Sanni M. Ali, Sileshi Lulseged, Girmay Medhin. *Ethiop Med J, 2018,Vol. 56, No. 3*

TEACHING ARTICLE

EMJ SERIES ON STATISTICS AND METHODS: NORMAL DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

Sanni M. Ali, DVM, MSs, PhD[1,2], Sileshi Lulseged, MD, MMed,[2*] Girmay Medhin, MSc, PhD[3]

INTRODUCTION

This second article in the EMJ *Series on Statistics and Methods* dwells on the basics of sampling distribution of' variables', which are presented in detail in the preceding article in this Issue of the Ethiopian Medical Journal (EMJ). The present article highlights recommended routines that need to be undertaken in order to understand information collected in a particular study before embarking on doing complex statistical analyses. It underscores the importance of descriptive statistics as a means to getting insights into data quality and learn about the scale and distribution of different variables in a data set. The article emphasizes the need for assessing the sampling distributions of variables as a prerequisite to making decide on selection of appropriate statistical techniques for in a data set. It describes salient features of a normally distributed random variable and touches on some other probability distributions commonly used in epidemiological studies. The article also describes the central limit theorem highlighting salient points on its conceptual basis in understanding sampling distributions of sample means and the implications of using normal distribution to make inference about the population based on summary measures from a sufficiently large sample.

NORMAL DISTRIBUTION

It is a wise and recommended routine to use descriptive summary measures such as means and proportions as well as various charts and graphs to understand information collected from individual study participants before diving into running statistical analysis of various complexity. This give us insight into the quality of the data to be analyzed, the scale and distributions of different variables included in the data set, the implications of these conditions on the statistical methods to be employed during data analysis, and any other data processing methods including transformation of variables and ways of handling missing data. Many statistical techniques such as t-tests, regression analyses, and analysis of variance require that the outcome variables should follow a distribution of a particular kind (1).

Let us assume a collection of systolic blood pressure measurements (in mmHg) taken on 1,000 individuals attending a medical centre, representing a randomly selected sample from a larger population visiting the medical centre. As an example, we used a computer generated dataset to simulate the 1,000 systolic blood pressure measurements (in mmHg) with a mean of 120 and standard deviation of 12. For a better visual representation of such data, we often use frequency distributions in a form of histograms (Figure 1A) or density plots (Figure 1B-C). In Figure 1, the x-axis represents the actual blood pressure measurements in mmHg and the y-axis represents the frequency or number of occurrences of measurements within each bar (Figure 1A) or the frequency density (Figure 1 B-D). The frequency density is calculated in two steps: first, relative frequency is calculated by dividing the actual counts in each bar by the total number of measurements (n=1,000 in the above example); then, density is calculated by dividing the relative frequency by the width of the bar. In figure 1B-D, the area under the entire histogram (Figure 1A-B) or the curve (Figure 1B-D) equals one. Such a plot is known as the probability density function or pdf (2).

[1,2]Faculty of Epidemiology and Population Health, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK and [2]Center for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
[4]School of Medicine, Addis Ababa University, Addis Ababa, Ethiopia
[5]Aklilu Lemma Institute of Pathobiology, Addis Ababa University Addis Ababa, Ethiopia
* Corresponding author: sileshilulseged@gmail.com

