

CHAPTER 1

Summarizing Data

Reporting Numbers and Descriptive Statistics

If choosing one summary statistic rather than another can even occasionally affect the clinical judgment of physicians reading a published article, then scrupulous attention must be paid to the use of summary statistics in the medical literature.

L. FORROW, W. C. TAYLOR, AND R. M. ARNOLD (1)

Descriptive statistics are numerical summaries of collections of data. Generating summary statistics is usually the first step in analyzing and presenting the results of a study because they reduce large amounts of data to a few, more manageable numbers. For example, listing the pulse rates of 5000 patients is seldom practical or desirable, but reporting the average pulse rate and perhaps the maximum and minimum pulse rates for the group is both practical and desirable. Here, the average, maximum, and minimum pulse rates are three descriptive statistics that summarize 5000 data points in three numbers.

We present here guidelines for reporting 1) numerical precision, 2) percentages, 3) categorical data, 4) continuous data, 5) paired data, 6) transformed data, and 7) data from small samples.

NUMERICAL PRECISION

1.1 Report all numbers with the appropriate degree of precision.

False (“spurious”) precision is undesirable and can be misleading. Reporting that the mean life expectancy is 22.085 years adds nothing to the fact that, for all practical purposes, mean life expectancy is 22 years. Ehrenberg (2) points out that readers actually can deal effectively only with numbers that contain no more than two significant digits. Thus, numbers should be rounded to two significant digits unless more precision is truly necessary. Compare these three statements (adapted from Ehrenberg):

1. The number of women physicians in training increased from 29,942 to 94,322, and that of men from 13,410 to 36,061.
2. The number of women physicians in training increased from 29,900 to 94,300, and that of men from 13,400 to 36,100.
3. The number of women physicians in training increased from 30,000 to 94,000, and that of men from 13,000 to 36,000.

The three-fold increase in physicians is less apparent in statement 1 because it is difficult to compare two five-digit numbers. Rounding to three significant digits in statement 2 works better, but the third digit still draws attention. In statement 3, however, the numbers have been rounded to two digits, and the approximate one-to-three relationship between them is much clearer.



Numerical data should be rounded when presented, not when analyzed (3). Information is lost when numbers are rounded, and this loss can affect the quality of the results. In the example above, the exact numbers of physicians in training may need to be reported for any number of reasons. Rounding helps readers to see the overall pattern of the results but should not be used when more accurate descriptions of data are necessary.



In most clinical and many biological studies, numbers with three or more decimal places should be examined for possible unnecessary precision. Some measurements can be made with a great deal of precision, and such precision is sometimes necessary to report. In biomedical research, however, highly precise measurements may be of little value. The smallest P value that need be reported, for example, is $P < 0.001$.

REPORTING PERCENTAGES

1.2 When reporting percentages, always give the numerators and denominators of the calculations.

The advantage of percentages is that they allow groups of different sizes to be compared on a common measure. The disadvantage is that perspective can be lost if only the percentages are given. Thus, a statement that 20% of patients were treated successfully is true for 1 of 5 patients, as well as for 1000 of 5000 patients. The numerator and denominator of the percentage can be given in parentheses or vice versa: 25% (650/2598); 33% (30 of 90 patients); 12 of 16 rabbits (75%).



Verify numerators and denominators and recalculate each percentage. A typical problem occurs when percentages are reported not for the entire sample but for subgroups of the sample. For example, “Of 1000 men

with heart disease, 800 (80%) had high serum cholesterol levels; of these 800, 250 (31%) were sedentary.” The 31% is $250/800$, not $250/1000$.

1.3 When the sample size is *greater* than 100, report percentages to no more than one decimal place. When sample size is *less* than 100, report percentages in whole numbers. When sample size is less than, say, 20, consider reporting the actual numbers rather than percentages.

The cutpoint of 20 to indicate a small sample is reasonable but arbitrary. Especially in small samples, percentages can be misleading because the size of the percentage can be so much greater than the number it represents: “In this experiment, 33% of the rats lived, 33% died, and the third one got away.”

1.4 When reporting changes in data as percent change, use the following formula: $[(\text{Final value} - \text{Initial value})/\text{Initial value}]$; multiply the result by 100 to obtain the percent increase or decrease.

