1 Probability Proportional to Size (PPS) Sampling

A population $\mathcal{U}$ has N units each with some value $y_i \in \mathbb{R}$. You can choose n of them to label in order to estimate $Y := \sum_{i\in\mathcal{U}} y_i$. Ideally, you sample without-replacement, with fixed sample size, and with $\mathbb{V}[\widehat{Y}]$ minimizing inclusion probabilities π_i – how likely each item is to show up in your sample $\boldsymbol{s}$. The Horvitz-Thompson (HT) estimator maps $\boldsymbol{\pi}, \boldsymbol{s}$ to an unbiased estimate of the population total[1]:

$$\widehat{Y} = \sum_{i\in\boldsymbol{s}} \frac{y_i}{\pi_i} \qquad \text{(HT AKA expansion estimator)}$$

$$\mathbb{E}[\widehat{Y}] = \mathbb{E}[\sum_{i\in\boldsymbol{s}} y_i/\pi_i] = \mathbb{E}[\sum_{i\in\mathcal{U}} s_i y_i/\pi_i] = \sum_{i\in\mathcal{U}} \mathbb{E}[s_i] y_i/\pi_i = \sum_{i\in\mathcal{U}} \pi_i y_i/\pi_i = Y \qquad \text{(requires } \pi_i > 0\text{:)}$$

We give theoretical justification that **Probability Proportional to Size** (PPS) gives reasonable $\boldsymbol{\pi}$. Given some x_i related to y_i, PPS submits $\pi_i \propto x_i$.

$$\pi_i = n\frac{x_i}{X} \qquad \text{(PPS inclusion probs with } X := \textstyle\sum_k x_k\text{)}$$

We ensure $\mathbb{E}[|\boldsymbol{s}|] = n$ because $\sum_i \pi_i = n$, and fixed-size algorithms satisfy $|\boldsymbol{s}| = n$ for all $\boldsymbol{s}$ by definition. For any sampling design[2], derive the variance by observing that y_i, π_i are fixed constants, and the random variable is s_i, the indicator on whether unit i is included in the sample $\boldsymbol{s}$.

$$\mathbb{V}[\widehat{Y}] = \mathbb{V}[\sum_{i\in\mathcal{U}} \frac{s_i y_i}{\pi_i}] = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}} \frac{y_i y_j}{\pi_i \pi_j}\mathbb{C}[s_i, s_j] = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}} \frac{y_i y_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i\pi_j)$$

Now we make a strong assumption called "superpopulation modeling"

Assumption 1 (Linear Labels)**.** $\exists a \in \mathbb{R} : \forall i \;\; y_i = ax_i + x_i\epsilon_i$

For example, hospital costs are directly proportional to number of patients, and one may reasonably claim that the randomness is the costs scales with the number of patients. Under PPS, we have

$$\frac{y_i}{\pi_i} = \frac{X(ax_i + x_i\epsilon_i)}{n \cdot x_i} = \frac{X}{n}(a + \epsilon_i)$$

$$\mathbb{V}[\widehat{Y}] = \frac{X^2}{n^2}\sum_i\sum_j (a+\epsilon_i)(a+\epsilon_j)\mathbb{C}[s_i, s_j] = \frac{X^2}{n^2}\mathbb{V}[\sum_i \epsilon_i s_i]$$

because

$$a^2\mathbb{V}[\sum_i s_i] = c^2\mathbb{V}[|\boldsymbol{s}|] = 0 \qquad \text{(fixed-size designs always have } |\boldsymbol{s}| = \text{n)}$$

$$a\sum_i\sum_j d_i\mathbb{C}[s_i, s_j] = c\sum_i \epsilon_i\mathbb{C}[s_i, \sum_j s_j] = 0 \qquad \text{(} \textstyle\sum_j s_j = \mathbb{E}[\boldsymbol{s}] \text{ because it is fixed-size)}$$

For an arbitrary fixed-size design, PPS depends only on irreducible noise. The most common variable-size design is Poisson Sampling, which flips an independent π_i biased coin at each unit. For that design, the optimal policy is $\pi_i \propto y_i$ for any superpopulation $y_i = f(x_i)$ because $\pi_{ij} = \pi_i\pi_j$ and thus $\mathbb{V}[\widehat{Y}] = \sum_i \frac{y_i^2}{\pi_i} - y_i^2$. A common strategy is to learn $\widehat{f} \approx f(x)$ and set $\pi_i \propto \widehat{f}(x_i)$, though this does not account for estimation errors in $\hat{f}$. When the estimate $\widehat{f}(x_i)$ thinks y_i is small it sets π_i small, which blows up the variance because a large y_i is divided by a small number. This issue extends to any PPS implementation, regardless of design.

[1] For the mean, just divide by N

[2] A design is a map from π to a distribution over possible samples. For example, Simple Random Sampling (SRS) AKA uniform-weight without-replacement fixed-size sampling is the WoR FS design that assigns equal probability to all samples. It happens to be easy to implement, but some designs are not.