Reviewer's report

Title: Statistical modelling of volume of alcohol exposure for epidemiological studies of population health: the example of the US

Version: 2 Date: 16 July 2009

Reviewer number: 1

Reviewer's report:

Reviewer Report

Manuscript:

Summary:
The authors consider the problem of estimating the overall volume of exposure of alcohol in the adult (>=18 years of age) in the US by way of triangulating survey and per capita consumption derived from sales data using three types of distributions, the log-normal, gamma, and Weibull distributions. Triangulation is necessary since survey data generally underestimate total consumption and available data from sales, export, and import, or production inventories are not stratified by sex, age, race and/or other demographic characteristics.

The three selected distributions are applied to the 2001-2002 NESARC, a large representative US survey which samples the civilian, non-institutionalized adult US population in face-to-face interviews. Using recall data on alcohol consumption in the previous 12 months stratified by age, sex, and ethnicity, the authors fit the three distributions. The variance of the shifted gamma distribution, i.e., shifted such that total consumption equals sale of alcohol, was estimated from regression models of alcohol consumption survey data from 66 countries. The resulting prevalence distributions were assessed for goodness of fit using Chi-Square tests. The fitted exposure data are then used in an example to estimate the attributable fraction of alcohol consumption on liver cirrhosis. While all three distributions fit the data reasonably well, the gamma distribution has the added advantage that the estimated mean equals the empirical mean.

Is the question posed by the authors new and well defined?
1. The problem of estimating exposure in the presence of under-coverage of survey data, especially in the context of estimating overall alcohol volume exposure, is not new and triangulation is a means to solving the inconsistency between accurate but aggregated administrative data and individual level survey data with incomplete coverage. Per-capita estimates obtained from aggregated data are of limited use in estimating attributable fractions in burden of disease studies because the proportion of the disease often varies by sex, age, and
Are the methods appropriate and well described, and are sufficient details provided to replicate the work?

2. The reviewer understands that the mean of the shifted gamma distribution can be obtained from per-capita consumption based on sales data. Then, the relationship between the expected value and the variance of a gamma distribution, which is discussed in the first paragraph of the section Method for shifting the distribution, namely $E(X) = k*\theta$ and $Var(X) = k*\theta^2$ could be used to directly obtain an estimate of the shifted gamma distribution’s variance. Instead, the authors use survey data from 66 countries to build a linear multiple regression model to estimate the association between the mean and the standard deviation of gamma distributions fit to the survey data from these countries. Is this done to account for differences between men and women or to get uncertainty estimates for the empirical mean-variance relationship that could be used in a sensitivity analysis? Could the authors explain the underlying rationale for their approach a bit more in detail (although reference is made to previous publications)?

3. Regarding the truncation of extreme values exceeding 150g/day. The process as described is known as winsorization whereas truncation refers to the exclusion of the values exceeding a specified threshold. While estimators based on winsorized can be more robust because the technique reduces the influence of outliers, in the context of alcohol consumption, the outliers, i.e., right tail of the distribution, is of importance. The three selected distributions are all very similar (for appropriately selected shape parameters) but differ in their tails. It might thus be of interest to model how well each distribution accommodates the extreme right tails of the empirical data (grouped by age, sex, and ethnicity).

4. The formula for the standard deviation of a gamma random variable should read $\sigma = \sqrt{k*\theta^2}$.

5. The authors could elaborate a bit more on their efforts to test the two assumptions made by shifting the survey distribution to fit the per-capita level from the sales data. For example, would it be possible to obtain the percentage of abstainers and coverage rates by age, sex, and ethnicity from medical epidemiological studies and evaluate them for their consistency?

6. The estimated regression model linking the estimates of the standard deviation to the estimated mean does not have an intercept, was it not fitted or was it estimated to be zero?

Are the data sound and well controlled?

Yes.

Does the manuscript adhere to the relevant standards for reporting and data deposition?
Yes.

Are the discussion and conclusions well balanced and adequately supported by the data?

7. The three distributions considered were selected because they are unimodal and can be used to fit right-skewed data, which generally holds for alcohol consumption since a small fraction of the population consumes a large amount of alcohol while the majority of people consume little to modest amounts. Thus, the three distributions are mainly evaluated with respect to how well they fit the tails of the empirical distribution and their ability to be shifted to fit the aggregated total consumption.

8. The Weibull distribution outperforms the other two distributions but is more complicated to shift than the gamma distribution. The authors argue that the gamma distribution provides superior fit compared to the log-normal for the given data given the Chi-Square goodness of fit statistics. However, the reviewer finds the differences between the log-normal and the gamma to be more subtle. For example, upon simple enumeration of Table 2, the log-normal outperforms the gamma distribution in 6 out of 12 alcohol consumption groups for White Non-Hispanic men and in 8 out of 12 groups for White Non-Hispanic women. In Table 3, the log-normal does better than the gamma in 23 of 30 sex-age-ethnicity groups for all exposure levels and in 21 of 30 for exposures of up to 100g/day. The gamma distribution is more flexible than the log-normal to fit different shapes and scales and together with its ease of fitting may still be the distribution of choice but the reviewer thinks that a more detailed discussion is called for, including an explanation of why and how the Chi square statistic was used as a measure of the goodness of fit.

Is the writing acceptable?

Yes, with some specific suggestions:

9. Abstract: Background: is the term ‘consistent’ used in the statistical sense or to mean that individual exposure levels from surveys are adjusted to aggregate to the total level obtained from sales data?

10. Methods (p.8): “… based on data summer over a separate series…”

11. Methods (p.9): “… and results in sufficiently accurate…”

12. Methods (p.10): “… of the most widely used distributions in life data analysis…”


14. Methods (p. 12): “As described above, the shifted …”

15. Discussion (p.15): “…risk relations usually are also be derived…”

**Level of interest:** An article whose findings are important to those with closely
related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.