Sanni Ali, Sileshi Lulseged, Tewodros Getinet, Girmay Medhin. *Ethiop Med J, 2019, Vol. 57, No. 1*

TEACHING ARTICLE

EMJ SERIES ON STATISTICS AND METHODS PART V: ESTIMATION WITH CONFIDENCE: CONFIDENCE INTERVALS MADE SIMPLE

Sanni Ali, DVM, MSc, PhD[1,2,3], Sileshi Lulseged, MD, MMed[4]*, Tewodros Getinet, MSc[3], Girmay Medhin, MSc, PhD[5]

SUMMARY

Confidence intervals have been routinely reported in medical journals since the mid-1980s; however, their interpretation is not well understood among clinical researchers with limited statistical training. In this article interpretation of Confidence intervals are clearly explained and supplemented with an example for a single sample using binary and continuous outcomes.

INTRODUCTION

In the previous teaching article on "Population and Sample" (1), It was discussed that the basic idea of sampling is to draw inference about the population of all individuals from which the sample is drawn. The interest in clinical research is primarily the population, not the sample. For example, one might be interested in estimating mean systolic blood pressure measurements (in mmHg) or proportion of individuals with type 2 diabetes mellitus (T2DM, as percentage) in a population of hypertensive patients. A well chosen sample will contain most of the information about the population, hence called representative sample, such that true (valid or unbiased) inferences about the population can be made. However, if the sample is not representative of the population, then the inferences drawn about the population may be misleading even if the sample size is large and statistical procedures cannot help to make any adjustment (2).

Statistical inference in medical research often revolves around hypothesis testing and parameter estimation. Hypothesis testing, sometimes called null hypothesis significance testing (NHST), refers to the formal statistical procedures used to reject or accept (preferably, fail to reject) a statistical hypotheses, an assumption about a population parameter. For example, one might be interested to test a hypothesis that states "the mean systolic blood pressure measurement (in mmHg) in a certain population is 120 mmHg". On the other hand, parameter estimation refers to the process by which inferences about population parameters such as mean or proportions are made, based on information obtained from a sample drawn from the population. It is important to underline that estimation can also be used, and in fact strongly recommended in most medical journals as opposed to significance testing using p-values (2,3).

Both hypothesis testing and parameter estimation require drawing a random sample from the population of interest for practical reasons. If the research question involves comparison of two populations, for example comparison of clinical outcomes such as mean systolic blood pressure measurements (in mmHg) between treated and untreated hypertensive patient populations, the researcher needs to draw two different random samples from respective populations. Hence, hypothesis testing in this example involves rejecting or not rejecting the assumption that there is no difference in mean systolic blood pressure measurements (in mmHg) between treated and untreated populations. Alternatively, this hypothesis can be stated as follows: the difference in the mean systolic blood pressure measurements (in mmHg) between the two groups is zero. Detailed discussion of hypothesis testing using p-values and confidence intervals will be covered in the next issue of the Ethiopian Medical Journal (EMJ).

[1]Faculty of Epidemiology and Population Health, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK; [2]Center for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.; [3]Public Health Department, St. Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia. [4]Department of Pediatrics and Child Health, School of Medicine, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia; [5]Aklilu Lemma Institute of Pathobiology, Addis Ababa University, Addis Ababa, Ethiopia.

In the above example, the research interest could be estimation of mean systolic blood pressure measurements (in mmHg) in each of the two populations (treated and untreated) or estimation of the difference in the mean systolic blood pressure measurements between the two populations, hence, estimation of two means or difference in two means, respectively. This procedure of estimation may be expressed using a point estimation or an interval estimation. And a value obtained from a point estimation is called point estimate and values obtained from interval estimation is called interval estimates.

A point estimate of a population parameter such as population mean or population proportion is a single value of the sample statistic, sample mean or sample proportion, respectively. For example, the sample mean of systolic blood pressure measurement (in mm Hg) $\bar{x}$ is a point estimator of the population mean of systolic blood pressure measurement (in mm Hg), μ. Similarly, for binary outcome variables such as presence of T2DM, the sample proportion (of individuals with T2DM) p is a point estimator of the population proportion π. The values that the point estimators $\bar{x}$ and p assumed are called point estimates.

