Modelling the true intake distribution from recall data with measurement/response errors

Alanko, Timo
Statistics Finland, Statistical Methods
Työpajakatu 13
Helsinki FI-00022 Finland
E-mail: timo.alanko@stat.fi

The present paper is concerned with modelling and estimating the 'true' intake distribution on the basis of survey recall scores from a short reference period. There is a long and highly sophisticated tradition for the problem (see e.g. Nusser et al., 1997, Dodd et al. 2006, Kipnis et al. 2009). We treat specifically the empirical case of estimating the true alcohol intake distribution from a series of Finnish alcohol consumption surveys. This paper is a revised and very condensed version, based on Alanko(1997). Space permits only a small proportion of the results to be presented.

Overview of the models

Two unobservable/latent variates Λ and Ξ representing the 'true' rate of consumption and the 'true' mean amount consumed on a single occasion are introduced.Observable variates L and $\bar{X}$ representing, respectively, the number of occasions in the reference period and the average amount per occasion, are defined at the individual level to follow the condensed Poisson and the gamma sampling distributions. The corresponding marginal distributions are derived as continuous mixtures of the individual level distributions by parametric prior distributions. 'True' volume consumed by a randomly chosen individual is defined as the product product $\Upsilon = \Lambda \cdot \Xi$ of the latent rate distribution and the latent mean amount consumed. The distribution of the true intake volume is derived and estimated by ML from the observed marginal distributions.

Modelling: the intake frequency

In our abundant empirical data, it was noticed that the average interoccasion lengths had a roughly linear relation to interoccasion standard deviations (of individual respondents). It was thus natural to examine individual consumption processes with constant CV's i.e. gamma distributed interoccasion lenghts. Moving to counts and examining the Gamma parameters, it was concluded that the Condensed Poisson distribution with parameter 2 would probably be a good approximation to an individual with a given process rate, say λ.Thus we would model the count distribution of consumption occasions of an individual with rate λ as:

$$
\begin{aligned}
\Pr(0 \text{ occasions in period}) &= P(0;\lambda) + \frac{1}{2}P(1;\lambda) = e^{-\lambda} + \frac{1}{2}e^{-\lambda}\lambda, \\
\Pr(l \text{ occasions in period}) &= \frac{1}{2}P(2l-1;\lambda) + P(2l;\lambda) + \frac{1}{2}P(2l+1;\lambda) \\
&= e^{-\lambda}\left[\frac{1}{2}\frac{\lambda^{2l-1}}{(2l-1)!} + \frac{\lambda^{2l}}{(2l)!} + \frac{1}{2}\frac{\lambda^{2l+1}}{(2l+1)!}\right],
\end{aligned} \tag{1}
$$

As individuals have very different rates of consumption, a standard technique is to treat the rate as a random variate with a flexible mixing distribution. For instance, we assumed that the rates over the population of individuals would vary according to the inverse Gaussian distribution, with

parameters $\mu_2 > 0, \beta_2 > 0$. The density of Λ, the rate distribution, would then be:

$$f_\Lambda(\lambda;\mu_2,\beta_2) = \sqrt{\frac{\beta_2}{2\pi\lambda^3}}\exp\left\{\frac{-\beta_2(\lambda-\mu_2)^2}{2\mu_2^2\lambda}\right\}. \tag{2}$$

and integrating over λ, one would get the marginal distribution of the number of occasions as

$$\begin{aligned} f_L(l;\mu_2,\beta_2) &= \frac{1}{2}(1-\delta_{l,0})f_K(2l-1;2\mu_2,2\beta_2) \\ &\quad + f_K(2l;2\mu_2,2\beta_2) + \frac{1}{2}f_K(2l+1;2\mu_2,2\beta_2). \end{aligned} \tag{3}$$

where $f_K(\cdot)$ is the probability function of the Poisson-inverse Gaussian distribution, PIG, defined for $k = 0, 1, \ldots$ and $\mu > 0, \beta > 0$ as

$$f_K(k;\mu,\beta) = \sqrt{\frac{2\beta}{\pi}}\frac{1}{k!}\exp\left\{\frac{\beta}{\mu}\right\}\left(\frac{\beta\mu^2}{\beta+2\mu^2}\right)^{\frac{1}{2}(k-\frac{1}{2})} K_{k-\frac{1}{2}}\left[\frac{\beta}{\mu}\sqrt{1+\frac{2\mu^2}{\beta}}\right], \tag{4}$$

with K_ν being the modified Bessel function of the third kind of order ν. The marginal distribution (3) could then be used for estimating the unknown parameters $\mu_2 > 0, \beta_2 > 0$ from consumption survey data.