Using this formula, if the result is a negative number, the minus sign is removed and the change is called a *decrease*. If the result is a positive number, the change is called an *increase*.

EXAMPLE

- A 10° *increase* in body temperature from 30°C to 40°C is a 33% increase: $(40 - 30)/30 = 0.33$. The 10° is one-third of the 30° .
- A 10° *decrease* in body temperature from 40°C to 30°C is a 25% decrease: $(30 - 40)/40 = -0.25$. The 10° is one-fourth of the 40° .

SUMMARIZING CATEGORICAL DATA

Sample Presentation

Of the 25 tumors, only 5 were malignant.

Here,

- The *ratio* of malignant to non-malignant tumors is 5:25.
- The *proportion* of malignant tumors is $5/25$, or 0.2.
- The *percentage* of malignant tumors is $(5/25) \times 100\%$, or 20%.
- After 5 years of follow-up, the tumor was malignant in 5 of the 25 patients, giving a 5-year recurrence rate of 20%. (A *rate* is associated with a time factor.)

1.5 Specify the denominators of rates, ratios, proportions, and percentages.

Categorical data (nominal or ordinal data) are counts of the number of participants or observations in each category. Such data are often described with percentages or other ratios. For example, if a sample is divided into four nominal categories on the basis of blood type, the number of patients in each category might be presented as four percentages, which total 100%. Although the numerators of ratios are often easily identified, the denominators may reflect the total group or a subgroup, so it is important to specify which group is being used as the denominator. Blood group AB may constitute 15% of all patients in the sample (say, 15 of 100) but 67% (12 of 18) of 18 patients with a certain condition.



Summarize categorical data in the text unless the number of categories is large enough to justify the use of a figure.

1.6 If continuous data have been separated by “cutpoints” into ordinal categories, identify the cutpoints and the rationale for choosing them.

Measurements of height for, say, 100 men may be treated as a continuous distribution on a scale of meters, or these measurements may be divided into three ordinal groups: short, medium, and tall men. Because, statistically, ordinal data are handled differently than continuous data, it helps to know when and why these categories were used. Dividing continuous data into ordinal categories can be undesirable because information is lost when individual values are lumped into fewer, more general categories. Dividing continuous data into ordinal categories can be desirable, however, if it simplifies calculations. A common example is the practice of analyzing age as a series of ordinal categories, rather than as a continuous variable.



Be cautious when interpreting ordinal data that have been treated as continuous data (4). A common but sometimes questionable practice is to treat a small number of ordinal categories as though they were continuous data. For example, a rating of the severity of disease might use a four-point scale: 1 = absent, 2 = mild disease, 3 = moderate disease, and 4 = severe disease. Scores from several patients might be combined to yield an average severity score of, say, 2.3. But these scores may not be realistic because the conceptual “distance” between the categories is not uniform. The “distance” between no disease and mild disease might be much “greater” than that between moderate and severe disease. Reporting the number of responses for each category, or the category that received the most responses (the **modal** score), may be a better way to summarize these data.

On the other hand, it is sometimes useful to average ordinal scores. For a seven-point scale used to rate satisfaction with a hospital stay, few people would object to an average presented as a fraction, such as 3.2 or 5.3. Even here, however, the average score is appropriate only if the scores are more or less normally distributed. If the scores are skewed, the median score (the score that divides the distribution into an upper and a lower half) is the most appropriate to report, and if the scores have a bimodal distribution, the two modal scores (the values of the two peaks of the bimodal distribution) are the most appropriate (*see Guideline 1.7*).

SUMMARIZING CONTINUOUS DATA

Sample Presentation

- *Antibody titers ranged from 25 to 347 ng/mL and had a mean (SD) of 110 ng/mL (43 ng/mL). If the data are approximately normally distributed, they are appropriately described with the mean and standard deviation.*
- *Antibody titers ranged from 25 to 347 ng/mL, with a median (interquartile range) of 110 ng/mL (61 to 159 ng/mL). If the data are markedly non-normally distributed, they are appropriately described with the median and interquartile range.*

1.7 Provide appropriate measures of central tendency and dispersion when summarizing data that have a continuous distribution.

Continuous data are data that, when graphed, form a distribution of values along a continuum. Such distributions can be summarized by appropriate measures of central tendency and dispersion. **Measures of central tendency**, such as the mean, median, or mode, indicate where on the continuum the data tend to cluster. **Measures of dispersion**, on the other hand, such as the standard deviation, range, or interquartile range, indicate the spread of the data over the continuum.

