Topic: Exploratory Data Analysis (EDA)

Nonlinear Transformations

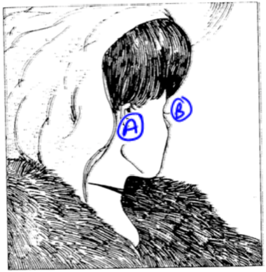
School of Mathematics and Applied Statistics

UNIVERSITY OF WOLLONGONG AUSTRALIA

Why Transform Data?

- Convert units (a linear transformation) $y_i = a + bx_i$
- To see data from different perspectives
- Spread out dense clusters
- Contract gaps between values in one tail
- Reduce asymmetry and make numerical summaries representative of data
- To make curved lines straight
- To fulfill assumptions underlying statistical tests about data

Different Perspectives: What do you see?

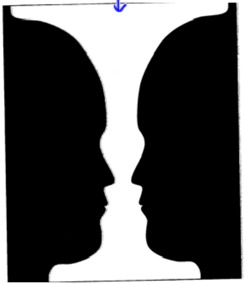


Weiten, 1989, p125

Different Perspectives: What do you see?

vase

Weiten, 1989, p125



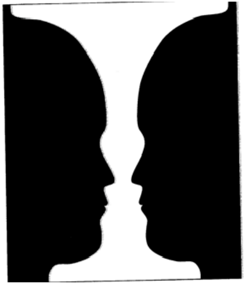
Weiten, 1989, p123

profiles

Transformations allow us to see data from different perspectives

Different Perspectives: What do you see?

Weiten, 1989, p125



Weiten, 1989, p123

Premonition, 2007 http://www.finerminds.com/metaphysical/premonition/

Transformations allow us to see data from different perspectives

Nonlinear Transformations

Transform each data point x_i by taking the square root i.e. $y_i = \sqrt{x_i}$

Discuss: What is the effect of taking the square root for these data points?

It ____________________ a tail of high values

In reverse:

Discuss: What is the effect of squaring each value? Consider $x_i = y_i^2$

It ____________________ the upper tail values

Nonlinear Transformations

Transform each data point x_i by taking the square root i.e. $y_i = \sqrt{x_i}$

Discuss: What is the effect of taking the square root for these data points?

It ___contracts___ a tail of high values
or pulls in

In reverse:

Discuss: What is the effect of squaring each value? Consider $x_i = y_i^2$

It ___stretches___ the upper tail values ⇒ larger spread.

Example: Brain Weights of mammals

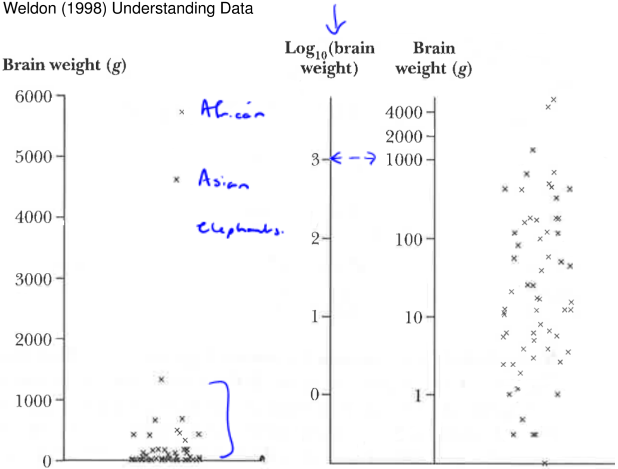
The average brain weights of 62 species of mammals

Species	Brain weight (g)	Log_{10} (brain wt)	Species	Brain weight (g)	Log_{10} (brain wt)
Arctic fox	44.50	1.648	Human	1320	3.121
Owl monkey	15.50	1.190	African elephant	5712	3.757
Mountain beaver	8.100	0.908	Water opossum	3.900	0.591
Cow	423.0	2.626	Rhesus monkey	179.0	2.253
Gray wolf	119.5	2.077	Kangaroo	56.00	1.748
Goat	115.0	2.061	Yellow-bellied marmot	17.00	1.230
Roe deer	98.20	1.992	Golden hamster	1.000	0.000
Guinea pig	5.500	0.740	Mouse	0.400	−0.398
Vervet	58.00	1.763	Little brown bat	0.250	−0.602
Chinchilla	6.400	0.806	Slow loris	12.50	1.097
Ground squirrel	4.000	0.602	Okapi	490.0	2.690
Arctic ground squirrel	5.700	0.756	Rabbit	12.10	1.083
African giant pouched rat	6.600	0.820	Sheep	175.0	2.243
Lesser short-tailed shrew	0.140	−0.854	Jaguar	157.0	2.196
Star-nosed mole	1.000	0.000	Chimpanzee	440.0	2.643
Nine-banded armadillo	10.80	1.033	Baboon	179.5	2.254
Tree hydrax	12.30	1.090	Desert hedgehog	2.400	0.380
N. American opossum	6.300	0.799	Giant armadillo	81.00	1.908
Asian elephant	4603	3.663	Rock hyrax (*P. habess.*)	21.00	1.322
Big brown bat	0.300	−0.523	Raccoon	39.20	1.593
Donkey	419.0	2.622	Rat	1.900	0.279
Horse	655.0	2.816	E. American mole	1.200	0.079
European hedgehog	3.500	0.544	Mole rat	3.000	0.477
Patas monkey	115.0	2.061	Musk shrew	0.330	−0.481
Cat	25.60	1.408	Pig	180.0	2.255
Galago	5.000	0.699	Echidna	25.00	1.398

