Determining true difference between treatment groups

Lucien J. Cardinal, MD*

Internal Medicine Residency Program, Department of Medicine, Stony Brook Medicine – Mather Hospital, Port Jefferson, NY, USA

In this article, the author reviews the P value and how it is used to determine true difference of outcome in treatment groups. P value, standard deviation, standard error of the mean, bias, and confidence interval are discussed in common language, with a minimum of jargon and with clinical examples.

Keywords: P value; standard error of the mean; central limits theorem; standard deviation; normal distribution; statistics; parametric statistics; bias; randomization

*Correspondence to: Lucien J. Cardinal, Internal Medicine Residency Program, Department of Medicine, Stony Brook Medicine – Mather Hospital, 75 North Country Road, Port Jefferson, NY 11777, USA, Email: LCardinal@matherhospital.org

Received: 3 November 2015; Accepted: 8 December 2015; Published: 17 February 2016

This article is a continuation of the series on basic concepts in research and contains terminology defined and discussed in prior articles (1, 2). The focus of this article is the P value and how it is used to differentiate groups, most commonly treatment groups. The P value as described in the last article may be used to characterize the probability of extreme measures in a single group of measures at hand. However, P values are most often encountered in the medical literature when comparing two treatment groups, in which case the P value serves as an arbiter of true difference between the two groups. This leads to a discussion of the standard error of the mean (SEM), the null hypothesis, inference, and the random sample.

Standard error of the mean

On choosing a sample from a large group (population), one can make calculations to determine the center value (mean) and the variability (standard deviation, SD) of the sample. These values can be interpreted as estimates of the true mean and SD of the parent group (population). However, performing the same calculations on a second sample will result in slightly different estimations of the mean and SD. Repeating this ad infinitum will generate a large number of estimates of the mean and SD that can be plotted and examined more closely. Examination of these estimates would reveal some interesting facts (3) and result in the following conclusions:

1. The distribution of the estimates of the means will be a normal distribution (or almost normal) even if the distribution of the parent group is not normal. The SD of the estimates of the mean is referred to as the SEM.

2. The SEM can be simply calculated from the SD of the sample and the size of the sample (n).²

3. As the sample gets larger, the SEM will become smaller.

4. As with all normal distributions, the true mean of the population will be the mean of the estimates, will be the most frequent value (mode), and will lie at the center of the distribution (median). Consequently, 50% of the estimates will lie below the true mean and 50% will lie above the true mean.

Because the SEM is normally distributed and related directly to the sample SD and sample size, the researcher can predict the likelihood of the true mean lying within specified bounds (see article 2). By increasing the size of the sample, the researcher can make the estimate more precise (the bounds tighter). The calculated upper and lower bounds, that will contain the true mean 95% of the time, is the ‘confidence interval’. Recall that 95% of values in a normal distribution lie within 2 SD of the mean.

In summary, using the SD of the sample and the sample size, the researcher can calculate the likelihood of the true population mean lying within specified bounds.

¹This is a simplified restatement of the Central Limit Theorem.
²SEM = SD/√n.
Furthermore, by progressively increasing the size of the sample, the range between the upper and lower bounds will become progressively smaller.

In the last article, we demonstrated that based on a given SD and the mean we could predict the distribution of red blood cell (RBC) volume in a specimen. Assume for example that the mean corpuscular volume (MCV) is 90 fL and that the SD is 15 fL. We can deduce that 95% of RBCs in the given specimen have volumes between 2 SD (2 \times 15 = 30 fL) above and below the mean of 90 fL. Thus 95% of RBCs in the original specimen will have volumes falling between 60 and 120 fL (Fig. 1).

Using the information presented under the SEM above, we can additionally quantify the accuracy of the estimate of the mean (90 fL) and we can examine how accuracy changes based on the sample size. The SD of the sample is 15 fL. The SEM is equivalent to the SD/\sqrt{n}.

In chart 1, we can observe how the ‘confidence interval’ changes with sample size, becoming more precise with increasing sample size. When the sample size is 10,000 (row 3), the SD of the mean (SEM) is 100 times less (0.15 fL) than the SD of the sample (15 fL).

