Doc Ophthalmol (2010) 121:215–222
DOI 10.1007/s10633-010-9249-7

REVIEW ARTICLE

A primer on common statistical errors in clinical ophthalmology

Karen Holopigian · Michael Bach

Received: 22 April 2010 / Accepted: 6 October 2010 / Published online: 23 October 2010

Abstract Although biomedical statistics is part of any scientific curriculum, a review of the current scientific literature indicates that statistical data analysis is an area that frequently needs improvement. To address this, we here cover some of the most common problems in statistical analysis, with an emphasis on an intuitive, tutorial approach rather than a rigorous, proof-based one. The topics covered in this manuscript are whether to enter eyes or patients into the analysis, issues related to multiple testing, pitfalls surrounding the correlation coefficient (causation, insensitivity to patterns, range confounding, unsuitability for method comparisons), and when to use standard deviation (SD) versus standard error of the mean (SEM) "antennas" on graphs.

Keywords Correlation coefficient · Independence · Statistics · Type I error · Variability

K. Holopigian (✉)
Department of Ophthalmology, New York University
School of Medicine, 462 First Avenue, New York,
NY 10016, USA
e-mail: kh19@nyu.edu

M. Bach
University Eye Hospital, University of Freiburg,
Killianstraße 5, 79106 Freiburg, Germany
e-mail: michael.bach@uni-freiburg.de

Introduction

Why do we use statistics? In a perfect world, we would simply conduct experiments and if we obtained differences among our groups, we would conclude that our manipulations caused an effect. However, there is variability in the world, which is reflected in our data. Because of this variability, we need a method of determining which aspects of variations in the data are due to true differences and which parts are due to variability. This is why we use statistics, to help us sort out the underlying sources of variability and to aid us in correctly attributing experimental change from random error, without bias. The purpose of this primer is to provide a general overview to those unfamiliar with basic statistical principles. Our aim is to provide a fundamental understanding on how to use these statistical tools appropriately to maximize accuracy in data interpretation. The following is a partial list of statistical concerns that commonly contain errors as well as some suggestions on the methods of overcoming these difficulties. A more comprehensive review can be found, for instance, in Strasak et al. [1]. Also recommended are Bland and Altman's series of tutorial-style editorials in the British Medical Journal in 1996.

Eyes versus patients

Perhaps the most common statistical error in the ophthalmic literature is the confusion between eyes and patients. Since one eye is good, two eyes are better, correct? Correct when looking at a visual scene, not necessarily correct when analyzing data. The difficulty with including both eyes of a patient into a data set is that the two eyes are not independent of each other, since the data from the right and left eyes are more highly correlated than are the data from the two eyes of different subjects. However, including both eyes has the benefit of increasing the sample size and therefore the power of any statistical test that is applied to the data. Should two eyes of a subject be included into a data set and how does one avoid confounding the statistical interpretations?

The most obvious case of when both eyes are included is when the relevant comparison is between the eyes (e.g., the affected vs. non-affected eyes in patients). In this case, it is necessary to use both eyes, and the statistical analysis only needs to account for the fact that the data sets are correlated in some way. However, when both eyes of a patient are included because both eyes have the disease of interest or when both eyes of a control subject are used, the non-independence of the data must be addressed. This is because the data sample has both independent observations (from different subjects) and potentially non-independent observations (for instance, in glaucoma there is a high correlation of disease markers between eyes) [2]. Since most statistical tests assume the observations are independent, including non-independent observations will increase the probability of making a Type I error (concluding that there is statistical significance among the groups when the observed differences are due to chance). Therefore, the more non-independent observations that are included, the more likely it is that one will erroneously conclude that there are statistical differences among the measures. To put it most simply, adding the fellow eye data can nearly be the same as duplicating each data point.

