Global Journal of Science Frontier Research
Mathematics and Decision Sciences
Volume 12 Issue 7 Version 1.0 June 2012
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Biostatistics : A Probable Integration of Mathematics in Biological Systems. *An Aided Philosophical Account*

By Rishan Singh
University of Kwa Zulu-Natal, South Africa

Abstract - The information supplied below is a philosophical account of the importance of statistics in biology. It is aimed at learners, who aspire to further their careers in biology through research. This paper brings recognition to **UKZN** and the Republic of South Africa.

Keywords : *Probability, Statistics, Bioprobability, Biomathematics.*

GJSFR-F Classification MSC 2010: *92B15, 62P10.*

Strictly as per the compliance and regulations of :

Biostatistics : A Probable Integration of Mathematics in Biological Systems. *An Aided Philosophical Account*

Rishan Singh

Abstract - The information supplied below is a philosophical account of the importance of statistics in biology. It is aimed at learners, who aspire to further their careers in biology through research. This paper brings recognition to UKZN and the Republic of South Africa.

Keywords : *Probability, Statistics, Bioprobability, Biomathematics.*

I. Introduction

On a daily basis, a life scientist would encounter times when the only way of understanding important scientific data that is used to write scientific papers, would be through the use of statistics. Therefore, statistics is, as stated by Neuhauser (2004), an indispensible tool for life scientists. However, although one may think that statistics is merely mathematics, in actual fact it is not. By philosophical definition, statistics is a branch of mathematics and it is not confined to its own usage. By this, I mean that statistics is often associated with probability. The implications of this is that we would, therefore, be faced with the possibility of 'guess-work', in order to solve simple biological problems.

Statistics and therefore probability are both reliant on calculus, even though they are not a part of calculus. Therefore, biological systems would have to incorporate such methods in scientific tests so as to test the validity of the results. And because, biostatistics would be testing validity, we can say that it is an objective tool for problem solvers.

II. Statistics Verses Probability

It is possible to separate statistics and probability when defining its usage. The reason for this has to do with the fact that they are both reliant on calculus, as mentioned, and therefore the principle behind the exploitation of these techniques are essentially the same. They can however, be separated on the basis of the type of data that is used and the results that are obtained.

It is important to understand that in biological systems, the sampling of data has to occur randomly for zero bias to be available in the results. The idea of randomness would ensure that no single variable would be favoured, knowingly. Moreover, this would ensure that the results take into account all the possible outcomes, all of which are uncertain but can be repeated. If this rule were violated in any way, the results that would be obtained, without statistical considerations, would be subjective, making the

Author : *University of Kwa Zulu-Natal, South Africa. E-Mail : rshnsingh1@webmail.co.za*

point of the investigation – pointless. Therefore, during statistical calculations, the probability theory, which forms the foundation of statistics, provides the essential tools for randomness to be modelled. Furthermore, it is for this reason that biologists turn to statistics prior to the design of experiments, and subsequently, the setting up of hypotheses to be tested.

However, statistics can be a very 'heavy' tool to employ to determine the objectivity of results obtained from non-deterministic phenomena. One of the main reasons for this is because statistics needs a huge sample set of data in order to be feasible. Some examples of such phenomena, as stated by Neuhauser (2004) are 'the number of eggs laid by a bird, the lifespan of an organism, the inheritance of genes' and even 'the number of people infected during an outbreak of a disease.'

At this point of the article, I would explain the importance of statistics by using bioprobability as an example.

Suppose a black dog (B) mates with a white female dog (W). After one mating (of adult dogs), the possible outcomes of the F_1 offspring are B and W and therefore the sample space can be defined as:

$$\Omega = \{B, W\}$$

However, suppose the F1 generation, which are all fertile grow to reach reproductive age, mate to produce the F2 generation of offspring, the outcome of the first mating followed by the outcome of the second mating, could be **BW**, which essentially brown male(s) and/or female(s). Now since,

Parents mate : $B \times W$
F1 generation : $BW \times BW$
(fertile)
F2 generation : $BB\ BW\ WB\ WW$

Therefore, the sample size is now given by: $\Omega = \{BB, BW, WB, WW\}$

The case of Fatality – the Rb genes – tumour suppressor genes!

The first case of these genes were identified in a type of eye cancer, called hereditary retinoblastoma, which occurs in young children due to family inheritance of defective gene/disease - children who inherit a deletion in a specific region of chromosome 13. However, the disease only becomes distinct when cell division of defective chromosome 13 occurs to form a duplicate of itself.

