4

Mean value and standard deviation Linear combinations of variables

4.1 Mean value and standard deviation

4.2 Linear combinations of variables

Most people have an at least basic knowledge about the idea of mean or average value although there are several such measures used in e.g. newspapers or other media.

However, for the professional statistical work a vague notion is not enough and therefore we put a lot of emphasis on this subject. Also the idea of a measure for spread is accepted but not so commonly seen in daily life. This measure is more mathematical but necessary for a proper analysis.

As soon as statistics is applied to practical problems the need for *combination of variables* becomes a reality. This is not so obvious to the inexperienced user but still a fact. Because of this, we describe the theoretical and practical consequences and illustrate with a number of real examples.

4.1 Mean value and standard deviation

When working with theoretical as well as practical statistical analysis, there is often a need to reduce the data or the process to just a few, well chosen numbers. We introduce two such numbers, the *mean value* and the *standard deviation* where the first is a value of location and the second is a value of variation.

These two measures are referred to as *parameters* and we will discuss their theoretical features as well as ways to estimate them from a set of data. In order to set the scene we first introduce some important statistical concept namely *random variable*, *probability*, and *estimation* along together with some vocabulary.

The idea to use just the mean and standard deviation to describe a process does not imply that we throw away, or do not use, other features of the process. Furthermore, calculating the *average value* (in order to estimate the mean value) does not imply that we are not interested in the original data. On the contrary, the original data will be used over and over again.

%XbarS
%AveStd

ECS – Ex: 1.1 – 1.7

4.1.1 Mean value

Random variable

In advanced books about statistics the expression random variable is described as a function. However, we prefer to introduce it by some examples. Imagine for example that we study the number of phone calls per hour to a telephone exchange. Obviously this number varies from hour to hour. If we study the weight of people we will find that it varies from person to person and if we study the time it takes to perform a certain task on the computer we will also find a variation.

There is a lot of things to be said about each example. Firstly, it must of course be possible to measure the variation, i.e. the measuring device must be accurate enough. Maybe a lot of the variation can be explained quite easily: the number of phone calls per hour is most likely different in the middle of the night compared with working hours, the weight depends on the height, age, and sex of the person, the traffic load on the computer can be different in the morning compared to the evening etc.

Despite these explainable differences (that have to be proven) there is still a variation in e.g. length of grown men of the same age. Maybe it is possible to argue that even these differences are explainable: different heritage, different food, different living conditions, etc. Sometimes it is interesting to science to try to find (and prove) such background causes, but sometimes we have to be satisfied with describing the random variable by its mean value and standard deviation. (Usually a random variable has to be described in more ways but we leave this to chapter 7 and 8.)

Vocabulary

We will use upper case letters to designate random variables e.g. X, Y, Z, W. Behind each such letter there is a hopefully well defined text string that fully describes the random variable such as "number of incorrect items in a box of 200 items", "number of incorrect statements in a program of 2 000 statements", "number of phone calls during 10 minutes", "time for an order to go from point A to point B" etc. The first examples designate what we later will call discrete variables and the last example is called a continuous variable.

%Prob1
%Prob2
%Prob3

Probability

In chapter 6 we will introduce the concept of probability and we will then write things like $P(X = x)$, $P(Y \leq y)$, $P(Z > z)$ etc. These expressions will be read "the probability that the variable X will be equal to the value x", "the probability that the variable Y will be equal to or less than the value y", "the probability that the variable Z will be greater than the value z". We will give mathematical formulas to calculate the probability in several cases and then, of course, we will use numbers instead of x, y and z.

Once given the probability of certain events, we will be able to calculate many useful pieces of information about the actual situation. These formulas will include some *parameters* (another word is *constants*) and these will generally be designated by Greek letters like λ, σ, μ, etc. However, behind these Greek letters there is a real number. The vocabulary described here is fairly universal in the literature and should not give any trouble after some practice.

Estimation

In statistics the word *estimation* is used frequently. The subject is an enormously large area, with many difficulties of statistical and mathematical nature, but the basic ideas are fairly simple even if it takes a bit to get used to. E.g. the true mean value of a random variable is something that we never will be able to calculate from sample data. When we calculate the average of a sample we *estimate* the true mean value. All parameters, as mentioned above, has to be estimated in some way or another by calculations on data.

However, because of the randomness we will never be able to state that we *know* the value of a certain parameter. The mathematical and statistical difficulties concerns things like 'what constitutes a good estimate?' or 'how should the data be used in the most efficient way?'. We will not discuss such features here. We will only show, by simulations, that there are different ways, with different good or bad sides, for estimation.

Example 1 – Time to completion

An organisation governs the time it takes to complete a certain task. The time necessary is measured in hours and the main interest is to find the mean time. The mean value can be utilised in several ways. One way is to monitor the mean value to see that it does not increase and another way is to use the mean to estimate costs in other similar tasks.

