OSIR, March 2017, Volume 10, Issue 1, p. 22-26

Outbreak, Surveillance and Investigation Reports

Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625901734-5, Fax: +6625918581, Email: osireditor@osirjournal.net, http://www.osirjournal.net

The Grammar of Science: Are You Confident to Say So?

Jaranit Kaewkungwal*

Mahidol University, Thailand

* Corresponding author, email address: jaranit.kae@mahidol.ac.th

The Grammar of Science

The Grammar of Science is a book written by Karl Pearson and was first published in 1892.[1] It is the book that was read and had impact on young Albert Einstein in creating many greatest scientific theories. In the first chapter, Pearson wrote about definitions of science while explaining about requirements and inquiries to be scientific in nature. I like one of the Pearson's definitions regarding distinctive features of scientific method - discovery of scientific laws by aid of the "creative imagination" and "self-criticism".[1] Later on, Pearson had a classic quote "Statistics is the grammar of science." What does he mean by the word "Grammar"? I opened up an online Oxford dictionary and one of the definitions of "grammar" is "the basic elements of an area of knowledge or skill"[2]. Thus, this has become the name of this column.

We will take a look at basic elements in doing research covering research methodology, epidemiology and statistics. There are times that we take it for grant, thinking that we know this and that, and then explain it the way that we think it is or should be. But we sometimes forget the origin or even the true definition or meaning of the terms that we use. Several authors will take turn writing up in this column with the expectation to reflect "back to basics" of what have been commonly used among researchers.

References

1. Pearson K. The Grammar of Science, Dover 2004 edition. New York: Dover Publications Inc; 2004.
2. Oxford online dictionary. "Grammar" [cited 2016 Nov 10]. <https://en.oxforddictionaries.com/definition/grammar>.

Are you confident to say so?

I would like to start the column with the concept of "confidence" in statistics. I just bought a new book, "A Field Guide to Lies and Statistics"[1] and enjoyed reading it a lot. The author started his chapter one that – because it is about numbers so statistics seems to represent hard facts given to us by nature. But - is it so? The argument is that – it is people who decide what to count, how to go about counting, how to group or analyze the numbers, and how to describe, present and interpret them. So statistics are not facts – they are interpretations! I agree with the author. Back to my first question – how do you interpret the numbers that you see in your study results? In the other word – how confident you are to claim that numbers are the facts in nature?

First of all - Back to basics

When we conduct a research, we do not have to collect data from the entire "population". We simply collect data from "samples" with expectation that they are good representatives of our population of interest and we have enough sample size to estimate the value that could be in that population. We hope that we can generalize or infer the value from samples, so-called "statistics", to the value in the population, so-called "parameter". That is why the statistics that we learned is called "Inferential Statistics". (Additional note: We usually use Greek symbol for "parameter" like μ σ ρ π to represent value that we never know (because we hardly or never collect data from the whole population) and we use English symbol for "statistics" like μ σ ρ π to represent the value that we know (because it comes from the samples that we collect by ourselves)[2-5].

What is "parameter estimation"?

When the researchers want to estimate the value in population from the value that they get from the samples, this is called "parameter estimation". For example, researchers want to estimate the mean score of quality of life among the patients with cancer stage 3 (μ), they do not have to collect data from all cancer stage 3 patients in the whole world or from all patients in the hospital, but simply collect data from the random or representative samples of the patients at that stage and get the sample statistics as ($\bar{X}$ and SD). Then they can estimate μ from that $\bar{X}$ and SD.

What we usually see as the estimate of the parameter is not only a single value, so-called "point estimate" but also the "interval estimate", also-called the "confidence intervals" (CI) around the value[2-5]. For example, say when analyzing an estimate of mean score for quality of life in a sample of 100 patients with cancer stage 3 we produce a mean result of 30 and SD of 5. From these statistics we can calculate a 95% confidence interval of +/- 1.96 (SE) for the population mean estimate. Our point estimate is 30 and interval estimates presenting as confidence interval is (30-1.96x0.5) to (30+1.96x0.5), or we can say that the confidence interval is (29.02 to 30.98)[6-8].

So - What is a "confidence interval"?

