Sampling of commonly used population characteristics: Is a normal approximation valid?

Jason A. Davis, Virginie Mittard, Rhodri Saunders

1. Coreva Scientific, Freiburg, Germany.

OBJECTIVES: Health economic models use a basecase that is generally representative of a subpopulation rather than the whole population. During sensitivity analysis, extrapolation of the model to other subpopulations or the whole population is estimated via sampling. Sampling is performed using summary statistics (e.g. mean and standard deviation) to inform generation of a distribution from which to draw values at random. Key population characteristics for healthcare include age, height, weight, and body mass index (BMI); all of which are commonly assumed to approximate to a normal distribution. Here the plausibility of this common assumption is tested. **METHODS:** Full data (N=451,075) were obtained from the 2010 Behavioral Risk Factor Surveillance System (BRFSS), a national, US, health-related, telephone survey. Data collected include age, gender, height and weight, with BMI being a calculated variable. Summary statistics and distributions were produced from the whole population. A sample of 2,500 records were extracted for in-depth analysis. Of these, 2,365 had complete data for age, gender, height, and weight. Analyses performed in R and Microsoft Excel® included subsampling, normality and Cullen-Frey plots. **RESULTS:** None of the data assessed were normally distributed. Cullen-Frey plots indicate that the best distributions to approximate the data are Beta, Log-normal, Beta, Log-normal for age, weight, height and BMI, respectively. Taking 1,000 subsamples of 236 patients, 39% of samples had a mean age falling outside of the 99% confidence interval for the population. For BMI the percentage was 38%. The ability of progressively smaller subsamples to represent the population was progressively worse. **CONCLUSIONS:** Many population characteristics of interest to healthcare do not follow a normal distribution. In the BRFSS dataset, the most descriptive distributions are the log-normal for BMI and the Beta distribution with negative skew for age. Age distribution skew may represent the aging population in the US setting.

Introduction

- Small samples of individuals are commonly taken to represent the characteristics of a larger population.
- Most health economic models are based on the assumption that the data are normally distributed (for example using mean ± sd in sensitivity analyses).
- Inspection of population samples can point to skewed distribution.
- Other, non-normal models may better describe certain parameters.

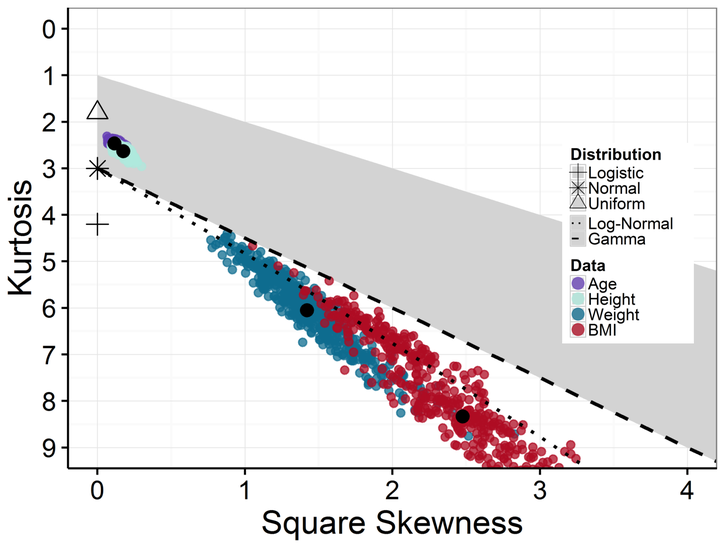
Methods

- Analysis of 2010 Behavioral Risk Factor Surveillance System data from the US Centers for Disease Control and Prevention
- Sample the dataset to create a test population on which to assess candidate distributions.
- Focusing on age data, compare goodness of fit for possible distributions.
- Compare fits for subsampled populations of varying sizes.
- Apply beta sampling for age to a model of capnography.

Results

- The test population of 2,365 complete cases was analysed according to distribution shapes after Cullen and Frey[1] (Figure 1).
- None of age, height, weight, or the derived body mass index (BMI) appears consistent with a normal distribution.

