你永远不知道一个强迫症能干出什么事情

倪兴程[1]

2024 年 10 月 18 日

[1]Email: 19975022383@163.com

Todo list

目录

第一部分

统计学

0.1 方差

0.1.1 全方差公式

定理 0.1. 对于方差的分解，有如下公式：

$$Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$$

证明. 由方差与期望的关系：

$$\begin{aligned} E[Var(X|Y)] &= E[E(X^2|Y) - E^2(X|Y)] \\ &= E[E(X^2|Y)] - E[E^2(X|Y)] \\ &= E(X^2) - E[E^2(X|Y)] \end{aligned}$$

$$\begin{aligned} Var[E(X|Y)] &= E[E^2(X|Y)] - E^2[E(X|Y)] \\ &= E[E^2(X|Y)] - E^2(X) \end{aligned}$$

于是：

$$E[Var(X|Y)] + Var[E(X|Y)] = E(X^2) - E^2(X) = Var(X) \qquad \square$$

理解：

1. $E[Var(X|Y)]$ 是每个划分下方差的均值，刻画了样本内差异的均值。
2. $Var[E(X|Y)]$ 是不同划分下均值的方差，刻画了样本间差异的程度。

即：方差刻画了样本内和样本间差异的和。

0.2 Delta method

Delta method 可以给出随机变量函数的近似方差。

定理 0.2. 设随机向量 $\mathbf{X}$ 的均值为 $E(\mathbf{X})$，方差为 $Var(\mathbf{X})$，现有另一随机变量 $g(\mathbf{X})$，则该随机变量有如下近似方差：

$$Var[g(\mathbf{X})] \approx \nabla g\left[E(\mathbf{X})\right]^\top Cov(\mathbf{X}) \nabla g\left[E(\mathbf{X})\right]$$

证明. 将 $g(\mathbf{X})$ 在 $g\left[E(\mathbf{X})\right]$ 处进行泰勒展开：

$$g(\mathbf{X}) \approx g\left[E(\mathbf{X})\right] + \nabla g\left[E(\mathbf{X})\right]^\top \left[\mathbf{X} - E(\mathbf{X})\right]$$

对此式求方差：

$$Var\left[g(\mathbf{X})\right] \approx Var\left[\nabla g\left[E(\mathbf{X})\right]^\top \mathbf{X}\right] = \nabla g\left[E(\mathbf{X})\right]^\top Cov(\mathbf{X}) \nabla g\left[E(\mathbf{X})\right] \qquad \square$$

推论 0.1. 设随机变量 X 的均值为 $E(X)$，方差为 $Var(X)$，现有另一随机变量 $g(X)$，则该随机变量有如下近似方差：

$$Var\left[g(X)\right] \approx g'\left[E(X)\right]^2 Var(X)$$

0.3 列联表独立性问题

检验列联表的行列变量之间是否是独立的。

假设

$$H_0：\text{行变量与列变量是独立的} \Leftrightarrow H_1：\text{行变量与列变量是相关的}$$

0.3.1 Pearson 近似检验

原理

假设行变量有 r 个取值，列变量有 c 个取值，列联表中频数记为 $n_{ij},\ i=1,2,\ldots,r,\ j=1,2,\ldots,c$，行频数总和 $n_{i\cdot}=\sum_j n_{ij}$，列频数总和 $n_{\cdot j}=\sum_i n_{ij}$，频数总和 $n=\sum_{i,j} n_{ij}$，列联表中第 ij 个格子的理论频数为 E_{ij}，行变量取第 i 个值的概率为 $p_{i\cdot}$，列变量取第 j 个值的概率为 $p_{\cdot j}$，一个观测值被分配到列联表中第 ij 个格子的理论概率为 p_{ij}。

写完概率论把独性链接到这里来

若行变量与列变量独立，由随机变量的独立性，有：

$$p_{ij}=p_{i\cdot}p_{\cdot j},\ E_{ij}=p_{\cdot j}n_{i\cdot}$$

但由于 $p_{\cdot j}$ 是理论值无法预知，用 $\frac{n_{\cdot j}}{n}$ 来代替，那么 $E_{ij}=\hat{p_{\cdot j}}n_{i\cdot}=\frac{n_{i\cdot}n_{\cdot j}}{n}$。由此构建以下 Pearson$\chi^2$ 统计量：

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$$

如果零假设不成立，那么 n_{ij} 与 E_{ij} 的值相差就会比较大，统计量的值也会偏大。若统计量的值过大，则有里有怀疑零假设。由此可看出这里只考虑上侧的单侧检验问题。

大样本近似

在大样本的情况下，若零假设成立，有如下近似分布：

$$Q\sim\chi^2_{(r-1)(c-1)}$$

0.3.2 低维列联表的 Fisher 精确检验

假定五个边际频数都是固定的，在零假设成立的情况下，这个具体的列联表出现的条件概率（给定边际频数的情况下，因此是条件概率）只依赖于四个频数中的任意一个，且该概率满足超几何分布：

$$P=\frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

若零假设成立，那么任何一个关于 n_{ij} 的尾概率 $P(n_{ij}\leqslant a),\ P(n_{ij}\geqslant b),\ 1-P(a<n_{ij}<b)$ 都不应该太小或太大。若尾概率太小或太大，则有理由怀疑零假设。由此可以看出这里其实是涉及备择假设的方向问题的，可以单侧也可以双侧。

有机会看看单侧实际意义，不懂么会考虑单侧

	B1	B2	总和
A1	n_{11}	n_{12}	$n_{1\cdot}$
A2	n_{21}	n_{22}	$n_{2\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	n