Example 2 – Transportation

A furniture retailer offers the customers a transportation of their new furniture for a certain fee. Of course the fee must be calculated in such a way that it covers the costs. The retailer is therefore interested in the mean value of the transportation cost.

Example 3 – Insurance

Every insurance company is interested in the mean values of several random variables such as mean number of traffic accidents during a year or mean number of injured people in an accident. The idea is of course to be able to set the fees properly.

Example 4 – The quality process

The quality of a process or a product is measured in many ways. E.g. in the manufacturing of electronic equipment there is a large number of variables that the manufacturer and the customer find important. These variables are often described by their mean values.

Example 5 – The development of software

The development of software is also full of questions that can be raised and handled in the context of random variables:

- mean time to completion of certain jobs
- mean time of testing
- mean number of remaining faults after test
- mean time for fault finding, mean backlog of trouble reports

The main purpose is of course to follow up results, to see that actions taken really improve quality and productivity, etc.

Example 6 – The maximum value, a counter example

%MinMax

Sometimes one is not interested in the mean value of a random variable but rather some other characteristic such as the maximum value. One common example is the 'taxi-problem'. A professor arrives to the railway station in a city. He stands outside the main building and sees the taxis driving past. He notes the number of the taxi on the special number plate used only by taxis. After a while he produces an estimate of the total number of taxis in the city. As a matter of fact, this technique was used to estimate the number of enemy aircraft during the 2nd World War.

Calculation of the average value as an estimate of the mean value

%XbarS
%AveStd
%CreDist

The true mean value of a random variable is calculated on theoretical grounds. We will defer that calculation to chapter 7 when we start dealing with distributions. The literature uses the Greek letter μ with some index to designate the (true) mean value. Another common way is to use E (for expectation). Thus the vocabulary will be:

$$\mu_Y = E(Y) \qquad \mu_X = E(X) \qquad \mu_Z = E(Z)$$

Here we will introduce the *average value*, calculated from a sample. The average will be at least approximate equal to the true mean:

$$\bar{x} \approx \mu$$

The average value $\bar{x}$ (pronounced "x-bar") is a simple calculation: $\bar{x} = \dfrac{\sum x_i}{n}$

In words this formula reads: "Sum all the figures and divide by the sample size n". Assume that we have the following 12 values measured in hours:

2.41 2.39 2.42 2.37 2.40 2.35 2.36 2.38 2.37 2.34 2.38 2.39

What will the average be? Using the formula we get the following:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2.41 + 2.39 + 2.42 + 2.37 + 2.40 + \ldots + 2.38 + 2.37 + 2.34 + 2.38 + 2.39}{12} = 2.38$$

Thus the average value of the 12 values is 2.38 hours.

Input				Answer
$x_1 = 5.2$	$x_2 = 2.2$	$x_3 = 9.8$	$x_4 = 3.6$	$\bar{x} = 5.2$
$x_1 = 0.25$	$x_2 = 0.27$	$x_3 = 0.38$	$x_4 = 0.32$	$\bar{x} = 0.305$
$x_1 = 12.2$	$x_2 = 13.4$	$x_3 = 12.8$	$x_4 = 11.6$	$\bar{x} = 12.5$

Some concluding notes

To the beginner there is a bit of confusion about the difference between the average and the mean values. The average is calculated from a limited number of numeric values (limited as opposite to unlimited) while the true mean is calculated on theoretical grounds (chapter 7). The average is an estimate of the true mean and we will later have a discussion how good the estimates are (chapter 9).

However, in statistical work the true mean is used for e.g. modelling purposes. When modelling, one can use a numerical value for the mean value obtained from a large sample. Of course, it would be possible to assign any realistic value to the mean just for modelling or simulating purposes:

- "What happens to the total testing time if the mean fault rate is such and such and the mean block size is such and such?"
- "Let's say that the mean incoming number of customer complaints per week is μ_1 and the mean number of customer complaints handled per week is μ_2. How will the mean back log (i.e. unanswered customer complaints) change?"

Statistical work is full of such ideas and questions.

4.1.2 Standard deviation

%XbarS
%AveStd
%CreDist

The standard deviation is the most common way to measure the spread of a random variable. From a sample of data we calculate the *sample standard deviation* according to the formula below. The sample standard deviation, or *standard deviation* only, is designated by *s* and is an estimate of the true standard deviation, that is designated by σ. The true standard deviation is calculated on theoretical grounds and we leave that to chapter 7.

Actually, there is another measure of spread that is the very base for all statistical analysis namely the *variance*. The variance is the measure with all the nice mathematical features but unfortunately for us human beings it gives us the numerical results in a strange way: if the studied variable concerns *number of faults*, the average is given in *number of faults* but the variance is given in *number of faults squared*, or if the variable is reported in *hours*, the average is given in *hours* but the variance will be stated in *hours*^{2}. If we take the square root of the variance we get the standard deviation.