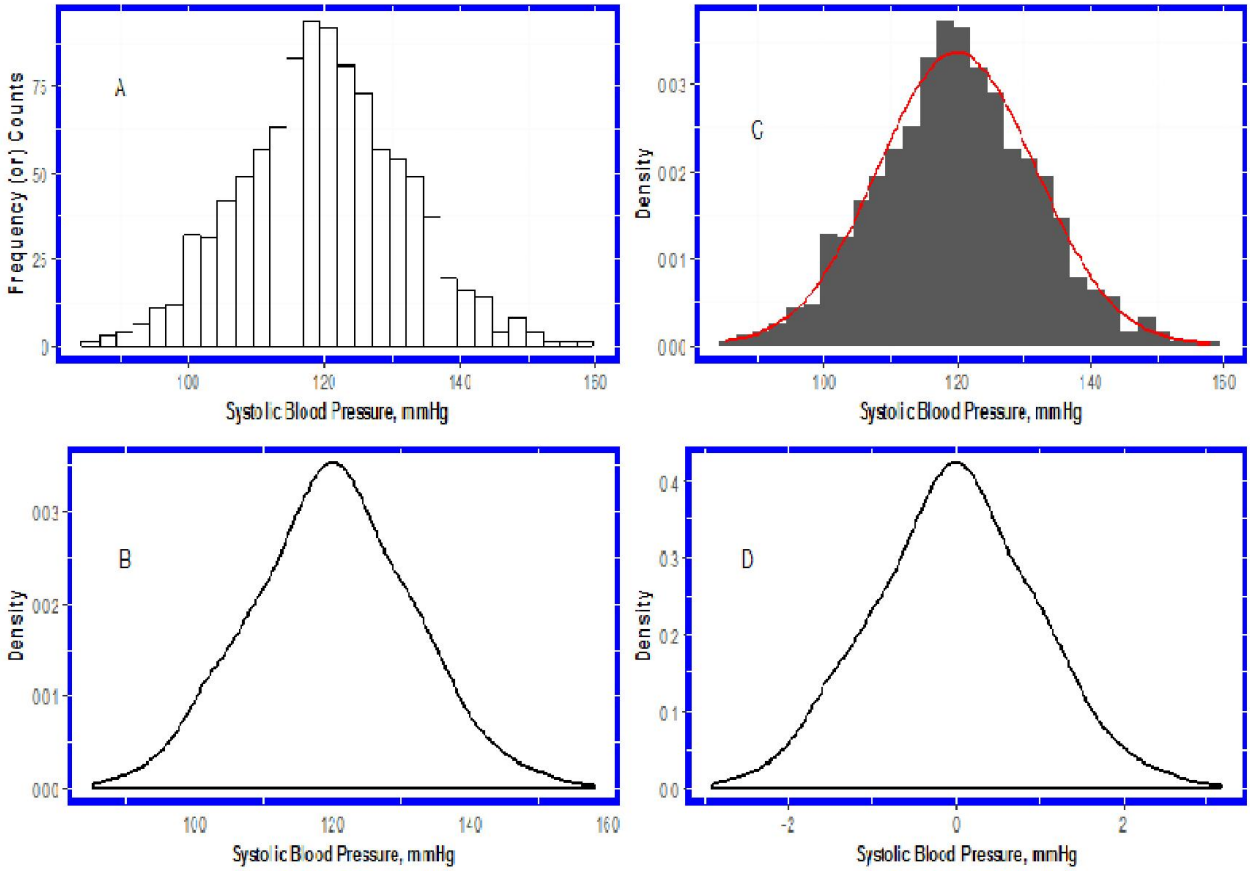

Figure 1. Plots of blood pressure measurements from 1,000 individuals, using frequency histograms (A); frequency density histogram overlaid with density plots (B); density plots (C), and standard normal distribution with mean of zero and standard deviation of one (D).

Like several quantitative clinical variables, the distribution of systolic blood pressure measurements is symmetrical about a single peak indicating more frequent blood pressure measurements in the middle of the distribution. It also has equal tails on either side of the peak (hence, a zero skewness) indicating very few measurements far from the peak, producing a characteristic bell-shaped curve known as the Normal or the Gaussian distribution (1,3,4). Often a capital N is used to emphasize that Normal is just a name of a distribution and does not necessarily imply normality (1). The shape of the distribution (or the curve) is determined by the variability in the data quantified by the standard deviation. The peak is tall and the shape of the curve is narrow for small variation in the data. Similarly, the pick is short and the shape of the curve is wide for larger variation in the data (Figure 2B). The normal distribution is the most common distribution that researchers encounter and it has several useful properties. Hence, Normal distribution is central to many inferential statistical techniques such as hypothesis testing and constructing confidence intervals (2,5).

