher's (Anderson, 1935; Fisher, 1936) data set contains measurements on three North American species of iris: *Iris setosa canadensis*, *Iris versicolor*, and *Iris virginica* (Fig. 2.8a-c). The 4-dimensional measurements on each of the species consist of sepal and petal length and width.

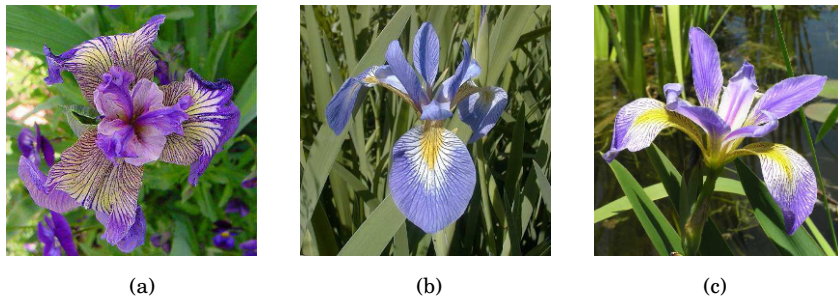


Fig. 2.8 (a) *Iris setosa*, C. Hensler, The Rock Garden, (b) *Iris virginica*, and (c) *Iris versicolor*, (b) and (c) are photos by D. Kramb, SIGNA.

The data set `fisheriris` is part of the MATLAB distribution and contains two files: `meas` and `species`. The `meas` file, shown in Fig. 2.9a, is a 150×4 matrix and contains 150 entries, 50 for each species. Each row in the matrix `meas` contains four elements: sepal length, sepal width, petal length, and petal width. Note that the convention in MATLAB is to store variables as columns and observations as rows.

The data set `species` contains names of species for the 150 measurements. The following MATLAB commands plot the data and compare sepal lengths among the three species.

```
load fisheriris
s1 = meas(1:50, 1); %setosa,      sepal length
s2 = meas(51:100, 1); %versicolor, sepal length
s3 = meas(101:150, 1); %virginica,  sepal length
s = [s1 s2 s3];
figure;
imagesc(meas)
```

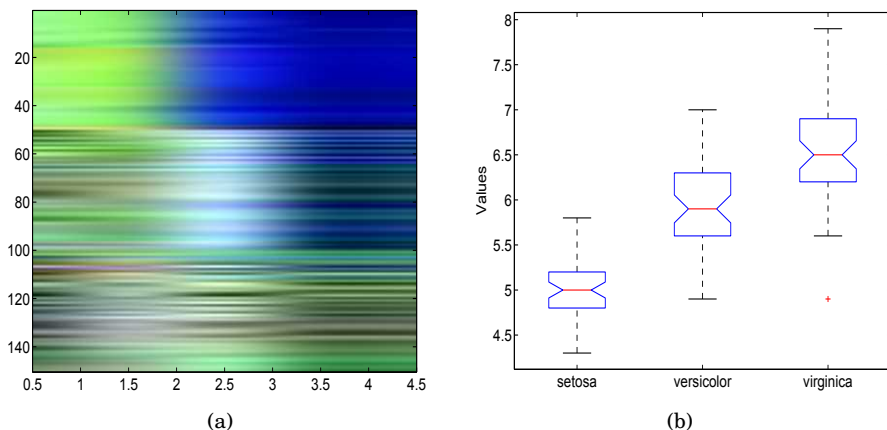


Fig. 2.9 (a) Matrix means in fisheriris, (b) Box plots of Sepal Length (the first column in matrix means) versus species.

```
figure;
boxplot(s,'notch','on',...
        'labels',{'setosa','versicolor','virginica'})
```

Correlation in Paired Samples. We will briefly describe how to find the correlation between two aligned vectors, leaving detailed coverage of correlation theory to Chap. 15.

Sample correlation coefficient r measures the strength and direction of the linear relationship between two paired samples $X = (X_1, X_2, \dots, X_n)$ and $Y = (Y_1, Y_2, \dots, Y_n)$. Note that the order of components is important and the samples cannot be independently permuted if the correlation is of interest. Thus the two samples can be thought of as a single bivariate sample (X_i, Y_i) , $i = 1, \dots, n$.

The correlation coefficient between samples $X = (X_1, X_2, \dots, X_n)$ and $Y = (Y_1, Y_2, \dots, Y_n)$ is

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

The summary $\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \left(\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right)$ is called the sample covariance. The correlation coefficient can be expressed as a ratio:

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y},$$

where s_X and s_Y are sample standard deviations of samples X and Y .

Covariances and correlations are basic exploratory summaries for paired samples and multivariate data. Typically they are assessed in data screening before building a statistical model and conducting an inference. The correlation ranges between -1 and 1 , which are the two ideal cases of decreasing and increasing linear trends. Zero correlation does not, in general, imply independence but signifies the lack of any linear relationship between samples.

To illustrate the above principles, we find covariance and correlation between sepal and petal lengths in Fisher's iris data. These two variables correspond to the first and third columns in the data matrix. The conclusion is that these two lengths exhibit a high degree of linear dependence as evident in Fig. 2.10. The covariance of 1.2743 by itself is not a good indicator of this relationship since it is scale (magnitude) dependent. However, the correlation coefficient is not influenced by a linear transformation of the data and in this case shows a strong positive relationship between the variables.



```
load fisheriris
X = meas(:, 1);    %sepal length
Y = meas(:, 3);    %petal length
cv = cov(X, Y); cv(1,2) %1.2743
r = corr(X, Y)     %0.8718
```

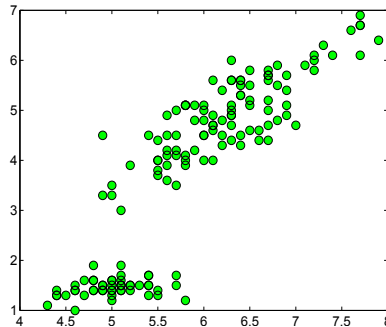



Fig. 2.10 Correlation between petal and sepal lengths (columns 1 and 3) in iris data set. Note the strong linear dependence with a positive trend. This is reflected by a covariance of 1.2743 and a correlation coefficient of 0.8718 .

In the next section we will describe an interesting multivariate data set and, using MATLAB, find some numerical and graphical summaries.

Example 2.4. Body Fat Data. We also discuss a multivariate data set analyzed in Johnson (1996) that was submitted to  <http://www.amstat.>


org/publications/jse/datasets/fat.txt and featured in Penrose et al. (1985). This data set can be found on the book’s Web page as well, as  fat.dat.



Fig. 2.11 Water test to determine body density. It is based on underwater weighing (Archimedes’ principle) and is regarded as the gold standard for body composition assessment.

Percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) were recorded for 252 men. Percent of body fat is estimated through an underwater weighing technique (Fig. 2.11).

The data set has 252 observations and 19 variables. Brozek and Siri indices (Brozek et al., 1963; Siri, 1961) and fat-free weight are obtained by the underwater weighing while other anthropometric variables are obtained using scales and a measuring tape. These anthropometric variables are less intrusive but also less reliable in assessing the body fat index.

–		Variable description
3–5	casen	Case number
10–13	broz	Percent body fat using Brozek’s equation: $457/\text{density} - 414.2$
18–21	siri	Percent body fat using Siri’s equation: $495/\text{density} - 450$
24–29	densi	Density (gm/cm^3)
36–37	age	Age (years)
40–45	weight	Weight (lb.)
49–53	height	Height (in.)
58–61	adiposi	Adiposity index = $\text{weight}/(\text{height}^2)$ (kg/m^2)
65–69	ffwei	Fat-free weight = $(1 - \text{fraction of body fat}) \times \text{weight}$, using Brozek’s formula (lb.)
74–77	neck	Neck circumference (cm)
81–85	chest	Chest circumference (cm)
89–93	abdomen	Abdomen circumference (cm)
97–101	hip	Hip circumference (cm)
106–109	thigh	Thigh circumference (cm)
114–117	knee	Knee circumference (cm)
122–125	ankle	Ankle circumference (cm)
130–133	biceps	Extended biceps circumference (cm)
138–141	forearm	Forearm circumference (cm)
146–149	wrist	Wrist circumference (cm) “distal to the styloid processes”

Remark: There are a few false recordings. The body densities for cases 48, 76, and 96, for instance, each seem to have one digit in error as seen from

the two body fat percentage values. Also note the presence of a man (case 42) over 200 lb. in weight who is less than 3 ft. tall (the height should presumably be 69.5 in., not 29.5 in.)! The percent body fat estimates are truncated to zero when negative (case 182).



```
load('\your path\fat.dat')
casen = fat(:,1);
broz = fat(:,2);
siri = fat(:,3);
densi = fat(:,4);
age = fat(:,5);
weight = fat(:,6);
height = fat(:,7);
adiposi = fat(:,8);
ffwei = fat(:,9);
neck = fat(:,10);
chest = fat(:,11);
abdomen = fat(:,12);
hip = fat(:,13);
thigh = fat(:,14);
knee = fat(:,15);
ankle = fat(:,16);
biceps = fat(:,17);
forearm = fat(:,18);
wrist = fat(:,19);
```

We will further analyze this data set in this chapter, as well as in Chap. 16, in the context of multivariate regression.



