Michael C. Costanza, Bruna Galobardes, Alfredo Morabia

Division of Clinical Epidemiology, Geneva University Hospital

Using percentiles to summarise data instead of means and standard deviations

Consider the distributions of daily total alcohol consumption (beer, wine, or hard liquor) by Geneva adults who reported drinking alcohol from 1993–2000 in Figures 1a and 1b. Most people (especially women) drink alcohol in relatively small quantities, but there are a few people (especially men) who drink larger quantities of alcohol. Because each distribution is skewed (has a long tail) to the right, the usual (arithmetic) mean would not necessarily be the most useful summary measure to characterise "typical" alcohol consumption. For example, the means for men and women are 23.1 and 10.7 g/day. If instead we use the geometric means (GMs, see below) of 13.7 and 5.5 g/day for men and women, we see they are very different from the means. Assuming for the moment that the GMs are preferable to the means in this case, it is nonetheless true that explaining exactly what these GMs *mean* to someone who needs to know and possibly do something about it but who is not particularly well-versed in statistical methodology (e.g., a public health official) may be difficult.

Alternatively, one can obtain valid and much more readily communicable results using *percentiles* such as those shown in Figure 2. Clearly, men consume much more alcohol (their 95% confidence (CI) for the median is 16.1 ± 0.8 g/day) than women (median 6.2 ± 0.5 g/day). Moreover, these two 95% CIs are not even close to overlapping.

In this Hinks & Kinks we briefly review summary measures of centre and spread, explain what percentiles are, why they may be more useful than means, and explain how to apply them.

"Typical" values: measures of centre (and spread)

The mean based on a random sample of observations from a population is commonly used to estimate a "typical" value (the "centre") of the distribution. (The standard deviation (SD) is a measure of typical "spread" of the data around the mean.) There are *many* measures of centre (also of spread). For example, the mode (= most frequent value) and the median (= 50th percentile, which separates the data into halves lying above or below it) are other centres.

Whether a particular centre is valid depends on the measurement scale. For example, the mode (but *neither* the median nor mean) is valid for categorical data (e.g., gender = {male, female}), while both the mode and median (but *not* the mean) are valid for categorical data that can also (but *only*) be ordered (e.g., clinical stage of breast cancer). Moreover, the shape of the distribution may indicate which measure of centre is most *useful*. For (approximately) symmetric distributions with a single maximum value (= the mode), such as height, weight, or systolic blood pressure, all three (sample) measures are (about) equal (although not equally precise). However, not all distributions are of this type. For example, the distributions of human mortality from all causes ("U-shaped"), and of alcohol consumption (not symmetric, see again Fig. 1) are not of this type.

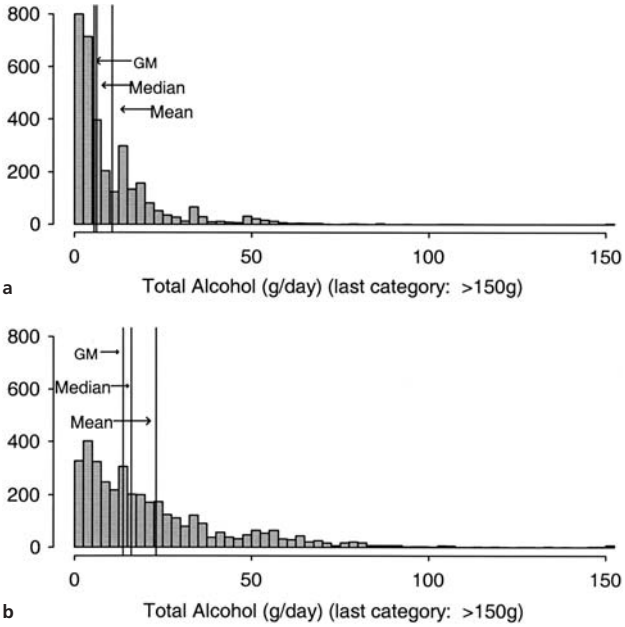

Figure 1 **a** Daily total alcohol consumption by 3820 Geneva MEN, 1993–2000; **b** Daily total alcohol consumption by 3280 Geneva WOMEN, 1993–2000

Soz.- Präventivmed. 47 (2002) 427–429
0303-8408/02/060419–03

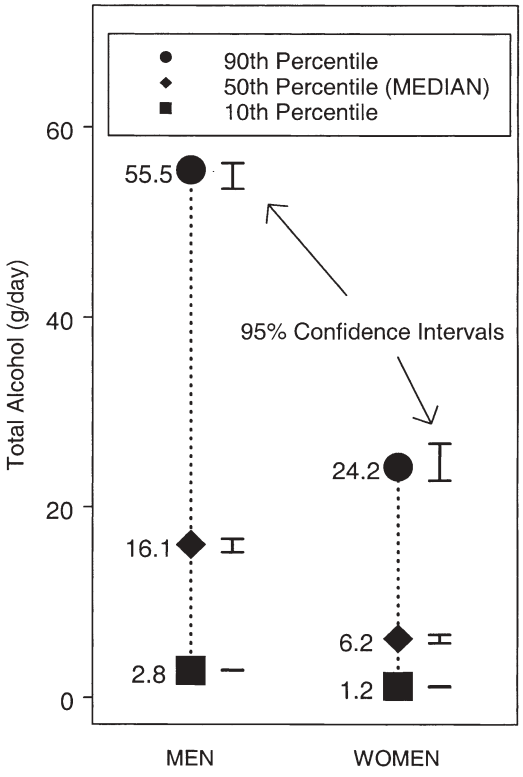

Figure 2 Daily total alcohol percentiles by Gender, Geneva, Switzerland, 1993–2000

