HHS Public Access

Author manuscript

J Pancreatol. Author manuscript; available in PMC 2021 May 18.

Published in final edited form as:

J Pancreatol. 2019 September ; 2(3): 69–71. doi:10.1097/jp9.0000000000000024.

A note on error bars as a graphical representation of the variability of data in biomedical research: Choosing between standard deviation and standard error of the mean

Li Tang, PhD[1], **Hui Zhang, PhD**[1], **Bo Zhang, PhD**[2,*]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, U.S.A.

[2]Department of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA 01605, U.S.A.

Abstract

Standard deviation and standard error of the mean have been applied widely as error bars in scientific plots. Unfortunately, there is no universally accepted principle addressing which of these two measures should be used. Here we seek to fill this gap by outlining the reasoning for choosing standard error of the mean over standard deviation and hope to shed light on this unsettled disagreement among the biomedical community. The utility of standard error of the mean and standard deviation as error bars is further discussed by examining the figures and plots published in two research articles on pancreatic disease.

Keywords

standard error of the mean; standard deviation; error bars; confidence interval; pancreatic disease

Introduction

Error bars are frequently used in biomedical and clinical publications to describe the variation in observed data, with standard deviation (SD) and standard error of the mean (SEM) being the most common measures of variability. Both SD and SEM are important concepts in statistical inference; however, they are not interchangeable. The SD describes the spread of a population from which the sample was drawn and represents an inherent feature of the cohort being studied. In contrast, the SEM indicates how precisely the mean of the population can be estimated from the sample that was drawn. Thus, SD is a constant that is independent of the sampling process, and SEM is random and influenced by sampling, especially by the sample size (n). In most cases, the relation between SD and SEM is expressed as $SEM = \widehat{SD}/\sqrt{n}$, where the circumflex (^)" represents estimation.

[*]**Corresponding author.** Bo.Zhang@umassmed.edu.

Author contributions: L.T. and H.Z. conceived of the presented idea and conducted the review. B.Z. verified the methods and results and led the review of exemplar articles of pancreatic research. All authors discussed the results and contributed to the writing of final manuscript.

Disclosure: The authors declare no conflicts of interest.

Choosing between standard deviation and standard error of the mean for error bars

Although several articles have discussed error bars in the last decade, [1–7] whether SD or SEM should be used in scientific plots remains controversial. [2,6,7] A recent publication in *Nature Methods* [9] discussed various types of error bars but did not provide clear suggestions on which error bar to pick in general. Table 1 summarizes the types of error bars reported in articles from representative scientific journals with high-impact factors. Issues published from January to March 2019 were reviewed. The data suggest that many scientific investigators are still uncertain about which type of error bar to present, thus underlining the need to establish a "universal" choice for the scientific community. From a biostatistics point of view, we favor the use of SEM over that of SD, for describing scientific results under most circumstances.

In most scientific data presentations with error bars, the goal is often to compare two or more population means. Although the population means are unknown, for the purpose of making a reliable inference, it is of more interest how far the estimated mean (not an individual observation) is from the true population mean. Therefore, the variability of the estimated means (i.e., SEM) suits the situation better than the SD.

The use of SEM also may enable one to make simple conclusions by visual inspection, because SEM is closely related to the confidence interval and *p*-value. For example, when comparing means, consider the popular 2-sample Student's *t*-test. If the SEM bars of two groups touch when plotted as box plots side-by-side, it usually implies that the test statistic *t* is 1.41 or less, corresponding to a *p*-value greater than 0.15.[1] For a visual display, if the sample size is 10 or more and both groups have similar SEMs, a gap of 1*SEM corresponds to $p \approx 0.05$ and 2*SEM corresponds to $p \approx 0.01$.[6] For smaller sample sizes, larger gaps are needed to get the same *p*-values. In contrast, error bars using SD cannot easily suggest these conclusions visually.

Sample size is crucial for obtaining a precise estimation and making a reliable inference. The larger the sample size, the more precise the estimation of the population mean (i.e., smaller SEM) and the greater the chance of identifying a difference in the means of multiple groups. In contrast, SD is not affected by sample size. Thus, by plotting SEM error bars, a sufficiently large sample size will be appropriately credited by showing a sharpened bound of the estimated population mean, which also facilitates the statistical demonstration.

