Urvalsmetoder och Estimation 4

Sampling and Estimation 4

2011-02-07

5. Estimation and sampling with general inclusion probabilities

5.1 Introduction

- Inclusion probabilities are central in (design-based) inference.
- First order inclusion probabilities:

$$\pi_i = P(i \in S) = P(I_i = 1)$$

- Second order inclusion probabilities: .

$$\pi_{ij} = P(i,j \in S) = P(I_{i,j} = 1)$$

in particular: $\pi_{ii} = \pi_i$

- Third, fourth, … order are defined accordingly but less used

- Both second order inclusion probabilities and covariances describe the relation between inclusion indicators

$$\pi_{ij} = E(I_i*I_j) = Cov(I_i,I_j) + \pi_i*\pi_j$$

- But second order inclusion probabilities are used more in the theory of designbased survey-sampling.

- If two objects are likely to be similar the probability to get both in the sample should be small
- e.g. stratified sample with only two in each stratum

$\pi_{ij} = 4/(N_h N_g)$ if in different strata i.e. Corr = 0

$\pi_{ij} = 2/(N_h*(N_h-1))$ if in the same i.e. Corr = $-1/(N_h-1)$

Example of the computation

- Two-stage cluster sampling with unequal cluster sizes (not a simple formula in standard text-books):

- Consider N blocks (or villages) with N_k households in each (N_k unknown in advance).
- n blocks are chosen with inclusion probabilities ρ_k. The households in them are listed and counted, N_k for k = 1, …n (and second order inclusion probabilities ρ_{kl}).
- Select n_k households by SRS from the selected n blocks. n_k may depend on N_k but not on N_i for i # k

- Inclusion probabilities:

$$\pi_{i(k)} = \rho_k \, n_k/N_k$$

$$\pi_{i(k)j(k)} = 2 \, \rho_k \, n_k(n_k -1)/(N_k \, (N_k-1))$$

$$\pi_{i(k)j(l)} = \rho_{kl} \, n_k n_l/(N_k N_l) \qquad k \# l$$

πps sampling

- Sampling with varying inclusion probabilities is usually called πps sampling (inclusion probabilities proportional to size. Nowadays it maybe proportional to anything, not necessarily size).

- Internationally πps sampling is often done in connection with multistage sampling (as we saw above). (Not quite that often in Sweden with our good frames).

- Common esamples. Farms and environmental statistics proportional to area of the unit. Enterprises proportional to numer of employees or turnover (last year)

5.2 Estimation with general inclusion probabilities

- Horvitz-Thompson (HT) estimator:

$$t_{y,HT} = \Sigma_S\, y_i/\pi_i = \Sigma_U\, I_i y_i/\pi_i$$

- Often written: $t_{y,HT} = \Sigma_S\, \omega_i y_i$, where $\omega_i=1/\pi_i$ are called design-weights
- Unbiased (Why?)
- Variance: $E(t_{y,HT}^2) – E^2(t_{y,HT}) =$

$$E(\Sigma\Sigma_{UU}\, I_i I_j y_i y_j/(\pi_i\pi_j)) - (\Sigma_U\, y_i)^2 =$$
$$\Sigma\Sigma_{UU}\, \pi_{ij} y_i y_j/(\pi_i\pi_j)) - \Sigma\Sigma_{UU}\, y_i\, y_j =$$
$$\Sigma\Sigma_{UU}\, ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j))\, y_i\, y_j$$

- This method always works for probability samples (if $\pi_i > 0$)!

But both second order inclusion probabilities and the double sum for large n may be difficult/complicated to compute

- Note: The second order inclusion probabilities decide the variance
- When designing a sampling procedure you should worry about them
- People often talk about πps-sampling without worrying about the second order inclusion probabilities. To them they are just a nuisance when estimating the variance, not a tool to obtain efficient samples

Variance estimation

- The variance is: $\Sigma\Sigma_{UU} ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j)) y_i y_j$