Newcombe and Duff [3] used computer simulations to quantify the increase in the statistical error rate from analyzing data sets that include both eyes of individuals using unpaired (i.e., not accounting for the two eyes) two-sample *t*-tests. They examined statistical outcomes of two groups of subjects using four methods: comparing the right eyes only; comparing the left eyes only; comparing the averaged values of the right and left eyes; and including the right and left eyes of all subjects. The simulations were done 200 times for each method with a statistical significance level of 0.05, which would yield 'significant differences' 10 times out of 200 by chance alone. For the four methods outlined above, the simulation yielded 'significant' differences of 8, 6, 9, and 39, respectively, clearly highlighting the error of including data from two eyes without accounting for it. Of the remaining three approaches to data management, the authors conclude that methods 1–3 are all valid, but that method 3 (averaging the data from two eyes) is preferable, because it includes more information and thus is likely to reduce variability.

These simulation results emphasize that one needs to be cautious when including two eyes of a subject. However, as mentioned above, including both eyes will increase the statistical power of a test, thereby increasing the likelihood of detecting true significant effects. Another method of dealing with this issue is the use of regression techniques to parcel out the independence of the eyes. There are different approaches to this (e.g., the use of linear and logistic regression models vs. the estimating equation approach) [4–6] and they all serve to utilize the entire data set. Glynn and Rosner [7] evaluated these approaches using separate data from patients with RP and from patients with glaucoma and provide guidelines for their use. They also pointed out that additional correlations, such as multiple family members within a data set, will require more complex approaches to maintaining the independence of the data.

Multiple measures and multiple comparisons

When analyzing electrophysiological and/or psychophysical data, it is not uncommon to have multiple measures. For example, let's imagine that we are conducting a study on whether open-angle glaucoma (OAG) has any effect on the electroretinogram (ERG). A priori, we do not know which component(s) of the ERG might be affected, so it is feasible that we might examine a number of measures of full-field flash and pattern ERG (e.g., scotopic amplitudes and peak times, photopic amplitudes and peak times, Naka–Rushton parameters, and P50 pattern ERG amplitude). Furthermore, we might be interested in comparing certain patient characteristics with our ERG parameters to explore the possibility that these

Table 1 An illustration of the increase in Type I errors (erroneous significance effects) due to multiple testing

Measure	Group 1, mean ± SD	Group 2, mean ± SD	*P* value (*t* test)
Age [years]	42 ± 23	54 ± 31	0.232
Acuity [logMAR]	0.49 ± 0.38	0.71 ± 0.54	0.203
Color score	70 ± 67	120 ± 60	0.0387*
IQ	105 ± 38	113 ± 30	0.526
MD [dB]	0.27 ± 1.0	0.71 ± 0.96	0.233
IOP [mm/Hq]	18 ± 5.6	17 ± 6.5	0.835
Dark-adapted 0.01 a-wave [μV]	12 ± 19	1.4 ± 22	0.164
Dark-adapted 0.01 b-wave [μV]	227 ± 51	250 ± 35	0.173
Dark-adapted 3.0 a-wave [μV]	251 ± 95	204 ± 72	0.137
Dark-adapted 3.0 b-wave [μV]	368 ± 118	322 ± 106	0.273
Dark-adapted 10.0 a-wave [μV]	265 ± 102	256 ± 103	0.805
Dark-adapted 10.0 b-wave [μV]	353 ± 86	360 ± 98	0.838
Light-adapted 3.0 ERG [μV]	47 ± 21	37 ± 18	0.176
Light-adapted 3.0 flicker ERG [μV]	20 ± 0.34	20 ± 0.58	0.755
Acuity [logMAR]	0.43 ± 0.63	0.39 ± 0.61	0.877
Acuity [logMAR]	0.5 ± 0.35	0.63 ± 0.44	0.371
EOG ratio	1.6 ± 0.27	1.9 ± 0.3	0.00559**
PERG amplitude [μV]	4.8 ± 3.5	4.7 ± 2.2	0.894
VEP amplitude [μV]	10 ± 10	12 ± 13	0.700
VEP peak time [ms]	102 ± 10	98 ± 12	0.245

