Planning Statistical Analysis: Wrong and Right Approaches Explained Using an Entertaining Example from Everyday Life

Chittaranjan Andrade, Nilesh B. Shah

ABSTRACT

Inferential statistical tests are used to examine hypotheses in research data but can also be applied to information in everyday life. Using data from a cricket tournament as an example, this article describes a plausible but wrong plan of analysis and explains what a correct method of analysis might be. Testing a hypothesis that was set after visual inspection of data or indiscriminate analysis can both result in false-positive conclusions. Making incorrect assumptions in statistical tests will also result in incorrect conclusions. Statistics is not merely about crunching numbers. It is also about knowing how to plan and execute the analysis.

Key words: Chi-square goodness-of-fit test, data mining, expected frequencies, false-positive errors, plan of analysis, primary hypothesis, probability

Twenty-two persons participated in a game in which they tested their ability to predict the outcomes of cricket matches (n = 60) during the course of the Indian Premier League (IPL), 2019. There were 59 matches played; 1 was washed out due to rain. Among the 22 participants, 1 dropped out after predicting the outcomes of four matches.

On all occasions, the participants made predictions on the morning of the match, before the start of the match; that is, in a uniform way. The participant with the highest prediction accuracy correctly identified the outcomes of 37 (63.8%) of the 58 matches for which she had offered predictions.

A question that might be asked is whether this participant had foresight that was greater than what could be accounted for by chance. Consider: each match had only two teams contesting, and so assuming that there are only two outcomes possible for a match, she had a 50:50 chance of being right, or a 29:29 chance
A Chi-square test was applied to examine the goodness of fit for the observed frequencies of 37:21 versus the expected frequencies of 29:29. The Chi-square value was 4.41 (df = 1), and the $P$ value was 0.036. In other words, the participant appears to have had statistically significant foresight.

There are flaws in this approach to the analysis and in the actual analysis as well. Readers are invited to consider what these flaws might be before reading the rest of this article.

**AN A PRIORI PLAN OF ANALYSIS**

In research, ideally, the primary and secondary hypotheses and the plan of analysis should be outlined in advance.[1] It is particularly fallacious to do a study and then test the statistical significance of an association only because the association “looks significant” on visual inspection of the data. This is because random variations in values in a sample may create what appear to be meaningful associations, although these associations are absent in the population from which the sample was drawn. Hence, applying statistical tests after spotting such “associations” in a sample may identify relationships that are statistically significant in that sample alone; the relationships may not be significant in other samples drawn from the same population and may not exist in the population.

In the IPL example, the precognition of the participant who scored 37/58 was tested statistically only because she had the best prediction score. She was not a priori chosen for testing. If some other participant had performed better than her, then that participant would have been tested. In other words, by visual inspection of the results, the best-performing participant, whoever s/he might have been, would have been selected to examine whether his or her prediction accuracy was statistically significant. If statistical significance was confirmed, it would only be because it was fated to be so; that is, because that performance had been chosen specially, after looking at the data.

If the statistical plan must be outlined in advance, can we plan to test every participant’s prediction accuracy against a 50:50 expected frequency of their being right versus wrong? No. This is because such a plan of analysis has no hypothesis; rather, it is a data mining exercise in which everything is tested in the hope that something will turn out to be statistically significant. As already explained, there are random variations in the data in all samples, and some of these variations may throw up spurious associations. When a large number of statistical tests are performed, spurious associations are more likely to be picked up. That is, many of the statistically significant results in data mining exercises may be false-positive (Type 1) errors.[2]

This does not mean that it is always wrong to test hypotheses that are set after examining the data, or that we should never embark on data mining exercises. However, this does mean that when we perform such analyses, we must be aware of the risk of false-positive findings that are due to how and why the hypotheses were set, or how the analysis was planned and executed. Therefore, in these situations, the results of such analyses must be considered hypothesis-generating, not hypothesis-confirming. Based on the results of the IPL example, we might hypothesize that the participant who was studied does have foresight and that her prediction abilities may merit prospective study as a primary outcome in another context.

**A SUGGESTED PLAN OF ANALYSIS**

So, what might be the correct approach to the analysis of the IPL data? One possibility is to calculate the percentage of correct predictions for every participant, including the one who dropped out after four guesses. The distribution of these percentages is expected to be normal. Outliers in this distribution, defined, for example, as those whose prediction accuracy is two or more standard deviations away from the mean, could be hypothesized to be either very good or very poor at prediction. With this taken as the hypothesis-generating study, the abilities of outlier participants can be tested in a subsequent hypothesis-confirming study.

Importantly, good or poor prediction ability does not necessarily mean good or poor precognition; it could also mean good or poor knowledge and good or poor ability to apply this knowledge in the field of cricket. Precognition refers to extrasensory powers, whereas knowledge and the application thereof describe conscious cognitive attributes and processes.

**THE EXPECTED FREQUENCIES**

In the example of the participant for whom the Chi-square goodness-of-fit results were computed, the observed frequencies for the test were 37 and 21 for right and wrong predictions, respectively. It was assumed that because her predictions could have been either right or wrong, she had a 50:50 chance of being right, and that the expected frequencies for right and wrong predictions, therefore, would be 29 and 29. This assumption is wrong. It cannot be assumed that her...
predictions were based on chance. It is almost certain that she had at least a passing awareness of which teams were playing and which team had better players. She would also almost certainly have known the outcomes of previous matches, and this knowledge of current form would have guided her predictions for subsequent match outcomes. So, the expected frequency for correct answers would almost certainly have been >29 and not 29.

This applies to all the participants; so, most, if not all, of the participants could be expected to have a prediction accuracy of above 50%. Because different participants would have different degrees of knowledge and different abilities to apply this knowledge, the distribution of prediction accuracy would still be expected to be normal. The mean of this distribution could be taken as the expected frequency for prediction accuracy if a Chi-square goodness-of-fit test is planned.

**AFTERNOTE**

Assume that cricket is a game of chance and that the expected frequencies for right and wrong predictions are truly 50:50. Did the participant who correctly predicted the results of 37 out of 58 matches exhibit precognition? To answer this question, we must phrase it in a different way: Assuming that the outcome of each match is random (50:50), like the toss of a coin, what is the probability that somebody would correctly predict 37 outcomes out of a total of 58 matches?

Probability is calculated as the number of favorable events divided by the total number of events. The number of favorable events is 37 correct predictions out of 58 matches; this number can be obtained in \(58C37\) possible ways. The total number of events is all the possible numbers of correct predictions, which can be obtained in \(58C0 + 58C1 + 58C2 \ldots + 58C58\) ways. The quotient of the two numbers provides the required probability.

What if, instead, we wish to calculate the probability that somebody will correctly predict the outcomes of at least 37 random-outcome matches instead of exactly 37 matches? The numerator here becomes \(58C37 + 58C38 \ldots + 58C58\); the denominator remains the same. Online permutation and combination calculators are available to perform these otherwise laborious calculations quickly.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

There are no conflicts of interest.

**REFERENCES**

1. Andrade C. The primary outcome measure and its importance in clinical trials. J Clin Psychiatry 2015;76:e1320-3.
2. Andrade C. Multiple testing and protection against a type 1 (false positive) error using the Bonferroni and Hochberg corrections. Indian J Psychol Med 2019;41:99-100.