- Using that $E(I_iI_j)= \pi_{ij}$ we easily see that

  $$E[\Sigma\Sigma_{SS} ((\pi_{ij} - \pi_i\pi_j)/(\pi_{ij} \pi_i\pi_j)) y_i y_j]$$

  is an unbiased variance estimator

$$= E[\Sigma\Sigma_{UU} ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j)) (I_i I_j /\pi_{ij}) y_i y_j] =$$

$$\Sigma\Sigma_{UU} ((\pi_{ij} - \pi_i\pi_j)/(\pi_i\pi_j)) E[(I_i I_j /\pi_{ij})] y_i y_j$$

- Another expression is the Sen-Yates-Grundy (SYG) estimator:

  $$½ \Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2$$

  which is unbiased for fixed sample sizes

$\frac{1}{2} \sum\sum_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2$

- This Sen-Yates-Grundy formula is good to look at for advice on how to choose the inclusion probabilities
- The first and second order inclusion probabilities decide the variance
- Suppose that you have a constant sample size (often good for efficiency reasons but also for planning reasons)
  - First, try to chose the π_i proportional to y_i. (Makes the second bracket small)
  - Second, try to get π_{ij} close to $\pi_i\pi_j$ (i.e. independent) when y_i/π_i differs much from y_j/π_j (Makes first bracket small)
- Stratified sampling e.g. Put similar units in the same strata and different units in different strata which are sampled independently

Sketch of proof of the unbiasedness of the Sen-Yates-Grundy (SYG) estimator

Expand the square

$$½ \Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (y_i/\pi_i - y_j/\pi_j)^2 =$$

$$½ \Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) [(y_i/\pi_i)^2 + (y_j/\pi_j)^2 - 2(y_jy_j/\pi_j\pi_j)]$$

The expected values of the two terms with the square $((y_i/\pi_i)^2)$ are both 0 since (if the sums include i=j)

. $E(\Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (y_i/\pi_i)^2) = E(\Sigma\Sigma_{UU} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) I_i I_j (y_i/\pi_i)^2) =$

$\Sigma\Sigma_{UU} (\pi_i\pi_j - \pi_{ij}) E(I_i I_j/ \pi_{ij}) (y_i/\pi_i)^2) = \Sigma\Sigma_{UU} ((\pi_i\pi_j - \pi_{ij})(y_i/\pi_i)^2 =$

$\Sigma_U ((\pi_i \Sigma_{j€U} \pi_j – \Sigma_{j€U} \pi_{ij})) (y_i/\pi_i)^2) = \Sigma_U n(\pi_i - \pi_i) (y_i/\pi_i)^2 = 0$

Here we used that the sample size is fixed, which implies that $\Sigma_j\pi_{ij} = n\pi_i$ and $\Sigma_{j€U} \pi_j = n$.

Thus only the cross product remains giving the result

$$\Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (y_jy_j/\pi_j\pi_j)$$

- These formulae (HT and SYG) give design-unbiased estimates for mean and variance, respectively, for any probability sample (SYG requires fixed sample size)

- This requires
  - strictly positive second order inclusion probabilities
  - computable inclusion probabilities for all units in the sample.

- Not always simple to compute the second order inclusion probabilities.

- The SYG-estimator is not always the best estimator/variance estimator but mostly fairly good or at least acceptable for all cases with fixed sample sizes

Optimal sampling

- Model for the data: $E(y_i \mid x_i) = \mu(x_i)$; $Var(y_i \mid x_i) = \sigma^2(x_i)$ and independence.

- Then if you use an asymptotically optimal estimator you should draw samples with the inclusion probability π_i proportional to $\sigma(x_i)$
- The second order inclusion probabilities are unimportant (if the sampling mechanism is mixing)
- The last conclusion is a consequence of the independence assumption

Optimal sampling

- Very often the variance depends on the unit size.
  - If Var(Y/X) is constant then Var(Y) = x^2 (This situation is very common in economic statistics. E.g. the relative change in, turnover say, does not depend on the size)
  - if Y€Po(X=x) then Var(Y)=x
- In the first case it is optimal to use precisely proportional to <u>size</u>