Other quantitative variables such as serum cholesterol, CD4 counts, and the biceps skin-fold measurements in Tuberculosis (TB) tests may have a skewed, far from normal distributions (Figure 2C). If the distribution of a continuous variable is skewed, it is not appropriate to use inferential statistical methods such as confidence interval estimation or significance testing that require the outcome variable to have a Normal distribution (2-5). It is a common practice to transform such skewed continuous variables in to other scales such as logarithmic scale and analyse on that scale rather than on its original untransformed scale (3,4). The logarithmic (or log for short) transformation is frequently used because of its several advantages. First, multiplicative relationships such as interactions between variables may become additive. Second, skewed distributions may become symmetrical in such a way that assumption about Normal distribution would be reasonable. Third, results become interpretable (and thus useful) after transformations back to the original scale. Fourth, curves, if exist, may become straight lines hence linear relationships, for example, in linear regression, can be a reasonable assumption (2-4).

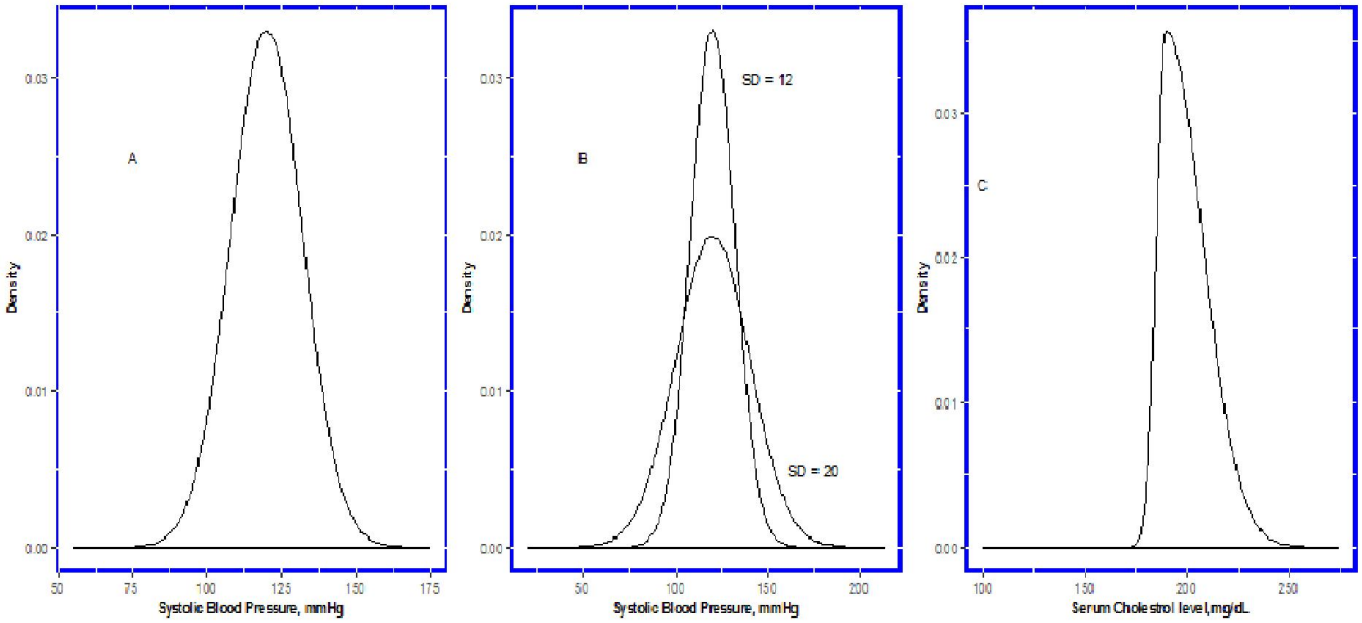

Figure 2. Density plots of normally distributed blood pressure measurements (A) for small and large standard deviations (B), and skewed distribution of blood cholesterol level measurements in mg/dL (C).

