Epistemic restrictions in population biology

CARLOS Y VALENZUELA

Departamento de Biología Celular y Genética, Facultad de
Medicina, Universidad de Chile, Santiago, Chile

Biologists have believed that the application of statistical or mathematical models to population biology has always been a correct and helpful tool to acquire knowledge. The present article demonstrates that the standard interpretation of statistical results yielded by the application of mathematical models to some populational processes, not only hides knowledge, but may lead to wrong knowledge. These epistemic restrictions are completely different from the known statistical restrictions (type I and II errors). A new more versatile conditional interpretation of statistical results is proposed.

Key terms: *epistemology, Hardy-Weinberg equilibrium, genetics, population biology, statistical restrictions*

INTRODUCTION

In August 1991, Science published (Cohen *et al.*, 1991) a controversy on forensic DNA tests and Hardy-Weinberg equilibrium (HWE). The main subject was the resolution of a test for an overall excess or dearth of heterozygotes. The authors and population biologists, in general, seem unaware of the great epistemic restrictions inherent to methods that measure deviation from HWE (Valenzuela, 1985). These restrictions occur not only in HWE studies, but in any similar statistical procedure. Moreover, most authors are not aware that these errors and restrictions are different from the type II statistical error. I shall show, with two examples, the loss of knowledge and some fallacies (misleading knowledge) of such statistical and mathematical procedures. Here, "epistemology" means the scientific study of the acquisition and validity of scientific knowledge. More specifically, this article deals with the consistency and validity of the knowledge acquired by the application of statistics and mathematics to biological processes.

METHODS

Two simple simulated examples have been devised to show that the application of mathematics or statistics to biology can produce both wrong knowledge or the lack of it. Both examples begin with a population with classes, where individuals from one class are removed. The analysis of the sensibility of a standard statistical test to show the direction and amount of the removal is performed. The first example deals with the simple mortality and birth rates in a subdivided population. The second one adds the Hardy-Weinberg model to the first example, maintaining the proportions of the classes, to show the new epistemic restrictions that a mathematical model can add by itself.

RESULTS

Example 1

Imagine a country with 1,000,000 inhabitants; 90,000 (9%) belonging to the high socioeconomic stratum (HSS) and 910,000

Correspondence to: Dr Carlos Y Valenzuela, Departamento de Biología Celular y Genética, Facultad de Medicina, Universidad de Chile, Independencia 1027, Casilla 70061, Santiago 7, Chile. Fax (56-2) 737-3158.

(91%) to the low socioeconomic stratum (LSS). Some HSS persons decide to kill secretly a random proportion of complete families from the LSS*. An international organization sends an observer (IOO) to study whether a demographic deviation from the expected proportions occurs in this country.

Changes can be produced only by an increase or a decrease in birth rate or mortality. The HSS persons allow the IOO to take a random sample of 400 individuals. The IOO decides to test an excess or dearth of LSS by a χ^2 test with a type I statistical error equal to 0.01 (χ^2 = 6.635). The sample size, the probability of type I error, the expected proportion, the distribution and the test have been fixed, in order to fix the power of the test. Thus, restrictions and errors we shall see are not related to the power of the test or the probability of type II error.

How many LSS individuals or what fraction of LSS must HSS kill to allow the IOO to see a significant result? The reader should give a figure, before knowing the correct answer. Since the maximal length of a 95% confidence interval for a proportion estimated from 400 individuals is close to 5%, a statistician would say around 50,000 individuals.

Table I presents the initial conditions and increasing numbers of LSS removals (the program works with fractional observed numbers). I assumed that the proportions in the sample are exactly those in the population after the removal of individuals (to hold the statistical power very near to 0.5). The withdrawal of 30% (273,000) LSS does not reach a significant χ^2. It is necessary a removal of 31.9281% to get a χ^2 equal to 6.635. The removal of 31.9281% of LSS is also a removal of 29.0546% of the total; but, the overall removal is barely 3.686% [(91-29.0546) / (100-29.0546) = 0.87314 = 87.314%; the statistician's expectancy]. The huge removal of 32% LSS can only be known as a fall from 91% to 87.314%. The IOO cannot know that an apparent fall of 3.7% implies really a fall of 29.1%. Moreover, this fall of 3.7% is a decrease of 3.686 / 91.000 = 4.05% of LSS, but it is an increase of 3.686 / 9.000 = 40.96% of HSS, that is, the less frequent class is much more distorted than the more frequent class.

