Error analysis in biology

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

Errors, like straws, upon the surface flow;
He who would search for pearls must dive below

John Dryden (1631-1700)
Error bars are 95% confidence intervals.
Previously on errors...

**How to make a good plot**
- Clarity of presentation!
- Good lines and symbols
- Clear labels
- Logarithmic plots

**Box plots**
- Good alternative for SD
- Show distribution of data

**Bar plots**
- Only for additive quantities
- Baseline must be zero
- Never in log space
- Careful with continuous variable
- Always show upper and lower error bars!

**Significant figures**
- Carry meaningful information
- Quote only significant figures
- Use error in the error

Number: 1.23457456
Error: 0.02377345
Example

- Measure positions of two fluorescent dots under a microscope (in μm)

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dot 1</td>
<td>3.68</td>
<td>3.12</td>
<td>5.44</td>
</tr>
<tr>
<td>Dot 2</td>
<td>3.90</td>
<td>3.86</td>
<td>4.02</td>
</tr>
</tbody>
</table>

- Measurement errors for x – y and z direction:
  - $\Delta_{xy} = 120$ nm
  - $\Delta_z = 200$ nm

- Find distance between the dots

$$R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} = 1.62 \, \mu m$$

- What is the error of $R$?
- We need to **propagate** errors of $x$, $y$ and $z$
7. Error propagation

“If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is”

John von Neumann
Derivative

Consider a function $y = f(x)$

Derivative of $f$

$$\frac{df}{dx} \approx \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x}$$

for small $\Delta x$

Derivative = slope

A few derivatives

$$\frac{d}{dx} ax = a$$

$$\frac{d}{dx} x^r = rx^{r-1}$$

$$\frac{d}{dx} x^2 = 2x$$

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a}$$
Error propagation (single variable)

- Consider a quantity \( x \pm \Delta x \)

- Transform to a new variable \( y = f(x) \)

- For example
  - \( y = ax \)
  - \( y = \log(x) \)
  - \( y = \sqrt{x} \)

- Find error of \( y \), \( \Delta y \)
Error propagation (single variable)

- Consider a quantity \( x \pm \Delta x \)
- Transform to a new variable \( y = f(x) \)
- Find error of \( y \), \( \Delta y \)
- If errors are small
  \[
  \frac{\Delta y}{\Delta x} \approx \frac{df}{dx}
  \]
- Hence
  \[
  \Delta y \approx \left| \frac{df}{dx} \right| \Delta x \quad \text{or} \quad \Delta y^2 \approx \left( \frac{df}{dx} \right)^2 \Delta x^2
  \]
Error propagation: one variable

\[ \Delta y = \left| \frac{df}{dx} \right| \Delta x \]

**Scaling**

\[ y = f(x) = ax \]

\[ \Delta y = \left| \frac{df}{dx} \right| \Delta x = |a| \Delta x \]

\[ \Delta y = |a| \Delta x \]

\[ \frac{d}{dx} ax = a \]

\[ y = 10x \]

\[ x = 5 \pm 3 \]

\[ y = 50 \pm 30 \]

**Logarithm**

\[ y = f(x) = \log_2 x \]

\[ \Delta y = \left| \frac{df}{dx} \right| \Delta x = \left| \frac{1}{x \ln 2} \right| \Delta x \]

\[ \Delta y \approx 1.44 \frac{\Delta x}{x} \]

\[ \frac{d}{dx} \log_a x = \frac{1}{x \ln a} \]

\[ y = \log_2 x \]

\[ x = 4 \pm 0.5 \]

\[ y = 2 \pm 0.2 \]
Error propagation: logarithm

\[ y = \log_{10} x \Rightarrow \Delta y = \frac{1}{\ln 10} \frac{\Delta x}{x} \approx 0.43 \frac{\Delta x}{x} \]

\[ \Delta x \propto x \]

\[ \Delta \log x = \text{const} \]
Error propagation: many variables

- Consider \( n \) independent (not correlated) variables \( x_1, x_2, ..., x_n \)

