In the other category, model-dependent statistics, we lump the whole subject of fitting data to a theory, parameter estimation, least-squares fits, and so on. Those subjects are introduced in Chapter 15.

Section 14.1 deals with so-called measures of central tendency, the moments of a distribution, the median and mode. In §14.2 we learn to test whether different data sets are drawn from distributions with different values of these measures of central tendency. This leads naturally, in §14.3, to the more general question of whether two distributions can be shown to be (significantly) different.

In §14.4–§14.7, we deal with measures of association for two distributions. We want to determine whether two variables are “correlated” or “dependent” on one another. If they are, we want to characterize the degree of correlation in some simple ways. The distinction between parametric and nonparametric (rank) methods is emphasized.

Section 14.8 introduces the concept of data smoothing, and discusses the particular case of Savitzky-Golay smoothing filters.

This chapter draws mathematically on the material on special functions that was presented in Chapter 6, especially §6.1–§6.4. You may wish, at this point, to review those sections.

CITED REFERENCES AND FURTHER READING:

\section{14.1 Moments of a Distribution: Mean, Variance, Skewness, and So Forth}

When a set of values has a sufficiently strong central tendency, that is, a tendency to cluster around some particular value, then it may be useful to characterize the set by a few numbers that are related to its \textit{moments}, the sums of integer powers of the values.

Best known is the \textit{mean} of the values $x_1, \ldots, x_N$,

$$
\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \tag{14.1.1}
$$

which estimates the value around which central clustering occurs. Note the use of an overbar to denote the mean; angle brackets are an equally common notation, e.g., $\langle x \rangle$. You should be aware that the mean is not the only available estimator of this
quantity, nor is it necessarily the best one. For values drawn from a probability
distribution with very broad “tails,” the mean may converge poorly, or not at all, as
the number of sampled points is increased. Alternative estimators, the median
and the mode, are mentioned at the end of this section.

Having characterized a distribution’s central value, one conventionally next
characterizes its “width” or “variability” around that value. Here again, more than
one measure is available. Most common is the variance,

$$\text{Var}(x_1 \ldots x_N) = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \bar{x})^2 \quad (14.1.2)$$

or its square root, the standard deviation,

$$\sigma(x_1 \ldots x_N) = \sqrt{\text{Var}(x_1 \ldots x_N)} \quad (14.1.3)$$

Equation (14.1.2) estimates the mean squared deviation of $x$ from its mean value.
The variance and standard deviation depend on the second moment. It is not uncommon, in real
life, to be dealing with a distribution whose second moment does not exist (i.e., is infinite). In this case, the variance or standard deviation is useless as a measure
of the data’s width around its central value: The values obtained from equations
(14.1.2) or (14.1.3) will not converge with increased numbers of points, nor show
any consistency from data set to data set drawn from the same distribution. This can
occur even when the width of the peak looks, by eye, perfectly finite. A more robust
estimator of the width is the average deviation or mean absolute deviation, defined by

$$\text{ADev}(x_1 \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} |x_j - \bar{x}| \quad (14.1.4)$$

One often substitutes the sample median $x_{\text{med}}$ for $\bar{x}$ in equation (14.1.4). For any
fixed sample, the median in fact minimizes the mean absolute deviation.

Statisticians have historically sniffed at the use of (14.1.4) instead of (14.1.2),
since the absolute value brackets in (14.1.4) are “nonanalytic” and make theorem-
proving difficult. In recent years, however, the fashion has changed, and the subject
of robust estimation (meaning, estimation for broad distributions with significant
numbers of “outlier” points) has become a popular and important one. Higher
moments, or statistics involving higher powers of the input data, are almost always
less robust than lower moments or statistics that involve only linear sums or (the
lowest moment of all) counting.
Chapter 14. Statistical Description of Data

That being the case, the **skewness** or third moment, and the **kurtosis** or fourth moment should be used with caution or, better yet, not at all.

