# Weighted data

Christopher A. Sims
Princeton University
sims@princeton.edu

February 26, 2018

# The model

We have i.i.d. data $\{y_i, w_i\}_{i=1}^n$, where observation $i$ is of a type that has been over- or under-sampled. That is, we can think of the sampling process as first drawing a potential $(y_i, w_i)$ pair, then with probability $w_i$ including $(y_i, w_i)$ in the sample. We consider both the case where the rejected draws, along with their $w_i$ values, are observed (while the corresponding $y_i$ is not) and the case where the data include only the retained $(y_i, w_i)$ pairs.

This framework can also accommodate oversampling, where we know that some types of observations $i$ have been overrepresented in the sample. This corresponds to allowing $w_i > 1$.

# Objective of inference

- We are interested in estimating the mean over the "population" of some function of $y_i$, or more generally in the population distribution of $y_i$.

- The interpretation of "population" may differ:
  - Non-random, finite, and much larger than our sample ("design-based" theory) or
  - a probability model for generating $y_i$ draws ("model-based" theory).

# Dependence between $y$ and $w$

- The simplest case is where we know that $y_i$ and $w_i$ are independent. Then we can carry out inference for the population distribution of $y_i$ by using the $y$ data alone, ignoring the observations on $w$.

- This setup is interesting only because we often think $w$ and $y$ are not independent. We might for example over-sample some ethnic, regional, or gender subgroups, expecting that the subgroups will have differing distributions of $y$. Or in the selection bias ($w < 1$) case, that people with different tendencies to drop out of a study have different distributions of $y$.

## Independence between selection and $y$, conditional on $w$

- We do maintain the assumption that, conditional on $w_i$, the distribution of $y_i$ is independent of whether the observation is included in the sample or not.

- Example: suppose women are twice as likely as men to have dropped out of the sample, so $w_F = \frac{1}{2} w_M$, where $F, M$ indexes women and men. We assume that the women in the sample for whom $y$ is observed, have the same $y$ distribution as those women who dropped out.

- If actually young women were more likely to drop out than old women, and the distribution of $y$ differs between young and old women, our assumptions are violated. We can't then proceed without obtaining data on which people are young and which old, and on the dropout probabilities of these subgroups.

## Common sense use of weights with only a few groups

- Suppose there are $k$ groups, that within group $j$ all $w_i$ values are the same, that for each group $j$ we know $\pi_j$, the proportion of the population that is in group $j$, and that every group $j$ is represented by $n_j > 1$ observations in our sample.

- Then the natural estimate of the population mean $E[y]$ of $y$ is

$$\sum_{j=1}^{k} \bar{y}_j \pi_j \ , \text{ where } \bar{y}_j = \frac{1}{n_j} \sum_{i \in S_j}^{n_j} y_i \ ,$$

  and $S_j$ is the set of $i$ indexes in the sample that fall in group $j$,

## Distribution theory for this simplest case

- Each $\bar{y}_j$ is a sample mean. If the $y_i$ distribution within each group is normal, standard small-sample distribution theory is available for the group sample means.

- Bayesian inference, with a prior on the group distributions, will allow easy sampling from the posterior, even if the variances of the $y$'s differ across groups.

- If the $y$'s within groups are non-normal but finite variance, and there are many observations in each group, frequentist asymptotic distribution theory will apply.

- Under usual regularity conditions, Bayesian inference based on the normality assumption will in large samples have the approximately the same frequentist properties as does the "natural estimate".

## What if the $\pi_j$ are unknown, but $N$ is known?

$N$ is the total number of observations, including those for which $y_i$ is not observed.

$$E\left[\frac{n_j}{N}\right] = \pi_j w_j \; .$$

In large samples, $n_j/N$ should estimate $\pi_j w_j$ with small error, and we know $w_j$. So why not substitute the implied estimate of $\pi_j$ into our "natural estimate":

$$\sum_{j=1}^{k} \bar{y}_j \frac{n_j}{w_j N} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_j}{w_{j(i)}} \; ,$$

where $j(i)$ is the group number for the $i$'th observation. Note that, if we use all the observations, including those for which $y_i$ is missing, inserting zeros for $y_i$ wherever it is missing, the right hant side above is simply the average over all $N > n$ observations of $y_i/w_i$. This is what is known as the Horwitz-Thompson estimator.