Continuous variables such as age and blood pressure measurements could also be stratified or categorized into few classes or dichotomized in to two groups (e.g. age can be dichotomized in to "young" and "old" and blood pressure measurements can be dichotomized in to "normal" and "high"). In few circumstances, this approach might improve interpretations of findings or avoid analytic restrictions imposed by requiring some assumptions to be fulfilled. For example, when a variable does not follow Normal distribution it is not acceptable to use mean and standard deviation as summary measures or t -test to compare the means of that variable (blood pressure measurement) between two categories of a binary variable (e.g. sex: "male" and "female"). In such cases, the use of Chi-square test to compare proportions of "normal" and "high" blood pressure measurements or "young" and "old" individuals, between groups ("male" and "female") is feasible. However, it is important to note that gains from categorizing a variable comes at a cost of loss of information particularly if a continuous variable is dichotomized in to two groups. If we use two age categories, "young" and "old", we cannot say much about other age groups as the information is already collapsed in to two categories. Hence, if stratification is deemed necessary, cut-off points are preferred to have been commonly used in the field of investigation and have useful biological or clinical meaning, e.g. 45-years for women as a cut-off might have clinical meaning as on average menopause starts around that age (6,7).

Categorical variables such as sex and the presence of a disease, and count variables such as number of admissions to a hospital and number of deaths from a traffic accident are other types of phenomena that occurs naturally but do not have a Normal distribution. Such variables are described using other classes of distributions different from Normal distribution. For example, Binomial distribution is used to describe binary variables and Poisson distribution is used to describe count variables. Interestingly, when the sample size increases, probability distributions of Binomial and Poisson distributions tend towards the Normal distribution. Hence, if the sample size is large Normal approximations are often used for such distributions (2, 5).

One unique feature of a Normal distribution is that it can be described using two quantities, the mean and standard deviation. The mean describes the "centre" of the data generated from individual study participant (e.g. blood pressure measurement) and the standard deviation reflects the amount of variation that exists in the collected data. In other words, the standard deviation measure the average distance of all observations relative to the mean value in the same units as the original data (8). A small standard deviation implies that the values of the variable in the data (e.g. blood pressure measurements) are close to the mean of the variable, on average, and a large standard deviation implies that the individual values of the variable are far away from the mean, on average. Standard deviation cannot be negative as it is the summary measure of distance between individual data points and the mean. However, it can be zero if the value of every observation is exactly equal to the mean.

For a Normally distributed random variable we could tell the percentage of observations that lie within a certain distance (e.g. standard deviations) from the mean. For example, 68.26% of the observations lie within one standard deviation (1SD) of the mean, 95.44% of the observations lie within 2SD of the mean, 99.7% the observations lie within 3SD of the mean, and 95% of the observations lie within 1.96SD, i.e. between $\overline{X} - 1.96SD$ and $\overline{X} + 1.96SD$ (Figure 3A-D) (9).

Figure 3 is the graphical presentation of a computer-generated data to simulate blood pressure measurements of 1,000 individuals with a mean of 120 and a standard deviation of 12. From the dataset 68.26% lie within 1.0 SD (1x12 = 12mmHg) distance from the mean (i.e. within 120±12 mmHg or between 120-12 and 120+12 mmHg, i.e., between 108mmHg and 132mmHg). Similarly, 95% of the measurements lie within 1.96 SD (i.e. 1.96x12 mmHg = 23.52mmHg) from the mean, or within 20±1.96x12 mmHg (i.e., between 96.48 and 143.52 mmHg). We used range of values or intervals to describe the percentage of measurements which lie within a certain range of SD from the mean. These ranges of values are known as reference ranges (5).

