**Data Transformation**

Transformations: an introduction

-------------------------------------------------------------------------------

In data analysis **transformation** is the replacement of a variable by a

function of that variable: for example, replacing a variable x by the

square root of x or the logarithm of x. In a stronger sense, a

transformation is a replacement that changes the shape of a distribution

or relationship.

**Reasons for using transformations**

There are many reasons for transformation. The list here is not

comprehensive.

1. Convenience

2. Reducing skewness

3. Equal spreads

4. Linear relationships

5. Additive relationships

If you are looking at just one variable, 1, 2 and 3 are relevant, while

if you are looking at two or more variables, 4 and 5 are more important.

However, transformations that achieve 4 and 5 very often achieve 2 and 3.

1. **Convenience** A transformed scale may be as natural as the original

scale and more convenient for a specific purpose (e.g. percentages

rather than original data, sines rather than degrees).

One important example is **standardisation**, whereby values are adjusted for

differing level and spread. In general

value - level

standardised value = -------------.

spread

Standardised values have level 0 and spread 1 and have no units: hence

standardisation is useful for comparing variables expressed in different

units. Most commonly a **standard score** is calculated using the mean and

standard deviation (sd) of a variable:

x - mean of x

z = -------------.

sd of x

Standardisation makes no difference to the shape of a distribution.

2. **Reducing skewness** A transformation may be used to reduce skewness. A

distribution that is symmetric or nearly so is often easier to handle and

interpret than a skewed distribution. More specifically, a normal or

Gaussian distribution is often regarded as ideal as it is assumed by many

statistical methods.

To reduce right skewness, take roots or logarithms or reciprocals (roots

are weakest). This is the commonest problem in practice.

To reduce left skewness, take squares or cubes or higher powers.

3. **Equal spreads** A transformation may be used to produce approximately

equal spreads, despite marked variations in level, which again makes data

easier to handle and interpret. Each data set or subset having about the

same spread or variability is a condition called **homoscedasticity**: its

opposite is called **heteroscedasticity**. (The spelling **-sked-** rather than

**-sced-** is also used.)

4. **Linear relationships** When looking at relationships between variables,

it is often far easier to think about patterns that are approximately

linear than about patterns that are highly curved. This is vitally

important when using linear regression, which amounts to fitting such

patterns to data. (In Stata, regress is the basic command for

regression.)

For example, a plot of logarithms of a series of values against time has

the property that periods with **constant rates of change** (growth or

decline) plot as straight lines.

5. **Additive relationships** Relationships are often easier to analyse when

additive rather than (say) multiplicative. So

y = a + bx

in which two terms a and bx are added is easier to deal with than

y = ax^b

in which two terms a and x^b are multiplied. **Additivity** is a vital issue

in **analysis of variance** (in Stata, anova, oneway, etc.).

In practice, a transformation often works, serendipitously, to do several

of these at once, particularly to reduce skewness, to produce nearly

equal spreads and to produce a nearly linear or additive relationship.

But this is not guaranteed.

**Review of most common transformations**

The most useful transformations in introductory data analysis are the

reciprocal, logarithm, cube root, square root, and square. In what

follows, even when it is not emphasised, it is supposed that

transformations are used only over ranges on which they yield (finite)

real numbers as results.

*Reciprocal*

The **reciprocal**, x to 1/x, with its sibling the **negative reciprocal**, x to

-1/x, is a very strong transformation with a drastic effect on

distribution shape. It can not be applied to zero values. Although it

can be applied to negative values, it is not useful unless all values are

positive. The reciprocal of a ratio may often be interpreted as easily as

the ratio itself: e.g.

population density (people per unit area) becomes area per person;

persons per doctor becomes doctors per person;

rates of erosion become time to erode a unit depth.

(In practice, we might want to multiply or divide the results of taking

the reciprocal by some constant, such as 1000 or 10000, to get numbers

that are easy to manage, but that itself has no effect on skewness or

linearity.)

The reciprocal reverses order among values of the same sign: largest

becomes smallest, etc. The negative reciprocal preserves order among

values of the same sign.

*Logarithm*

The **logarithm**, x to log base 10 of x, or x to log base e of x (ln x), or

x to log base 2 of x, is a strong transformation with a major effect on

distribution shape. It is commonly used for reducing right skewness and

is often appropriate for measured variables. It can not be applied to

zero or negative values. One unit on a logarithmic scale means a

multiplication by the base of logarithms being used. Exponential growth

or decline

y = a exp(bx)

is made linear by

ln y = ln a + bx

so that the response variable y should be logged. (Here exp() means

raising to the power e, approximately 2.71828, that is the base of

natural logarithms.)

An aside on this **exponential growth or decline** equation: put x = 0, and

y = a exp(0) = a,

so that a is the amount or count when x = 0. If a and b > 0, then y grows

at a faster and faster rate (e.g. compound interest or unchecked

population growth), whereas if a > 0 and b < 0, y declines at a slower

and slower rate (e.g. radioactive decay).