Distributions that form a “bell-shaped” curve are said to be “approximately normally distributed”; all other distributions are non-normally distributed. Approximately normal distributions can correctly be described with the mean and standard deviation; other distributions are better described with the median and the range or interquartile range.

The classic Tukey box plot (**Figure 1.1**) and a variation of the box plot rendered as a Cleveland dot chart (**Figure 1.2**) (5) are excellent for presenting either normally or non-normally distributed data (6). They can show the mean or median, standard deviation or interquartile range, 90% to 10% range, outlying values, and so on (*see Guideline 21.17 and Figures 21.13 and 21.15*).

It may also be useful to provide small histograms of the actual data to show the general shape of the distributions (**Figure 1.3**).

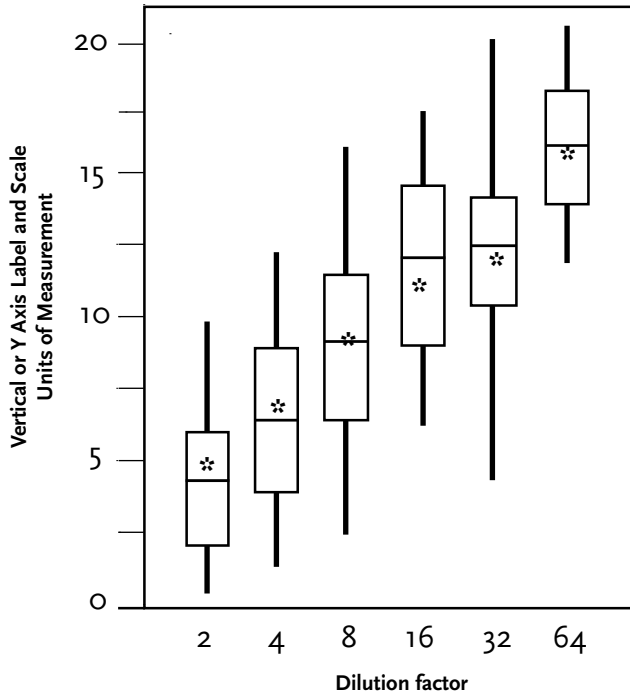


Figure 1.1

Tukey's box plot or (box-and-whisker plot) can summarize entire distributions in little space. Here, the box indicates the interquartile range; the horizontal line in the box, the median; and the asterisk, the mean. The "whiskers" indicate the range of the distribution. In other variations, the whiskers may indicate the range from, say, the 5th percentile to the 95th percentile, and the individual values at the extremes of the distribution will be graphed individually so that outliers can be identified.

1.8 Do not summarize continuous data with the mean and the standard error of the mean (SEM).

The standard error of the mean (SEM) is a measure of precision for an estimated population mean, whereas the standard deviation (SD) indicates the variability of the actual data around the mean of a single sample of a population. Unlike the SD, the SEM is not a descriptive statistic and should not be used as such. However, many authors incorrectly use the SEM as a descriptive statistic to summarize the variability in their data because it is always smaller than the SD, implying, incorrectly, that their measurements are more precise.

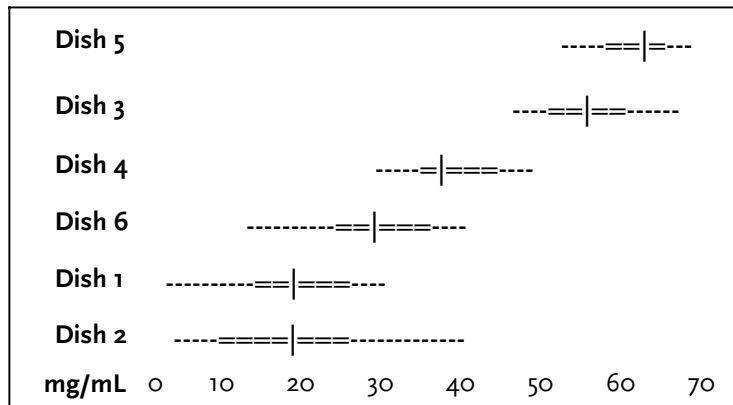


Figure 1.2

The classic Tukey box plots shown in Figure 1.1 can also be rendered using a variation of Cleveland's dot chart. Here, the median is indicated by a vertical line, the interquartile range by double lines, and the full range by the dotted lines.

