Introduction to Statistics for Research

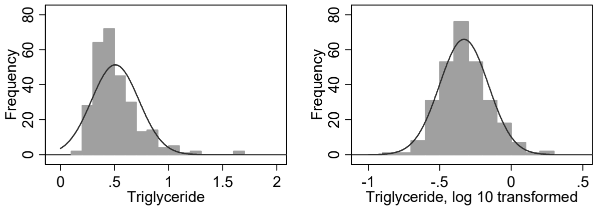
Transformations

Martin Bland
Emeritus Professor of Health Statistics
University of York
http://martinbland.co.uk/

The need for transformations

Instead of analysing the data as observed, we can carry out a mathematical transformation first. These can make data more suitable for analysis.

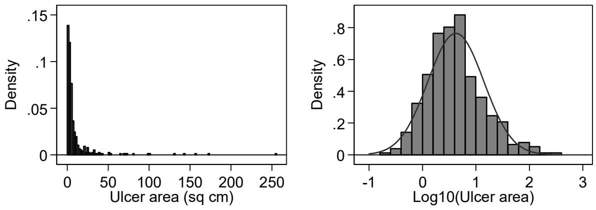
Serum triglyceride from cord blood



The need for transformations

Instead of analysing the data as observed, we can carry out a mathematical transformation first. These can make data more suitable for analysis.

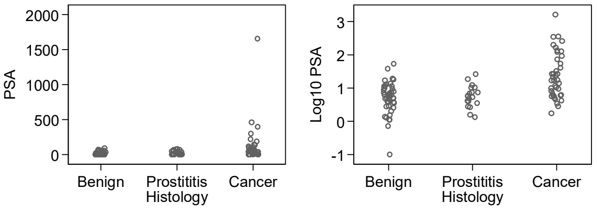
Area of venous ulcer at recruitment, VenUS I trial



The need for transformations

Instead of analysing the data as observed, we can carry out a mathematical transformation first. These can make data more suitable for analysis.

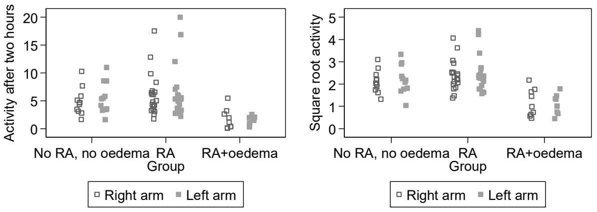
Prostate specific antigen (PSA) by prostate diagnosis



The need for transformations

Instead of analysing the data as observed, we can carry out a mathematical transformation first. These can make data more suitable for analysis.

Arm blood flow in rheumatoid arthritis with oedema

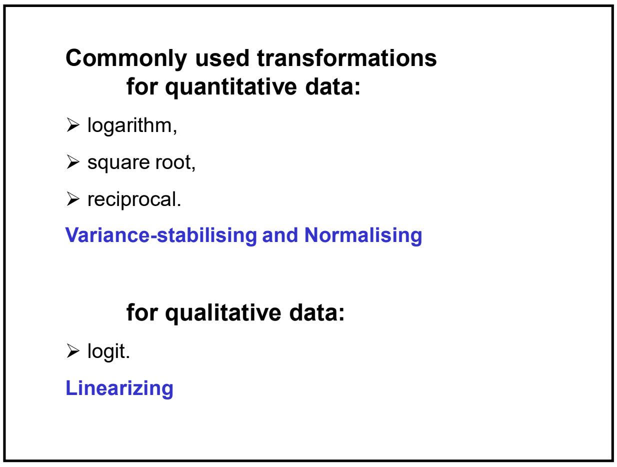


The need for transformations

It can be shown that if we take several samples from the same population, the means and variances of these samples will be independent if and only if the distribution is Normal.

Thus uniform variances tend to go with a Normal Distribution.

Addition and Normal distribution go together, so transformations which Normalize will often linearize.



Commonly used transformations for quantitative data:

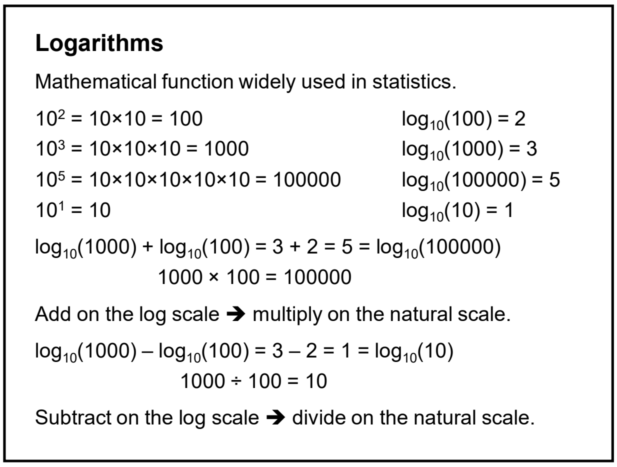
- logarithm,
- square root,
- reciprocal.

Variance-stabilising and Normalising

for qualitative data:

- logit.

Linearizing



Logarithms

Mathematical function widely used in statistics.

