03/0048/32

บทความ

Confidence Intervals and P-Values

ช่วงความเชื่อมั่นและค่าความน่าจะเป็นของการทดสอบทางสถิติ

*Chansuda Wongsrichanalai**

บทคัดย่อ : **จันทน์สุดา วงศ์ศรีชนาลัย. 2532. ช่วงความเชื่อมั่นและค่าความน่าจะเป็นของการทดสอบทางสถิติ. วารสารวิจัยวิทยาศาสตร์การแพทย์ 3 (2) : 133-137**

การรายงานผลการวิจัยทางวิทยาศาสตร์การแพทย์ โดยใช้ช่วงความเชื่อมั่น (confidence interval) ประกอบกำลังเป็นที่ยอมรับกันมากขึ้น ในประเทศไทยเรา การนำช่วงความเชื่อมั่นมาใช้ ยังไม่แพร่หลาย สาเหตุอย่างหนึ่งก็คือ นักวิจัยส่วนใหญ่ไม่มีโอกาสที่จะได้ฟื้นฟูความรู้ทางชีวสถิติมาเป็นเวลานาน จึงยังคงใช้วิธีเดิมในการทดสอบนัยสำคัญทางสถิติ และรายงานค่าความน่าจะเป็น (p-value) เป็นหลัก บทความฉบับนี้ได้อธิบายถึงความหมายที่แท้จริง ของช่วงความเชื่อมั่นและคุณสมบัติต่าง ๆ ของมัน โดยเฉพาะในแง่เปรียบเทียบกับค่าความน่าจะเป็น เพื่อที่ว่านักวิจัยจะได้นำไปใช้ในการตีความหมายผลงานวิจัยที่ตีพิมพ์ในวารสารได้ถูกต้อง และสามารถใช้ช่วงความเชื่อมั่น ในการรายงานผลงานวิจัยของตนเองได้อย่างเหมาะสม นอกจากนี้ยังเป็นที่หวังว่า บทความฉบับนี้คงจะช่วยลดช่องว่างระหว่างนักวิจัยและนักชีวสถิติได้ไม่มากก็น้อย หากวัตถุประสงค์เหล่านี้เป็นที่บรรลุถึงจุดหมายแล้ว เราคงตั้งความหวังที่จะได้ยินคำถามในทำนองที่ว่า "มีนัยสำคัญทางสถิติ (significant) ไหม ?" น้อยลง แต่กลับได้ยินคำถามในทำนอง "แล้วช่วงความเชื่อมั่น ล่ะ เท่ากับเท่าไร ?" มากขึ้นในไม่ช้า

* *Dept of Immunology and Biochemistry, US Army Medical Component, Armed Forces Research Institute of Medical Sciences (AFRIMS), Rajvithi Road, Bangkok 10400.*

แผนกอิมมูนวิทยาและชีวเคมี สถาบันวิจัยวิทยาศาสตร์การแพทย์ทหาร (ฝ่ายสหรัฐ) ถนนราชวิถี กทม. 10400

01/0042/32

Abstract : **Chansuda Wongsrichanalai. 1989. Confidence intervals and p-values. Thai J Hlth Resch 3 (2) : 133-137**

Reporting the results of medical research studies with confidence intervals (CIs) is becoming increasingly recognized. In Thailand, CI has not been widely used. This is partly due to the fact that the majority of medital investigators have had little chance to refresh their biostatistics or research methodology in recent years. As a result the conventional hypothesis testing, which leads to the report of p-values is still predominantly practices. In this paper the meaning of CI is explained and its characteristics described, particularly in comparison to p-values. This paper aims at helping researchers to interpret CIs correctly when reading journal articles, and to know when to apply the CI to report their own studies appropriately. It is also directed at bridging the gap between medical investigators and biostatisticians to some extent. Once these goals are achieved, we shell expect to hear less of such a question as "Is that statistically significant ?" and more of "..... what's the confidence interval associated with that estimate ?".

INTRODUCTION

Many of us who attended training courses on research methodology or biostatistics a decade or so ago in Thailand probably remembered what we learned about statistical tests of significance and p-values. We were told that reporting the result of a statistical test merely as "p = n.s. (not significant)" or "p = significant" was not informative enough and were encouraged to report the commonly-used levels of significance. For example, a p-value should be specified as $p < 0.01$, $p < 0.05$, or $0.01 < p < 0.05$. To our understanding at that time, test of significance was the key mathematical aspect of medical research.

Later on we began to notice in medical and scientific journals that there was a tendency towards reporting actual p-values rather than a range of them. $p = 0.049$ is more informative than $p < 0.05$. This is logical because a true difference between $p = 0.049$ and $p = 0.012$ is hidden if both are reported as $p < 0.05$.