In some biomedical studies, the primary interest is to compare percentages, and each subject is observed with a binary response (e.g., yes/no or 0/1). In this scenario, the percentage is the mean of responses, and its margin of error is the most important statistical feature of the results, which can only be represented by SEM and not SD.

Using standard deviation versus standard error of the mean as error bars in the presentation of pancreatic disease research

The pancreas plays a key role in metabolism and is involved in the pathogenesis of several diseases. To describe the utility of SD and SEM in pancreatic research, we evaluated the figures and usage of SD and SEM by two research articles, one preclinical study and one biological study.

In one research article entitled, “Morphine worsens the severity and prevents pancreatic regeneration in mouse models of acute pancreatitis,” [9] the authors elucidated the roles of morphine in the progression of acute pancreatitis (AP), which had not been rigorously tested. Opioid analgesics, including morphine, hydromorphone, and fentanyl, are commonly used to alleviate pain caused by AP. [10] Opioids affect the immune system and regulate inflammatory pathways in nonpancreatic diseases. [11–13] Therefore, whether opioids should be used for analgesia of AP was controversial in past decades.

In this article, the authors induced AP in wild-type or Mu opioid receptor knockout mice by using caerulein, *L*-arginine, or ethanol–palmitoleic acid. Mice were then treated with placebo or morphine. To evaluate the effect of morphine, various tissues were collected. To determine tissue function, the intestinal permeability was evaluated, the regeneration was detected by 5-bromo-2’ deoxy uridine incorporation, and myeloperoxidase activity was analyzed. Immunohistochemical analysis was done to show the morphology of tissues and quantify necrosis. Immunofluorescence and qPCR (quantitative polymerase chain reaction) were used to capture the expression of target genes on the protein and nucleic acid levels.

The figures with error bars in this article (Figures 1 and 3–6) were configured by statistical analysis, to describe the quantification of necrosis, the infiltration of pancreatic macrophage, the expression of protein, and the proliferative response in the injured pancreas. The error bars were all calculated from the SEMs of data obtained from histologic, cytologic morphologic, or molecular biologic experiments. For these types of data, the dispersion of the sample mean should be well considered, because the mean value is the key characteristic that differs between study groups. The size of the sample is then directly related to the soundness of the scientific inference.

SD is an inherent measure that quantifies the dispersion of an experimental sample that was drawn from a population. When the goal is to demonstrate the population-level mean and variation, rarely are the SDs used to plot the error bars.

In another research article entitled, “Comprehensive characterization of compartment-specific long non-coding RNAs associated with pancreatic ductal adenocarcinoma,” [14] the authors used systematic, experimental methods to study the function of long noncoding epithelial RNAs associated with genetic characteristics and clinical outcomes in pancreatic ductal adenocarcinoma (PDA). PDA is a highly metastatic disease with limited treatment choices. [15] Genomic and transcriptomic analyses have identified signaling pathways and cancer-driving genes that can inform treatment stratification and targeted therapy, but these analyses were often carried out in large samples and focused on coding genes, which make

up only a small portion of the genome. In this article, the authors developed a computational framework for reconstructing the noncoding transcriptome from cross-sectional RNA sequencing, integration of somatic copy number changes. They investigated the function of epithelial long noncoding RNA related to genetic characteristics and clinical outcomes in PDA by using systematic and experimental biological methods.

In the figures with error bars in this article (Figures 3–5), the authors displayed the error bars as the graphical representations of measured gene expression levels by using $\log_2$-transformed RPKM (reads per kilobase of transcript per million mapped reads) values from RNA-sequencing data. The delta-CT values from the qRT-PCR analysis were commonly considered approximately normally distributed at the population level. The gene expression values could be affected by many factors, such as batch effects. In some cases, the values were standardized or inversely transformed. It might be the intention of the investigators to present the error bars to show the spread of expression of various genes at the population level, rather than mean expression values estimated by a certain study sample. Thus, SDs were used as error bars in these figures. However, the authors should have emphasized that the SDs reflect the variation but not the errors in the gene expression levels. The sample size in this study was 147, a number that was considered large enough. It should be emphasized that, unlike SEMs, the SDs do not shrink as the study sample size increases.

Conclusion

Our arguments support the use of SEM rather than SD as the “universal” error bar in scientific publications. When there is a need to show the dispersion of individuals in the population, a box plot with interquartile range should be shown. Nevertheless, we urge investigators to clearly state whether their error bars are SEMs or SDs in all biomedical research publications.