$10^2 = 10 \times 10 = 100$	$\log_{10}(100) = 2$
$10^3 = 10 \times 10 \times 10 = 1000$	$\log_{10}(1000) = 3$
$10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100000$	$\log_{10}(100000) = 5$
$10^1 = 10$	$\log_{10}(10) = 1$

$\log_{10}(1000) + \log_{10}(100) = 3 + 2 = 5 = \log_{10}(100000)$

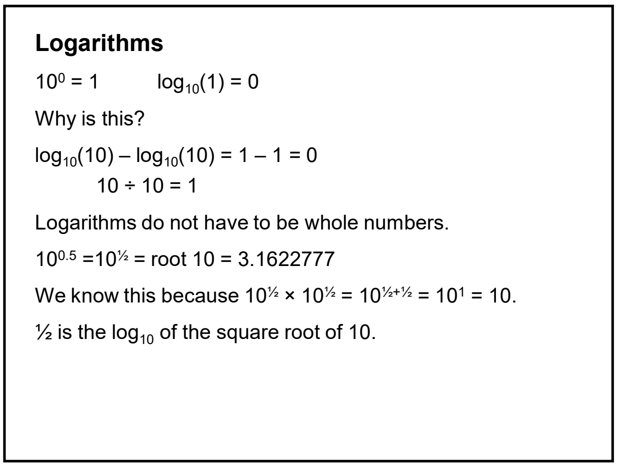
$1000 \times 100 = 100000$

Add on the log scale ➔ multiply on the natural scale.

$\log_{10}(1000) - \log_{10}(100) = 3 - 2 = 1 = \log_{10}(10)$

$1000 \div 100 = 10$

Subtract on the log scale ➔ divide on the natural scale.



Logarithms

$10^0 = 1$ $\log_{10}(1) = 0$

Why is this?

$\log_{10}(10) - \log_{10}(10) = 1 - 1 = 0$

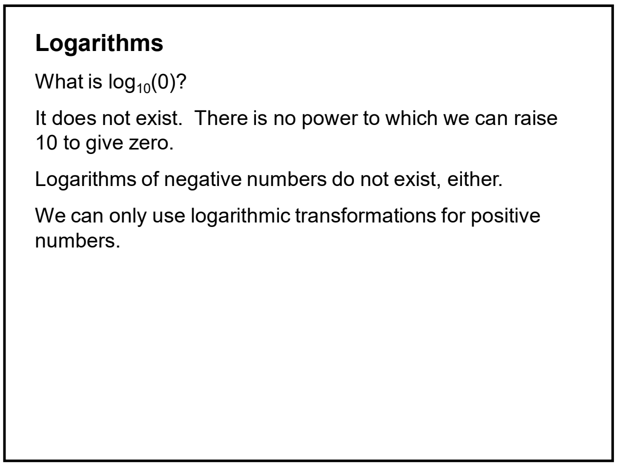
$10 \div 10 = 1$

Logarithms do not have to be whole numbers.

$10^{0.5} = 10^{½} =$ root 10 = 3.1622777

We know this because $10^{½} \times 10^{½} = 10^{½+½} = 10^1 = 10$.

½ is the $\log_{10}$ of the square root of 10.



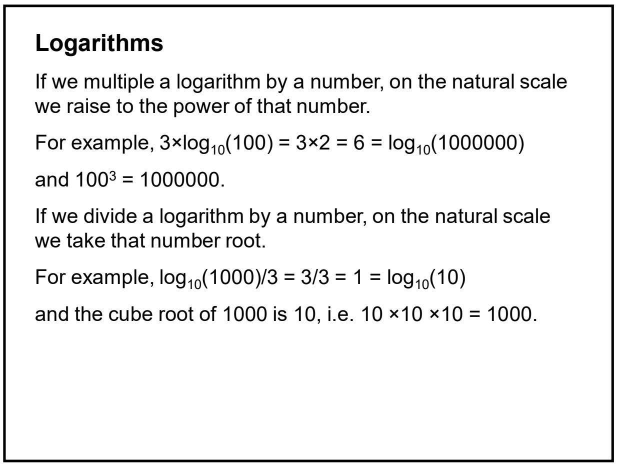
Logarithms

What is $\log_{10}(0)$?

It does not exist. There is no power to which we can raise 10 to give zero.

Logarithms of negative numbers do not exist, either.

We can only use logarithmic transformations for positive numbers.



Logarithms

If we multiple a logarithm by a number, on the natural scale we raise to the power of that number.

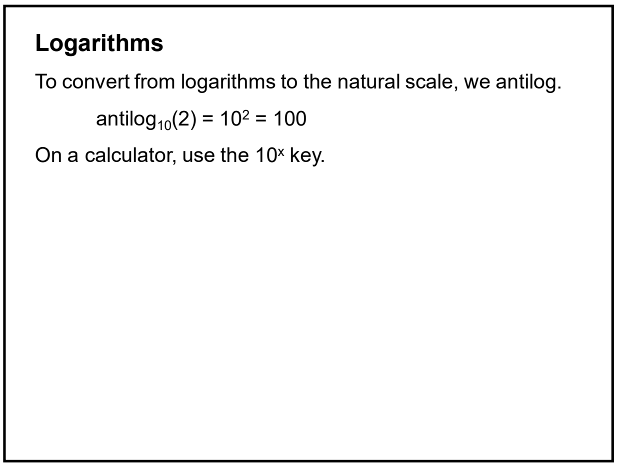
For example, $3\times\log_{10}(100) = 3\times2 = 6 = \log_{10}(1000000)$

and $100^3 = 1000000$.

If we divide a logarithm by a number, on the natural scale we take that number root.

For example, $\log_{10}(1000)/3 = 3/3 = 1 = \log_{10}(10)$

and the cube root of 1000 is 10, i.e. $10 \times 10 \times 10 = 1000$.



