Log Scales

- Often data are better analyzed in a log scale.
- If we transform the data to a log scale, we analyze X where maybe
  - Log scale
    - $X = \log_{10}(Y)$
    - $X = \log_2(Y)$
    - $X = \log_e(Y) = \ln(Y)$
  - Reciprocal scale
    - $X = 1/Y$
    - $Y =$ time, $X$=rate
- If we calculate means and confidence intervals in the log scale, it helps to translate, back-transform, these back to the original scale for presenting results.

**Log Scale: Geometric Means**
- If $\bar{X} =$ mean of $\log_{10}(Y)$
- Back-transformed to original scale: $10^{\bar{X}} =$ geometric mean.
- Example
  - $Y = 1, 10, \text{and } 100.$
  - $\bar{Y} = 37.$ Arithmetic mean
  - $X = 0, 1, 2$
  - $\bar{X} = 1$
  - Geometric mean = : $10^1 = 10$
  - Always: Geometric mean $\leq$ Arithmetic mean

**Reciprocal Scale: Harmonic Means**
- If $\bar{X} =$ mean of $1/Y$
- Back-transformed to original scale: $\frac{1}{\bar{X}} =$ harmonic mean.
- Example
  - $Y = 1, 10, \text{and } 100.$
  - $\bar{Y} = 37.$ Arithmetic mean
  - $X = 0, 0.1, 0.01$
  - $\bar{X} = 0.37$
  - Harmonic mean = : $\frac{1}{0.37} = 2.70$
  - Always: Harmonic mean $\leq$ Geometric mean $\leq$ Arithmetic mean
For log scales

- If $\ln(Y) = X \sim \text{Normal}(\mu, \sigma^2)$ then $Y$ has a lognormal distribution.
  - The median of $X$ is $\mu$.
  - $Y = e^X$, so the median of $Y$ is $e^\mu = \text{geometric mean}$.
  - For $Y$: $\text{mean} = e^{\mu+0.5\sigma^2} \leq e^\mu = \text{geometric mean}$
  - (To show $\text{mean}$ of $Y = e^{\mu+0.5\sigma^2}$, use the moment generating function for a normal distribution.)
  - Sometimes people want to estimate the mean of $Y$ after analysis of Log($Y$).
    - For example fish biologists want total biomass = $n\times\text{mean}$.
    - Mean of $Y \approx e^{X+0.5\sigma^2}$

**Differences in the log scale correspond to (back-transform as) ratios.**
- $\log_{10}a - \log_{10}b = 2 \Rightarrow 10^{\log_{10}a - \log_{10}b} = 10^{\log_{10}(a/b)} = \frac{a}{b}$
- For example
  - $\text{before} = 1000$, $\log_{10}\text{before} = 3$
  - $\text{after} = 10$, $\log_{10}\text{after} = 1$
  - $\log_{10}\text{after} - \log_{10}\text{after} = -2$
  - $\frac{\text{after}}{\text{before}} = \text{ratio} = 10^{-2} = 0.01$
- Similarly
  - $\text{before} = 2$, $\log_{2}\text{before} = 1$
  - $\text{after} = 16$, $\log_{10}\text{after} = 4$
  - $\log_{10}\text{after} - \log_{10}\text{after} = 2$
  - $\frac{\text{after}}{\text{before}} = \text{ratio} = 2^3 = 8$

**Ratios of geometric means and confidence intervals for ratios**
- Suppose $Y$ = stopping distance.
- $\bar{X}_1$ and $\bar{X}_2$ are means of Ln(distance) at speeds 45 and 60 mph
- Back-transformed difference
  - $e^{\bar{X}_1 - \bar{X}_2} = \frac{e^{\bar{X}_1}}{e^{\bar{X}_2}} = \frac{\text{45 mph geo mean distance}}{\text{60 mph geo mean distance}} = \text{ratio of geometric means}$
- 95% confidence interval for $\mu_1 - \mu_2$ for Ln(distance) is
  - $\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2} \times SE_{\bar{X}_1 - \bar{X}_2} = (\text{lower, upper})$
- $\text{lower} \leq \ln(\text{distance}_1) - \ln(\text{distance}_2) \leq \text{upper}$
- $\text{lower} \leq \ln(\text{distance}_1/\text{distance}_2) \leq \text{upper}$
- $e^{\text{lower}} \leq \frac{\text{distance}_1}{\text{distance}_2} \leq e^{\text{upper}}$
- For example if 95% confidence interval for means of Ln(distance)

<table>
<thead>
<tr>
<th>Speed</th>
<th>Method</th>
<th>Mean</th>
<th>95% CL</th>
<th>Mean</th>
<th>95% CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff  (1-2)</td>
<td>Pooled</td>
<td>-0.3231</td>
<td>-0.3920</td>
<td>-0.2541</td>
<td></td>
</tr>
</tbody>
</table>

- Ratio of geometric means
  - $e^{-0.3231} = 0.724$
  - On average the stopping distance at 45 mph is 72.4% of the stopping distance at 60 mph.
- 95% confidence interval for ratio of geometric means
  - $(e^{-0.3920}, e^{-0.2541}) = (0.676, 0.776)$
  - Between 67.6% and 77.6% shorter stopping distance at 45 mph
If the log scale values are lognormal with same variances, then

- ratio of means \( = \frac{e^{\mu_1 + 0.5 \cdot \sigma^2}}{e^{\mu_2 + 0.5 \cdot \sigma^2}} = \frac{e^{\mu_1} e^{0.5 \cdot \sigma^2}}{e^{\mu_2} e^{0.5 \cdot \sigma^2}} = \frac{e^{\mu_1}}{e^{\mu_2}} = \text{ratio of geometric means} \)
- A confidence interval for ratio of geometric means is also a confidence interval for ratio of means.

