The Analysis of Experimental Data:

The Appreciation of Tea and Wine

Dennis V Lindley
University of Warwick, England.

KEYWORDS:
Teaching;
Significance level;
Likelihood;
Alternatives;
Bayes formula.

Summary
A classical experiment on the tasting of tea is used to show that many standard methods of analysis of the resulting data are unsatisfactory. A similar experiment with wine is used to show how a more sensible method may be developed.

The experiment with tea

One afternoon in the 1920's at Rothamsted Experimental Station, the statistician, R.A.Fisher, made Muriel Bristol a cup of tea. She protested when he put the tea infusion into the cup before adding the milk, claiming that she could discriminate whether the milk had been added first or second, preferring the former. Fisher then devised a classic experiment that is beautifully discussed in chapter 2 of his book, Fisher (1935). The principles developed there are today widely used in the design and analysis of many types of experiment. Because the original experiment leads to technical difficulties in its analysis, we shall here consider a modified form that avoids them, yet retains all the essential qualities of the original.

In the modified form, the lady is presented with a pair of cups of tea and told truthfully that one has had the milk put in first, whilst the other has had it added to the tea infusion. She is required to identify which is which. The only possible results are right, denoted by R and wrong W. The experiment is to be repeated with 6 pairs of cups in all. Suppose that the result is RRRRRW with only the last pair wrong. Fisher's analysis goes as follows.

First suppose the lady is completely unable to do what she claims so that she is effectively guessing which cup of the pair is which. The hypothesis of her inability to perform the task is called the **null hypothesis**. In Fisher's view the purpose of this, and many other, experiments is to provide an opportunity of discrediting the null hypothesis that she is guessing. Here the null hypothesis means that each pair is R with probability 1/2 or W with probability 1/2, independently of the others. The observed result has probability $(1/2)^6 = 1/64$.

Fisher then argued that either

(a) the null hypothesis is true and an event of small probability has occurred, or

(b) the null hypothesis is false and the lady has discriminatory powers.

In this case the small probability is 1/64. Since events of small probability only rarely occur, we might favour (b) as the more reasonable explanation in getting all correct except one. The result is said to be **significant** with probability 1/64 and the probability is the **significance level**. The key idea is that if something which is unusual on the null hypothesis happens, then the null hypothesis is discredited. Nowadays it is common to use a value of 1/20, or 5%, as a benchmark and to say the result is **significant at 5%** if the small probability is less than this value, as with our result.

Fisher immediately realized that this argument fails because every possible result with the 6 pairs has probability $(1/2)^6 = 1/64$, so **every** result is significant at 5%. Fisher avoided this absurdity by saying that any outcome with just 1 W and 5 R's, no matter where that W occurred, is equally suggestive of discriminatory powers and so should be included. There are 6 such possibilities, including the actual outcome, so the relevant probability for (a) above is $6(1/2)^6 = 6/64 = .094$, so now the result is **not** significant at 5%.

Fisher's amended argument for a general situation replaces the probability of the outcome on the null hypothesis by the probability of that and similar outcomes; here the probability of 1 error in 6. Fisher realized that even this would not work. For what is the most probable result with pure guessing? Clearly one half of the pairs right and one half wrong. For 128 pairs of cups with 64 R and 64 W, the probability is ${}^{128}C_{64}(1/2)^{128}$ which is about .05. This is for the most probable outcome, every other outcome has smaller probability. So for 128 pairs we are back to the difficulty that every result is significant at 5%. To overcome this, Fisher ingeniously argued that if 1 error in 6 is significant, so surely is no error, or 6 R's. In other words, cases that more strongly suggest discriminatory powers than in the case observed should also be included when calculating the probability to be judged against 5%. Outcomes that suggest powers as, or more, strongly than the outcome observed are said to be as, or more, **extreme**.

The upshot of this is that Fisher's simple, either (a) or (b) above, has to be amended to read: either

(a) the null hypothesis is true and the probability of events as, or more, extreme than that observed is small, or

(b) the null hypothesis is false and the lady has discriminatory powers.

This form is accepted by most statisticians and the scientific literature is full of 5% significances, where the 5% refers to the probability of all results *as, or more, extreme* than that observed. It is the italicised words that distinguish the accepted form from that first given by Fisher. With the outcome RRRRRW of probability $(1/2)^6$, there are 5 others as extreme and 1, with no errors, more extreme, giving 7 cases in all and a total probability of $7(1/2)^6 = .109$, not significant at 5%.