The above analysis shows three types of restrictions: a) A big loss of knowledge; the IOO cannot know the true processes of killing LSS individuals and will only see huge cataclysms. b) If the HSS kills a proportion of the LSS sufficient to reach the significant 6.635, the IOO will see it in little more than half of the samples (type II error); because the observed number of LSS is a random variable, 51 or more yields a significant result and less than 51 implies a non-significant result (51 is the most probable result of this binomial distribution). c) The least frequent class is always the most affected one; the HSS always contributes more than the LSS to the χ^2 test; the IOO will be prone to interpret the results as an increase in HSS births (*i.e.*, wrong knowledge, misleading results).

Let K be the proportion of removed LSS, N the sample size, Y and Z the proportions of HSS and LSS, respectively; the expected numbers of HSS and LSS individuals are NY and NZ, and the observed numbers are NY/(1-KZ) and N(Z-KZ)/(1-KZ), respectively. With observed and expected numbers, the χ^2 value was calculated as:

$$\chi^2 = NZY\ [K / (1 - KZ)]^2\ (Z + Y)\quad \{A1\}.$$

TABLE I

REMOVALS OF INDIVIDUALS FROM THE LOW SOCIOECONOMIC STRATUM

VARIABLE	INITIAL	-20% LSS	-30% LSS	-32% LSS
Nº OF HSS	90,000	90,000	90,000	90,000
Nº OF LSS	910,000	728,000	637,000	618,000
TOTAL	1,000,000	818,000	727,000	708,800
PROP HSS	0.09000	0.11002	0.12380	0.12698
OBSERVED HSS	36	44	50	51
EXPECTED HSS	36	36	36	36
HSS χ^2 CONT	0.0000	1.7821	5.0764	6.0763
PROP LSS	0.91000	0.88998	0.87620	0.87302
OBSERVED LSS	364	356	350	349
EXPECTED LSS	364	364	364	364
LSS χ^2 CONT	0.0000	0.1763	0.5021	0.6010
TOTAL χ^2	0.0000	1.9584	5.5785	6.6772

PROP = proportion; **CONT** = contribution

* The author, referees of this paper and editor of this journal explicitly condemn the atrocity involved in the behavior here figured.

In the last parenthesis, Z and Y refer to the HSS and LSS contributions to the χ^2, respectively. {A1} demonstrates that the χ^2 value is directly proportional to N and the product ZY (since, Y = 1-Z, ZY = Z-Z^2), which tends to 0 as Z or Y tend to 0; and it is inversely proportional to (1-KZ). Thus, the χ^2 value tends to 0 as Z or Y tend to 0, no matter the sample size. {A1} also shows that the less frequent class (HSS = Y) yields the larger contribution to χ^2 (Z in parenthesis).

If K = 0, there is no epistemic restriction and we have the current statistical interpretation (deviations only from sampling). The lower the value of Y, the larger the epistemic restriction. The minimal restrictions will be produced when Y = Z = 0.5; but they will be still important (HSS could kill 114,000 before reaching a significant χ^2).

Either HSS killed or (exclusively) they did not kill LSS. Only one of these statements is true, but the IOO cannot know which is the correct situation. The same is true for an increase or decrease in births or migrations. The other formulas for an increase or decrease in HSS or LSS are:

χ^2 = NZY $[K(1+KY)]^2$ (Z + Y) {A2}; increase of HSS

χ^2 = NZY $[K(1+KZ)]^2$ (Z + Y) {A3}; increase of LSS

χ^2 = NZY $[K(1-KY)]^2$ (Z + Y) {A4}; decrease of HSS

Statistics cannot help the IOO to know the truth (either there is or there is not a deviation). However, this analysis opens a new interpretation of statistical tests. A χ^2 test of 1.9584 (1 d.f.) has been currently interpreted as a non-significant deviation (0.2 > P > 0.1). After the present analysis, it must be interpreted in two conditions: I) if there is no change in LSS or HSS, it is not a significant deviation (sampling deviation); II) if there is a change in the proportion of HSS or LSS (true + sampling deviations), this χ^2 may imply the following: i) a withdrawal of 19.28% (20% with integer numbers) of LSS; ii) an excess of 23.88% of HSS; iii) an excess of 29.52% of LSS; or iv) a removal of 22.79% of HSS. The equivalence between the χ^2 and the percentage of removal can be obtained by isolating K from equations {A1}, {A2}, {A3}, {A4}. Naturally, a mixed change can be produced (its analysis is beyond the scope of this article).