2.7 Multivariate Samples and Their Summaries*

Multivariate samples are organized as a data matrix, where the rows are observations and the columns are variables or components. One such data matrix of size $n \times p$ is shown in [Fig. 2.12](#).

The measurement x_{ij} denotes the j th component of the i th observation. There are n row vectors $\mathbf{x}_1', \mathbf{x}_2', \dots, \mathbf{x}_n'$ and p columns $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(n)}$, so that

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = [\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \dots, \mathbf{x}_{(n)}].$$

Note that $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ is a p -vector denoting the i th observation, while $\mathbf{x}_{(j)} = (x_{1j}, x_{2j}, \dots, x_{nj})'$ is an n -vector denoting values of the j th variable/component.

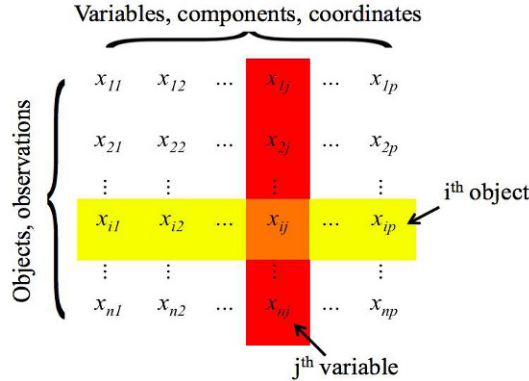


Fig. 2.12 Data matrix X . In the multivariate sample the rows are observations and the columns are variables.

The mean of data matrix X is a vector \bar{x} , which is a p -vector of column means

$$\bar{x} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_{i1} \\ \frac{1}{n} \sum_{i=1}^n x_{i2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n x_{ip} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}.$$

By denoting a vector of ones of size $n \times 1$ as $\mathbf{1}$, the mean can be written as $\bar{x} = \frac{1}{n} X' \cdot \mathbf{1}$, where X' is the transpose of X .

Note that \bar{x} is a column vector, while MATLAB's command `mean(X)` will produce a row vector. It is instructive to take a simple data matrix and inspect step by step how MATLAB calculates the multivariate summaries. For instance,



```
X = [1 2 3; 4 5 6];
[n p]=size(X)  %[2 3]: two 3-dimensional observations
meanX = mean(X)'  %or mean(X,1), along dimension 1
               %transpose of meanX needed to be a column vector
meanX = 1/n * X' * ones(n,1)
```

For any two variables (columns) in X , $x_{(i)}$ and $x_{(j)}$, one can find the sample covariance:

$$s_{ij} = \frac{1}{n-1} \left(\sum_{k=1}^n x_{ki} x_{kj} - n \bar{x}_i \bar{x}_j \right).$$

All s_{ij} s form a $p \times p$ matrix, called a *sample covariance matrix* and denoted by S .

A simple representation for \mathbf{S} uses matrix notation:

$$\mathbf{S} = \frac{1}{n-1} \left(\mathbf{X}'\mathbf{X} - \frac{1}{n} \mathbf{X}'\mathbf{J}\mathbf{X} \right).$$

Here $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is a standard notation for a matrix consisting of ones. If one defines a *centering matrix* \mathbf{H} as $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{J}$, then $\mathbf{S} = \frac{1}{n-1} \mathbf{X}'\mathbf{H}\mathbf{X}$. Here \mathbf{I} is the identity matrix.



```
X = [1 2 3; 4 5 6];
[n p]=size(X);
J = ones(n,1)*ones(1,n);
H = eye(n) - 1/n * J;
S = 1/(n-1) * X' * H * X
S = cov(X) %built-in command
```

An alternative definition of the covariance matrix, $\mathbf{S}^* = \frac{1}{n} \mathbf{X}'\mathbf{H}\mathbf{X}$, is coded in MATLAB as `cov(X,1)`. Note also that the diagonal of \mathbf{S} contains sample variances of variables since $s_{ii} = \frac{1}{n-1} (\sum_{k=1}^n x_{ki}^2 - n\bar{x}_i^2) = s_i^2$.

Matrix \mathbf{S} describes scattering in data matrix \mathbf{X} . Sometimes it is convenient to have scalars as measures of scatter, and for that purpose two summaries of \mathbf{S} are typically used: (i) the determinant of \mathbf{S} , $|\mathbf{S}|$, as a generalized variance and (ii) the trace of \mathbf{S} , $\text{tr}\mathbf{S}$, as the total variation.

The sample correlation coefficient between the i th and j th variables is

$$r_{ij} = \frac{s_{ij}}{s_i s_j},$$

where $s_i = \sqrt{s_i^2} = \sqrt{s_{ii}}$ is the sample standard deviation. Matrix \mathbf{R} with elements r_{ij} is called a sample correlation matrix. If $\mathbf{R} = \mathbf{I}$, the variables are uncorrelated. If $\mathbf{D} = \text{diag}(s_i)$ is a diagonal matrix with (s_1, s_2, \dots, s_p) on its diagonal, then

$$\mathbf{S} = \mathbf{D}\mathbf{R}\mathbf{D}, \quad \mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}.$$

Next we show how to standardize multivariate data. Data matrix \mathbf{Y} is a standardized version of \mathbf{X} if its rows \mathbf{y}_i' are standardized rows of \mathbf{X} ,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix}, \quad \text{where } \mathbf{y}_i = \mathbf{D}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, \dots, n.$$


\mathbf{Y} has a covariance matrix equal to the correlation matrix. This is a multivariate version of the z-score. For the two-column vectors from \mathbf{Y} , $\mathbf{y}_{(i)}$ and $\mathbf{y}_{(j)}$, the correlation r_{ij} can be interpreted geometrically as the cosine of angle φ_{ij} between the vectors. This shows that correlation is a measure of similarity

because close vectors (with a small angle between them) will be strongly positively correlated, while the vectors orthogonal in the geometric sense will be uncorrelated. This is why uncorrelated vectors are sometimes called orthogonal.

Another useful transformation of multivariate data is the Mahalanobis transformation. When data are transformed by the Mahalanobis transformation, the variables become decorrelated. For this reason, such transformed data are sometimes called “sphericized.”

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix}, \quad \text{where } \mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, \dots, n.$$

The Mahalanobis transform decorrelates the components, so $\text{Cov}(\mathbf{Z})$ is an identity matrix. The Mahalanobis transformation is useful in defining the distances between multivariate observations. For further discussion on the multivariate aspects of statistics we direct the student to the excellent book by Morrison (1976).

Example 2.5.  The Fisher iris data set was a data matrix of size 150×4 , while the size of the body fat data was 252×19 . To illustrate some of the multivariate summaries just discussed we construct a new, 5 dimensional data matrix from the body fat data set. The selected columns are `broz`, `densi`, `weight`, `adiposi`, and `biceps`. All 252 rows are retained.



```
X = [broz densi weight adiposi biceps];
varNames = {'broz'; 'densi'; 'weight'; 'adiposi'; 'biceps'};

varNames =
    'broz'    'densi'    'weight'    'adiposi'    'biceps'

Xbar = mean(X)

Xbar = 18.9385 1.0556 178.9244 25.4369 32.2734

S = cov(X)

S =
    60.0758   -0.1458   139.6715   20.5847   11.5455
   -0.1458    0.0004   -0.3323   -0.0496   -0.0280
   139.6715   -0.3323   863.7227   95.1374   71.0711
    20.5847   -0.0496    95.1374   13.3087    8.2266
    11.5455   -0.0280    71.0711    8.2266    9.1281

R = corr(X)
```

```

R =
    1.0000    -0.9881     0.6132     0.7280     0.4930
   -0.9881     1.0000    -0.5941    -0.7147    -0.4871
     0.6132    -0.5941     1.0000     0.8874     0.8004
     0.7280    -0.7147     0.8874     1.0000     0.7464
     0.4930    -0.4871     0.8004     0.7464     1.0000

% By 'hand'
[n p]=size(X);
H = eye(n) - 1/n * ones(n,1)*ones(1,n);
S = 1/(n-1) * X' * H * X;
stds = sqrt(diag(S));
D = diag(stds);
R = inv(D) * S * inv(D);
%S and R here coincide with S and R
%calculated by built-in functions cov and cor.

Xc= X - repmat(mean(X),n,1); %center X
%subtract component means
%from variables in each observation.

%standardization
Y = Xc * inv(D); %for Y, S=R

%Mahalanobis transformation
M = sqrtm(inv(S)) %sqrtm is a square root of matrix

%M =
%    0.1739     0.8423    -0.0151    -0.0788     0.0046
%    0.8423   345.2191    -0.0114     0.0329     0.0527
%   -0.0151    -0.0114     0.0452    -0.0557    -0.0385
%   -0.0788     0.0329    -0.0557     0.6881    -0.0480
%    0.0046     0.0527    -0.0385    -0.0480     0.5550

Z = Xc * M; %Z has uncorrelated components
cov(Z)      %should be identity matrix

```

Figure 2.13 shows data plots for a subset of five variables and the two transformations, standardizing and Mahalanobis. Panel (a) shows components *broz*, *densi*, *weight*, *adiposi*, and *biceps* over all 252 measurements. Note that the scales are different and that *weight* has much larger magnitudes than the other variables.