Pseudo-likelihood-function

- Likelihood estimation is model-based but what about sampling with varying inclusion probabilities (e.g. a medical study, where some patients are more likely to be included than others)
- Ordinary likelihood: $L(\Theta,y) = \Pi_{i€S}\ f(y_i|\Theta)$
  or ordinary loglikelihood: $l(\Theta,y) = \Sigma_{i€S}\ \mathit{ln}(f(y_i|\Theta))$
  Maximising the likelihoodfunction gives a good estimate.
- Now use a form of pseudo-likelihood, instead, where one weights with the inverse inclusion probabilities

  $L(\Theta,y) = \Pi_{i€S}\ f(y_i|\Theta)^{1/\pi i}$

  or for the logarithm

  $l(\Theta,y) = \Sigma_{i€S}\ (1/\pi_i)\ \mathit{ln}\ f(y_i|\Theta)$

  Maximising this pseudo-likelihoodfunction gives a good estimate
- But the variance can no longer be estimated using the Fisher information (1/ (second derivative of the likelihood function))
- Of course, the full likelihood can be used. But often much more difficult and extremely model-sensitive. This approach is mostly much more robust.

5.3 Sampling with varying probabilities

5.3.1 Intentional

There are two main lines on how to draw samples with given inclusion probabilities

- Mimick simple random sampling. Introduce as little extra structure as possible i.e. getting π_{ij} close to $\pi_i * \pi_j$. If the procedure is used with equal inclusion probabilities SRS should be obtained. (I.e. get the highest possible entropy).
- Use all the information you have. Model second order inclusion probabilities. Even the inclusion probabilities may contain <u>extra</u> information.

Mimicking SRS
How to draw πps? Problems!

- Not easy if the sample fraction is not negligible!
- E.g. draw 2 units among 5 with probabilities 0.7, 0.3, 0.5, 0.2 and 0.3
- First try. Draw one with half this probability and second with probability proportional to the remaining ones.

  E.g. $\pi_1 =$ 0.35+2*0.15*0.7/1.7+0.25*0.7/1.5+0.1*0.7/1.8 = 0.629

(small π will be too large and vice versa)

Mimicking SRS

- E.g. draw 2 units among 5 with probabilities 0.7, 0.3, 0.5, 0.2 and 0.3
- Second try. Conditional Poisson (also called 3P-sampling). Go through the units one by one and include the units independently with correct probabilities. If the sample contains less or more than two units reject it and try again.

E.g. π_1 = P(1€S | 2 units in sample) = 0.7 * (2*0.3*0.5*0.8*0.7 + 0.7*0.5*0.8*0.7 + 0.7*0.5*0.2*0.7) / (0.7*(2*0.3*0.5*0.8*0.7 + 0.7*0.5*0.8*0.7 + 0.7*0.5*0.2*0.7) + 2*0.3*0.3*(0.5*0.8*0.7 + 0.5*0.2*0.7 + 0.5*0.8*0.3) + 0.3*0.7*0.5*0.2 *0.7) = 0.744

(Small π_i will become too small). May take a long time until a sample of correct size is obtained

Mimicking SRS

Third try.

- Order sampling (distributional Poisson)
  - Draw a uniform random number Z_i for each unit.
  - Transform it in some way $g(Z_i, \pi_i)$ so that $P(g(Z_i, \pi_i) > 1) = \pi_i$.
  - Pick the n units with the largest transformed numbers.

  Not exactly correct, but almost, for a suitably chosen g.

- Illustrative example (for the sake of illustration)
  - Let $g(Z_i, \pi_i) = Z_i /(1-\pi_i)$
  - Then $P(g(Z_i, \pi_i) >1) =P(Z_i /(1-\pi_i) >1) = P(Z_i >1-\pi_i) = \pi_i$

Mimicking SRS

Third try.

- Order sampling (distributional Poisson). This choice of $g(z_i ,\pi_i)$ above was not particularly good. The best choice in this category is Pareto Poisson, which is used at Statistics Sweden.

