THE VERY BEGINNING OF A CLASS ON INFERENCE: CLASSICAL VS BAYESIAN

Lisbeth Kaiserlian Cordani
University of São Paulo (*retired*), Brazil
lisbeth@ime.usp.br

Although the original Bayesian theory was settled in the 18th century, due to various previous computational difficulties, only in the last 20 years has the Bayesian method grown substantially. This may explain why only the classical approach has been offered at the educational level. In our view, it is important to present both approaches to undergraduate students, enlarging their vision not only about statistical tools but also about the inherent philosophy of those schools. Both approaches offer tools to solve practical problems. The students may have a quite different background and come from different courses, and can be encouraged to select, in the future, the best method for their purposes. Our suggestion is to present the main ideas of inference, starting with some issues about conditional logic and its influence on the inferential conclusions, comparing both classical and Bayesian approaches. Examples will be presented.

INTRODUCTION

In several areas of science, statistics is a powerful tool to analyze data both from controlled experiments (mainly in the natural sciences) and from observational studies (mainly in the human sciences). Such empirical observations are used to reinforce scientific theories, evaluated through hypotheses.

Even being aware that there is no definitive way to prove theories, such reinforcement becomes stronger if we can find some evidence to go from a particular view to a general one, through inferential procedures. So researchers expect to find methods which can provide means to judge a population from a subset of it, a sample. Over several decades, statistics has developed many different theories to be applied in different situations, and all of them have a characteristic in common: given the uncertainty, they try to find the best strategy to answer scientists' queries.

Almost every course at university includes a basic unit called Introduction to Statistics and the great evasion rate it produces among students is well known. Part of the content is dedicated to Statistical Inference usually presented in a very instrumental way, with no remarks about epistemological issues or discussions about different approaches.

The objective of this work is two-fold: a) to encourage the learning of some aspects of conditional logic, which supports the classical theory of statistical inference; b) to introduce Bayesian reasoning in a very simple way, in order to show how to deal with probability assignment in inferential problems. As pointed out by Barnett (1973), it is very important to expose the students to different philosophical and conceptual approaches.

IDEAS OF INFERENCE

The theory of logic describes formally what an argument must present to be valid. Inference is a process through which we arrive at a proposition, based on one or more propositions accepted at the starting point of the process (Copi, 1978).

Reasoning can be considered an inferential process in which new knowledge is generated from evidence of a well-established knowledge. From a statistical point of view, we could say that making an inference is to generate knowledge for the population from observing a sample, and such a process is mediated by probability.

In this process we need to highlight two types of reasoning: deductive and inductive. The first one, deductive, is called logical reasoning by excellence, because from true premises we arrive at a true conclusion–in this case, the conclusion is contained in the premises. The second one, inductive reasoning, gives a more general conclusion not contained in the premises. Then, we can get to a false conclusion even if the premises are true. So, deductive inference is always conclusive and inductive inference is always inconclusive. Of course we are talking about incomplete induction, where the trivial case of enumeration (simple induction) and mathematical induction are not included.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

Many philosophers say that only deductive inference is valid because of its inherent logical justification. Hume (XVIII century) seems to be the first to treat the induction problem in a systematic way, although his main interest was centered in causal inference, without reference to probabilities. He argued that induction, even without logical or rational evidence, could be useful to humanity. Popper (1975) rejected induction, saying that it is possible to refute hypotheses but it is impossible to prove them. He proposed the hypothetico-deductive method: if an initial hypothesis is true something will happen; through an experiment we plan to refute it and if the results are not compatible with the premises, the hypothesis will be rejected; otherwise the hypothesis is corroborated (but not confirmed). This reasoning could have inspired the classical theory of inference.

According to Lindley (1990), formal logic does not help to prove a hypothesis: "Scientific inference is essentially the usage of observed data in the past to predict future data not observed yet." The fundamental problem of statistical inference is to express an opinion about a non observable set y from an observable set x".

Usually, scientific inference turns into statistical inference when the relationship between the unknown state of nature and the observed object is translated into probabilistic terms.

The formal study of Statistical Inference began in the XX century and in almost all the basic texts of statistics there are many chapters dedicated to this subject. Most of them present, in an instrumental way, the so-called classical inference, but nothing is written about inference from a logical point of view nor about alternative views of inference, for example, the Bayesian school (there are few exceptions to this).