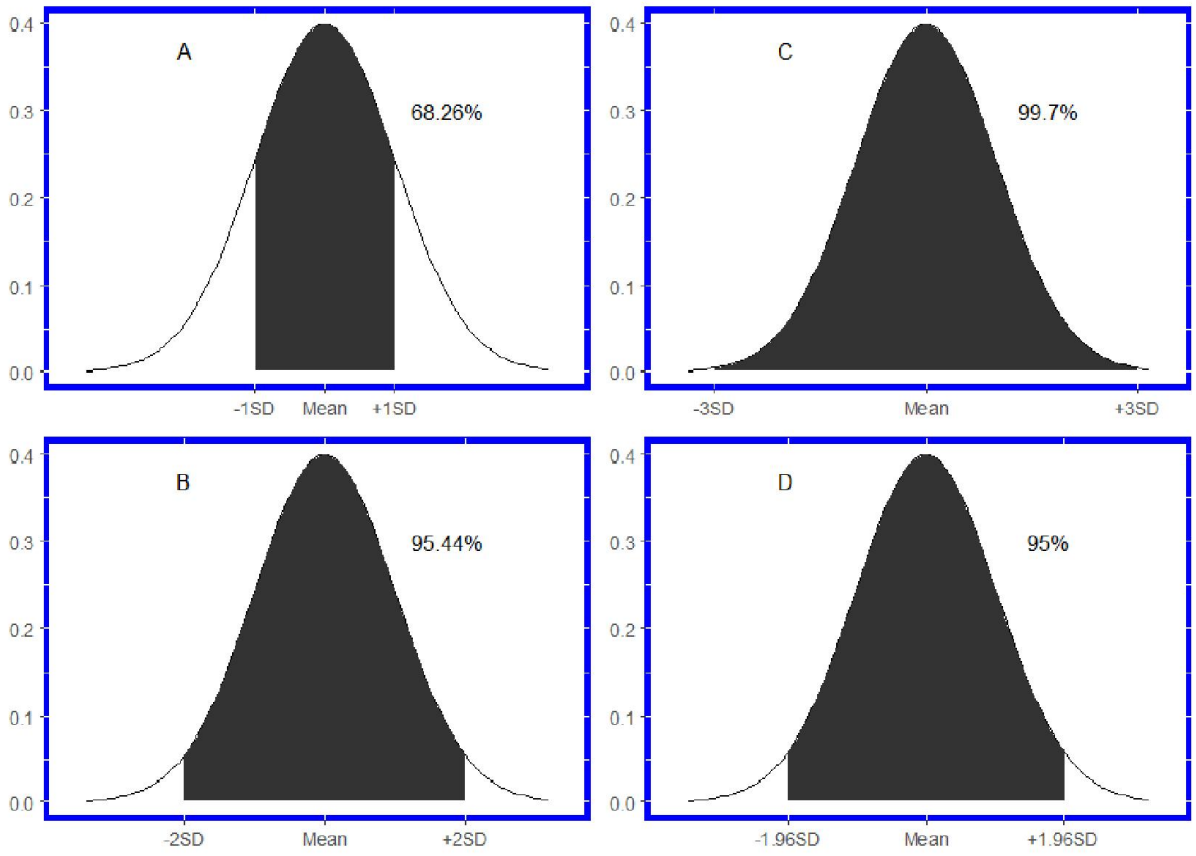

Figure 3. Plots for standardized Normal distribution of blood pressure measurements that lie within various multiples of standard deviation from the mean.

We might also consider to standardize the blood pressure measurements by subtracting the mean (20mmHg) from each of the measurements and by dividing these differences by the standard deviation (12mmHg) of the measurements $\left(Z_i = \frac{X_i - 20}{12}\right)$. This process generate a random variable (i.e. Z) which has a distribution with a mean of zero and a standard deviation of one. This distribution is known as standard Normal distribution (8). In other words, standard Normal distribution is a Normal distribution having a mean of zero and a standard deviation of 1.0 (Figure 2D).In simple terms, the measure of how many standard deviation units below or above the mean that a certain proportion or percentage of the observations (measurements) lie is called z-score or the standard score or the standard Normal deviate (SND) (9). In the previous example, the values of z were 1.0 and 1.96 in demarcating the region that cover 68.26% and 95%, respectively, of the systolic blood pressure measurements. The values of z-score and the corresponding proportions are often read from standard Normal distribution tables or generated from corresponding probability distributions using computer programs. In

order to use the z-score, we need to know the population mean, μ, and the population standard deviation, σ.

The **t-distribution** (also called **Student's t-distribution**) is another class of probability distribution commonly used in statistics and epidemiology. The t-distribution looks almost identical to the Normal distribution but has a bit shorter and fatter peak with long tails (2). When the sample size is (n<30) or when population standard deviation is unknown, the t-distribution is used instead of the Normal distribution in inferential data analysis. Unlike the Normal distribution, the shape of t-distribution and the number of observations that lie within a certain SD from the mean varies depending on the degrees of freedom (df). The t-distribution and associated t-score, in contrast to (standard) Normal distribution and z-score, respectively, are used in hypothesis testing (using t-test) and confidence interval estimation. Similar to the z-score, t -score can be read from t-distribution tables for a given degrees of freedom or generated from t-distribution using computer programs. As the sample size increase, the t-distribution looks more and more like the standard Normal distribution.