表 1: 二维列联表

代码

Pearson 近似检验

直接将列联表矩阵输入以下函数即可。

```
chisq.test(x)
```

Fisher 精确检验

直接将列联表矩阵输入以下函数即可。

```
fisher.test(x, alternative="two.sided")
```

0.4 似然比检验

目的

有两个对未知分布 F 参数 θ 的点估计 θ_1 和 θ_2，需要去判断哪个估计更好。

假设

$$H_0 : \theta = \theta_1 \quad H_1 : \theta = \theta_2$$

原理

令 L 表示最大似然函数，则检验统计量为：

$$\lambda = \frac{L(X|\theta_1)}{L(X|\theta_2)}$$

如果 λ 偏小，则说明在 θ_2 下样本出现的概率更大，那么就应该拒绝零假设。当样本量 n 足够大时，有：

$$-2\ln\lambda \sim \chi^2_p, \quad p = dim(\Theta_0 \cup \Theta_1) - dim(\Theta_0)$$

第二部分

抽样调查

符号说明

符号		说明
N		总体中个体的总数
n		样本单元个数
Y_k	$k=1,2,\ldots,N$	总体中个体的值
y_k	$k=1,2,\ldots,n$	样本中样本单元的值
$p(s)$		样本概率
π_k	$k=1,2,\ldots,N$	个体的入样概率
w_k	$k=1,2,\ldots,N$	个体的抽样权重
τ		总量
μ		均值
σ^2		方差
s^2		样本方差
Z_i	$i=1,2,\ldots,N$	表示个体是否入样的示性变量
Q_i	$i=1,2,\ldots,N$	个体在样本中出现的次数
p		阳性率
d		误差幅度
X_k	$k=1,2,\ldots,N$	总体中个体对应的辅助变量值
x_k	$k=1,2,\ldots,n$	样本中样本单元对应的辅助变量值
B		比例估计系数
H		分层抽样总层数
h	$h=1,2,\ldots,H$	层下标
S_h	$h=1,2,\ldots,H$	某一层中所有个体构成的集合
N_h	$h=1,2,\ldots,H$	层中个体的总数
n_h	$h=1,2,\ldots,H$	层中的样本数
Y_{hj}	$h=1,2,\ldots,H,\ j=1,2,\ldots,N_h$	层中个体的值
y_{hj}	$h=1,2,\ldots,H,\ j=1,2,\ldots,n_h$	层中样本单元的值
c_h	$h=1,2,\ldots,H$	层中抽样的平均成本

表 2: 符号说明表

Chapter 1

抽样调查基本概念

1.1 抽样调查的目的

抽样调查最直接的主要任务，就是根据测得的样本（样本需要能够反应总体的差异）数量指标 $\{y_1, y_2, \ldots, y_n\}$，对总体 $\{Y_1, Y_2, \ldots, Y_N\}$ 的一些数字特征进行估计。如估计：

1. 总体均值$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$, 总体总量$\tau = \sum_{i=1}^{N} Y_i$。

2. [1]总体方差$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2$。

3. 总体中满足某一特征的单元所占比例 p。

4. 总体分布的分位数。

1.2 目标群体、源群体、研究群体

- 目标群体：研究者希望对其进行描述或推断的群体。

- 源群体：研究者实际用于选择样本的群体，通常是目标群体的一个子集或一个接近目标群体的群体。

- 研究群体：抽样后的样本

[1]这里指的是有限总体的方差，我们在抽样调查中认为有限总体也相当于一个样本，以样本方差公式计算其方差。在无限总体的情况下，分母取 N。

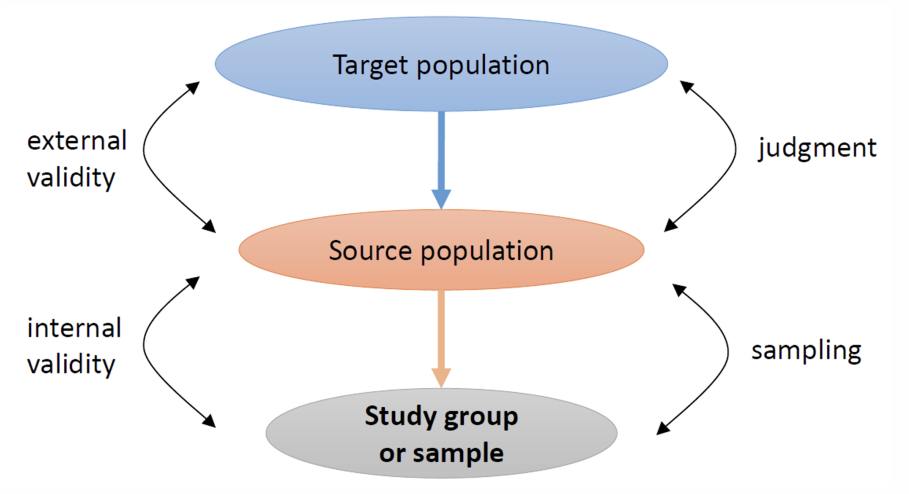

图 1.1: 群体

1.3 抽样设计

- 抽样设计
  - 概率抽样
    - 简单随机抽样
    - 分层抽样
    - 整群抽样
  - 非概率抽样
    - 方便抽样
    - 雪球抽样
    - 判断抽样

- 概率抽样样本是随机产生的，非概率抽样样本不是随机产生的。
- 非概率抽样无法判断样本是否具备代表性，也就无法确定抽样误差。但在一定程度上可以说明总体的特征。
- 方便抽样是选择最容易获得的样本作为研究对象。
- 雪球抽样是先选择一个较小的样本，再让这个样本中的样本单元去提供一些样本。
- 判断抽样是研究人员从总体中选择自己认为最能代表总体的个体作为样本。