A third type of error bar in biomedical research publications is based on the confidence interval, an interval estimate indicating reliability of a measurement. The confidence interval and the SEM are both depending on the sample size and are related by the *t*-statistic. In large samples, the SEM bar is approximately equal to a confidence interval of 67%, and twice of the SEM bar is approximately equal to a confidence interval of 95%. [3]

Acknowledgement

We thank Ms. Fang Wang from St. Jude Children’s Research Hospital for assisting in the review of error bars used in articles from representative scientific journals and thank Dr. Wei Zhang from the University of Massachusetts Medical School for assisting the presentation of two exemplar articles in pancreatic research.

Funding: Dr. Bo Zhang’s research was supported, in part, by the National Institutes of Health grant U24 AA026968 and the University of Massachusetts Center for Clinical and Translational Science grants UL1TR001453, TL1TR01454, and KL2TR01455.

References

[1]. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. Am Stat 2001; 55:182–186.

[2]. Nagele P Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. Br J Anaesth 2003; 90:514–516. [PubMed: 12644429]

[3]. Vaux DL. Error message. Nature 2004; 428:799.

[4]. Altman DG, Bland JM. Standard deviations and standard errors. BMJ 2005; 331:903. [PubMed: 16223828]

[5]. Belia S, Fidler F, Williams J, Cumming G. Researchers misunderstand confidence intervals and standard error bars. Psychol Methods 2005; 10:389–396. [PubMed: 16392994]

[6]. Cumming G, Fidler F, Vaux DL. Error bars in experimental biology. J Cell Biol 2007; 177:7–11. [PubMed: 17420288]

[7]. Carter RE. A standard error: distinguishing standard deviation from standard error. Diabetes 2013;62: e15. [PubMed: 23881207]

[8]. Krzywinski M, Altman N. Points of significance: error bars. Nat Methods 2013; 10:921–922. [PubMed: 24161969]

[9]. Barlass U, Dutta R, Cheema H, George J, Sareen A, Dixit A, et al. Morphine worsens the severity and prevents pancreatic regeneration in mouse models of acute pancreatitis. Gut 2018; 67:600–602.

[10]. Basurto Ona X, Rigau Comas D, Urrutia G. Opioids for acute pancreatitis pain. Cochrane Database Syst Rev 2013;7:CD009179.

[11]. Ammori BJ. Role of the gut in the course of severe acute pancreatitis. Pancreas 2003; 26:122–129. [PubMed: 12604908]

[12]. Hotz HG, Foitzik T, Rohweder J, Schulzke JD, Fromm M, Runkel NS, et al. Intestinal microcirculation and gut permeability in acute pancreatitis: early changes and therapeutic implications. J Gastrointest Surg 1998; 2:518–525. [PubMed: 10458730]

[13]. Meng J, Yu H, Ma J, Wang J, Banerjee S, Charboneau R, et al. Morphine induces bacterial translocation in mice by compromising intestinal barrier function in a TLR-dependent manner. PLoS One 2013;8: e54040. [PubMed: 23349783]

[14]. Arnes L, Liu Z, Wang J, Maurer HC, Sagalovskiy I, Sanchez-Martin M, et al. Comprehensive characterisation of compartment-specific long non-coding RNAs associated with pancreatic ductal adenocarcinoma. Gut 2019; 68:499–511. [PubMed: 29440233]

[15]. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 2014; 74:2913–2921. [PubMed: 24840647]

Table 1.

Counts of articles by types of error bars published in representative scientific journals from January 1, 2019, to March 31, 2019.

Journals	Counts of Articles by Error Bar Types				Total Counts[b]
	SD	SEM	Others[a]	Unidentified	
Science	20	29	15	7	71
Nature	43	47	19	5	114
Cell	30	34	4	3	71
New England Journal of Medicine	0	4	9	2	15
Journal of the American Medical Association	0	2	14	0	16
The Lancet	1	1	17	2	21

[a]Other measures shown as error bars.

[b]These data represent the total number of articles that appeared in the publication during the review period that used error bars in figures. The articles using two or more types of error bars were counted in each category but only once in the total category.

Abbreviations: SD, standard deviation; SEM, standard error of the mean.