The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean, standard deviation, and average deviation are *dimensional* quantities, that is, have the same units as the measured quantities \( x_j \), the skewness is conventionally defined in such a way as to make it *nondimensional*. It is a pure number that characterizes only the shape of the distribution. The usual definition is

\[
\text{Skew}(x_1 \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} \left[ \frac{x_j - \bar{x}}{\sigma} \right]^3 \tag{14.1.5}
\]

where \( \sigma = \sigma(x_1 \ldots x_N) \) is the distribution’s standard deviation (14.1.3). A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive \( x \); a negative value signifies a distribution whose tail extends out towards more negative \( x \) (see Figure 14.1.1).

Of course, any set of \( N \) measured values is likely to give a nonzero value for (14.1.5), even if the underlying distribution is in fact symmetrical (has zero skewness). For (14.1.5) to be meaningful, we need to have some idea of its standard deviation as an estimator of the skewness of the underlying distribution. Unfortunately, that depends on the shape of the underlying distribution, and rather critically on its tails! For the idealized case of a normal (Gaussian) distribution, the standard deviation of (14.1.5) is approximately \( \sqrt{15}/N \). In real life, it is good practice to believe in skewnesses only when they are several or many times as large as this.

The kurtosis is also a nondimensional quantity. It measures the relative peakedness or flatness of a distribution. Relative to what? A normal distribution, what else! A distribution with positive kurtosis is termed *leptokurtic*; the outline of the Matterhorn is an example. A distribution with negative kurtosis is termed *platykurtic*; the outline of a loaf of bread is an example. (See Figure 14.1.1.) And, as you no doubt expect, an in-between distribution is termed *mesokurtic*.

The conventional definition of the kurtosis is

\[
\text{Kurt}(x_1 \ldots x_N) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[ \frac{x_j - \bar{x}}{\sigma} \right]^4 \right\} - 3 \tag{14.1.6}
\]

where the \(-3\) term makes the value zero for a normal distribution.
The standard deviation of (14.1.6) as an estimator of the kurtosis of an underlying normal distribution is $\sqrt{96/N}$. However, the kurtosis depends on such a high moment that there are many real-life distributions for which the standard deviation of (14.1.6) as an estimator is effectively infinite.

Calculation of the quantities defined in this section is perfectly straightforward. Many textbooks use the binomial theorem to expand out the definitions into sums of various powers of the data, e.g., the familiar

$$\text{Var}(x_1\ldots x_N) = \frac{1}{N-1} \left( \sum_{j=1}^{N} x_j^2 \right) - N\bar{x}^2 \approx \bar{x}^2 - \bar{x}^2$$  \hspace{1cm} (14.1.7)

but this can magnify the roundoff error by a large factor and is generally unjustifiable in terms of computing speed. A clever way to minimize roundoff error, especially for large samples, is to use the \textit{corrected two-pass algorithm}\cite{1}: First calculate $\bar{x}$, then calculate $\text{Var}(x_1\ldots x_N)$ by

$$\text{Var}(x_1\ldots x_N) = \frac{1}{N-1} \left\{ \sum_{j=1}^{N} (x_j - \bar{x})^2 - \frac{1}{N} \left[ \sum_{j=1}^{N} (x_j - \bar{x}) \right]^2 \right\}$$  \hspace{1cm} (14.1.8)

The second sum would be zero if $\bar{x}$ were exact, but otherwise it does a good job of correcting the roundoff error in the first term.

SUBROUTINE moment(data,n,ave,addev,sdev,var,skew,curt)
INTEGER n
REAL ave,addev,sdev,var,skew,curt,data(n)

Given an array of data(1:n), this routine returns its mean ave, average deviation addev, standard deviation sdev, variance var, skewness skew, and kurtosis curt.

INTEGER j
REAL p,s,ep

if(n.le.1)pause 'n must be at least 2 in moment'

s=0. \hspace{1cm} \text{First pass to get the mean.}
do : j=1,n
s=s+data(j)
enddo:
ave=s/n
addev=0.
var=0.
skew=0.
curt=0.
ep=0.
do : j=1,n
s=data(j)-ave
ep=ep+s
addev=addev+abs(s)
p=s
var=var+p
p=p*s
skew=skew+p
p=p*s
curt=curt+p
enddo:

addev=addev/n

var=(var-ep**2/n)/(n-1) \hspace{1cm} \text{Corrected two-pass formula.}
sdev=sqrt(var)

Put the pieces together according to the conventional definitions.
Semi-Invariants