1.4 抽样框架

对可以选择作为样本单元的群体单位列出的名册或排序编号，用以确定总体的抽样范围和结构。

1.5 抽样调查随机性的来源

1.5.1 基于模型的方法

绝大多数统计学课程是基于模型的 (model-based approch)，它将总体看作是背后随机变量的实现，随机性来源于随机变量的取值。

1.5.2 基于设计的方法

抽样调查是基于设计的 (design-based approch)，它将总体中个体的值看作为定值，而随机性来源于是否抽到该个体。

1.6 概率抽样概述

1.6.1 样本概率

记一个可能的样本为 s，记在抽样设计下出现这个样本的概率[2]为 $p(s)$，应有 $\sum_s p(s) = 1$。

1.6.2 入样概率与抽样权重

对于任意个体 Y_k，该个体的入样概率 π_k 记为 $\sum_{Y_k \in s} p(s)$，该个体的抽样权重 w_k 定义为 $\frac{1}{\pi_k}$，表示该个体可以代表总体里的多少个个体。若每个个体的抽样权重都一样，则称产生的样本为自加权 (self-weighting)样本。

[2]样本出现的概率未必是等可能的。

Chapter 2

简单随机抽样

简单随机抽样分为不放回型简单随机抽样 (simple random sampling, SRS)与放回型简单随机抽样 (simple random sampling with replacement, SRSWR)两种。我们通常更喜欢不放回抽样，因为同一个个体在样本中多次出现并不能提供额外的信息，同时有放回抽样会导致估计量的方差更大。

简单随机抽样意味着每个样本出现的概率是一样的，即 $p(s)$ 一致，那么每个个体的入样概率也是一致的（样本出现概率一致时每个个体的入样概率服从古典概型）。

2.1 SRS 的参数估计

SRS 中 Z 的相关性质

定理 2.1. *SRS 中表示个体是否入样的示性变量具有如下性质：*

$$
\begin{aligned}
E(Z_i) &= \frac{n}{N} \\
Var(Z_i) &= \frac{n}{N}\left(1-\frac{n}{N}\right) \\
Cov(Z_i, Z_j) &= \frac{-n}{N(N-1)}\left(1-\frac{n}{N}\right)
\end{aligned}
$$

证明. 每个个体被抽到的概率为：

$$
\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}
$$

由此可知个体示性变量的期望：

$$
E(Z_i) = 1 \times P(Z_i = 1) = 1 \times \pi_i = \frac{n}{N}
$$

注意到 $E(Z_i^2) = 0 \times P(Z_i^2 = 0) + 1 \times P(Z_i^2 = 1) = P(Z_i = 1) = E(Z_i)$，即有：

$$
Var(Z_i) = E(Z_i^2) - E^2(Z_i) = E(Z_i)\left(1 - E(Z_i)\right) = \frac{n}{N}\left(1-\frac{n}{N}\right)
$$

注意到 $E(Z_iZ_j)=P(Z_i=1,\ Z_j=1)$，所以：

$$Cov(Z_i,Z_j)=E(Z_iZ_j)-E(Z_i)E(Z_j)=\frac{\binom{N-2}{n-2}}{\binom{N}{n}}-E^2(Z_i)=\frac{-n}{N(N-1)}\left(1-\frac{n}{N}\right)\quad\square$$

2.1.1 SRS 总体均值 μ 的估计

点估计及点估计的性质

定理 2.2. *在 SRS 中中利用样本均值可对总体均值 μ 给出如下点估计：*

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{N}Y_iZ_i$$

该点估计具有如下性质：

$$E(\hat{\mu})=\mu,\ Var(\hat{\mu})=\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}$$

$$\widehat{Var}(\hat{\mu})=\left(1-\frac{n}{N}\right)\frac{s^2}{n}\text{是关于}Var(\hat{\mu})\text{的无偏估计。}$$

其中方差的计算见下文总体总数方差的推导，由方差的性质即可推导出总体均值的方差。[1]

区间估计

定理 2.3. *由大数定律，大样本下有：*

$$\frac{\hat{\mu}-\mu}{\sqrt{Var(\hat{\mu})}}\sim N(0,\ 1)$$

所以 SRS 中总体均值 μ 的区间估计如下：

$$\hat{\mu}\pm u_{1-\frac{\alpha}{2}}\times\sqrt{Var(\hat{\mu})}$$

由于 $Var(\hat{\mu})$ 的计算中涉及未知参数 σ^2，以 $\widehat{Var}(\hat{\mu})$ 代替，因此置信度为 $(1-\alpha)$ 的估计的双侧置信区间为：

$$\hat{\mu}\pm u_{1-\frac{\alpha}{2}}\times\sqrt{\widehat{Var}(\hat{\mu})}$$

2.1.2 SRS 总体总数 τ 的估计

点估计

利用样本均值进行估计

定义 2.1. *在 SRS 中利用样本均值可对总体总数 τ 给出如下点估计：*

$$\hat{\tau}=N\hat{\mu}=\frac{N}{n}\sum_{i=1}^{n}y_i=\frac{N}{n}\sum_{i=1}^{N}Y_iZ_i$$

[1] $\left(1-\frac{n}{N}\right)$ 被称之为有限群体校正分数 (finite population correction fraction, FPC)，有放回抽样或 N 远大于 n 时不需要 FPC。

Horvitz-Thompson 估计量（HT 估计量）

定义 2.2. 在 SRS 中引入抽样权重可给出关于 τ 的 HT 估计：

$$\hat{\tau} = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{N} w_i Y_i Z_i$$

点估计的性质

定理 2.4. 关于 SRS 总体总数 τ 的点估计有如下性质：

$$E(\hat{\tau}) = \tau$$

$$Var(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}, \quad \widehat{Var}(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

下给出点估计估计量方差公式的证明。

证明. 将方差展开可得到：

$$\begin{aligned} Var(\hat{\tau}) &= Var\left(\sum_{i=1}^{N} w_i Y_i Z_i\right) \\ &= \sum_{i=1}^{N} Var(w_i Y_i Z_i) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} Cov(w_i Y_i Z_i, w_j Y_j Z_j) \\ &= \sum_{i=1}^{N} w_i^2 Y_i^2 Var(Z_i) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_i Y_i w_j Y_j Cov(Z_i, Z_j) \end{aligned}$$