- Each of them with error \( \Delta x_1, \Delta x_2, \ldots, \Delta x_n \)

- New variable \( y = f(x_1, x_2, ..., x_n) \)

- It can be shown that

\[
\Delta y^2 \approx \left( \frac{\partial f}{\partial x_1} \right)^2 \Delta x_1^2 + \left( \frac{\partial f}{\partial x_2} \right)^2 \Delta x_2^2 + \cdots + \left( \frac{\partial f}{\partial x_n} \right)^2 \Delta x_n^2
\]
Sum or difference

\[ y = f(x_1, x_2) = x_1 + x_2 \]

\[ \Delta y^2 = \left( \frac{\partial f}{\partial x_1} \right)^2 \Delta x_1^2 + \left( \frac{\partial f}{\partial x_2} \right)^2 \Delta x_2^2 \]

\[ \frac{\partial f}{\partial x_1} = 1 \]

\[ \frac{\partial f}{\partial x_2} = 1 \]

\[ \Delta y^2 = \Delta x_1^2 + \Delta x_2^2 \]

errors add in quadrature

\[ x_1 = 8 \pm 3 \]
\[ x_2 = 10 \pm 4 \]
\[ x + y = 18 \pm 5 \]
Ratio or product

\[ y = f(x_1, x_2) = \frac{x_1}{x_2} \]

\[ \Delta y^2 = \left( \frac{\partial f}{\partial x_1} \right)^2 \Delta x_1^2 + \left( \frac{\partial f}{\partial x_2} \right)^2 \Delta x_2^2 \]

\[ \frac{\partial f}{\partial x_1} = \frac{1}{x_2}, \quad \frac{\partial f}{\partial x_2} = -\frac{x_1}{x_2^2} \]

\[ \Delta y^2 = \left( \frac{1}{x_2} \right)^2 \Delta x_1^2 + \left( -\frac{x_1}{x_2^2} \right)^2 \Delta x_2^2 = \left( \frac{x_1}{x_2} \right)^2 \left( \frac{\Delta x_1}{x_1} \right)^2 + \left( \frac{x_1}{x_2} \right)^2 \left( \frac{\Delta x_2}{x_2} \right)^2 \]

\[ = \left( \frac{x_1}{x_2} \right)^2 \left[ \left( \frac{\Delta x_1}{x_1} \right)^2 + \left( \frac{\Delta x_2}{x_2} \right)^2 \right] = y^2 \left[ \left( \frac{\Delta x_1}{x_1} \right)^2 + \left( \frac{\Delta x_2}{x_2} \right)^2 \right] \]

\[ \left( \frac{\Delta y}{y} \right)^2 = \left( \frac{\Delta x_1}{x_1} \right)^2 + \left( \frac{\Delta x_2}{x_2} \right)^2 \]

Geometrical interpretation

Multiply and divide by \( x_1^2 \neq 0 \)

\[ \frac{\Delta x_1}{x_1} = 25 \pm 2.5 \]
\[ x_2 = 10 \pm 1 \]
\[ \frac{x_1}{x_2} = 2.5 \pm 0.4 \]

10% error in \( x_1 \) and \( x_2 \) gives 14% error in \( x_1/x_2 \)

fractional errors add in quadrature
Fluorescent dots

- Two dots: \((x_1, y_1, z_1)\) and \((x_2, y_2, z_2)\)

\[
R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}
\]

- Propagated error of \(R(x_1, y_1, z_1, x_2, y_2, z_2)\):

\[
\Delta R^2 = \left(\frac{\partial R}{\partial x_1}\right)^2 \Delta^2_{xy} + \left(\frac{\partial R}{\partial y_1}\right)^2 \Delta^2_{xy} + \left(\frac{\partial R}{\partial z_1}\right)^2 \Delta^2_z 
+ \left(\frac{\partial R}{\partial x_2}\right)^2 \Delta^2_{xy} + \left(\frac{\partial R}{\partial y_2}\right)^2 \Delta^2_{xy} + \left(\frac{\partial R}{\partial z_2}\right)^2 \Delta^2_z
\]