In the classical approach the method of Neyman and Pearson is used (it could be interpreted as an extension of the Fisher method) and it is based on the conditional reasoning *modus tollens* of formal logic. It is a hypothetico-deductive method, although in their original paper they considered it an *inductive conduct*. From another point of view, Good (1989) said that scientific induction is the change of the probability of a hypothesis based on evidence, and this problem is partially solved by the Bayesian theory (based upon Bayes' rule, XVIII century). The adepts of Bayesian theory advocate it as being an inductive procedure, though this has been disdained by members of the other school due to its subjective approach. We will discuss both approaches.

CONDITIONAL LOGIC

Let us begin with some logical principles of conditional logic, used by inference in the classical approach. The conditional reasoning appears in two forms:

Modus ponens *(affirmative way)*	Modus tollens *(negative way)*
If **P** then **Q**	If **P** then **Q**
P occurs ∴ **Q**	**~ Q** (**Q** does not occur) ∴ ~ **P**

These are the only ways to have valid conclusions. In probability terms, both valid conclusions could be considered as P(Q | P)=1 and P(~P| ~Q)=1. But if we assume that Q occurred, then we could have P or ~P (inconclusive).

The Neyman-Pearson theory uses a *modus tollens* form, that is:

If the Hypothesis H is true (P) then the data should have the behavior Q. If the data have no such behavior (~Q) then H should be rejected (~P).

This theory does not allow us to establish probabilities of the hypothesis–the only answer is reject/non-reject H (with an associated risk, defined in an appropriate way). The problematic point occurs when the data actually have the expected behavior **Q**: students tend to accept H (fallacious argument, because **if Q** we cannot conclude **P**!).

Modus tollens is generally difficult for people at first, because it is reasoning by negation. Nor is it easy for undergraduate students either. I have confirmed this last statement by applying to my students, for several years, the well known Wason task (Anderson, 2000).

The experiment uses 4 cards: one side has a number and the other side has a letter. The statement is: If a card has a vowel (**P**) then it has an even number on the other side (**Q**). You show the students the following arrangement:

U J 6 9

and ask them which card must be turned in order to confirm or disprove the statement.

The easiest way is to apply *modus ponens* first (direct way–if vowel then even) and card U must be turned, and then *modus tollens* (reverse way–if not even then not vowel) which means to see the other side of card 9. The first movement is used by 80% of the students but the second is more difficult–less than 25% of correct answers. This shows that students have difficulties with the *modus tollens* form. Therefore, it is not surprising they tend to accept the null hypothesis H when the data have the desired behavior(**Q**): it is a fallacy!

Students behave as if the conditional statement were bi-conditional (if and only if) or as if they had introduced a probabilistic reasoning (**P** has a probability to occur if **Q** occurs). A discussion with medical students about bioequivalence (which means that two different drug products are therapeutically equivalent) could help them to understand the difficulties of the *modus tollens* approach. As Hauck and Anderson (1996) have pointed out, “We often see examples where statisticians and non-statisticians are testing the wrong hypotheses, apparently stuck in a mode of thinking based on null hypotheses of no-difference.” To accept that two drugs are bioequivalent, what is needed are tests of alternative hypotheses of negligible difference.

Of course classical inference is very useful in a large number of situations, but we must stimulate the discussion about its role among our students.

BAYESIAN APPROACH

The Bayesian theory deals with probability statements which are conditional on the observed value. This conditional feature introduces the main difference between Bayesian inference and classical inference. Despite the difference, in many simple analyses we get to superficially similar conclusions from the two approaches. According to Gelman et al. (1995), analyses obtained using Bayesian methods can be easily extended to more complex problems.

Let θ be a parameter and y the observed data. Let $P(\theta)$ be the prior distribution for θ and $P(y \mid \theta)$ be the likelihood. So Bayes’ rule uses the conditional probability to produce the posterior distribution

$P(\theta \mid y) = P(\theta) \,.\, P(y| \theta)/ P(y)$, where $P(y)$ considers all possible values of θ.

Gelman et al. (1995) give a simplified expression which “encapsulates” the technical core of Bayesian inference:

$$P(\theta| y) \approx P(\theta)\, p(y \mid \theta)$$

We propose to introduce students to the Bayesian theory by presenting two examples of inference: the first one is an estimation problem and the second is a testing problem.

Estimation problem (Cordani, 2001 - adapted from Albert, 1996)

A football coach wants to assess the proportion (**p**) of goals (success) scored by a new applicant for the team.