Logarithms

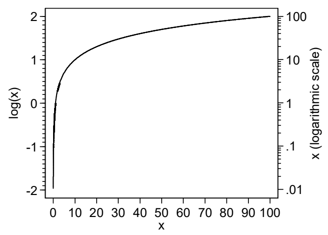
To convert from logarithms to the natural scale, we antilog.

$\text{antilog}_{10}(2) = 10^2 = 100$

On a calculator, use the 10^x key.

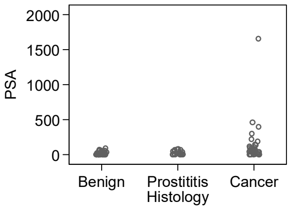
Logarithms

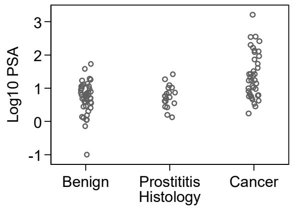
The logarithmic curve and logarithmic scale

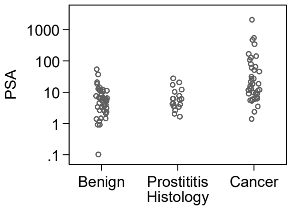


Logarithms

Logarithmic scales







Logarithms

We can use logarithms to multiply or divide large numbers.

Logarithms to the base 10 are called **common logarithms**.

They were used for calculation before the age of cheap electronic calculators.

Mathematicians find it convenient to use a different base, called ‘e’, to give **natural logarithms**.

Logarithms

Mathematicians find it convenient to use a different base, called ‘e’, to give **natural logarithms**.

‘e’ is a number which cannot be written down exactly, like π.

e = 2.718281 . . .

They use this because the slope of the curve

$y = \log_{10}(x)$

is $\log_{10}(e)/x$. The slope of the curve

$y = \log_e(x)$

is $1/x$.

Using natural logs avoids awkward constants in formulae.

When you see ‘log’ written in statistics, it is the natural log.

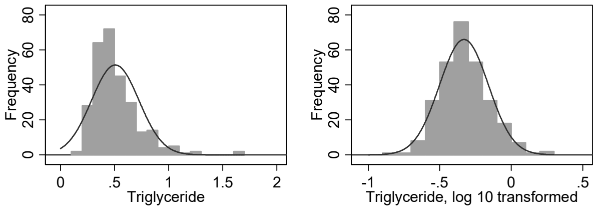
Logarithms

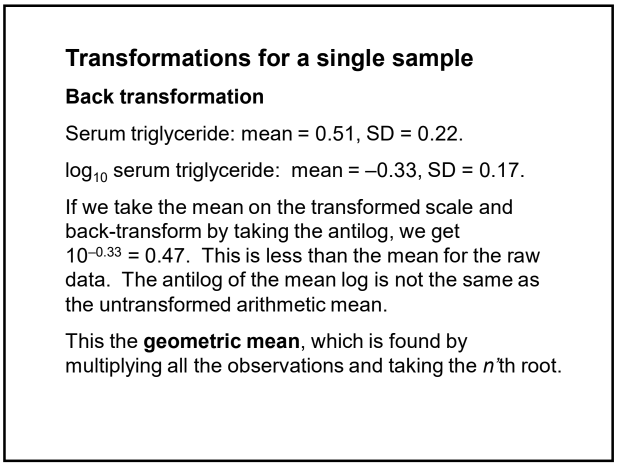
To antilog from logs to base e on a calculator, use the key labelled ‘e^x’ or ‘exp(x)’.

Transformations for a single sample

Serum triglyceride and $\log_{10}$ serum triglyceride in cord blood for 282 babies, with corresponding Normal Distribution curve.

The log transformation gives a good fit to the Normal.





Transformations for a single sample

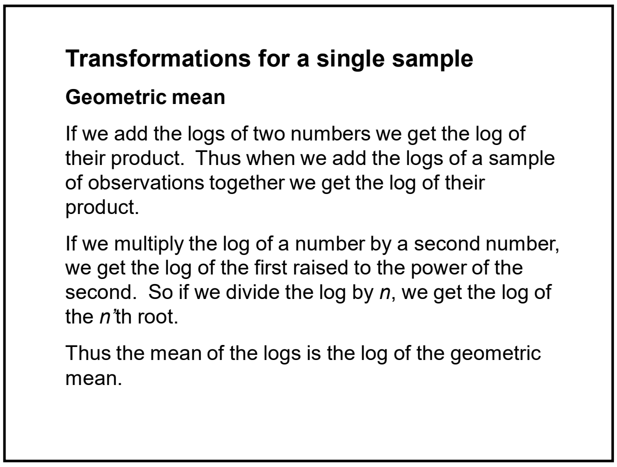
Back transformation

Serum triglyceride: mean = 0.51, SD = 0.22.

$\log_{10}$ serum triglyceride: mean = –0.33, SD = 0.17.

If we take the mean on the transformed scale and back-transform by taking the antilog, we get $10^{-0.33} = 0.47$. This is less than the mean for the raw data. The antilog of the mean log is not the same as the untransformed arithmetic mean.

This the **geometric mean**, which is found by multiplying all the observations and taking the *n*’th root.