To decide among the four situations, we look at the relationships between expected and observed values. There are only two possibilities: i) relative excess of HSS or lack of LSS; ii) relative lack of HSS or excess of LSS. The fixed "all or none" statistical interpretation has been changed to a "versatile" conditional one. Every χ^2 value is significant (shows the amount of the conditional change) and meaningful (shows the relative direction of the change). However, neither the current statistical procedure nor the present one can full access reality, but they can partly access it; the difference is a matter of interpretational richness.

Example 2

Turning to Hardy-Weinberg equilibrium (HWE), let AA, AB and BB be the genotypes for a gene locus with two alleles (A and B) in a population of 1,000,000 inhabitants; and D, 2H and R, the respective observed genotypic frequencies in a sample of size N. Let us assume that R = D + W; that the gene frequency of A is P = 0.3, and that of B is Q = 0.7; and that the population is in HWE. Then, the initial genotypic frequencies are: D = 0.09; 2H = 0.42; R = 0.49. We have equated AA to HSS and AB + BB to LSS, to see the distortion that an algorithm (HWE) can add to the already present epistemic restrictions.

The procedure to test HWE estimates first the gene frequencies from the sample (P = D + H; Q = R + H). The expected genotypic frequencies are found by squaring the polynomial of the gene frequencies $[(P + Q)^2]$. The χ^2 test is:

$$\chi^2 = 4N\,(DR\text{-}H^2)^2\,[1/(1\text{-}W)^2 + 1/(1\text{-}W)(1+W) + 1/(1+W)^2] \quad \{A5\}$$

as shown in Valenzuela (1985). However, D, 2H and R are the observed genotypic frequencies; thus, they include the previous possible deviations. If S, T and V are the respective proportions of AA, AB and BB removals, and we denote by D_0 and D_1, $2H_0$ and $2H_1$, R_0 and R_1 the respective genotypic frequencies (now in the population) before and after the removal, the relation between both sets of genotypic frequencies are:

$$D_1 = (D_0\text{-}SD_0)/(1\text{-}SD_0\text{-}2TH_0\text{-} VR_0);$$
$$2H_1 = 2\,(H_0\text{-}TH_0)\,/\,(1\text{-}SD_0\text{-}2TH_0\text{-}VR_0);$$
$$R_1 = (R_0\text{-}VR_0)/(1\text{-}SD_0\text{-}2TH_0\text{-}VR_0) \quad \{A6\}$$

These transformations ({A6}) must be inserted into {A5} if we want a complete picture of restrictions. As in the first example, the AA individuals begin to kill AB and BB individuals. The IOO has an additional restriction to estimate the deviations, the one due to the circular process of calculating the expected values with estimates of gene frequencies which include the deviations, because the IOO does not know the original gene frequencies [other restrictions and the contribution of genotypes to the χ^2 test were dealt with in Valenzuela (1985)].

Table II shows removals of AB + BB. Since the epistemic power (not the statistical power) is reduced in relation to the first example, we begin with the removal of 32% of AB and BB. In fact AA can kill 42% of BB + AB and the IOO will not find a significant deviation. Given a χ^2 value, we can determine its significance and meaning as in the first example, by using {A5} and {A6}.

TABLE II

REMOVALS OF AB AND BB GENOTYPES

VARIABLE	INITIAL	-32% (AB+BB)	-42% (AB+BB)	-43% (AB+BB)
Nº AA	90,000	90,000	90,000	90,000
Nº AB	420,000	285,600	243,600	239,400
Nº BB	490,000	333,200	284,200	279,300
TOTAL	1,000,000	708,800	617,800	608,700
P	0.300000	0.328443	0.342829	0.344505
PROP AA	0.090000	0.126975	0.145678	0.147856
OBSERVED AA	36	51	58	59
EXPECTED AA	36	43.15	47.01	47.47
AA χ^2 CONT	0.0000	1.3528	2.6961	2.8683
PROP AB	0.420000	0.402935	0.394302	0.393297
OBSERVED AB	168	161	158	157
EXPECTED AB	168	176.45	180.24	180.66
AB χ^2 CONT	0.0000	1.3233	2.8130	3.0149
PROP BB	0.490000	0.470090	0.460020	0.458847
OBSERVED BB	196	188	184	184
EXPECTED BB	196	180.40	172.75	171.87
BB χ^2 CONT	0.0000	0.3236	0.7337	0.7923
TOTAL χ^2	0.0000	2.9997	6.2429	6.6755

P = frequency of A; PROP = proportion; CONT = contribution

DISCUSSION

The Hardy-Weinberg equilibrium is an algebraic structure useful only as a first approach to the genetics of populations. No population can fit it. Naturally, it is not a law of nature. The HWE requires no mutation (this case is never seen), no drift or an infinite population (impossible), no selection, random mating and no migration (very difficult).