A simulated hypothetical experiment comparing multiple measures for two groups ($n = 15$ per group). All values were randomly drawn from normal distributions with roughly typical characteristics for each measure "tested". There should be approximately one significant effect per 20 *t*-tests (5% = 1/20). This was the second simulation and it yielded one significant effect at the 5% level (*, $P < 0.05$) and one at the 1% level (**, $P < 0.01$). Such a data set should either be avoided (since too many measures will diminish the power) or be analyzed with recognition of multiple testing, e.g., using an ANOVA. In addition, another error in this table is the use of differing significant digits for the different values: the same number of significant digits should be used throughout

factors might have a strong relationship with our dependent measures. Again, there are a variety of parameters that we might examine here, such as age, visual acuity, intraocular pressure, color vision scores, visual field mean deviation, and cup/disk ratio. All of these measures are worthwhile to examine in this study, and all of them are legitimate outcome measures for our question of interest. But, once again, the possibility of greatly increasing Type I errors rears its ugly head again.

It is sometimes the case that individuals faced with this data set will simply run a whole series of *t*-tests (e.g., control vs. OAG) comparing measure after measure using a statistical significance level of 0.05 (see Table 1). Presumably, the assumption is that since there are different parameters being examined, there is no issue with using *t*-tests for each comparison. The problem is, even if OAG has no effect on the ERG, with this number of comparisons there will be at least a few ERG parameters which appear to be significantly different between the control and OAG groups just by chance. Using the traditional level of $P = 0.05$ to accept a significant difference, for every twenty comparisons that are made, there will be, on average, one significant difference identified in your sample, even if there really is no true effect in the population (see example in Table 1).

This is a difficult issue, since some statisticians argue that it is acceptable to perform multiple comparisons on data sets, since it increases the likelihood of detecting a statistical effect. And certainly this approach is appropriate for generating hypotheses, to be rigorously tested in subsequent finely targeted experiments. However, more conservative approaches suggest using tools to account for the number of comparisons, such as using analysis of variance (ANOVA) with post hoc comparisons (this involves pair-wise comparisons only among measures that have already been shown to have statistical significance) and Bonferroni adjustments for multiple comparisons. The Bonferroni adjustment is simple to understand and to perform (the *P* value is adjusted for the number of possible comparisons among the groups being compared). A drawback to the Bonferroni adjustment is that it can greatly diminish the likelihood of finding significant differences even if a true effect is present. However, there is a modified version (the Bonferroni–Holm adjustment) which is less stringent and uses a graduated series of *P* values to determine significance [8]. This can be used as a reasonable compromise to the problem of multiple comparisons.

Correlation coefficients and causation

Correlation coefficients are useful measures that provide information about relationships between measures but, unfortunately, they can easily be misused. The most commonly used correlation coefficients are the Pearson correlation coefficient (r) and the Spearman rank order correlation (r_s), which is used for ranked pairs. Correlation coefficients are the indicators of how much the variability in one parameter corresponds to the variability in a second parameter. In other words, the correlation coefficient indicates how two measures covary. There is not any implication about causality that can be derived from calculating a correlation coefficient; it is not correct to attribute cause-and-effect relationships between variables based on significant correlation coefficients. A rather obvious example of this fallacy is the strong positive correlation between crime and ice cream sales. Obviously, one would not conclude that increased ice cream sales cause crime or conversely, that crime causes an increase in ice cream sales. Most likely, common factor(s), such as warmer weather and longer days in summer, are responsible for the strong correlation.

The interpretation of correlation coefficients

Typically, a Pearson correlation coefficient with an r value of 0.60 would be considered significantly