On the other hand, an interval estimate is defined by two numbers, between which a population parameter is said to lie. In the systolic blood pressure measurement example, $a < \bar{x} < b$ is an interval estimate of the population mean μ, where $\bar{x}$ is the sample mean systolic blood pressure measurement (in mm Hg). The interval estimate can also be expressed as $\bar{x} \pm c$, where $a = \bar{x}$-c and $b = \bar{x}$+c are the lower and upper interval limits within which the population mean μ lies with a given level of confidence. When this interval is associated with a certain level of confidence, it is called confidence interval (we will come back to this later).

Variation between samples

Theoretically, one can draw several random samples from a defined population. In practice, however, researchers make estimation of parameters based only on one random sample of a given size unless it involves comparison of two or more populations as described before. A well designed simple random sampling assures that every member in the population has equal chance of being included in the sample particularly if one can enumerate every member of the population to construct a sampling frame. Even in this sampling procedure, series of samples drawn from this population will not be identical because of random variation in the samples. Hence, parameter estimates from a single sample are subject to uncertainty due to sampling variability (2). This uncertainty in the point estimate using one sample can be communicated using interval estimates. Hence, in research both point estimates and interval estimates are used in combination.

Whether parameter estimates from a series of samples are close to the truth (unbiasedness or validity) is determined by how representative the samples are of the population and it is less influenced by the size of the sample. On the other hand, the sampling variability, which describes how close to each other (precision) the parameter estimates from several samples drawn from a population are, is strongly inversely related to sample size and directly related to the amount of variation between individuals in the population. The sampling variability decreases with increasing sample size or decreasing variation in the population, and increases with decreasing sample size or increasing variation in the population.

Note that the true population parameter, for example the mean systolic blood pressure measurement (in mm Hg) μ of medical professionals working in Afar Regional State of Ethiopia, is a fixed but unknown quantity. This quantity is estimated from a random sample of size n using sample statistic $\bar{x}$. These two quantities, the true population parameter μ and sample statistic $\bar{x}$, will not necessarily be identical. The difference between the sample statistic obtained from random sample and used to estimate a population parameter and the true but unknown value of the parameter is called sampling error. Information on the amount of the sampling error is incorporated when intervals estimates are reported compared to point estimates. This is the rationale for using confidence interval rather than merely reporting point estimate (2,4).

CONFIDENCE INTERVALS

Now one might wonder why such an interval estimate is called a "confidence interval", CI for short, often reported as 95% CI and rarely 99% CI. As defined before, an interval estimate is the range of values which is "likely" to include the population parameter of interest. But one cannot be 100%" confident" or "certain" (note the terms "confidence interval" and "uncertainty") that estimate from a single random sample describes the population parameter precisely. This holds true even with several random samples. In other words, the interval estimate would not always include the population parameter between them. Hence, they should be constructed at a certain level of confidence. With serial random samples, some interval estimates would include the population parameter while others would miss it. The likelihood for the interval estimate to contain the true population parameter is often described as a percentage of confidence such as 95% CI; 95% is the level of confidence set. Confidence level is often denoted as 100*(1-*á*)% where *á* refers to the significance level (will be discussed in detail in the next issue on "Hypothesis testing") set at 5% for a 95% confidence level.

Theoretically, the statistical explanation of confidence intervals is based on repeated sampling from the same population. Assume that one draws 100 random samples using the same sampling method from a population (note that 100 is chosen to describe the level of "confidence" as percentage) and computes interval estimate in each of the 100 samples. This results in 100 individual statistic (i.e. means) and their corresponding interval estimates; the true population mean would fall within the intervals 95% of the time. In other words, out of the 100 intervals estimated, 95 of them would include the true population parameter (i.e. mean in our example) and five of them would miss it; 99% confidence interval means that 99% of the intervals contain the true population parameter; and so on.