A Criticism

For many years the argument went largely unchallenged and was supported by alternative, more mathematical, approaches due to Neyman, Pearson and Wald. But recently doubts, originally advanced by Jeffreys, have crept in and the argument is increasingly being attacked. Let us see how the criticism works for the outcome RRRRRW. Fisher has to consider what results are as, or more, extreme and to do this he takes other possibilities with 6 pairs of cups. But why fix 6? The value 6 may have arisen by chance. Perhaps Dr. Muriel Bristol had a meeting to go to after tea and had to leave after 6 pairs. Had the cups not been prepared so efficiently, she might have done fewer. Another possible form of experiment, suggested and used by J. B. S. Haldane in the context of cats rather than tea-tasting, is to go on until the first mistake is made. Dr. Bristol's result is comp-atible with this type of experiment-ation. So let us use Fisher's argument for Haldane's experiment. The probability of the sequence RRRRRW is still $(1/2)^6$. More extreme sequences are those in which the first mistake occurs after the sixth pair. Thus at the seventh, probability $(1/2)^7$; eighth, probability $(1/2)^8$; and so on. The probability of the observed result and more extreme ones is therefore

$(1/2)^6 + (1/2)^7 + (1/2)^8 + \ldots$

$= (1/2)^6/(1 - 1/2) = (1/2)^5 = .031.$

Before we had .109, yet now we have significance at 5%. This is surprising.

Let us see where we stand. If the experiment consisted of 6 pairs of cups being tested and the result was RRRRRW, the relevant probability is .109. If the experiment consisted of pairs being tested until the first error, with the same result, the relevant probability is .031, less than a third of the previous value. And lack of significance in the first case changes to significance in the second. Is not this absurd? Here are 6 pairs of cups honestly being tested, resulting in RRRRRW; what does it matter what might have happened (for example RRRWRR in one case, RRRRRRRW in the other) but did not? What would be the probability if Dr. Bristol had stopped because of the meeting?

Let us pinpoint the difficulty with Fisher's either/or argument. It lies in deciding just what results are as, or more, extreme than that observed. (We have seen that the extreme results must be included since there are experiments in which every result is unusual.) In the case of a fixed number, 6, of cups, the extreme values are different from those in the case where one continues until a mistake is made. Let us call these two experiments the **fixed** and the **sequential** respectively. It might be argued that the judgements should depend on whether the fixed or sequential experiment was used. But, if you feel that, consider the following experiment. A fair coin is tossed, if it comes down heads, the fixed experiment with 6 cups is used; if tails, the sequential one is adopted. The result RRRRRW has probability associated with it equal to the average of the two experiments, namely $(.109 + .031)/2 = .070$, and the result is not significant at 5%. But if the coin came down tails and the sequential form used (with a natural probability of .031) should we really quote .070 merely because the coin might have shown heads? The suggestion seems strange. Attempts have been made to define exactly what is meant by more extreme but without success.

So we have to abandon the use of more extreme outcomes. This leaves us only with the probability of what happened and we have seen that is unsatisfactory because in some experiments all probabilities are small. So what are we to do?

An Alternative Analysis

Fisher's approach only considers probabilities on the null hypothesis. It does not consider probabilities were Muriel Bristol to have discriminatory powers. Of course, if she had perfect power then R would have probability 1 and the sequential experiment never end. But even the most enthusiastic supporter of the thesis that the milk must go in first would admit to occasional lapses. We saw that on the null hypothesis each R had probability P, independent of the others, with $P = 1/2$. A reasonable indication of discriminatory power would admit a value of P in excess of 1/2. The higher the value of P, the greater is the lady's ability. The values of P above 1/2 are called **alternative hypotheses**. The result RRRRRW has probability $P^5(1-P)$, $P = 1/2$ giving $(1/2)^6$ as before. This is called the **likelihood** function of P, the probability of correct classification, for the observed result. In general, it describes the probability of the observed result as a function of P. Modern work says that it is this function that is required, not any consideration of more extreme cases. The probability of what actually happened is considered under various hypotheses, rather than the probability of several outcomes solely under the null hypothesis.

What has to be done is to compare the probability on the null hypothesis with probabilities for other values of P, the alternative hypotheses. But which value of P? To answer this consider another lady.