Assuming the defective gene is indicated by Rb, and the genetic disorder is due to a X-linked recessive inheritance with the mother being the carrier on one of her X-chromosomes (recessive allele), then:

Carrier mother RbX — Normal father **X**Y

(Meiosis, gamete formation)

Rb eggs X — **X** sperm Y

Possible outcome :

XX (Normal Daughter)	XY (Normal Son)
RbX (Carrier Daughter)	RbY (Affected Son)

Now, assume that **X**, i.e. both mother and father are carriers, taking into account the maternal contributor first and then the paternal consideration. The following probable outcomes would be applicable in the situation:

$$\Omega = \{(Rb, Rb), (Rb, Y), (Rb, X), (X, Y)\}$$

Notes

According to Mendel's Law of Inheritance, since gametes form at random, all possible outcomes of Ω are likely to occur equally. Now since, there are four possible outcomes, the absolute sample space i.e. $|\Omega| = 4$ and the probability of each outcome would be 25 % or ¼.

It is important to note the difference between this example and the pea situation with which Mendel experimented. In Mendel's experiments with peas, there were 4 possible outcomes, but essentially only 3 genotypes. In this case there are, however, four genotypes with four possible outcomes.

Therefore,

P(RbRb) = ¼
P(RbY) = ¼
P(RbX) = ¼
P(XY) = ¼

a) What is the probability that a child not having the genetic disorder be born?

Since only one genotype would result in a perfectly normal child (son), it follows that,

$$P(\text{normal}) = P(X,Y) = ¼$$

Hence, the probability of a child born, developing a deletion and duplication of it later in age would be:

$$P(\text{disorder}) = \{(Rb,Rb), (Rb,Y), (Rb,X)\}$$
$$= ¾$$

The Mark Recapture Method and Probability

A good way of illustrating a probability example of greater complexity is to explain in the lights of the mark recapture technique. The mark recapture technique is often used by zoologists and environmental conservationists so as to estimate the size of a population and in certain cases; this technique is also used when the aim of the conservationist is to conserve genetic diversity.

I would now illustrate this method by using an example of introducing a new population of peppered moths (*Biston betulana*) into an existing population. Suppose M moths, where M is unknown, are present in the area of interest. In order to evaluate the number of M or how large the M population is, the moths are captured and marked/painted (O), by using paint, for example, that would not cause any harm to the moths, and then released back into the population, subsequently. Once released, the moths are allowed to interact and mix with those moths that had already pre-existed before their introduction. This essentially means that the sample size is larger than expected. *m* moths are thereafter captured. Suppose that *m* of the *n* moths are marked, while assuming that *m* ¿ *0*, and then released to mix with the population again, the ratio of marked to unmarked moths in the sample size, *n* should be approximately equal to the ratio of marked is to unmarked moths in the area i.e.,

$$\frac{m}{n} = \frac{O}{M}$$

Therefore the size of M is given by,

$$M = O_{n/m}$$

moths in the area.

a) What is the probability of finding m marked moths in a sample of size n?

There are M moths in the area, O of which are marked. Choosing *n* as the sample size, each outcome becomes a subset of *n*, which are equally likely.

$$|\Omega| = \binom{M}{n}$$

Let R denote the event that the sample of size *n* contains exactly *m* marked moths. Select *m* moths from M marked moths and, *n-m* moths from M-O unmarked one. Selecting the *m* marked moths can be done in $\binom{O}{m}$ ways.

Similarly, *n-m* unmarked moths may be selected in $\binom{O-M}{n-m}$ ways.

The multiplication principle can be used to find the total number of ways of obtaining a sample size of *n* with exactly *m* marked moths. This is because each choice of *m* marked moths can be combined with any choice of the *n-m* unmarked moths. Therefore,

$$|R| = \binom{O}{m}\binom{M-O}{n-m}$$

Thus, $P(R) = \frac{|R|}{|\Omega|}$

$$= \frac{\binom{O}{m}\binom{M-O}{n-m}}{\binom{M}{n}}$$

b) Why can the total number of moths in the area be estimated using the formula $M = O\,{}^{n}/_{m}$?

Obtaining a sample of size *n* and observing *m* marked moths in the area, it is possible to show that the value of M that maximises the probability of finding *m* marked moths in a sample size *n* is the largest integer less than or equal to $M = O\,{}^{n}/_{m}$. This is used as the estimate for the population size M. This is referred to as the maximum likelihood estimate, because its objective is to maximise the probability of what is observed.

From, $P(R) = \frac{|R|}{|\Omega|}$

$$= \frac{\binom{O}{m}\binom{M-O}{n-m}}{\binom{M}{n}}$$

Notes

Notes

Now, considering that P(R) is a function of M, it is possible to express this relationship as p_M. In order to evaluate the value of M that maximises P_M, the ratio of p_M / p_{M-1}. Hence, p_M cannot be differentiated to find its maximum since it is not continuous and can only be defined by integer values of M, the ratio is stated as follows:

$$\frac{p_M}{p_{M-1}} = \frac{\dfrac{\binom{O}{m}\binom{M-O}{n-m}}{\binom{M}{n}}}{\dfrac{\binom{O}{m}\binom{M-1-O}{n-m}}{\binom{M-1}{n}}}$$

$$\frac{p_M}{p_{M-1}} = \frac{\binom{M-O}{n-m}\binom{M-1}{n}}{\binom{M-1-O}{n-m}\binom{M}{n}}$$

$$= \frac{(M-O)!(n-m)!(M-1-O-n+m)!n!(M-n)!}{(n-m)!(M-O-n+m)!(M-1-O)!n!(M-1-n)!M}$$

Cancellation gives,

$$\frac{p_M}{p_{M-1}} = \frac{M-O}{M-O-n+m} \quad \frac{M-n}{M}$$

In order to find the values of M at which p_M exceeds p_{M-1}, it is important to obtain the ratio of p_M/p_{M-1} is greater than or equal to. This is because the local maxima are the values of M at which p_M exceeds both p_{M-1} and p_M.