Vocabulary

In the literature, the most common designation for the standard deviation and the variance will be seen in the following table:

	Common	Less common
Sample standard deviation	s	
True standard deviation	σ	$S(X)$
Sample variance	s^2	
True variance	σ^2, $V(X)$	v

Calculation of the variance and standard deviation

The true standard deviation and the true variance of a random variable is calculated on theoretical grounds. We will show some examples in chapter 8 when dealing with the different features of *statistical distributions*. However, the sample standard deviation s, calculated via the formula below, is an estimate of the true standard deviation σ:

$$s \approx \sigma$$

The formula used to calculate the standard deviation looks a bit awkward at first sight:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

In words it can be described as follows.

After calculating the average, we do as follows:

- calculate the distance from each value to the common average ⇒ $(x_i - \bar{x})$
- square this value ⇒ $(x_i - \bar{x})^2$
- Sum all the squared values ⇒ $\sum (x_i - \bar{x})^2$
- Divide by $n - 1$ (n = sample size) to get the sample variance ⇒ $\dfrac{\sum (x_i - \bar{x})^2}{n-1}$
- The square root gives the sample standard deviation ⇒ $\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$

Thus the complete formula for the sample standard deviation is:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Assume that we use the same 12 figures as for the calculation of the average:

2.41 2.39 2.42 2.37 2.40 2.35 2.36 2.38 2.37 2.34 2.38 2.39

What will the standard deviation be? Using the formula we get the following (remember that the calculated average is 2.38 hours):

$$s = \sqrt{\frac{(2.41-2.38)^2 + (2.39-2.38)^2 + ... + (2.39-2.38)^2}{12-1}} = \sqrt{\frac{0.0062}{11}} \approx 0.0237 \text{ hours}$$

Thus the standard deviation, calculated from the 12 values, is 0.0237 hours.

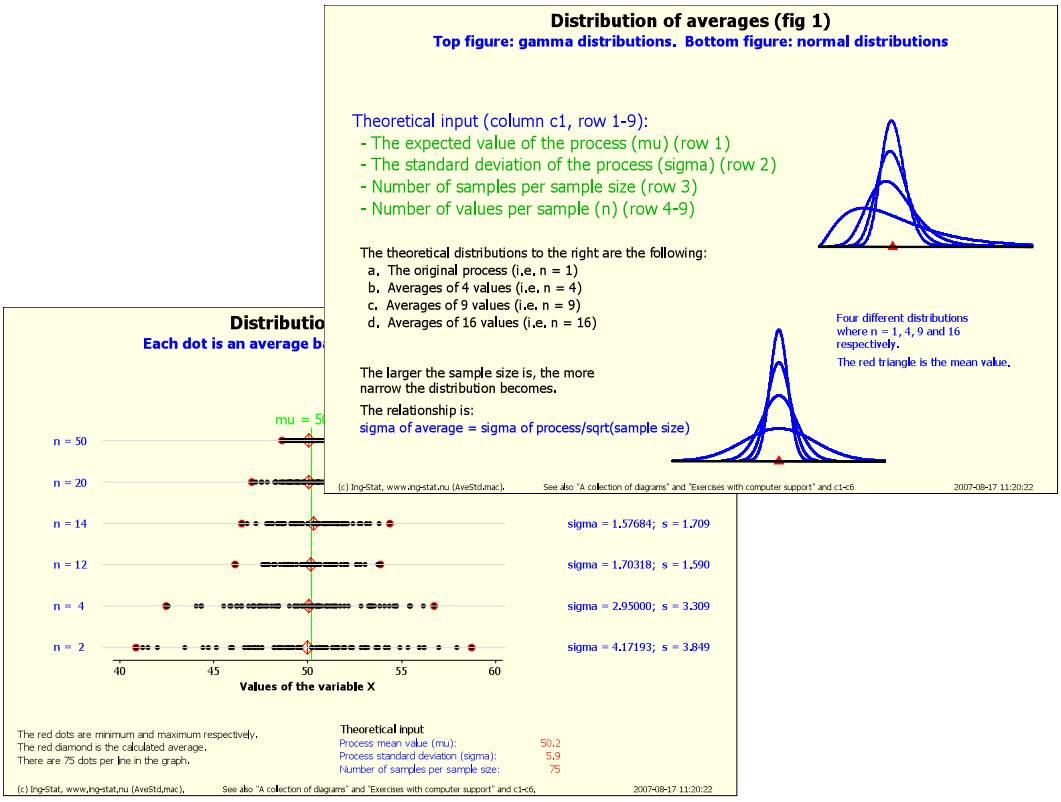
Input				Answer
$x_1 = 5.2$	$x_2 = 2.2$	$x_3 = 9.8$	$x_4 = 3.6$	$s = 3.30$
$x_1 = 0.25$	$x_2 = 0.27$	$x_3 = 0.38$	$x_5 = 0.32$	$s = 0.058$
$x_1 = 12.2$	$x_2 = 13.4$	$x_3 = 12.8$	$x_5 = 11.6$	$s = 0.775$