By varying the sample size, the researcher can make the estimate of the mean as precise as desired.

**Null hypothesis and P value**
In the medical literature, the P value is commonly used to define the likelihood of occurrence of an observed difference in means between two samples, assuming that the two samples come from the same group (population). The assumption is that the difference observed is the result of an error related to sampling. The P value quantifies how commonly such a magnitude of difference is predicted to occur when samples come from the same population. This assumption that the samples come from the same parent group is referred to as the ‘null hypothesis’. A P value equal to or below 0.05 indicates that the observed magnitude of difference would occur by chance only 5% of the time if in fact the samples come from the same parent group. A P value of 0.05 or less is said to be ‘significant’ and is conventionally accepted as evidence that the samples come from different groups.

Suppose a group of patients taking a certain medication have an observed RBC MCV of 90 fL and SD of 15 fL. A comparison group of nonmedicated patients has an MCV established in prior research to be 89 fL. The examination of 100 RBCs from the medicated group might not be sufficient to establish a difference because 89 fL lies within the confidence interval around the 90 fL mean of the medicated group at the sample size of 100 (Chart 1, row 2). On the other hand by expanding the sample size to 10,000 cells, 89 fL would lie below the lower bound (89.7 fL, Chart 1, row 3) of a 90 fL mean, thus establishing that a true difference is likely to exist (Fig. 2). This disproves the ‘null hypothesis’ that the two samples come from the same group, consistent with the fact that they come from two separate groups: the medicated group and the nonmedicated group.

---

**Chart 1. Relationship Between Sample Size and Confidence Interval**

| SD (fL) | Sample size | SEM (fL) | 95% confidence interval of the mean (fL) |
|---------|-------------|---------|----------------------------------------|
| 15      | 25          | 15/5 = 3| 90 ± (2)(3) = 84 – 96                  |
| 15      | 100         | 15/10 = 1.5| 90 ± (2)(1.5) = 87 – 93               |
| 15      | 10,000      | 15/100 = 0.15| 90 ± (2)(0.15) = 89.7 – 90.3         |

The confidence interval extends 2 SD above and below the estimated mean of 90 fL.
Inference

Within the medical literature, the goal of research is often to make an inference about all patients with a given disease or characteristic. However, ultimately the inference about the larger group comes down to an examination of the sample at hand. When the science of statistics is used to make predictions about a broad group, beyond the actual cases examined, it is referred to as ‘inferential statistics’.

When using inferential statistics, the researcher is attempting to predict an unknown value (e.g., the mean) that applies to all such patients. The true value for the entire group of all such patients is initially unknown and remains unknown. The nature of inferential statistics is the calculation of an estimate and qualification of the estimate in terms of its accuracy.

Because the researcher attempts to make a generalization based on the sample at hand, it is extremely important that the sample be representative of the population of interest. If the sample is not representative, then the estimated value will likely be incorrect. Errors in statistical prediction that are related to the way the study is constructed or carried out are said to be the result of ‘bias’ or ‘systematic error’. Statistical use of the word ‘bias’ does not imply any intention or prejudice. Error resulting from bias is contrasted with error resulting from mathematically predictable variability.

Random sample

In most experiments or surveys, a subset of the entire group of interest is examined. The examination of all cases is usually not done for the following reasons: it is difficult to identify all cases, it is resource intensive, and from a statistical point of view it provides little added certainty. Additionally, examination might involve destruction of the biological sample (e.g., estimating the weight of biologic organs on a scale is limited to pathologic specimens removed as warranted by surgery or at autopsy). For these reasons, the researcher is likely to use a sample in order to estimate values related to the greater population. In order to assure an accurate estimation, that sample must be representative of the population.

One important technique to promote a representative sample is random selection. The researcher will use charts or random number generators to select members from the population for inclusion in the study. The use of “convenience sampling” such as selecting every 10th person, persons presenting on certain days, or persons with a prespecified first letter to the surname may seem random but have been shown to introduce bias (4).