Recently we have observed in medical research articles, an increasing number of study results presented in the form of an estimate, e.g., mean, mean difference, relative risk, etc., and a 95% confidence interval (CI). Reviewers have also begun to encourage authors to report CIs when the data justify their calculation. Campaigns for reporting of CIs have appeared in both the British Medical Journal (Gardner and Altman, 1986 ; Langman, 1986) and the Lancet (Anonymous, 1987 ; Bulpitt, 1987). Does this shift of interest from hypothesis testing reflect technical progress in the use of statistical methods in medical research ? What is a CI? Why has it become increasingly recognized and is it now outdate to report a p-value alone?

This paper describes the meaning and characteristics of the CI, particularly its strength relative to the p-value. The paper does not intend to demonstrate the calculation techniques for CI, but rather to help readers to understand its concept, be better able to interprete published articles and to make an extrapolation with confidence. It will also help medical investigators to effectively communicate with biostatisticians when seeking assistance in data analyses. Methods for the calculation of CI vary, depending on the parameters of interest (e.g., mean, relative risk, odds ratio, etc.) and the applied statistical assumptions. Methods can be found in textbooks of statistics and epidemiology (Rothman, 1986 ; Woolson, 1987) as well as in medical journals (Gardner and Altman, 1986 ; Morris and Gardner, 1988).

STATISTICAL TESTS OF SIGNIFICANCE AND P-VALUES

Comparison of a certain characteristic of two or more groups is frequently encountered in medical research. To answer whether babies born to normal mothers and antenatally malaria-infected mothers were

different in terms of birth weight, an investigator did a statistical test to compare the mean birth weights of the two baby groups and then reported the p-value. Suppose the difference between the two means were 0.25 kilogram (kg). P-value means that if the mean birth weights of the two groups of babies are the same, in other words if malaria in pregnancy has no effects on birth weight (i.e., if the two groups of babies are born to mothers who have been randomly selected from populations that are similar in terms of factors contributing to birth weight) the chance we will observe the difference in mean birth weights as big as 0.25 kg equals p. A small p-values, say < 0.05, indicates that given the null hypothesis is true (e.g., no difference in the mean birth weights of the two baby groups) the probability that the difference as big as or bigger than 0.25 kg will occur is less than 5%. Traditionally, this leads to the conclusion that the difference is real or the populations from which the subjects have been selected are dissimilar. A p-value of < 0.05 is intuitively taken by many as "statistically significant." To some clinicians, this means medically significant and to other scientists, it is of biological importance. This is not always true. With a large sample size, even a small difference can be statistically significant. Conversely, a large difference may be statistically non-significant if the number of subjects studied is small.

POINT ESTIMATES

So much emphasis was made on hypothesis testing in the past that often, only the word "significant" or "non-significant" or a p-value was reported in a paper. In comparing, say two means, the mean of each group and the difference between them should be reported. The sample statistic "0.25 kg" in the above example is a point estimate and is the kind of information readers always look for. Rather than saying that babies born to antenatally malaria-infected mothers have a "significantly" lower mean birth weight than those born to normal mothers, it is more informative to say what the mean weight of each group was and to give the size of the weight difference between the two groups. This allows the reader to independently assess the importance of the results and to make an individual extrapolation. For example, a housewife sought advice on where to make a good buy of grocery products. Friend A pointed out to her that things at Store x were "significantly" (whatever that means !) less expensive than those at Store Y. Friend B, however was more explicit and informed her that on the average a grocery item cost 50 Satang more at Store Y. If the housewife was to listen to Friend A only, she would no doubt shop at Store x Following Friend B's suggestion, she would be able to make her own reasonable judgement. If she was in a hurry, carrying other heavy belongings and had to risk her life crossing a busy street of Bangkok to get to Store x just to save a Baht or so on the few things she wanted to buy, she might decide to make a purchase at Store Y. We have always been told that hypothesis testing helps in decision making. The above example shows how a point estimate can be even more helpful.

CONFIDENCE INTERVALS (CIs)

There is no question that the point estimate is a useful statistic. Statistical practice, however, is not as easy as deciding on where to buy groceries. To simplify the following discussion, the statistic "sample mean" will be used to represent the point estimate of interest. In making a point estimate, we first sample a certain number of subjects from the population we want to study. The "population mean" or "true mean" is theoretically unknown. We know that each time we sample the same size of subjects, we do not have to get the same sample mean. The mean that we report is based on only one sampling. If we repeat the study 99 times, the 100 sample means estimated may not be the same at all. Statisticians try to add precision to the only reported point estimate and the CI serves this purpose. The CI makes the point estimate more meaningful.