Back-transformed differences work similarly for paired differences.
- From the herpes example.

\[ \text{Log}_{10} \text{Before} \quad \text{Log}_{10} \text{After} \quad \text{Difference} \]

<table>
<thead>
<tr>
<th>Mean</th>
<th>Log(_{10}) Before</th>
<th>Log(_{10}) After</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.307</td>
<td>2.182</td>
<td>-0.125</td>
<td></td>
</tr>
</tbody>
</table>

A Herpes Oncolytic Virus Can Be Delivered Via the Vasculature to Produce Biologic Changes in Human Colorectal Cancer

**Changes in carcinoembryonic antigen (CEA) levels.** (a) Maximum drop in CEA before chemotherapy in the first 28 days after NV1020 administration for each individual subject. Drops varied between 13 and 74% (median 24%). (b) Box plots of these data. \( P = 0.02 \) by paired t-test.

\[ \text{mean} \left[ \log_{10} \left( \frac{\text{CEA After}}{\text{CEA Before}} \right) \right] = -0.125 \]

\[ \text{Geometric mean of } \frac{\text{CEA After}}{\text{CEA Before}} = 10^{-0.125} = 0.75. \]

\[ \text{or} \]

\[ \text{mean} [\log_{10} (\text{CEA After}) - \log_{10} (\text{CEA Before})] = \text{mean} [\log_{10} (\text{CEA After})] - \text{mean} [\log_{10} (\text{CEA Before})] = 2.182 - 2.307 = -0.125 \]

\[ \text{ratio of geometric means} = \frac{10^{2.182 - 2.307}}{10^{2.182}} = \frac{10^{2.182}}{10^{2.182}} = 152.13 \]

If 95% confidence interval for mean \( \log_{10}(\text{CEA After}) - \log_{10}(\text{CEA Before}) \) is
-0.149 to -0.110
  - The 95% confidence interval for \( \frac{\text{CEA After}}{\text{CEA Before}} \) or ratio of geometric means is \( 10^{-0.149} \) to \( 10^{-0.110} = 0.71 \) to 0.79.
  - The after values are 71% to 79% less than the before CEA values.

- Here is a published example of using ratios of geometric means.
  - Often bioinformatics data are analyzed in a log scale, often \( \log_2 \).

**Effects of Allergen Challenge on Airway Epithelial Cell Gene Expression**

Craig M. Lilly, Hiroki Tateo, Tsuyoshi Oguma, Elliot Israel and Larry A. Sonna


"The resulting expression ratios from each of the five paired sets of samples were then used to determine whether there had been a statistically significant change in expression, by computing means and 95% CIs (derived using the t-distribution) on natural log-transformed data, as described previously (11)."

<table>
<thead>
<tr>
<th>Sequence</th>
<th>GenBank Identifier</th>
<th>Expression Ratio Allergen/Control (( n = 5 ))</th>
<th>Geometric Mean (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala consensus sequence</td>
<td>U14573</td>
<td>1.0 (0.91–1.2)</td>
<td></td>
</tr>
<tr>
<td>( \beta )-Actin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' region</td>
<td>X00351</td>
<td>1.7 (0.78–3.6)</td>
<td></td>
</tr>
<tr>
<td>Middle region</td>
<td>X00351</td>
<td>1.6 (0.84–3.1)</td>
<td></td>
</tr>
<tr>
<td>3' region</td>
<td>X00351</td>
<td>1.6 (0.91–2.7)</td>
<td></td>
</tr>
</tbody>
</table>
• Standard deviations of Ln(Y) and coefficients of variation

  o If \( CV(Y) = \frac{\text{St Dev}}{\text{Mean}} \) is small, \( \text{St Dev}(LnY) \approx CV(Y) \)
  o See below for the justification for this.
  o For example with the 3 minute sand timer

<table>
<thead>
<tr>
<th>Time</th>
<th>Ln Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>271</td>
<td>5.602</td>
</tr>
<tr>
<td>317</td>
<td>5.759</td>
</tr>
<tr>
<td>314</td>
<td>5.749</td>
</tr>
<tr>
<td>337</td>
<td>5.820</td>
</tr>
<tr>
<td>315</td>
<td>5.753</td>
</tr>
<tr>
<td>309</td>
<td>5.733</td>
</tr>
<tr>
<td>297</td>
<td>5.694</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time Mean</th>
<th>Ln Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>308.57</td>
<td>5.730</td>
</tr>
</tbody>
</table>

| St Deviation | 20.39 | 0.068 |
| CV           | 0.066 |

If different treatments have similar CV's where the standard deviations are similar percents of the means, then log scale Y values have similar variances or standard deviations.

• To show this relationship Use "propagation of error" or "delta method"

  o \( g(Y) \approx g(\mu_Y) + g'(\mu_Y) * (Y - \mu_Y) \)
  o \( E[g(Y) - g(\mu_Y)]^2 \approx g'(\mu_Y)^2 * E(Y - \mu_Y)^2 \)
  o \( \text{Var}[g(\mu_Y)] \approx g'(\mu_Y)^2 * \text{Var}(Y) \)
  o \( \text{Var}[\ln(Y)] \approx \frac{\text{Var}(Y)}{(\mu_Y)^2} = CV^2 \)
  o \( \text{St Dev}[\ln(Y)] \approx CV \)

Since any log scale is a constant times Ln(Y), if different treatments have similar CV’s where the standard deviations are similar percents of the means, then log scale Y values in any log scale have similar variances or standard deviations.