Statistics and Quantitative Biology

Monday, September 28, 2009

Overview

Objectives

To appreciate and feel comfortable with uncertainty

To assimilate the principles of statistical inference, estimation and hypothesis testing

To get a primer in doing statistics with R

To better plan your experiments

To appreciate how much you can benefit by talking to a statistician or even to a “young apprentice” before you do an experiment

Not in the objectives

That you learn a list of statistical recipes and procedures
We do not see causes directly

A bit of epistemology

Natural Sciences

Explaining Nature

Explaining biological systems
Expectation and Prediction

Ptolemaic astronomy and Newtonian astronomy

Mendelian genetics and Modern synthesis

Statistical inference and causal inference

Ptolemy
deferents and epicycles

“After the discovery of Uranus, it was noticed that its orbit was not as it should be in accordance with Newton’s laws. It was therefore predicted that another more distant planet must be perturbing Uranus’ orbit. Neptune was first observed by Galle and d’Arrest on 1846 Sept 23 very near to the locations independently predicted by Adams and Le Verrier from calculations based on the observed positions of Jupiter, Saturn, and Uranus.”

Isaac Newton
Reification of the gene

Morgan chromosomes
Muller X-ray gene mutation
Delbrück variants precede selection
James Watson & Francis Crick DNA structure

Modelling Nature, Natural Law

At least two distinct levels of observation

“Structure and function”
The Modelling Relation

Measurement, Interpretation, Prediction

Inference → Model → Natural System → Cause

Measurement


Inferential structure is mapped to causality structure

Models in Biology

Conceptual, Propositional Models in Natural Language

(Cartoons)

Experimental models

Mathematical Models

Cartoons

The lac-operon

Monod’s “model”
Maturana & Varela

Cartoons

What about time?
Cartoons

What about dynamics?

Why are mathematical descriptions and models better?

Through mathematics one can be cold-bloodedly objective
“It is often held that science and common sense are closely linked. Thomas Henry Huxley, Darwin’s brilliant colleague, spoke of science as being nothing more than trained common sense. (…) However reasonable they may sound, such views are, alas, quite misleading. In fact, both the ideas that science generates and the way in which science is carried out are entirely counter intuitive and against common sense. (…) Science does not fit with our natural expectations.”

– Lewis Wolpert

Earth moves around the Sun
Earth is round

“If one bullet is dropped from your hand and another is fired horizontally from a gun at exactly the same time, which will hit the ground first?”

Uniform motion without force (Newton’s first law)
Mathematics is generative

Postulates

Consequences

More Postulates

Consequences

Consequences

Consequences

Consequences

… consequences of consequences of consequences...

Mathematics in communication and intersubjectivity

“Then, 11 months later (ample time for chimerism to develop), we measured chimerism in the blood (..). Unexpectedly, there were no donor-derived B cells and seldom any APCs; the chimeric cells were almost entirely T-cell receptor (TCR)ab +, CD4 or CD8 T cells (called ‘chimeric T cells’ here) (...).

A cataclysmic inflammatory event, for example, from a microbial incursion, might initiate an autoimmune disease process, and any readily processed determinant which gains ascendance in either the class I or class II presentation systems will stimulate ambient T cells. In this case, the first T cell to be activated will either be the most abundant or have the most avid receptors:”

\[
\frac{dx}{dt} = a \cdot x - b
\]

\[
\dot{x} = a \cdot x - b
\]

Occam’s Razor

“Plurality non est ponenda sine necessitate (Plurality should not be posited without necessity)”

William of Occam, XIV century

“Everything should be made as simple as possible,
But not simpler”

Albert Einstein, XX century

Complicate only when simplicity fails

JC
The Modelling Relation

Measurement, Interpretation, Prediction

Inference

Model

Natural System

Cause

Measurement


Inferential structure is mapped to causality structure

Some readings

Pestana & Velosa. Introdução à probabilidade e à estatística. Edições Gulbenkian

Whitlock & Schluter. The analysis of biological data. Roberts and Company Publishers


Cobb. Introduction to design and analysis of experiments. Springer

Stochastic processes, probability, and the central limit theorem
Random events

Probability and probability distributions

Normal distribution and the central limit theorem

Randomness rules the world

Fortuna Imperatrix Mundi

What is Probability?

Probability is a numerical measure of uncertainty about an event.
**Laplace’s Concept**

\[
\text{probability} = \frac{\text{# favorable events}}{\text{# all possible events}}
\]

Problems:
Every event is equally likely to occur
Finite number of events

Pierre Simon Laplace (1749-1827)

**Frequency-based Concept**

It is the relative frequency of occurrence of an event if an experiment is repeated ad infinitum.