By definition, a 95% CI is the range of values in which there is 95% chance that the true value, in this case the population mean, will fall. If we sample the population a hundred times, each with an equal number of subjects and calculated its mean and 95% CI, in the long run, 95 out of 100 of such

CIs will include the (unknown) population mean. The 95% CI that we report is, however, based on a single sampling ; so it may be one of the 95 which include the population mean or one of the 5 which do not. Since it may or may not include the true mean, we say there is 95% chance that it will.

The CI is closely related to hypothesis testing. A CI covering the value of no effect or value according to the statement of null hypothesis,[1] in risk estimation corresponds to the p-value at that level of CI being "non-singificant." Conversely, a CI that does not include the value of no effect corresponds to the p-value at that level of the CI being "statistically significant." For example, a 95% CI for a mean difference of -0.9 to 12.1 means that the corresponding p-value is greater than 0.05 ; a 90% CI of 2.0 to 10.5, on the other hand, infers a p-value of less than 0.10. However, a CI should not be used as a substitute for hypothesis testing. Once a CI is calculated, the investigator should report and explain the full meaning of it rather than interpreting it merely as an equivalence of hypothesis testing.

CI for a sample statistic is a function of its standard error, so a CI takes into account both the standard deviation and the sample size.[2] CI thus reflects the variability of the data as well as the study size. A wide CI suggests a small sample size or a highly variable set of data. CI also conveys information on the magnitude of the study result on the effect estimate. For example, reporting a relative risk (RR) (Ho : RR = 1) becomes more meaningful when accompanied by a CI whether that be 2.0 - 3.6 or 1.0 - 5.7 for an RR of 2.8 (arbitary) infers quite a different outcome. The first CI tells us that the size of the risk is fairly large, in the range of 2 to 3.6 and is likely to be above 1. The latter CI, however, shows that the estimate is compatible with an even larger effect (upper bound = 5.7) as well as no effect (lower bound = 1.0). The RR of 2.8 with the former CI thus indicates a stronger effect than that with the latter.

A CI is wider at a higher degree of significance. A 99% CI is thus wider than a 95% CI for the same point estimate. This is self-explanatory because in order to have a higher degree of confidence, we must allow for the interval to include more extreme values, both at the lower and the upper ends.

The term "confidence limits" is sometimes used to refer to the CI. It has been suggested, however, that "confidence interval" is a preferable term since "limits" suggests that values smaller or larger are not possible and thus contradicts the actual meaning of CI (Gardner and Altman, 1986).

Presentation of the CI is not suggested here to replace p-values. In fact, the actual p-value does convey information in addition to the CI. Therefore the point estimate, the CI, as well as the p-value should be reported (Gardner and Altman, 1986). The need for hypothesis testing in medical research still certainly exists. Interested readers should refer to the summary of valid applications of tests of significance in epidemiological research by Fleiss (1986).

SUMMARY

It has been generally accepted that presenting a study result as a point estimate and a CI is more advantageous than the conventional way of reporting a p-value alone. CI adds precision to the point estimate and reflects both the variability and the size of the data. It is hoped that this paper will help researchers in Thailand to become acquainted with the use of CIs and to choose to report them when appropriate. Perhaps it is time to start giving our readers more "freedom to think" by presenting to them the study results with CIs instead of telling them what they must think by reporting just the p-values.

---

[1] *For a risk difference (RD) this value is equal to zero ; for a relative risk (RR), it is equal to one.*

[2] $SE = SD/\sqrt{n}$, where SE = standard error,
SD = standard deviation and n = sample size.

REFERENCES

Anomymous. 1987. Report with confidence, editorial. Lancet **1** : 488.

Bulpitt CJ. 1987. Confidence intervals. Lancet **1** : 494-497.

Fleiss JL. 1986. Significance tests have a role in epidemiologic research : reactions to AM Walker. Am J Public Health **76**(5) : 559-560.

Gardner MJ, and Altman DG. 1986. Confidence intervals rather than p-values : estimation rather than hypothesis testing. British Med J **292** : 746-750.

Langman MJS. 1986. Towards estimation and confidence intervals, editorial. British Med J **292** : 716.

Morris JA, and Gardner MJ. 1988. Calculating confidence intervals for relative risks (odds ratios) and standard ratios and rates. British Med J **296** : 1313-1316.

Rothman KJ. 1986. Modern Epidemiology. Little, Brown and Company, Boston.

Woolson RF. 1987. Statistical methods for analysis of biomedical data. John Wiley & Sons, Inc, New York.