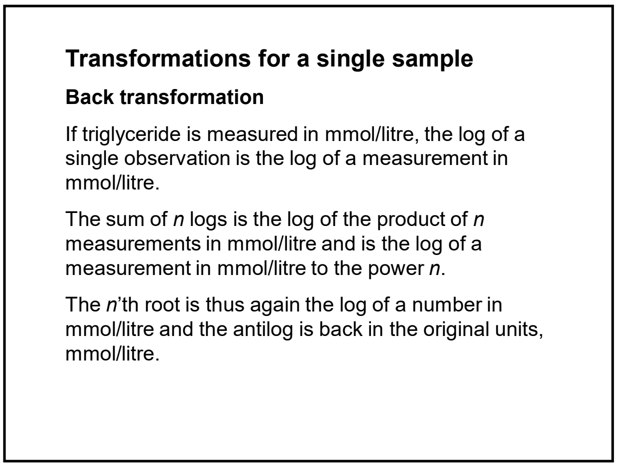
Transformations for a single sample

Geometric mean

If we add the logs of two numbers we get the log of their product. Thus when we add the logs of a sample of observations together we get the log of their product.

If we multiply the log of a number by a second number, we get the log of the first raised to the power of the second. So if we divide the log by *n*, we get the log of the *n*’th root.

Thus the mean of the logs is the log of the geometric mean.



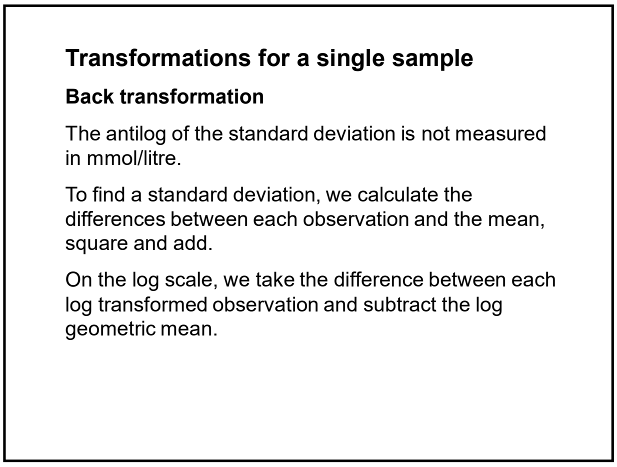
Transformations for a single sample

Back transformation

If triglyceride is measured in mmol/litre, the log of a single observation is the log of a measurement in mmol/litre.

The sum of *n* logs is the log of the product of *n* measurements in mmol/litre and is the log of a measurement in mmol/litre to the power *n*.

The *n*’th root is thus again the log of a number in mmol/litre and the antilog is back in the original units, mmol/litre.



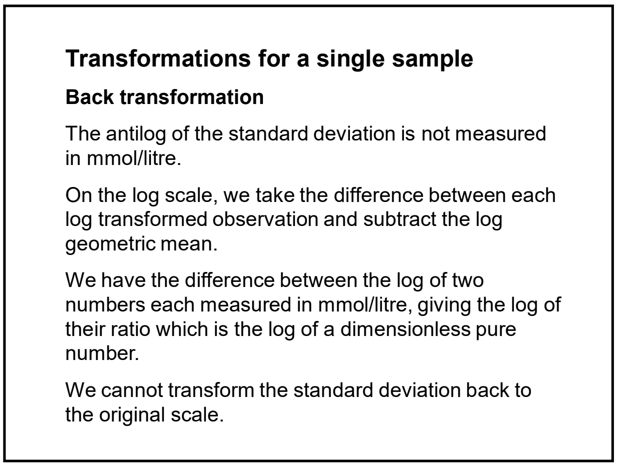
Transformations for a single sample

Back transformation

The antilog of the standard deviation is not measured in mmol/litre.

To find a standard deviation, we calculate the differences between each observation and the mean, square and add.

On the log scale, we take the difference between each log transformed observation and subtract the log geometric mean.



Transformations for a single sample

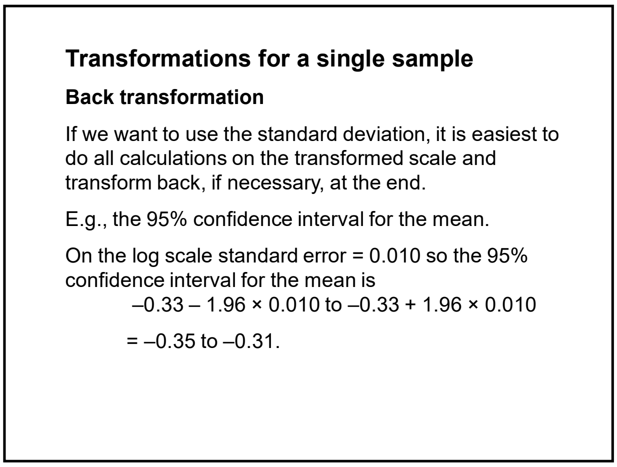
Back transformation

The antilog of the standard deviation is not measured in mmol/litre.

On the log scale, we take the difference between each log transformed observation and subtract the log geometric mean.

We have the difference between the log of two numbers each measured in mmol/litre, giving the log of their ratio which is the log of a dimensionless pure number.

We cannot transform the standard deviation back to the original scale.



Transformations for a single sample

Back transformation

If we want to use the standard deviation, it is easiest to do all calculations on the transformed scale and transform back, if necessary, at the end.

E.g., the 95% confidence interval for the mean.