Example: Brain Weights of mammals

The average brain weights of 62 species of mammals

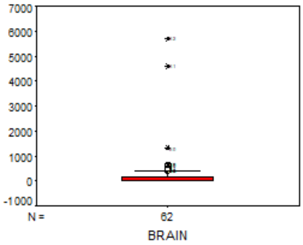
Source: p.39 Griffiths, Stirling & Weldon (1998) Understanding Data

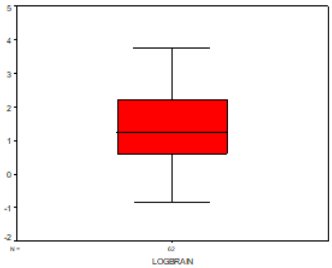


Example: Brain Weights of mammals

The average brain weights of 62 species of mammals:

Original scale (g) and after applying a log (base 10) transformation:





Discuss: What effect has the transformation had?

Power Transformations of Data

p		Effect on whole positive values
2	square	Pushes out upper tail and contracts lower tail
1	identity	leaves unchanged
1/2	square root	Contracts upper tail of whole numbers, pulls in outliers & spreads out lower tail
1/3	Cube root	Contracts upper tail of whole numbers
-1	reciprocal	Order of magnitude is reversed
-1/2	Reciprocal square root	Order of magnitude is reversed
-2	reciprocal square	Order of magnitude is reversed

$$\ln\left(\frac{x}{1-x}\right)$$

Source: p.40 Griffiths, Stirling & Weldon (1998) Understanding Data

Transformations - Reporting

It is important to:

- Always state clearly that a transformation has been applied
- Data values should usually be transformed back when results are reported.
- Example: Brain weights
  Median of approximately 1.2 is in log base 10 units
  So to get the median in the original units we have $10^{1.2}$ = 15.85g

Example: Richter scale

An example of a **nonlinear transformation** is the Richter scale

- transforms the measured intensity of earthquakes to a logarithmic scale

Video: How does the Richter Scale work? 4.56mins

https://www.youtube.com/watch?v=NaNw9LHq9dc

Notes: 25 April 2015 Nepal Earthquake

- __________ on Richter scale
- ________________ on Modified Mercalli scale

This complicated cocktail of factors have to be considered to help determine how earthquakes are experienced & categorized.

Z score transformation

This **transforms** the variable X into variable Z such that

$$z_i = \frac{x_i - \bar{x}}{s_x} = \frac{x_i}{s_x} - \frac{\bar{x}}{s_x} = \left(-\frac{\bar{x}}{s_x}\right) + \left(\frac{1}{s_x}\right)x_i$$

$a + b x_i$

so that the transformed variable Z has

Mean=0 and Standard deviation $s_z = 1$

When two sets of scores are transformed in this manner, each with their own mean and standard deviation we can **compare the standardised scores.**

This is especially useful if the data follow an approximately **normal distribution.**

Activity: Standardised Scores

NAPLAN Year 7 Numeracy

Figure 7.N1: Achievement of Year 7 Students in Numeracy, by State and Territory, 2015.

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT	Aust
Mean scale score / (S.D.)	546.7 (74.4)	548.4 (66.1)	538.9 (62.9)	538.3 (67.3)	532.7 (60.7)	528.8 (60.3)	549.4 (65.7)	484.7 (81.2)	542.5 (68.6)

Leah sat the Year 7 NAPLAN test in Qld and Kate sat it in ACT and Diana sat it in NT. Each achieved a mark of 590.

Relative to the state they live in, who performed better in Numeracy?

Activity: Standardised Scores

	NSW	Vic	Qld	WA	SA	Tas	ACT	NT	Aust
Mean scale score / (S.D.)	546.7 (74.4)	548.4 (66.1)	538.9 (62.9)	538.3 (67.3)	532.7 (60.7)	528.8 (60.3)	549.4 (65.7)	484.7 (81.2)	542.5 (68.6)

Leah (Qld)

$$z_L = \frac{x_L - \bar{x}_Q}{s_Q}$$

$$= \frac{590 - 538.9}{62.9}$$

$$= 0.8124$$

2

Kate (ACT)

$$z_K = \frac{590 - 549.4}{65.7}$$

$$= 0.6180$$

3

Diana (NT)

$$z_D = \frac{590 - 484.7}{81.2}$$

$$= 1.2968$$

1

Overview: Specific transformations

Linear transformations

- Add and subtract constants
- Multiply and divide by constants
- Z scores to standardise data

Nonlinear transformations

- Square root (cube root, . . .)
- Square (cube, . . .)
- Logarithm
- Reciprocal
- Logit
- Helpful when statistical tests require assumptions about population from which data has been drawn such as Normal distribution.