A confidence interval or CI is defined as a range of values that describes the uncertainty surrounding an estimate[6-8]. In the "Biostatistics for Dummies"[9] defines it in simple words informally that a CI indicates a range of values that's likely to encompass the true value in population; and a more formally as a specified chance of surrounding (or "containing") the value of the corresponding population parameter. The interval represents by two numbers as lower and upper bounds or limits of the confidence interval; sometimes they are written as CI_L and CI_U, respectively.

It should be noted that the confidence interval itself is also an estimate from the samples in our study as it depends on how we do sampling, measuring, and modeling the numbers that we collected. It could be said that confidence interval is the uncertainty between the true value of what we are estimating and our estimate of that value[6].

How do we calculate confidence interval?

The most commonly used term in research report is "95% Confidence Interval" or "95% CI". In fact, you can see that 95% CI is reported along with different parameter estimates, say 95% CI for mean, proportion, relative risk (RR), odds ratio (OR) and several others. (Note that there might be some studies reporting other level of CI such as 90% CI or 99% CI.) In general, we can interpret 95% CI around any estimate somewhat the same way. But let's look into basic concept from the 95% CI of mean as an example.

When we conduct a study to estimate mean in population (μ), we draw a sample and calculate $\bar{X}$ and SD. What we get are only values from that sample. The question is - will the value that we get from that sample be the value in population? It may or may be not, and most likely maybe not. Now assume that if we can repeat the study again and again, we will get several samples from the same population and get several $\bar{X}$s and SDs. The distribution of different $\bar{X}$s is called sampling distribution as the scatter of $\bar{X}$s is due to sampling that we keep repeatedly doing it. Thus, we can calculate the "mean of the means" (mean of $\bar{X}$s = $\bar{\bar{x}}$). The $\bar{\bar{x}}$ could be said as the estimate of μ. The distribution of $\bar{X}$s around the μ (or $\bar{\bar{x}}$) is thus called "standard error" (SE). But in real life, we never conduct the study again and again, so we simply say that the estimated μ is the $\bar{X}$ that we get from our one time sample. And the SE is also estimated from the "standard deviation" (SD) that we get from that sample as relative to the sample size (n). The simple formula in this case is: $SE_{\bar{X}} = \frac{SD}{\sqrt{n}}$. Based on the concept of area under normal cure, the cut offs for the middle 95% area under curve is +1.96 (we may revisit this concept of area under curve at some other time). Thus, the 95% CI of the mean estimate is usually reported around $\bar{X}$+ 1.96*SE. As shown in Figure 1 - an example of the estimate of mean[2-5].

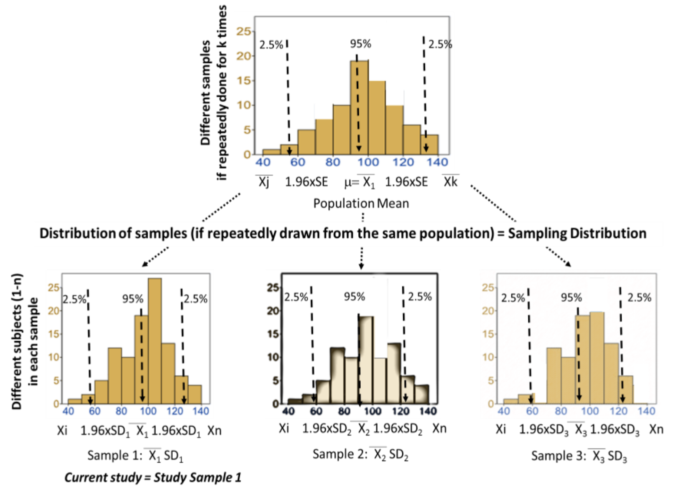

Figure 1. Estimation of μ in population from the $\bar{X}$ and SD of a sample

Similarly, we can calculate SE for different other statistics. For example, to estimate proportion of HIV infection among teenagers (π), the researchers collect data among a sample and get a proportion (p). Then estimate π from p; and they will have to estimate SE of p from the formula $SE_{(P)} = \sqrt{p(1-p)/n}$ and then report the 95% CI of the proportion estimate around p ± 1.96*SE[10].

