Transformations of Data

A transformation of data is the application of some function to all values of a particular variable. Transformations are used in statistical analysis for a few reasons:

1. There is no good natural distribution that models a response variable. In this case, a transformation may be applied to this variable so that the transformed values can be modelled by a normal distribution.
2. The relationship between a predictor and a response variable is not fit well by a linear relationship. In this case, either the predictor or response variable may be transformed to bring the relationship closer to linear.
3. The residuals from a model are heteroscedastic, that is, the spread of the residuals from the model depend on the values of the predictor. In this case, a transformation of the response variable may correct for this problem.

Often, more than one of these issues will be present at once, and an appropriate transformation will correct all of them.

Different transformations work for different data types. Below is a list of examples of possible transformations for different data types:

- **Logarithms:** Log transforms are used if the relationship between a predictor and a response follows an exponential curve. For instance, growth of a system as a function of time is often exponential. Log transforms are also appropriate if the variance of a distribution increases with the mean, or if a distribution is highly skewed to the right. The base of the logarithm used for the transformation will affect the result, with higher bases (such as 10) pulling in a long tail on a distribution more than a lower base log (such as e). Less severe skews are better corrected with a square root transformation.
  Commonly occurring response variables that need a log transformation include: Response time data in linguistic or psychology experiments; income; concentrations of chemicals, e.g. Phosphorous concentration in soil.
  When using this transformation it may be necessary to know the basic algebra of logs. Listed below are some basic principles:

  - Addition (Multiplication): Log_c(a) + Log_c(b) = Log_c(ab)
  - Subtraction (Division): Log_c(a) – Log_c (b) = $\frac{Log_c(a)}{Log_c(b)}$
  - Log_c (a^b) = b Log_c(a)
  - $Log_c(a) = \frac{1}{Log_a(c)}$

  The inverse of the logarithm function is the exponential function e. The basic principles of exponentials are listed below:

- ❖ Multiplication: $e^a \times e^b = e^{a+b}$
- ❖ Division: $e^a \div e^b = e^{a-b}$
- ❖ $e^{-a} = \frac{1}{e^a}$
- ❖ $e^0 = 1$
- ❖ $e^{\frac{a}{b}} = \sqrt[b]{e^a}$

- **Reciprocal:** If a log transform does not normalise the distribution of your response variable, you could try a reciprocal $\frac{1}{x}$ transformation. This is often used for enzyme reaction rate data.

- **Square root:** This transform is sometimes used for count data, e.g. blood cells on a haemocytometer or woodlice in a garden. Carrying out a square root transform will convert data with a Poisson distribution to a normal distribution. Generally, however, it is considered better to use Poisson inference procedures for count data than to normalise the data and use normal distribution methods.

- **Arcsine:** This transformation is also known as the angular transformation and is especially useful for data given as percentages and proportions, such as percentage or proportion of students in various schools requiring additional tuition. (Note that the percent of students is simply 100 times the fractional proportion of students expressed as a decimal, so, e.g. a proportion of ¼ of students is the same as 25% of the students.)

Important: Transformations need to be approached with caution!

Even though a transformation may transform data to the normal distribution, there are inference procedures for the other distributions that will provide more accurate conclusions regarding your sample data than you would obtain by normalising the data and computing normal parametric tests on this data. In addition, it can be difficult to interpret the results of inference on transformed data, as the results are not expressed on the same scale as the original data. It is always a wise plan to check the literature in your discipline to see how data such as what you are using has been analysed by other researchers.

Tukey's ladder of transformations

There are a variety of transformations that can be used to correct for skewing to a greater or lesser extent in a distribution with a single maximum (called a unimodal distribution). The correct transformation to use will depend on both the direction and extent of skew. It is possible to over-correct by using too powerful a transformation and change the direction of the skew. For example, a small amount of downward

skewing (a long tail to the left) might be over-balanced by squaring the measurements and result in an upward skew distribution (a long tail to the right).