注意到 $w_i = \frac{N}{n}$ 并代入 Z_i 相关性质的公式可以得到：

$$\begin{aligned} Var(\hat{\tau}) &= \sum_{i=1}^{N} w_i^2 Y_i^2 Var(Z_i) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_i Y_i w_j Y_j Cov(Z_i, Z_j) \\ &= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left[(N-1)Y_i^2 - \sum_{j=i+1}^{N} 2Y_i Y_j\right] \\ &= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}(N-1)Y_i^2 - \sum_{i=1}^{N}\sum_{j=i+1}^{N} 2Y_i Y_j\right] \\ &= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[\sum_{i=1}^{N}(N-1)Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2 + \sum_{i=1}^{N} Y_i^2\right] \\ &= \frac{N}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left[N\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2\right] \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\mu^2\right) \end{aligned}$$

$$
\begin{aligned}
Var(\hat{\tau}) &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left(\sum_{i=1}^{N}Y_i^2 - 2N\mu^2 + N\mu^2\right) \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left(\sum_{i=1}^{N}Y_i^2 - 2\mu\sum_{i=1}^{N}Y_i + N\mu^2\right) \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\mu)^2 \\
&= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sigma^2
\end{aligned}
$$

□

区间估计

定理 2.5. 由大数定律，大样本下有：

$$\frac{\hat{\tau}-\tau}{\sqrt{Var(\hat{\tau})}} \sim N(0,\ 1)$$

所以 *SRS* 中总体总数 τ 的区间估计如下：

$$\hat{\tau} \pm u_{1-\frac{\alpha}{2}} \times \sqrt{Var(\hat{\tau})}$$

由于 $Var(\hat{\tau})$ 的计算中涉及未知参数 σ^2，以 $\widehat{Var}(\hat{\tau})$ 代替，因此置信度为 $(1-\alpha)$ 的估计的双侧置信区间为：

$$\hat{\tau} \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{Var}(\hat{\tau})}$$

2.1.3 SRS 总体方差 σ^2 的估计

定义 2.3. 在 *SRS* 中利用样本方差可对总体方差 σ^2 给出如下点估计：

$$\hat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{N}(Y_i-\hat{\mu})^2 Z_i$$

2.2 SRS 阳性率问题

阳性率问题是前述问题的一种特殊形式，Y_i 只能在 0 和 1 中取值。因此阳性率 p 即为总体均值 μ。

2.2.1 阳性率 p 的估计

点估计

定义 2.4. 在 *SRS* 中利用样本阳性率可给出阳性率 p 的点估计如下：

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n}y_i = \frac{1}{n}\sum_{i=1}^{N}Y_i Z_i$$

该点估计具有如下性质：

$$E(\hat{p}) = pVar(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}, \quad \widehat{Var}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

区间估计

定理 2.6. 由大数定律，大样本下有：

$$\frac{\hat{p}-p}{\sqrt{Var(\hat{p})}} \sim N(0,\ 1)$$

所以 SRS 中阳性率 p 的区间估计如下：

$$\hat{p} \pm u_{1-\frac{\alpha}{2}} \sqrt{Var(\hat{p})}$$

由于 $Var(\hat{p})$ 的计算中涉及未知参数 σ^2，以 $\widehat{Var}(\hat{p})$ 代替，因此置信度为 $(1-\alpha)$ 的估计的双侧置信区间为：

$$\hat{p} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{p})}$$

上述计算公式需要满足 $n\hat{p} \geqslant 5$ 和 $n(1-\hat{p}) \geqslant 5$，即大样本条件。

2.2.2 总体方差 σ^2 的估计

总体方差 σ^2 与总体均值 p 的关系

由于 $Y_i^2 = Y_i$，可得：

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-p)^2 = \frac{N}{N-1}p(1-p)$$

点估计

定义 2.5. 在 SRS 中利用样本方差可对总体方差 σ^2 给出如下点估计：

$$\hat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

2.3 SRSWR 的参数估计

SRSWR 中 Q 的相关性质

定理 2.7. *SRSWR* 中表示个体在样本中出现次数的变量 Q_i 具有如下性质：

$$\overrightarrow{Q} = (Q_1, Q_2, \ldots, Q_N) \sim Multi\left(n,\ \overbrace{\left(\frac{1}{N},\ \frac{1}{N}, \ldots, \frac{1}{N}\right)}^{N\text{个}\ \frac{1}{N}}\right)$$

$$Q_i \sim Binom\left(n,\ \frac{1}{N}\right)$$

$$E(Q_i) = \frac{n}{N},\ Var(Q_i) = \frac{n}{N}\left(1-\frac{n}{N}\right),\ Cov(Q_i,\ Q_j) = -\frac{n}{N^2}$$

证明. 由多项分布性质：多项分布随机变量的一维边际分布是二项分布，因此上第二、三式成立。下证第四式，将 Q_i 分解为独立二项分布随机变量的和，有：

$$\begin{aligned} Cov(Q_i,\ Q_j) &= Cov\left[\sum_{k=1}^{n} I_i(k),\ \sum_{l=1}^{n} I_j(l)\right] \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} Cov[I_i(k),\ I_j(l)] \\ &= \sum_{k=l} Cov[I_i(k),\ I_j(l)] + \sum_{k\neq l} Cov[I_i(k),\ I_j(l)] \end{aligned}$$

由二项分布随机变量的独立性，后一项为 0：

$$Cov(Q_i,\ Q_j) = \sum_{k=1}^{n} Cov[I_i(k),\ I_j(k)] = \sum_{k=1}^{n} \{E[I_i(k)I_j(k)] - E[I_i(k)]E[I_j(k)]\}$$

