Data transformations

Why: To make graphs more symmetric or linear, to get data onto better scale

When: All values are positive (amounts), dotplots skewed, large range ($\frac{\text{largest}}{\text{smallest}} > 3$?)

How: Most common transformation is $x \longrightarrow \log(x)$.

Recall: $\log_a b = c$ if $a^c = b$. "c is the power to raise a to to get b."

$\log_{10} 1 =$, $\log_{10} 10 =$, $\log_{10} 10^{106} =$, $\log_{10} 0 =$

Use either base 10 (common) or e (natural) logs.

Properties:

$\log(a \cdot b) =$

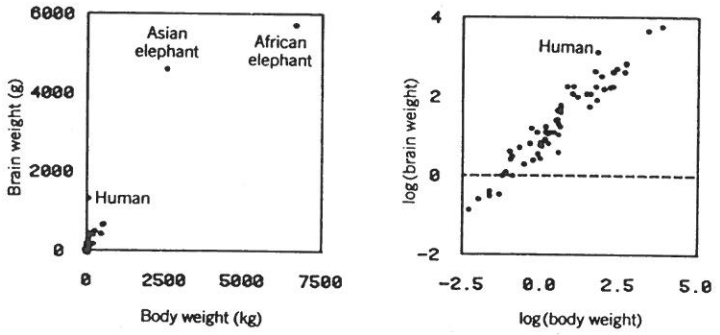
$\log(a/b) =$

$e^{\log_e(x)} =$; $\log_e e^x =$

Other possibilities:

$x \rightarrow 1/x$ (reciprocal, good for ratios)

$x \rightarrow \sqrt{x}$ (good for counts or areas)



Inference with transformed data

We sometimes transform data by square root, log, reciprocal, or some other transformation. The object is to achieve distributions that are more symmetric, (and scatterplots that are straighter). Some tip-offs that a transform will help are:

-
-
-
-

When we decide to transform data, the new data will be more normally distributed. So, inference (confidence intervals, tests) that require normality will be more valid.

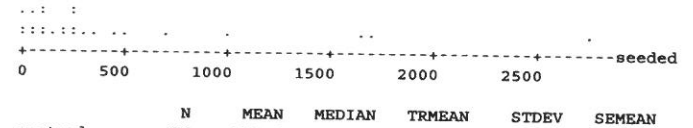
Often, results are just presented on the transformed scale. Common ~~examples~~ of log scales:

Sometimes results are transformed back to the original scale. For CI's, we inverse-transform the edges of the CI. This gives a CI for the median of the original quantity. (P-values are not transformed back.)

Example: Rainfall and cloud seeding

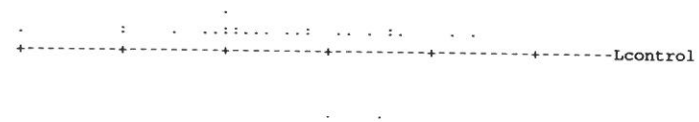
Chambers, Cleveland, Kleiner and Tukey (1983) Graphical methods for data analysis give rainfall in acre-feet from 52 clouds, of which 26 were chosen randomly and seeded with silver oxide.

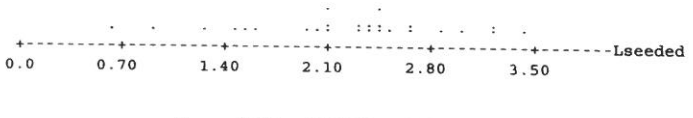
Rainfall from control clouds				Rainfall from seeded clouds			
4.9	830.1	36.6	87.0	198.6	302.8	242.5	1656.0
95.0	26.1	21.7	372.4	32.7	274.7	31.4	4.1
29.0	345.5	41.1	4.9	274.7	2745.6	200.7	40.6
28.6	321.2	1202.6	163.3	978.0	7.7	119.0	334.1
11.5	17.3	26.3	244.3	430.0	129.6	255.0	703.4
1.0	147.8	24.4		1697.8	118.5	17.5	
81.2	47.3	68.5		115.3	489.1	92.4	



	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
control	26	164.6	44.2	128.2	278.4	54.6
seeded	26	442	222	364	651	128

Log10 rainfall





	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
Lcontrol	26	1.733	1.644	1.749	0.713	0.140
Lseeded	26	2.230	2.344	2.247	0.695	0.136

