

ST 432 Notes on Weighting, Imputation and Variance Calculations

April 2009

Weighting Methods

Basic Form of Population Total Estimator

The Horvitz Thompson estimator of the population total can be represented by

$$\hat{t}_{\pi} = \sum \left(\frac{y_i}{\pi_i} \right),$$

where π_i is the inclusion probability and is a known quantity by the design of the survey..

Survey researchers often rewrite this as

$$\hat{t}_{HT} = \sum w_i y_i$$

where $w_i = 1/\pi_i$

Generalisation to Allow for Non-Response

$$\hat{t}_{HT} = \sum w_i^* y_i,$$

where now

$$w_i^* = 1/\pi_i \hat{\phi}_i, \text{ and}$$

where $\hat{\phi}_i$ is an estimate of the probability of getting a response for unit i given that they are in the sample. The main issue is how to obtain a reasonable estimate of this probability. We discussed a couple of methods for doing that in class.

Note- Further weighting adjustments such as for incomplete frames are possible but we wont discuss in detail here. Now

$$w_i^{**} = 1/\pi_i \hat{\phi}_i \hat{\gamma}_i \text{ where } \hat{\gamma}_i \text{ is an estimate of the probability of being on incomplete frame.}$$

Note- Because units are often unequally weighted there are complex analysis issues. Standard software like regression packages usually assume equal weighting and thus estimates obtained can suffer severe biases! An exception is the special SUDAAN software written specifically for complex survey use by RTI. Gordon Brown discussed this topic.

ST 432 Imputation Methods Summary

Reference: Section 8.6 in Lohr (1999)

So far we have considered design based methods of dealing with nonresponse. We then also considered how weights might be used to adjust for nonresponse. Now we consider another method of dealing with nonresponse at the analysis stage.

Imputation methods are methods used to fill in missing values remaining after a survey is completed. Types of Imputation methods used are:

- Deductive Imputation
- Cell Mean Imputation
- Hot Deck Imputation
- Regression Imputation.
- Cold Deck Imputation

Here a brief summary of each method will be given.

Deductive Imputation- this uses logical relationships between the variables measured. (Very simple example –a respondent states they are not a crime victim but later leaves blank if they are not a violent crime victim –assign them a no! Limited usefulness)

Cell Mean Imputation-here respondents are divided into classes. Then the cell mean for a particular class is used for all missing values in that class. Sometimes a stochastic error term is added. This method obviously has the assumption that the values are missing at random.

Hot Deck Imputation-this method extends from cell mean imputation. For a particular class some other value than the mean is chosen to be the imputed value. One variation is sequential hot deck imputation which uses the value from nearest previous respondent that had data on this variable. Random hot deck imputation uses a cell class member chosen at random. Nearest Neighbor hot deck imputation uses the value for the respondent with the complete data “most like” the respondent with the incomplete data in the survey.

Regression Imputation-this method uses a regression model to predict the value to be imputed.

Cold Deck Imputation-this uses information from a previous survey or historical information.

Here is a very small data set useful for showing implementation of different imputation methods.

TABLE 8.3
Small Data Set Used to Illustrate Imputation Methods

Person	Age	Sex	Years of Education	Crime Victim?	Violent-Crime Victim?
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

Summary Points on Imputation Methods

Why Use Imputation?

- To reduce item non response bias.
- To provide a data set that is “complete” and hence easier to analyse with some packages.

Dangers of Imputation

- There are very strong assumptions made because it is a model based approach. For example, in the regression model, the assumption is that only age affects non response and in a linear manner.
- Variances calculated ignoring that the imputed values were really missing are biased low. The degree of negative bias depends on how many values had to be imputed.
- Correct variance calculations can be very complex. Some kind of parametric bootstrap procedure (see next section) would have to be used.

ST 432 Methods of Calculating Variances

Class Notes

I am reminded of Gordon Brown's comment in his lecture which was that correct variance calculations are what keep statisticians in business. Some methods are:

- Standard Approaches
- Taylor Series Approach
- Replicated Sampling Approach
- Resampling Approaches
 - Jackknife
 - Bootstrap
- Simulation or Parametric Bootstrap Approach.

1. Standard Approaches

We used this approach for many exact variance calculations early in the semester. The problem is that for many nonlinear functions these kind of calculations are impossible

2. Taylor Series Approach

Now we move to somewhat more complex situations where calculation of exact variances may not be possible for a function of random variable(s) (log(x), ratio(x/y) etc). Consider the simplest possible situation where we want to estimate γ by $\hat{\gamma}$ a consistent estimator with finite variance and $\hat{\gamma} = f(x)$, with x , a random variable and $E(x) = \theta$.

Let us note the approximation based on a Taylor series expansion that

$$f(x) \approx f(\theta) + f'(\theta)(x - \theta) + f''(\theta)(x - \theta)^2 / 2! + \dots$$

Note that the derivatives are evaluated at θ and it follows that

$$E[\hat{\gamma}] = E[f(x)] \approx f(\theta)$$

with a better approximation ,

$$E[\hat{\gamma}] = E[f(x)] \approx f(\theta) + f''(\theta)Var(x) / 2!$$

and

$$\text{Var}[\hat{\gamma}] = \text{Var}[f(x)] \approx [f'(\theta)]^2 \text{Var}(x).$$

These large sample approximate results can be generalized to the situation where x is a vector valued random variable. Also the Taylor series approach may be used to obtain approximate biases as well. Some simple examples will be given in class.

Gordon Brown in his lecture earlier in the semester mentioned that the special RTI software package called SUDAAN uses variances based on this general approach.

3. Replicated Sampling Approach

In some complex sampling situations it may be possible to take replicate samples to solve the problem of an intractable or impossible variance calculation. (Recall that I used this approach for systematic random sampling earlier in the semester. I suggested one take several random starts rather than just one.)

For each of R replicates calculate $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_R$ and therefore

$$\hat{\theta} = \frac{\sum_{i=1}^R \hat{\theta}_i}{R} \text{ and}$$

$$\text{Var}(\hat{\theta}) = \frac{\sum_{i=1}^R (\hat{\theta}_i - \hat{\theta})^2}{R(R-1)}$$

$$SE(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

We could also get an approximate 95% confidence interval as

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta}).$$

4. Resampling Approaches

These are computer intensive approaches that involve resampling on the computer to generate a large number of “replicates”.

Jackknife-this older approach which goes back to Tukey involves dropping out units one at a time but this approach won't be described here.

Bootstrap –this now very widely used approach in its simplest form involves resampling from the actual sample with replacement on the computer. One obtains a set of R

“replicates”. For each replicate one estimates $\hat{\theta}_i$ and from $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_R$ can then use the equations given in Section 3. A good reference on the bootstrap is Efron and Tibshirani (1993). See also Monahan (2001).

One very important issue is that because R is large then one can estimate the complete sampling distribution and use more robust percentile based confidence intervals that do not assume normality. Suppose R was 1000 then one could order the values of the estimates and just take the 25 th and 975 th values as the percentiles.

5. Simulation or Parametric Bootstrap Approach

Here one estimates parameters of interest using a model. Then one again obtains “replicates” samples on the computer using a monte carlo simulation method. When running the simulation one assumes that the parameter estimate is the true parameter. For each replicate one estimates $\hat{\theta}_i$ using the model and from $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_R$ can then use the equations given in Section 3. Again more robust percentile based confidence intervals are often used. Monte Carlo methods are described in many statistical computing books such as Monahan (2001).

References

- Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.
- Monahan, J. F. (2001). Numerical Methods of Statistics, Cambridge University Press.