由于同一次伯努利实验中不可能出现两个结果，所以前一项为 0：

$$Cov(Q_i,\ Q_j) = -\sum_{k=1}^{n} E[I_i(k)]E[I_j(k)] = -np_ip_j = -\frac{n}{N^2}$$ □

2.3.1 SRSWR 总体总数 τ 的估计

定理 2.8. *在 SRS 中中利用样本均值可对总体总数 τ 给出如下点估计：*

$$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{N} Y_iQ_i$$

该点估计具有如下性质：

$$E(\hat{\tau}) = \tau,\ Var(\hat{\tau}) = \frac{N(N-1)}{n}\sigma^2$$

证明. 将方差展开可得到：

$$\begin{aligned} Var(\hat{\tau}) &= Var\left(\frac{N}{n}\sum_{i=1}^{N} Q_iY_i\right) \\ &= \frac{N^2}{n^2}\left[\sum_{i=1}^{N} Y_i^2 Var(Q_i) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} Y_iY_jCov(Q_i,Q_j)\right] \\ &= \frac{N^2}{n^2}\left[\sum_{i=1}^{N} Y_i^2\frac{n}{N}\left(1-\frac{1}{N}\right) + 2\sum_{i=1}^{N}\sum_{j=i+1}^{N} Y_iY_j\frac{-n}{N^2}\right] \\ &= \frac{N}{n}\left[\sum_{i=1}^{N} Y_i^2\left(1-\frac{1}{N}\right) - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} 2Y_iY_j\right] \end{aligned}$$

此处使用平方和公式（该技巧常用）：

$$\begin{aligned} Var(\hat{\tau}) &= \frac{N}{n}\left\{\sum_{i=1}^{N} Y_i^2\left(1-\frac{1}{N}\right)-\frac{1}{N}\left[\left(\sum_{i=1}^{N} Y_i\right)^2-\sum_{i=1}^{N} Y_i^2\right]\right\} \\ &= \frac{N}{n}\left[\sum_{i=1}^{N} Y_i^2-\frac{1}{N}\left(\sum_{i=1}^{N} Y_i\right)^2\right] \end{aligned}$$

此处使用 $N\mu^2$ 并进行加减凑项（该技巧常用）：

$$Var(\hat{\tau}) = \frac{N}{n}\left(\sum_{i=1}^{N} Y_i^2 - N\mu^2\right) = \frac{N}{n}\left(\sum_{i=1}^{N} Y_i^2 - 2N\mu^2 + N\mu^2\right)$$

此处使用 $N\mu = \sum_{i=1}^{N} Y_i$（该技巧常用）：

$$Var(\hat{\tau}) = \frac{N}{n}\left(\sum_{i=1}^{N} Y_i^2 - 2\mu\sum_{i=1}^{N} Y_i + N\mu^2\right) = \frac{N}{n}\sum_{i=1}^{N}(Y_i-\mu)^2 = \frac{N(N-1)}{n}\sigma^2 \qquad \square$$

2.4 样本容量的选择

抽样分布的方差决定置信区间的长度，样本容量增大的时候抽样分布的方差减小，置信区间变窄。

注意事项

通过控制总体均值的置信区间长度去选择样本容量而不从总体总数来考虑，因为总体总数的方差计算中还会涉及到 N^2，这对于无穷总体是无法进行计算的。无穷总体时忽略被 FPC。

误差幅度 MOE

置信区间的半径称之为误差幅度 (margin of error, MOE)。

2.4.1 总体均值样本容量公式

定理 2.9. *待估参数为总体均值时有如下样本容量公式：*

$$n_{SRS} = \frac{1}{\frac{d^2}{u^2\sigma^2}+\frac{1}{N}}, \quad n_{SRSWR} = \frac{u^2\sigma^2}{d^2}$$

其中 u 为求解区间估计过程中选择的正态分布分位数。

上式中二者有关系（又称为两步法）：

$$\frac{1}{n_{SRS}} = \frac{1}{n_{SRSWR}} + \frac{1}{N}$$

公式涉及到总体方差真实值 σ^2，解决方案：

1. 使用历史数据的样本方差代替 σ^2。
2. 由正态分布的性质，在 $\mu \pm 2\sigma$ 范围内应包含了 97.7% 的样本，因此，我们使用样本的极差来近似 4σ：用样本极差除 4 替代 σ。但是这个时候又涉及到极差从何而来的问题，因为是先确定样本容量再去做抽样，没有样本怎么来的极差呢？查阅资料得到样本的大致分布范围。

2.4.2 总体阳性率样本容量公式

定理 2.10. *待估参数为总体阳性率时有如下样本容量公式：*

$$n_{SRS} = \frac{Np(1-p)}{\frac{d^2}{u^2}(N-1)+p(1-p)}, \quad n_{SRSWR} = \frac{u^2 p(1-p)}{d^2}$$

其中 u 为求解区间估计过程中选择的正态分布分位数。

但这里需要注意，阳性率问题两步法不能用，与理论公式不等。
公式涉及到阳性率的真实值 p，解决方法：

1. 有根据的推测，使用历史数据来替代真实阳性率。
2. 取 $p = 0.5$，最大化样本容量，进行保守估计。
3. 如果获取额外样本的代价大于开始一次抽样的代价（也就意味着最大化样本容量带来的代价大于去做一次抽样来估计一下真实值的代价），那在没有历史数据的情况下可以自己去抽样。

阳性率问题下理论公式难以推导的情况

蒙特卡罗，代码如下：

```
sample_size_p <- function(p, N, n, conf, repeat.times=10000) {
  moe_mc_help <- function(p, N, n, conf, repeat.times) {
    alpha <- 1 - conf
    u <- qnorm(1 - alpha / 2)
    X <- rep(c(0, 1), times = c(round(N * (1 - p)), round(N * p)))
    phat <- NULL
    for (i in 1:repeat.times) {
      x <- sample(X, n)
      phat <- append(phat, mean(x))
    }
    v <- var(phat)
    moe <- sqrt(v) * u
```

```
    moe
  }
  data.frame(n, sapply(n, moe_mc_help, p=p, N=N, conf=conf,
                       repeat.times=repeat.times))
}
set.seed(1234)
sample_size_p(0.15, 3000, 860:880, 0.95)
```