The experiment with wine

This lady is a wine expert, testified by her being a Master (sic) of Wine, MW. Instead of tasting tea, she tasted wine. She was given 6 pairs of glasses (not cups). One member of each pair contained some French claret. The other had a Californian Cabernet Sauvignon, Merlot blend. In other words, both wines were made from the same blend of grapes, one in France, the other in California. She was asked to say which glass had which. That is, she did the same experiment as Dr. Bristol but with the two wines instead of the two preparations of tea. Suppose she got the same result RRRRRW and consequently the same likelihood function $P^5(1 - P)$, P now referring to the probability of classifying the pairs of wines correctly.

At this point I can only speak for myself though I hope that many will agree with me. You may freely disagree and still be sensible. I believe that Masters of Wine can distinguish the Californian imitation from the French original. Mathematically I think that $P > 1/2$. Yet I think it doubtful that ladies can distinguish the two methods of teamaking. $P = 1/2$ seems quite reasonable to me there though I admit that $P > 1/2$ is possible. So what I want to do is to put something into the analysis that incorporates my belief that tea is different from wine. Notice that the likelihood is the same for both though the meaning of P is different.

The way this is done is to introduce probability distributions for P appropriate for tea and for wine. Let me give you my distributions to illustrate the ideas. For wine, I chose the expression

$48(1 - P)(P - 1/2)$,

for $1/2 < P < 1$, (1)

having the form illustrated in Figure 1 and labelled prior. This expresses the fact that I think that she can discriminate but can make mistakes. The value 48 makes the total probability 1. For tea I took

0.8 for the probability that $P = 1/2$

and $1.6\,(1 - P)$ for $P > 1/2$,

having the form illustrated in Figure 2 and labelled prior. This expresses my personal probability of 0.8 that she cannot discriminate. (Fisher may have had such a value since he expressed surprise at Dr. Bristol's claim, reportedly saying "Nonsense, surely it makes no difference", Box (1978).) This allows a probability of 0.2 that she can, thinking that having good discriminatory power (P near 1) is less likely than modest ones (P near 1/2). These formulae reflect my own views. You may freely insert your own. More details will be found in Lindley (1984).

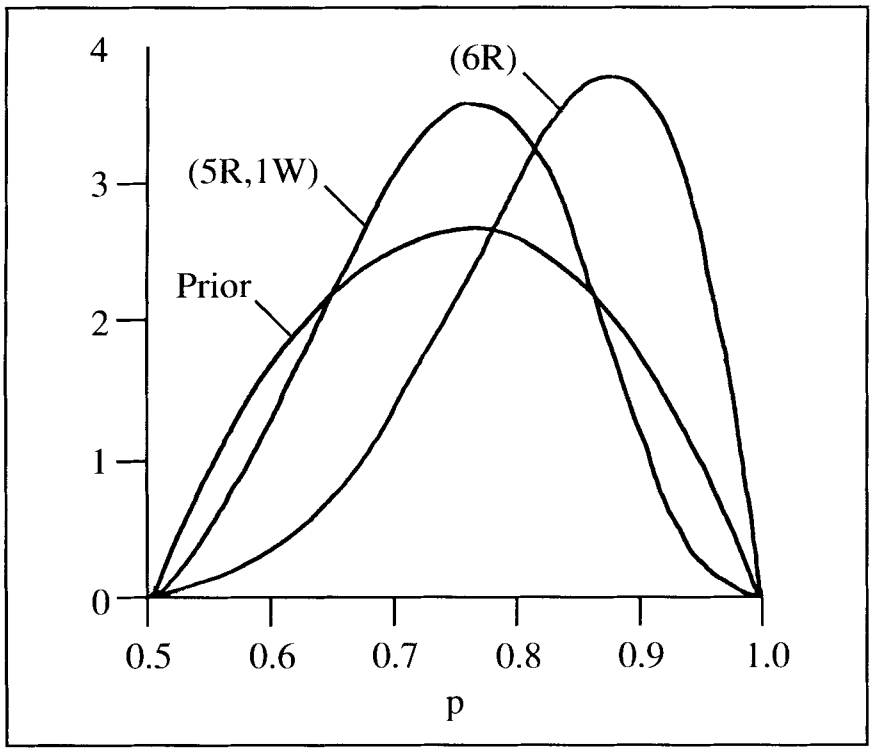

Figure 1. Wine.

Bayes Formula

It is next necessary to combine these personal opinions with the evidence of the data expressed through the likelihood function. The calculus of probability tells us how this is to be done, namely by multiplying the original probability by the likelihood function. For the lady tasting wine we have

$48(1 - P)(P - 1/2)P^5(1 - P)$,

for $1/2 < P < 1$.