When,

$(M-O)(M-n) \geq M\,(M-O-n+m)$, the ratio of p_M / p_{M-1} is greater than or equal to . Factorising this equation, we find that

$$M^2 - Mn - OM + Mn \geq M^2 - MO - Mn + Mm$$

Simplifying yields,

$$On \geq m\,M \text{ or/ } M \leq O\ {}^{n}/_{m}$$

As long as $M \leq O\ {}^{n}/_{m}$, $p_M \geq p_{M-1}$. If $M \leq O\ {}^{n}/_{m}$ is an integer, then $p_M = p_{M-1}$ for $M = O\ {}^{n}/_{m}$ and both $O\ {}^{n}/_{m}$ and $O\ {}^{n}/_{m-1}$ maximises the probability of observing m moths in the sample size n and so both values can be chosen as estimates for the number of moths in the area. If $O\ {}^{n}/_{m}$ is not an integer less than $O\ {}^{n}/_{m}$ maximises the probability, p_M. To arrive at just one value, we will always use the largest integer less than or equal to $O\ {}^{n}/_{m}$ to estimate the total number of moths in the area.

c) Assume that there are 32 marked moths in the area. We take a sample of size 20 and observe 10 marked moths. From this data and the information gathered from (b), it is possible to find an estimate of the number of moths in the area.

The estimate of the number of moths in the area is denoted by M, which is the largest integer less than or equal to $O\,{}^{n}/_{m}$, where $O = 32$, $n = 20$, $m = 10$,

Since $O\,{}^{n}/_{m} = 32 \times 20/10 = 64$,

It can be estimated that there are 64 moths in the area.

Notes

III. Potential Impacts of Statistics for Future Biologists

The merger of mathematics and life sciences is possible through employing techniques (such as statistics) that can be used to explain phenomena that cannot be seen, generally, through the naked eye. The marriage of these two sciences has enabled the evaluation of a number of facets. Some of these facets have been explained in the examples that are illustrated in this article, for example; it is possible to make an estimation of the number of organisms occupying a particular area, even when an ecological niche has been filled to capacity on introduction of new/different species of the same kind of organisms. Also it is able to estimate the outcome of offspring percentages of normal and defective origin that is unknown. Therefore, the idea of statistics can be exploited as an invaluable tool to estimate lineages that constitute ancestral and descendant traits in evolutionary context, as well. This bridges the gap between statistics and evolutionary biology.

One of the major advantages for understanding the workings of statistics, is that it would provide life scientists with the depth and knowledge to make suitable deductions by knowing the mathematics involved and the terminology associated with it. It would, more importantly, enable scientists to point out important concepts of biology nature through a mathematical medium.

Most scientists often associate biological concepts in isolation. As in the example of the male and female dog, probability deductions of biological nature can be deduced by simply looking at the answers of the F_2 generation. However, when one looks at mathematical probability, this would also be possible, but in this case the sample size would be expressed in the answer. A biologist, on the basis of looking at the total number of offspring and then distinguishing between traits to get ratios, would in most cases miss out 'sample size' in the final answer. Although mathematics and biology are at two different ends of the scientific spectrum, statistics could provide the language to bridge the link between the two sciences. By doing this, biologists, would be able to critically evaluate their findings and express them biomathematically.

Statistics is therefore an objective tool to both mathematicians and other scientists, its aims are to ultimately eliminate subjectivity and to focus on reliable results. Furthermore, this means that our daily operations are confined to mathematics because we are surrounded by biology, which depends on mathematical deductions.

The major concluding remark is that statistics should not be viewed as an effort in identifying unidentifiable problems. Instead, students should be objective in selecting career opportunities that explore a variety of avenues; and by understanding statistics, it would make this exploration an impressive endeavour!

References Références Referencias

1. W.M. Becker, L.J. Kleinsmith, J. Hardin, World of the Cell (fifth edition), San Francisco, Benjamin Cummings, Pearson Education, Inc (2003).

2. C. Neuhauser, Calculus for biology and medicine (second edition), New Jersey, Pearson Education, Inc (2004).
3. C. Starr, R Taggart, Biology: The unity and diversity of life (ninth edition), USA, Brooks/Cole, Thomson Learning, Inc (2001).

Notes

Notes

This page is intentionally left blank