Practically, however, researchers rely only on one random sample (and not repeated random samples) to compute confidence interval for the population parameter. Hence, the interpretation of confidence interval is as follows: the researcher is 95% confident that the true population parameter (i.e. mean in the systolic blood pressure example) will fall within the range of estimated interval. In simple terms, the estimated interval from one random sample is one of the 95 intervals which contain the true population parameter (e.g. mean) and is not one of the five intervals which miss the true population parameter (e.g. mean).

Note that a 95% CI does not mean that there is a 95% probability that the interval contains the true population parameter. The interval computed from a given random sample either contains the true population parameter or it does not; there is no probability associated with it because the population parameter is a fixed but unknown quantity. The two limits of a confidence interval are random quantities that vary from one random sample to another random sample of the same size taken from the same underlying population. However, the population parameter from which the random samples are selected is fixed quantity, which is not known to the researcher unless each and every member of the population is investigated and the parameter value is determined.

Components of a confidence interval

A confidence interval consists of three parts: a confidence level, a statistic, and a margin of error. The confidence level, 100*(1-α)%, describes the uncertainty of a sampling method and it is the probability part of the CI expressed as percentage. It describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter; it is not affected by the margin of error but set by design often at 95% meaning that $\alpha = 5\%$. It is important to distinguish interval estimates and confidence intervals as they may not be the same; both have a margin of error associated with the point estimate as in formula 1 but confidence intervals are always described using a certain confidence level. Research finding with higher confidence level (say, 95%) are more convincing to readers than findings with low confidence level (say, 90%).

$$\text{sample statistic} \pm \text{margin of error} \quad (1)$$

By central limit theorem, the sequence of the 100 sample means computed from the 100 samples (of size ≥ 30) conform to a Normal distribution, even if the observations from which they were obtained do not (5). The distribution of the 100 sample means is the sampling distribution of the mean and can be described using z-distribution (t-distribution if the sample size is < 30 and the population standard deviation is unknown) as discussed in the previous teaching article (4). One can estimate mean of the 100 sample means (which equals the true population mean) and the variation among the sample means. This variation between the sample means is what we call the standard error of the mean in case of one sample mean estimation. Note that standard deviation describes the variation of observations in a single sample of a give size where as standard error describes uncertainty in the sample mean in estimating the population mean. The standard error of the sample mean obtained from one sample can also be thought of as the estimate of the standard deviation that would be obtained from the means of large number of repeated samples (100 samples in our example) drawn from that population.

The interval estimate of a confidence interval is defined by the sample statistic and margin of error, the range of values above and below the sample statistic that describes the precision of the estimate. When the sample size is < 30 and the population standard deviation is unknown, we use t-distribution to compute CI for population as;

$$\text{sample mean} \pm \text{t-multiplier} * \text{standard error} \quad (2)$$

In this case the standard error of the sample mean $\bar{x}$ is estimated using standard deviation of the sample and sample size n as $S/\sqrt{n}$

$$\bar{x} \pm t_{á/2,n-1} * S/\sqrt{n} \quad (3)$$

where the "t-multiplier" which is denoted as $t_{á/2,\,n-1}$, depends on the sample size through n-1 (called the "degrees of freedom") and the confidence level $100*(1-á)\%$ through $á/2$. And it can be obtained from a t-distribution table.

When the population variance is known or the sample size is large ($n \geq 30$), we use standard Normal distribution (z-value which does not depend on sample size, unlike t-distribution and can be obtained from a Z-distribution table) to compute CI for population as;

$$\text{sample mean} \pm \text{z-multiplier} * \text{standard error} \quad (4)$$

$$\bar{x} \pm Z_{\alpha/2} * S/\sqrt{n} \quad (5)$$

$$\bar{x} \pm Z_{á/2} * \sigma/\sqrt{n} \quad (6)$$

In this case the standard error of the sample mean $\bar{x}$ is estimated as in formula 5 when $n \geq 30$ and population variance is unknown or as in formula 6 when population variance is known.

