STANDARD STATISTICAL CONCEPTS: CAN THEY PRODUCE INCOHERENCE?

Carlos Alberto de Bragança Pereira
University of São Paulo, Brazil
cpereira@ime.usp.br

The present article concerns statistical concepts that are usually presented in the statistical classroom. Examples are presented in a way that simple applications of these concepts produce incoherent conclusions. The examples illustrate that: iid random variables are in fact strongly dependent; conditional probabilities may depend on how the conditioning arguments were learned; confidence intervals may have the property of diminished precision when information is increasing; and significance tests may not reject impossible hypotheses.

INTRODUCTION

Learning of basic disciplines such as mathematics or physics has become a hurdle in the education of professionals who are to deal with scientific and technological challenges of modern society. Students' apathy for technical disciplines, which demand a more logical and abstract reasoning, has become universal. On the other hand, modern society has been modifying its values at the same pace at which new technologies are incorporated to our everyday life.

Curiously enough, the superb power of adaptation of young people to new technology is not matched by any interest in how such technology is developed nor in the concepts used for its creation. Young people rapidly become excellent users of new technology, as opposed to adults who were educated under a totally different cultural paradigm. Today's adult questioned the necessity of the new available technology when he was young. It was important to know how and why technological apparatus were built and used. At present, it seems that there is no time to be "wasted" on such conceptual questioning. This may produce the (possibly mistaken) idea that young people are only interested in knowledge of immediately application and, in many instances, of easy learning.

How a teacher should convince a student that analysis, and not only calculus, is important for the statistician's logical thinking? In the past, it was not common to a student to question a teacher about the usefulness of a subject being taught. A teacher used to be a Master! Today a teaching professional is being evaluated by her/his productivity, that it is sometimes judged by the number of approved students. The Master used to look for a strong student background. Today teachers may only look for the minimum learning that a student requires to advance to a next college step.

The demand for statisticians has also changed and the market is now looking for a professional with an MSc degree. A regular graduate diploma is not enough anymore! As expected, one looks for a statistician with strong knowledge. Maybe in the near future a PhD diploma will be necessary for the new professional in statistics. These additional degrees used to be needed only for scientific and academic rather than professional purposes.

Instead of criticizing students the objective of this paper is to question concepts that are usually introduced in statistical classes. The author's opinion is: the fact that students are not interested in foundations is a consequence of the imprecise way in which important statistical concepts are introduced. These concepts are discussed using examples throughout the text. In the second section we discuss probabilistic and statistical independence. The third section is devoted to conditional independence and we call attention to how conditioning arguments should be learned. The fourth section is about incoherent practical solutions when using confidence intervals. Finally we show how p-values provide misleading responses.

PROBABILISTIC AND STATISTICAL INDEPENDENCE

In statistics, the concept of dependence is closely related to the concept of association. It is common sense that two phenomena are associated if one influences the other. For example, the occurrence of one will increase the chance of the other occurring; this is the case of positive association. There are also cases of negative association in which the occurrence of one event decreases the chance of the other occurring. It is important to recall that the equivalent concepts

of association or dependence of two random quantities, x and y, have a symmetric property; that is, if x depends on y then, y depends on x at the same level.

The lack of dependence is known as *independence*. There are no distinct levels of independence! Two random quantities are independent whenever the observation of one does not alter the probability distribution of the other. Equivalently, x and y are independent random quantities if $p(y|x) = p(y)$ for every possible value of $(x;y)$. That is, the conditional probability of y given x is equal to the marginal probability of y, for every possible value of $(x;y)$. A good homework for the reader is to prove the equivalence of the following three equalities:

1. $p(y) = p(y|x)$,
2. $p(x) = p(x|y)$, and
3. $p(x)p(y) = p(x;y)$.

