Some Useful Formulas From the Statistics of Survey Sampling, I

Equally-Likely Outcomes & Counting

- If K outcomes $O_1, \ldots, O_K$ are equally likely, then the probability of any one of them is $1/K$.
- Consider taking a sample of n objects from a population of N objects.
  - Sampling with replacement, there are N^n possible samples of size n; the probability of any one of them is $1/N^n$.
  - Sampling without replacement, there are $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible samples of size n [where $N! = N \cdot (N-1) \cdot (N-2) \cdots 3 \cdot 2 \cdot 1$], so the probability of any one of them is $1\Big/\binom{N}{n}$.

Discrete Random Variables

Let X and Y be random variables with sample spaces $\{x_1, \ldots, x_K\}$ and $\{y_1, \ldots, y_K\}$ and distributions

$$P[X = x_i, Y = y_j] \;=\; p_{ij} \;\;,\quad P[X = x_i] \;=\; p_{i\cdot} \;=\; \sum_{j=1}^{K} p_{ij} \;\;,\quad P[Y = y_j] \;=\; p_{\cdot j} \;=\; \sum_{i=1}^{K} p_{ij}$$

Then, for example

$$E[X] \;=\; \sum_{i=1}^{K} x_i p_{i\cdot} \;\;,\quad Var(X) \;=\; \sum_{i=1}^{K} (x_i - E[X])^2 p_{i\cdot} \quad,\quad Cov(X,Y) \;=\; \sum_{i=1}^{K} (x_i - E[X])(y_i - E[Y]) p_{ij}$$

$$P[X = x_i | Y = y_j] \;=\; p_{ij}/p_{\cdot j} \;\;,\quad E[X|Y = y_j] \;=\; \sum_{i=1}^{K} x_i P[X = x_i | Y = y_j] \quad,\quad E[aX + bY + c] \;=\; aE[X] + bE[Y] + c$$

Random Sampling From a Finite Population

Consider a population of size N and a sample of size n. Let y_i be the (fixed) values of some variable of interest in the population (such as a person's age, or whether they would vote for Obama). Let

$$Z_i = \begin{cases} 1, & \text{if } i \text{ is in the sample} \\ 0, & \text{else} \end{cases}$$

be the random sample inclusion indicators, and let Y_i be the random observations in the sample. Then the sample average can be written

$$\overline{Y} \;=\; \frac{1}{n} \sum_{i=1}^{n} Y_i \;=\; \frac{1}{n} \sum_{i=1}^{N} Z_i y_i$$

The Z_i's are Bernoulli random variables with

$$E[Z_i] \;=\; \frac{n}{N} \;\;,\quad Var(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) \;\;,\quad Cov(Z_i, Z_j) = -\frac{1}{N-1}\frac{n}{N}\left(1 - \frac{n}{N}\right)$$

Confidence Intervals and Sample Size

(a) A CLT-based $100(1-\alpha)\%$ confidence interval for the population mean is $(\overline{Y} - z_{\alpha/2} SE \;,\; \overline{Y} + z_{\alpha/2} SE)$.

(b) For sampling with replacement from an infinite population, $SE = SD/\sqrt{n}$.

(c) For sampling without replacement from a finite population, the SE has to be multiplied by the finite population correction (FPC).

(d) For a given margin of error (ME, half the width of the CI) and confidence level $1-\alpha$, we can find the sample size by solving

$$z_{\alpha/2} SE < ME$$

for n. The same approach works for both SRS with replacement (using the SE in (b)) and SRS without replacement (using the SE in (c)).

Some Useful Formulas From the Statistics of Survey Sampling, II

Stratified Sampling

Consider H strata with population counts $N = \sum_{h=1}^{H} N_h$ and sample counts $n = \sum_{h=1}^{H} n_h$. Let $f_h = n_h/N_h$; $W_h = N_h/N$; and $\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$ in each stratum, and let $s_h^2 = \frac{1}{n_h-1}\sum_i (y_{ih} - \overline{y}_h)^2$ be the sample variance in each stratum. Then

$$\overline{y}_{st} = \sum_{h=1}^{H} W_h \overline{y}_h \ , \quad \mathrm{Var}\,(\overline{y}_{st}) \approx \sum_{h=1}^{H} W_h^2 (1 - f_h)\frac{s_h^2}{n_h} \ , \quad DEFF = \frac{\mathrm{Var}\,(\overline{y}_{st})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{\sum_{h=1}^{H} W_h^2 (1 - f_h)\frac{s_h^2}{n_h}}{(1-f)\frac{s^2}{n}}$$

Cluster Sampling

Consider a population of N clusters. We take an SRS $\mathcal{S}$ of n clusters, and all units within each sampled cluster (one-stage clustering). Assume clusters all have same size M. Let $\overline{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ in each cluster. Then

$$\overline{y}_{cl} = \frac{1}{n}\sum_{i\in\mathcal{S}} \overline{y}_i \ , \quad \mathrm{Var}\,(\overline{y}_{cl}) \approx \left(1 - \frac{n}{N}\right)\frac{1}{n} s_{\overline{y}_i}^2 = \left(1 - \frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}} (\overline{y}_i - \overline{y}_{cl})^2\right]$$

and

$$DEFF = \frac{\mathrm{Var}\,(\overline{y}_{cl})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{M s_{\overline{y}_i}^2}{s_{y_{ij}}^2} \approx 1 + (M-1)\rho$$

where $s_{\overline{y}_i}^2$ is the sample varance of the cluster means, $s_{y_{ij}}^2$ is the sample variance of the individual observations, and ρ is the intraclass (intracluster) correlation, or ICC.

Post-Stratification Weights and Means

As part of survey data collection it is a good idea to get general demographic information (e.g. in our surveys: sex, age, class, major, hometown, etc.). After data collection we compare the proportions in each of these categories in our sample with the same proportions in the population. If they agree, great. If not, calculate

$$w_i = (N_h/N)/(n_h/n) \text{ for each } i \text{ in post-stratum } h \ , \quad \textit{and} \ \ \overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

Post-Stratification Variance Calculations

<u>*Taylor series*</u>:

$$\mathrm{Var}_{TS}\,(\overline{y}_w) \approx \frac{1}{(\sum_i w_i)^2}\left[\mathrm{Var}\,(\sum_i w_i y_i) - 2\overline{y}_w \mathrm{Cov}\,(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 \mathrm{Var}\,(\sum_i w_i)\right]$$

where $\overline{y}_w$ is as above, $\overline{w} = \frac{1}{n}\sum_i w_i$, $\overline{wy} = \frac{1}{n}\sum_i w_i y_i$,

$$\mathrm{Var}\,(\textstyle\sum_{i=1}^n w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i - \overline{w})^2, \quad \mathrm{Var}\,(\sum_{i=1}^n y_i w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})^2,$$

$$\mathrm{Cov}\,(\textstyle\sum_{i=1}^n y_i w_i, \sum_{i=1}^n w_i) \approx n \cdot \frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})(w_i - \overline{w})$$

<u>*Jackknife*</u>:

- Replicate n times (by removing one obs. each time and recalculating weights):

$$\overline{y}_w^{(r)} = \frac{\sum_{i=1}^n w_i^{(r)} y_i^{(r)}}{\sum_{i=1}^n w_i^{(r)}}$$

- Calculate

$$\overline{y}_{JK} = \frac{1}{n}\sum_{r=1}^{n} \overline{y}_w^{(r)} \ , \quad Var_{JK}(\overline{y}_w) \approx \frac{n-1}{n}\sum_{r=1}^{n} (\overline{y}_w^{(r)} - \overline{y}_{jk})^2$$