Sampling: Design and Analysis

Hao Wang

The College of Statistics

April 8, 2020

1. Introduction

2. Basic Concepts used in Sampling

3. Simple Random Sampling

4. Stratified Sampling

Work of Jerzy Neyman, 1934

- Census and survey
- Application of the Representative Method in Statistics
- Two different aspects of the representative method: random sampling and purposive selection
- Report of the Commission appointed by the International Statistical Institute (1926): In the selection of that part of the material which is to be the object of direct investigation, one or the other of the following two principles can be adopted: in certain instances it will be possible to make use of a combination of both principles. The one principle is characterized by the fact that the units which are to be included in the sample are selected at random. This method is only applicable where the circumstances make it possible to give every single unit an equal chance of inclusion in the sample...

Work of Jerzy Neyman, 1934

- ... The other principle consists in the samples being made up by purposive selection of groups of units which it is presumed will give the sample the same characteristics as the whole. There will be especial reason for preferring this method, where the material differs with respect to composition from the kind of material which is the basis of the experience of games of chance, and where it is therefore difficult or even impossible to comply with the aforesaid condition for the application of selection at random. Each of these two methods has certain advantages and certain defects
- Choice of collective characters for constructing Confidence Intervals:
  1. They must follow a frequency distribution which already tabled or may be easily calculated
  2. The resulting confidence intervals should be as narrow as possible

Work of Jerzy Neyman, 1934

- $\hat{\theta} = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \lambda_{ij} x_{ij}$

  (a) $E[\hat{\theta}] = \theta$
  (b) Standard error of $\hat{\theta}$ is less than that of any other linear function, satisfying (a)
- biased (systematic error): $E[\hat{\theta}_1] = \theta + \Delta$
- Stratified sampling (Bowley): Before drawing the random sample from the population II, this is divided into several "strata", say $P_1, P_2, \cdots P_k \cdots$, and the sample Υ is composed of k partial samples, say $\Upsilon_1, \Upsilon_2, \cdots \Upsilon_k \cdots$, each being drawn (with replacement or not) from one or other of the strata
- SRS: special case of stratified sampling with the number of strata $k = 1$ or an individual be considered as a group the size of which is 1

Work of Jerzy Neyman, 1934

- multi-stage stratified sampling
- purposive selection consists
  (a) in dividing the population of districts into second order strata according to values of "control" and size
  (b) in selecting randomly from each stratum a definite number of districts
- generally representative sample, a representative method of sampling and a consistent method of estimation -
  Thus, if we are interested in a collective character X of a population and use methods of sampling and of estimation, allowing us to ascribe to every possible sample, Υ, a confidence interval $X_1(\Upsilon)$,$X_2(\Upsilon)$ such that the frequency of errors in the statement $X_1(\Upsilon) \leq X \leq X_2(\Upsilon)$ does not exceed the limit $1 - \epsilon$ prescribed in advance, whatever the unknown properties of the population

基本概念

- "stratify": comes from Latin words meaning "to make layers"
- 实施分层抽样的可能原因:
  - 总体规模与样本容量大，总体单元之间的差异较大
  - 掌握先验信息，可以将总体按一定指标划分为子总体(strata)
  - We want to be protected from the possibility of obtaining a really bad sample (100 from 1000 male and 1000 female)
  - We may want data of known precision for subgroups of the population (many more male than female)
  - A stratified sample may be more convenient to administer and may result in a lower cost for the survey (large and small firms, households in urban strata and in rural strata)
  - Stratified sampling often gives more precise (having lower variance) estimates for population means and totals. Stratification works for lowering the variance because the variance within each stratum is often lower than the variance in the whole population. Prior knowledge can be used to save money in the sampling procedure

分层抽样/例I

接例8：美国1982、1987、1992年农业普查的数据

```
## R Code
library(SDaA)
# population with size 3078
data(agpop)
# regions
table(agpop$region)
# sample with size 300
data(agstrat)
# Northeast / North Central / South / West
stratNE=subset(agstrat,agstrat$region=="NE")
stratNC=subset(agstrat,agstrat$region=="NC")
stratS=subset(agstrat,agstrat$region=="S")
stratW=subset(agstrat,agstrat$region=="W")
```

定义与符号I

- 层：把一个包含N个单位的总体分成H个子总体，其中总体的每个单元属于且仅属于一个子总体，这样的子总体称为层。$N = N_1 + N_2 + \cdots + N_H$
- 分层抽样：在每一层中独立进行抽样，总的样本由各层样本组成，总体参数则根据各层样本参数的汇总做出估计，这种抽样称为分层抽样，所得样本称为分层样本$n = n_1 + n_2 + \cdots + n_H$
- 分层随机抽样：如果每层中的抽样都是独立地按照简单随机抽样进行的，那么这样的分层抽样称为分层随机抽样，所得的样本称为分层随机样本
- y_{hj}:第h层中第j个样本单元取值
- $t_h = \sum_{j=1}^{N_h} y_{hj}$，$t = \sum_{h=1}^{H} t_h$，$\bar{y}_{hU} = \sum_{j=1}^{N_h} y_{hj}/N_h$
- $\bar{y}_U = \frac{t}{N} = \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}/N$，$S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{y}_{hU})^2}{N_h - 1}$