Degrees of freedom refer to the information we have to estimate a parameter. Some parameter estimates are based on more information than others. For example, in the previous example, there is more information to estimate the mean blood systolic pressure measurements with the sample size of 1,000 than with a sample size of 100. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based. In estimation of the mean using 1,000 and 100 sample sizes, there are 1,000 and 100 degrees of freedom, respectively, assuming that the individuals in each sample are independent. Detailed description of degrees of freedom will be provided in the coming Issue of the EMJ.

Other commonly used probability distributions are the Chi-squared distribution which is used in Pearson's Chi-squared tests and the Fisher's F-distribution which is used in the analysis of variances (ANOVA) and in the analysis of covariance (ANCOVA). Unlike the (standard) Normal and t-distributions, which are symmetrical around their mean values, both the Pearson Chi-squared and Fisher's F-distributions are positively skewed, the tail of the distribution on the right hand side is longer than on the left hand side. Similar to the t-distribution, the shapes of Pearson's chi-squared and Fisher's F-distributions vary with the degrees of freedom (df). Pearson's Chi-squared and Fisher's F-distributions are related in that F -distribution is a ratio of two random variables each of them having independent Pearson's Chi-squared distributions scaled by their respective degrees of freedom. $\left(i.e.\ F = \frac{\chi^1_{df1}}{df_1} \Big/ \frac{\chi^2_{df2}}{df_2} \right)$

THE CENTRAL LIMIT THEOREM

Consider the previous medical centre we mentioned where a sample of 1,000 individuals were randomly selected from a computerized record of all systolic blood pressure measurements on every individual who visited the centre for medical reasons. It is possible to take as many samples of the same size (n) if the researcher want using the electronic record (size N) as the sampling frame. It is possible to do sampling of individuals with or without replacement. The process of sampling is quite easy to implement with statistical computer programs. For simplicity, let us consider 100 different samples with each sample having a size of 1,000 individuals, sampled with replacement, and calculate the mean for each sample. When we say "sampling with replacement," we mean that every individual record or measurement have a chance to be included in several samples. After each sampling of a record, that record is put back to the sampling frame and given equal chance with other records of being included in the subsequent sampling processes.

Now, the 100 means calculated from each of the 100 samples (of size 1,000) will have a distribution known as the sampling distribution of the means. The centre of the sampling distribution of the sample means is at the mean of the entire 100 sample means, mean of the sample means, which is equal to the population mean, μ $\left(i.e.\ \mu = \sum_{i=1}^{100} \overline{X}_i \Big/ 100 \right)$. Note the difference between sample mean (mean of individual measurements in each sample) and mean of the sample means (mean of the 100 sample means, as we have 100 samples as an example). In our example, the population mean, μ, would be the mean of all blood pressure measurements (of size N) in the record of the medical centre. The sampling distribution of the sample means is Normal distribution; similarly, the sampling distribution of proportions (for binary variables such as sex) is called the Binomial distribution. As stated before, the Binomial distribution becomes very close to Normal distribution as the sample size increases (9).

On the other hand, the standard deviation (SD) of the sampling distribution of the means (i.e. the 100 sample means), also called the standard error (SE) of the means, is not equal to the standard deviation (SD) of the population

$\left(i.e. \sigma \neq \sum_{i=1}^{100} SD_i \Big/ 100 \right)$. The standard error of the mean is a measure of the variability that exists in the sample means relative to the population mean. In other words, SE of the mean summarizes the deviations of the individual sample means, $\overline{X}_i$ from the population mean, μ. Similar to how SD is interpreted, larger values of SE indicates higher degree of variability between the sample means (5, 9).