- Homework: do the calculations and confirm that

\[
\Delta R = \frac{\sqrt{2}}{R} \sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2] \Delta^2_{xy} + (z_1 - z_2)^2 \Delta^2_z}
\]
When error propagation is not necessary

- Experiment to measure $IC_{50}$ of a drug
- 5 replicates
- Logarithmic version: $pIC_{50} = -\log \frac{IC_{50}}{1 \text{ M}}$

- Two ways of finding mean $pIC_{50}$ and its error
  - propagate from $IC_{50}$
  - direct calculation

- Results are not identical
- Log of the mean is not mean of the logs!

- Errors are large (~30%), so error propagation formula does not work well
- Do not use error propagation if you can calculate errors directly from replicated data

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$IC_{50}$ (nM)</td>
<td>25</td>
<td>85</td>
<td>43</td>
<td>108</td>
<td>12</td>
</tr>
<tr>
<td>$pIC_{50}$</td>
<td>7.6</td>
<td>7.2</td>
<td>7.4</td>
<td>6.7</td>
<td>7.9</td>
</tr>
</tbody>
</table>

$M = 54.6 \text{ nM}$
$SE = 18.2 \text{ nM}$

$IC_{50} = 50 \pm 20 \text{ nM}$

$pIC_{50} = -\log \frac{54.6 \text{ nM}}{1 \text{ M}} = 7.26$

error propagation $\Delta pIC_{50} = 0.43 \frac{SE}{M} = 0.14$

$pIC_{50} = 7.3 \pm 0.1$

$M_p = 7.39$
$SE_p = 0.17$

calculating directly from logarithmic data:

$pIC_{50} = 7.4 \pm 0.2$
Error propagation summary

- When a quantity is transformed, its error must be propagated
- Single variable

\[ y = f(x) \]

\[ \Delta y \approx \left| \frac{df}{dx} \right| \Delta x \]

- Multiple variables

\[ y = f(x_1, x_2, ..., x_n) \]

\[ \Delta y^2 \approx \left( \frac{\partial f}{\partial x_1} \right)^2 \Delta x_1^2 + \left( \frac{\partial f}{\partial x_2} \right)^2 \Delta x_2^2 + \cdots + \left( \frac{\partial f}{\partial x_n} \right)^2 \Delta x_n^2 \]
<table>
<thead>
<tr>
<th>Function</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = ax )</td>
<td>( \Delta y = a \Delta x )</td>
</tr>
<tr>
<td>( y = ax^b )</td>
<td>( \frac{\Delta y}{y} = \frac{b \Delta x}{x} )</td>
</tr>
<tr>
<td>( y = a \log_b cx )</td>
<td>( \Delta y = \frac{a \Delta x}{\ln b \cdot x} )</td>
</tr>
<tr>
<td>( y = ae^{bx} )</td>
<td>( \frac{\Delta y}{y} = b \Delta x )</td>
</tr>
<tr>
<td>( y = 10^{ax} )</td>
<td>( \frac{\Delta y}{y} = a \ln(10) \Delta x )</td>
</tr>
<tr>
<td>( y = ax_1 \pm bx_2 )</td>
<td>( \Delta y = \sqrt{a^2 \Delta x_1^2 + b^2 \Delta x_2^2} )</td>
</tr>
<tr>
<td>( y = x_1 x_2, \quad y = \frac{x_1}{x_2} )</td>
<td>( \frac{\Delta y}{y} = \sqrt{\left(\frac{\Delta x_1}{x_1}\right)^2 + \left(\frac{\Delta x_2}{x_2}\right)^2} )</td>
</tr>
</tbody>
</table>
8. Simple linear regression