Apart from the fact that the total probability is not 1, this is a probability distribution. Simple, but tedious, calculations enable us to find a constant K such that

$K(1 - P)^2 P^5 (P - 1/2)$,

for $1/2 < P < 1$ (2)

is a probability distribution, having integral from 1/2 to 1 of 1. The first probability distribution (1) is called the **prior** distribution (prior, that is, to the data). The one just obtained, (2), is called the **posterior** distribution. The formula says

posterior = K × prior × likelihood,

where K is a number chosen to make the integral of the right-hand side 1. It is called **Bayes** formula and the method is termed **Bayesian**. The only complication in its calculation is the determination of K.

Figure 1 shows for the wine-tasting (i) the prior distribution (1), (ii) the posterior distribution (2) for the case of 6 pairs yielding 1 error, and (iii) the same with no errors. Initially I thought

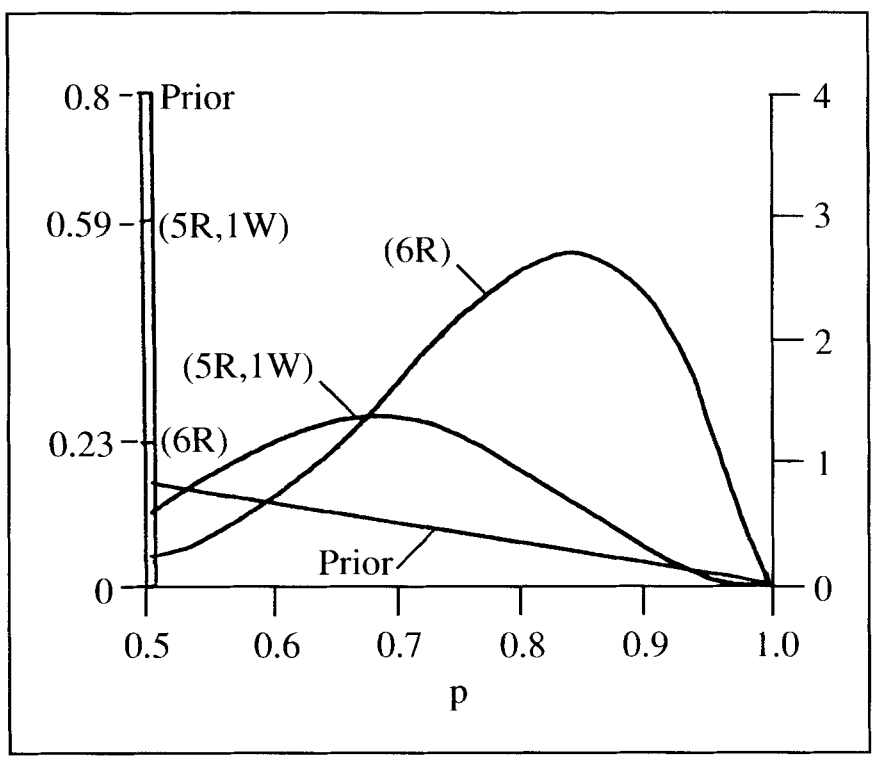

Figure 2. Tea.

$P = 3/4$ was the most probable value but there was substantial uncertainty expressed by the large spread about that value. With 1 error, there has hardly been any shift in the most probable value but I am slightly more confident that P is near 3/4 as expressed by a smaller spread. To understand the spread, consider the area under these curves between say .6 and .9, .15 either side of $P = 3/4$. The area, and hence the probability, is a little larger for the posterior distribution than for the prior. With no errors, the situation changes and the most probable value has risen to around .87 and the spread is substantially lower. For example, the probability that P is less than .75 is about .2 whereas

originally it was .5.

The situation with tea is subtler because I had initially a probability that she could not discriminate, P = 1/2, which was not entertained with wine. The similar graphs are shown in Figure 2. The prior value of this probability was .8, which drops to .59 when 1 error is made in 6 pairs, and to .23 with no errors. It is these values that can be contrasted with the significance levels, the probabilities of results as, or more, extreme than the actual results on the null hypothesis. The latter are .109 and .016 respectively. Notice that in both cases the significance probability is substantially lower than the posterior probability. A partial reason for this is the high value of the prior probability at .8. But the statement is still true even if one thinks that the lady is just as likely to have the power as not, expressed through a prior probability of .5. For example, with 1 error in 6 pairs, the posterior probability is .26 compared with a significance level of .109. It is typically true that the posterior probability of the null hypothesis exceeds the significance level, though there is no logical connection between the two values. The behaviour of the curves for the distributions for $P > 1/2$ is similar to those for wine.