Ex: a 95% CI for median rainfall from control clouds is

$$\left(10^{1.733 + t_{25} \times \frac{.713}{\sqrt{26}}},\ 10^{1.733 + t_{25} \times \frac{.713}{\sqrt{26}}}\right)$$

(27.5, 106.4) (median rainfall = 44.2)

Logarithms

Logarithms are not simply a method of calculation dating from before the computer age, but a set of fundamental mathematical functions. Because of their special properties they are much used in statistics. We shall start with logarithms (or logs for short) to base 10, the common logarithms used in calculations. The log to base 10 of a number x is y where

$$x = 10^y$$

We write $y = \log_{10}(x)$. Thus for example $\log_{10}(10) = 1$, $\log_{10}(100) = 2$, $\log_{10}(1\,000) = 3$, $\log_{10}(10\,000) = 4$, and so on. If we multiply two numbers, the log of the product is the sum of their logs:

$$\log(xy) = \log(x) + \log(y)$$

For example,

$$100 \times 1\,000 = 10^2 \times 10^3 = 10^{2+3} = 10^5 = 100\,000$$

Or in log terms:

$$\log_{10}(100 \times 1\,000) = \log_{10}(100) + \log_{10}(1\,000) = 2 + 3 = 5$$

Hence, $100 \times 1\,000 = 10^5 = 100\,000$. This means that any multiplicative relationship of the form

$$y = a \times b \times c \times d$$

can be made additive by a log transformation:

$$\log(y) = \log(a) + \log(b) + \log(c) + \log(d)$$

This is the process underlying the fit to the Lognormal Distribution described in §7.4.

There is no need to use 10 as the base for logarithms. We can use any number. The log of a number x to base b can be found from the log to base a by a simple calculation:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$$

Ten is convenient for arithmetic using log tables, but for other purposes it is less so. For example, the gradient, slope or differential of the curve $y = \log_{10}(x)$ is $\log_{10}(e)/x$, where $e = 2.718\,281\ldots$ is a constant which does not depend on the base of the logarithm. This leads to awkward constants spreading through formulae. To keep this to a minimum we use logs to the base e, called natural or Napierian logarithms after the mathematician John Napier. This is the logarithm usually produced by LOG(X) functions in computer languages.

Figure 5.11 shows the log curve for three different bases, 2, e and 10. The curves all go through the point (1,0), i.e. $\log(1) = 0$. As x approaches 0, $\log(x)$ becomes a larger and larger negative number, tending towards minus infinity as x tends to zero. There are no logs of negative numbers. As x increases from 1, the curve becomes flatter and flatter. Though $\log(x)$ continues to increase, it does so more and more slowly. The curves all go through $(base, 1)$ i.e. $\log(base) = 1$. The curve for log to the base 2 goes through (2,1), (4,2), (8,3) because $2^1 = 2$, $2^2 = 4$, $2^3 = 8$. We can see that the effect of replacing data by their logs will be to stretch out the scale at the lower end and contract it at the upper.

We often work with logarithms of data rather than the data themselves. This may have several advantages. Multiplicative relationships may become additive, curves may become straight lines and skew distributions may become symmetrical.

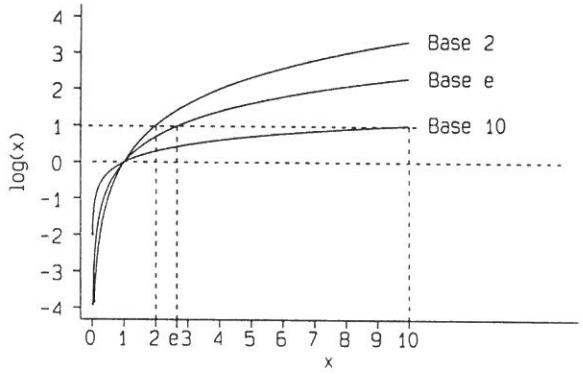


Fig. 5.11. Logarithmic curves to three different bases

We transform back to the natural scale using the **antilogarithm** or **antilog**. If $y = \log_{10}(x)$, $x = 10^y$ is the antilog of y. If $z = \log_e(x)$, $x = e^z$ or $x = \exp(z)$ is the antilog of z. If your computer program doesn't transform back, most calculators have e^x and 10^x functions for this purpose.