The central limit theorem states that if we have a population with a given mean and a given standard deviation and if sufficiently large samples (of each sample size n ≥30) are randomly drawn with replacement from the population, then the distribution of the sample means will be approximately Normal regardless of whether the distribution of the population is Normal or skewed. If the distribution of the population is Normal, then the theorem holds true even for small sample size (2,5,8). The central limit theorem enables researchers to use Normal distribution to make inference about the population mean based on summary measures obtained from one sufficiently large sample without the need to take many samples of the same size described above. Hence, it is possible to quantify uncertainty while making inferences about a population mean based on a single sample mean using Confidence Intervals.

The sample mean is considered to be unbiased estimate of the population mean provided the sample is randomly selected from the population and the sample size it is sufficiently large. If many random samples of the same size are drawn, the standard deviation of the sample means (i.e. standard error of the mean) is related to the SD of individual measurements in the sample by $SE = SD/\sqrt{n}$ where n is the sample size. From the formula, we can see that the SE is always smaller than the SD meaning that there is more variability between individual measurements within a sample than between sample means (5). Hence, the bigger the sample size, the less uncertainty in using the sample mean as estimate of the population mean.

It is now possible to use similar reasoning to construct confidence intervals (CI) for the single sample mean described above to reflect the uncertainty around it. If we do a repeated sampling, we know that the sample means will have a Normal distribution. It is also known that for a normally distributed random variable, 68.26% of the values lie within one SE of the mean, 95.44% of the values lie within 2 SE of the mean, and 99.7% of the values lie within 3 SE of the mean. Similarly, 95% of the sample means will lie within 1.96*SE from the mean (i.e. Sample mean - 1.96*SE to sample mean + 1.96*SE, note the symbol "*" is multiplication). This interval is known as the 95% confidence interval (95%CI) for the mean of the sample means.

The interpretation of 95% CIs is based on the concept of multiple repeated sampling we described before. If we draw multiple (let us say100) random samples of a given sample size (n), calculate the mean for each sample, and construct 95%CI around each of the 100 sample means, where 95% is called the confidence level. Note the difference between 95% (without CI) which is the proportion or percentage of the confidence interval and 95% CI that is the confidence level. In our example of 100 samples, 95 of the 100 95%confidence intervals (i.e. 95% of the 95%CIs) will contain the population mean, μ. The remaining 5% of the 95%CI (i.e. 5 out of the 100 95% CIs) will not cover the population mean. Hence, the population mean will not be covered within the two limits of five of the 95% confidence intervals.

In actual research work, the practice is to take only one random sample of a given size and estimate population parameters of interest. Hence, a researcher does not know if the 95%CI constructed based on the research data is among the ninety-five 95% CIs that contain the true population mean or it is among the five 95% CIs that miss the population mean. It means that researcher is 95% confident that the estimated 95%CI is one of the ninety-five 95% CIs that contains the population mean. It can also be restated as follows, the researcher can be 95% confident that the true population mean lies within the sample 95% confidence interval (5,9). In future series, CIs for other statistics will be discussed in detail.

REFERENCES

1. Altman D, Bland M. Statistics notes: the normal distribution. BMJ 1995;310:298.
2. Bland M. An introduction to medical statistics, 4th ed. Oxford University Press, Oxford, UK 2015. 448pp.
3. Bland M, Altman D. Transformations, means, and confidence intervals. BMJ 1996;312: 1079.
4. Bland M, Altman D. Statistics notes: the use of trans-formation when comparing two means". BMJ 1996;312:1153.
5. Whitley E, Ball J. Statistics review 2: Samples and populations. Critical Care 2002;6:143-8.

6. Rothman K, Greenland G, Lash T. Modern epidemiology, 3rd ed. Lippincott Williams and Wilkins, Philadelphia, PA 2008. 758pp.
7. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes A, Moons KGM. Adjustment for continuous confounders: an example of how to prevent residual confounding. Canadian Medical Association Journal 2013;185:401-6.
8. Van Belle G, Fisher L, Heagerty B, Lumley T. Biostatistics: a methodology for the health sciences. Vol. 519. John Wiley & Sons, Inc., Hoboken, New Jersey 2004. 879pp.
9. Kirkwood B, Sterne JC. Essential medical statistics. Blackwell Science Ltd, Oxford, UK 2003.501pp.