This concept of independence is naturally extended to *conditional independence*, whenever another random quantity, z, is considered. The random quantities x and y are conditionally independent given z if, for every possible value of $(x;y;z)$, $p(x,y|z) = p(x!z)p(y|z)$. That is, if z is to be known, x and y are independent. Again, the reader should check the equivalence among the next tree equalities:

4. $pIy|z) = p(y|x,z)$,
5. $p(x|z) = p(x|y,z)$, and
6. $p(x|z)p(y|z) = p(x,y|z)$.

In statistics one may say that x and y are iid with common distribution p that depend on an unknown parameter, say z. Hence, the *statistical independence* between x and y should be understood as the conditional independence with the parameter playing the role of z in the above discussion. In the next three examples the differences between probabilistic and statistical independence are illustrated.

Example 1 (Standard Normal Distribution): Let x and y be independent random quantities with standard normal common distribution, N(0,1). It is most probable that neither variable will assume values outside of the interval [-6;6], although we know they may assume any value in the whole real line. Knowing, for example, that $x = 3$ will not change at all our expectation about the value y assumes in [-6,6]. This is an illustration of probabilistic independence.

Example 2 (Normal Distribution with Unknown Mean): Let x and y be independent and identically distributed random quantities with normal distribution, N(z,1), where z is the unknown parameter. Now, there are no intervals with higher expectations of occurrence than others of the same length. Both, x and y, can assume values in the real line without any preference. Now, the observation $x = 3$ is should carry a lot of *information* about y. Recall that, before the observation of x, the only expectation one has is that y is any real number. The other important fact is that the value of x is most probable to be in the interval [z-6;z+6]. Since $x = 3$, one could say that, with high probability, $3 < z+6$ and $3 > z-6$: i. e., $-3 < z < 9$. Consequently, after observing $x = 3$, it is improbable that y lies outside of the interval [-9;15]. In the beginning, y could be any real number but after observing that $x = 3$, one could say that with high probability $-9 \leq y \leq 15$. Hence, x and y are statistically independent but strongly dependent!

To really pique the reader a more extreme example is presented next.

Example 3 (Uniform with Unknown Mean): Let x and y be independent and identically distributed random quantities with uniform distribution in the interval [z-1;z+1]. Again, z is the unknown parameter belonging to the real line. As before, x and y can assume any value in the real line. In particular, assume $x = 3$ is the actual observation. For this case, with probability one, y belongs to the finite interval [1;5]. This is a consequence of $z-1 < 3 < z+1$ and $z-1 < y < z+1$. Again, x and y are statistically independent but strongly dependent!

The last two examples show that statistical independence involves in fact strong dependence in most cases. Whenever a probabilistic structure is given to the parameter, one can identify statistical independence with conditional independence. For the case of identically distributed random quantities, the common parameter is the element that causes the dependence

among the quantities. Instead of statistical independence one could say, x and y are independent if z is to be known.

CONDITIONAL INDEPENDENCE AND INFORMATION

The example discussed in this section is equally important to emphasize the differences between probabilistic and statistical work. However, the reader must understand that one complements, rather than replaces, the other.

The example highlights the role of likelihood functions. It also tries to show the difference between sampling distribution and likelihood function. Both are obtained from the statistical model, which is a function of two quantities, $f(x|z)$, x assuming values in the sample space and z in the parameter space. The sampling model is the set of probability functions defined in the sample space for each possible value of the parameter. The likelihood function is the probability of the actual observation of x calculated to every value of the parameter z. For the uniform example 3, the likelihood for $x = 3$ is $L(z|3) = f(3|z) = ½$ for every z in the interval of possibilities for z, [2;4].