2.5 简单随机抽样适用条件

1. 可使用的额外信息较少。

2. 研究多元关系，没有特别特殊的理由使用别的抽样方法。

Chapter 3

回归估计与比例估计

本章介绍比例估计 (regression estimate)与比例估计 (ratio estimate)，回归估计认为个体值 y 与某个变量 x（即辅助变量 (auxiliary variable)）之间存在如下关系：

$$y = B_1 x + B_0$$

比例估计中取 $B_0 = 0$，因此它可以看作为回归估计的一个特例。先来介绍辅助变量。

3.1 辅助变量

在回归估计中我们通过辅助变量去得到我们想要的估计值。

3.1.1 辅助变量选择要求

辅助变量需要满足以下条件：

1. 辅助变量的获得需要简单快捷。如果它的值都很难得到或者得不到那根本没办法作回归估计。
2. 辅助变量需要和个体值之间存在高度的线性相关性，在比例估计中我们还要求二者线性方程中的截距为 0（如果存在截距那截距就会是一个偏倚）。

选择好辅助变量并获取数据后，可以去做辅助变量与样本单元值的线性回归，来检验是否满足高度线性相关性（这里其实假设了数据满足正态分布）。在比例估计中我们还要去检验截距是否为 0。如果线性回归后的截距很小但不显著，我们可以认为满足要求；如果截距较大但不显著，我们认为是样本随机性带来的问题，可以认为截距满足为 0 的条件。

3.1.2 辅助变量与个体值的相关性

定理 3.1. *对于辅助变量与个体值的相关性，有如下结论：*

$$Corr(\bar{x}, \bar{y}) = Corr(X, Y)$$

证明. 将协方差进行展开：

$$
\begin{aligned}
Cov(\bar{x},\bar{y}) &= Cov\left(\frac{1}{n}\sum_{i=1}^{N}X_iZ_i, \frac{1}{n}\sum_{j=1}^{N}Y_jZ_j\right)\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N}X_iY_iVar(Z_i) + 2\sum_{i=1}^{N}\sum_{j\neq i}^{N}X_iY_jCov(Z_i,Z_j)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{1}{nN}\frac{1}{N-1}\left[(N-1)\sum_{i=1}^{N}X_iY_i - \sum_{i=1}^{N}\sum_{j\neq i}^{N}X_iY_j\right]\\
&= \left(1-\frac{n}{N}\right)\frac{1}{nN}\frac{1}{N-1}\left[(N-1)\sum_{i=1}^{N}X_iY_i - \left(\sum_{i=1}^{N}\sum_{j=1}^{N}X_iY_j - \sum_{i=1}^{N}X_iY_i\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{1}{nN}\frac{1}{N-1}\left(N\sum_{i=1}^{N}X_iY_i - \sum_{i=1}^{N}\sum_{j=1}^{N}X_iY_j\right)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{nN}\frac{1}{N-1}\left(N\sum_{i=1}^{N}X_iY_i - \sum_{i=1}^{N}X_i\sum_{j=1}^{N}Y_j\right)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\left(\sum_{i=1}^{N}X_iY_i - N\mu_X\mu_Y\right)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\left(\sum_{i=1}^{N}X_iY_i - 2N\mu_X\mu_Y + N\mu_X\mu_Y\right)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\left(\sum_{i=1}^{N}X_iY_i - \sum_{i=1}^{N}X_i\mu_Y - \sum_{i=1}^{N}Y_i\mu_X + N\mu_X\mu_Y\right)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{N-1}\sum_{i=1}^{N}(X_iY_i - X_i\mu_Y - Y_i\mu_X + \mu_X\mu_Y)\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{\sum_{i=1}^{N}(X_iY_i - X_i\mu_Y - Y_i\mu_X + \mu_X\mu_Y)}{N-1}\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y)}{N-1}\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}Cov(X,Y)
\end{aligned}
$$

所以：

$$
\begin{aligned}
Corr(\bar{x},\bar{y}) &= \frac{Cov(\bar{x},\bar{y})}{\sqrt{Var(\bar{x})Var(\bar{y})}}\\
&= \frac{(1-\frac{n}{N})\frac{1}{n}Cov(X,Y)}{\sqrt{(1-\frac{n}{N})^2\frac{1}{n^2}Var(X)Var(Y)}}\\
&= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}\\
&= Corr(X,Y)
\end{aligned}
$$

□

3.2 估计量

3.2.1 回归估计量

定义 3.1. 比例估计量有如下计算公式（其中 μ_X 和 τ_X 是已知的）：

$$\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{s_y\hat{R}}{s_x}, \quad \hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$$

$$\hat{\mu}_{Y_{reg}} = \hat{B}_1\mu_X + \hat{B}_0 = \hat{B}_1(\mu_X - \bar{x}) + \bar{y}, \quad \hat{\tau}_{Y_{reg}} = \hat{B}_1\tau_X + \hat{B}_0$$

3.2.2 比例估计量

定义 3.2. 比例估计量有如下计算公式（其中 μ_X 和 τ_X 是已知的）：

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\tau_y}{\tau_x}$$

$$\hat{\mu}_{Yr} = \hat{B}\mu_X, \quad \hat{\tau}_{Yr} = \hat{B}\tau_X$$

其中 $\hat{B}$ 是比例系数 B 的估计[1]，对于 B 的真实值，应有 $B = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}$。[2]