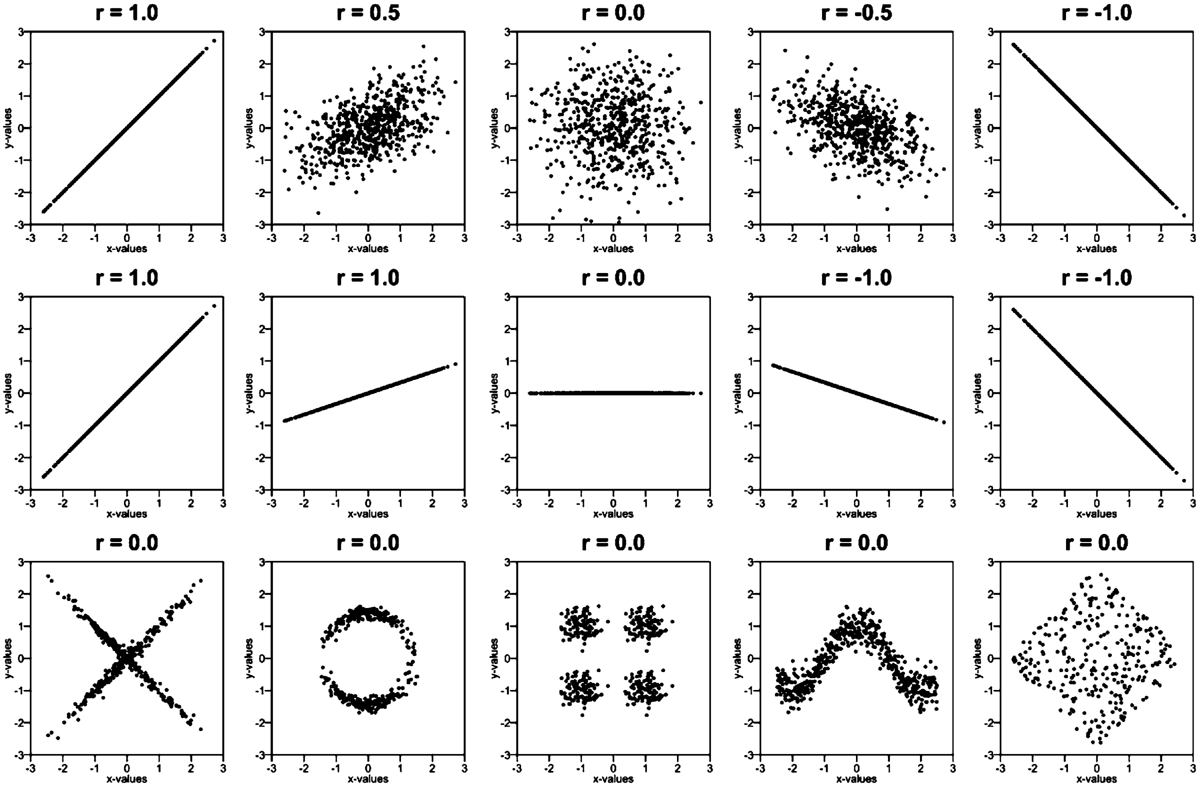

Fig. 1 Strikingly different data relations can result in the same correlation coefficient. Furthermore, non-significant correlations do not mean there exists no relationships within the data set, as shown in the examples in the bottom row

different from zero (i.e., there is a significant relationship), depending on the number of observations. However, an $r = 0.60$ indicates that only 36% of the variance in a is accounted for by the variance in b (% variance accounted for is the square of the correlation coefficient times 100). Although considered significant (depending on n), this is not an impressive relationship between these variables. On the other hand, even if a correlation coefficient is zero, there can be some very strong structure in the data. The bottom row of Fig. 1 demonstrates this—obviously contrived, but it demonstrates the basic principle. The point is that one must be careful in interpreting a significant correlation coefficient, since a significant relationship is not always a meaningful relationship.

The effects of range on correlation coefficients

There are other issues that need to be considered when using correlation coefficients to evaluate data. One is the concept of extreme values (e.g., the results from a person with no visual function vs. the results from a control observer). The size of the correlation coefficient (and hence the strength of the implied relationship) is directly related to the sample range. A larger range will produce a larger r value, while a narrower range will produce a smaller one. Therefore, inclusion of extreme values will often elevate the size of an r value, possibly erroneously. By way of example consider Fig. 2.