Modelling: amounts consumed

In a similar vein to the number of occasions, the amounts consumed were modelled as compound distributions, with a Gamma distributed individual amount variation, and inverse Gaussian distributed inter-individual distribution for the individual mean parameter. Then the mean amount per occasion would be $\Xi \sim IG(\nu_2,\ \gamma_2)$ but the parameters would have to be estimated from a distribution of observables, in this case the observed average consumption $\bar{x}$ of an individual, given the number of occasions l The following is an example: For l observed amounts, the density of $\bar{X}|L$ which we will refer to as GIG, is, for $\bar{x} \geq 0, l = 1, 2, \ldots$ and $\nu_2 > 0, \gamma_2 > 0, \delta > 0$,

$$\begin{aligned} f_{\bar{X}|L}(\bar{x}|l;\nu_2,\gamma_2,\delta) &= \sqrt{\frac{\gamma_2}{2\pi}}\frac{2}{\Gamma(l\delta)}(l\delta)^{l\delta}\left(\frac{2l\nu_2^2\delta\bar{x}+\nu_2^2\gamma_2}{\gamma_2}\right)^{-\frac{1}{2}[l\delta+\frac{1}{2}]} \\ &\quad \times\exp\left(\frac{\gamma_2}{\nu_2}\right)\bar{x}^{l\delta-1}K_{l\delta+\frac{1}{2}}\left[\sqrt{\frac{2l\delta\gamma_2\bar{x}+\gamma_2^2}{\nu_2^2}}\right]. \end{aligned} \tag{5}$$

True intake distribution

Given the distributions of Ξ and Λ above, the population distribution of the volume consumed is expressed as the density of the product variable $\Upsilon = \Lambda \cdot \Xi$ on the condition $\Lambda \perp\!\!\!\perp \Xi$. The density of the product of the two variables is of the form

$$f_\Upsilon(\upsilon;\theta_\Xi,\theta_\Lambda) = \int_0^\infty f_\Xi(\frac{\upsilon}{\lambda};\theta_\Xi)\ f_\Lambda(\lambda;\theta_\Lambda)\frac{1}{\lambda}d\lambda, \tag{6}$$

where $f_\Lambda(\cdot)$, $f_\Xi(\cdot)$ and the parameter vectors $\theta_\Lambda = \{\mu_i, \beta_i\}$ and $\theta_\Xi = \{\nu_i, \gamma_i\}$ as defined previously. In practice, we are interested in the distribution function of Υ

$$F_\Upsilon(a) = \int_0^a f_\Upsilon(\upsilon)d\upsilon, \tag{7}$$

where a is a given consumption threshold or, alternatively, in the exceedance proportions $1- F_{\Upsilon}(a)$.Choosing, for instance the combination $\Lambda \sim IG(\mu_2, \beta_2)$ and $\Xi \sim IG(\nu_2, \gamma_2)$, the density for the volume consumed, or, more generally, the density of the product of two independent inverse Gaussian variates with different parameters is given by

$$\begin{aligned} f_{\Upsilon}(\upsilon; \mu_2, \beta_2, \nu_2, \gamma_2) &= \sqrt{\frac{\beta_2\gamma_2}{\pi^2\upsilon^3}} \exp\left(\frac{\beta_2}{\mu_2}\right) \exp\left(\frac{\gamma_2}{\nu_2}\right) \\ &\quad \times K_0\left[\frac{\sqrt{\upsilon(\beta_2\upsilon + \mu_2^2\gamma_2)(\gamma_2\upsilon + \nu_2^2\beta_2)}}{\upsilon\mu_2\nu_2}\right]. \end{aligned} \tag{8}$$

Empirical

In Alanko (1997) several combinations of mixing distributions and models are derived, estimated and tested. Some of the goodness-of-fit results are given in the oral presentation. The main empirical results of this study are the 'true' intake distributions estimated from Finnish alcohol consumption surveys samples for several consequtive surveys starting from 1976, separately for males and females.

Additional topics

In addition to the main results, the modelling enables prediction and inference, i.e. scoring concerning the respondents in the sample. For that purpose regression techniques, based on the Bayesian approach, are developed. The modelling approach employed gives also the opportunity to model (in a somewhat speculative sense) the mechanisms behind the survey measurement/response errors. Some suggestions for this are explored empirically.

REFERENCES

Alanko T. (1997), Statistical Models for Estimating the Distribution Function of Alcohol Consumption; A Parametric Approach, The Finnish Foundation for Alcohol Studies, Vol 44, Helsinki, 109 pages, ISBN 951-9192-61-1.

Dodd, K., et al. (2006), Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. Journal of the American Dietetic Association 106, 1640–1650.

Kipnis, V., et al. (2009), Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships between Episodically Consumed Foods and Health Outcomes. Biometrics, 65: 1003–1010.

Nusser, S.M., Fuller, W.A. and Guenther, P. (1997) Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-Hour Food Intake Data, in: Lyberg, L. et al., Survey Measurement and Process Quality, John Wiley&Sons, NeYork, pp. 689-709