Besides the above conditions, the HWE includes an insuperable logical and factual contradiction: random mating cannot occur in an infinite population; it is thermodynamically impossible. An infinite population implies infinite distances among individuals, and an infinite amount of energy, so far as everyone can meet everyone with the same probability. The argument that populations may be "so close to HWE as to be dealt with as if in HWE" may be misleading. Biological populations are not in HWE, but we are unable to detect deviations with our procedures. In the example, if AA kills 10% of AB and BB, IOO will see the gene B disappear in 50 generations, in a perfect HWE; IOO will conclude that this is an example of gene fixation (the A allele) by drift.

At least, seven different sources of restrictions for knowing or errors in getting knowledge can be seen from the present analysis:

1) The impossibility to know the original process or population (ontic restriction). If we can only know a population which includes a deviation from the original one, our knowledge of the amount of deviation is always less than the original one. In Example 1, if we remove 20% of the 91% of LSS, the IOO will see a removal of 2.25% instead. We study evolution from what we can perceive. A real theory of evolution should be constructed from all what has been produced or, better, from all what could be produced. The study of this kind of restrictions can show a conditional access to the universe that we cannot perceive. Formulas {A1} to {A6} include deviations to calculate the χ^2 values; they show that it is not possible to find a test completely independent of deviations.

2) The wrong knowledge due to the change in the direction of deviations that mathematical algorithms produce. In the latter example, a removal of 20% of LSS is barely seen as a removal of 2.25/0.91 = 2.47% of the 91% of LSS. However, it will be seen as the removal of 2.25/0.09 = 25% of the 9% of HSS. The simple algorithm used to fit the new total leads the researcher to be prone to interpret the removal of LSS as an addition of HSS instead. Knowing the distortion that an algorithm produces, we can create some populational monsters. Let us imagine a sample of 10,000 individuals, typed for the ABO system, and taken from a population where p (frequency of gene A) = 0.2, q (B) = 0.1, and r (O) = 0.7. The most probable sample in HWE will be: 400 AB, 3200 A (400 AA + 2800 AO), 1500 B (100 BB + 1400 BO) and 4900 O (OO). Now, we create in the population a huge heterozygote AB advantage; we add 10% AB and we remove 30% of homozygotes AA, BB and OO. The most probable sample will be: 522 AB, 3658 A, 1746 B and 4074 O. The HWE analysis, performed with the maximum likelihood method shows: p = 0.2376; q = 0.1209 and r = 0.6415. The expected numbers for phenotypes are: 574.6 AB; 3612.5 A; 1697.6 B and 4115.3 O. The χ^2 with 1 d.f. is 7.17 ($P < 0.01$); however, only the lack of AB is lonely significant ($\chi^2 = 5.1$, $P < 0.025$). So, a huge heterozygote AB advantage is regarded as a huge heterozygote AB disadvantage! (see also Valenzuela and Cifuentes, 1989). The knowledge of the distortions that are produced by algorithms will allow us to better interpret the true processes which occur.

3) The circular procedure which is created when using estimates that include deviations, to evaluate such deviations. Example 1 cannot have this restriction, because the proportion of LSS is given without error (0.91). In HWE studies we estimate gene frequencies from the sample taken from the population, which includes deviations. The mostly used procedure is the maximum likelihood (ML) method. It estimates the gene frequencies that fit best the HWE distribution, because it assumes that the sample was taken from a population in HWE and the likelihood function is constructed according to this assumption. So, if the population is not in HWE, the ML searches for those gene frequencies that produce the least deviation from HWE, in a sample from that population. In example 2, the removal of 42% of AB+BB changes the (estimate of) gene frequency of B from 0.7 to 0.657. However, the observer cannot know this frequency change, and uses 0.657 to evaluate the deviation of the sample from HWE. In the case of ABO, dominance adds another source of errors and restrictions.

4) Degrees of freedom. The most probable biological case is that in which each of the **n** classes of a population has its own coefficient of fitness; that is, **n** different fitness or selection coefficients (S for AA, T for AB, V for BB in Example 2). Besides that, in the simplest problem, we need to estimate one gene frequency or distribution parameter. The number of degrees of freedom is the number of classes minus the number of parameter estimates. So, we have at most –1 degree of freedom (n classes - n fitness coefficients –1 frequency), for every evolutionary problem where classes are involved. We can get valid results only when several of the fitness coefficients are so similar that our method treats them as being equal.