定义与符号II

- $\bar{y}_h = \frac{1}{n_h}\sum_{j\in\Upsilon_h} y_{hj}$, $\hat{t}_h = \frac{N_h}{n_h}\sum_{j\in\Upsilon_h} y_{hj} = N_h\bar{y}_h$, $s_h^2 = \sum_{j\in\Upsilon_h}\frac{(y_{hj}-\bar{y}_h)^2}{n_h-1}$
- $\hat{t}_{str} = \sum_{h=1}^H \hat{t}_h = \sum_{h=1}^H N_h\bar{y}_h$, $\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^N \frac{N_h}{N}\bar{y}_h$
- 无偏性: $E[\sum_{h=1}^H \frac{N_h}{N}\bar{y}_h] = \sum_{h=1}^H \frac{N_h}{N}E(\bar{y}_h) = \sum_{h=1}^H \frac{N_h}{N}\bar{y}_{hU} = \bar{y}_U$
- 估计量方差:

  $$V(\hat{t}_{str}) = \sum_{h=1}^H V(\hat{t}_h) = \sum_{h=1}^H (1-\frac{n_h}{N_h})N_h^2\frac{S_h^2}{n_h}$$

  $$\hat{V}(\hat{t}_{str}) = \sum_{h=1}^H (1-\frac{n_h}{N_h})N_h^2\frac{s_h^2}{n_h}$$

  $$\hat{V}(\bar{y}_{str}) = \frac{1}{N^2}\hat{V}(\hat{t}_{str}) = \sum_{h=1}^H (1-\frac{n_h}{N_h})(\frac{N_h}{N})^2\frac{s_h^2}{n_h}$$
- 置信区间:
  在每层样本容量足够大，或者有大量的抽样层的条件下，
  $\bar{y}_U$ 的近似$100(1-\alpha)\%$的置信区间为$\bar{y}_{str} \pm z_{\alpha/2}SE(\bar{y}_{str})$

分层抽样/例II

接例8:

- 相关结果:

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2	$\hat{t}=N_h\bar{y}_h$	$(1-\frac{n_h}{N_h})N_h^2\frac{s_h^2}{n_h}$
Northeast	220	21	97629.8	7,647,472,708	21,478,558.2	1.59432×10^{13}
North Central	1054	103	300,504.2	29,618,183,543	316,731,379.4	2.88232×10^{14}
South	1382	135	211,350.0	53,587,487.856	292,037,390.8	6.84076×10^{14}
West	422	41	662,295.5	396,185,956,266	279,488,706.1	1.55365×10^{15}
total	3078	300			909,736,034.4	2.5419×10^{15}
$\sqrt{total}$						50,417,248

- 置信区间：t的95%近似置信区间为$\hat{t}_{str}\pm z_{\alpha/2}\hat{SE}(\hat{t}_{str})$，即$[1,008,553,820,\ \ 810,918,207]$
- 分层抽样和简单随机抽样方差比较

$$\frac{\textit{estimated variance from stratified sample with } (n=300)}{\textit{estimated variance from SRS with } (n=300)}=\frac{2.5419\times10^{15}}{3.3837\times10^{15}}=0.75$$

意味着在样本容量相同的情况下，分层抽样的精度更高

各层样本的分配/比例分配I

- Designing the survey is the most important part of using a survey in research
- 比例分配：在分层抽样中，若每层的样本量n_h都与层的大小N_h成比例，即入样概率$\frac{n_h}{N_h}=\frac{n}{N}$，则称样本量的这种分配方式为比例分配
- 自加权：若总体总量（或总体均值）的一个无偏估计量可以表示成样本基本单元的变量值总量（或均值）的一个常数倍，则称这种估计量为自加权或等加权
- 用SSW表示层内离差平方和
  $SSW=\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_{hU})^2=\sum_{h=1}^{H}(N_h-1)S_h^2$,

  用SSB表示层间离差平方和
  $SSB=\sum_{h=1}^{H}\sum_{j=1}^{N_h}(\bar{y}_{hU}-\bar{y}_U)^2=\sum_{h=1}^{H}N_h(\bar{y}_{hU}-\bar{y}_U)^2$,

  用SST表示总离差平方和
  $SST=\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_U)^2=(N-1)S^2$