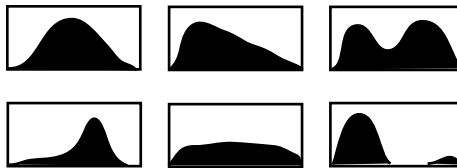


Figure 1.3

Small histograms can also show the general shape of a distribution of data without taking up much space. When descriptive statistics do not describe the distribution well or when they might be misleading, such histograms may provide a better sense of the data.

The SEM is correctly used only to indicate the precision of the estimated mean of a population. Even then, however, a 95% confidence interval (that is, the range of values encompassed by about two SEMs above and below the sample mean) is preferred (*see Chapter 3*).

EXAMPLE

- If the mean weight of a sample of 100 men is 72 kg and the SD is 8 kg, then (assuming a normal distribution), about two-thirds of the men (68%) are expected to weigh between 64 and 80 kg. Here, the mean and SD are used correctly to describe this distribution of weight for men.

However, the mean weight of the sample, 72 kg, is also the best estimate of the mean weight of all men in the population from which the sample was drawn. Using the formula $SEM = SD/\sqrt{n}$, where $SD = 8$ kg and $n = 100$, the $SEM = 0.8$. The interpretation here is that if (random) samples of 100 were repeatedly drawn from the same population of men, about two thirds (68%) of these samples would be expected to have mean values between 71.2 and 72.8 kg (the range of values between one SEM above and below the estimated mean). In this example, the preferred expression for the estimate of the mean and its precision is the mean and the 95% confidence interval (the range of values between about *two* SEMs above and below the mean). Here, the expression would be “The mean weight was 72 kg (95% CI = 70.4 to 73.6 kg),” meaning that if (random) samples of 100 were repeatedly drawn from the same population of men, about 95% of these samples would be expected to have mean values between 70.4 and 73.6 kg.

To summarize, for these data,

- The preferred presentation of the descriptive statistics is: mean (SD) = 72 kg (8 kg).
- The preferred presentation of the estimate and its precision is: mean (95% CI) = 72 kg (70.4 to 73.6 kg).

Presenting the estimate and its precision as the mean and SEM is discouraged because it is commonly confused with the mean and SD.



The SEM is often inappropriately used 1) instead of the SD to describe variability in a set of data and 2) instead of the 95% confidence interval to indicate the precision of an estimate.

Summarizing Normally Distributed Data

1.9 Use the mean and standard deviation (SD) only when describing approximately normally distributed data.

The mean and SD can be computed for any distribution of continuous data. For the average reader of the medical literature, however, the normal (Gaussian) distribution or “bell-shaped” curve is the only easily visualized distribution for which the mean and SD have meaning. That is, most readers know that about 68% of the distribution lies within the mean and plus and minus one SD; that about 95% lies within plus and minus two SDs; and that about 99% lies within plus and minus three SDs.

The mean and SD can be used correctly to describe other known distributions, such as the Poisson and chi-square, but such descriptions do not mean much to nonstatisticians. Thus, the mean and SD should be used only to describe data that are approximately normally distributed. Markedly non-normal distributions should be described with the median and range or interquartile range (*see Guideline 1.12*).

- ❗ **Most biological characteristics are not normally distributed (4,7–12).** Because most biological characteristics are non-normally distributed, the median and range or interquartile range, not the mean and standard deviation, should probably be the most common descriptive statistics in medical science.
- ↩ **Report means and standard deviations to no more than one decimal place more than the data they summarize (3,13–15).** As always, round to two significant digits when possible.
- ✓ **Data described with a standard deviation that exceeds one-half the mean are not normally distributed (assuming that negative values are impossible) and should be described with the median and range or interquartile range (10,11,16–18).** “Mean (SD) plasma values were 45 (25) mg/dL.” By definition, 95% of a sample of normally distributed data falls within about two SDs above and two SDs below the mean. Here, 95% of the range would run from -5 to 95 mg/dL, which is not possible [$45 - (25 + 25) = -5$; $45 + (25 + 25) = 95$], indicating that plasma values are not normally distributed.
- ✓ **Subtracting the median from the mean produces a crude estimate of the skewness of the data: the larger the difference, the greater the skewness (19,20).** In a normal distribution, the mean and the median values are about equal. When the mean is markedly greater than the median, the data are “right-skewed,” usually because a few high values increase the mean.

