Sampling: A Brief Review

"Workshop on Respondent-driven Sampling Analyst Software"

Hanoi

2013

Purpose

- To review some of the influences on estimates in design-based inference in classic survey sampling methods

- To give context to some of the challenges of estimation in respondent driven sampling

Reprise Lessons from Statistics 101

- Suppose we tossed a coin **100** times and it came up heads **20** times.
- We would estimate the probability(head) as **0.2**
- And using a formula we could look up in our textbook, and calculate a confidence interval.

Figure 1

- The width of confidence intervals decrease with increasing sample size (all other things being equal)

Figures 2 and 3

- There is more variability as the probability approaches 0.5

Lessons from Statistics 101

Another view: Figure 4

- Now instead of rigging things so I find p = .2 in the sample (for which p = .2 in the population is the best point estimate); I set P = .2 in the population (so to speak)

- So if I toss a coin with probability(heads) = .2, 100 times, I will not get 20 heads every time. Instead there will be variability.

- I did this 1000 times for each sample size.

- The black dot shows the median for these 1000 trials of Sample Size coin tosses

- The box shows the 25th and 75th percentiles -- 50% of the 1000 different estimates of P fell within this box and the other 50% outside of it.

- The two blue lines show the 2.5th and 97.5th percentiles: 95% of the 1000 different estimates fell within these lines

Design-based Inference (beyond Statistics 101)

- The population is viewed as fixed.
- The values of the variables of interest are fixed.
- Sampling is the only stochastic process, the only source of uncertainty.
- This is in contrast to model-based inference that posits an underlying data-generating model. Hence the variables of interest in the population are viewed as random variables.
- Survey sampling methods are typically design-based.
- To date, published RDS methods are design-based.

Probability Sample (necessary for design-based inference)

1. Every individual in the population must have a non-zero probability of ending up in the sample (written π_i for every individual i)

2. The probability π_i must be know for every individual who does end up in the sample

3. Every pair of individuals in the sample must have a non-zero probability of both ending up in the sample (written π_{ij} for every pair of individuals (i, j))

4. The probability π_{ij} must be known for every pair that does end up in the sample.

Lumley(2010)

This is the minimal requirement for estimating the mean and variance of any statistic. Ideally one would like to know the probability for any set of individuals in the sample, not just pairs.

Simple Random Sampling (random sampling without replacement)

is a sampling design in which *n* distinct units are selected from the *N* units in the population in such a way that every possible combination of *n* units is equally likely to be the sample selected. *Thompson(2002)*

In without replacement sampling, the sampling probability for an individual changes at each draw (draw-wise probability).

The probability that a particular individual *z* is chosen on the first draw is $p(z) = n/N$.

If *z* is chosen on the first draw, then the probability that *z* is chosen on the second draw is 0.

If *z* not chosen on the 1st or 2nd draw then the probability that *z* is chosen on the third draw is $p(z) = (n-2)/(N-2)$.

However, the *list-wise* inclusion probability for individual z, the probability that *z* will be in any given sample regardless of its size

$p(z \text{ in sample}) = n/N$

for a simple random sample

Repeated Simple Random Sampling from a Finite Population

Figure 5

- The line at Sampling Fraction Percent = 50 is the result of creating a population of size 1000, that is, a population size where 500 is 50%.
- In this population, P = .2, that is, 20% of the individuals are assigned a value of 1 (disease) and 80% the value 0 (non-disease).
- Then I sample 500 individuals without replacement from this population
- Calculate the proportion in the sample with the disease as this is my best estimate of the proportion in population.
- I repeat this 1000 times and record this estimate each time.
- As before, the black dot shows the median for these 1000 trials 500 draws without replacement from the finite population. The box shows the 25th and 75th percentiles. And the two blue lines show the 2.5th and 97.5th percentiles: 95% of the 1000 different estimates fell within these lines

Repeated Simple Random Sampling from a Finite Population

Figure 5

- As the sampling fraction increases, holding the sample size constant, the variability in the estimate due to sampling decreases.

- This makes sense as in the limit, a sampling fraction of 100% would be the entire population and then the variability would be 0.

- So what does a sampling fraction of 0 mean? Conceptually, an infinite population. This is what we were doing in Figures 1 through 4. In some sense, we generated these figures using a data-generating process, not a sampling process.

- In real life, we need to sample from a known finite population. So the 0 sampling fraction on this plot stands in for a very large population relative to the sample size. That is, a small enough sampling fraction as to provide negligible reduction in the variability of the estimate.

Repeated Simple Random Sampling from a Finite Population

- The standard error of the mean is reduced by

$$\sqrt{\frac{N-n}{N}}$$

- This is called the **finite population correction**.

Weighted Sampling

- An individual's sampling weight is the inverse of that individual's sampling probability $1/\pi_i$

- The fundamental statistical idea behind all of design-based inference is that an individual sampled with a sampling probability of π_i represents $1/\pi_i$ individuals in the population. *Lumley (2010)*

- Equal probability sampling (equal weights) results in lower variance than unequal weight sampling (all other things being equal)

- Higher variability in the weights in a population results in higher variance of sample estimators

Weighted Sampling

Figure 6

- This plot was created the same way as the other, that is, with 1000 repeated samplings of each sample size from a population of twice the sample size (50% sample fraction).

