Logarithmic transformation

• Special case of a nonlinear transformation.
  – A logarithm (to base \( b \)) of a number \( x \) is the exponent \( y \) that satisfies \( x = b^y \).
  – Decreases skewness to the right, increases skewness to the left.
    • Many biological data distributions are slightly right-skewed.
  – But is even more useful because:
    • Transforms power relationships or geometric relationships into linear relationships.
    • Converts multiplicative factors into additive factors.
• Many concentration and intensity scales expressed as logarithms; e.g.,
  – $\text{pH} = -\log \left[H^+\right]$.
  – Decibels = $10 \log \frac{I_1}{I_0}$, where $I_0$ is a threshold intensity.

– Richter scale:
Transforming to increase skewness to the left

\[ X' = X \]
\[ X' = X^{0.5} \]
\[ X' = \ln X \]

\[ X' = -X^{-0.5} \]
\[ X' = -X^{-1} \]
\[ X' = -X^{-2} \]
• Logarithms of some selected values:

<table>
<thead>
<tr>
<th>Common logarithm (log = \log_{10})</th>
<th>Natural logarithm (\ln = \log_e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\log_{10} (1) = 0</td>
<td>\log_e (1) = 0</td>
</tr>
<tr>
<td>\log_{10} (10) = 1</td>
<td>\log_e (10) = 2.3</td>
</tr>
<tr>
<td>\log_{10} (100) = 2</td>
<td>\log_e (100) = 4.6</td>
</tr>
<tr>
<td>\log_{10} (500) = 2.7</td>
<td>\log_e (500) = 6.2</td>
</tr>
<tr>
<td>\log_{10} (1000) = 3</td>
<td>\log_e (1000) = 6.9</td>
</tr>
</tbody>
</table>
- Logarithmic functions:

Any log-transformation (to any base) is a linear transformation of any other log-transformation (to any other base).

If you know \( \log_b X \)

then \( \log_a X = \frac{\log_b X}{\log_a b} \)
Logarithmic transformation

- Useful properties, based on the arithmetic of exponents:

\[
\begin{align*}
\log(x \cdot y) &= \log x + \log y \\
\log(x / y) &= \log x - \log y \\
\log(1 / x) &= -\log x \\
\log(x^c) &= c \log x \\
\log(\sqrt[n]{x}) &= \log(x^{1/n}) = \frac{1}{c} \log x \\
\log 1 &= 0 \\
\log 0 &= -\infty
\end{align*}
\]

- \(10^{\log x} = x\)
- \(e^{\ln x} = x\)

Example: factorials –

\[
N! = \prod_{i=1}^{N} i \\
\log(N!) = \sum_{i=1}^{N} \log(i) \\
10^{\log(N!)} = N!
\]
• The log transformation is the *most generally useful transformation* in biological applications, for many reasons.
  – Three primary reasons, associated with variation:

(1) **Univariate** case (one variable):
  • For many biological variables, the level of variation (*standard deviation*) is proportional to the *mean*.
    – The “mouse–elephant” effect: elephants are much more variable than mice because they’re much larger.
  • For this reason, the *coefficient of variation* is useful in biology:
    \[ CV = \frac{s}{\bar{X}} \]
    – Measure of *relative variation*, relative to the mean.
    – Tends to be constant for a particular variable across a large size range.
• Wright showed that: \[ s_{\log X}^2 \approx \log\left(1 + CV_X^2\right) \]
  - Variance of the log-transformed data is approximately equal to the log-transformed coefficient of variation (squared) of the original variables.

  - For \( CV < 30\% \) or so (almost always the case for biological variables), a closer approximation is:

    \[
    s_{\log X}^2 \approx CV^2 \\
    s_{\log X} \approx CV
    \]

    » Standard deviation of the log-transformed data is approximately equal to the coefficient of variation.

  - Thus: the standard deviation and variance of log-transformed data are measures of *relative variation* rather than absolute variation.
(2) **Bivariate** case (two variables):

- Many biological processes are inherently exponential or multiplicative in their variation and change.

(a) A *single variable* at any particular time tends to be proportional to its value at a previous time: \( X_{t+1} = k \cdot X_t \)

- *Ex:* populations (of organisms or cells) have multiplicative growth, approximately doubling their size over some time period rather than increasing by constant numbers (=additive growth).

(b) Values of *different variables* tend to be proportional to one another: \( Y = k \cdot X \)

<table>
<thead>
<tr>
<th>Raw data</th>
<th>Log-transformed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_Y^2 = s_{kX}^2 = k^2 \cdot s_X^2 )</td>
<td>( s_{\log Y}^2 = s_{\log(kX)}^2 )</td>
</tr>
<tr>
<td>( s_Y = k \cdot s_X )</td>
<td>( = s_{(\log k + \log X)}^2 )</td>
</tr>
<tr>
<td></td>
<td>( s_{\log Y} = s_{\log X}^2 )</td>
</tr>
<tr>
<td></td>
<td>( s_{\log Y} = s_{\log X} )</td>
</tr>
</tbody>
</table>
• In other words: if two variables have proportional values (with a constant of proportionality $k$), then:
  - The variances of their original values differ by a factor of $k^2$.
  - The standard deviations of their original values differ by a factor of $k$.
  - But the variances of their log-transformed values are equal.
  - And the standard deviations of their log-transformed values are equal.
• Thus, using log-transformed data eliminates the differences in variability due to differences in magnitude of the variables:
  - Produces homogeneity of variances (homoscedasticity), to the extent that the individual $CV$’s are equal.
  - Again, variation is relative rather than absolute.
(3) Third reason relevant to bivariate and multivariate statistics:

- Many biological variables are related by *power (scaling) relationships (allometry)*: \( Y = bX^k \)
  - Typical form is *linearized* by log-transformation:
    \[ \ln Y = \ln b + k \ln X \]
    - Important because most statistical procedures assume linear relationships among variables.