Error theory

Introduction to Synthetic Biology

E Navarro
A Montagud
P Fernandez de Cordoba
JF Urchueguía

Overview

- Introduction to measurement theory.
- Error theory description.
- Types of errors.
- Evaluation of the error in an experimental measurement.
- Minimal square fit.
- Obtaining relations from experimental data.
Statistics: A collection of procedures and processes to enable researchers in the unbiased pursuit of Knowledge

Statistics is an important part of the Scientific Method

State a Hypothesis

Design a Study and Collect Data

Interpret the Results—Draw Conclusions

Analyze the Data

State a Hypothesis: The OBJECTIVE or OBJECTIVES of the Study

A HYPOTHESIS OR SET OF HYPOTHESES should state exactly what you want to DO or LEARN or STUDY

SHOULD ANSWER

What are the factors to be studied and what relationships are to be investigated? What is the experimental material? Etc.?
The area of STATISTICS would not be needed if each time you measured an experimental unit you obtained the same response or value.

BUT, THE RESPONSES ARE NOT THE SAME SINCE THERE IS VARIABILITY or NOISE IN THE SYSTEM.

STATISTICAL METHODS EXTRACT THE SIGNAL FROM THE NOISE TO PROVIDE INFORMATION.

---

DESIGN VS. ANALYSIS

The PURPOSE OF DATA COLLECTION is to GAIN INFORMATION OR KNOWLEDGE!!

Collecting Data does not guarantee that information is obtained.

INFORMATION ≠ DATA

At best:

INFORMATION=DATA+ANALYSIS
If data are collected such that they contain NO information in the first place, then the analysis phase cannot find it!!!

The best way to ensure that appropriate information is contained in the collected data is to DESIGN (plan) and Carefully Control the DATA COLLECTION PROCESS.

The measured variables must be in accordance to the stated OBJECTIVES of the study.

What is error theory?

- It is a theory of measurements.
  - There are errors associated to instruments.
  - There are errors associated with human beings.
  - There are errors associated to our mathematical limitations.

\[ 3.56 \pm 0.23 \]
\[ 3.563 \pm 0.236 \]
The error is important to:
- The precision of the obtained results.
- The number of numbers that we should take into account.
- To decide which could be the best measurement strategy.

**Classification of errors**

- **Systematic errors**
  - Instrument malfunction.
    - Soluble: calibrating the device

The devices used to calibrate are more expensive. However, sometimes it is also more difficult to work with them.
Classification of errors

- Human error.

As these errors are difficult to detect, it is convenient to proceed to a recalibration of the devices periodically.

Fidelity — In a device, it is the systematic error that we make when we use the device.

The repeatness is also a very important factor in the measurement of a device.

Classification of errors

- Accidental errors:
  - All the measurements have an imprecision because it is impossible to control everything in all experiments.
  - These fluctuations are taken into account with the absolute error $\varepsilon$.
  - There is another way to express the fluctuations with the magnitude called $\varepsilon_r$.

$$\varepsilon_r = \frac{\varepsilon}{V}$$
The error theory is basically the study of $\varepsilon$.

How to express the magnitude of a measurement?

3.418 ± 0.123  
6.3 ± 0.085  
46.288 ± 1533  
428.351 ± 0.27  
0.01683 ± 0.0058

3.49±0.01  
3.49±0.01  
3.49±0.01  
3.48±0.01

The theory of errors is a measurement theory to reduce the value of them.

The arithmetic media of all the measurements has a lower error than the different individual measurements

\[ x_1 = x \pm \varepsilon_1; \quad x_2 = x \pm \varepsilon_2; \ldots; \quad x_n = x \pm \varepsilon_n \]

\[ x = \frac{x_1 + x_2 + \ldots + x_n}{n} + \frac{\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n}{n} \]

\[ x_m = \frac{x_1 + x_2 + \ldots + x_n}{n} \]
It can be seen that the mean value has the lower error.

\[
\sum e_i^2 = nx^2 \pm \sum x_i^2 - 2x \sum x_i
\]

\[\text{Derivative}\]

\[2nx - 2 \sum x_i = 0 \rightarrow x = \sum \frac{x_i}{n} = x_{\text{ma}}\]

---

**How to estimate \( \varepsilon \)**

- **Direct measurements**

\( \varepsilon \) of a direct measurement is the sensibility of the measurement device.
How to estimate $\varepsilon$

- How many measurements should be performed?