Panel (b) shows the standardized data. All column vectors are centered and divided by their respective standard deviations. Note that the data plot here shows the correlation across the variables. The variable *density* is negatively correlated with the other variables.

Panel (c) shows the decorrelated data. Decorrelation is done by centering and multiplying by the Mahalanobis matrix, which is the matrix square root of the inverse of the covariance matrix. The correlations visible in panel (b) disappeared.



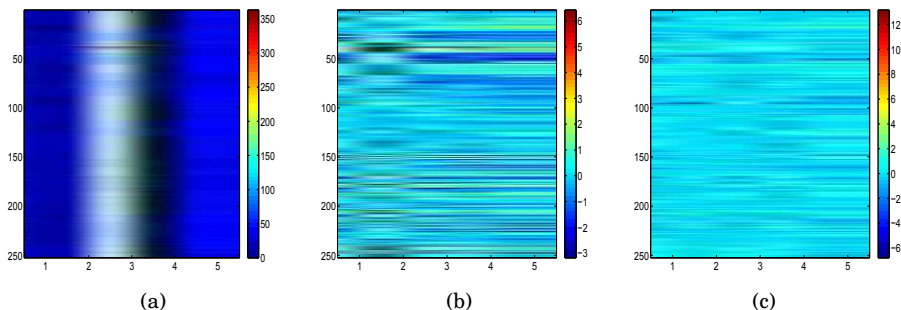


Fig. 2.13 Data plots for (a) 252 five-dimensional observations from Body Fat data where the variables are *broz*, *densi*, *weight*, *adiposi*, and *biceps*. (b) \mathbf{Y} is standardized \mathbf{X} , and (c) \mathbf{Z} is a decorrelated \mathbf{X} .

2.8 Visualizing Multivariate Data

The need for graphical representation is much greater for multivariate data than for univariate data, especially if the number of dimensions exceeds three.

For a data given in matrix form (observations in rows, components in columns), we have already seen a quite an illuminating graphical representation, which we called a data matrix.

One can extend the histogram to bivariate data in a straightforward manner. An example of a 2-D histogram obtained by m-file `hist2d` is given in Fig. 2.14a. The histogram (in the form of an image) shows the sepal and petal lengths from the `fisheriris` data set. A scatterplot of the 2-D measurements is superimposed.

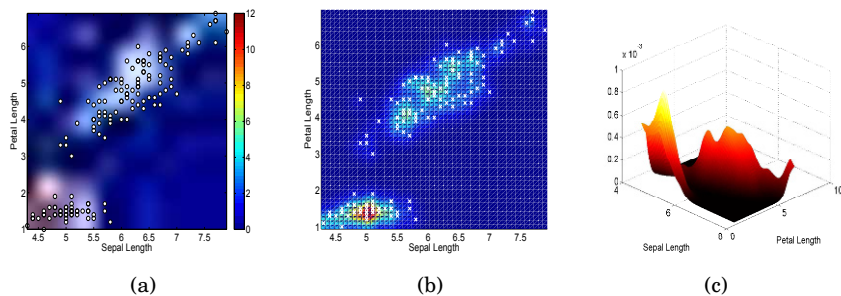





Fig. 2.14 (a) Two-dimensional histogram of Fisher's iris sepal (X) and petal (Y) lengths. The plot is obtained by `hist2d.m`; (b) Scattercloud plot – smoothed histogram with superimposed scatterplot, obtained by `scattercloud.m`; (c) Kernel-smoothed and normalized histogram obtained by `smoothhist2d.m`.

Figures 2.14b-c show the smoothed histograms. The histogram in panel (c) is normalized so that the area below the surface is 1. The smoothed histograms are plotted by  `scattercloud.m` and  `smoothhist2d.m` (S. Simon and E. Ronchi, MATLAB Central).

If the dimension of the data is three or more, one can gain additional insight by plotting pairwise scatterplots. This is achieved by the MATLAB command `gplotmatrix(X,Y,group)`, which creates a matrix arrangement of scatterplots. Each subplot in the graphical output contains a scatterplot of one column from data set X against a column from data set Y .

In the case of a single data set (as in body fat and Fisher iris examples), Y is omitted or set at $Y=[]$, and the scatterplots contrast the columns of X . The plots can be grouped by the grouping variable `group`. This variable can be a categorical variable, vector, string array, or cell array of strings.

The variable `group` must have the same number of rows as X . Points with the same value of `group` appear on the scatterplot with the same marker and color. Other arguments in `gplotmatrix(x,y,group,clr,sym,siz)` specify the color, marker type, and size for each group. An example of the `gplotmatrix` command is given in the code below. The output is shown in Fig. 2.15a.

```
 X = [broz densi weight adiposi biceps];
varNames = {'broz'; 'densi'; 'weight'; 'adiposi'; 'biceps'};
agegr = age > 55;
gplotmatrix(X,[],agegr,['b','r'],['x','o'],[],'false');
text([.08 .24 .43 .66 .83], repmat(-.1,1,5), varNames, ...
     'FontSize',8);
text(repmat(-.12,1,5), [.86 .62 .41 .25 .02], varNames, ...
     'FontSize',8, 'Rotation',90);
```

Parallel Coordinates Plots. In a *parallel coordinates plot*, the components of the data are plotted on uniformly spaced vertical lines called component axes. A p -dimensional data vector is represented as a broken line connecting a set of points, one on each component axis. Data represented as lines create readily perceived structures. A command for parallel coordinates plot `parallelcoords` is given below with the output shown in Fig. 2.15b.



```
 parallelcoords(X, 'group', age>55, ...
               'standardize','on', 'labels',varNames)
set(gcf,'color','white');
```

Figure 2.16a shows parallel cords for the groups $\text{age} > 55$ and $\text{age} \leq 55$ with 0.25 and 0.75 quantiles.

```
 parallelcoords(X, 'group', age>55, ...
               'standardize','on', 'labels',varNames,'quantile',0.25)
set(gcf,'color','white');
```

Andrews' Plots. An *Andrews plot* (Andrews, 1972) is a graphical representation that utilizes Fourier series to visualize multivariate data. With an

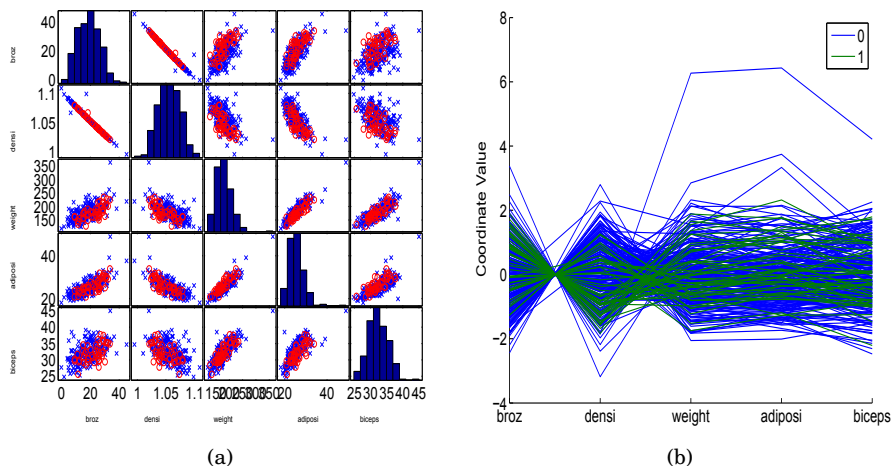


Fig. 2.15 (a) `gplotmatrix` for `broz`, `densi`, `weight`, `adiposi`, and `biceps`; (b) `parallelcoords` plot for X , by `age>55`.

observation (X_1, \dots, X_p) one associates the function

$$F(t) = X_1/\sqrt{2} + X_2 \sin(2\pi t) + X_3 \cos(2\pi t) + X_4 \sin(2 \cdot 2\pi t) + X_5 \cos(2 \cdot 2\pi t) + \dots,$$

where t ranges from -1 to 1 . One Andrews' curve is generated for each multivariate datum – a row of the data set. Andrews' curves preserve the distances between observations. Observations close in the Euclidian distance sense are represented by close Andrews' curves. Hence, it is easy to determine which observations (i.e., rows when multivariate data are represented as a matrix) are most alike by using these curves. Due to the definition, this representation is not robust with respect to the permutation of coordinates. The first few variables tend to dominate, so it is a good idea when using Andrews' plots to put the most important variables first. Some analysts recommend running a principal components analysis first and then generating Andrews' curves for principal components. The principal components of multivariate data are linear combinations of components that account for most of the variability in the data. Principal components will not be discussed in this text as they are beyond the scope of this course.