# Frequentist properties of Horwitz-Thompson

- When $\pi$ is in fact known, H-T is worse than the natural estimator, because it does not use our knowledge of $\pi$, but instead replaces $\pi$ with a noisy estimate.

- It is unbiased: $E[\bar{y}_j] = \mu_j$ (where $\mu_j = E[y_i \mid i \in S_j]$) regardless of the value of $n_j$, and as we already noted $E[n_j] = w_j \pi_j N$, which delivers the unbiasedness conclusion.

- To say anything useful about its small-sample distribution beyond unbiasedness, we have to make assumptions about the distribution of $y$ within each group.

## Pretending we can't see $w$

- The numerator of $H - T$ is a sum of $n$ i.i.d. variables of the form $y_i/w_{j(i)}$. They all have finite mean, and their variance may well be finite.

- Finiteness of the variance depends on $w_i$'s near zero being extremely rare, which may sometimes be hard to be certain of.

- With finite variance, the sum will be asymptotically normal, and the usual methods to form an asymptotic approximation to the distribution of a sum of i.i.d. variables will apply.

# H-T wastes information

- Consider what happens if all the groups have mean $y$ values that are very similar, but the weights vary over a wide range.
- Then uncertainty about the mean of $y$ must be small, since the groups have nearly the same means.
- The H-T estimator can nonetheless show very high, even infinite, variance.
- This is because H-T ignores the fact that we can see $w_{j(i)}$ and $y_i$ separately, not just their ratio, and this throws away information.

# Likelihood function

With a small number of groups, a natural parameterization and likelihood is

$$\prod_{i=1}^{N} \big(q(y_i \mid \theta_{j(i)}) w_{j(i)}\big)^{\delta_i} \big(1 - w_{j(i)}\big)^{1-\delta_i} \pi_{j(i)} \ ,$$

where as before $\pi_j$ is the population frequency of group $j$ and $\delta_i$ is an indicator function that is one when the $y_i$ is observed, zero otherwise. If we don't see non-selected observations, then the expression has to condition on $\delta_i = 1$, so it becomes

$$\prod_{i=1}^{n} q(y_i \mid \theta_{j(i)}) \frac{\pi_{j(i)} w_{j(i)}}{\sum_{j'} w_{j'} \pi_{j'}} \ .$$

# Inference based on likelihood

- Both these likelihood functions factor into one piece dependent only on the data and the $\theta_j$ parameters and another that depends only on the $\pi$ parameters and the observed sequence of $w_j$ values.

- Furthermore the $\theta$ part of the likelihood breaks into separate pieces for each group.

- So if our prior makes $\pi$ and $\theta$ independent, and the $\theta_j$'s independent across $j$, inference on theta can proceed group by group, and inference about $\pi$ will involve only the $w_i$'s and counts of numbers of observations in each group.

- Of course if we want to do inference about, say, the population mean of $\theta_{j(i)}$, we will need to form a joint posterior over $\theta_j$'s and $\pi_j$'s, but this is in principle straightforward.

# Large numbers of groups

- When the number of groups is large, the number of obsevations in each group may be small, at least for some groups. Indeed there may be groups with no observations. In fact it can easily happen that *most* groups have no observations.

- We have been assuming $w_j$ is constant within groups. In some applications the weights vary over a wide range, with most observations having a unique value of $w$ and many possible values of $w$ not represented in the sample.

- If we stick to a prior that makes $\theta_j$ independent across groups, the posterior on $\theta_i$ for those groups not represented in the sample is entirely determined by the prior. If most groups are not represented in the sample, inference is dominated by the prior, with little influence from the sample.

# A more reasonable prior

- With a large number of groups, many not represented in the sample, a prior on $\{\theta_j\}$ that makes $\theta_j$ independent across groups makes no sense.

- If we are estimating the population mean of $y_i$ or of $\theta_{j(i)}$ from a random sample of the total population, we must believe that our random sample contains information about the rest of the population. A prior that makes them independent across groups rules this out.

- But we might make the $\theta_j$'s **exchangeable**, rather than independent, by taking the population to itself have been drawn i.i.d. from, e.g., a $N(\bar{\theta}, 1)$, where $\bar{\theta}$ is unknown and has some prior distribution.

- With this prior, our sample of $y_i$ values carries information about $\bar{\theta}$, and thereby about $\theta_j$'s for unobserved groups.