3.3 偏差

3.3.1 回归估计的偏差

由回归估计量计算公式显然可知回归估计是有偏的。

3.3.2 比例估计的偏差

定理 3.2. 比例估计量是有偏的，偏差如下：

$$bias(\hat{\mu}_{Yr}) = E(\hat{\mu}_{Yr}) - \mu_Y = Cov(-\hat{B}, \bar{x})$$

$$bias(\hat{\tau}_{Yr}) = E(\hat{\tau}_{Yr}) - \tau_Y = Cov(-\hat{B}, \bar{x})N$$

上式不便于计算，可以用下式进行估计：

$$bias(\hat{\mu}_{Yr}) \approx \frac{1}{\mu_X}\left[BVar(\bar{x}) - Cov(\bar{x},\bar{y})\right] = \left(1-\frac{n}{N}\right)\frac{1}{n\mu_X}[B\sigma_X^2 - Corr(X,Y)\sigma_X\sigma_Y]$$

$$bias(\hat{\tau}_{Yr}) \approx \frac{\tau_X}{\mu_X^2}\left[BVar(\bar{x}) - Cov(\bar{x},\bar{y})\right] = \left(1-\frac{n}{N}\right)\frac{\tau_X}{n\mu_X^2}[B\sigma_X^2 - Corr(X,Y)\sigma_X\sigma_Y]$$

[1]这是一个有偏估计!!!

[2]依据辅助变量的选择原则，X 与 Y 之间需要满足线性关系且截距为 0，B 其实就是线性关系中的斜率，同时证明了 $Corr(\bar{x},\bar{y}) = Corr(X,Y)$，因此 $\hat{B}$ 和 B 既可以由均值来表示也可以由总体总数来表示。

证明. 将协方差分解：

$$\begin{aligned} Cov(-\hat{B},\bar{x}) &= -Cov(\hat{B},\bar{x}) \\ &= -\left[E(\hat{B}\bar{x}) - E(\hat{B})E(\bar{x})\right] \\ &= -\left[E\left(\frac{\bar{y}}{\bar{x}}\bar{x}\right) - E(\hat{B})\mu_X\right] \\ &= E(\hat{\mu}_{Yr}) - E(\bar{y}) \\ &= E(\hat{\mu}_{Yr}) - \mu_Y \end{aligned}$$

下证近似公式：

$$\begin{aligned} bias(\hat{\mu}_{Yr}) &= E(\hat{B}\mu_X) - \mu_Y \\ &= \mu_X E\left(\frac{\bar{y}}{\bar{x}} - \frac{\mu_Y}{\mu_X}\right) \\ &= \mu_X E\left(\frac{\bar{y}}{\mu_X}\frac{\mu_X}{\bar{x}} - \frac{\mu_Y}{\mu_X}\right) \\ &= \mu_X E\left(\frac{\bar{y}}{\mu_X} - \frac{\bar{y}}{\mu_X}\frac{\bar{x}-\mu_X}{\bar{x}} - \frac{\mu_Y}{\mu_X}\right) \\ &= \mu_X E\left(\frac{\bar{y}}{\mu_X}\frac{\mu_X-\bar{x}}{\bar{x}}\right) \\ &= \mu_X E\left(\frac{\bar{y}}{\mu_X}\frac{\mu_X-\bar{x}}{\bar{x}}\frac{\mu_X}{\bar{x}}\frac{\bar{x}}{\mu_X}\right) \\ &= \mu_X E\left[\frac{\bar{y}}{\mu_X^2}(\mu_X-\bar{x})\frac{\mu_X}{\bar{x}}\right] \\ &= \mu_X E\left[\frac{\bar{y}}{\mu_X^2}(\mu_X-\bar{x})\left(1-\frac{\bar{x}-\mu_X}{\bar{x}}\right)\right] \\ &= \mu_X E\left\{\frac{\bar{y}}{\mu_X^2}\left[(\mu_X-\bar{x}) + \frac{(\bar{x}-\mu_X)^2}{\bar{x}}\right]\right\} \\ &= \mu_X E\left[\frac{-\bar{y}(\bar{x}-\mu_X)}{\mu_X^2} + \frac{\bar{y}(\bar{x}-\mu_X)^2}{\mu_X^2\bar{x}}\right] \\ &= \frac{1}{\mu_X}E\left[\frac{\bar{x}}{\bar{y}}(\bar{x}-\mu_X)^2 - (\bar{x}-\mu_X)\bar{y}\right] \\ &= \frac{1}{\mu_X}\left\{E[\hat{B}(\bar{x}-\mu_X)^2] - E[\bar{y}(\bar{x}-\mu_X)]\right\} \\ &= \frac{1}{\mu_X}\left\{E[(\hat{B}-B+B)(\bar{x}-\mu_X)^2] - Cov(\bar{x},\bar{y})\right\} \\ &= \frac{1}{\mu_X}\left\{E[(\hat{B}-B)(\bar{x}-\mu_X)^2 + B(\bar{x}-\mu_X)^2] - Cov(\bar{x},\bar{y})\right\} \end{aligned}$$

由于 $E\left[(\hat{B}-B)(\bar{x}-\mu_X)^2\right]$ 极小（严谨证明不提供），因此:

$$\begin{aligned} bias(\hat{\mu}_{Yr}) &\approx \frac{1}{\mu_X}\left\{E[B(\bar{x}-\mu_X)^2] - Cov(\bar{x},\bar{y})\right\} \\ &= \frac{1}{\mu_X}\left[BVar(\bar{x}) - Cov(\bar{x},\bar{y})\right] \end{aligned}$$

□

由此，在以下情况比例估计的偏倚会较小：

1. n 较大。
2. $\frac{n}{N}$ 较大。
3. μ_X 较大。
4. σ_x 较小。
5. R 接近 ± 1。

3.4 均方误差

3.4.1 回归估计量的均方误差

定理 3.3. *回归估计量的均方误差为：*

$$\begin{aligned} MSE(\hat{\mu}_{Y_{reg}}) &= E\left\{\left[\bar{y} + \hat{B}_1(\mu_X - \bar{x}) - \mu_Y\right]^2\right\} \\ &\approx Var(\bar{d}) \\ &= \left(1 - \frac{n}{N}\right)\frac{\sigma_d^2}{n} \end{aligned}$$