An example of Andrews' plots is given in the code below with the output in Fig. 2.16b.



```
andrewsplot(X, 'group', age>55, 'standardize','on')
set(gcf,'color','white');
```

Star Plots. The star plot is one of the earliest multivariate visualization objects. Its rudiments can be found in the literature from the early nineteenth

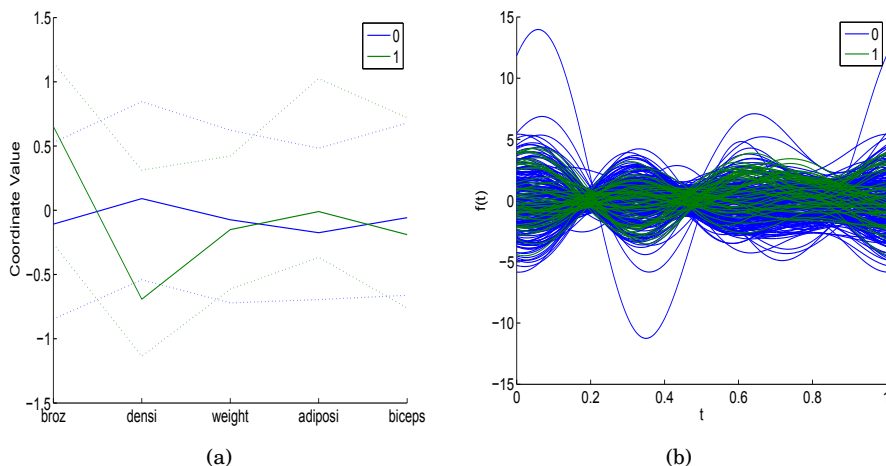


Fig. 2.16 (a) X by `age>55` with quantiles; (b) `andrewsplot` for X by `age>55`.

century. Similar plots (rose diagrams) are used in Florence Nightingale's *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army* in 1858 (Nightingale, 1858).

The star glyph consists of a number of spokes (rays) emanating from the center of the star plot and connected at the ends. The number of spikes in the star plot is equal to the number of variables (components) in the corresponding multivariate datum. The length of each spoke is proportional to the magnitude of the component it represents. The angle between two neighboring spokes is $2\pi/p$, where p is the number of components. The star glyph connects the ends of the spokes.

An example of the use of star plots is given in the code below with the output in Fig. 2.17a.



```
ind = find(age>67);
strind = num2str(ind);
h = glyphplot(X(ind,:), 'glyph','star', 'varLabels',...
    varNames,'obslabels', strind);
set(h(:,3),'FontSize',8); set(gcf,'color','white');
```

Chernoff Faces. People grow up continuously studying faces. Minute and barely measurable differences are easily detected and linked to a vast catalog stored in memory. The human mind subconsciously operates as a super computer, filtering out insignificant phenomena and focusing on the potentially important. Such mundane characters as `:`, `:`, `:`, `:`, `:`, and `>:p` are readily linked in our minds to joy, dissatisfaction, shock, or affection.

Face representation is an interesting approach to taking a first look at multivariate data and is effective in revealing complex relations that are not visible in simple displays that use the magnitudes of components. It can be used

to aid in cluster analysis and discrimination analysis and to detect substantial changes in time series.

Each variable in a multivariate datum is connected to a feature of a face. The variable-feature links in MATLAB are as follows: variable 1 – size of face; variable 2 – forehead/jaw relative arc length; variable 3 – shape of forehead; variable 4 – shape of jaw; variable 5 – width between eyes; variable 6 – vertical position of eyes; variables 7–13 – features connected with location, separation, angle, shape, and width of eyes and eyebrows; and so on. An example of the use of Chernoff faces is given in the code below with the output in Fig. 2.17b.



```
ind = find(height > 74.5);
strind = num2str(ind);
h = glyphplot(X(ind,:), 'glyph','face', 'varLabels',...
varNames,'obslabels', strind);
set(h(:,3),'FontSize',10); set(gcf,'color','white');
```

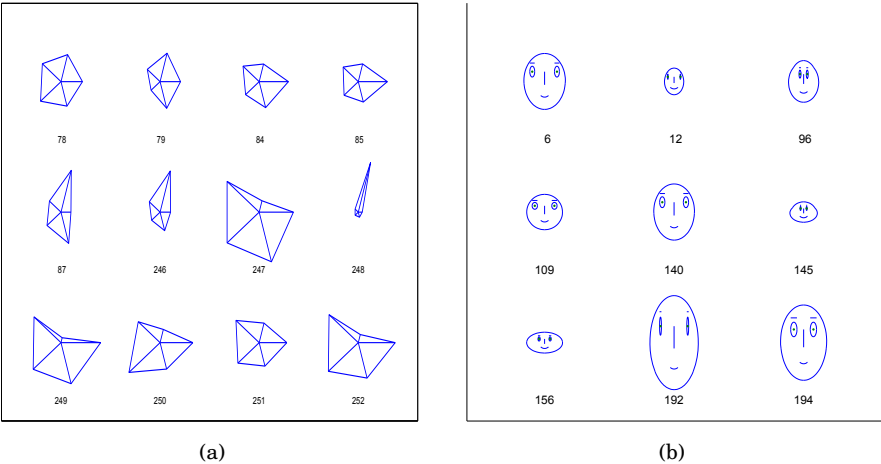




Fig. 2.17 (a) Star plots for X ; (b) Chernoff faces plot for X .

2.9 Observations as Time Series

Observations that have a time index, that is, if they are taken at equally spaced instances in time, are called *time series*. EKG and EEG signals, high-frequency bioresponses, sound signals, economic indices, and astronomic and geophysical measurements are all examples of time series. The following example illustrates a time series.

Example 2.6. Blowflies Time Series. The data set  `blowflies.dat` consists of the total number of blowflies (*Lucilia cuprina*) in a population under controlled laboratory conditions. The data represent counts for every other day. The developmental delay (from egg to adult) is between 14 and 15 days for insects under the conditions employed. Nicholson (1954) made 361 bi-daily recordings over a 2-year period (722 days), see [Fig. 2.18a](#). 

In addition to analyzing basic location, spread, and graphical summaries, we are also interested in evaluating the degree of autocorrelation in time series. Autocorrelation measures the level of correlation of the time series with a time-shifted version of itself. For example, autocorrelation at lag 2 would be a correlation between $X_1, X_2, X_3, \dots, X_{n-3}, X_{n-2}$ and $X_3, X_4, \dots, X_{n-1}, X_n$. When the shift (lag) is 0, the autocorrelation is just a correlation. The concept of autocorrelation is introduced next, and then the autocorrelation is calculated for the blowflies data.

Let X_1, X_2, \dots, X_n be a sample where the order of observations is important. The indices $1, 2, \dots, n$ may correspond to measurements taken at time points $t, t + \Delta t, t + 2\Delta t, \dots, t + (n-1)\Delta t$, for some start time t and time increments Δt . The autocovariance at lag $0 \leq k \leq n-1$ is defined as

$$\hat{\gamma}(k) = \frac{1}{n} \sum_{i=1}^{n-k} (X_{i+k} - \bar{X})(X_i - \bar{X}).$$

Note that the sum is normalized by a factor $\frac{1}{n}$ and not by $\frac{1}{n-k}$, as one may expect.