5) Restrictions coming from the logic of the scientific method. When the "modus (ponendo) ponens" is used in factual science, a positive result does not mean that the hypothesis is true. If the proposition X implies the proposition Y, to find Y does not mean that X is true. If a model (based on an hypothetical explanation of the deviation) implies a mathematical relationship of data, to find this relationship in the data does not mean that the model is valid. If a population is in HWE, the genotypic frequencies must be the square of the polynomial of gene frequencies. If we find a population where the genotypic frequencies are equal to the square of gene frequencies, this does not mean that the population is in HWE. An example of an infinite set of non HWE populations that simulate HWE populations was presented by Li (1988).

6) Non-applicability of mathematical or statistical assumptions to biological processes. Continuity, linearity, homosedasticity, gaussianity, equal weight of variables are some assumptions that cannot be applied to

living beings, without error. Biological processes include genes and individuals which are discrete unities related to one another and to the environment by non linear relationships. Populational or gene frequencies move in rational numbers, while maximum likelihood estimates need also irrational numbers.

7) Statistical errors. The previous errors are deterministic errors or restrictions. They are not due to random variation in sampling, but to "mathematization" or "statistization" of biological processes. A sample from a non deviated population may show a significant deviation by the simple random variation in sampling; this is the type I statistical error. On the other hand, a deviated population may be regarded as not deviated, because a sample from it resulted not to be deviated by the random process of sampling; this is the type II statistical error. These stochastic errors must be added to the former ones to have a view of the general picture of epistemological restrictions.

The main problem lies on the validity of the application of mathematical or statistical models or thoughts to natural processes or biological conceptions. We do not have a transdisciplinary metatheorem which can prove that validity. It seems that the richness of the interaction between the researcher and nature cannot be reduced to mathematics without error. Mathematics and statistics should be regarded as useful tools to put our scientific experience into an operational and conventional frame, rather than intellectual tools to explain biotic processes. That is, these disciplines are operational languages that can translate biological ideas into their languages. However, a translation may be a betrayal. As teachers, we know that our students learn much more from their field or laboratory experiences, or seeing us in the arena of research, than from textbooks or from our speeches.

In practice, since in most cases we cannot measure deviations, I agree with those geneticists that use empirical probabilities (which include all deviations) to calculate paternity or identity assignment (for VNTRs). To test deviations by comparing samples, as proposed by Lewontin and Hartl (1991), could increase the errors or restrictions. Those procedures have their own restrictions. If the IOO in example 1 tests a difference in the birth rate or mortality between HSS and LSS, a non significant difference should be the most probable result. Thus, a significant χ^2 value should be dismissed, because the assumed more precise tests of samples comparisons give non significant results.

ACKNOWLEDGEMENTS

I am grateful to Prof Catherine Connelly for reviewing the manuscript. Work supported from FONDECYT grants 91-1118 and 92-921. This subject was presented to the 5th International Congress of Biomathematics (Santiago, 14-19 October, 1991).

REFERENCES

COHEN JE, LYNCH M, TAYLOR CE, GREEN P, LANDER E, DEVLIN B, RISCH N, ROEDER K (1991) Forensic DNA tests and Hardy-Weinberg equilibrium. Science 253: 1037-1041

LEWONTIN RC, HARTL D (1991) Population genetics in forensic DNA typing. Science 254: 1745-1750

LI CC (1988) Pseudo-random mating populations. In celebration o the 80th anniversary of the Hardy-Weinberg law. Genetics 119: 731-737

VALENZUELA CY (1985) Algebraic and epistemological restrictions in studies on Hardy-Weinberg equilibrium. Am Nat 125: 744-746

VALENZUELA CY, CIFUENTES L (1989) Interpretation of ABO (genetic) segregation distortions. Brazil J Genet 12: 659-663

ADDENDUM

When this article was in the last phase of the printing process, the author received information that a previous version, rejected in first instance (letter of June 2, 1993), had been published in *Evolución Biológica*, volume 7, pages 71-80, August 1993 (although that date does not correspond to the more recent date of circulation of the issue). The author was not opportunely informed of the later decision of the Editor of that journal to publish the article.

The first version of this Ms was received by the Editorial Office of **Biological Research** on November 3, 1993, being reviewed by three different referees, and returned to the author on January 15, 1994, for resubmission after considering the referees' comments. The second version of this article was accepted on April 26, 1994.