Why is the issue of outliers significant? Is it recommended that the outliers be ignored or removed from the data set? Obviously, one cannot remove data from the data set but must analyze the data present in a series, regardless of where the values fall. However, this information may sometimes be relevant when interpreting the meaningfulness of the correlations among measures. If it does appear that there are only one or two sets of extreme data points that are primarily important for the significance of the correlation coefficients, it is possible that the relationship between the two measures may be not meaningful in the 'real' world.

Correlation coefficients not suitable for test–retest or instrument comparisons

Since correlation coefficients are often used to examine the associations between two variables, it has been assumed that they provide a valid method for assessing test–retest goodness or can be used to validate how well a new instrument replicates the measurements of an older instrument. However, a high correlation between two measures merely indicates that they are related and, as Bland and Altman [9] pointed out, it would be amazing if two methods designed to measure the same quantity were not related. Instead, what is required to assess test–retest goodness is a measure of agreement, which correlational measures do not provide. In addition, the strength of the correlation coefficient depends on the range of the two data sets compared, not just on the association between the variables, since the correlation is always normalized with respect to variance. Figure 3 demonstrates an example of how the use of the correlation coefficient can be misleading. Figure 3A demonstrates a test–retest assessment of an acuity measure based on 74 eyes, where the correlation coefficient is large (0.93) and there is a high level of significance. In Fig. 3B, the correlation coefficient is much lower and is not statistically significant (the latter, in part, is due to the fact that there are fewer data points). Closer inspection reveals that the data set in Fig. 3B is just a subset of the data

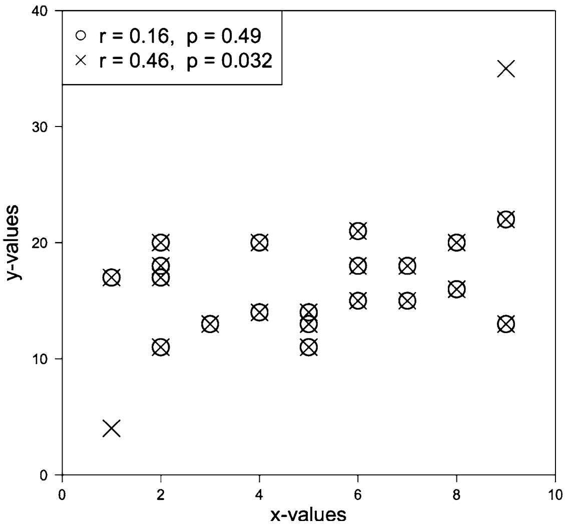

Fig. 2 The effect of outliers on correlation coefficients. The *circles* represent twenty data pairs for which the Pearson correlation coefficient is 0.16, which is not significant at the 0.05 level ($P = 0.490$). When just two data points are added at extreme positions (the two data points shown as crosses without overlapping circles), the correlation coefficient is increased to 0.46, which is statistically significant at the 0.05 level ($P = 0.032$)

in Fig. 3A. Fig. 3B represents a blowup of the lower left portion of Fig. 3A, comprising only eyes with good vision. This illustrates that test–retest assessment should not be based on the correlation coefficient, since just adding some very good scores and a single very poor score nearly guarantees an impressive correlation coefficient, independent of the quality of the measure being evaluated. The present example is rather extreme, but the principle applies to any comparison situation; in vision adding end-point disease cases and control subjects will nearly always lead to a—possibly irrelevant—significant correlation. If these acuity data were quantified as logMAR values, a better test–retest quality measure would be the 95% confidence interval of the differences which is actually a little lower in Fig. 3B (0.32 logMAR) compared to Fig. 3A (0.44 logMAR) [9].

Graphing data: standard deviation versus standard error of the mean

When graphing data, some investigators show the range of their data using standard deviation values while others use standard error of the mean values. What is the difference between these and when should each be used? The SD (standard deviation) represents the spread of the data set. With increasing sample size, it does not systematically change its value, rather it becomes more correct in representing the population spread. The SEM (or SE, standard error of the mean) is given by the formula (SEM $= \mathrm{SD}/\sqrt{n}$). Thus, the value of SEM decreases with increasing sample size.

