


Error theory

Introducción to Synthetic Biology


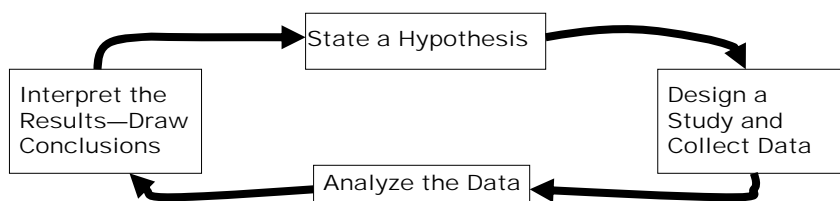
Overview

- ▣ 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: 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 \neq 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!!!

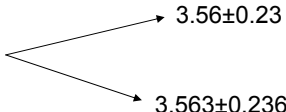
The best way to ensure that appropriate information is contained in the collected data is to DESIGN (plan) and Carefully Control the DATA COLLECTION PROCESS

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.

Measurement 

3.56±0.23

3.563±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



$f(T)$



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 ε .

There is another way to express the fluctuations with the magnitude called ε_r .

$$\varepsilon_r = \frac{\varepsilon}{V}$$

The error theory is basically the study of ϵ
 How to express the magnitude of a measurement?

$3,418 \pm 0,123$
 $6,3 \pm 0,085$
 46288 ± 1533
 $428,351 \pm 0,27$
 $0,01683 \pm 0,0058$

3.49 ± 0.01
 $3.49 \pm 0.01 \longrightarrow 3.49 \pm 0.01 \longrightarrow 3.487 \pm 0.010$
 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

$$x_1 = x \pm \epsilon_1; \quad x_2 = x \pm \epsilon_2; \quad \dots; \quad x_n = x \pm \epsilon_n$$

$$x = \frac{x_1 + x_2 + \dots + x_n}{n} + \frac{\epsilon_1 + \epsilon_2 + \dots + \epsilon_n}{n}$$

$$x_m = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\pm\epsilon_1 = x - x_1; \quad \pm\epsilon_2 = x - x_2; \quad \pm\epsilon_n = x - x_n$$

It can be seen that the mean value has the lower error

$$\sum \epsilon_i^2 = nx^2 - 2x \sum x_i + \sum x_i^2$$

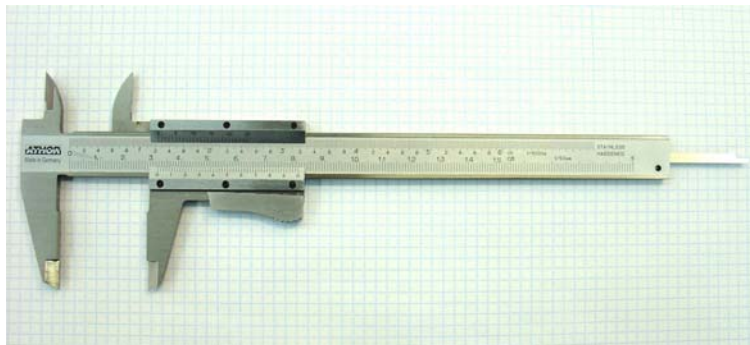
Derivative

$$2nx - 2 \sum x_i = 0 \rightarrow x = \frac{\sum x_i}{n} = x_m$$

How to estimate ϵ

▣ Direct measurements

ϵ of a direct measurement is the sensibility of the measurement device.



How to estimate ε

- How many measurements should be performed?

Develop 3 measurements and calculate the relative dispersion:

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

T	N° of meas.
$T < 2\%$	3
$2\% < T < 8\%$	6
$8\% < T < 15\%$	15
$15\% < T$	50

How to estimate ε

- How to estimate ε

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

$$D = \frac{|x_{\max} - x_{\min}|}{4} \quad m = \frac{|x_i - \bar{x}|}{n}$$

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 ε

- ▣ 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 ε

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

The velocity is usually determined measuring space and time

$$\longrightarrow V = \frac{S}{t}$$

Some expressions can have some irrational numbers which also introduce an error

How to estimate ε

□ There are three cases:

- Direct problem
- Indirect problem
- Inverse problem

How to estimate ε

□ 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.

$$z = f(x, y, \dots) \quad dz = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \dots$$

$$\varepsilon(z) = \left| \frac{\partial f}{\partial x} \varepsilon(x) \right| + \left| \frac{\partial f}{\partial y} \varepsilon(y) \right| + \dots$$

How to estimate ε

▣ A particular case:

$$z = Ax^a y^b \dots$$

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

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

$$\epsilon_{r,z} = |a \epsilon_{r,x}| + |b \epsilon_{r,y}| + \dots$$

How to estimate ε

▣ Semidirect problem.

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

$$z = f(x, y, \dots, \mu, \nu, \dots, a, b)$$

$$\epsilon_z = \left| \frac{\partial f}{\partial x} \epsilon_x \right| + \left| \frac{\partial f}{\partial y} \epsilon_y \right| + \dots + \left| \frac{\partial f}{\partial \mu} \epsilon_\mu \right| + \left| \frac{\partial f}{\partial \nu} \epsilon_\nu \right| + \dots + \left| \frac{\partial f}{\partial a} \epsilon_a \right| + \left| \frac{\partial f}{\partial b} \epsilon_b \right| + \dots$$

How to estimate ε

$$\epsilon_z = \left| \frac{\partial f}{\partial x} \epsilon_x \right| + \left| \frac{\partial f}{\partial y} \epsilon_y \right| + \dots + \underbrace{\left| \frac{\partial f}{\partial \mu} \epsilon_\mu \right| + \left| \frac{\partial f}{\partial \nu} \epsilon_\nu \right| + \dots + \left| \frac{\partial f}{\partial a} \epsilon_a \right| + \left| \frac{\partial f}{\partial b} \epsilon_b \right| + \dots}_{\mathbf{T}}$$

$$\begin{array}{ccc} nT & 0.1A & z & 1.1A \\ z & A & nT & \\ T & \frac{0.1A}{n} & z & \end{array}$$

How to estimate ε

▣ 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.