On the log scale standard error = 0.010 so the 95% confidence interval for the mean is

$$-0.33 - 1.96 \times 0.010 \text{ to } -0.33 + 1.96 \times 0.010$$

$$= -0.35 \text{ to } -0.31.$$

Transformations for a single sample

Back transformation

On the log scale standard error = 0.010 so the 95% confidence interval for the mean is

–0.33 – 1.96 × 0.010 to –0.33 + 1.96 × 0.010

= –0.35 to –0.31.

To get these we took the log of something in mmol/litre and added or subtracted the log of a pure number, so we still have the log of something in mmol/litre.

To get back to the original scale we antilog to give a 95% confidence interval for the geometric mean (0.47) of 0.45 to 0.49 mmol/litre.

Transformations for a single sample

Back transformation

To get back to the original scale we antilog to give a 95% confidence interval for the geometric mean (0.47) of 0.45 to 0.49 mmol/litre.

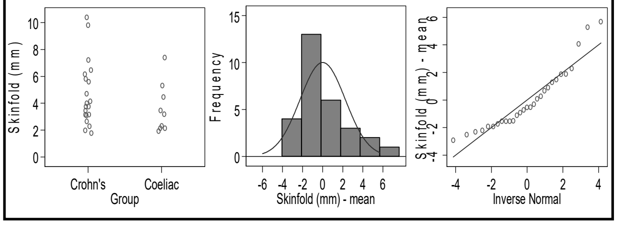
For the arithmetic mean, using the raw, untransformed data we get 0.48 to 0.54 mmol/litre. This interval is wider than for the geometric mean.

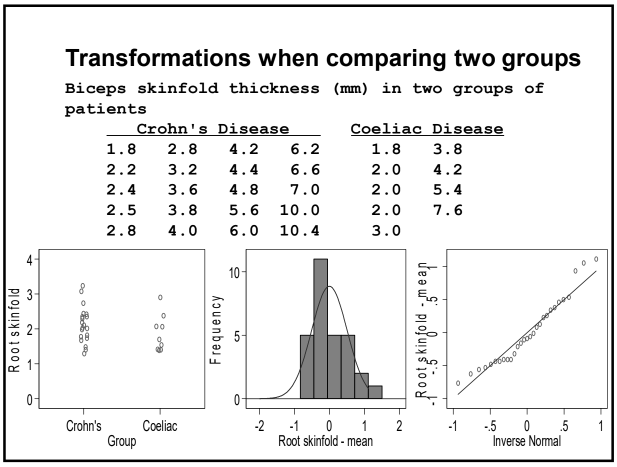
In highly skew data the extreme observations have a large influence on the arithmetic mean, making it more prone to sampling error.

Transformations when comparing two groups

Biceps skinfold thickness (mm) in two groups of patients

Crohn's Disease				Coeliac Disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	

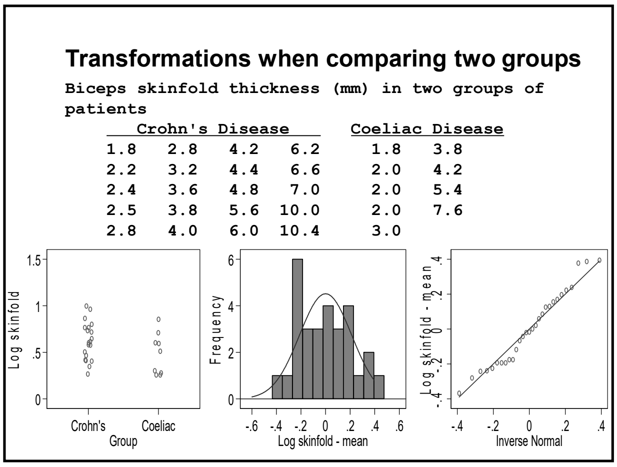




Transformations when comparing two groups

Biceps skinfold thickness (mm) in two groups of patients

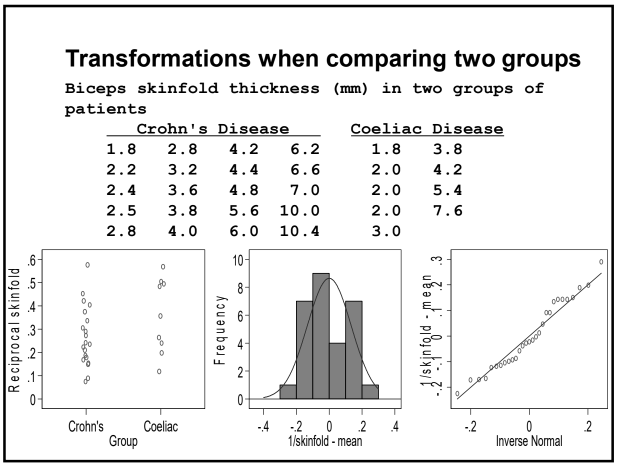
Crohn's Disease				Coeliac Disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	



Transformations when comparing two groups

Biceps skinfold thickness (mm) in two groups of patients

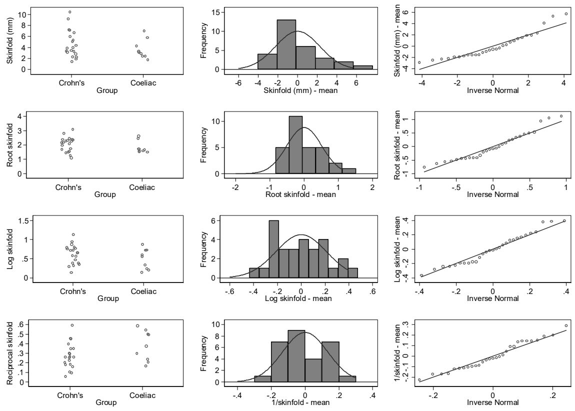
Crohn's Disease				Coeliac Disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	