So, which measure is the appropriate one? Naïvely, one could lean toward using SEM, because it is smaller, and therefore the graphs look nicer. Or one could be conservative and lean toward using SD, since then one is certain to be on the safe side. One could also (correctly) argue that it does not matter, since the statistical outcome does not depend on the size of an antenna in a graph. It does, however, affect the ability to use “inference by eye” [10, 11]. One glance at a figure should allow the reader to determine to what degree the result is significant or relevant. So, which measure (SD or SEM) is appropriate depends on the question asked or aspect to be highlighted.

Figure 4A depicts a situation where the SD, representing the spread of the population, is appropriate. Assume we have a single measure from a patient—say, the ERG 3.0 b-wave amplitude—and we want to assess whether it is ‘normal’. We will compare it with the normal range (the control mean ± 2·SDs), or the non-parametric equivalent

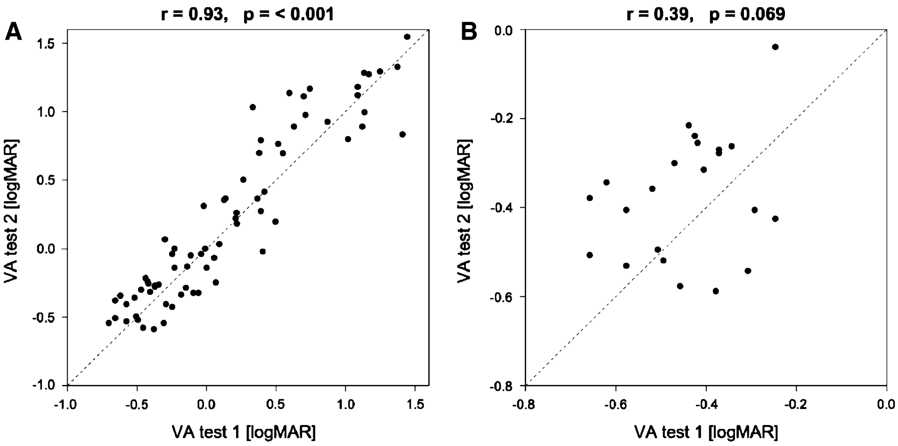

Fig. 3 Scatterplots showing the test–retest scores of a measure, where the *x*-axis shows the results of test 1 and the *y*-axis shows the results of the retest (test 2). Points on the dashed line have the identical outcomes for the test and retest; a good test method would have very similar outcomes on test and retest. Evaluation of the test–retest quality using the correlation coefficient, however, may lead to misleading conclusions. In this example, there is a high correlation coefficient for the data in **A**, and a lower correlation coefficient in **B**, suggesting that the data in **B** represent a poorer test method. In reality, the data in **B** are a subset of the data in **A** ($-0.7 < \mathrm{VA1} < -0.24$); the correlation coefficient is lower in **B** because it also depends on the *range* of the compared data sets, not just on their *association*. As a side note: when the *x*- and *y*-axes for scatterplots cover an identical range (as here), they should be square and not rectangular as offered by some software. This makes “inference by eye” [10] easier