Some concluding notes

In most textbooks one can find other formulas to calculate the standard deviation. Although these stem from the formula above and give exactly the same result, they are considered to be simpler to use when doing the calculations by hand or using only a simple pocket calculator. However, when using these formulas, all resemblance with the original definition is lost. From the formula used above, one can, at least after some guidance, see that the variance is the average squared deviation from the mean value. Thus the variance is in some sense an average value. The small examples above are just for the reader to test the understanding of the mechanics of the calculation of the standard deviation.

The average value and the standard deviation are two figures calculated from the data. However, before these values can come into practical use, they must be put into some further context, i.e. the context of statistical distributions. We return to this most important idea in chapter 7. We just want the reader to have some patience and ask him to make sure that he knows and understands how to calculate the two values.

%XbarS
%AveStd
%CreDist



4.1.3 Summary of mean value and standard deviation

The average value

$$\bar{x} = \frac{\sum x_i}{n}$$

$\bar{x}$ estimates μ

The standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

s estimates σ

x_i = the individual values
n = the number of values
μ = the true mean value of the process
σ = the true standard deviation of the process

μ and σ are calculated on theoretical grounds. See e.g. '7.1.2 Summary of…'

The dictionary is the only place where success comes before work.
Mark Twain

Do not let what you cannot do interfere with what you can do.
John Wooden

Success usually comes to those who are too busy to be looking for it.
Henry David Thoreau

4.2 Linear combinations of variables

In (4.1) we introduced the idea of random variables, mean value and standard deviation. However, as soon as one starts with practical use of statistical analysis, life becomes more difficult and interesting. One will come across the problem of combining variables in the same problem. The type of combination that we treat here is linear, i.e. sums and differences of variables. Such a sum is of course also a variable and its mean and standard deviation can be calculated from the terms in the combination.

There are of course other types of combinations of variables such as non-linear or a mixture of linear or non-linear or combinations where also the *number* of terms is a variable. We show some short examples of these latter types in this section but we concentrate on common linear combinations.

%LinC
%Die
%CLT

ECS – Ex: 2.1 – 2.10

4.2.1 Mean value of linear combinations of variables

Example 1 – Humidity and temperature

In the manufacturing of electronic components there are many photo processes in order to develop the fine pattern used in the components of electronic hardware. In the booklet KODAK ACUMAX Products (appendix D page 33) there is a table that shows how the length of the photo negative changes when the humidity and temperature changes. Thus the change in length is a combination of the two variables humidity and temperature. With some arithmetic (we used regression analysis) this table can be reduced to a simple formula where

- Y is the change in thousands of an inch (over a distance of 24 inches),
- Rh is the change in relative humidity in per cent,
- T is the temperature change in degrees Fahrenheit.

The formula becomes

$$Y = 0.264 \cdot Rh + 0.240 \cdot T$$

and we see that Y is a linear combination of the two variables Rh and T. However, if we want the formula to express the change in length in millimetres and the temperature in centigrade, the formula has to be recalculated. It then becomes

$$Y = 0.00671 \cdot Rh + 0.0110 \cdot T$$

The question is: what will the mean value and the standard deviation be for Y if these two values are known for Rh and T? Before we enter that discussion we have some more examples.

Example 2 – Total time (I)

If we study the total time T it takes to perform a certain task, whether it be to manufacture an item or to develop a software program, we often find that the times vary. One explanation is that the task is made up by a number of subtasks S_i connected in series. T can then be regarded as a linear combination of subtasks. If the time for each subtask is constant then of course the total time is just a constant without variation. But most likely the subtasks have a variation and thus we can express it using the ideas of random variable, mean value and standard deviation. Let us assume that we have four subtasks. We can then state the relationship between T and S_i in the following way:

$$T = S_1 + S_2 + S_3 + S_4$$

Example 3 – Total number of lines per day (I)

Suppose that a department that produces software has a fixed number of programmers. Let us call this number n. Assume that we want to model the total number of lines of code Y that is produced during a day. Every programmer does not produce, by some reason or other, the same number of lines per day. Therefore, let us consider the number of lines per day per programmer as a random variable X. This means that the total number of lines in a certain day is a linear combination (in this case a sum) of n random variables. The model is then

$$Y = X_1 + X_2 + X_3 + \ldots + X_n$$

Example 4 – Total number of lines per day (II)

Suppose that the number of available programmers varies (maybe more realistic) and we regard this number as a random variable N. N is the number of available programmers on a given day. Here Y is a linear combination of a random number of random variables:

$$Y = X_1 + X_2 + X_3 + \ldots + X_N$$

This model is also valid for the situation describing manufacture where N is the number of orders per day and X is the number of items per order and thus Y is total number of items to be delivered per day. A common servicing situation can also be described with this model. N is the number of customers per day and X is the time spent by each customer. Y is thus he total time spent for service per day.