The mean and variance of independent random variables are additive: If $x$ and $y$ are drawn independently from two, possibly different, probability distributions, then

$$ (x + y) = \mu_x + \mu_y \quad \text{Var}(x + y) = \text{Var}(x) + \text{Var}(x) $$ \hspace{1cm} (14.1.9)

Higher moments are not, in general, additive. However, certain combinations of them, called semi-invariants, are in fact additive. If the centered moments of a distribution are denoted $M_k$,

$$ M_k \equiv \langle (x_i - \mu)^k \rangle $$ \hspace{1cm} (14.1.10)

so that, e.g., $M_2 = \text{Var}(x)$, then the first few semi-invariants, denoted $I_k$ are given by

- $I_2 = M_2$
- $I_3 = M_3$
- $I_4 = M_4 - 3M_2^2$
- $I_5 = M_5 - 10M_2M_3$
- $I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3$ \hspace{1cm} (14.1.11)

Notice that the skewness and kurtosis, equations (14.1.5) and (14.1.6) are simple powers of the semi-invariants,

$$ \text{Skew}(x) = I_3/I_2^{3/2} \quad \text{Kurt}(x) = I_4/I_2^2 $$ \hspace{1cm} (14.1.12)

A Gaussian distribution has all its semi-invariants higher than $I_2$ equal to zero. A Poisson distribution has all of its semi-invariants equal to its mean. For more details, see [2].

Median and Mode

The median of a probability distribution function $p(x)$ is the value $x_{\text{med}}$ for which larger and smaller values of $x$ are equally probable:

$$ \int_{-\infty}^{x_{\text{med}}} p(x) \, dx = \frac{1}{2} = \int_{x_{\text{med}}}^{\infty} p(x) \, dx $$ \hspace{1cm} (14.1.13)

The median of a distribution is estimated from a sample of values $x_1, \ldots, x_N$ by finding that value $x_i$ which has equal numbers of values above it and below it. Of course, this is not possible when $N$ is even. In that case it is conventional to estimate the median as the mean of the unique two central values. If the values $x_j$, $j = 1, \ldots, N$ are sorted into ascending (or, for that matter, descending) order, then the formula for the median is

$$ x_{\text{med}} = \begin{cases} x_{(N+1)/2}, & N \text{ odd} \\ \frac{1}{2}(x_{N/2} + x_{(N/2)+1}), & N \text{ even} \end{cases} $$ \hspace{1cm} (14.1.14)
If a distribution has a strong central tendency, so that most of its area is under a single peak, then the median is an estimator of the central value. It is a more robust estimator than the mean is: The median fails as an estimator only if the area in the tails is large, while the mean fails if the first moment of the tails is large; it is easy to construct examples where the first moment of the tails is large even though their area is negligible.

To find the median of a set of values, one can proceed by sorting the set and then applying (14.1.14). This is a process of order $N \log N$. You might rightly think that this is wasteful, since it yields much more information than just the median (e.g., the upper and lower quartile points, the deciles, etc.). In fact, we saw in §8.5 that the element $x_{(N+1)/2}$ can be located in of order $N$ operations. Consult that section for routines.

The mode of a probability distribution function $p(x)$ is the value of $x$ where it takes on a maximum value. The mode is useful primarily when there is a single, sharp maximum, in which case it estimates the central value. Occasionally, a distribution will be bimodal, with two relative maxima; then one may wish to know the two modes individually. Note that, in such cases, the mean and median are not very useful, since they will give only a “compromise” value between the two peaks.

**CITED REFERENCES AND FURTHER READING:**

### 14.2 Do Two Distributions Have the Same Means or Variances?

Not uncommonly we want to know whether two distributions have the same mean. For example, a first set of measured values may have been gathered before some event, a second set after it. We want to know whether the event, a “treatment” or a “change in a control parameter,” made a difference.

Our first thought is to ask “how many standard deviations” one sample mean is from the other. That number may in fact be a useful thing to know. It does relate to the strength or “importance” of a difference of means if that difference is genuine. However, by itself, it says nothing about whether the difference is genuine, that is, statistically significant. A difference of means can be very small compared to the standard deviation, and yet very significant, if the number of data points is large. Conversely, a difference may be moderately large but not significant, if the data