Tukeys ladder of transformations (shown below) gives several common transformations to correct skew in each direction and illustrates the relative effectiveness of these.

$$\frac{-1}{x^2}, \quad \frac{-1}{x}, \quad \log(x), \quad \ln(x), \quad \sqrt{x}$$

Correct Upward Skew

$$x^2, \quad x^3, \quad antilog(x)$$

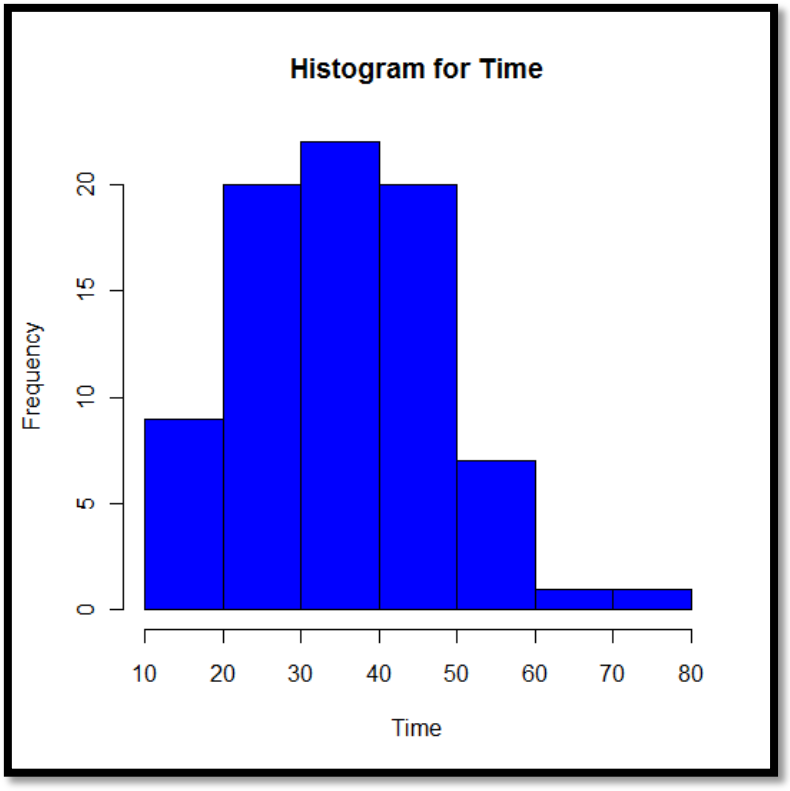
Correct Downward Skew

Example:

Data set of waiting time in minutes of 80 patients in a hospital:

11,15,18,10,19,19,20,20,20,21,25,26,24,23,29,25,28,21,26,25,24,25,28,26,30,30,30,30,30,31,32,32,36,39,31,35,34,36,38,39,35,32,36,36,36,32,31,31,35,40,40,41,43,47,48,48,49,47,46,45,41,42,45,46,43,46,48,49,45,41,50,51,58,56,53,54,52,60,65,78.

Represented in a histogram:

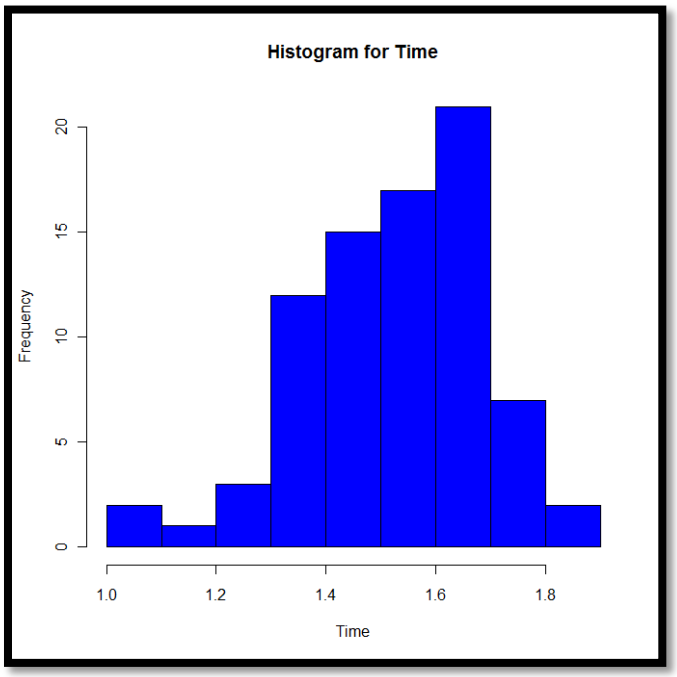


As you can see above the data is skewed upwards (long tail to the right). If we apply the log transformation to our data it will correct the upwards skew. Taking the logs of our data:

x	Log(x)
10	1
11	1.041393
15	1.176091
18	1.255273
19	1.278754
19	1.278754
20	1.30103
20	1.30103
20	1.30103
21	1.322219
21	1.322219
23	1.361728
24	1.380211
24	1.380211
25	1.39794
25	1.39794
25	1.39794
25	1.39794
26	1.414973
26	1.414973
26	1.414973
28	1.447158
28	1.447158
29	1.462398
30	1.477121
30	1.477121
30	1.477121
30	1.477121
30	1.477121
31	1.491362
31	1.491362
31	1.491362
31	1.491362
32	1.50515
32	1.50515
32	1.50515
32	1.50515
34	1.531479
35	1.544068
35	1.544068
35	1.544068
36	1.556303
36	1.556303
36	1.556303
36	1.556303
36	1.556303
36	1.556303
38	1.579784
39	1.591065
39	1.591065
40	1.60206
40	1.60206
41	1.612784
41	1.612784
41	1.612784
42	1.623249
43	1.633468
43	1.633468
45	1.653213
45	1.653213
45	1.653213
46	1.662758
46	1.662758
46	1.662758

47	1.672098
47	1.672098
48	1.681241
48	1.681241
48	1.681241
49	1.690196
50	1.69897
51	1.70757
52	1.716003
53	1.724276
54	1.732394
56	1.748188
58	1.763428
60	1.778151
65	1.812913
78	1.892095

The following histogram is produced:

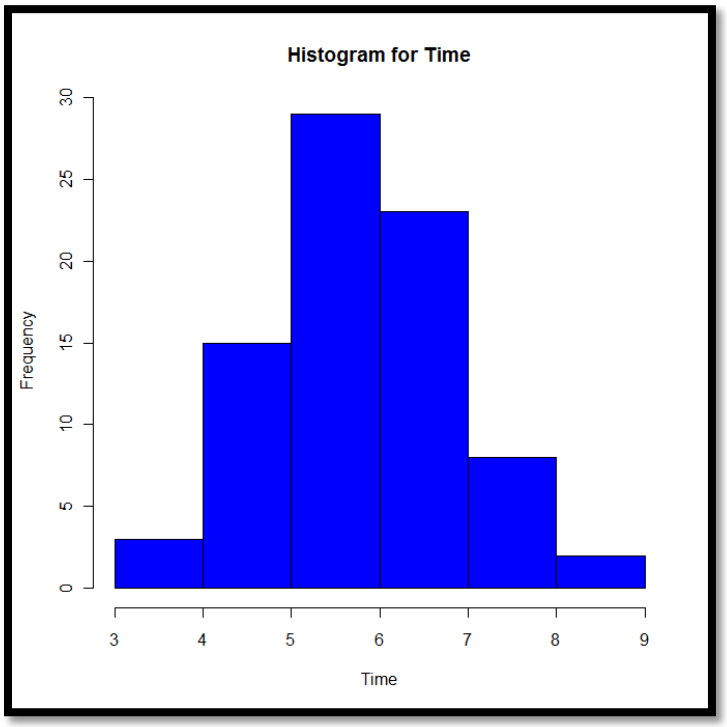


We can see in the histogram above, the log transformation is an over balance, and the new transformed data is now skewed downwards (long tail to the left). Therefore the more appropriate transformation method may be square root. Taking the square root of our original data:

x	$\sqrt{x}$
10	3.162278
11	3.316625
15	3.872983
18	4.242641
19	4.358899
19	4.358899
20	4.472136
20	4.472136
20	4.472136
21	4.582576
21	4.582576
23	4.795832
24	4.898979
24	4.898979
25	5

25	5
25	5
25	5
26	5.09902
26	5.09902
26	5.09902
28	5.291503
28	5.291503
29	5.385165
30	5.477226
30	5.477226
30	5.477226
30	5.477226
30	5.477226
31	5.567764
31	5.567764
31	5.567764
31	5.567764
32	5.656854
32	5.656854
32	5.656854
32	5.656854
34	5.830952
35	5.91608
35	5.91608
35	5.91608
36	6
36	6
36	6
36	6
36	6
36	6
38	6.164414
39	6.244998
39	6.244998
40	6.324555
40	6.324555
41	6.403124
41	6.403124
41	6.403124
42	6.480741
43	6.557439
43	6.557439
45	6.708204
45	6.708204
45	6.708204
46	6.78233
46	6.78233
46	6.78233
47	6.855655
47	6.855655
48	6.928203
48	6.928203
48	6.928203
49	7
50	7.071068
51	7.141428
52	7.211103
53	7.28011
54	7.348469
56	7.483315
58	7.615773
60	7.745967
65	8.062258
78	8.831761

The following histogram is produced:



We can see in the histogram above, the square root transformation has normalised the data, indicating that this is the appropriate transformation method for normalising this data set. Choosing a correct transformation method can be a matter of trial and error.