- Once again, I fix the number of individuals with disease as 20% of the population.

- What makes this weighted sampling is that each individual in the population has a sampling weight that influences the probability of selection into the sample for that individual

Weighted Sampling

Figures 7a through 7f

- The first step in creating the weights was to start with a distribution that was similar to the degree distributions found in some RDS samples.

- These degrees were selected from a Conway-Maxwell Poisson distribution, a distribution that produces counts, but has a separate parameter for the standard deviation, unlike the Poisson.

  That is, it can be under-dispersed or over-dispersed relative to a Poisson distribution

- Then the sampling was done with probabilities to proportional to degree.

- The sampling without replacement of half of the population was done with individual sampling probabilities proportional to that individual's degree.

- The weights are the inverse of the sampling probability normalized so they sum to the population total.

Weighted Sampling

Figure 6

- The first thing to notice is that the medians do not all lie on the population value of 0.2 and that they are not identical to the means (the dashes).

- Asymptotically, that is, with enough repeated samplings, the medians and means will be equal to the population value. But apparently 1000 is not enough, unlike in the simple random sampling case.

Weighted Sampling

Figure 8

- This is another view of the same weighted sampling process for the sample size of 500 and the population size of 1000.
- Each line shows the 95% confidence interval of the estimate (dot) of the population proportion of diseased individuals for one sample.
- There are 100 such samples shown on this plot. How many confidence intervals would you expect to not contain the true value of 0.2?
- The interesting thing about weighted sampling is that the same point estimate does not necessarily produce the same confidence interval as it would in a simple (unweighted) sample.
- Because of the weights, there are many different samples with different sets of weights and outcomes that can result in the same point estimate of disease proportion in the population.
- The more variability in these sets of weights that produce this sample mean, the larger the confidence interval.

Cluster Sampling

- In both the simple and weighted sampling described thus far, sampled units (individuals) are statistically independent.
- In cluster sampling, dependence between observations is introduced. This effectively reduces the sample size, that is, reduces the power to detect effects, that is, increases the variance of the estimators.
- In this example, there are 20 clusters of size 100 each, for a total population size of 2000.
- In each cluster, there is a different proportion of the disease ranging from 0.1 to 0.3 with the proportion in the total population being 0.2.
- 10 of the 20 clusters are sampled without replacement with equal probability.
- 50 of the 100 individuals from each sampled cluster are sampled without replacement with equal probability.
- The overall sample fraction is 25%.

Cluster Sampling

Figure 9a

- This shows the variability in the point estimates and confidence intervals in 100 different cluster samples

- In this case, the proportion with disease in each cluster varied from .10 to .30 in increments of .01 (except .20, in order to make 20 clusters with an overall population proportion of .20)

- Note the variability in the size of the confidence intervals even when the point estimates are similar.

Cluster Sampling

Figure 9b

- Same conditions as Figure 9a except for the distribution of the disease proportion across clusters
- In this case, half (10) of the clusters had a disease proportion of .10 and the other half had a disease proportion of .20.
- Note that the confidence intervals in Figure 9a are smaller than those in Figure 9b
- This is a reflection of the difference in the variability of the proportion of disease between the clusters.
- The standard deviation of the cluster disease proportions is .064 for Figure 9a and .103 for Figure 9b

Sampling Methods Comparison

Figure 10

- Simple (equal weights) random sampling from a fixed population does best
- Unequal weights adds to the variability in the estimate
- Cluster sampling with equal weights may do better or worse than weighted sampling depending on the variability in the weights and the between cluster variability
- Clusters sampling with unequal weights would have higher variability (not shown)

Summary

All other things being equal, the following increase the uncertainty in the design-based estimation of population quantities.

- Small sample sizes
- Population proportions approaching 0.5
- Small sample fractions
- High variability in sampling weights
- Clustering (dependence)
- High variability between cluster populations (in the quantities of interest)

Samples from RDS

- Small sample sizes

  Yes, especially for subpopulations

- Population proportions approaching 0.5

  Sometimes "hidden populations" are studied because of their relatively high proportion of the disease being studied

- Small sample fractions?

  Sometimes large sample fractions, but population size is typically unknown so that estimating N adds to the uncertainty

- High variability in sampling weights

  Sometimes high variability in the degrees (network size) of respondents

- Clustering (dependence)

  Dependence between egos and alters? respondents from the same seed?

Samples from RDS

- High variability between cluster populations (in the quantities of interest)

  Might this be a result of trying to start with diverse seeds?

Recommended Books

Complex Surveys: A Guide to Analysis Using R by Thomas S. Lumley, Wiley Series in Survey Methodology, 2010.

Sampling (2nd edition) by Steven K. Thompson, Wiley Series in Probability and Statistics, 2002.