J. Bernoulli (1664-1705)

**Classical Statistics**

R. A. Fisher (1890-1962)

K. Pearson (1857-1936)

**Subjective Concept**

It describes the degree of belief of an individual on the occurrence of a certain event.

Rev. Thomas Bayes (1701-1761)

**Bayesian Statistics**

J. Haldane (1892-1964)

H. Jeffreys (1891-1989)
Random Variable

Random experiment:
Every experiment that the outcome is not deterministic.
e.g. tossing a dice, inheritance of a paternal allele

Event space:
The set of all possible outcomes of the random experiment.
tossing a dice = \{1, 2, 3, 4, 5, 6\},
paternal allele = \{grandfather allele, grandmother allele\}

Random variable \(X\):
A mapping that assigns a probability to every possible outcome.

\[
P(X = x) = \frac{1}{6}, \text{ where } x = 1, 2, 3, 4, 5, 6
\]

A few properties of probability and its computation

Probability of an event

(Venn diagram)
Probability that the “outcome is 2”

\[ P[X=2] = \frac{1}{6} \]

Probability that the “outcome is 6”

\[ P[X=6] = \frac{1}{6} \]

Addition of probability of mutually exclusive events: either this or that

\[ P[X=1 \text{ AND } X=2] = 0 \]

\[ P[X=1 \text{ OR } X=2] = P[X=1] + P[X=2] = \frac{2}{6} \]
Probability of all possible mutually exclusive events adds to 1

\[ P[X \neq 1] = 1 - P[X = 1] \]

Probability that the “outcome is an even number”

\[ P[X \text{ is even}] = \frac{3}{6} \]

Complementary events

\[ P[X \text{ is even}] + P[X \text{ is odd}] = 1 \]
\[ P[X \text{ is even AND odd}] = 0 \]
Adding probabilities of non-mutually exclusive events

Probability that the outcome "is an even number"

- $P[X \text{ is even}] = \frac{3}{6}$

Probability that the outcome "is equal to or greater than 3"

- $P[X \geq 3] = \frac{4}{6}$
Probability that the outcome “is equal to or greater than 3” OR “is an even number”

\[ P[X \geq 3 \text{ OR } X \text{ is even}] = P[X \geq 3] + P[X \text{ is even}] - P[X \geq 3 \text{ AND } X \text{ is even}] \]

**Probability Addition Rule**

\[ P[A \text{ OR } B] = P[A] + P[B] - P[A \text{ AND } B] \]

**Independence and the multiplication rule**
Probability Multiplication Rule

If two events A and B are independent

\[ P[A \text{ AND } B] = P[A] \times P[B] \]
What events are interdependent?

**Conditional probability**

Is the probability of an event given that another event occurs, i.e. the probability of an event given that a condition is met

\[ P(A|B) \]

**Law of total probability**

The probability of an event \( X \) is:

\[ P(X) = \sum_Y P(Y) \cdot P(X|Y) \]

where \( Y \) represents all possible mutually exclusive values of the conditions.
**Probability Multiplication Rule**

\[ P(A \text{ AND } B) = P(A|B) \times P(B) = P(B|A) \times P(A) \]

**Bayes’ Theorem**

\[ P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \]

**Example**

Bayes’ Theorem and the diagnosis of Down syndrome

Down syndrome is a chromosomal condition that occurs in about 1:1000 pregnancies. The most accurate test for DS requires amniocentesis, which is not devoid of risk of miscarriage (1:200). It would be good to have an accurate non-invasive test without risks. Recently a method has been proposed involving the quantification of three enzymes in the blood.

This so-called triple test does not always identify a fetus DS (false negative), and some time incorrectly identifies a fetus with a normal set of chromosomes (false positive). Under normal conditions, the detection rate of the triple test (the probability that a fetus with DS is identified correctly) is 0.60. The false positive rate is 0.05 [Newberger et al. 2000 American Family Physician].

For most people’s intuition these numbers are acceptable. Make a guess of the rate at which a fetus positively identified has DS.

Adapted from Whitlock & Schluter: The analysis of biological data. Roberts and Company Publishers.
\[ P[DS] = 0.001 \]
\[ P[TT+ | DS] = 0.60 \]
\[ P[TT+] = 0.001 \times 0.60 + (1 - 0.001) \times 0.05 = 0.05055 \]
\[ P[DS | TT+] = \frac{P[TT+ | DS] \times P[DS]}{P[TT+]} = \frac{0.60 \times 0.001}{0.05055} = 0.012 \]