Example 5 – Pythagoras' Theorem

Numerous situations in industry and science can be handled using Pythagoras' theorem. If we have two random variables X and Y and apply the theorem we get

$$Z = \sqrt{X^2 + Y^2}$$

and Z is certainly not a linear combination of X and Y.

Example 6 – Electronic circuitry

In electronic circuits there are a lot of statistics and mathematics. Lets say that we have three resistors A, B and C connected according to figure 4.2.1. We regard their resistance R_A, R_B, and R_C as random variables with different parameters. What parameters will the total resistance R obtain?

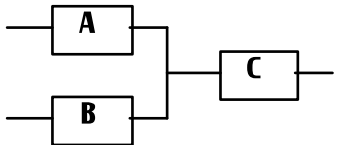


Figure 4.2.1 Three resistors in a circuit. What parameters will the total resistance get?

First we construct the proper formula (model) for the assumptions. Some arithmetic gives that R is the following:

$$R = \frac{R_A \cdot R_B}{R_A + R_B} + R_C \qquad R = \underbrace{\frac{R_A \cdot R_B}{R_A + R_B}}_{R_D} + R_C = R_D + R_C$$

R is not a linear combination of R_A, R_B, and R_C but if we can replace the first term by, say R_D, and find the parameters of R_D we find that R is a linear combination of R_D and R_C as in the expression to the right. Finding the parameters of R_D can be more or less easy.

Example 7 – Total time (II)

In industry there are many models that describe the times of operations necessary to complete a task. Suppose that we have the operations A, B, C, D, and E. T_A to T_E are also the random variables that describe the time it takes for each operation. Figure 4.2.2 shows how the operations are connected i.e. B, C, and D are parallel and the flow can not continue until all the operations of B, C, and D are finished.

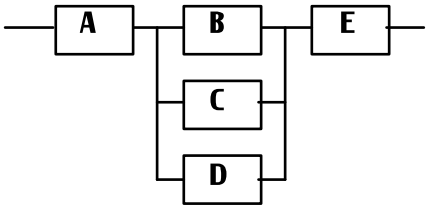


Figure 4.2.2 Five operations. What parameters will the total time get?

The total time T then becomes: $$T = T_A + \max(T_B, T_C, T_D) + T_E$$

'max(T_B, T_C, T_D)' means that the maximum of the three variables 'sets the pace'. In this case we have to find the mean for this maximum but after doing that, we are back to a linear combination of *three* variables. (To find the mean of max(T_B, T_C, T_D) is not an easy task. We can use tricky mathematics or we can simulate the model.)