The autocorrelation is defined as normalized autocovariance,

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}.$$

Autocorrelation is a measure of self-affinity of the time series with its own shifts and is an important summary statistic. MATLAB has the built-in functions `autocov` and `autocorr`. The following two functions are simplified versions illustrating how the autocovariances and autocorrelations are calculated.



```
function acv = acov(ts, maxlag)
%acov.m: computes the sample autocovariance function
%          ts      = 1-D time series
%          maxlag = maximum lag (< length(ts))
%usage: z = autocov (a,maxlag);
n = length(ts);
ts = ts(:) - mean(ts); %note overall mean
suma = zeros(n,maxlag+1);
suma(:,1) = ts.^2;
for h = 2:maxlag+1
    suma(1:(n-h+1), h) = ts(h:n);
```

```

    suma(:,h) = suma(:,h) .* ts;
end
acv = sum(suma)/n; %note the division by n
                    %and not by expected (n-h)

function [acrr] = acorr(ts , maxlag)
    acr = acov(ts, maxlag);
    acrr = acr ./ acr(1);

```

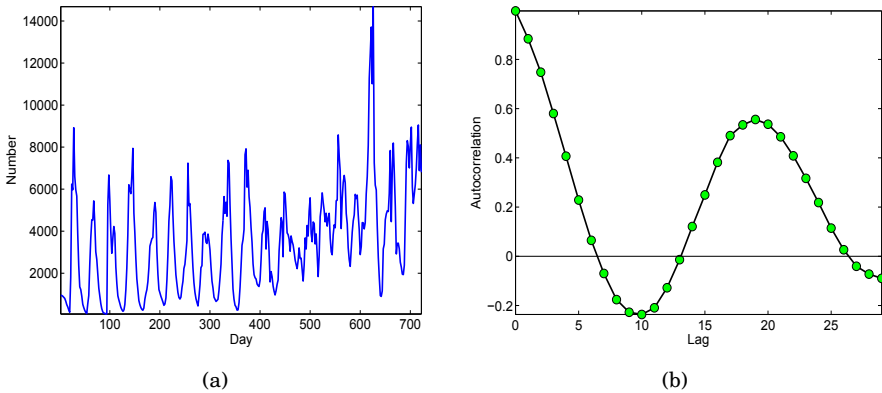


Fig. 2.18 (a) Bi-daily measures of size of the blowfly population over a 722-day period, (b) The autocorrelation function of the time series. Note the peak at lag 19 corresponding to the periodicity of 38 days.

Figure 2.18a shows the time series illustrating the size of the population of blowflies over 722 days. Note the periodicity in the time series. In the autocorrelation plot (Fig. 2.18b) the peak at lag 19 corresponding to a time shift of 38 days. This indicates a periodicity with an approximate length of 38 days in the dynamic of this population. A more precise assessment of the periodicity and related inference can be done in the frequency domain of a time series, but this theory is beyond the scope of this course. Good follow-up references are Brillinger (2001), Brockwell and Davis (2009), and Shumway and Stoffer (2005). Also see Exercise 2.12.

2.10 About Data Types

The cell data elaborated in this chapter are *numerical*. When measurements are involved, the observations are typically *numerical*. Other types of data encountered in statistical analysis are categorical. Stevens (1946), who was influenced by his background in psychology, classified data as nominal, ordinal, interval, and ratio. This typology is loosely accepted in other scientific cir-

cles. However, there are vibrant and ongoing discussions and disagreements, e.g., Veleman and Wilkinson (1993). *Nominal data*, such as race, gender, political affiliation, names, etc., cannot be ordered. For example, the counties in northern Georgia, Cherokee, Clayton, Cobb, DeCalb, Douglas, Fulton, and Gwinnett, cannot be ordered except that there is a nonessential alphabetical order of their names. Of course, numerical attributes of these counties, such as size, area, revenue, etc., can be ordered.

Ordinal data could be ordered and sometimes assigned numbers, although the numbers would not convey their relative standing. For example, data on the Likert scale have five levels of agreement: (1) Strongly Disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree; the numbers 1 to 5 are assigned to the degree of agreement and have no quantitative meaning. The difference between Agree and Neutral is not equal to the difference between Disagree and Strongly Disagree. Other examples are the attributes “Low” and “High” or student grades A, B, C, D, and F. It is an error to treat ordinal data as numerical. Unfortunately this is a common mistake (e.g., GPA). Sometimes T-shirt-size attributes, such as “small,” “medium,” “large,” and “x-large,” may falsely enter the model as if they were measurements 1, 2, 3, and 4.

Nominal and ordinal data are examples of *categorical* data since the values fall into categories.

Interval data refers to numerical data for which the differences can be well interpreted. However, for this type of data, the origin is not defined in a natural way so the ratios would not make sense. Temperature is a good example. We cannot say that a day in July with a temperature of 100°F is twice as hot as a day in November with a temperature of 50°F. Test scores are another example of interval data as a student who scores 100 on a midterm may not be twice as good as a student who scores 50.

Ratio data are at the highest level; these are usually standard numerical values for which ratios make sense and the origin is absolute. Length, weight, and age are all examples of ratio data.

Interval and ratio data are examples of *numerical* data.

MATLAB provides a way to keep such heterogeneous data in a single structure array with a syntax resembling C language.

Structures are arrays comprised of structure elements and are accessed by named fields. The fields (data containers) can contain any type of data. Storage in the structure is allocated dynamically. The general syntax for a structure format in MATLAB is `structurename(recordnumber).fieldname=data`

For example,




```
patient.name = 'John Doe';
patient.agegroup = 3;
patient.billing = 127.00;
patient.test = [79 75 73; 180 178 177.5; 220 210 205];
patient
%To expand the structure array, add subscripts.
patient(2).name = 'Ann Lane';
```

```
patient(2).agegroup = 2;
patient(2).billing = 208.50;
patient(2).test = [68 70 68; 118 118 119; 172 170 169];
patient
```

2.11 Exercises

- 2.1. Auditory Cortex Spikes.** This data set comes from experiments in the lab of Dr. Robert Liu of Emory University⁴ and concerns single-unit electrophysiology in the auditory cortex of nonanesthetized female mice. The motivating question is the exploration of auditory neural differences between female parents vs. female virgins and their relationship to cortical response.

Researchers in Liu's lab developed a restrained awake setup to collect single neuron activity from both female parent and female naïve mice. Multiple trials are performed on the neurons from one mother and one naïve animal.

The recordings are made from a region in the auditory cortex of the mouse with a single tungsten electrode. A sound stimulus is presented at a time of 200 ms during each sweep (time shown is 0–611 and 200 is the point at which a stimulus is presented). Each sweep is 611 ms long and the duration of the stimulus tone is 10 to 70 ms. The firing times for mother and naïve mice are provided in the data set  `spikes.dat`, in columns 2 and 3. Column 1 is the numbering from 1 to 611.

(a) Using MATLAB's `diff` command, find the inter-firing times. Plot a histogram for both sets of interfiring times. Use `biplot.m` to plot the histograms back to back.

(b) For inter-firing times in the mother's response find descriptive statistics similar to those in the cell area example.

- 2.2. On Average.** It is an anecdotal truth that an average Australian has less than two legs! Indeed, there are some Australians that have lost their leg(s); thus the number of legs is less than twice the number of people. In this exercise, we compare several sample averages.

A small company reports the following salaries: 4 employees at 20K, 3 employees at 30K, the vice-president at 200K, and the president at 400K. Calculate the arithmetic mean, geometric mean, median, harmonic mean, and mode. If the company is now hiring, would an advertising strategy in which the mean salary is quoted be fair? If not, suggest an alternative.

- 2.3. Contraharmonic Mean and f -Mean.** The contraharmonic mean for X_1, X_2, \dots, X_n is defined as

⁴ <http://www.biology.emory.edu/research/Liu/index.html>

$$C(X_1, \dots, X_n) = \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n X_i}.$$

(a) Show that $C(X_1, X_2)$ is twice the sample mean minus the harmonic mean of X_1, X_2 .

(b) Show that $C(x, x, x, \dots, x) = x$.

The generalized f -mean of X_1, \dots, X_n is defined as

$$X_f = f^{-1} \left(\frac{1}{n} \sum_{i=1}^n f(X_i) \right),$$

where f is suitably chosen such that $f(X_i)$ and f^{-1} are well defined.

(c) Show that $f(x) = x, \frac{1}{x}, x^k, \log x$ gives the mean, harmonic mean, power k mean, and geometric mean.

2.4. Mushrooms. The unhappy outcome of uninformed mushroom picking is poisoning. In many cases, such poisoning is due to ignorance or a superficial approach to identification. The most dangerous fungi are Death Cap (*Amanita phalloides*) and two species akin to it, *A. verna* and Destroying Angel (*A. virosa*). These three toadstools cause the majority of fatal poisoning.

One of the keys to mushroom identification is the spore deposit. Spores of *Amanita phalloides* are colorless, nearly spherical, and smooth. Measurements in microns of 28 spores are given below:

9.2	8.8	9.1	10.1	8.5	8.4	9.3
8.7	9.7	9.9	8.4	8.6	8.0	9.5
8.8	8.1	8.3	9.0	8.2	8.6	9.0
8.7	9.1	9.2	7.9	8.6	9.0	9.1

(a) Find the *five-number summary* (Min, Q_1, Me, Q_3, Max) for the spore measurement data.

(b) Find the mean and the mode.

(c) Find and plot the histogram of z -scores, $z_i = (X_i - \bar{X})/s$.

2.5. Manipulations with Sums. Prove the following algebraic identities involving sums, useful in demonstrating properties of some sample summaries.