The analysis just presented depends heavily on my opinion of the two ladies' abilities. Your opinion may be different. This seems sensible to me. On the slender evidence of 12 cups or glasses it is not surprising that our views might differ, just as scientists currently differ over the greenhouse effect because the evidence is inadequate. But had we evidence on 1200 cups, perhaps with 100 ladies, the different initial opinions would be swamped by the evidence of the data and we would essentially agree. Technically, the likelihood dominates the prior with a large sample. This happens in science. 20 years ago many of us were suspicious of the claims made that lead affected intelligence. The evidence now overwhelms the original opinions. All evidence does is to change opinions: it does not create them.

Conclusions

There are four lessons that can be learned from this analysis.

(a) Since the significance level is typically less than the posterior probability of the null hypothesis and a small value of the former, like 5%, is going to cast doubt on the null hypothesis, it follows that null hypotheses will be more easily discounted using Fisher's method rather than the Bayesian approach. When it is remembered that a typical null hypothesis is that a drug is of no use, or that a treatment is ineffective, it will be seen that the plethora of significance tests that are used today will encourage specious beliefs in the efficacy of drugs or treatments. Whenever you read of some effect having been detected, remember that it probably refers to significance, which too easily suggests an effect when none exists.

(b) The Bayesian analysis provides the scientist with what he requires. He is interested either whether or not the null hypothesis is true (as with tea) or in the magnitude of the effect being investigated (as with wine) or both. He requires a measure of belief in either of these and probability provides such a measure. For the null hypothesis directly; for the magnitude, in our example expressed through P, by a probability distribution illustrated by the curves in the figures. This is in marked contrast to the significance level which provides a probability for something that did not happen on a hypothesis that may not be true.

(c) The Bayesian analysis distinguishes between tea and wine. Fisher's analysis used only probabilities assuming guessing, and guessing is the same for both, as the word 'guessing' implies. The Bayesian view recognizes that one's opinion of tasting the two liquids may be different or that the ladies may have different skills.

(d) This is easily the most important point of the four. The Bayesian method is **comparative**. It compares the probabilities of the observed event on the null hypothesis and on the alternatives to it. In this respect it is quite different from Fisher's approach which is absolute in the sense that it involves only a single consideration, the null hypothesis. All our uncertainty judgements should be comparative: there are no absolutes here. A striking illustration of this arises in legal trials. When a piece of evidence E is produced in a court investigating the guilt G or innocence I of the defendant, it is not enough merely to consider the probability of E assuming G; one must also contemplate the probability of E supposing I. In fact, the relevant quantity is the ratio of the two probabilities. Generally if evidence is produced to support some thesis, one must also consider the reasonableness of the evidence were the thesis false. Whenever courses of action are contemplated, it is not the merits or demerits of any course that matter, but only the comparison of these qualities with those of other courses.

Summary

The main points are now summarized. Fisher argued in the form of a dichotomy; either (a) an event of small probability on the null hypothesis has occurred, or (b) the null hypothesis is false. This did not work and the probability had to include events that did not occur but were as, or more, extreme. This did not work because of the ambiguity over what is 'more extreme'. The way out of this difficulty is to compare the probabilities of what actually occurred on the null hypothesis and on alternatives to it. Since there are ordinarily several alternative hypotheses, they have to be weighted. This is done by expressing personal beliefs about the situation before experimentation. These prior beliefs, also in the form of probabilities, are then modified by the experimental data, using Bayes' formula, to give posterior beliefs. One compares the various possible explanations for what has happened, and compares one's posterior beliefs with those held initially. All the analysis is comparative.

Note

This paper is based on a Collingwood lecture presented by Professor Lindley at the University of Durham.

References

Box, J.F. (1978). *R.A.Fisher, the Life of a Scientist.* New York; Wiley.

Fisher, R.A. (1935). *The Design of Experiments.* Edinburgh; Oliver and Boyd.

Lindley, D.V. (1984). A Bayesian lady tasting tea. In *Statistics: an Appraisal.* Ed. H.A. David and H.T. David. Ames; Iowa State University Press, 455-485 (with discussion).