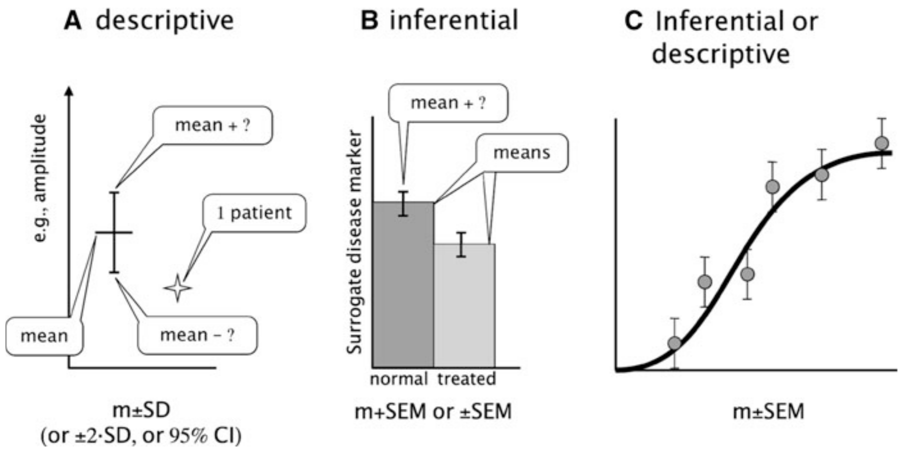

Fig. 4 **A**, **B**, **C** show cartoons of when to use standard deviation as an indicator of variance and when to use standard error of the mean

($\approx$the 95% confidence interval). In this example, as a *descriptive* statistic, the SD is useful because it provides information about the population distribution that would not be provided by the SEM (since the latter is "confounded" by sample size).

Figure 4B illustrates the SEM as an *inferential* statistic: it tells us how close the mean of a given sample or sub-population is to the "real" population mean. As an example, assume we have an experiment with two groups, treated versus untreated. We want to know whether the difference in the two means is statistically significant. Plotting SEM "antennas" allows immediate inference by eye by applying Cumming and Finch's [10] rules of thumb. Assuming similar variance in the two groups to be compared; if the SEM antennas overlap, then there is no significant difference. If the gap between the ends of the antennas is $\geq 1 \cdot$SEM bar, then the difference is significant at the 5% (0.05) level; if this gap $\geq 2 \cdot$SEM bars, then the difference is significant at the 1% level (0.01).

Figure 4C illustrates a situation where either SD or SEM units can be used to describe the data. In this case, we have a function derived from a model and we want to assess how well it fits the data. We have shown the data with SEM indicators, which give an immediate indication of the goodness of fit of the model: In the situation of a correct fit, only the mean of every third data point would be off (away from the curve) by more than the distance of one SEM bar (because ± the SEM covers 67% of the likelihood of the true population mean). If SEM indicators are longer, this indicates that a common variance source has not been factored out and therefore the model has missed accounting for this.

Conclusions

Statistics are tools that allow us to make inferences about our data. Every statistic is based on a series of assumptions about how the data set is being used to evaluate was derived. If these assumptions are violated, then the statistic is no longer a valid measure of the variability we are using it to assess and it becomes invalid. Therefore, it is important that investigators take the care with their data analysis that they take setting up their research protocols and obtaining their data. Statistics can greatly enhance our ability to learn from our results but when not carefully applied can also lead to misinterpretation. This basic primer was designed to help researchers avoid some common mistakes with data analysis. Much more sophisticated summaries exist and should be used for more complex topics.

Acknowledgments This work was supported by a grant from the Foundation Fighting Blindness, Columbia, Maryland and a grant from the Allene Reuss Memorial Trust, New York, New York.

References

1. Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H (2007) Statistical errors in medical research—a review of common pitfalls. Swiss Med Wkly 137:44–49
2. Murdoch IE, Morris SS, Cousens SN (1998) People and eyes: statistical approaches in ophthalmology. Br J Ophthalmol 82:971–973
3. Newcombe RG, Duff GR (1987) Eyes or patients? Traps for the unwary in the statistical analysis of ophthalmological studies. Br J Ophthalmol 71:645–646

4. Rosner B (1984) Multivariate methods in opthalmology with application to other paired data situations. Biometrics 40:1025–1036
5. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
6. Katz J, Zeger S, Liang KY (1994) Appropriate statistical methods to account for similarities in binary outcomes between fellow eyes. Invest Ophthalmol Vis Sci 35:2461–2465
7. Glynn RJ, Rosner B (1992) Accounting for the correlation between fellow eyes in regression analysis. Arch Ophthalmol 110:381–387
8. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
9. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
10. Cumming G, Finch S (2005) Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60:170–180
11. Cumming G, Fidler F, Vaux DL (2007) Error bars in experimental biology. J Cell Biol 177:7–11