Transformations when comparing two groups

Biceps skinfold thickness (mm) in two groups of patients

Crohn's Disease				Coeliac Disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	



Transformations when comparing two groups

Biceps skinfold thickness compared for two groups of patients, using different transformations

Transform- ation	Two sample t test, 27 d.f. t	 P	95% confidence interval for difference on transformed scale	Variance ratio larger/ smaller
None	1.28	0.21	-0.71mm to 3.07mm	1.52
square root	1.38	0.18	-0.140 to 0.714	1.16
logarithm	1.48	0.15	-0.114 to 0.706	1.10
reciprocal	-1.65	0.11	-0.203 to 0.022	1.63

Transformations when comparing two groups

Trans	t	P	95% CI for diff	Var ratio
None	1.28	0.21	-0.71mm to 3.07mm	1.52
square root	1.38	0.18	-0.140 to 0.714	1.16
logarithm	1.48	0.15	-0.114 to 0.706	1.10
reciprocal	-1.65	0.11	-0.203 to 0.022	1.63

The transformed data clearly gives a better test of significance than the raw data.

Confidence intervals for the transformed data are more difficult to interpret.

Confidence limits for the difference cannot be transformed back to the original scale.

Transformations when comparing two groups

```
Trans          t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

Confidence limits for the difference cannot be transformed back to the original scale.

The lower (negative limit) for the square root transformation is undefined.

Transformations when comparing two groups

```
Trans          t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

Confidence limits for the difference cannot be transformed back to the original scale.

The upper limit for the reciprocal is very small (0.022) with reciprocal 45.5.

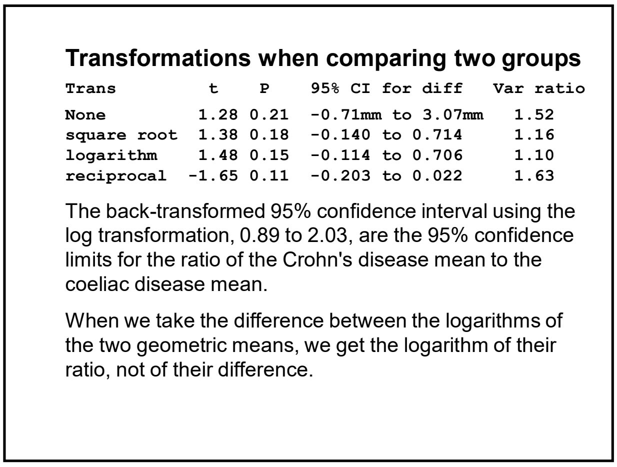
Transformations when comparing two groups

```
Trans          t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

Confidence limits for the difference cannot be transformed back to the original scale.

The log gives interpretable results (0.89 to 2.03) but these are not limits for the difference in millimetres.

They do not contain zero yet the difference is not significant.

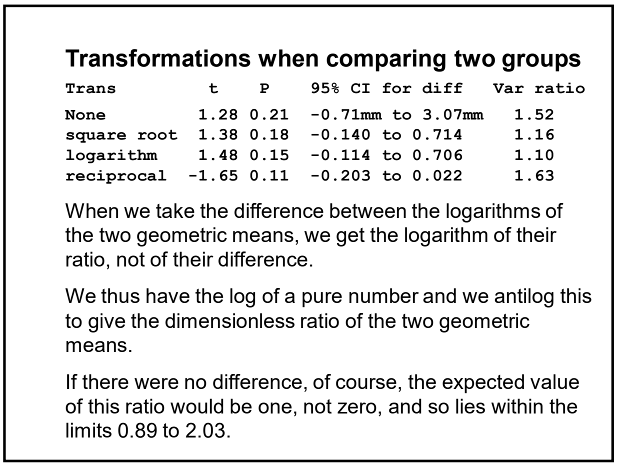


Transformations when comparing two groups

```
Trans         t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

The back-transformed 95% confidence interval using the log transformation, 0.89 to 2.03, are the 95% confidence limits for the ratio of the Crohn's disease mean to the coeliac disease mean.

When we take the difference between the logarithms of the two geometric means, we get the logarithm of their ratio, not of their difference.