Develop 3 measurements and calculate the relative dispersion:

$$T = 100 \frac{V_{\text{max}} - V_{\text{min}}}{V}$$

<table>
<thead>
<tr>
<th>$T$</th>
<th>Nº of meas.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T &lt; 2%$</td>
<td>3</td>
</tr>
<tr>
<td>$2% &lt; T &lt; 8%$</td>
<td>6</td>
</tr>
<tr>
<td>$8% &lt; T &lt; 15%$</td>
<td>15</td>
</tr>
<tr>
<td>$15% &lt; T$</td>
<td>50</td>
</tr>
</tbody>
</table>

If $T < 8\%$ the error associated to the mean value will be given by the maximum of these two errors:

$$d = \frac{|x_{\text{max}} - x_{\text{min}}|}{4}$$

$$m = \frac{|x_i - \bar{x}|}{n}$$

If $8\% < T < 15\%$ the error associated to the mean value will be given by:

$$m = \sqrt{\frac{(x_i - \bar{x})^2}{n(n - 1)}}$$
How to estimate $\varepsilon$

- If $T$ was larger then we should perform enough measurements to obtain a representative distribution function and the error would be obtained from that distribution.

How to estimate $\varepsilon$

- How to determine the error of a magnitude measured indirectly.

The velocity is usually determined measuring space and time.

\[ V = \frac{S}{t} \]

Some expressions can have some irrational numbers which also introduce an error.
How to estimate $\varepsilon$

- There are three cases:
  - Direct problem
  - Indirect problem
  - Inverse problem

**Direct problem**

When we have a straightforward relation between our magnitude $z$ and magnitudes that can be measured and we do not have irrational numbers.

$$z = f(x, y...)$$

$$dz = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + ...$$

$$\varepsilon(z) = \left| \frac{\partial f}{\partial x} \varepsilon(x) \right| + \left| \frac{\partial f}{\partial y} \varepsilon(y) \right| + ...$$
How to estimate $\varepsilon$

- A particular case:

$$z = Ax^a y^b ...$$

$$\ln z = \ln A + a \ln x + b \ln y + ...$$

$$\frac{dz}{z} = a \frac{dx}{x} + b \frac{dy}{y} + ...$$

$$\varepsilon_{z} = |a \varepsilon_{x}| + |b \varepsilon_{y}| + ...$$

How to estimate $\varepsilon$

- Semidirect problem.

When we have a direct relation but the error of some of the variables are unknown or there are irrational numbers

$$z = f(x, y, ..., a, b)$$

$$\varepsilon_{z} = |\frac{\partial f}{\partial x}\varepsilon_{x}| + |\frac{\partial f}{\partial y}\varepsilon_{y}| + ... + |\frac{\partial f}{\partial \mu}\varepsilon_{\mu}| + |\frac{\partial f}{\partial v}\varepsilon_{v}| + ... + |\frac{\partial f}{\partial a}\varepsilon_{a}| + |\frac{\partial f}{\partial b}\varepsilon_{b}| + ...$$
How to estimate $\varepsilon$

$$
\varepsilon_z = \left| \frac{\partial f}{\partial x} \varepsilon_x \right| + \left| \frac{\partial f}{\partial y} \varepsilon_y \right| + \cdots + \left| \frac{\partial f}{\partial \mu} \varepsilon_\mu \right| + \left| \frac{\partial f}{\partial v} \varepsilon_v \right| + \cdots + \left| \frac{\partial f}{\partial a} \varepsilon_a \right| + \left| \frac{\partial f}{\partial b} \varepsilon_b \right| + \cdots
$$

$\frac{nT}{0.1A}$ $z$ $1.1A$

$z$ $A$ $nT$

$T$ $\frac{0.1A}{n}$ $Z$

---

How to estimate $\varepsilon$

- **Inverse problem**

Sometimes it is desirable to obtain the value of $z$ with a certain error. In those cases that error is the one which imposes conditions in the other magnitudes.

$z$ $A$ $nT$
How to estimate $\varepsilon$

- How to estimate the error from a graph

\[ x \pm \varepsilon \]
\[ z_{\text{max}} = z_{(x+\varepsilon)} \]
\[ z_{\text{min}} = z_{(x-\varepsilon)} \]
\[ z = \frac{z_{\text{max}} - z_{\text{min}}}{2} \]

- How to determine the error from a table

\[ z \quad z, \quad \frac{z_2 - z_1}{x_2 - x_1} = \frac{z_1}{x_1} \quad x \begin{array}{|c|c|} \hline z_1 & x_1 \\ \hline z_2 & x_2 \\ \hline \end{array} \]

The assumption of linearity is present in interpolation.