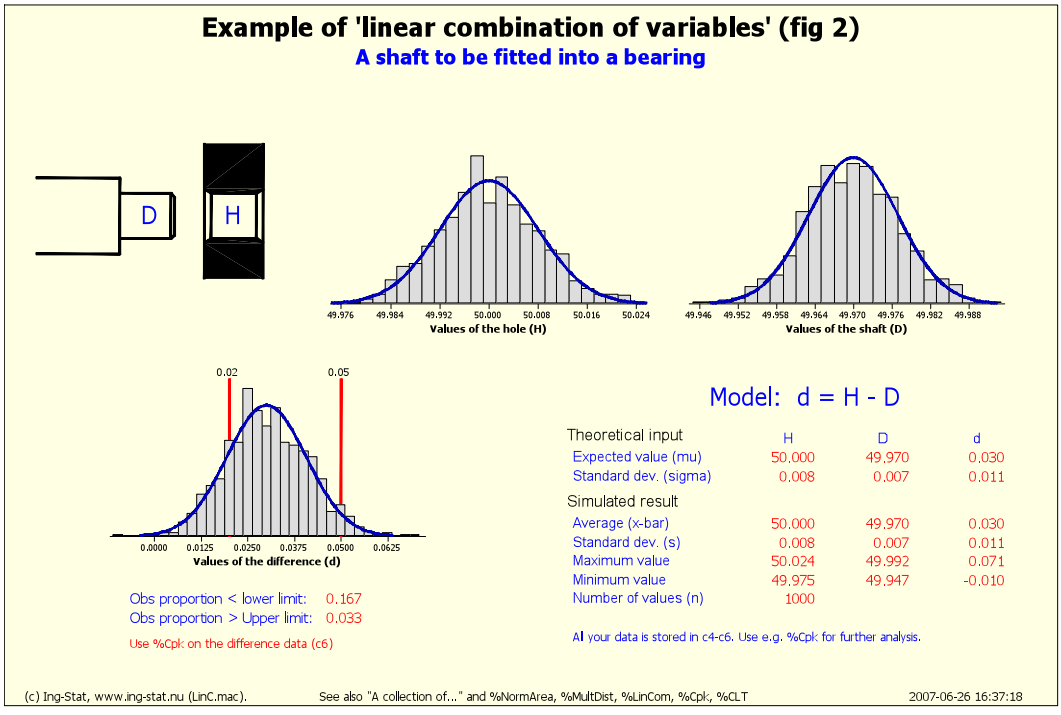
Example 8 – Measurements

A measurement always contains an error and very often we can consider the error as a random variable. If we assume the error to have the same mean and standard deviation irrespective of the size of the value that we want to measure (the real value), we have the following model:

$$\textit{Measured value} = \textit{Real value} + \textit{error}$$

Here we regard *Real value* as a random variable. We also assume that the variables are independent. If we have knowledge of the parameters of *Measured value* and the *error*, then we can estimate the variation in the *Real value*, an often interesting question.

%LinC



A linear combination of variables. (See the macro %LinC.)

Example 9 – Trouble reports

A department receives trouble reports from the field regarding malfunctions of some software or electronic equipment. If we regard the stream of incoming reports per week as a random variable X and the number of handled reports per week as another random variable Y then we can state that the difference D is a random variable:

$$D = X - Y$$

Let us assume that the department starts with a fixed number k of trouble reports. After four weeks we expect the following number T of unanswered reports:

$$T = k + X_1 - Y_1 + X_2 - Y_2 + X_3 - Y_3 + X_4 - Y_4 = k + D_1 + D_2 + D_3 + D_4$$

If the expected value of Y is larger than the expected value of X, then the true mean of T will decrease week by week. In addition, the model above is only valid as long as the number of unanswered trouble reports is more than zero. (After that, the situation becomes more complicated).

Summary of the examples

Before going into the details of calculating the parameters of combination of variables we summarise the examples.

Example	Type	Further info
1	Linear	Ch 4
2	Linear	
3	Linear	
8	Linear	
9	Linear	
4	Non-linear	"Combinations of variables.doc"
5	Non-linear	
6	Mixture	The macro %Mix, "A mixture of variables.doc"
7	Mixture	

All the examples above relate to what we call *models* (see chapter 5). In that chapter we will discuss *deterministic models* as well as *statistical models*. In some of these models we will state clearly what variables will be included. In some situations however, we do not know exactly what variables should be a part of the model; the analysis will show this. We maybe have to state that the model includes an unknown number of unknown variables. This 'unknown number of unknown variables' is then combined into one piece, often called the *error term*.

This situation arises e.g. when we have measurements on one interesting variable and we want to investigate if, and how, this variable is related to other variables. Thus we try to 'fit' different models to measurements by some technique (most often regression analysis) in order to develop a model that we believe in. The fit will most likely not be perfect and the difference between the measurements and model we call the error or the error term.

Calculation of the mean value of a linear combination

As mentioned in (4.1) we use upper case letters to designate random variables and now we introduce lower case letters to designate constants. If we use this convention, a general form of a linear combination T of the random variables W, X, Y and then becomes:

$$T = a\cdot W + b\cdot X + c\cdot Y + d\cdot Z$$

If we use $E(\ )$ as the expected value (true mean) of a variable, we can give the following general expression for the mean of a linear combination:

$$E(T) = a\cdot E(W) + b\cdot E(X) + c\cdot E(Y) + d\cdot E(Z)$$

The constants a to d can be either positive or negative e.g. in example 9 some constants are +1 (in front of the X's) and some are –1 (in front of the Y's). If we apply this to example 1 we get:

$$E(Y) = 0.264\cdot E(Rh) + 0.240\cdot E(T)$$

$E(Y)$ is then the true mean of the change of length of the photo expressed in mils, i.e. thousands of an inch. If we prefer to use centigrade and give the answer in millimetres we have to write

$$E(Y) = 0.00671\cdot E(Rh) + 0.0110\cdot E(T)$$

where the temperature T now is given in centigrade. If we know that $E(Rh)$ = 6.3% and $E(T)$ = 3°C then we can calculate $E(Y)$:

$$\begin{aligned} E(Y) &= 0.00671\cdot E(Rh) + 0.0110\cdot E(T) \\ &= 0.00671\cdot 6.3 + 0.0110\cdot 3 = 0.075 \end{aligned}$$

We then expect that the mean value of the change in length (over 24 inches) is 0.075 mm. Note that Rh and T are measured as deviations from some target. If the deviation from target were zero the expression would, as expected, give the answer of zero deviation. Note also that the above is a theoretical result. When we later simulate this problem, we will not get this exact result. The reason is of course that a simulation is based on a limited number of measurements (not limited in the sense of small but as opposite to unlimited) but it will come close to the result above.