First of all, we should define a *prior distribution* for p. We can assign some discrete values to **p**, according the coach’s opinion: **p**=0.40 (weak player), **p**=0.50 (regular player), **p**=0.75 (good player) and **p**=0.90 (excellent player). To complete the distribution we can assign the same probability (0.25) to each value of **p** (a non-informative criterion) or different probabilities, according to the coach’s knowledge about the applicant. This procedure has an essential difference to the classical theory, because here **p**, the unknown parameter, has a distribution, which summarizes our previous opinion about **p**.

Beginning with the non-informative approach, the prior distribution for **p** is a uniform discrete distribution as follows:

Values of p	0.40	0.50	0.75	0.90	(*)
Probabilities	0.25	0.25	0.25	0.25	

The second step is to verify how many *successes* were *observed* in **n** trials. Suppose there were 12 goals in 20 independent shots. So the associated likelihood is $\mathbf{p}^{12}$ (1- $\mathbf{p}$)^{8} .

The third step is to calculate the *posterior distribution* for **p**, for which we must apply Bayes' rule, through the following scheme:

(posterior) ≈ (prior) x (likelihood).

In this example,
P(p=0.40 | 12 successes) =
P(p=0.40) . P(12 successes | p=0.40) / P(12 successes) =
$0.25 . 0.40^{12}\ 0.60^{8} / (0.25 . (\sum(p^{12}\ (1\text{-}p)^{8}) = 0.1637$,
for each value of **p** shown in (*).

So, this posterior probability (16.37%) corresponds to the value **p**=0.40, and is less than the value (0.25) assigned in the prior distribution, which means that the sample evidence decreased the chance of the value 0.40. In other words, with 12 goals, the player has less chance to be considered weak. Other values (with analogous calculations) for the posterior distribution are:

p	**prior**	**posterior**
0.40	0.25	0.1637
0.50	0.25	0.5537
0.75	0.25	0.280
0.90	0.25	0.0018

Mean of the prior distribution = 0.6375
Mean of the posterior distribution = 0.5546

Initially, every **p** could be feasible, since each one has the same probability (0.25). The mean of the prior distribution is 0.6375, which could be interpreted as the expectation of the coach that the player is more than regular (if not, why to hire him?). Looking at the posterior distribution, the most probable value for **p** is 0.50 (regular player) with a probability of 0.5537. The associated probability for an exceptional player is very low (0.0018), less than the initial one (0.25). All the comments here are taking into account the sample results (12 successes in 20 trials) and the prior distribution as well. It is expected that it corresponds to the coach's experience and satisfies coherence rules (Gelman et al., 1995).

It is usual to consider the mean (or median or mode) of the posterior distribution as an estimate of **p**. So, in this example, a Bayesian estimate for **p** could be the mean of the posterior distribution, which is equal to 0. 5546, showing that his hopes with the new player after the evidence are less than his initial ones, (0.5546 *vs* 0.6375). After the test, the new player is closer to being a regular player than a good player – so the decision is up to the coach.

A different prior could produce different results. Suppose the coach has a more informative prior distribution, as follows:

p	**prior**	**posterior**
0.40	0.10	0.0681
0.50	0.15	0.3457
0.75	0.50	0.5844
0.90	0.25	0.0018

Mean of the prior distribution = 0.715
Mean of the posterior distribution = 0.640

In this case, the coach bets on the player, assigning the highest prior probability to **p**=0.75 (a good player). The expected value of the prior distribution is 0.715, which indicates more confidence in this player now. Based on the posterior distribution, the most probable value for **p** is 0.75 (probability = 0.5844) but the expected value for the posterior is 0.64, which can be chosen as the estimate for **p**, suggesting to the coach that the player is closer to regular than he expected initially.

As we can see, the posterior distribution depends on the chosen prior. On the other hand, the likelihood depends on the observed data, and is equal in both situations. We can argue that more confidence in the player in the second prior pushed the posterior distribution towards p=0.75, differently from the first situation (uniform opinion).

The comparison of this approach with the classical one is interesting to the student: 20 trials and 12 successes give an estimate $\hat{p}$ = 0.60 (12/20). This result does not consider the previous opinion and the result is closer to both situations seen in the Bayesian approach.

Testing example

Taking advantage of the previous example, we can transform it into a hypothesis testing problem. Consider the interest knowing whether the applicant is a regular player (**p**=0.50) or a good player (**p**=0.75). We also keep the previous result, so there are 12 successes in 20 trials. The hypotheses can be written as:

$$H_0 : \mathbf{p} = 0.50 \quad \textit{vs} \quad H_a : \mathbf{p} = 0.75$$

Let G be the random variable which represents the number of goals (successes) in **n**=20 trials, and let **p** be the probability of success (scoring a goal). For independent trials we have a Binomial distribution for G (G ~ B (**n**, **p**)).