Recapitulating

**Addition**
\[ P[A \text{ OR } B] = P[A] + P[B] - P[A \text{ AND } B] \]

**Multiplication**

**Total probability**
\[ P[X] = \sum_Y P[Y] \times P[X|Y] \]

**Bayes' Theorem**
\[ P[A|B] = \frac{P[B|A] \times P[A]}{P[B]} \]

Probability Distributions
Probability Distributions

Discrete Probability Distributions

Event space has a countable number of possible outcomes

- **Bernoulli**
  - Tossing a coin
  - Inheritance of a paternal allele

- **Binomial**
  - Tossing a coin n times
  - Number of cured animals after treatment

- **Poisson**
  - Number of births in a certain period
  - Number of mutations in a generation

Continuous Probability Distributions

Event space has a non-countable number of possible events

- **Normal**
  - Height of human population

- **Log-normal**
  - Age at disease onset
  - Survival after cancer diagnosis
  - Specific protein content per cell

- **Exponential**
  - Time between births
  - Lymphocyte life span

Probability is an area under the curve
Cumulative probability functions

\[ F(x) = P[X \leq x] = \sum_{x_i \leq x} P[X = x_i] \]

\[ F(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(x) \, dx \]

Probability functions

- Mass probability function
- Probability Density Function

\[ P[X = x] \]

\[ f_X(x) \]

Normal distribution

Gauss (1777-1855)

Symmetric around the mean
Mean=Median=Mode
Why is Normal Distribution so important?

Can you single out a major concept, experiment, or result in science as THE one?

The Central Limit Theorem
Let $X_1$, $X_2$, $X_3$, ..., $X_n$ be identical and independently distributed random variables with mean $m$ and variance $s^2$. Let $\sum X$ be the sum of the values of the $n$ variables. When $n$ is very large $\sum X$ follows approximately a Normal distribution with mean $n\cdot m$ and variance $n\cdot s^2$. 
As a corollary

A sample mean tends to be distributed as a Normal irrespective of the distribution of the population one is sampling from.
What about some more implications of the Central Limit Theorem?

\[
\frac{d[X]}{dt} = -k_{on}[X][Y] + k_{off}[XY]
\]

Collision Theory of Reaction

Collision frequency (collisions per molecule per unit of time)

\[
z = \frac{2^{1/2} \sigma \bar{v} N}{V} = \frac{2^{1/2} \sigma \bar{v} p}{kT}
\]

Collision density (total collisions per unit of volume)

\[
Z_{AA} = \sigma \left( \frac{4kT}{\pi m} \right)^{1/2} N_A[\Lambda]^2
\]
Collision Theory of Reaction

\[ \frac{d[X]}{dt} = -k_{on}[X][Y] + k_{off}[XY] \]

Valid when we have many, many molecules

**WHAT IS LIFE?**

ERWIN SCHRODINGER

PHYSICAL LAWS REST ON ATOMIC STATISTICS AND ARE THEREFORE ONLY APPROXIMATE

And why could all this not be fulfilled in the case of an organism composed of a moderate number of atoms only and sensitive already to the impact of one or a few atoms only? Because we know all atoms to perform all the time a completely disorderly heat motion, which, so to speak, opposes itself to their orderly behavior and does not allow the events that happen between a small number of atoms to enrol themselves according to any recognizable laws. Only in the cooperation of an enormously large number of atoms do statistical laws begin to operate and control the behavior of these assemblies with an accuracy increasing as the number of atoms involved increases. It is in that way that the events acquire truly orderly features. All the
Schrodinger on diffusion laws …

Being based on pure chance, its validity is only approximate. If it is, as a rule, a very good approximation, that is only due to the enormous number of molecules that co-operate in the phenomenon. The smaller their number, the larger the quite haphazard deviations we must expect and they can be observed under favourable circumstances.

How many copies of each molecule does a cell have?
How are the data produced?

Cell Tissue → Blending → Homogenate
Noise in protein expression scales with natural protein abundance

Xiren Bar-Evan1, Johan Paulsson2,3, Narendra Maheshri4, Miri Carmi5, Erin O'Shea4, Yitzhak Pilpel1 & Naama Barkai1,5

Figure 1: Single cell distributions of homogeneous levels (left). (b) Histograms showing protein abundance after one hour of growth on glucose. Each line represents a different cell. The histograms show the distribution of protein abundance in the single-cell samples. (c) CDF of the log-normal distribution. The CDF is shown in blue, green, red, and orange, respectively. (d) Normalized protein distributions. The histograms show the distribution of protein abundance in the single-cell samples. The distributions are shown in blue, green, red, and orange, respectively. (e) Normalized protein distributions. The histograms show the distribution of protein abundance in the single-cell samples. The distributions are shown in blue, green, red, and orange, respectively. (f) Normalized protein distributions. The histograms show the distribution of protein abundance in the single-cell samples. The distributions are shown in blue, green, red, and orange, respectively.

Any (more) questions before the break?