Estimation of other statistics which is not a single parameter estimate also follows the same algorithm. For example in the estimate of confidence interval for the difference in means (μ1-μ2) from two independent samples, the CI of the difference could be $(\bar{x}_1-\bar{x}_2) \pm z\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ where z is the confidence level desired (it does not have to be fixed at 95% or 1.96) and Sp is the pooled estimate of the common standard deviation, $S_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$. Another example, in estimating a risk ratio (RR) or prevalence ratio (PR) from two independent samples, RR = p_1/p_2, the CI for RR could be calculated as $Ln\left(\widehat{RR}\right) \pm z\sqrt{\frac{(n_1-x_1)/x_1}{n_1}+\frac{(n_2-x_2)/x_2}{n_2}}$ and then antilog or take exp[lower limit of Ln (RR)] and exp[upper limit of Ln (RR)] to get the CI_L and CI_U for RR. Similarly, the CI for an odds ratio (OR) can be calculated from $Ln(OR) \pm z\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. Note that these are formulas for larger samples[11-12].

How do we interpret a confidence interval?

The true value for the population does exist and it is a fixed number, but we just do not know exactly what it is. Although we may conduct a perfect study collecting data from the samples that are well (or even perfect) representatives of the population; the very good estimate of the value in the population that we get from our sample may not be the exact value of the population parameter[11-13]. However, we want to be somewhat certain about the value that we get from our sample so that we can say or make inference about the population value. That is, CI allows us to say what the true value in population could be[13]. In other words, we may simply explain that if we can repeat the studies many times, 95% percent of the CIs would contain the true population mean[14-16]. As shown in figure 2, the true value in population, μ does exist but we do not know; however, if we repeated the studies in the same population again and again 100 (or 20 in figure 2) times, our 95% confidence interval generated from each sample will cover μ in 95 studies (95/100 or 19/20) but we may miss that true value for about 5 times (5/100 or 1/20)[11,15,16].

Back to the example of the estimation of mean score for quality of life in patients with cancer stage 3, suppose the true μ is 29.67; and from a sample of 100 patients with cancer stage 3 we have got a mean result of 30 and SD of 5. For these estimates we can calculate a 95% CI as: (30-1.96*0.5) to (30+1.96*0.5), or we can say that the 95% CI is (29.01 to 30.99).That would mean this range of 95%CI does cover the true μ of 29.67.And if we repeat the studies again 100 times, 95% of the times the ranges would still cover 29.67. The interpretation of a 95% CI as indicating a range within which we can be 95% certain that the true population parameter lies is a loose interpretation, but is useful as a rough guide[17]. The strictly-correct interpretation of a CI is based on the hypothetical notion of considering the results that would be obtained if the study were repeated many times; and if a study were repeated infinitely often, and on each occasion a 95% CI calculated, then 95% of these intervals would contain the true value in population [8,14,17].

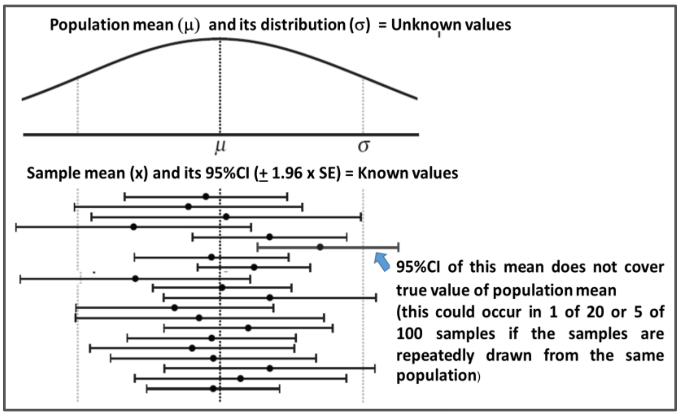

Figure 2. Estimation of population mean with 95% confidence

Confidence interval and p-value

When the study compares outcomes of different groups, the report could be presented with an estimate of the difference (say mean difference, risk difference, risk ratio, odds ratio, hazard ratio) and its CI along with p-value. Some studies, however, skip CI or p-value. In fact, there is logical correspondence between the CI and the p-value. In general, the 95% CI for the estimate will exclude the null value (i.e., null for RR, OR or HR is 1.0; and null for mean difference or risk difference is 0) if and only if the test of significance yields a p-value < 0.05; and either the upper or lower limit of the 95% CI will be at the null value if the p-value is exactly 0.05[15,17,18], given that the 95% CI and p-value are both calculated from the same method.