First of all, we should assign prior probabilities to both hypotheses H_0 and H_a. Let us choose the same probability for each, $P(H_0) = P(H_a) = ½$, that are the prior probabilities for the null and the alternative hypothesis respectively. How can the data modify the prior opinion about the hypotheses? Taking into account Bayes' rule, we can find the posterior distribution for each hypothesis:

$$P(H_0 \mid G) = P(H_0 \text{ and } G) / P(G) = P(H_0)\, P(G \mid H_0) / P(G)$$

$$P(H_a \mid G) = P(H_a \text{ and } G) / P(G) = P(H_a)\, P(G \mid H_a) / P(G)$$

Comparing both hypotheses using their ratio (Bayes Factor, see Gelman et al., 1995) we have

$$\underset{\textbf{(1)}}{\frac{P(H_0 \mid G)}{P(H_a \mid G)}} = \underset{\textbf{(2)}}{\frac{P(H_0)}{P(H_a)}} \; \underset{\textbf{(3)}}{\frac{P(G \mid H_0)}{P(G \mid H_0)}} \qquad (**)$$

where **(1)** is the ratio of the posterior probabilities (that is, the probability of each hypothesis given the sample evidence), **(2)** is the ratio of prior probabilities of the same hypotheses and (**3)** is considered the ratio between the likelihoods [P(G|H), sometimes written as L(H), is not a probability, although many authors have used it in a standardized form that sums to 1]. If the quotient between the posterior probabilities is greater than unity, we say the evidence supports the null hypothesis H_0. Otherwise, the evidence supports the alternative hypothesis H_a.

As the observed value for G was 12, in 20 trials, and both priors are equal, (**) can be calculated as:

$$0.5^{12} \;.\; 0.5^{8} \;/\; 0.75^{12} \;.\; 0.25^{8} = 1.97$$

Such a value, being much greater than one, shows that the posterior probability of H_0 is superior to that of H_a, given the same evidence, so the suggestion is to consider the player as a regular one.

If we have $P(H_0)=2/3$ and $P(H_a)=1/3$ as a prior distribution, the calculation for (**) will result in 3.94, offering even more support to H_0, given that the weight was initially greater.

The teacher can carry out the classical test and it will show there is no evidence to reject H_0. This theory allows the researcher to give just a single answer: reject/not reject (with an associated probability of error), while the Bayesian approach offers probabilities of the hypotheses.

CONCLUSION

The purpose of this paper is two-fold. On the one hand, showing the difficulty inherent in the logic that is used in the classical inference theory and on the other hand showing that could be easy to find motivating examples for an initial usage of the Bayesian principles. With this initial help, the student should have more confidence to choose his or her own way in the research field.

The Bayesian theory develops skills that can be used to deal with uncertainty, represented by assignment of probabilities to our previous knowledge of the phenomenon. It is very important to learn the classical theory also, if one knows how to extract the best of it and its limitations. From the examples presented, we can see that the Bayesian approach provides a rich discussion, giving to people a kind of culture for thinking probabilistically. The well-known tension between the two schools might be introduced in the pedagogical field, given that Education is a process in construction, a movement which prepares the spirit for criticism.

REFERENCES

Albert, J. H. (1996). *Bayesian computation using Minitab*. Belmont: Duxbury Press.

Barnett, V. D. (1973). Comparative statistical inference. Wiley. Chichester.

Copi, I. M. (1978). Introduction to logic (translated). Editora Mestre Jou, São Paulo.

Cordani, L. K. (2001). O ensino de estatística na universidade e a controvérsia sobre os fundamentos da inferência. Unpublished PhD thesis.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Good, I. J. (1989). The interface between statistics and the philosophy of science. In J. Fenstad et al., (Eds.), *Logic, Methodology and Philosophy of Science, VIII*. Elsevier Science Publishers. Amsterdam.

Hauck, W. H., & Anderson, S. (1996). Comment to “Bioequivalence trials, intersection-union tests and equivalence confidence sets” by Berger, R. L. & Hsu J. C. *Statistical Science, 1996, 11*(4).

Lindley, D.V. (1990). The 1988 Wald Memorial Lectures: The present position in Bayesian statistics. *Statistical Science, 5*(1).

Popper, K. (1975). *The logic of scientific discovery* (translated). Cultrix & EDUSP. São Paulo.