If we are trying to estimate the values outside the table we make an extrapolation.
Logarithmic scale

This representation smooths the representation of experimental data which increases highly in certain periods of time.

The exponential functions are transformed to linear functions in a logarithmic plot.

\[ y = a_0 e^{-\alpha t} \rightarrow \log y = \log a_0 - t \alpha \log e \]
Minimal square fitting

It is very common to have experimentally a set of data \((x_i, y_i)\) which are related by a mathematical expression:

\[
y = mx + n
\]

Lineal case

Unknown variables

MSF estimates the values of \(m\) and \(n\) which allows the best correlation between \(x\) and \(y\)

\[
\sigma^2(a, b) = \frac{1}{N} \sum_{i=1}^{N} (y_i - y(x_i))^2
\]

Minimize

\[
y(x_i) = mx + n
\]

Minimal square fitting

\[
\frac{\partial \sigma^2}{\partial a} = 0 \quad \text{Pa + Qb = R}
\]

\[
\frac{\partial \sigma^2}{\partial b} = 0 \quad \text{Qa + Nb = S}
\]

\[
P = \sum_{i=1}^{N} x_i^2 \quad Q = \sum_{i=1}^{N} x_i \quad R = \sum_{i=1}^{N} x_i y_i \quad S = \sum_{i=1}^{N} y_i
\]

\[
a = \frac{RN - QS}{PN - Q^2} \quad b = \frac{PS - QR}{PN - Q^2}
\]

\[
\epsilon(a) = \left| \frac{\partial a}{\partial P} \epsilon(P) \right| + \left| \frac{\partial a}{\partial Q} \epsilon(Q) \right| + \left| \frac{\partial a}{\partial R} \epsilon(R) \right| + \left| \frac{\partial a}{\partial S} \epsilon(S) \right|
\]

\[
\epsilon(b) = \left| \frac{\partial b}{\partial P} \epsilon(P) \right| + \left| \frac{\partial b}{\partial Q} \epsilon(Q) \right| + \left| \frac{\partial b}{\partial R} \epsilon(R) \right| + \left| \frac{\partial b}{\partial S} \epsilon(S) \right|
\]
Minimal square fitting

Correlation factor

\[ r = \frac{(n R) - QS}{\sqrt{(n P - Q^2)\left(n \sum_{i=1}^{n} y_i^2 - S^2\right)}} \]

- 0.75 < |r| < 1: good correlation.
- 0.25 < |r| < 0.75: There is a slight linear tendency.
- 0 < |r| < 0.25: It is very improbable a linear tendency between x,y

Usefulness

- To know if there is a functional dependence between two variables.
- To reduce the error in the determination of a variable which is known that it is related to other two by a physical relation.

\[ y = k \Delta x \]

- The results of a fit could be used to estimate the values of not measured y
- To calibrate devices.
Example

\[ y_{eq} = l_0 + \frac{s}{k} m \]

Obtaining a relation from data.
Step 1: Model Variables

- Goal: \( y = f(x_1, x_2, x_3, \ldots) \)

- Identify performance parameter (\( y \)) and design variables (\( x_1, x_2, x_3, \ldots \))
  - Design variables = control variables
  - \( n = \# \) of design variables

- Note any noise variables (things which you cannot control)
Step 2: Variable Targets & Boundaries

- Specify target for performance parameter
  - Use QFD for this
- Determine bounds on design variables
  - Step 6 of RE & Redesign Methodology

Step 3: Experimental Plan

- Design the experiment
  - Specify levels of design variables
    - Scale DV to [-1, +1]
  - Calculate number of trials, \( N = (\text{levels})^n \)
  - Decide on the number of replicates
    - Replicate – repeat trial of an experiment

- Plan how to measure DV and performance parameter
Step 3: Experimental Plan (2)

- For a basic linear model
  - factorial experiments \( (2^n) \)
  - \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \)
- Collect data to determine \( \beta \) coefficients
  - Use an *experimental matrix* to show all permutations of DV
- For example, consider a \( 2^3 \) experiment:
  - \( N = 2^3 = 8 \) trials

Step 3: Experimental Plan (3)