1.10 Do not use the “±” symbol when presenting the mean and standard deviation.

The “±” symbol is unnecessary because, by definition, the normal distribution is symmetrical and because, by definition, the SD extends an equal distance on both sides of the mean.

EXAMPLE

- Data are presented as “means and standard deviations” (not “means ± SDs”)
- The mean (SD) was “12 mL (2 mL)” (not “12 ± 2”).

A common source of confusion in the medical literature is the meaning of the interval defined by the “±” symbol. For example, “12 ± 2 mL” may be used to represent the mean and standard deviation (SD), the mean and the standard error of the mean (SEM), or even the estimate and the 95% confidence interval (95% CI) around the estimate. The “±” sign does not always imply that the number following it is the SD, which is why it should be replaced with a statement identifying the statistic as the SD or 95% CI. Unlike the SD and SEM, confidence intervals are not always symmetrical about the mean, so the “±,” even if it were an appropriate expression, would be inaccurate in many instances.



Do not report the standard error of the mean. The 95% confidence interval is the preferred form for describing the precision of an estimate, and its use requires that the upper and lower limits of the interval be given. For example, “The difference was 12 mL (95% CI = 10 to 14 mL).” (*See also Chapter 3.*)

1.11 When comparing the variability of two or more sets of normally distributed data, use the coefficient of variation instead of the standard deviation.

The variability of biological measures typically increases with the magnitude of the measures. For example, the variation in weight at birth is less than the variation of weight at death because as weight increases, so does the range over which it can vary. As a result, examining the variation between samples by comparing their standard deviations may be misleading. The coefficient of variation (CV) is useful because it incorporates both the mean and the standard deviation in a single measure.

The CV is simply the standard deviation expressed as a percentage of the mean. Thus, it gives a measure of dispersion relative to the size of the mean. So, for a mean of 12 and a standard deviation of 3, the CV is 25%.

EXAMPLE

- In **Table 1.1**, measure 1 has the least variation because it has the lowest CV.

The CV is particularly useful for comparing the variability of two or more sets of data with different units of measurement because it is expressed as a percentage, rather than in the units measured. For example, one diagnostic test might be reported as the area of an image, measured in square millimeters, and a competing test might measure the uptake of a radioactive tracer, measured in milliliters per minute. The relative dispersion of these two measurements could be assessed by comparing the CVs.



Check the CV with the formula: $CV = (SD/\text{mean}) \times 100\%$.

Table 1.1 Coefficient of Variation* and Standard Deviation for Comparing Variability among Measurements

Measure	Mean (SD), mm	Coefficient of Variation, %
1	90 (15)	16.7†
2	45 (15)	33.3
3	33 (13)	39.4

*The coefficient of variation is the standard deviation as a percentage of the mean.

† Measure 1, with the lowest coefficient of variation, has the least variability.

SD = standard deviation.

Summarizing Non-Normally Distributed Data

The mean and standard deviation are often incorrectly used to summarize all data, irrespective of whether the distribution is approximately normal or not, and even when the sample is too small to determine whether the distribution is normal. When a distribution cannot be said to be approximately normally distributed, it should be summarized with statistics other than the mean and standard deviation, as described below.

Data should be summarized appropriately not only to describe their distribution but for other statistical reasons as well. Approximately normally distributed data can be analyzed with what are called “parametric” statistical tests, but markedly non-normally distributed data should be analyzed with “non-parametric” statistical tests. In some cases, markedly non-normally distributed data can be “transformed” into a more normal distribution and analyzed with parametric tests (*see Guideline 1.14*), but both the non-normality of the distribution and the transformation should be reported. Many authors inappropriately use parametric tests on markedly non-normally distributed data.

1.12 Describe markedly non-normally distributed (skewed) data with the median and range (actually, the minimum and maximum values) or the interquartile range (actually, the values at the 25th and 75th percentiles).