Back to our example in an estimation of risk ratio between teenagers and adults in getting infection with HIV, suppose the RR=3.2 and the 95% CI is shown as (0.8 to 5.4); that would mean the true risk ratio between the populations of teenagers vs. adults might not be 3.2 but could be somewhere in this range of 0.8 to 5.4. Since 95% CI includes 1, we will also see that p-value >0.05; thus we cannot conclude that there is a statistically significant risk difference between the two groups. If you want to interpret from the 95% CI without looking at the p-value (which the researcher may decide not to present), we could still say that the risk ratio is not absolute and not significant. In our sample we found that the teenagers have higher risk than adults (3.2 vs. 1) but the estimates of the true risk ratio in population could be that the teenagers have lower risk (0.8 vs. 1) or they may have even higher risk (5.4 vs. 1). In contrast, suppose the results from the same study show the estimate of RR=3.2 and 95% CI (1.9 to 4.5). Since 95% CIexcludes 1, we will also see that p-value <0.05; thus we can conclude that there is a statistically significant risk difference between the two groups. If you want to interpret from the 95% CI without looking at the p-value, we could say that the risk ratio is absolutely shown in one direction. In our sample we found that the teenagers have higher risk than adults (3.2 vs.1) and the estimates of the true risk ratio in populations of the two groups would always be that the teenagers have higher risk which might be not at (3.2 vs. 1) but could be as low as (1.9 vs. 1) or as high as (4.5 vs. 1).

What is "good" or "not good" CI estimates?

CI could technically tell us how "good" an estimate is; it is an important reminder of the limitations of the estimates such that the larger a CI for a particular estimate, the more caution is required when using the estimate.[6,7,19] As CI represents margin of error (or the width of the interval), a larger margin of error (wider interval) is indicative of a less precise estimate[12,15,19]. As an example, in an estimation of risk ratio between teenagers and adults in getting infection with HIV, suppose the RR=3.2 (i.e., teenagers are more likely to get infected 3.2 times than adults) and the 95% CI is shown as (1.5 to 60.7); that would mean the true risk ratio in the populations of teenagers and adults might not be 3.2 but could be somewhere in this range which is so wide.

The width of the CI of a study is usually related to the sample size; study with large sample size tends to give more precise estimates (or narrow CI)[13,17,19]. For the estimate of continuous variable, the CI might depend on the variability (or SD); but for the estimate of dichotomous variable, it depends on the chance (or proportion) of the event that could occur; and for the estimate of time-to-event outcome, it depends on the number of events observed[17].When the CI is wide, there are a number of methods we can use to reduce it. In attempt to improve the precision of our results (having narrower CI), we could increase our sample size (if possible),[5,8,11,13]. However, as larger sample sizes would result in narrow CI, but if you increase the sample size to a certain number then it won't help that much anymore. As shown in one reference, increasing the sample size from 100 to 500 reduces the CI from 9.8 to 4.3, but when sample size is 1,000, the CI will reduce down to only to 3.1 which may not worth doing it, comparing to what you have to collect the data from 1,000 rather than 500 subjects[13].

In the study that compares the outcomes between groups, when the estimates come with a wide CI, it may not be that the sample size is too small but it may indicate that the underlying data are disparate, including too few events occurring in one group or another or both, or too many outliers and oddball data points[20]. For example, in an estimation of risk ratio between teenagers and adults in getting infection with HIV, suppose the RR=3.2 and the 95% CI is shown as (1.5 to 132.6); that would mean the true risk ratio between the populations of the two groups could be somewhere in this wide range. If this is the case, the researcher should not emphasize this statistically significant result that much even though we may have a large enough sample size in total but it might be that we have too few subjects in one group or another, or there might be too few infection incidences relative to the sample sizes of one of both groups. In fact, when this wide range is shown, the researcher should look back at the descriptive information about the two groups. It may help explain why so.

So, the question then is - how wide is too wide? As a rule of thumb, the researcher should be cautioned to oneself and to the readers of that study results if a CI is wider than the magnitude of the estimate[20]. For example, when you see a RR=3.2 and the 95% CI (1.5-132.6), the width of the CI thus is 131.1 which is too much higher than the size of the RR. But when you have narrower CI, say RR=3.2 and 95% CI(1.9 to 4.5), thus the width of the CI is 2.6 which is a fraction of the size of the RR; then one can be quite confident in the population estimate.