“It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest”

S. den Hartog
Linear regression

- Sample linear regression:
  \[ y(x) = ax + b \]
  - \( a \) – slope
  - \( b \) – intercept
  - \( x \) – explanatory variable
  - \( y \) – response variable

- Population true regression
  \[ \bar{y} = \alpha x + \beta \]
  - \( \bar{y} \) – mean response to \( x \)
  - \( \alpha \) – true slope
  - \( \beta \) – true intercept

- We want to find estimators of \( \alpha \) and \( \beta \) and their uncertainties from a sample

Data from Hong Kong Growth Survey (25,000 adolescent youths). Body mass (\( m \)) is plotted against squared height (\( h^2 \))
Linear regression

- Data are usually noisy

\[ y = \alpha x + \beta + r \]

- In biology there is often an underlying simple relationship between the observables, even if it is badly affected by noise

Data from Hong Kong Growth Survey (25,000 adolescent youths). Body mass \( m \) is plotted against squared height \( h^2 \).
Simple linear fit

- Consider a sample \((x_i, y_i), i = 1, \ldots, n\)

- Actual response:
  \[ y_i = ax_i + b + R_i \]

- \(R_i = y_i - y(x_i)\) are residuals

- For best-fitting model minimise
  \[ Q = \sum_{i=1}^{n} R_i^2 = \text{min} \]

- Minimum: \(\frac{\partial Q}{\partial a} = 0\) and \(\frac{\partial Q}{\partial b} = 0\)

Random selection of 16 points from the Hong Kong Growth Survey
Simple linear fit: the solution

- Minimise the sum of squared residuals and find:
  \[ a = \frac{S_{xy}}{S_{xx}} \]
  \[ b = M_y - aM_x \]

- These are the estimators of true unknown slope, \( \alpha \), and intercept, \( \beta \)

\[
M_x = \frac{1}{n} \sum_{i=1}^{n} x_i \quad M_y = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

\[
S_{xx} = \sum_{i=1}^{n} (x_i - M_x)^2 \quad S_{yy} = \sum_{i=1}^{n} (y_i - M_y)^2
\]

\[
S_{xy} = \sum_{i=1}^{n} (x_i - M_x)(y_i - M_y)
\]

The best-fitting line always passes through data centroid, \((M_x, M_y)\)
Analogy to standard deviation

- Sample $y_i$ is scattered around the mean, $M$
- Standard deviation is calculated from squared residuals
  \[ R_i = y_i - M \]
  \[ SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} R_i^2} \]

- Sample $(x_i, y_i)$ is scattered around the regression line, $y(x)$
- Standard deviation is calculated from squared residuals
  \[ R_i = y_i - y(x_i) \]
  \[ SD_R = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} R_i^2} = \sqrt{\frac{S_{yy} - aS_{xy}}{n-2}} \]
Confidence intervals on fit parameters

- We can quantify scatter around the regression line by $SD_R$
- Assume the scatter is due to uncertainty (noise/variability) in $y$
- Let’s use $SD_R$ as a common uncertainty of $y_i$
  \[ \Delta y_i = SD_R \]
- This is our guessed error of $y_i$
- It estimates a typical error in response
Confidence intervals on fit parameters

- Fit parameters depend on data points
  \[ a = a(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) \]
  \[ b = b(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) \]
  \[ \Delta x_i = 0 \quad \Delta y_i = SD_R \]

- Hence, we can propagate estimated errors \( \Delta y_i = SD_R \):
  \[ \Delta a^2 = \sum_{i=1}^{n} \left( \frac{\partial a}{\partial y_i} \right)^2 \Delta y_i^2 \]
  \[ \Delta b^2 = \sum_{i=1}^{n} \left( \frac{\partial b}{\partial y_i} \right)^2 \Delta y_i^2 \]

- After some rather straightforward calculations we get
  \[ \Delta a^2 = \frac{SD_R^2}{S_{xx}} \]
  \[ \Delta b^2 = SD_R^2 \left[ \frac{1}{n} + \frac{M_x^2}{S_{xx}} \right] \]

\[ a = \frac{S_{xy}}{S_{xx}} \]
\[ b = M_y - aM_x \]
Confidence intervals on fit parameters

- $\Delta a$ and $\Delta b$ represent the width of their sampling distributions.