Example	Input	Answer
2	$E(S_1) = 5.2$ $E(S_2) = 2.2$ $E(S_3) = 9.8$ $E(S_4) = 3.6$	$E(T) = 20.8$
3	$E(X_i) = 525$ lines/day 12 programmers	$E(Y) = 6300$
9	$k = 200$ $E(X) = 23.5$ $E(Y) = 28.6$	$E(T) = 179.6$

In the last example we can also note that although X, Y, and T are necessarily integers, the corresponding expected values need not to be integers. In (4.2.4) there are several more exercises.

The average – a linear combination

In (4.1) we showed how the average is calculated. The average is of course also a random variable. Let us show this linear combination; it will be of outmost importance later on when we go into the concept of variation.

$$\overline{X} = \frac{\sum X_i}{n} = \frac{1}{n} \cdot \sum X_i = \frac{1}{n} \cdot (X_1 + X_2 + X_3 + ... + X_n)$$
$$= \frac{1}{n} \cdot X_1 + \frac{1}{n} \cdot X_2 + \frac{1}{n} \cdot X_3 + ... + \frac{1}{n} \cdot X_n$$

If we replace all $1/n$ by the constants a_1, a_2, a_3, ... we get

$$\overline{X} = a_1 \cdot X_1 + a_2 \cdot X_2 + a_3 \cdot X_3 + ... + a_n \cdot X_n$$

Now we have the average more clearly as a linear combination. We will return to this rather special combination in the next section.

Some concluding notes

In (4.1) we showed how the average of a number of values was calculated. We have also stated that this average is another random variable. However $E(T)$, $E(X)$, $E(Z)$ etc are constants (fixed values) and we still have not discussed how these values are calculated in a given situation. We will wait until we have started the discussion of distributions in chapter 7.

Once the idea of the expected value of a random variable is accepted, the concept of linear combinations and their properties is easy to understand. What we have stated so far in this chapter about the expected value does not depend on such things as *dependence* between variables or any special *distribution*.

4.2.2 Standard deviation of linear combinations of variables

In (4.2.1) we discussed the concept of mean value of a variable that itself is a linear combination of other variables. Here we add the very important idea of standard deviation of such combinations. In (4.1) we introduced this by some examples and also pointed out the difference between the standard deviation calculated from a sample of data and the standard deviation of a random variable. The most pragmatic way to illustrate this difference is to say that the latter is the theoretical framework and the former is theory set in practice. (This difference we find everywhere in science: in physics there are a lot of theories and a lot of measures, based on a limited number of measurements, and it works well.) The general form of the linear combination of variables is the following:

$$T = a \cdot W + b \cdot X + c \cdot Y + d \cdot Z + ...$$

Calculation of the variance

The variance $V(T)$ is then calculated using the following formula:

$$V(T) = a^2 \cdot V(W) + b^2 \cdot V(X) + c^2 \cdot V(Y) + d^2 \cdot V(Z) + ...$$

The squaring of the constants comes from the original definition of the variance, a fact that we do not proof. However, it is important to know that when a constant is negative (see example 9) its square becomes positive. The expression above assumes that the variables are uncorrelated; otherwise there are extra terms to be included.

Example 1 (again)

This example is rather good because it describes a real problem and it is easily understood by most people. The model, using centigrade (T) and millimetres (Y), was

$$Y = 0.00671 \cdot Rh + 0.0110 \cdot T$$

If the variance of Rh, $V(Rh)$, is 4.5 percent^2 and the variance $V(T)$ of T is 2.3 °C^2 the variance of Y, $V(Y)$, becomes

$$\begin{aligned} V(Y) &= 0.00671^2 \cdot V(Rh) + 0.0110^2 \cdot V(T) \\ &= 0.00671^2 \cdot 4.5 + 0.0110^2 \cdot 2.3 = 0.000480 \end{aligned}$$

Example 9 (again)

If we apply the above reasoning on the model for the number of trouble reports after four weeks we get the following model:

$$T = k + X_1 + X_2 + X_3 + X_4 - Y_1 - Y_2 - Y_3 - Y_4$$

and if $V(X_i) = 5.6$ and $V(Y_i) = 7.8$ we get

$$V(T) = 4 \cdot V(X_1) + 4 \cdot V(Y_1) = 4 \cdot 5.6 + 4 \cdot 7.8 = 53.6$$

If we prefer the standard deviation we take the square root of the variance. There are at least two things to note. First, the constant k disappeared when calculating the variance because a constant has no variance. The second thing is that, as mentioned before, the minus sign was squared and became a plus sign.