Example 4 (Pair of siblings): The neighborhood dance school participates yearly in a national competition with its group of 10 girls. Lily, the teacher, has learnt that the family of the two sisters that belong to the group will emigrate soon to another country. She heard that a new family, with two kids is moving to the apartment of the two ballerinas. Lily thinks that she could have another two dancers if both new kids are females. She could train them to join the group. Lily assigned a probability of ¼ to the event of two new girls in the neighborhood. For her, the sample space was {(m;m),(m;f),(f;m),(f;f)}, m and f representing male and female. Clearly, for her, every sample point had the same probability ¼. When talking to Jony, her brother who handles the rent of the apartment to the new family, she learned that at least one of the kids is female. In fact he was at the telephone talking to the kids' mother when she shouted to someone saying "be quiet girl" saying that she was talking to one of her kids. Lily then was happy since her space now become a set of e equally probable sample points, {(m;f),(f;m),(f;f)}; i.e., she now considers that her probability of (f;f) is 1/3. As soon Lily reached this conclusion Jony's wife, Mary, enters the office and said that a daughter of the new family was downstairs in her car. Lily then runs to the car and was happier since her probability now became ½. This conclusion arose from the fact that she looked for the gender of the other kid and the sample space became {m,f}, a set of two equally probable points. To accept Lily´s analysis one should agree that the event "at least one sibling is a female" produces different probabilities depending on whether it is learned by different channels: auditory or visual.

The statistician's eyes allow the incorporation of other kinds of information in the learning process. Let us consider all the nuances of the process. The parameter of interest takes the values $z = 1$ if the state of nature is (f;f) and $z = 0$ otherwise. Considering the three point sample space, let the prior probability be $p(z=1) = 1/3$. As the sample observation, let $x = 0$ if Mary had brought a male kid and $x = 1$ a female. The likelihood function *is* $L(z|1) = p(1|z)$ that takes the values $L(1|1) = 1$ and $L(0|1) = q \in [0;1]$. The reader should confirm that the posterior probability of interest is $p(z=1|x=1) = (1+2q)^{-1}$. Lily was correct when considering that $q = ½$ producing ½ as the posterior probability. However, after listening to the Lily's arguments, Mary reported that the mother had asked her to bring that girl to try out a skirt in the female clothing store next door. In that case she should consider that $q = 1$ and then the posterior would be 1/3 just like the prior.

The reader must understand that a statistician has to consider a model for the data generator; the likelihood. A probability expert usually looks at the future, considering the present. The statistician, being interested in understanding the present, looks at the past to ascertain how the information was obtained.

CONFIDENCE AND CREDIBILITY

This section is about interval estimation. The most controversial statistical concept, for the author, is the confidence interval. Most of the time, the statistician has to explain to scientists that the confidence of an interval is not the probability that the true value of the parameter

belongs to the interval. Confidence of an interval is the frequency of intervals that contain the true value of the parameter whenever repetitions of the experiment are performed. However, most of the experiments cannot be repeated and for some scientists confidence factors do not make much sense. To better understand these arguments the following example illustrates the difference between confidence and *credibility*. By interval *credibility* we understand the posterior probability that the interval contains the parameter value.

Example 5 (Uniform with unknown mean): Let (*u;v;x;y*) be a vector of four independent and identically distributed random quantities with uniform distribution in the interval [*z*-½;*z*+½]. Again, it is a case of statistical independence where *z* is the unknown parameter belonging to the real line. The interest now is to obtain a confidence interval for *z*. Let *m* and *M* represent the minimum and the maximum of the sample*: m = min*(*u;v;x;y*) and *M = max*(*u;v;x;y*). It is not difficult to prove that the interval [*m;M*] is an interval with 87.5% of confidence: i.e, using the sample distribution, $p([m;M] \supset z \mid z) = .875$. In fact, $(½)^4$ is the probability that all four observations are smaller (or larger) than *z*. Hence, $(½)^3 = .125$ is the chance that all four observations lie in the same half side of the interval and finally the chance of $\{[m;M] \supset z\}$ for any fixed *z* should be $1- (½)^3 = .875$. After observing a sample, it is not unusual for a statistician to present to his client the observed interval attached to the confidence level, 87.5% in this case. For example, suppose that the observed sample vector was (1.11;1.27;1.43;1.59). In this case, a careless statistician would present to his client the interval [1.11;1.59] and the number 87.5% to indicate his high confidence.