(a) $\sum_{i=1}^n (x_i - \bar{x}) = 0$	(b) If $y_1 = x_1 + a, y_2 = x_2 + a, \dots, y_n = x_n + a$, then $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$
(c) If $y_1 = c \cdot x_1, y_2 = c \cdot x_2, \dots, y_n = c \cdot x_n$, then $\sum_{i=1}^n (y_i - \bar{y})^2 = c^2 \sum_{i=1}^n (x_i - \bar{x})^2$	(d) If $y_1 = c \cdot x_1 + a, y_2 = c \cdot x_2 + a, \dots, y_n = c \cdot x_n + a$, then $\sum_{i=1}^n (y_i - \bar{y})^2 = c^2 \sum_{i=1}^n (x_i - \bar{x})^2$
(e) $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$	(f) $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n(\bar{x})(\bar{y})$
(g) $\sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - a)^2$	(h) For any constant a , $\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$

- 2.6. Emergency Calculation.** Graduate student Rosa Juliusdottir reported the results of an experiment to her advisor who wanted to include them in his grant proposal. Before leaving to Reykjavik for a short vacation, she left the following data in her advisor's mailbox: sample size $n = 12$, sample mean $\bar{X} = 15$, and sample variance $s^2 = 34$.

The advisor noted with horror that the last measurement X_{12} was wrongly recorded. It should have been 16 instead of 4. It would be easy to fix \bar{X} and s^2 , but the advisor did not have the previous 11 measurements nor the statistics training necessary to make the correction. Rosa was in Iceland, and the grant proposal was due the next day. The advisor was desperate, but luckily you came along.

- 2.7. Sample Mean and Standard Deviation After a Change.** It is known that $\bar{y} = 11.6$, $s_y = 4.4045$, and $n = 15$. The observation $y_{12} = 7$ is removed and observation y_{13} was misreported; it was not 10, but 20. Find \bar{y}_{new} and $s_{y(new)}$ after the changes.

- 2.8. Surveys on Different Scales.** We are interested in determining whether UK voters (whose parties have somewhat more distinct policy positions than those in the USA) have a wider variation in their evaluations of the parties than voters in the USA. The problem is that the British election survey takes evaluations scored 0–10, while the US National Election Survey gets evaluations scored 0–100. Here are two surveys.

UK	6	7	5	10	3	9	9	6	8	2	7	5
US	67	65	95	86	44	100	85	92	91	65		

Using CV compare the amount of variation without worrying about the different scales.

- 2.9. Merging Two Samples.** Suppose \bar{X} and s_X^2 are the mean and variance of the sample X_1, \dots, X_m and \bar{Y} and s_Y^2 of the sample Y_1, \dots, Y_n . If the two samples are merged into a single sample, show that its mean and variance are


$$\frac{m\bar{X} + n\bar{Y}}{m + n} \quad \text{and} \quad \frac{1}{m + n - 1} \left[(m - 1)s_X^2 + (n - 1)s_Y^2 + \frac{mn}{m + n}(\bar{X} - \bar{Y})^2 \right].$$

- 2.10. Fitting the Histogram.** The following is a demonstration of MATLAB's built-in function `histfit` on a simulated data set.



```
dat = normrnd(4, 1,[1 500]) + normrnd(2, 3,[1 500]);
figure; histfit(dat(:));
```

The function `histfit` plots the histogram of data and overlays it with the best fitting Gaussian curve. As an exercise, take Brozek index `broz` from

the data set  `fat.dat` (second column) and apply the `histfit` command. Comment on how the Gaussian curve fits the histogram.

- 2.11. **QT Syndrome.** The QT interval is a time interval between the start of the Q wave and the end of the T wave in a heart's electrical cycle (Fig. 2.19). It measures the time required for depolarization and repolarization to occur. In long QT syndrome, the duration of repolarization is longer than normal, which results in an extended QT interval. An interval above 440 ms is considered prolonged. Although the mechanical function of the heart could be normal, the electrical defects predispose affected subjects to arrhythmia, which may lead to sudden loss of consciousness (syncope) and, in some cases, to a sudden cardiac death.

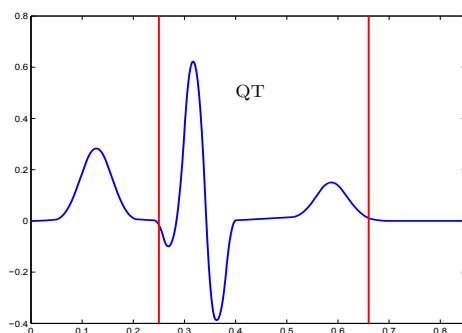



Fig. 2.19 Schematic plot of ECG, with QT time between the red bars.

The data set  `QT.dat` |mat was compiled by Christov et al. (2006) and is described in <http://www.biomedical-engineering-online.com/content/5/1/31>. It provides 548 QT times taken from 293 subjects. The subjects include healthy controls (about 20%) and patients with various diagnoses, such as myocardial infarction, cardiomyopathy/heart failure, bundle branch block, dysrhythmia, myocardial hypertrophy, etc. The Q-onsets and T-wave ends are evaluated by five independent experts, and medians of their estimates are used in calculations of the QT for a subject.

Plot the histogram of this data set and argue that the data are reasonably “bell-shaped.” Find the location and spread measures of the sample. What proportion of this sample has prolonged QT?

- 2.12. **Blowfly Count Time Series.** For the data in Example 2.6 it was postulated that a major transition in the dynamics of blowfly population size appeared to have occurred around day 400. This was attributed to biological evolution, and the whole series cannot be considered as representative of the same system. Divide the time series into two data segments with in-

dices 1–200 and 201–361. Calculate and compare the autocorrelation functions for the two segments.

- 2.13. **Simpson's Diversity Index.** An alternative diversity measure to Shannon's in (2.1) is the Simpson diversity index defined as

$$D = \frac{n^2}{\sum_{i=1}^k f_i^2}.$$

It achieves its maximum k when all frequencies are equal; thus Simpson's homogeneity (equitability) index is defined as $E_D = D/k$.

Repeat the calculations from Example 2.3 with Simpson's diversity and homogeneity indices in place of Shannon's. Is the Brazilian sample still the most homogeneous, as it was according to Shannon's E_H index?

- 2.14. **Speed of Light.** Light travels very fast. It takes about 8 min to reach Earth from the Sun and over 4 years to reach Earth from the closest star outside the solar system. Radio and radar waves also travel at the speed of light, and an accurate value of that speed is important to communicate with astronauts and orbiting satellites. Because of the nature of light, it is very hard to measure its speed. The first reasonably accurate measurements of the speed of light were made by A. Michelson and S. Newcomb. The table below contains 66 transformed measurements made by Newcomb between July and September 1882. Entry 28, for instance, corresponds to the actual measurement of 0.000024828 s. This was the amount of time needed for light to travel approx. 4.65 miles.

28	22	36	26	28	28	26	24	32	30	27
24	33	21	36	32	31	25	24	25	28	36
27	32	34	30	25	26	26	25	−44	23	21
30	33	29	27	29	28	22	26	27	16	31
29	36	32	28	40	19	37	23	32	29	−2
24	25	27	24	16	29	20	28	27	39	23

You can download  light.data|mat and read it in MATLAB.

If we agree that outlier measurements are outside the interval $[Q_1 - 2.5 IQR, Q_3 + 2.5 IQR]$, what observations qualify as outliers? Make the data “clean” by excluding outlier(s). For the cleaned data find the mean, 20% trimmed mean, real MAD, std, and variance.

Plot the histogram and kernel density estimator for an appropriately selected bandwidth.

- 2.15. **Limestone Formations in Jamaica.** This data set contains 18 observations of nummulitid specimens from the Eocene yellow limestone formation in northwestern Jamaica ( limestone.dat). The use of faces to represent points in k -dimensional space graphically was originally illustrated on this

data set (Chernoff, 1973). Represent this data set graphically using Chernoff faces.

ID	Z ₁	Z ₂	Z ₃	Z ₄	Z ₅	Z ₆	ID	Z ₁	Z ₂	Z ₃	Z ₄	Z ₅	Z ₆
1	160	51	10	28	70	450	45	195	32	9	19	110	1010
2	155	52	8	27	85	400	46	220	33	10	24	95	1205
3	141	49	11	25	72	380	81	55	50	10	27	128	205
4	130	50	10	26	75	560	82	70	53	7	28	118	204
6	135	50	12	27	88	570	83	85	49	11	19	117	206
41	85	55	13	33	81	355	84	115	50	10	21	112	198
42	200	34	10	24	98	1210	85	110	57	9	26	125	230
43	260	31	8	21	110	1220	86	95	48	8	27	114	228
44	195	30	9	20	105	1130	87	95	49	8	29	118	240

2.16. **Duchenne Muscular Dystrophy.** *Duchenne muscular dystrophy* (DMD), or Meryon’s disease, is a genetically transmitted disease, passed from a mother to her children (Fig. 2.20). Affected female offspring usually suffer no apparent symptoms and may unknowingly carry the disease. Male offspring with the disease die at a young age. Not all cases of the disease come from an affected mother. A fraction, perhaps one third, of the cases arise spontaneously, to be genetically transmitted by an affected female. This is the most widely held view at present. The incidence of DMD is about 1 in 10,000 male births. The population risk (prevalence) that a woman is a DMD carrier is about 3 in 10,000.