When data are markedly non-normally distributed, the mean and standard deviation, although they may be mathematically correct, do not accurately communicate the shape of the distribution. The median (the 50th percentile) and interquartile range (the range of the values between the 25th and 75th percentiles of the distribution) provide a better summary of the distribution because they are not affected by extreme values. Other interpercentile ranges are sometimes used, such as the 10th to the 90th percentile.

Technically, the range is the difference between the minimum and maximum values. In common usage, however, the “range” usually refers to the minimum and maximum values. The same is true of interpercentile ranges: technically, the range is the difference between, say, the values at the 25th and 75th percentiles but the values themselves are typically reported.

EXAMPLE

- Median weight was 72 kg (25th percentile = 60 kg; 75th percentile = 87 kg).
- Median weight was 72 kg (interquartile range = 60 to 87 kg).
- After 8 weeks, weight (median and interquartile range) was 72 kg (60 to 87 kg).

REPORTING PAIRED DATA**1.13 Report paired observations together.**

Paired or matched data are observations taken from the same participant (such as pre-test and post-test data or data from the right side and left side of the same participant), or from different participants *matched* on certain characteristics to control for the influence of these characteristics on the outcome. Paired observations should be reported together so that the relationship between them is preserved.

Figures 21.26 and **21.27** show the individual changes that would not be apparent if only group means were presented for the pre- and post-test data. Paired data can be presented in tables, but, if so, the *differences* or *changes in the pairs* should also be presented and summarized. For example, the distribution of the differences should be described with, say, the median and interquartile range.

REPORTING TRANSFORMED DATA**1.14 Indicate whether and how markedly non-normally distributed data were transformed into an approximately normal distribution.**

A skewed distribution can sometimes be mathematically “transformed” into an approximately normal distribution (**Figure 1.4**), which makes subsequent analysis with “parametric” tests possible. Common transformations in medical science are the logarithmic transformation, the square-root transformation, the exponential transformation, and the reciprocal transformation.

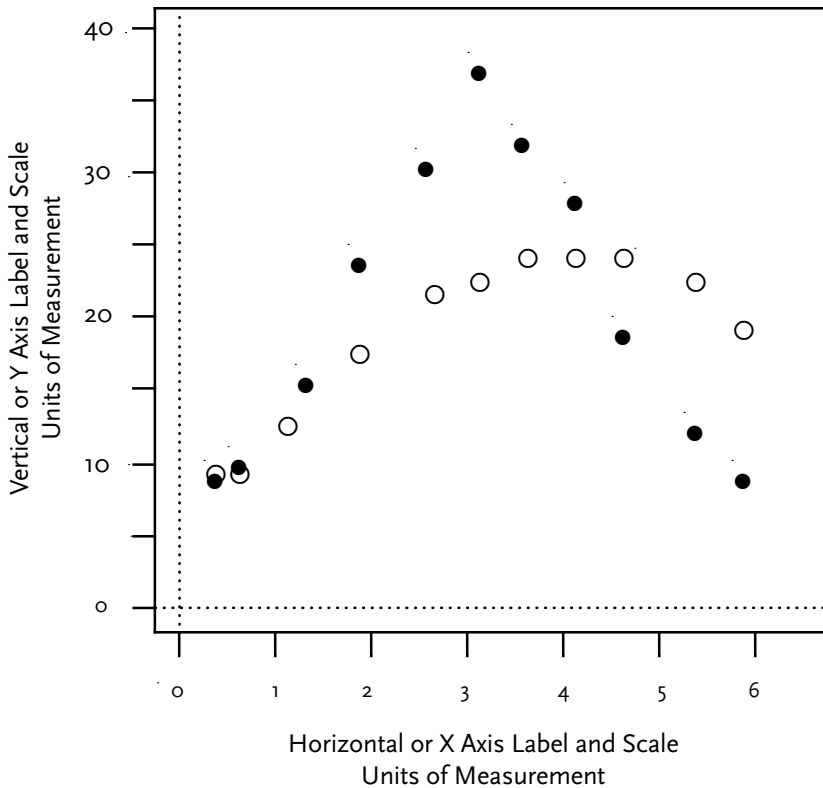


Figure 1.4

Non-normally distributed data before (*open circles*) and after (*closed circles*) mathematical transformation. After the analysis is complete, the results should be transformed back into their original scale so that the original units of measurement can be used. (The transformed distribution shown here is approximate; the transformation is not mathematically accurate.)

