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

Jason A. Davis, Virginie Mittard, Rhodri Saunders
1. Coreva Scientific, Freiburg, Germany.

Methods

- Analysis of 2010 Behavioral Risk Factor Surveillance System data from the US Centers for Disease Control and Prevention (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 2,365 records were extracted for in-depth analysis. Of these, 2,363 had complete data for age, gender, height, and weight. Analyses performed in R and Microsoft Excel included subsampling, normality and Cullen-Frey plots.

Results

- The test population of 2,365 complete cases was analysed according to distribution shapes after Cullen and Frey (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

- 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

- 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

- 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

- 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 capnography3 comparing two sampling methods

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:

**Figures**

- Figure 1: Shape comparison of age and BMI with common parametric distributions
- Figure 2: Distribution of bootstrapped sample means
- Figure 3: Shape comparison of age and BMI with common parametric distributions
- Figure 4: Comparison of normal and beta distributions to model population age
- Figure 5: Budget impact of capnography comparing two sampling methods