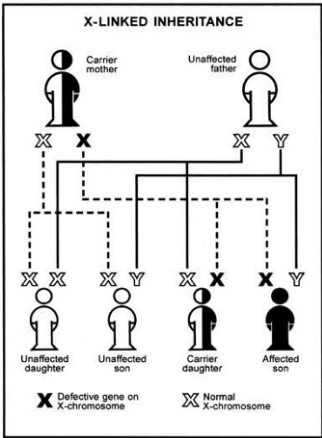


Fig. 2.20 Each son of a carrier has a 50% chance of having DMD and each daughter has a 50% chance of being a carrier.

From the text page download data set dmd.dat|mat|xls. This data set is modified data from Percy et al. (1981) (entries containing missing values

excluded). It consists of 194 observations corresponding to blood samples collected in a project to develop a screening program for female relatives of boys with DMD. The program was implemented in Canada and its goal was to inform a woman of her chances of being a carrier based on serum markers as well as her family pedigree. Another question of interest was whether age should be taken into account. Enzyme levels were measured in known carriers (67 samples) and in a group of noncarriers (127 samples). The first two serum markers, creatine kinase and hemopexin (*ck,h*), are inexpensive to obtain, while the last two, pyruvate kinase and lactate dehydrogenase (*pk,ld*), are expensive.

The variables (columns) in the data set are

Column	Variable	Description
1	<i>age</i>	Age of a woman in the study
2	<i>ck</i>	Creatine kinase level
3	<i>h</i>	Hemopexin
4	<i>pk</i>	Pyruvate kinase
5	<i>ld</i>	Lactate dehydrogenase
6	<i>carrier</i>	Indicator if a woman is a DMD carrier

- Find the mean, median, standard deviation, and *real MAD* of pyruvate kinase level, *pk*, for all cases (*carrier*=1).
- Find the mean, median, standard deviation, and *real MAD* of pyruvate kinase level, *pk*, for all controls (*carrier*=0).
- Find the correlation between variables *pk* and *carrier*.
- Use MATLAB's `gplotmatrix` to visualize pairwise dependencies between the six variables.
- Plot the histogram with 30 bins and smoothed normalized histogram (density estimator) for *pk*. Use `ksdensity`.

- 2.17. **Ashton's Dental Data.** The evolutionary status of fossils (Australopithecinae, Proconsul, etc.) stimulated considerable discussion in the 1950s. Particular attention has been paid to the teeth of the fossils, comparing their overall dimensions with those of human beings and of the extant great apes. As "controls" measurements have been taken on the teeth of three types of modern man (British, West African native, Australian aboriginal) and of the three living great apes (gorilla, orangutan, and chimpanzee). The data in the table below are taken from Ashton et al. (1957), p. 565, who used 2-D projections to compare the measurements. Andrews (1972) also used an excerpt of these data to illustrate his methodology. The values in the table are not the original measurements but the first eight *canonical variables* produced from the data in order to maximize the sum of distances between different pairs of populations.

A. West African	-8.09	0.49	0.18	0.75	-0.06	-0.04	0.04	0.03
B. British	-9.37	-0.68	-0.44	-0.37	0.37	0.02	-0.01	0.05
C. Au. aboriginal	-8.87	1.44	0.36	-0.34	-0.29	-0.02	-0.01	-0.05
D. Gorilla: male	6.28	2.89	0.43	-0.03	0.10	-0.14	0.07	0.08
E. Female	4.82	1.52	0.71	-0.06	0.25	0.15	-0.07	-0.10
F. Orangutan: Male	5.11	1.61	-0.72	0.04	-0.17	0.13	0.03	0.05
G. Female	3.60	0.28	-1.05	0.01	-0.03	-0.11	-0.11	-0.08
H. Chimpanzee: male	3.46	-3.37	0.33	-0.32	-0.19	-0.04	0.09	0.09
I. Female	3.05	-4.21	0.17	0.28	0.04	0.02	-0.06	-0.06
J. <i>Pithecanthropus</i>	-6.73	3.63	1.14	2.11	-1.90	0.24	1.23	-0.55
K. <i>pekinensis</i>	-5.90	3.95	0.89	1.58	-1.56	1.10	1.53	0.58
L. <i>Paranthropus robustus</i>	-7.56	6.34	1.66	0.10	-2.23	-1.01	0.68	-0.23
M. <i>Paranthropus crassidens</i>	-7.79	4.33	1.42	0.01	-1.80	-0.25	0.04	-0.87
N. <i>Meganthropus paleojavanicus</i>	-8.23	5.03	1.13	-0.02	-1.41	-0.13	-0.28	-0.13
O. <i>Proconsul africanus</i>	1.86	-4.28	-2.14	-1.73	2.06	1.80	2.61	2.48

Andrews (1972) plotted curves over the range $-\pi < t < \pi$ and concluded that the graphs clearly distinguished humans, the gorillas and orangutans, the chimpanzees, and the fossils. Andrews noted that the curve for a fossil (*Proconsul africanus*) corresponds to a plot inconsistent with that of all other fossils as well as humans and apes.

Graphically present this data using (a) star plots, (b) Andrews plots, and (c) Chernoff faces.

- 2.18. **Andrews Plots of Iris Data.** Fisher iris data are 4-D, and Andrews plots can be used to explore clustering of the three species (*Setosa*, *Versicolor*, and *Virginica*). Discuss the output from the code below.



```
load fisheriris
andrewsplot(meas,'group',species);
```

What species clearly separate? What species are more difficult to separate?

- 2.19. **Cork Boring Data.** Cork is the bark of the cork oak (*Quercus suber L.*), a noble tree with very special characteristics that grows in the Mediterranean. This natural tissue has unique qualities: light weight, elasticity, insulation and impermeability, fire retardancy, resistance to abrasion, etc. The data measuring cork boring of trees given in Rao (1948) consist of the weights (in centigrams) of cork boring in four directions (north, east, south, and west) for 28 trees. Data given in Table 2.1 can also be found in




cork.dat|mat.

- Graphically display the data as a data plot, pairwise scatterplots, Andrews plot, and Chernoff faces.
 - Find the mean \bar{x} and covariance matrix S for this data set. Find the trace and determinant of S .
 - Find the Mahalanobis transformation for these data. Check that the covariance matrix for the transformed data is identity.
- 2.20. **Balance.** When a human experiences a balance disturbance, muscles throughout the body are activated in a coordinated fashion to maintain an

Table 2.1 Rao’s data. Weights of cork boring in four directions (north, east, south, west) for 28 trees.

Tree	N	E	S	W	Tree	N	E	S	W
1	72	66	76	77	15	91	79	100	75
2	60	53	66	63	16	56	68	47	50
3	56	57	64	58	17	79	65	70	61
4	41	29	36	38	18	81	80	68	58
5	32	32	35	36	19	78	55	67	60
6	30	35	34	26	20	46	38	37	38
7	39	39	31	27	21	39	35	34	37
8	42	43	31	25	22	32	30	30	32
9	37	40	31	25	23	60	50	67	54
10	33	29	27	36	24	35	37	48	39
11	32	30	34	28	25	39	36	39	31
12	63	45	74	63	26	50	34	37	40
13	54	46	60	52	27	43	37	39	50
14	47	51	52	43	28	48	54	57	43

upright stance. Researchers at Lena Ting Laboratory for Neuroengineering at Georgia Tech are interested in uncovering the sensorimotor mechanisms responsible for coordinating this automatic postural response (APR). Their approach was to perturb the balance of a human subject standing upon a customized perturbation platform that translates in the horizontal plane. Platform motion characteristics spanned a range of peak velocities (5 cm/s steps between 25 and 40 cm/s) and accelerations (0.1g steps between 0.2 and 0.4 g). Five replicates of each perturbation type were collected during the experimental sessions. Surface electromyogram (EMG) signals, which indicate the level of muscle activation, were collected at 1080 Hz from 11 muscles in the legs and trunk.

The data in  `balance2.mat` are processed EMG responses to backward-directed perturbations in the medial gastrocnemius muscle (an ankle plan-tar flexor located on the calf) for all experimental conditions. There is 1 s of data, beginning at platform motion onset. There are 5 replicates of length 1024 each collected at 12 experimental conditions (4 velocities crossed with 3 accelerations), so the data set is 3-D $1024 \times 5 \times 12$.