- Pareto Poisson where $g(z_i ,\pi_i) = \pi_i (1-z_i)/(z_i (1-\pi_i))$ (Named since $g(z_i ,\pi_i)$ then follows a Pareto distribution)
  - Example: A sample n=2 with probabilities 0,7, 0,3 0,5, 0,2, 0,3
  - Take five random numbers: 0,175, 0,832, 0,746, 0,312, 0,098
  - Compute g(.): 0,50 (0,175*0,7/0,825*0,3), 2,12, 2,96, 0,14, 0,05
  - Select the two largest: In this case number 3 and 4

Sampford (often used internationally) E.g. draw 2 units among 5 with probabilities 0.7, 0.3, 0.5, 0.2 and 0.3

1. Take one unit with probability proportional to π_i (i.e with probability $\pi_i/\Sigma\pi_j$)

0,35, 0,15, 0,25, 0,1, 0,15

Say that we get nr 3

2. Take n-1 units <u>with replacement</u> with probabilities proportional to $\pi_i/(1-\pi_i)$ (i.e. with probability $\pi_i/(1-\pi_i)$ / Σ $(\pi_j/(1-\pi_j))$

Proportional to 2,33, 0,43, 1, 0,25, 0,43

i.e. 0,52, 0,10, 0,22, 0,06, 0,10

Say that we get nr 1

3. If the sample now contains exactly n different units, this is the sample.

Sample contains 3 and 1, that is two different

4. Otherwise restart from 1.

- It is not easy to show that this method will give the correct inclusion probabilities

Mimicking SRS

- Many other methods have been suggested e.g. by Hajek, Sunter. All rather complicated. Systematic πps-sampling with random order (see below) is simpler, but gives lower entropy.

- Some persons have the goal to get a high ”entropy”, i.e. put in maximum amount of randomness given the inclusion probabilities and sample size. (Sampford gives maximum entropy). (Pareto Poisson is very close)

Using more structure (not mimicking SRS)

- Systematic sampling was described earlier (and next page). Works and is good if an even spread is desired but not if one wants the probabilities to include close units to be substantial (or near $\pi_i*\pi_j$). (If the elements are ordered a good spread in that dimension is obtained. If the order is random this procedure once again mimicks SRS)

- Sometimes one may want to use the background knowledge more. In those cases where a good spread is desired, methods close to systematic πps (but where variance is possible to estimate) or a very fine stratification are better alternatives.

Systematic Πps-sampling

- Order the units in some sensible way using the information you want to employ (e.g. geografical, an important auxiliary variable or the probabilities themselves)

- Compute the cumulative inclusion probabilities $\Pi_k = \Sigma_{i<k+1}\ \pi_i$
- Choose an arbitary startingpoint, U, uniformly in (0, Π_N /n)
- Choose all units on the distance k*Π_N /n from it (i.e. All units, where the cumulative sum $\Pi_{i-1} < U + k\Pi_N /n < \Pi_i$) for some 0<k<n.

- This procedure gives a good spread over the variable you have ordered after and thus the variance is almost always less than for other πps-metoder (if there is no periods)

Systematic sampling

- The variance under systematic sampling can not be estimated without a bias
- But the variance you should have got with ordinary πps-methods can be estimated (even better than with πps design)
- Since this is known to be an overestimate, all uncertainty intervals will be conservative. (i.e. the coverage probabilities will be higher than the nominal confidence level).
- The variance may be estimated if you take e.g. M independent systematic samples i.e. M starting points < $Mk\Pi_N/n$ ond then units with the distance $Mk\Pi_N/n$.

Bulldozer method

- The units are ordered in some way (Real time sampling)
- Take the first with probability π_1
- Update the following inclusion probabilities, in a suitable way (the expected value of the updated value must equal the original)
  - Simplest example: π_2 <- $\pi_2 - (1-\pi_1)I_1 + \pi_1(1-I_1)$

  if the first unit is taken and if this expression is between 0 and 1.

  if it is not also π_3 (and π_4? …) must be updated.
- Repeat successively for units 2 (updating 3), 3 (updating 4) a.s.o
- It is easily seen that the sum of the remaining inclusion probabilities and the observed inclusion indicators is always exactly n, meaning a fixed sample of size n.
- Using a good updating procedure can yield almost the desired second order inclusion probabilities

Bulldozer method - Example

- Suppose the units pass the interviewer entering an amusement park. (An example of real time sampling)
- Take the first with probability π_1 (=0,7, suppose he is taken)
- Update the following inclusion probabilities, in a suitable way
  - New probabilities 1, 0, 0,5, 0,2, 0,3
- Repeat successively
  - unit 2 is not aken p=0
  - Take unit three with probability 0,5. Suppose not taken, update probabilities 1, 0, 0, 0,4, 0,6

