Bayes or chi-squared? Or does it not matter?

David J.C. MacKay July 12, 2005

Given a collection of characters plucked randomly from a document, how would you distinguish whether the document is an English-language document or a totally-random document?

How do we measure 'which hypothesis fits the data best'? Two approaches have been suggested:

Bayes' theorem, in which the log likelihood ratio is

$$\log \frac{P(\mathbf{x} \,|\, \mathcal{H}_P)}{P(\mathbf{x} \,|\, \mathcal{H}_Q)} = \sum_i F_i \log \frac{p_i}{q_i}, \tag{1}$$

where i runs over characters in the alphabet, F_i is the number of times character i actually occurred in the data string $\mathbf{x}$, and the two models $\mathcal{H}_P$ and $\mathcal{H}_Q$ state that the symbols come i.i.d. from the distributions $\mathbf{p}$ and $\mathbf{q}$ respectively.

Chi-squared. In a chi-squared approach, we see how close the observed character counts are to the expected counts according to the two hypotheses. (By the way, I'm not asserting that an expert non-Bayesian statistician would use this method; non-Bayesian statistics offers many ways to answer a question, instead of just one, and experts might well use another approach! I discuss the chi-squared approach here because my impression is that many undergraduates have been taught about using chi-squared as a way to fit the parameters of a model; I find that those students are likely to use chi-squared to compare two alternative values of the parameters of a model, too.) The goodnesses of fit are

$$\chi^2_P = \sum_i \frac{(F_i - p_i N)^2}{p_i N} \quad \text{and} \quad \chi^2_Q = \sum_i \frac{(F_i - q_i N)^2}{q_i N}, \tag{2}$$

where N is the number of characters received. We choose the hypothesis with smaller χ^2.

These two approaches do not always make the same decision. (Notice that the log likelihood ratio is a *linear* function of $\{F_i\}$ whereas $\chi^2_P - \chi^2_Q$ has a *quadratic* dependence on $\{F_i\}$.)

Bayes' theorem gives the correct answer, according to the axioms of rational inference: it tells you how probable the alternative hypotheses are. But our students have been taught to use chi-squared.

Does it matter? Is there much practical difference?

It would be surprising if chi-squared and the log likelihood ratio gave *opposite* answers *all* the time. We expect them typically to agree. But there must be examples where there are small differences, since, as we noted above, the functional dependence on the data $\{F_i\}$ is different in the two expressions. We will need to hunt a little to seek out examples that magnify the differences between these two approaches.

Perhaps we can find cases where there is a *big* difference, and it is evident to the intuition which of the two answers is correct.

The Galilean approach

Ed Jaynes calls hunting for example data sets that magnify the difference between alternative worldviews the Galilean approach. Having got his telescope pointed at Jupiter and its moons, Galileo encouraged his discussants to *look*.

Where best to point our telescope? We expect the biggest differences between the approaches will be found for data sets that are in some way atypical. And we would like to find examples where it is evident to the intuition which hypothesis should be the winner. I have satisfied these two constraints by choosing a fake data set that has atypically small fluctuations. The examples

that follow have counts that are *atypically uniform*, or, to put it another way *atypically typical* of one of the hypotheses. This extreme typicality makes it easy for the discussant to understand what is going on.

Let the alphabet size be $I = 26$; let the two hypotheses be

$$\mathbf{p} = \frac{1}{13 \times 14} (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13)$$
$$\mathbf{q} = \frac{1}{13 \times 14} (6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8).$$

Let the number of letters collected be $N = 14$. Imagine that the outcome of these 14 rolls of the dice is

$$\mathbf{F} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1) .$$

This outcome is 'extremely typical' of the distribution $\mathbf{p}$. The expected count in each of the right-hand bins, according to $\mathbf{p}$, is 1, and this is exactly what occurred; the expected sum of the counts in the left-hand bins is 1, and indeed there is exactly one count in those bins.

So, what's chi-squared?

$$\chi^2_P = \sum_i \frac{(F_i - p_i N)^2}{p_i N} = 12. \qquad \chi^2_Q = \sum_i \frac{(F_i - q_i N)^2}{q_i N} = 9.29. \tag{3}$$

Thus according to chi-squared, model Q has 'better fit' to the data than model P! But surely it is evident to the intuition that this data set favours model P? A data set of this size couldn't be more typical of P. Indeed, the log likelihood ratio agrees.

$$\log_e \frac{P(D \,|\, \mathbf{p})}{P(D \,|\, \mathbf{q})} = 4.5. \tag{4}$$

The data are 99-to-1 in favour of $\mathbf{p}$, according to Bayes' theorem.

Is there any need to discuss the issue further? This one example convinces me that chi-squared is capable of giving absurdly wrong answers. In contrast I have never seen an example where the log likelihood ratio favours a model that is evidently the wrong choice out of two.

Chi-squared gives different answers to equivalent problems

Another way of criticising an approach to inference is to find two situations that are equivalent, but for which the approach gives different answers.

The above 26-character situation provides a perfect starting point. Noticing the both models $\mathbf{p}$ and $\mathbf{q}$ assign uniform distributions over the first 13 characters of the alphabet (let's call them hearts), and both assign uniform distributions over the second 13 characters (let's call them clubs), a smart data-processor will observe: 'when comparing $\mathbf{p}$ with $\mathbf{q}$, the precise outcome within each suit is not relevant; all that matters is the frequency of each suit'. Thus the model comparison question is completely equivalent to the following situation. The two hypotheses are:

$$\begin{aligned} \mathbf{p} &= \tfrac{1}{14} (1, 13) \\ \mathbf{q} &= \tfrac{1}{14} (6, 8). \end{aligned} \tag{5}$$

The number of letters collected is $N = 14$. The outcome of these 14 rolls of the dice is

$$\mathbf{F} = (1, 13) \tag{6}$$

The outcome of the chi-squared comparison is completely changed.

$$\chi^2_P = 0. \qquad \chi^2_Q = 7.3. \tag{7}$$

In contrast, the log likelihood ratio is unchanged, in accordance with the data processor's intuition.

$$\log_e \frac{P(D \,|\, \mathbf{p})}{P(D \,|\, \mathbf{q})} = 4.5. \tag{8}$$

The chi-squared answer depends catastrophically on the choice of how the data is binned.