Just as it is important to use random selection so that the sample is representative of the disease group under consideration, it is also important that the assignment of sampled individuals to treatment group be random as well. This will decrease the likelihood of systematic error. A common misconception is that random assignment to treatment groups statistically guarantees that the groups are the same. Although it is true that the randomly assigned groups will not be subject to systematic error, they will be subject to statistical variation (chance difference).

For example, in assigning two groups at random it may occur by chance alone that there is some important difference between the two groups that could be associated with differences in the measurement of interest. A disproportionate (chance) assignment of patients taking iron supplements to one group might alter estimates of the MCV, because total body iron may impact the MCV (5).

Many studies that compare treatments will include a table comparing baseline characteristics between two groups assigned to different treatments. This allows

**Fig. 2.** Distribution of estimations of the mean in the treatment group for two sample sizes.
The horizontal axis is RBC size in femtoliter (fL). Sample size = 10,000 (red). Sample size = 100 (blue). Mean of comparison group = 89 fL (green). The sample size of 100 is insufficient to discriminate between the treatment group mean and the comparison mean. Reference Chart 1.
readers to examine if there are any important baseline differences between groups. In these tables, they often compare 20 or more characteristics. These tables also commonly include a $P$ value next to each characteristic to indicate the likelihood that the observed difference would occur by chance alone. It is not uncommon to encounter a characteristic with a ‘significant’ $P$ value. By definition, a significant $P$ value (0.05) will occur 5% of the time (0.05 = 5%) when the samples come from the same parent group, a frequency of one in 20. When a table lists many characteristics (e.g., 20 or more), it is not unusual to have one, two, or three values with a significant $P$ value. This may simply represent sampling error. An article with a baseline characteristic table demonstrating this phenomenon is cited (6). When reviewing a study, the reader must consider two questions. Does the baseline difference between groups suggest a systematic error in choosing samples or is it an artifact of sampling? Could the baseline differences between samples account for an observed difference in the outcome of interest and thus undermine the conclusions of the study?

**Main points**

- Standard deviation is a measure of variability of the population.
- Standard error of the mean is a measure of the variability of the estimate of the mean. It diminishes as the sample size increases.
- Confidence interval marks the bounds within which the true mean lies 95% of the time.
- The null hypothesis is the working assumption that two samples come from the same population, and that differences in the means are simply the result of chance differences in the samples.
- A $P$ value of $\leq 0.05$ is referred to as ‘significant’ and is accepted as evidence that two samples come from different groups (populations), disproving the ‘null hypothesis’.
- Systematic error may occur as a result of inappropriate sampling and is referred to as ‘bias’.
- Randomization is used to diminish the likelihood of bias.

**Acknowledgements**

The author gratefully acknowledges the contributions of Yashodhan Chivate and Lorraine Farrell, F.N.P., whose thoughtful reviews and constructive criticisms contributed to the completion of this document.

**Conflict of interest and funding**

The author has not received any funding or benefits from industry or elsewhere to conduct this study.

**References**

1. Cardinal LJ. Central tendency and variability in biological systems. J Community Hosp Intern Med Perspect 2015; 5(3): 27930, doi: http://dx.doi.org/10.3402/jchimp.v5.27930
2. Cardinal LJ. Central tendency and variability in biological systems: Part 2. J Community Hosp Intern Med Perspect 2015; 5(5): 28972, doi: http://dx.doi.org/10.3402/jchimp.v5.28972
3. Duncan RC, Knapp RG, Miller MC. Introductory biostatistics for the health sciences. 2nd ed. New York: Wiley; 1983, 249 p.
4. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract 2010; 115(2): c94–9.
5. Zhu A, Kaneshiro M, Kaunitz JD. Evaluation and treatment of iron deficiency anemia: A gastroenterological perspective. Dig Dis Sci 2010; 55(3): 548–59.
6. Soliman EZ, Howard G, Cushman M, Kissela B, Kleindorfer D, Le A, et al. Prolongation of QTc and risk of stroke: The REGARDS (REAsons for Geographic and Racial Differences in Stroke) study. J Am Coll Cardiol 2012; 59(16): 1460–7.