- Hence, they are standard errors:
  
  $$SE_a = \frac{SD_R}{\sqrt{S_{xx}}} \quad SE_b = SD_R \sqrt{\frac{1}{n} + \frac{M_x^2}{S_{xx}}}$$

- We can find confidence intervals:
  
  $$a - t^*SE_a \leq \alpha \leq a + t^*SE_a$$
  $$b - t^*SE_b \leq \beta \leq b + t^*SE_b$$

- where $t^*$ is the critical value from t-distribution with $n - 2$ degrees of freedom.

**Sampling distribution of the slope.** 100,000 samples of size $n$ were randomly drawn from the Hong Kong Survey, and fitted with a straight line.
Example, Hong Kong sample

\[ M_x = \frac{1}{n} \sum_{i=1}^{n} x_i \quad M_y = \frac{1}{n} \sum_{i=1}^{n} y_i \]

\[ S_{xx} = \sum_{i=1}^{n} (x_i - M_x)^2 \quad S_{yy} = \sum_{i=1}^{n} (y_i - M_y)^2 \]

\[ S_{xy} = \sum_{i=1}^{n} (x_i - M_x)(y_i - M_y) \]

\[ a = \frac{S_{xy}}{S_{xx}} \quad b = M_y - aM_x \]

\[ SD_R = \sqrt{\frac{S_{yy} - aS_{xy}}{n-2}} \]

\[ SE_a = \frac{SD_R}{\sqrt{S_{xx}}} \quad SE_b = SD_R \sqrt{\frac{1}{n} + \frac{M_x^2}{S_{xx}}} \]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>h (m)</td>
<td>1.66</td>
<td>1.70</td>
<td>1.64</td>
<td>1.74</td>
<td>1.72</td>
<td>1.82</td>
<td>1.78</td>
<td>1.74</td>
</tr>
<tr>
<td>m (kg)</td>
<td>50.9</td>
<td>56.5</td>
<td>54.0</td>
<td>57.5</td>
<td>55.0</td>
<td>64.5</td>
<td>62.6</td>
<td>54.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>h (m)</td>
<td>1.68</td>
<td>1.76</td>
<td>1.69</td>
<td>1.74</td>
<td>1.77</td>
<td>1.69</td>
<td>1.78</td>
<td>1.77</td>
</tr>
<tr>
<td>m (kg)</td>
<td>49.9</td>
<td>62.5</td>
<td>62.5</td>
<td>65.8</td>
<td>68.4</td>
<td>60.9</td>
<td>74.3</td>
<td>64.3</td>
</tr>
</tbody>
</table>
Example, Hong Kong sample

- Calculate
  \[ M_x = 2.995 \, \text{m}^2 \]
  \[ M_y = 60.24 \, \text{kg} \]
  \[ S_{xx} = 0.4439 \, \text{m}^4 \]
  \[ S_{yy} = 663.4 \, \text{kg}^2 \]
  \[ S_{xy} = 12.12 \, \text{kg m}^2 \]

- Find slope and intercept
  \[ a = 27.30 \, \text{kg m}^{-2} \]
  \[ b = -21.53 \, \text{kg} \]

- Find their standard errors
  \[ SD_R = 4.873 \, \text{kg} \]
  \[ SE_a = 7.314 \, \text{kg m}^{-2} \]
  \[ SE_b = 21.94 \, \text{kg} \]

- Critical \( t^* = 2.145 \) for \( n - 2 = 14 \) d.o.f.