Other examples

Here is a Galilean example with two components only: The two hypotheses are:

$$\begin{array}{rcl} \mathbf{p} & = & (0.001, 0.999) \\ \mathbf{q} & = & (0.070, 0.930). \end{array} \tag{9}$$

The number of letters collected is $N = 100$. The outcome of these 100 rolls of the dice is

$$\mathbf{F} = (1, 99) \tag{10}$$

The outcome of the chi-squared comparison is:

$$\chi^2_P = 8.1. \qquad \chi^2_Q = 5.5. \tag{11}$$

The winner, according to chi-squared, is $\mathbf{q}$. In contrast, the log likelihood ratio is:

$$\log_e \frac{P(D \,|\, \mathbf{p})}{P(D \,|\, \mathbf{q})} = 2.84, \tag{12}$$

corresponding to a posterior probability in favour of $\mathbf{p}$ of 94%, assuming equal priors on $\mathbf{p}$ and $\mathbf{q}$.

Questions and answers

Q: Why make a fuss about this issue? Chi-squared only gets the wrong answer in atypical cases. In practice, chi-squared will almost certainly work fine.

A: Why use a method that only *sometimes* works fine, when there is a method available that *always* works fine? Surely we are mis-educating our students if they come away from our courses using chi-squared and similar bodges instead of Bayes' theorem?

Furthermore, the chi-squared approach returns only a chi-squared difference, a number with no straightforward interpretation. In contrast, the log likelihood ratio really means something: the likelihood ratio is the factor by which the data should modify one's beliefs about the two hypotheses; a '99-to-1' likelihood ratio is something the man in the street understands correctly. The log likelihood ratio quantifies the difference in the compression that would be achieved if the two models were used to compress the data.

Q: I have an alternative to chi-squared, namely the sum-of-squared errors. Chi-squared gets wrong answers, but maybe sum-of-squares is better behaved?

A: Given any method whose answers are not *identical* to those given by the log likelihood, I expect the Galilean technique will be able to find situations where the answers differ substantially, and where intuition would side with one approach – no prizes for guessing which my money's on! As an illustration, here's a case where the sum-squared error criterion gives an answer different from both chi-squared and Bayes' theorem: The two hypotheses are:

$$\begin{array}{rcl} \mathbf{p} & = & (0.001, 0.999) \\ \mathbf{q} & = & (0.070, 0.930). \end{array} \tag{13}$$

The number of letters collected is $N = 100$. The outcome of these 100 rolls of the dice is

$$\mathbf{F} = (2, 98) \tag{14}$$

The chi-squareds are:

$$\chi^2_P = 36.1. \qquad \chi^2_Q = 3.8. \tag{15}$$

The sum-squared-errors are:

$$S_P = 7.2. \qquad S_Q = 50. \tag{16}$$

The log likelihood ratio is:

$$\log_e \frac{P(D \mid \mathbf{p})}{P(D \mid \mathbf{q})} = -1.5. \tag{17}$$

Thus according to sum-squared error, $\mathbf{p}$ is the winner; but according to χ^2 and Bayes' theorem, $\mathbf{q}$ is the better bet.

Q: So what are you recommending we do?

A: I recommend

- When discussing model comparison or parameter-fitting, we should cut all mention of chi-squared as a measure of goodness of fit. We should teach students about the likelihood function. We should train students to do model comparison using probability theory.
- Chi-squared need not be totally expunged; it is a helpful term for professionals to have in their vocabulary: I would mention chi-squared in two contexts:
  - Alternative-free model criticism. Imagine we want to measure 'whether model $\mathcal{H}_0$ seems to fit the data OK'. One of many ways to perform model criticism is to evaluate χ^2 and see if its value is typical.
  - Sketching and approximating likelihood functions. In *some* problems, the log likelihood function is exactly or approximately equal to the quantity $\chi^2/2$. It may be helpful to mention this connection.

Message to teachers: more Bayes' theorem, less chi-squared

Q: Aren't you being unfair to frequentist statistics, by making out that a professional frequentist statistician would use chi-squared in this way? Does anyone actually advocate using chi-squared to compare models in the way you describe? Since the two models have no parameters, frequentist theorists would treat it as a case where the null and alternative hypotheses are simple (ie, not composite), and recommend applying the Neymann-Pearson lemma, which results in a test based on the likelihood ratio.

A: As I said on page 1, I'm not asserting that an expert non-Bayesian statistician would use this method. Non-Bayesian statistics offers many ways to answer a question, instead of just one. I discuss the chi-squared approach here because my impression is that many undergraduates have been taught about using chi-squared as a way to fit the parameters of a model; I find that those students are likely to use chi-squared to compare two alternative values of the parameters of a model, too.

This is principally a note about educational strategy, rather than a criticism of professional frequentist statistics. I think that the feature of frequentist statistics, that it offers many ways to answer a question, is a defect, since a partially-trained student may end up using sub-optimal methods that no professional would use. In contrast, the Bayesian approach has the clear educational advantage that, once the assumptions have been defined, any well-posed question has a unique answer.