    (if unit three was taken updated prob would be 1, 0, 1, 0, 0 i.e. the full sample had been taken)
  - Take unit four with probability 0,4. Suppose taken, update probabilities 1, 0, 0, 1, 0.

    i.e. full sample is taken
- It is easily seen that the sum of the remaining inclusion probabilities and the observed inclusion indicators is always exactly n, meaning a fixed sample of size n.
- Using a good updating procedure can yield almost the desired second order inclusion probabilities. Getting an even spread over the day and seldom close customers

- The Bulldozer method is a special case of ”the splitting method” (Tillé and Deville).
- At each stage you make a random decision and then you change all inclusion probabilities. (keeping in mind that the expected unconditional inclusion probabilities are fixed.
- It is highly fashionable among sampling theorists nowadays

Systematic πps-sampling

- Order the units in some sensible way using the information you want to employ (e.g. geografical, an important auxiliary variable or the probabilities themselves)

- Compute the cumulative inclusion probabilities $\Pi_k = \Sigma_{i<k+1}\ \pi_i$
- Choose an arbitary startingpoint, U, uniformly in (0, Π_N /n)
- Choose all units on the distance k*Π_N /n from it (i.e. All units, where the cumulative sum $\Pi_{i-1} < U + k\Pi_N /n < \Pi_i$) for some 0<k<n.

- This procedure gives a good spread over the variable you have ordered after and thus the variance is almost always less than for other πps-metoder (if there are no periods)

Systematic sampling

- The variance under systematic sampling can not be estimated without a bias (within design-bases theory)
- But the variance you should have got with ordinary πps-methods can be estimated (even better than with πps design)
- Under mild restrictions this is known to be an overestimate. Thus all intervals will be conservative. (The coverage probability is higher than the nominal confidence level).
- The variance may be estimated if you take e.g. M independent systematic samples i.e. M starting points $< Mk\Pi_N/n$ ond then units with the distance $Mk\Pi_N/n$.

Real time sampling

- The units arrive in a steady stream. Each time a unit arrives you must decide whether it should be in the sample before you know the others
  - Magnetic datatapes
  - Tourist statistics
  - Sampling trees in a forest
  - Visitors to a national park
  - Customers leaving a shop
  - …

Some comments

- The second order inclusions are important
  - Trees very close to each other are probably quite similar due to the soil. It may be silly to take to many trees close to each other
  - Or farms in the same village
  - …
- The Intervjewer/sampler may have a tendency to select ”representative” persons/trees, which may affect the order they are counted in the list. Avoid giving some units the conditional inclusion probability 0.

Bulldozer method

- The units are ordered in some way (Real time sampling)
- Take the first with probability π_1
- Update the following inclusion probabilities, in a suitable way (the expected value of the updated value must equal the original)
  - Simplest example: $\pi_2 \leftarrow \pi_2 - (1-\pi_1)I_1 + \pi_1(1-I_1)$

  if the first unit is taken and if this expression is between 0 and 1.

  if it is not also π_3 (and π_4? …) must be updated.
- Repeat successively for units 2 (updating 3), 3 (updating 4) a.s.o
- It is easily seen that the sum of the remaining inclusion probabilities and the observed inclusion indicators is always exactly n, meaning a fixed sample of size n.
- Using a good updating procedure can yield almost the desired second order inclusion probabilities

Bulldozer method - Example

- Suppose the units pass the interviewer entering an amusement park. (An example of real time sampling. Take n=2 with probab 0,7, 0,3 0,5, 0,2, 0,3)
- Take the first with probability π_1 (=0,7, suppose he is taken)
- Update the following inclusion probabilities, in a suitable way
  - New probabilities 1, 0, 0,5, 0,2, 0,3
- Repeat successively
  - unit 2 is not aken p=0
  - Take unit three with probability 0,5. Suppose not taken, update probabilities 1, 0, 0, 0,4, 0,6