To discuss the method used in the above example, let us change the value of *M* and evaluate what could be given to the client. Let us consider three other possibilities for the maximum, say 1.50, 1.62, and 1.91. Let us now consider the four 87.5% confidence intervals with their lengths, *L*'s: I_1 = [1.11;1.59], L_1=.48*;* I_2 = [1.11;1.50], L_2=.39; I_3 = [1.11;1.62], L_3=.51; and I_4 = [1.11;1.91], L_4=.80. A smart statistician who believes that there is more thinking involved than the simple use of a method, would understand that $m > z$-½ and $M < z$+½ or, equivalently, for sure M-½ < z < m+½. This statistician could present to his client the following 100% confidence intervals together with their lengths: S_1 = [1.08;1.61], l_1=.53; S_2 = [1.00;1.61], l_2=.61; S_3 = [1.12;1.61], l_3=.49; and S_4 = [1.41;1.61] ,l_4=.20. Note that the smaller the interval is, the more precise an answer is given to the client. Comparing the answers of the two statisticians, the second presented two more precise intervals, besides giving 100% of confidence. The important fact to be noticed is that the larger $M - m$ is, the more informative is the sample. Hence the first statistician's answer has less precision whenever the sample is more informative. Finally, suppose that a probabilistic structure is given to *z*. For example, let *z* have a uniform prior distribution in the interval [-100;100]. Consequently, for the four samples, *z* would have uniform posterior distributions in the intervals S_1, S_2, S_3, and S_4. With these four posterior distributions one obtains the following four 87.5% credible intervals together with their lengths: C_1 = [1.12;1.58], l_1=.46; C_2 = [1.04;1.57], l_2=.53; C_3 = [1.15;1.58], l_3=.43; and C_4 = [1.42;1.60], l_4=.18. With this Bayesian method, the less (more) informative sample, the one with M = 1.50 (1.91), produces a less (more) precise interval. The classical method, used by most statisticians, went in the wrong direction. Hence, this method produces incoherence! Table 1 summarizes this discussion.

Table 1*:* Estimation intervals: 87.5% confidence, exact and 87.5% credible

	Intervals			Length		
sample	confidence	sure	credible	confidence	sure	credible
1	[1.11;1.59]	[1.08;1.61]	[1.12;1.58]	0.48	0.53	0.46
2	[1.11;1.50]	[1.00;1.61]	[1.04;1.57]	0.39	0.61	0.53
3	[1.11;1.62]	[1.12;1.61]	[1.51;1.58]	0.51	0.49	0.43
4	[1.11;1.91]	[1.41;1.61]	[1.42;1.60]	0.80	0.20	0.18

The above discussion is a more elaborated version of the one presented in page 400 of DeGroot (1986). See Basu (1988) for a discussion on information provided by an observed sample.

SIGNIFICANCE AND EVIDENCE

This section is about significance tests or p-values. The very simple example discussed here was originally presented by Pereira and Wechsler (1993). It is not uncommon that the definition of a p-value presented in classroom disregards the alternative hypothesis. There are cases were the teacher gives the idea that p-values are the tails of the null distribution, the sample distribution under the consideration that the null hypothesis holds. In the next example the central area, not tails, should be the p-value. In Pereira and Wechsler, another example shows that a p-value could be a composition of tail and central areas.

Example 6 (Urn and Balls): There is an urn with three balls of different colors; white, black and red. A sample of three balls is selected from the urn and one records the values of x and y; x being the sampling number of white balls and y of black balls. Both, x and y, can take values in the set {0;1;2;3}. With the observed sample, $(x;y) = (1;1)$, one wants to test the null hypothesis *"H: the sample was taken with replacement,"* against the alternative hypothesis *"A: the sample was taken without replacement."* Table 2 illustrates the null probability distribution multiplied by 27.