Power functions y = ax^b are made linear by log y = log a + b log x so

that both variables y and x should be logged.

An aside on such **power functions**: put x = 0, and for b > 0,

y = ax^b = 0,

so the power function for positive b goes through the origin, which often

makes physical or biological or economic sense. Think: does zero for x

imply zero for y? This kind of power function is a shape that fits many

data sets rather well.

Consider ratios y = p / q where p and q are both positive in practice.

Examples are

males / females;

dependants / workers;

downstream length / downvalley length.

Then y is somewhere between 0 and infinity, or in the last case, between

1 and infinity. If p = q, then y = 1. Such definitions often lead to

skewed data, because there is a clear lower limit and no clear upper

limit. The logarithm, however, namely

log y = log p / q = log p - log q,

is somewhere between -infinity and infinity and p = q means that log y =

0. Hence the logarithm of such a ratio is likely to be more symmetrically

distributed.

*Cube root*

The **cube root**, x to x^(1/3). This is a fairly strong transformation with

a substantial effect on distribution shape: it is weaker than the

logarithm. It is also used for reducing right skewness, and has the

advantage that it can be applied to zero and negative values. Note that

the cube root of a volume has the units of a length. It is commonly

applied to rainfall data.

Applicability to negative values requires a special note. Consider

(2)(2)(2) = 8 and (-2)(-2)(-2) = -8. These examples show that the cube

root of a negative number has negative sign and the same absolute value

as the cube root of the equivalent positive number. A similar property is

possessed by any other root whose power is the reciprocal of an odd

positive integer (powers 1/3, 1/5, 1/7, etc.).

This property is a little delicate. For example, change the power just a

smidgen from 1/3, and we can no longer define the result as a product of

precisely three terms. However, the property is there to be exploited if

useful.

*Square root*

The **square root**, x to x^(1/2) = sqrt(x), is a transformation with a

moderate effect on distribution shape: it is weaker than the logarithm

and the cube root. It is also used for reducing right skewness, and also

has the advantage that it can be applied to zero values. Note that the

square root of an area has the units of a length. It is commonly applied

to counted data, especially if the values are mostly rather small.

*Square*

The **square**, x to x^2, has a moderate effect on distribution shape and it

could be used to reduce left skewness. In practice, the main reason for

using it is to fit a response by a quadratic function y = a + b x + c

x^2. Quadratics have a turning point, either a maximum or a minimum,

although the turning point in a function fitted to data might be far

beyond the limits of the observations. The distance of a body from an

origin is a quadratic if that body is moving under constant acceleration,

which gives a very clear physical justification for using a quadratic.

Otherwise quadratics are typically used solely because they can mimic a

relationship within the data region. Outside that region they may behave

very poorly, because they take on arbitrarily large values for extreme

values of x, and unless the intercept a is constrained to be 0, they may

behave unrealistically close to the origin.

Squaring usually makes sense only if the variable concerned is zero or

positive, given that (-x)^2 and x^2 are identical.

*Which transformation?*

The main criterion in choosing a transformation is: what works with the

data? As examples above indicate, it is important to consider as well two

questions.

What makes physical (biological, economic, whatever) sense, for example

in terms of limiting behaviour as values get very small or very large?

This question often leads to the use of logarithms.

Can we keep dimensions and units simple and convenient? If possible, we

prefer measurement scales that are easy to think about. The cube root of

a volume and the square root of an area both have the dimensions of

length, so far from complicating matters, such transformations may

simplify them. Reciprocals usually have simple units, as mentioned

earlier. Often, however, somewhat complicated units are a sacrifice that

has to be made.

**Psychological comments** - **for the puzzled**

The main motive for transformation is greater ease of description.

Although transformed scales may seem less natural, this is largely a

psychological objection. Greater experience with transformation tends to

reduce this feeling, simply because transformation so often works so

well. In fact, many familiar measured scales are really transformed

scales: decibels, pH and the Richter scale of earthquake magnitude are

all logarithmic.

However, transformations cause debate even among experienced data

analysts. Some use them routinely, others much less. Various views,

extreme or not so extreme, are slightly caricatured here to stimulate

reflection or discussion. For what it is worth, I consider all these

views defensible, or at least understandable.

"This seems like a kind of cheating. You don't like how the data are, so

you decide to change them."

"I see that this is a clever trick that works nicely. But how do I know

when this trick will work with some other data, or if another trick is

needed, or if no transformation is needed?"

"Transformations are needed because there is no guarantee that the world

works on the scales it happens to be measured on."

"Transformations are most appropriate when they match a scientific view

of how a variable behaves."

Often it helps to transform results back again, using the reverse or

**inverse** transformation:

reciprocal t = 1 / x reciprocal x = 1 / t

log base 10 t = log\_10 x 10 to the power x = 10^t

log base e t = log\_e x = ln x e to the power x = exp(t)

log base 2 t = log\_2 x 2 to the power x = 2^t

cube root t = x^(1/3) cube x = t^3

square root t = x^(1/2) square x = t^2