    (if unit three was taken updated prob would be 1, 0, 1, 0, 0 i.e. the full sample had been taken)
  - Take unit four with probability 0,4. Suppose taken, update probabilities 1, 0, 0, 1, 0.

    i.e. full sample is taken
- It is easily seen that the sum of the remaining inclusion probabilities and the observed inclusion indicators is always exactly n, meaning a fixed sample of size n.
- Using a good updating procedure can yield almost the desired second order inclusion probabilities. Getting an even spread over the day and seldom close customers

Example

- To show that this is a flexible approach:
- Suppose you want the correlations $\rho_1 = \rho_2 = \rho_3$; $\Sigma_1{}^3\rho_1 = \rho_{sum} = -0{,}5$; $\rho_\kappa = 0$ for $k > 3$; (The summing condition gives a fixed sample size but is possible to obtain only if $\pi \geq 1/7$)
- We tried with $\omega_1 = 0.446$, $\omega_2 = 0{,}327$ and $\omega_3 = 0{,}226$, ($\omega_k = 0$ for $k>3$) in the updating equations $\pi_{k+l}{}^{k} = \pi_{k+l}{}^{k-1} – \omega_l{}^{k}(I_k – \pi_{ki}{}^{k-1})$

π	ρ_1	ρ_2	ρ_3	ρ_{sum}
0,5	-0,159	-0,178	-0,162	-0,499
0,45	-0,153	-0,170	-0,173	-0,496
0,4	-0,160	-0,173	-0,166	-0,499
0,35	-0,151	-0,171	-0,174	-0,499
1/3	-0,151	-0,180	-0,166	-0,497
0,3	-0,149	-0,170	-0,176	-0,495
1/4	-0,144	-0,169	-0,173	-0,486
0,20	-0,139	-0,152	-0,165	-0,486
0,15	-0,139	-0,152	-0,165	-0,456
0,10	-0,111	-0,111	-0,111	-0,333
0,05	-0,0525	-0,0525	-0,5250	-0,1575

Note that in the two last rows the sum is the highest possible.

ρ_k is exacly equal to 0 for k > 3.

- The Bulldozer method is a special case of ”the splitting method” (Cube method) (Tillé and Deville).

- At each stage you make a random decision and then you change all inclusion probabilities. (keeping in mind that the expected unconditional inclusion probabilities are fixed.

- It is highly fashionable among sampling theorists nowadays

5.3.2 Unintentionally varying probabilities

Occurs often naturally e.g.

- Frame of households but you want to sample individuals (e.g. Random digit dialling)
- or vica versa
- Calling people at evenings (ask about how many evenings they were home last week)
- Selecting all patients at a special date in a hospital (People with long convalescence periods have larger probabilities of being selected).
- Selecting customers/suppliers from a list of invoices sent/received during the year.
- Selecting customers of a shop (exit interviews)
- Visitors to an angling park/nature resort

What to do about unintentional πps?

((Inclusion) probability proportional to size. Nowadays size can be any number)

- In many cases you can find a proxy for the inclusion probability e.g. number of evenings home, members of households etc. If you can, use it.

- If the proxy is such that the study-variable and the inclusion indicator are conditionally independent given the proxy, use the proxy as inclusion probability (e.g. poststratify or use weighted ML) and everything will be (almost) correct (easy if sampling fraction is small).

- If you can't find a good proxy, you have to resort to modelling (e.g. propensity scores) and if possible do some sort of follow-up study.

- If $\pi_i = 0$ for some units, i, you are in real trouble! What to do? (State in quality declaration, Use model-based sampling)

A more complicated example

A market researcher will stand outside a shop to interview a sample of customers.

She has selected two mondays, two tuesdays, …, two sundays randomly during a four week period. During each selected day she selects customers independently with a probability of 1/10. (It is obvious that the inclusion probabilities will depend on the number of visits). Determine the inclusion probabilities. (Assume for simplicity that no customer visits the shop more than once during the same day).

Ask about how often the customers have visited the shop last month or which day during a typical week

Solution for some special cases:

- First look at a man visiting the shop once. He will be selected if that day is selected (1/2) and if he is selected that day (1/10). Thus he will be selected with probability π_1 = 1/20.