$$z \quad A \quad nT$$

How to estimate ε

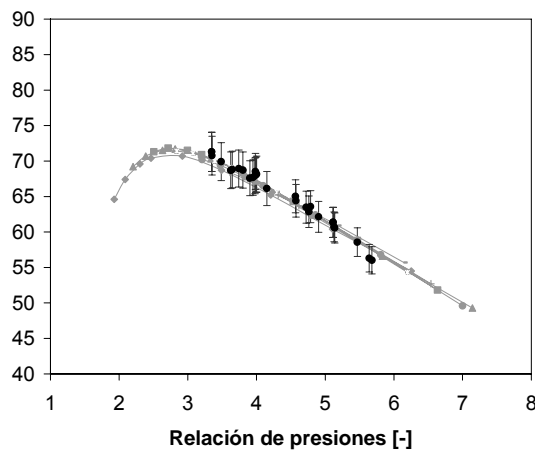
How to estimate the error from a graph

$$x \pm \varepsilon$$

$$z_{\max} = z_{(x+\varepsilon)}$$

$$z_{\min} = z_{(x-\varepsilon)}$$

$$z = \frac{|z_{\max} - z_{\min}|}{2}$$



How to estimate ε

How to determine the error from a table

$$z = z_1 + \frac{z_2 - z_1}{x_2 - x_1}(x - x_1)$$

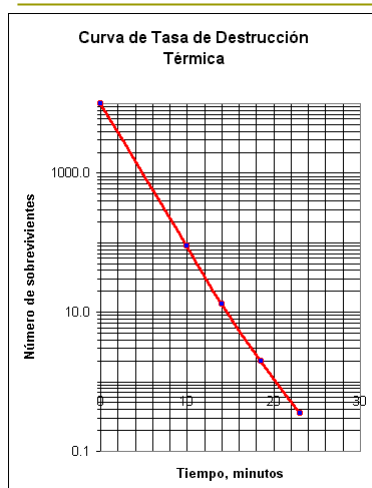
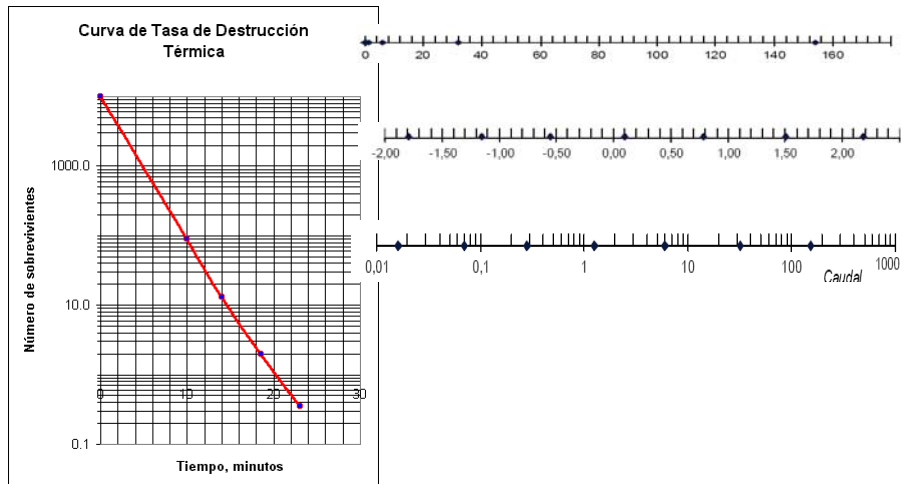
z1	x1
z2	x2

$$\left| \frac{z_2 - z_1}{x_2 - x_1} \right| \cdot x$$

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.

$$y = a_0 e^{-at} \rightarrow \log y = \log a_0 - ta \log e$$

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:

$$y = \overset{\text{Unknown variables}}{\underset{\text{Unknown variables}}{m}}x + \underset{\text{Unknown variables}}{n} \longrightarrow \text{Lineal case}$$

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

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

$$y(x_i) = mx + n$$

Minimal square fitting

$$\frac{\partial \sigma^2}{\partial a} = 0$$

$$Pa + Qb = R$$

$$\frac{\partial \sigma^2}{\partial b} = 0$$

$$Qa + Nb = S$$

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

$$a = \frac{RN - QS}{PN - Q^2}$$

$$b = \frac{PS - QR}{PN - Q^2}$$

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

$$\epsilon(b) = \left| \frac{\partial b}{\partial P} \epsilon(P) \right| + \left| \frac{\partial b}{\partial Q} \epsilon(Q) \right| + \left| \frac{\partial b}{\partial R} \epsilon(R) \right| + \left| \frac{\partial b}{\partial S} \epsilon(S) \right|$$

Minimal square fitting

Correlation factor

$$r = \frac{(nR) - QS}{\sqrt{(nP - Q^2)(n \sum_{i=1}^n y_i^2 - S^2)}}$$

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

Minimal square fitting

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.

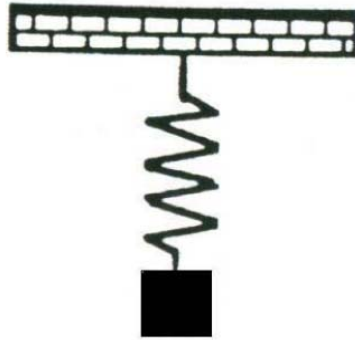
$$y = k\Delta x$$

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

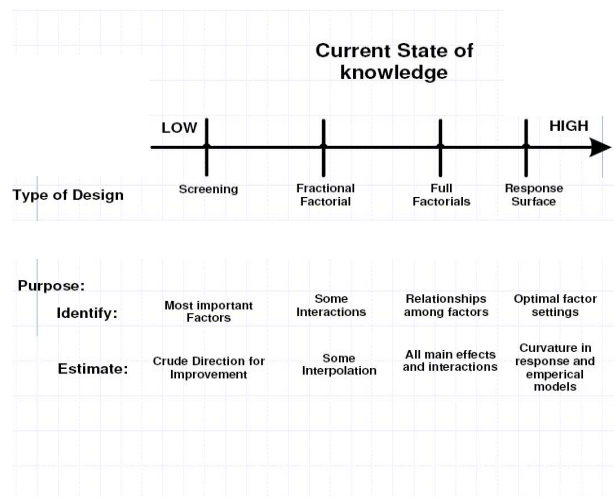
Example

$$y_{eq} = l_0 + \frac{g}{K}m$$

y	m



Obtaining a relation from data.