Table 2: Null Probabilities

x \ *y*	*0*	*1*	*2*	*3*
0	1	3	3	1
1	3	6	3	0
2	3	3	0	0
3	1	0	0	0

If the definition of p-value includes the observed sample, its value is 100% of significance in favor of H. If it does not include the observed sample then its value would be 77.78% corresponding to the probability value 21/27. The conclusion is that the sample evidence favors the null hypothesis, H, against A. Also it is interesting to note that the observation (1;1) is the sample point that most favors H. Again incoherence is present! The reason is that $(x;y) = (1;1)$ is the only sample point that can be observed under A. All other sample points are impossible under the alternative hypothesis A. The statistic $t = (x\text{-}1)^2+(y\text{-}1)^2$, known as the χ^2 statistic, could work as a classical test statistic.

Any statistician should look at the likelihood function! In this case the likelihood is $L(H|1;1) = p(x=1,y=1|H) = 6/27$ and $L(A|1;1) = p(x=1,y=1|A) = 1$. The likelihood ratio here is given by $LR(1;1) = L(A|1;1)/L(H|1;1) = 27/6 = 4.5$, favoring A against H. If a priori one considers *a priori* equal probabilities for the hypotheses, $p(H) = p(A) = ½$, then, *a posteriori*, the null hypothesis will have probability equal to $p(H|1;1) = (1+27/6)^{-1} = 2/11$. Clearly this favors A rather than H.

Consider in the above example an urn with 4 balls of different colors: black, white, red and blue. A sample of four balls is selected from this urn. Let x, y and z be, respectively, the sampling numbers of white, black and red balls. The sampling null distribution is the multinomial distribution $p(x,y,z|H) = [(32/3)(x!y!z!)]^{-1}$. The most probable sample point for this distribution is $(x;y;z) = (1;1;1)$. Again, applying the standard definition for the p-value, as before, we obtain a significance of 100% for (1;1;1), strongly favoring H against A. The likelihood ratio, in this case is $LR(1;1;1) = 32/3 = 10.67$ and the posterior probability of H for a flat prior would be 3/35, strongly favoring A. Application of the standard concept again produces incoherence!

FINAL REMARKS

This paper is a consequence of the many encouragements the author has received from his colleagues and friends, mostly classical statisticians. They knew that the author's former supervisor, Dev Basu, was the master of counter examples and had many public debates with Oscar Kempthorne who also liked the adversarial environment. The idea was to try to have rich conflicts also in our community. Whenever public debates about statistical concepts are present, Students would benefit from the time they would need to choose a side in the discussions. This is the moment when a student feels that he is not being trained, but instead challenged, to thinking and to having his own judgments. To better understand this discussion, readers are recommended to complement these ideas in Basu (1988), Good (1983), and Kempthorne and Folks (1971). For a Bayesian training the reader should also look for Barlow (1998), Blackwell (1969) and mainly de Finetti (1972).

In the classroom, the author tries to make the students choose a way among alternatives. He is against the training to use methodologies without discussing how and on which bases they were developed. For the author the lemma to be followed is: Nothing is just good!

REFERENCES

Barlow, R. (1998). *Engineering Reliability*. New York: SIAM.

Basu, D. (1988). *Statistical Information and Likelihood.* In J. K Gosh (Ed.), *A Collection Of Critical Essays by D. Basu*, Lecture Notes in Statistics #45, Berlin: Springer-Verlag

Blackwell, D. (1969). *Basic Statistics.* New York: McGraw-Hill.

de Finetti, B. (1972). *Probability, Induction, and Statistics.* New York: John Wiley.

DeGroot, M. (1986). *Probability and Statistics*. New York: Adison-Wesley.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Minneapolis: University of Minnesota Press.

Kempthorne, O. and Folks, F. (1971). *Probability, Statistics and Data Analysis*. Ames: The Iowa State University Press.

Pereira, C. A. de B. and Wechsler, S. (1993). On the concept of p-value. *Brazilian Journal of Probability and Statistics*, 7, 159-77.