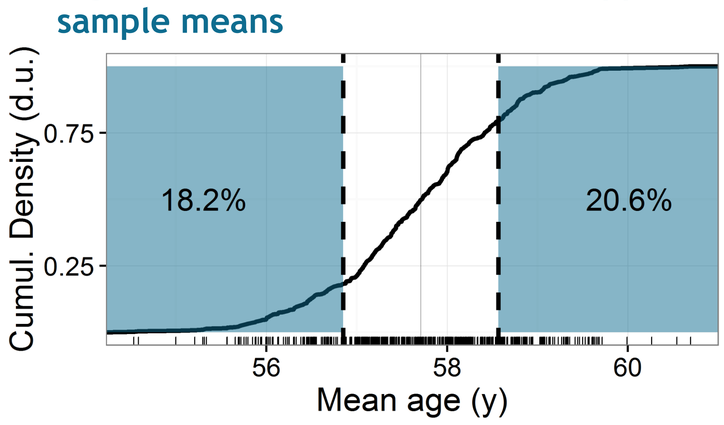
Figure 1: Shape comparison of age and BMI with common parametric distributions



Plot of kurtosis versus the square of skewness for a data sample of American adults. Points from data are in black, bootstrap replicates (500) are shown according to indicated colours

- Over one third of means from subsampled populations fall outside of the 99% confidence interval of the test population mean (Figure 2).

Figure 2: Distribution of bootstrapped sample means



A considerable percentage of sample means from bootstrap replicates fall outside of the 99% confidence interval of the true simulated population mean.

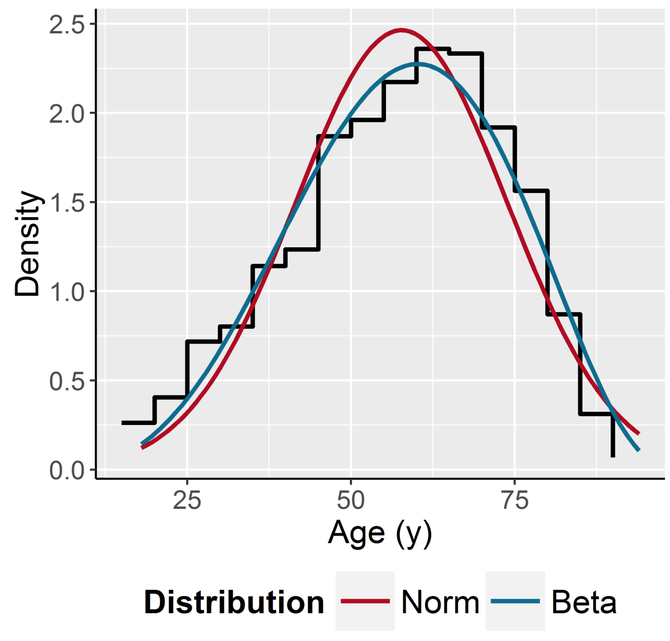
Is the normal distribution appropriate to model populations from samples?

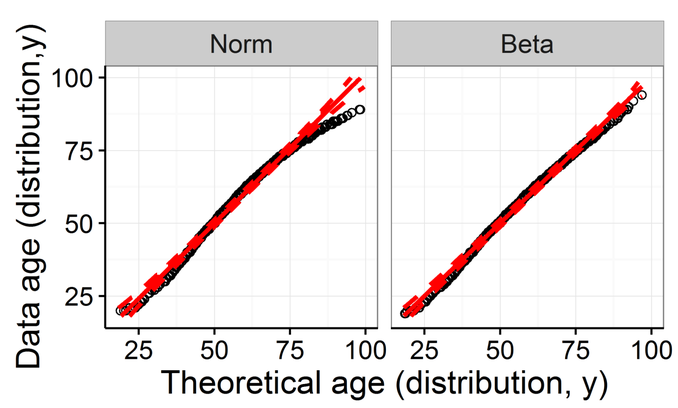
1. Self-reported US age data follow a beta distribution.
2. Height/weight (BMI) data follow a more complex skewed distribution.
3. Size of sample influences distribution model fitting.
4. Using the most suitable distribution may significantly change the accuracy of a model.

Alternatives to assumptions of normality warrant investigation to better describe data.