Final words – how confident you are to interpret your estimate(s)?

The confidence interval tells you more than just the possible range around the estimate but it also tells you about how stable the estimate is[21]. A stable estimate means that the value that you claim in your study result section is one that would be close to the true value in population that we never know. Wider CI in relation to the estimate itself indicates instability and less precision of your estimate. One of the nice things about presenting the estimate with 95% CI is that you never have to commit yourself 100% on anything in statistics. Claiming 100% confidence is impossible anyway since we do not conduct the study in the whole population. A classic quote (or joke?) about statistics is that "Statistics mean never having to say you are certain". This is quite right as you can always claim “I am under the 95% confidence limit”.

Suggested Citation

Kaewkungwal J. The grammar of science: are you confident to say so? OSIR. 2017 Mar;10(1):22-26.

References

1. Levitin D. A field guide to lies and statistics. UK: Penguin Random House; 2016.
2. Bernard R. Fundamentals of biostatistics. 5th ed. Duxbery: Thomson learning; 2000. p. 384-5.
3. Gardner WP. Statistics for the Biosciences NY: Prentice Hall Inc;1997.
4. Everitt B. Medical statistics: from A to Z. Cambridge: Cambridge University Press; 2006.
5. Daly LE, Bourke W, McGilvray J. Interpretation and uses of medical statistics. 4th ed. London: Blackwell Scientific Pub; 1991.
6. United State Census Bureau. A basic explanation of confidence intervals [cited 2016 Nov 10]. <https://www.census.gov/did/www/saipe/methods/statecounty/ci.html>.
7. Kalinowski P. Association for psychological science. Understanding confidence intervals (CIs) and effect size estimation [cited 2016 Nov 10]. <http://www.psychologicalscience.org/observer/understanding-confidence-intervals-cis-and-effect-size-estimation#.WJr3YW996cM>.
8. Cumming G, Finch S. A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. Educational and Psychological Measurement. 2001;61:530-72.
9. Pezzullo J. Biostatistics for dummies. New Jersey: John Wiley & Sons, Inc; 2013.
10. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New Jersey: John Wiley & Sons; 2003.
11. Ellis PR, Brumby PJ. The epidemiological approach to investigating disease problems [cited 2016 Nov 10]. <http://www.fao.org/Wairdocs/ILRI/x5436E/x5436e06.htm#>.
12. Sullivan L. Confidence intervals. Boston University School of Public Health [cited 2016 Nov 10]. <http://sphweb.bumc.bu.edu/otlt/MPHModules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals_print.html>.
13. Scottish Health Statistics. Confidence intervals [cited 2016 Nov 10]. <http://www.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey/ConfidenceIntervals>.
14. Cumming G, Finch S. Inference by eye: confidence intervals, and how to read pictures of data. American Psychologist. 2005;60:170-80.
15. Cumming G, Williams J, Fidler F. Replication and researchers’ understanding of confidence intervals and standard error bars. Understanding Statistics. 2004;3:299-311.
16. Fidler F. From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology. Department of History and Philosophy of Science, University of Melbourne. 2005 [cited 2016 Nov 10]. <http://www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf>.
17. Higgins JPT, Green L. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. March 2011 Mar [cited 2016 Nov 10]. <http://handbook.cochrane.org/chapter_12/12_4_1_confidence_intervals.htm>.
18. Krzywinski M, Altman N. Points of significance - error bars. Nature Methods. 2013;10(10):921-2.
19. Cochran S. Introduction to statistical reasoning. UCLA School of Public Health [cited 2016 Nov 10]. <http://www.stat.ucla.edu/~cochran/stat10/>.
20. Debunkosaurus. How to evaluate clinical trial [cited 2016 Nov 10]. <http://www.debunkosaurus.com/debunkosaurus/index.php/How_to_evaluate_a_clinical_trial>.
21. Department of Health. New York State. Confidence intervals - statistics teaching tools [cited 2016 Nov 10]. <https://www.health.ny.gov/diseases/chronic/confint.htm>.