各层样本的分配/比例分配II

- 分层随机抽样总体总量估计量的方差（比例分配）
  $V_{prop}(\hat{t}_{str}) = \sum_{h=1}^{H}(1-\frac{n_h}{N_h})N_h^2\frac{S_h^2}{n_h}$
  $= (1-\frac{n}{N})\frac{N}{n}\sum_{h=1}^{H} N_h S_h^2$
  $= (1-\frac{n}{N})\frac{N}{n}(SSW + \sum_{h=1}^{H} S_h^2)$
- 不放回简单随机抽样总体总量估计量的方差
  $V_{SRS}(\hat{t}) = (1-\frac{n}{N})N^2\frac{S^2}{n}$
  $= (1-\frac{n}{N})\frac{N^2}{n}\frac{SST}{N-1}$
  $= (1-\frac{n}{N})\frac{N^2}{n(N-1)}(SSW + SSB)$
- V_{prop}和V_{SRS}
  $V_{SRS}(\hat{t}) = V_{prop}(\hat{t}_{str}) + (1-\frac{n}{N})\frac{N}{n(N-1)}[N\cdot SSB - \sum_{h=1}^{H}(N-N_h)S_h^2]$
  $SSB > \sum_{h=1}^{H}(1-\frac{N_h}{N})S_h^2 \Leftrightarrow V_{SRS}(\hat{t}) > V_{prop}(\hat{t}_{str})$

各层样本的分配/最优分配

- 在分层随机抽样中，对于给定的费用，使估计量的方差$V(\bar{y}_{str})$达到最小，或者对于给定的估计量方差V，使得总费用达到最小的各层样本量的分配称为最优分配
- 总费用函数：$C = c_0 + \sum_{h=1}^{H} c_h n_h$，$C$为总费用，$c_0$为与样本量无关的固定费用，$c_h$为在第$h$层中抽取一个单元的平均费用
- 最优分配：对于分层随机抽样，若费用函数为$C = c_0 + \sum_{h=1}^{H} c_h n_h$，则最优分配为

$$n_h = n \cdot \frac{N_h S_h / \sqrt{c_h}}{\sum_{l=1}^{H} (N_l S_l / \sqrt{c_l})} \tag{1}$$

证明：加拉格朗日乘数或用Cauchy-Schwarz不等式，详见Appendix A5，《Sampling Methodologies with Application》2000, by Poduri S. R. S. Rao

拉哥朗日乘数

Lagrange Multipliers

Function: $f(x,y,z)$

Level Surface: $g(x,y,z)=C$

Lagrange Multiplier: λ

Lagrange Function (Lagrangian):
$\Lambda(x,y,\lambda)=f(x,y,z)+\lambda(g(x,y,z)-C)$

Solve: $\Lambda'(x,y,\lambda)_{x,y,\lambda}=0$

例1: $f(x,y)=xy;\ 3x^2+y^2=6$
例2: $f(x,y,z)=x+y+2z;\ x^2+y^2+z^2=3$

各层样本的分配/最优分配证明

证明：（拉格朗日乘数）

$V(\bar{y}_{str}) = \frac{1}{N^2}V(\hat{t}_{str}) = \sum_{h=1}^{H}(1-\frac{n_h}{N_h})(\frac{N_h}{N})^2\frac{S_h^2}{n_h}$

$= \sum_{h=1}^{H}\frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{H}\frac{W_h S_h^2}{N}$, $with\ W_h = \frac{N_h}{N}$

给定成本，使方差最小情况下的样本分配：

$\Delta_1 = V(\bar{y}_{str}) + \lambda(c_0 + \sum_{h=1}^{H} c_h n_h - C)$

给定方差，使成本最小情况下的样本分配：

$\Delta_2 = c_0 + \sum_{h=1}^{H} c_h n_h + \lambda(V(\bar{y}_{str}) - V)$

解法：对Δ_1和Δ_2中的n_h求导，让等式为0。

各层样本的分配

- 内曼最优分配：各层单位抽样费用相等情况下的最优分配
- ? 在内曼最优分配下，
  $V_{Neyman}(\hat{t}_{str}) = \frac{1}{n}(\sum_{h=1}^{H} N_h S_h)^2 - \sum_{h=1}^{H} N_h S_h^2$
- ? $V_{prop}(\hat{t}_{str}) - V_{Neyman}(\hat{t}_{str})$

  $= \frac{N^2}{n}[\sum_{h=1}^{H} \frac{N_h}{N} S_h^2 - (\sum_{h=1}^{H} \frac{N_h}{N} S_h)^2]$

  $= \frac{N^2}{n} \sum_{h=1}^{H} \frac{N_h}{N}(S_h - \sum_{l=1}^{H} \frac{N_l}{N} S_l)^2$
- ? $V_{prop}(\hat{t}_{str}) - V_{Neyman}(\hat{t}_{str}) = \frac{N_1 N_2}{n}(S_1 - S_2)^2, \ for\ H = 2$
- 样本容量：$V(\bar{y}_{str}) \leq \frac{1}{n} \sum_{h=1}^{H} \frac{n}{n_h}(\frac{N_h}{N})^2 S_h^2 = \frac{v}{n}$，$v = \sum_{h=1}^{H} \frac{n}{n_h}(\frac{N_h}{N})^2 S_h^2$
  若总体均值的95%置信区间为$\bar{y}_{str} \pm z_{\alpha/2}\sqrt{v/n}$，抽样设计的允许误差为$e = z_{\alpha/2}\sqrt{v/n}$，则$n = \frac{z_{\alpha/2}^2 v}{e^2}$