- Finally, the 95\% confidence intervals
  \[ a = 27 \pm 16 \, \text{kg m}^{-2} \]
  \[ b = -22 \pm 47 \, \text{kg} \]
Linear fit prediction errors

- Linear fit gives prediction for every $x$:
  \[ y(x) = ax + b \]

- Can we find uncertainty of $y(x)$?
- $y$ is a function of $a$ and $b$, so we can propagate errors:
  \[ SE_y^2 = \left( \frac{\partial y}{\partial a} \right)^2 SE_a^2 + \left( \frac{\partial y}{\partial b} \right)^2 SE_b^2 \]

This is wrong!
Linear fit prediction errors

- Linear fit gives prediction for every $x$:
  
  $$y(x) = ax + b$$

- Can we find uncertainty of $y(x)$?

- $y$ is a function of $a$ and $b$, so we can propagate errors

- Keep in mind that $a$ and $b$ are strongly correlated

$$SE_y^2 = \left( \frac{\partial y}{\partial a} \right)^2 SE_a^2 + 2 \frac{\partial y}{\partial a} \frac{\partial y}{\partial b} \text{Cov}(a, b) + \left( \frac{\partial y}{\partial b} \right)^2 SE_b^2$$

- After some derivations

$$SE_y = SD_R \sqrt{\frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}}}$$

Correlation between fit parameters. 1000 samples of size $n = 16$ were drawn from the Growth Survey, fitted with the linear regression, parameters $a$ and $b$ found.
Linear fit prediction errors

- Standard error of $y(x)$ is
  \[ SE_y(x) = SD_R \sqrt{ \frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}} } \]

- From this we can find confidence intervals
  \[ y(x) - t^*SE_y \leq \bar{y}(x) \leq y(x) + t^*SE_y \]

- $t^*$ is a critical value from t-distribution with $n - 2$ degrees of freedom

- We can find these errors at any $x$

Best-fitting regression line and its 95% confidence intervals

\[ a = 27 \pm 16 \text{ kg m}^{-2} \]
\[ b = -22 \pm 47 \text{ kg} \]
Linear regression summary

- Simple linear regression $y(x) = ax + b$
- Best-fitting parameters
  
  $a = \frac{S_{xy}}{S_{xx}} \quad b = M_y - aM_x$

- Assumption: $x$ is measured accurately, scatter is caused by error in $y$
- Guessed common error of $y$, $\Delta y_i = SD_R$
- Standard error of the slope and intercept can be propagated from $SD_R$

  $SE_a = \frac{SD_R}{\sqrt{S_{xx}}} \quad SE_b = SD_R \sqrt{\frac{1}{n} + \frac{M_x^2}{S_{xx}}}$

- Confidence intervals are $t^*SE$ for $n - 2$ degrees of freedom
- Standard error of $y$ can be propagated from $a$ and $b$

  $SE_y = SD_R \sqrt{\frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}}}$

- Confidence interval is $t^*SE_y$ for $n - 2$ degrees of freedom
Hand-outs available at http://is.gd/statlec

Please leave your feedback forms on the table by the door
General curve fitting

- We want to fit a model to x-y data
- Example: exponential decay
- Fit data points \((t_i, y_i, \Delta y_i)\) with a function
  \[
y(t) = A_0 + Ae^{-t/\tau}
  \]
- Find best-fitting time scale, \(\tau\), and find its error (95% CI)
- Minimize goodness of the fit:
  \[
  \chi^2 = \sum_{i=1}^{n} \left[ \frac{y_i - y(t_i)}{\Delta y_i} \right]^2
  \]
- From tables of \(\chi^2\) distribution we can find 95% probability corresponds to \(\chi^2 = 3.84\)
- Move the fitted parameter, \(\tau\), left and right from the best-fitting value until \(\chi^2\) increases by 3.84
- We find \(\tau = 2^{+1.1}_{-0.6}\)