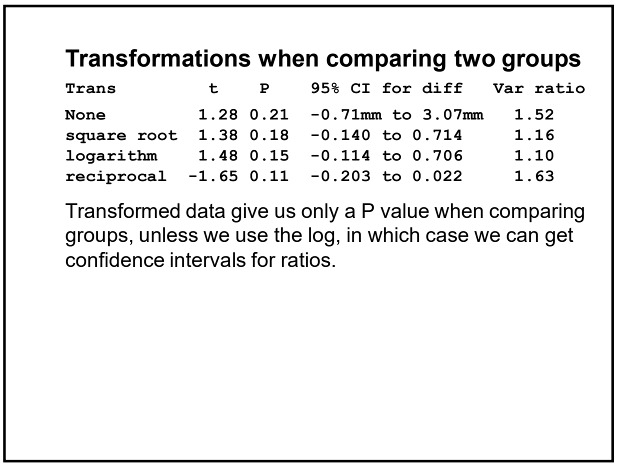
Transformations when comparing two groups

```
Trans         t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

When we take the difference between the logarithms of the two geometric means, we get the logarithm of their ratio, not of their difference.

We thus have the log of a pure number and we antilog this to give the dimensionless ratio of the two geometric means.

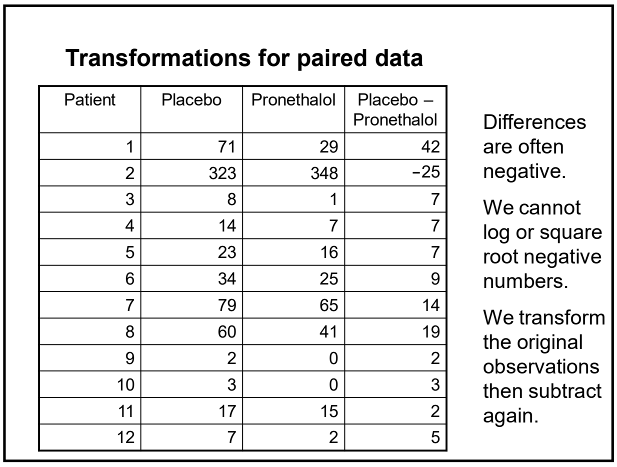
If there were no difference, of course, the expected value of this ratio would be one, not zero, and so lies within the limits 0.89 to 2.03.



Transformations when comparing two groups