For binary outcomes such as prevalence of T2DM, the parameter estimates are population proportions and the CI for population proportion. In the previous teaching article on Normal distribution (4), we described that in repeated sampling, the mean of the sample proportions would be approximated by a Normal distribution, hence we use z-values in formula 8. If r is the observed number of individuals with T2DM in a random sample of size n , then the estimated proportion is p = r/n and the proportion of individuals who do not have T2DM is q=1-p.
The standard error of the sample proportion p is estimated using the sample proportion itself and sample size n as;

$$SE(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} \quad (7)$$

And the $100(1-\alpha)\%$ confidence interval for the population proportion is computed using the following formula,

$$p \pm Z\alpha/2 * \sqrt{\frac{pq}{n}} = p \pm Z_{\alpha/2} * \sqrt{\frac{p(1-p)}{n}} \quad (8)$$

The values of $Z_{\alpha/2}$ can be found from statistical tables; for a 95% confidence interval $Z_{\alpha/2}$ is 1.96. Note that formula 8 is based on an approximation and should not be used for very low observed proportions. As a rule of thumb neither r nor n-r should be less than 5. When these assumptions are violated, there are other ways to compute confidence intervals for proportions (2,6).

Confidence intervals are preferred to point estimates for several reasons: they are readily interpretable, linked to familiar statistical significance tests, and also provide magnitude of the effect can encourage meta-analytic thinking, and give information about precision (3).

When comparing two populations, for example with respect to outcome, confidence intervals convey only the effects of sampling variation on the estimated means or proportions and their differences and cannot control for other non-sampling errors such as biases introduce because of in study design, during the conduct, or analysis. With large samples, you know the mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample. A narrow CI justifies the confidence we have in the reasonably precise knowledge about the effect under study.

Worked-out examples

I - Confidence interval for means

Nigussie et al. (7) conducted a case-control study, recruiting study participants using systematic random sampling, in Tikur Anbessa Specialized Hospital from Jun 2013 to March 2014 to determine serum levels of β-hCG in normotensive and preeclamptic pregnant women. Cases (n=38) were preeclamptic pregnant women and controls (n=38) were normotensive pregnant women identified by using their blood pressure level from the mother's record card at ANC clinic. Characteristics of preeclampsia and control study participants are summarized in Table 1 of the article (7). They reported higher level of Serum β-hCG in the preeclamptic group (34439.18±28223.67mIU/ml) than the normal group (20582.00± 17588.31 mIU/ml), and the mean difference was statistically significant (p=0.013). Note that the authors reported only point estimates (sample means) with standard deviations in each group. They did not compute standard errors of the sample means and confidence intervals.

For this exercise, we will focus on the point and interval estimates for both groups without comparing them. Since the outcome variable, serum β-hCG level, is measured on a continuous scale (in mIU/ml), they computed the mean serum β-hCG level in cases and controls to be 34439.18 and 20582.00mIU/ml. If the authors were to take many repeated samples (say 100) from the populations that each group is generated, compute the means, the mean of the 100 sample means would conform to a normal distribution since $n \geq 30$, using the central limit theorem. Hence we can use z-table to obtain a multiplier for the standard error then to compute confidence intervals.

First, the standard errors for the mean serum β-hCG level in the two samples can be calculated using the formula $S/\sqrt{n}$ where S is the standard deviation of the observations in each sample, hence, the SE is $28223.67/\sqrt{38}$ = 4578.49 in the sample of cases and $17588.31/\sqrt{38}$ = 2553.21 in the sample of controls

Then, the margin of error for the sample mean is computing using z-value of 1.96 (for a confidence level of 95%) in each sample. Hence, 1.96*SE = 1.96*4578.49 = 8973.84 in cases and 1.96*2553.21 = 5004.29 in controls.