Example	Input	Answer
2	$V(S_1) = 1.2$ $V(S_2) = 1.1$ $V(S_3) = 0.8$ $V(S_4) = 0.6$	$V(T) = 3.7$
3	$V(X_i) = 105$ lines/day 12 programmers	$V(Y) = 1260$

Standard deviation of the average

As we showed above the average is a linear combination of random variables. This formulation takes some time to get used to. However, it is of outmost importance to understand how its variance can be calculated from the original variables. If we apply the ordinary rules we get

$$V(\overline{X}) = \left(\frac{1}{n}\right)^2 \cdot V(X_1) + \left(\frac{1}{n}\right)^2 \cdot V(X_2) + \left(\frac{1}{n}\right)^2 \cdot V(X_3) + \ldots + \left(\frac{1}{n}\right)^2 \cdot V(X_n)$$

But because all observations come from the very same process their variances are equal so we get

$$V(\overline{X}) = n \cdot \left(\frac{1}{n}\right)^2 \cdot V(X) = \frac{1}{n} \cdot V(X)$$

and in different notations:

$$V(\overline{X}) = \frac{V(X)}{n} \qquad \text{that also can be stated as} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

This result reflects something quite natural, i.e. the average, when calculated from different samples, will show a smaller variation than the original values. The relationship between the variance (or standard deviation) of a process and the variance (or the standard deviation) of averages of samples from that process, will come back many times in statistics. Another point is worth a remark. Sometimes one can come across a value of the standard deviation that seems surprisingly small. A small standard deviation is of course very often something to be proud of. However, if the standard deviation is calculated on averages it might be that someone is trying to hide the true standard deviation!

A common trick to reduce the standard deviation, and thus the uncertainty, is to measure several times before calculating the result. This trick is used internally in many measuring devices. Let us imagine that we want to measure the weight of a certain item. The device we use does not give exactly the same answer every time. If this variation is stated as a standard deviation, we can calculate how many times we have to weigh the item to get a specified accuracy.

Some concluding notes

The additive feature of the variance is very important. To decrease the total variance the largest source of variance should be removed first. When introducing an operation into a system, one must be sure not to increase the total variance too much. Sometimes one finds engineers that in good spirit want to decrease the total variance by introducing another operation that is supposed to counteract a previous operation. Let's say e.g. that one operation stretches the item and to compensate this, a shrinking operation is introduced.

However, the shrinking operation most likely has a variance that adds to the total variance *unless there is a negative correlation between the stretching and shrinking operations*. A negative correlation would mean that whenever an item is stretched it is shrunk and whenever an item is not stretched it is not shrunk. In this way we decrease the total variance.

This is what happens at a final test or when we adjust e.g. some electrical feature of a circuit. By using some adjustable component we decrease the measured value if it is too high or increase it if too low. The inspector together with the component acts a random variable that is correlated with the circuit coming to the test. If the correlation is not exactly 1, the total variance will not be reduced to zero. This sounds quite natural. If the inspector does not read the value properly, he will either over or under compensate using the adjustable component and thus all circuits will not have the same electrical value. A final note on linear combinations is that all formulas above are correct irrespective of what statistical distribution that is considered (statistical distributions are discussed in chapter 8). However, the formulas for the variance have to include an extra term if there are correlations between variables.

4.2.3 Summary of linear combinations of variables

Let us assume that we have the following linear combination of random variables:

$$Y = a\cdot X_1 + b\cdot X_2 + c\cdot X_3 + d\cdot X_4$$

The mean value of the linear combination

$$E(Y) = a\cdot E(X_1) + b\cdot E(X_2) + c\cdot E(X_3) + d\cdot E(X_4)$$

or with a different notation

$$\mu_Y = a\cdot \mu_{X_1} + b\cdot \mu_{X_2} + c\cdot \mu_{X_3} + d\cdot \mu_{X_4}$$

The standard deviation of the linear combination

$$\sigma_Y = \sqrt{a^2\cdot V(X_1) + b^2\cdot V(X_2) + c^2\cdot V(X_3) + d^2\cdot V(X_4)}$$

or with a different notation

$$\sigma_Y = \sqrt{a^2\cdot \sigma_{X_1}^2 + b^2\cdot \sigma_{X_2}^2 + c^2\cdot \sigma_{X_3}^2 + d^2\cdot \sigma_{X_4}^2}$$

X_i = the random variables of the linear combination
μ_{X_i} = the true mean of the random variable X_i
σ_{X_i} = the true standard deviation of the random variable X_i

N.B. If there is a correlation between the X_i variables, the formula must contain the so-called covariance terms.

Important formulas for the average value

$$E(\overline{X}) = E(X) \quad \text{or, with a different notation,} \quad \mu_{\overline{X}} = \mu$$

$$\sigma_{\overline{X}} = \sqrt{\frac{V(X)}{n}} \quad \text{or, with a different notation,} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

n = the number of values in the sample
μ = the true mean value of the random variable X
σ = the true standard deviation of the random variable X

μ and σ are calculated on theoretical grounds. See chapter 7!