Step 1: Model Variables

- Goal: $y = f(x_1, x_2, x_3, \dots)$
- Identify performance parameter (y) and design variables (x_1, x_2, x_3, \dots)
 - Design variables = *control* variables
 - n = # of design variables
- Note any noise variables (things which you can not 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)
 - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$
- Collect data to determine β 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.:

Trial	Vect.	x_1	x_2	x_3	$y_1(d)$	$y_2(d)$
1	d₁	-1	-1	-1		
2	d₂	+1	-1	-1		
3	d₃	-1	+1	-1		
4	d₄	+1	+1	-1		
5	d₅	-1	-1	+1		
6	d₆	+1	-1	+1		
7	d₇	-1	+1	+1		
8	d₈	+1	+1	+1		



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 β coefficients
 - Regression analysis may be used
- β_s can be defined in terms of the *effect* a variable x_i has on the perf. parameter

$$E_i = \frac{\sum_{h: x_i = (+)} y_h - \sum_{h: x_i = (-)} y_h}{N/2}, \quad \text{where } h = 1 \dots N$$

$$= \frac{\sum \text{responses at high} - \sum \text{responses at low}}{N/2}$$

Step 5: Analysis (2)

- β coefficients are then determined as:

$$i \quad \frac{E_i}{2} \quad \text{and} \quad 0 \quad \frac{\sum_{h=1}^N y_h}{N}$$

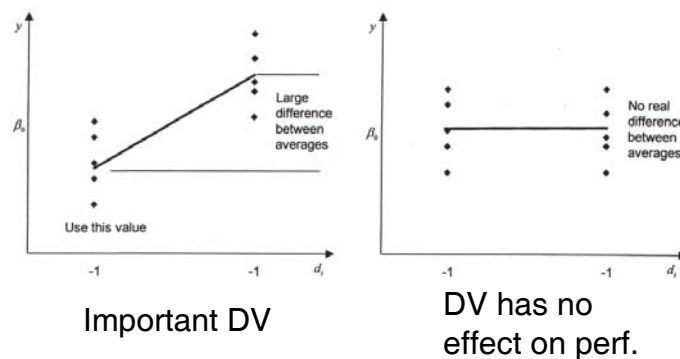
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
 - for a 2^3 factorial experiment

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

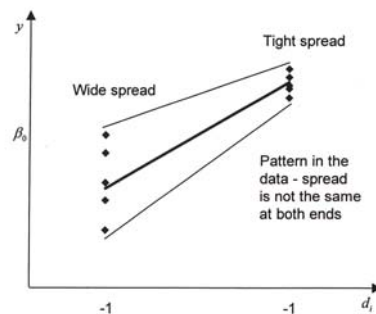
Using the Effects (2)

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



Using the Effects (3)

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



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:

$$s_i^2 = \frac{y_{i_1}^2 - \bar{y}_i^2 + \dots + y_{i_r}^2 - \bar{y}_i^2}{r - 1}$$

where r = # of replicates

$$\bar{y}_i = \frac{y_{i_1} + \dots + y_{i_r}}{r} \quad \text{(average response for trial } i \text{)}$$

Replicates (2)

□ For all trials:

- Calculate the standard deviation (exp. error) of experiment, s_T

$$s_T^2 = \frac{1}{N} \sum_{i=1}^N s_i^2 \quad (\text{average variance})$$

- 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

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \dots + \beta_{n-1,n} x_{n-1} x_n$$

- ### □ Same approach as linear model

- Define new DV as $x_{12} = x_1 x_2$, etc.
- β 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:

Trial	x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}
1	-1	-1	-1	+1	+1	+1	-1
2	+1	-1	-1	-1	-1	+1	+1
3	-1	+1	-1	-1	+1	...	
4	+1	+1	-1	...			
5	-1	-1	+1				
6	+1	-1	+1				
7	-1	+1	+1				
8	+1	+1	+1				

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