1.15 If data have been transformed, convert the units of measurement back to the original units for reporting.

When data are transformed, the unit in which they are expressed changes. For example, in a square-root transformation, “kilogram” becomes “square-root kilogram,” which has no real meaning. The results of the analysis, therefore, must be transformed back to be useful; that is, so that they can be correctly expressed in kilograms.

SUMMARIZING DATA FROM SMALL SAMPLES

1.16 If appropriate, present all the data when the number of observations is small or when descriptive statistics would be misleading.

Descriptive statistics are useful because they reduce large amounts of data to a few summary measures. If data do not need to be reduced or summarized, there is no need to use descriptive statistics.



Standard descriptive statistics (such as the mean and standard deviation) may not adequately summarize small data sets. Not enough data may be available, for example, to determine whether or not the distribution is normal. Means and standard deviations can be computed from as few as two data points, but these statistics mean little under these circumstances.

1.17 Avoid using percentages to summarize small samples.

When percentages are calculated for small samples, they can lose their meaning because only a few percentages are possible. For example, in a study of seven patients, one patient equals 14%, two equal 29%, three equal 43%, and so on. Thus, a table of adverse reactions may show only several entries of 14%, 29%, and 43%, which provides no more information than reporting that 1, 2, or 3 patients were affected.

A cutpoint of 20 to indicate a small sample is reasonable but arbitrary (*see Guideline 1.3*).

REFERENCES

1. **Forrow L, Taylor WC, Arnold RM.** Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med.* 1992;92:121-4.
2. **Ehrenberg AS.** The problem of numeracy. *Am Statistician.* 1981;35:67-71.
3. **Altman DG, Gore SM, Gardner MJ, Pocock SJ.** Statistical guidelines for contributors to medical journals. *BMJ.* 1983;286:1489-93.
4. **Haines SJ.** Six statistical suggestions for surgeons. *Neurosurgery.* 1981;9:414-8.
5. **McGill R, Tukey JW, Larsen WA.** Variation of box plots. *Am Statistician.* 1978;32:12-6.
6. **Simpson RJ, Johnson TA, Amara IA.** The box-plot: an exploratory analysis graph for biomedical publications. *Am Heart J.* 1988; 116:1663-5.
7. **Griner PF, Mayewski RJ, Mushlin AI, Greenland P.** Selection and interpretation of diagnostic tests and procedures: principles and applications. *Ann Intern Med.* 1981;94:553-600.
8. **Evans M, Pollock AV.** Trials on trial: a review of trials of antibiotic prophylaxis. *Arch Surg.* 1984;119:109-3.

9. **Feinstein AR.** X and ipr;: an improved summary for scientific communication [Editorial]. *J Chronic Dis.* 1987;40:283-8.
10. **Hall JC, Hill D, Watts JM.** Misuse of statistical methods in the Australasian surgical literature. *Aust N Z J Surg.* 1982;52:541-3.
11. **Hall JC.** The other side of statistical significance: a review of type II errors in the Australian medical literature. *Aust N Z Med.* 1982;12:7-9.
12. **Wulff HR, Andersen B, Brandenhoff P, Guttler F.** What do doctors know about statistics? *Stat Med.* 1987;6:3-10.
13. **Sumner D.** Lies, damned lies—or statistics? *J Hypertens.* 1992;10:3-8.
14. **Murray GD.** Statistical guidelines for the British Journal of Surgery. *Br J Surg.* 1991;78:782-4.
15. **Journal of Hypertension.** Statistical guidelines for the Journal of Hypertension. *J Hyper.* 1992;10:6-8.
16. **Brown GW.** Statistics and the medical journal [Editorial]. *Am J Dis Child.* 1985;139:226-8.
17. **Evans M.** Presentation of manuscripts for publication in the British Journal of Surgery. *Br J Surg.* 1989;76:1311-4.
18. **Gardner MJ.** Understanding and presenting variation [Letter]. *Lancet.* 1975;25:230-1.
19. **Oliver D, Hall JC.** Usage of statistics in the surgical literature and the 'orphan P' phenomenon. *Aust N Z J Surg.* 1989;59:449-51.
20. **Gore SM, Jones IG, Rytter EC.** Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *BMJ.* 1977;1:85-7.