A woman visiting twice during different weekdays. The probability of being selected the first time is 1/20 and similarly the second time. (The weekdays are independent). The inclusion probability is thus π_2 = 1/20 + 1/20 – 1/400 = 39/400.

A woman visiting twice during the same weekdays. The probability that she is selected the first (second) Monday is 1/20. The probability that she is selected twice is (2/4)*(1/3) times 1/100. Thus the probability is 1/20 + 1/20 – 1/600 = 59/600

General solution

- A person visiting 3 or 4 times during the same weekday will be selected on that day with probabilities
  - p_3 = P(both days with interviewer)*(2/10 – 1/100) + P(only one day with interviewer)*1/10 = 29/200
  - p_4 = 2/10-1/100 = 18/100.
- She will not be selected with probability $q_x = 1 – p_x$.
- A person visiting x_1 times on mondays, x_2 on tuesdays, a.s.o will not be selected with probability $q_{x1} * q_{x2} * \ldots * q_{x7}$. The first order inclusion probability is thus $\pi_{x1, x2,x3, x4, x5, x6, x7} = 1 – \Pi_i q_{xi}$
- Since different person are included independently of each other the second order inclusion probabilities are just the product of the first order probabilities.

5.4
Are sampling with varying probabilities and HT-estimation any good

?

Basu´s elefants

- A circus-owner arrives a railway station with his herd of five elephants. He must estimate the weight in order to pay for the freight. There is only one scale and each weighing costs 1 rupie, but the owner has only one coin. He decides to weigh Mumbo, the middle elphant and multiply by 5.
- The circus statistician protests. This is not a probability sample! Every elephant must have a positive probability.
- But he admits that it should be more sensible to weigh the middle elephant than the two extremes so they decide on a compromise: Weigh the biggest Colonel Hathi or the smallest Jumbo with probabilities 0.01 and the next biggest or smallest, Tvumbo and Dumbo with probability 0.04 each.
- After looking into a random number table Mumbo is chosen and the director is happy: What did I say? Mumbo is weighed and the result is 2 ton. The director estimates the total weight to 10 ton.
- The statistician, however, says No! No! That is not an unbiased estimate; use the HT-estimate 2/0.9=2.22 ton.

(What would the the estimate have been, if Colonel Hathi had been weighed (His weight is 5 ton?))

When does HT behaves badly?

1. If the sample size varies the HT-estimate may be severely affected (or if Σ_S 1/π_i is a bad estimator of the total population size).
   - One often uses a ratio estimate instead $N(\Sigma_S y_i/\pi_i)/(\Sigma_S 1/\pi_i)$.
   - As a rule of thumbs: Use HT-estimates only when the sample size is fixed or varies marginally.
   - Sample size may in this discussion be replaced by other size measures. E.g. the number of employees in a firm. Use ratio estimators also in this case (or regression or prediction estimators see below)
2. Model-based methods are usually preferable for small sample sizes (cf Basu’s elephants)

5.5 Other aspects

- Sampling with replacement is much easier than without.
- One may show that without replacement strategies are always better (= more efficient) than with replacement strategies. (But for small sampling fractions almost equivalent)
- Some books talk much about methods with replacement e.g. The cumulative - size method or Lahiri’s method (A rejective method) for cluster sampling).

Design - estimation formula

- With SRS we saw that much could be gained by ratio, regression or prediction, estimates

- For the designs stratification and fixed sample size πps, we suggested the Horvitz-Thompson estimator.

- But one may combine different types of estimators and designs in other ways, getting even better procedures.

Examples

A πps sample

1. Find a good regression/predictor, e.g.

$Y_i = f(X_i) + e_i$; E.g. $a + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$

Estimate the parameters/function e.g. by (suitably weighted) GLS or non-parametric kernel estimation

Estimate the total by $\Sigma_i f(X_i) + t_{eHT}$

2. Use a HT-weighted ratio estimator $T_X(\Sigma_S y_i/\pi_i)/(\Sigma_S x_i/\pi_i)$ (we saw this earlier with N as T_X.)

Remember

- A good sampling procedure involves the combination of a good sampling design and a suitable estimation formula.