- With a focus on modelling age data, the alternative beta distribution (as suggested by kurtosis/skewness plot) appears to better fit the underlying data (Figure 3).

Figure 3: Shape comparison of age and BMI with common parametric distributions

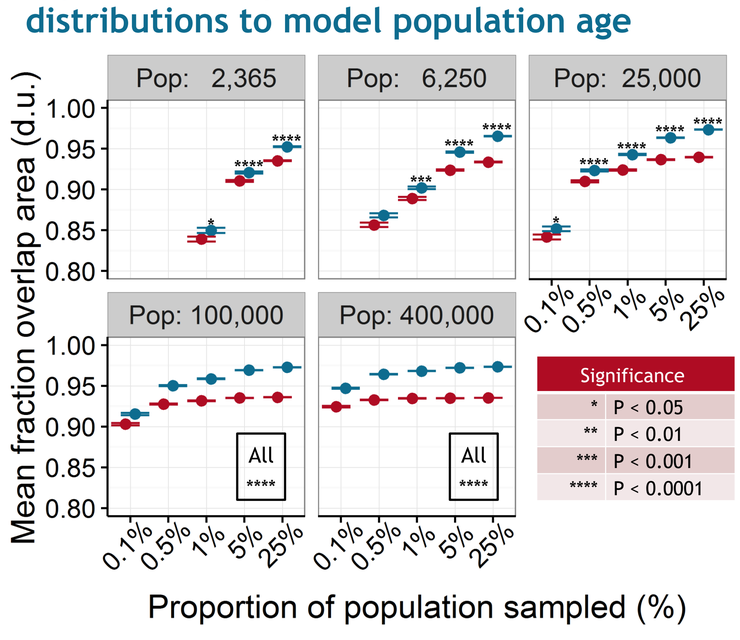




A density histogram is overlaid with best fits for normal and beta distributions (top). To better assess to quality of fit, quantile-quantile plots were generated of the two distributions (black circles) with the ideal (solid red) and a 99% confidence envelope (broken red curves) indicated.

- Using the multiple subsampled populations, goodness of fit was significantly better ($p < 0.01$) with a beta distribution compared with a normal distribution (Figure 4)

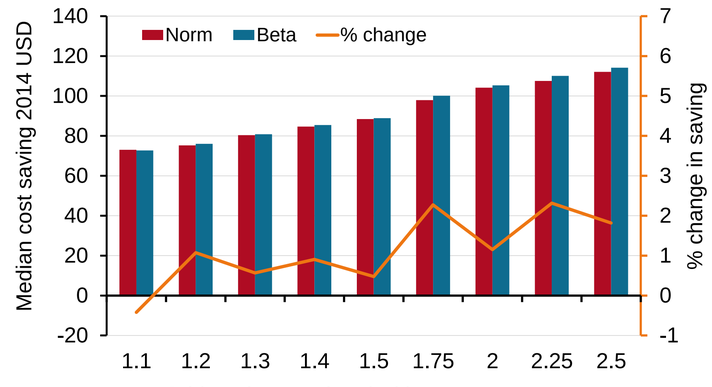
Figure 4: Comparison of normal and beta distributions to model population age



Means (±SEM) across 500 bootstrap replicates of subsampled populations were compared for the normal and beta distribution. Two-sided non-parametric (Mann-Whitney) tests of significance revealed a significant difference ($p<0.01$).

- Sensitivity analysis sampling age parameters using a beta distribution yields different results than the presumed standard normal distribution (Figure 5).

Figure 5: Budget impact of capnography[2] comparing two sampling methods



A sensitivity analysis of capnography yields slight differences in per patient annual savings after 500 bootstrap replicates.

Discussion

- The assumption of normality is common and convenient but more representative characterization may be achieved with an alternate distribution.
- The effect of the change in sampling to a better fitting distribution may impact results of health economic analyses, but depends on how different from normal most influential factors are.
- Greater application may be in more accurately modelling disease and disease burden across a population.

References: 1) Cullen and Frey. Probabilistic techniques in exposure assessment, 1999, pp.81-92; 2) Saunders et al., Endosc Int Open 2016; 04(03): E340-E351

jason@coreva-scientific.com
+49 (0)76 176 999 422