- Experimental matrix for \( 2^3 \) factorial exp.:

<table>
<thead>
<tr>
<th>Trial</th>
<th>Vect.</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y_1(d) )</th>
<th>( y_2(d) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( d_1 )</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>( d_2 )</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>( d_3 )</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>( d_4 )</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>( d_5 )</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>6</td>
<td>( d_6 )</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>7</td>
<td>( d_8 )</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>8</td>
<td>( d_8 )</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
</tbody>
</table>
Step 3: Experimental Plan (4)

- -1 = low value of design variable
- +1 = high value of design variable
- For the matrix shown, number of replicates is 2

Step 4: Testing

- Perform tests in random order
- Keep noise variables constant (as much as possible)
Step 5: Analysis

- Determine $\beta$ coefficients
  - Regression analysis may be used
- $\beta$s can be defined in terms of the effect a variable $x_i$ has on the perf. parameter

\[
E_i = \frac{\sum_{h,x_i=+} y_h - \sum_{h,x_i=-} y_h}{N/2}, \quad \text{where } h = 1 \ldots N
\]
\[
= \sum \text{responses at high} - \sum \text{responses at low} \quad \frac{N}{2}
\]

Step 5: Analysis (2)

- $\beta$ coefficients are then determined as:

\[
E_i \quad \frac{\sum_{h=1}^{N} y_{h}}{N}
\]

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
  - for a $2^3$ factorial experiment
Using the Effects

- Each $E_i$ holds useful information about each DV $x_i$
  - If $E_i$ is near zero, then $x_i$ has little effect on $y$
  - If $E_i$ is large, the $x_i$ significantly affects $y$

- A graphical representation of the response vs. each DV (called a response diagram) is helpful in showing these facts

Using the Effects (2)

- From a response diagram point of view:
  - Plot values of $y$ vs. $x_i(-)$ and $x_i(+)$. 

![Important DV](image1)

![DV has no effect on perf.](image2)
Using the Effects (3)

- Use a response diagram to check the validity of the linear model assumption.

![Diagram showing response spread from low to high, indicating weak linear model.]

Spread in the data from low to high indicates linear model is weak.

Replicates

- Def'n: Replicate is a repeat trial.
- Use to check the significance of results:
  - Replicates rarely produce the exact same responses.
- Calculate the variance for each trial:

\[
s_i^2 = \frac{\sum_{j=1}^{r} y_{ij}^2 - \left( \frac{\sum_{j=1}^{r} y_{ij}}{r} \right)^2}{r - 1}
\]

where \( r \) is the number of replicates, \( y_{ij} \) is the response for trial \( i \) and replicate \( j \), and \( \bar{y}_i \) is the average response for trial \( i \).
Replicates (2)

- For all trials:
  - Calculate the standard deviation (exp. error) of experiment, \( s_T \)

\[
s_T^2 = \frac{1}{N} \sum_{i=1}^{N} s_i^2
\]

  (average variance)

- If \( 3 \cdot s_T < E_i \), then \( x_i \) is significant (i.e., more than noise) within a 99.7% confidence level

Interactions of DV

- If a pure linear model is weak (from response diagram), use an interaction model
  - \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \ldots + \beta_{n-1,n} x_{n-1} x_n \)

- Same approach as linear model
  - Define new DV as \( x_{12} = x_1 x_2 \), etc.
  - \( \beta \) coefficients defined as \( \beta_{ij} = E_{ij} / 2 \)
Interactions of DV (2)

- For a $2^3$ factorial experiment, we add DV to the experimental matrix:

<table>
<thead>
<tr>
<th>Trial</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_{12}$</th>
<th>$x_{13}$</th>
<th>$x_{23}$</th>
<th>$x_{123}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DOE Summary

- DOE provides a methodical approach to develop an empirical model of a physical phenomenon

- A basic linear or interaction (nonlinear) model can be constructed by performing $[levels]^n$ experiments

- Significant DV are determined by the DOE analysis
Advanced DOE

- For a DOE with more than 2 levels, the number of experiments increases exponentially.

- Instead of completing a full factorial experiment, a fractional experiment may be performed.

Bibliography

- Design and analysis of experiments. A Dean and D Boss. Springer

- Design and analysis of experiments. DC Montgomery. Wiley and Sons

- Statistical design and analysis of experiments. RL Mason, LM Gunst and JL Guess. Wiley and Sons