For right skewed distributions, the median is often a more useful descriptor of centre than the mean. However, even in this case a common analytical approach is to log-transform the data and calculate the mean logarithm, which is then exponentiated (= "geometric mean" (GM)) to return to the original scale. (*N.B.*: GM ≤ mean.) This may lead to some confusion for at least two reasons: (1) interpretations of estimates (e.g., statistically significant?) on the log scale may not be (as) meaningful when transformed back to the original scale; and (2) the mean (*per se*, geometric, or otherwise) may *not* be a very meaningful center.

Percentiles as measures of centre (and spread)

One or more percentiles can be used to measure center (and spread), together with confidence intervals (CI), instead of the mean (and SD) – an approach long known but still rarely applied in health research.

The q^{th} percentile, P_q, is the value below which q% of the *ordered* data from the distribution occur. Thus, P_{50} (the median) is the value below which 50% of the ordered data occur, P_{75} has 75% of the ordered data to its left and 25% to its right, and so on. Besides the median, other percentiles at the extremes of the distribution (e.g., P_{10} and P_{90}) can provide a more complete description of the shape of a distribution. Moreover, differences between extreme percentiles can be used to measure the spread of the data. CIs for percentiles are based on the "order statistics", which are simply the n sample observations $\{ X_1, X_2, \ldots, X_n \}$transformed into { the minimum value = $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n-1)} \leq X_{(n)}$ = the maximum value }. (*N.B.:* Generally $X_1 \neq X_{(1)}$, $X_2 \neq X_{(2)}$, etc., and (at least) orderable data are needed to obtain the order statistics.)

In theory, a 95% (e.g.) CI for a percentile makes use of the binomial probability distribution (see Conover 1980). In principle, one can determine two order statistics, say $X_{(L)} \leq X_{(U)}$, such that the (exact binomial) probability is (around) 95% that $\{ X_{(L)} \leq$ population $P_q \leq X_{(U)} \}$ (in words: "for 95% of all possible samples the population P_q will be between $X_{(L)}$ and $X_{(U)}$). In practice, the values of L and U are well approximated (say for $n \leq 100$) by the relatively simple formulas: $L = nq - 1.96 \sqrt{\{nq(1-q)\}}$ and $U = nq + 1.96 \sqrt{\{nq(1-q)\}}$ (*N.B.*: in these formulas, q is a *proportion* (*not* %).)

For example, a 95% CI for the population median based on a sample of n = 400 would have $L = (400(.5)) - 1.96 \sqrt{400} \{.5(1-.5)\} = 200 - 19.6 = 180$ (after rounding down). Likewise, $U = 200 + 19.6 = 220$ (after rounding up). Thus, $(X_{(180)}, X_{(220)})$ is an approximate 95% CI for the population P_{50} (in words, "there is 95% confidence that the population median is between the 180th and 220th order statistics"). The (exact binomial)) procedure based on order statistics is *always* valid, even for small n, and it has at least two other advantages: (1) percentiles (and their CIs) are calculated and remain in the original measurement scale; and (2) percentiles are readily understood and easy to convey. (*N.B.*: a CI with confidence other than 95% is obtained by substituting another number for the "1.96" in the formulas for L and U (there is less (more) confidence if the number is less (more) than 1.96 (see Conover)).

Returning to the example of daily alcohol consumption by Geneva men and women, we see that in addition while the two P_{10}'s are very similar (albeit their very narrow CIs do not overlap), the P_{90} is much larger for men than for women. Also, the (very different) differences $(P_{90} - P_{50})$ and $(P_{50} - P_{10})$ clearly indicate that both distributions are right skewed. Finally, the differences, $(P_{90} - P_{10})$, indicate that alcohol consumption by men is much more spread around "the" center (the median) than is the case for women.

Recommendations and practical tips

Which and how many percentiles to use in order to characterise the data concisely and meaningfully depends on the particular objectives of the study. However, in most surveys attention is focused on estimating "the" centre (plus its spread = the confidence interval). As in the alcohol con-

sumption example, we frequently employ the median (50th percentile) for this purpose, as well as report the 10th and 90th percentiles to provide some idea of shape and of spread, *per se*.

For SAS Version 8 users (SAS Online Doc 1999), the UNIVARIATE procedure (through the new option CIPCTLDF) automatically provides the percentiles with confidence intervals (plus more). Another similar method makes use of so-called *box-plots* (Fisher & Van Belle 1993), which also facilitate identifying any "stray" (very extreme) values.

References

Conover WJ (1980). Nonparametric statistical inference, 2/e. New York: J. Wiley & Sons: 105–6.

Fisher LD, Van Belle G (1993). Biostatistics: a methodology for the health sciences. New York: J. Wiley & Sons: 53.

SAS OnlineDoc (1999). Version 8. Cary, NC: SAS Institute Inc.

Address for correspondence

Michael Costanza, PhD
Geneva University Hospital
Division of Clinical Epidemiology
Rue Micheli-du-Crest, 25
CH-1211 Geneva 14
Switzerland
Tel.: 41 22 372 95 52
Fax: 41 22 372 95 65
e-mail: Michael.Costanza@hcuge.ch