Finally, the 95% CI for the population means is computed using formula 4/5 as 34439.18 ± 8973.84 (25465.36mIU/ml, 43413.04mIU/ml) in cases and 20582.00 ± 5004.29 (15577.11mIU/ml, 25586.29mIU/ml) in controls, respectively.

In this particular study, the objective was to determine serum β-hCG level in the two groups and not to compare the two groups or test the hypothesis that the mean serum β-hCG level is not different between normotensive and preeclamptic pregnant women. However, they performed comparison and significance testing using p-value; they concluded that serum β-hCG level is higher in preeclamptic pregnant women compared to normotensive pregnant women and this was statistically significant (p=0.013). .

The ideal procedure would be to compute the mean difference in the two groups, construct 95% CI for the difference of the two population means and conduct hypothesis testing using confidence intervals rather than p-values. We will discuss in detail in the next issue on hypothesis testing.

II - Confidence interval for proportions

Woldetsadik and Kumie (8) conducted a cross-sectional study to determine the prevalence of symptoms of asthma and associated factors among primary school children in Addis Ababa. They randomly selected a total of 20 primary schools in Addis Ababa. Using questionnaire adapted from International Study of Asthma and Allergies in Childhood, they collected data from total of 1,259 primary school children aged 6-7 years. The questionnaires were completed by parents/ guardians of the children. Information collected include childhood wheeze, wheeze in the past 12 months, ever diagnosed asthma, exercise induced wheeze in the past 12 months, and dry cough at night in the past 12 months. They reported the prevalence of different symptoms in Table 1, For this example we consider diagnosed asthma (50 out of 1,259 children had diagnosed asthma), the prevalence was reported to be 4.1% (95% CI, 2.98% to 5.22%).

First, prevalence is calculated by dividing the number of children diagnosed with asthma (n1 = 50) by the total number of children included in the study (n = 1,259). This step resulted in the point estimate of 4.1%. To account for uncertainty with this single sample, they calculated 95% CI (they set confidence level at 95% hence assumed 5% error or significance level).
The standard error for proportion p is computed (using formula 7) as follows:

$$SE(p) = \sqrt{(0.041)(1 - 0.041)/1259} = 0.006$$

Using z-value of 1.96 for 95% confidence level, the margin of error is 1.96*0.006 = 0.012. Using this value, the 95% CI for the population proportion is computed using formula 1,

= sample statistic ±margin of error (z-multiplier ∗ standard error) = 0.041± 0.012

This results in a point estimate of 0.041 or 4.1% and 95% confidence interval of 0.0290 to 0.0530 or 2.9% to 5.3%. These figures are similar to the prevalence (and 95%CI) reported by the authors except some rounding errors introduced during calculation.

REFERENCES

1. Ali SM, Lulseged S, Medhin G. EMJ series on statistics and methods: variables, populations and samples. Ethiop Med J 2018; 56(3): 277-283.
2. Altman D, Machin D, Bryant T, Gardner M. Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd ed. John Wiley & Sons, 2013.
3. Cumming G, Finch S. A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are based on Central and Non central Distributions. Educational and Psychological Measurement. 2001; 61(4): 532–574.
4. Ali SM, Lulseged S, Medhin G. EMJ series on statistics and methods: normal distribution and the central limit theorem. Ethiop Med J 2018; 56(3): 285-291.
5. Campbell MJ, Swinscow TDV. Statistics at Square one, 11th ed. John Wiley & Sons, 2011.
6. Newcombe RG:.Confidence Intervals for Proportions and Related Measures of Effect Size: Chapman & Hall/ CRC Biostatistics Series, Illustrated ed. CRC Press, 2012.
7. Ayele A, Zewde T, G/Hiwot Y. Serum level of β-hCG in Normotensive and Preeclamptic Pregnant Women in Tikur Anbessa Specialized Hospital Addis Ababa, Ethiopia. Ethiop Med J 2018; 56(3).
8. Woldetsadik M, Kumie A. Prevalence of Symptoms of Asthma and Associated Factors among Primary School Children in Addis Ababa, Ethiopia. Ethiop Med J. 2018; 56(4).