```
Trans         t    P    95% CI for diff   Var ratio
None         1.28 0.21  -0.71mm to 3.07mm   1.52
square root  1.38 0.18  -0.140 to 0.714     1.16
logarithm    1.48 0.15  -0.114 to 0.706     1.10
reciprocal  -1.65 0.11  -0.203 to 0.022     1.63
```

Transformed data give us only a P value when comparing groups, unless we use the log, in which case we can get confidence intervals for ratios.



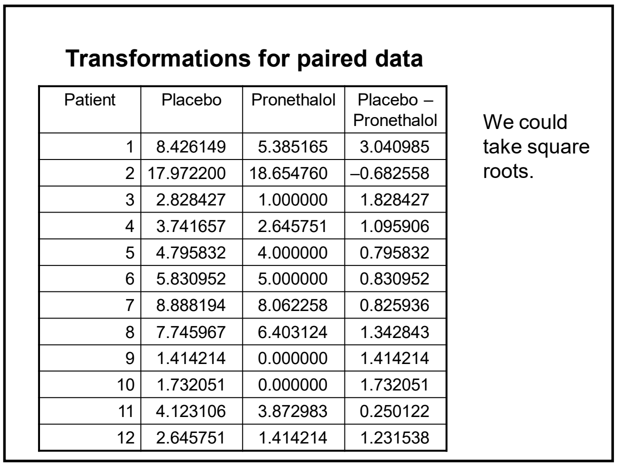
Transformations for paired data

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	71	29	42
2	323	348	−25
3	8	1	7
4	14	7	7
5	23	16	7
6	34	25	9
7	79	65	14
8	60	41	19
9	2	0	2
10	3	0	3
11	17	15	2
12	7	2	5

Differences are often negative.

We cannot log or square root negative numbers.

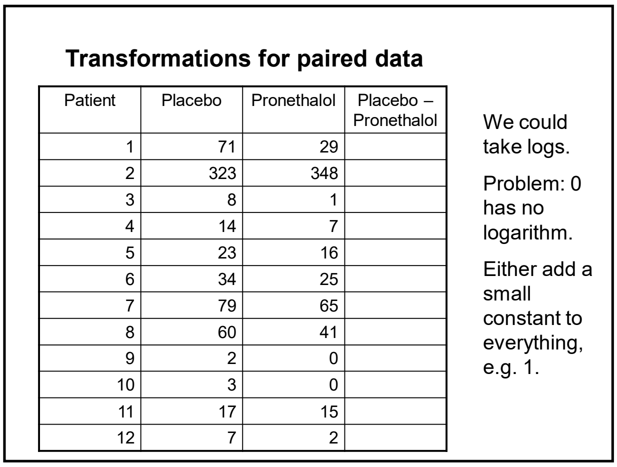
We transform the original observations then subtract again.



Transformations for paired data

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	8.426149	5.385165	3.040985
2	17.972200	18.654760	–0.682558
3	2.828427	1.000000	1.828427
4	3.741657	2.645751	1.095906
5	4.795832	4.000000	0.795832
6	5.830952	5.000000	0.830952
7	8.888194	8.062258	0.825936
8	7.745967	6.403124	1.342843
9	1.414214	0.000000	1.414214
10	1.732051	0.000000	1.732051
11	4.123106	3.872983	0.250122
12	2.645751	1.414214	1.231538

We could take square roots.



Transformations for paired data

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	71	29	
2	323	348	
3	8	1	
4	14	7	
5	23	16	
6	34	25	
7	79	65	
8	60	41	
9	2	0	
10	3	0	
11	17	15	
12	7	2	

We could take logs.

Problem: 0 has no logarithm.

Either add a small constant to everything, e.g. 1.

Transformations for paired data

Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	71+1=72	29 +1=30	
2	323+1=323	348+1=349	
3	8 +1=9	1 +1=2	
4	14 +1=15	7 +1=8	
5	23 +1=24	16 +1=17	
6	34 +1=35	25 +1=26	
7	79 +1=80	65 +1=66	
8	60 +1=61	41 +1=42	
9	2 +1=3	0 +1=1	
10	3 +1=4	0 +1=1	
11	17 +1=18	15 +1=16	
12	7 +1=8	2 +1=3	

We could take logs.

Problem: 0 has no logarithm.

Add a small constant to everything, e.g. 1.

Transformations for paired data

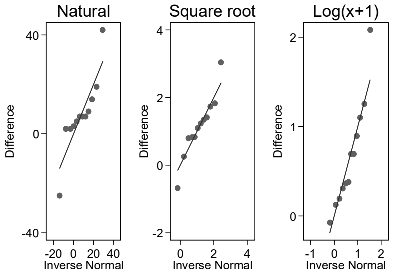
Patient	Placebo	Pronethalol	Placebo – Pronethalol
1	4.276666	3.401197	0.875469
2	5.780744	5.855072	–0.074328
3	2.197225	0.693147	1.504077
4	2.708050	2.079442	0.628609
5	3.178054	2.833213	0.344841
6	3.555348	3.258096	0.297252
7	4.382027	4.189655	0.192372
8	4.110874	3.737670	0.373204
9	1.098612	0.000000	1.098612
10	1.386294	0.000000	1.386294
11	2.890372	2.772589	0.117783
12	2.079442	1.098612	0.980830

We could take logs.

Log(*x*+1).

Transformations for paired data

Both square root and log seem to improve the fit to the Normal.



Paired t tests

Natural scale: P=0.11

Square root scale: P=0.0011

Log(x+1) scale: P=0.0012

Which transformation?

To make Normal

- counts: try square root,
- concentrations in blood: try log, then reciprocal,
- ratios: try log, then reciprocal.

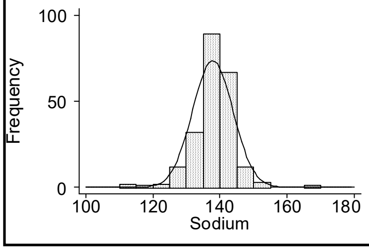
To make variability uniform

- variance proportional to mean: square root,
- standard deviation proportional to mean: log,
- standard deviation proportional to mean squared: reciprocal.

Can all data be transformed?

Sometimes we have very long tails at both ends of the distribution, which makes transformation by log, square root or reciprocal ineffective.

Blood sodium in ITU patients:



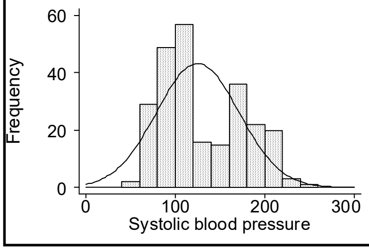
We can often ignore this departure from the Normal distribution.

It is possible to transform, but difficult to interpret afterwards.

Can all data be transformed?

Sometimes we have a bimodal distribution, which makes transformation by log, square root or reciprocal ineffective.

Systolic blood pressure in ITU patients:



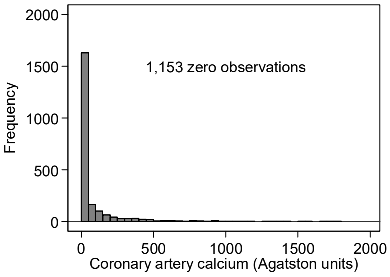
We should not ignore this departure from the Normal distribution.

It is possible to transform, but difficult to interpret afterwards.

Can all data be transformed?

Sometimes we have a large number of identical observations, usually at zero.

Coronary artery calcium:



Any transformation will leave half the observations with the same value, at the extreme of the distribution.

It is impossible to transform these data to a Normal distribution.

What can we do if we cannot transform data to a suitable form?

We can use non-parametric methods, such as the Mann-Whitney U test.

These will give us a significance test, but usually no confidence interval.

Are there data which should not be transformed?

Sometimes we are interested in the data in the actual units only.

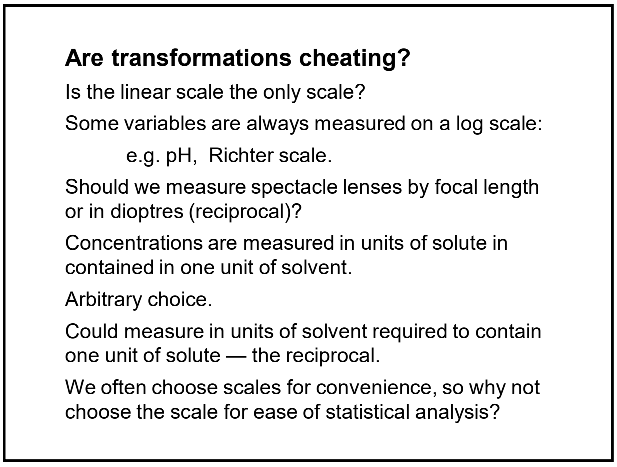
Cost data is a good example.

Costs of treatment usually have distributions which are highly skew to the right.

However, in a trial we need to estimate the difference in mean costs in pounds. No other scale is of interest.

We should not transform such data.

We rely on large sample comparisons or on methods which do not involve any distributions (e.g. bootstrap methods,)



Are transformations cheating?

Is the linear scale the only scale?

Some variables are always measured on a log scale:

e.g. pH, Richter scale.

Should we measure spectacle lenses by focal length or in dioptres (reciprocal)?

Concentrations are measured in units of solute in contained in one unit of solvent.

Arbitrary choice.

Could measure in units of solvent required to contain one unit of solute — the reciprocal.

We often choose scales for convenience, so why not choose the scale for ease of statistical analysis?