MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Editorial

Open Access

CrossMark

A very short and gentle review of mean and standard deviation as summary statistics of a sample

Volume 6 Issue 2 - 2017

Heon-Jae Jeong,[1] Wui-Chiang Lee[2]

[1]The Care Quality Research Group, Chuncheon, Korea

[2]Department of Medical Affairs and Planning, Taipei Veterans General Hospital & National Yang-Ming University School of Medicine, Taipei, Taiwan

Correspondence: Wui-Chiang Lee, Department of Medical Affairs and Planning, Taipei Veterans General Hospital & National Yang-Ming University School of Medicine, Taipei, Taiwan, Tel +886-2-28757120, Fax +886-2-28757200, Email leewuichiang@gmail.com

Received: June 21, 2017 | **Published:** June 27, 2017

When we are asked to analyze data, some of us just dive into calculating centrality and spread, usually with a mean and standard deviation (SD) combo; it is our innate nature. Why are we doing this? What is the use of it? Assume we have 100 observations; the most precise way to describe the data is just to show the 100 numbers as they are. However, this is not at all efficient, actually giving us little or no *information*. Thus, we summarize the data, and the most popular way to do so is to use the above-mentioned mean and SD combo. We successfully decrease the amount of data from 100 to 2 (one mean and one SD).

This huge shrinkage allows us to conduct many statistical comparisons, such as a t-test and analysis of covariance. Indeed, anyone who uses statistical analyses in daily life gets help from this summarizing measure. However, before we move on, let us make it clear that mean and SD are terminology for normal distribution.

Now, we have to ask ourselves how many times we check whether the data really follow normal distribution. We habitually believe so and that belief stems from the following misunderstanding: If the sample size is large enough (the magic number is usually 30), then the mean will be approximately the same as that of the population, and variance in the *samples* will follow normal distribution. Here, the most important word is *samples*, the plural. Actually, the above is not about the 100 observations in themselves, but the countless possible samples that can be derived from the population. Thus, what follows normal distribution is not the statistics of the single sample, but the distribution of possible statistics—so-called sampling distribution. We named this phenomenon the central limit theorem (CLT). The virtue of CLT is that it can be applied to a sample with any distribution, such as beta, gamma, and Weibull. How convenient it is!

Here, our most popular misunderstanding ambushes us. More often than not, people think of a sample, so long as its size is more than 30 or so, as normally distributed and use mean and SD as summary statistics, even when the actual distribution is largely skewed. Of course, outliers play an important role to twist the summary measures, albeit frequently removed beforehand. If a sample looks odd, it is actually odd, period. With the help of CLT, we can deal with only population-level parameters with the normality assumption.

Indeed, CLT is a savior for us. However, we must check the distribution of sample data on hand and understand how they look. That way, we won't make unsanctioned mistakes while saying mean and SD. We are statisticians, or at least statistics enthusiasts. We do not have the right to turn a blind eye to this elementary mistake.

Acknowledgements

None.

Conflicts of interest

None.