For example, `data(:,1,4)` is an array of 1024 observations corresponding to first replicate, under the fourth experimental condition (30 cm/s, 0.2g). Consider a fixed acceleration of 0.2g and only the first replicate. Form 1024 4-D observations (velocities 25, 30, 35, and 40 as variables) as a data ma-trix. For the first 16 observations find multivariate graphical summaries using MATLAB’s `gplotmatrix`, `parallelcoords`, `andrewsplot`, and `glyphplot`.

2.21. Cats. Cats are often used in studies about locomotion and injury recovery. In one such study, a bundle of nerves in a cat’s legs were cut and then surgically repaired. This mimics the surgical correction of injury in people. The recovery process of these cats was then monitored. It was monitored quantitatively by walking a cat across a plank that has force plates, as well

as by monitoring various markers inside the leg. These markers provided data for measures such as joint lengths and joint moments. A variety of data was collected from three different cats: Natasha, Riga, and Korina. Natasha (cat = 1) has 47 data entries, Riga (cat = 2) has 39 entries, and Korina (cat = 3) has 35 entries.

The measurements taken are the number of steps for each trial, the length of the stance phase (in milliseconds), the hip height (in meters), and the velocity (in meters/second). The researchers observe these variables for different reasons. They want uniformity both within and between samples (to prevent confounding variables) for steps and velocity. The hip height helps monitor the recovery process. A detailed description can be found in Farrell et al. (2009).

The data set, courtesy of Dr. Boris Prilutsky, School of Applied Physiology at, Georgia Tech, is given as the MATLAB structure file  `cats.mat`.

Form a data matrix

```
X = [cat.nsteps cat.stancedur cat.hipheight cat.velocity cat.cat];
```

and find its mean and correlation matrix. Form matrix Z by standardizing the columns of X (use `zscore`). Plot the image of the standardized data matrix.

- 2.22. **BUPA Liver Data.** The BUPA liver disorders database (courtesy of Richard Forsyth, BUPA Medical Research Ltd.) consists of 345 records of male individuals. Each record has 7 attributes,

Attribute	Name	Meaning
1	<code>mcv</code>	Mean corpuscular volume
2	<code>alkphos</code>	Alkaline phosphatase
3	<code>sgpt</code>	Alamine aminotransferase
4	<code>sgot</code>	Aspartate aminotransferase
5	<code>gammagt</code>	Gamma-glutamyl transpeptidase
6	<code>drinks</code>	Number of half-pint equivalents of alcoholic beverages drunk per day
7	<code>selector</code>	Field to split the database


The first five variables are all blood tests that are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.


The variable `selector` was used to partition the data into two sets, very likely into a training and validation part.

Using `gplotmatrix` explore the relationship among variables 1 through 6 (exclude the selector).

- 2.23. **Cell Circularity Data.** In the lab of Dr. Todd McDevitt at Georgia Tech, researchers wanted to elucidate differences between the “static” and “rotary” culture of embryonic bodies (EBs) that were formed under both conditions with equal starting cell densities. After 2, 4, and 7 days of culture, images of EBs were acquired using phase-contrast microscopy. Image analysis software was used to determine the circularity (defined as

$4\pi(\text{Area}/\text{Perimeter}^2)$) of each EB imaged. A total of $n = 325$ EBs were analyzed from three separate plates for both static and rotary cultures at the three time points studied. The circularity measures were used to examine differences in the shape of EBs formed under the two conditions as well as differences in their variability.

The data set  `circ.dat|mat` consists of six columns corresponding to six treatments (2d, rotary), (4d, rotary), (7d, rotary), (2d, static), (4d, static), and (7d, static). Note that this is not an example of multivariate data since the columns are freely permutable, but rather six univariate data sets.

(a) For rotation and static 2d measurements, plot back-to-back histograms () as well as boxplots.

(b) For static 7d measurements graph by pie chart (`pie`) the proportion of EBs with circularity smaller than 0.75.

MATLAB FILES AND DATA SETS USED IN THIS CHAPTER

<http://springer.bme.gatech.edu/Ch2.Descriptive/>



`acorr.m, acov.m, ashton.m, balances.m, bat.m, bihist.m, biomed.m, blowfliesTS.m, BUPAliver.m, carea.m, cats.m, cats1.m, circular.m, corkrao.m, crouxrouss.m, crouxrouss2.m, diversity.m, ecg.m, empiricalcdf.m, fisher1.m, grubbs.m, hist2d.m, histn.m, lightrev.m, limestone.m, mahalanobis.m, meanvarchange.m, multifat.m, multifatstat.m, mushrooms.m, myquantile.m, mytrimmean.m, piecharts.m, scattercloud.m, simple2comp.m, smoothhist2D.m, spikes.m, surveysUKUS.m`



`ashton.dat, balance2.mat, bat.dat, blowflies.dat|mat, BUPA.dat|mat|xlsx, cats.mat, cellarea.dat|mat, circ.dat|mat, coburn.mat, cork.dat|mat, diabetes.xls, dmd.dat|mat|xls, fat.dat, light.dat, limestone.dat, QT.dat|mat, raman.dat|mat, spikes.dat`

CHAPTER REFERENCES

- Anderson, E. (1935). The Irises of the Gaspé Peninsula. *Bull. Am. Iris Soc.*, **59**, 2–5.
- Andrews, F. D. (1972). Plots of high dimensional data. *Biometrics*, **28**, 125–136.
- Bowley, A. L. (1920). *Elements of Statistics*. Scribner, New York.
- Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. Classics Appl. Math. **36**, SIAM, pp 540.
- Brockwell, P. J. and Davis, R. A. (2009). *Introduction to Time Series and Forecasting*. Springer, New York.
- Brozek, J., Grande, F., Anderson, J., and Keys, A. (1963). Densitometric analysis of body composition: revision of some quantitative assumptions. *Ann. New York Acad. Sci.*, **110**, 113–140.
- Chernoff, H. (1973). The use of faces to represent points in k -dimensional space graphically. *J. Am. Stat. Assoc.*, **68**, 361–366.
- Christov, I., Dotsinsky, I. , Simova, I. , Prokopova, R., Trendafilova, E., and Naydenov, S. (2006). Dataset of manually measured QT intervals in the electrocardiogram. *BioMed. Eng. OnLine*, **5**, 31 doi:10.1186/1475-925X-5-31. The electronic version of this article can be found online at:
<http://www.biomedical-engineering-online.com/content/5/1/31>
- David, H. A. (1998). Early sample measures of variability. *Stat. Sci.*, **13**, 4, 368–377.
- Farrell B., Bulgakova M., Hodson-Tole E.F., Shah S., Gregor R.J., Prilutsky B.I. (2009). Short-term locomotor adaptations to denervation of lateral gastrocnemius and soleus muscles in the cat. In: Proceedings of the Society for Neuroscience meeting, 17–21 October 2009, Chicago.
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, Pt. II, 179–188.
- Gauss, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. *Zeitschrift Astron.*, **1**, 185–197.
- Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *J. Stat. Educ.*, **4**, 1.
<http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>
- Kaufman, L. and Rock, I. (1962). The moon illusion. *Science*, **136**, 953–961.
- Moors, J. J. A. (1988). A Quantile Alternative for Kurtosis. *Statistician*, **37**, 25–32.
- Morrison, D. F. (1976). *Multivariate Statistical Methods*, 2nd edn. McGraw-Hill, New York
- Nicholson, A. J. (1954). An Outline of the Dynamics of Animal Populations. *Aust. J. Zool.*, **2**, 1, 9–65.
- Nightingale, F. (1858). Notes on matters affecting the health, efficiency, and hospital administration of the British army. Founded chiefly on the experience of the late war. Presented by request to the Secretary of State for War. Privately printed for Miss Nightingale, Harrison and Sons.
- Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Med. Sc. Sports Exerc.*, **17**, 2, 189.
- Percy, M. E., Andrews, D. F., Thompson, M. W., and Opitz J. M. (1981). Duchenne muscular dystrophy carrier detection using logistic discrimination: Serum creatine kinase and hemopexin in combination. *Am. J. Med. Genet.*, **8**, 4, 397–409.
- Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.
- Shumway, R. H. and Stoffer, D. S. (2005). *Time Series Analysis and Its Applications*. Springer Texts in Statistics, Springer, New York.

- Siri, W. E. (1961). Body composition from fluid spaces and density: Analysis of methods. In *Techniques for Measuring Body Composition*, Eds. J. Brozek and A. Henzchel. National Academy of Sciences, Washington, 224–244.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 2684, 677–680. PMID 17750512.
- Sturges, H. (1926). The choice of a class-interval. *J. Am. Stat. Assoc.*, **21**, 65–66.
- Velleman, P. F. and Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *Am. Stat.*, **47**, 1, 65–72.



<http://www.springer.com/978-1-4614-0393-7>

Statistics for Bioengineering Sciences
With MATLAB and WinBUGS Support

Vidakovic, B.

2011, XVI, 753 p., Hardcover

ISBN: 978-1-4614-0393-7