其中：

$$d_i = y_i - \left[\mu_Y + B_1(x_i - \mu_X)\right]$$

$$\sigma_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}[y_i - \mu_Y - B_1(x_i - \mu_X)]^2 = (1 - R^2)\sigma_Y^2$$

证明. 下求 $E(d)$：

$$E(d) = \frac{1}{N}\left[\sum_{i=1}^{N} y_i - N\mu_Y - B_1\sum_{i=1}^{N} x_i + B_1 N\mu_X\right] = 0$$

因此：

$$\begin{aligned} \sigma_d^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left[y_i - \mu_Y - B_1(x_i - \mu_X)\right]^2 \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - \mu_Y)^2 + \sum_{i=1}^{N} B_1^2(x_i - \mu_X)^2 - \sum_{i=1}^{N} 2B_i(y_i - \mu_Y)(x_i - \mu_X)\right] \\ &= \sigma_Y^2 + B_1^2\sigma_X^2 - 2B_1 R\sigma_X\sigma_Y \end{aligned}$$

而：

$$B_1 = \frac{\sigma_Y R}{\sigma_X}$$

所以：

$$\sigma_d^2 = \sigma_Y^2 + B_1^2\sigma_X^2 - 2B_1R\sigma_X\sigma_Y = \sigma_Y^2 + \frac{\sigma_Y^2R^2}{\sigma_X^2}\sigma_X^2 - 2\frac{\sigma_YR}{\sigma_X}R\sigma_X\sigma_Y = \sigma_Y^2 - R^2\sigma_Y^2 = (1-R^2)\sigma_Y^2$$

□

由此，在以下情况 $MSE(\hat{\mu}_{Y_{reg}})$ 较小：

1. n 较大。
2. $\frac{n}{N}$ 较大。
3. σ_Y 较小。
4. R 接近 ± 1。

3.4.2 比例估计量的均方误差

定理 3.4. *比例估计量的均方误差为：*

$$\begin{aligned} MSE(\hat{\mu}_{Yr}) &\approx E\left[(\bar{y} - B\bar{x})^2\right] \\ &= \left(1 - \frac{n}{N}\right)\frac{\sigma_Y^2 - 2BR\sigma_X\sigma_Y + B^2\sigma_X^2}{n} \\ &\approx Var(\hat{\mu}_{Yr}) \end{aligned}$$

证明过程可见 David and Sukhatme, 1974。

3.5 方差

3.5.1 回归估计的方差

定理 3.5. *回归估计的方差为：*

$$Var(\hat{\mu}_{Y_{reg}}) = Var(\bar{d}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_d^2}{n} = \left(1 - \frac{n}{N}\right)\frac{(1-R^2)\sigma_Y^2}{n}$$

定义 $e_i = y_i - (\hat{B}_1x_i + \hat{B}_0)$ 可得：

$$\widehat{Var}(\hat{\mu}_{Y_{reg}}) = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}$$

这里 s_e^2 可取两种计算公式：

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 = \left(1 - \frac{n}{N}\right)\frac{(1-\hat{R}^2)s_y^2}{n}, \quad s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

第二种是考虑回归估计有两个待估参数，自由度为 $n-2$，这样子做修正了回归中自由度的问题。

由上述总结：

$$\widehat{Var_1}(\hat{\mu}_{Y_{reg}}) = \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left[y_i-(\hat{B}_1x_i+\hat{B}_0)\right]^2$$
$$\widehat{Var_2}(\hat{\mu}_{Y_{reg}}) = \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{1}{n-2}\sum_{i=1}^{n}\left[y_i-(\hat{B}_1x_i+\hat{B}_0)\right]^2$$

3.5.2 比例估计的方差

由 Delta method（见定理 0.2），注意到关系：

$$\hat{B}=\frac{\bar{y}}{\bar{x}}=g(\bar{x},\bar{y})$$
$$\hat{\mu}_{Yr}=\hat{B}\mu_X=g(\hat{B})$$
$$\hat{\tau}_{Yr}=\hat{\mu}_{Yr}\frac{\tau_X}{\mu_X}=\hat{\mu}_{Yr}N$$

可得以下比例估计量的近似方差 ($R=Corr(X,Y)$)：

$$Var(\hat{B})\approx\left(1-\frac{n}{N}\right)\frac{\sigma_Y^2-2BR\sigma_X\sigma_Y+B^2\sigma_X^2}{n\mu_X^2}=\left(1-\frac{n}{N}\right)\frac{\sigma_\varepsilon^2}{n\mu_X^2}$$
$$\widehat{Var}(\hat{B})\approx\left(1-\frac{n}{N}\right)\frac{s_y^2-2\hat{B}\hat{R}s_xs_y+\hat{B}^2s_x^2}{n\bar{x}^2}=\left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2}$$
$$Var(\hat{\mu}_{Yr})\approx\left(1-\frac{n}{N}\right)\frac{\sigma_Y^2-2BR\sigma_X\sigma_Y+B^2\sigma_X^2}{n}=\left(1-\frac{n}{N}\right)\frac{\sigma_\varepsilon^2}{n}$$
$$\widehat{Var_1}(\hat{\mu}_{Yr})\approx\left(1-\frac{n}{N}\right)\frac{s_y^2-2\hat{B}\hat{R}s_xs_y+\hat{B}^2s_x^2}{n}=\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$$
$$Var(\hat{\tau}_{Yr})\approx N\left(N-n\right)\frac{\sigma_Y^2-2BR\sigma_X\sigma_Y+B^2\sigma_X^2}{n}=N(N-n)\frac{\sigma_\varepsilon^2}{n}$$
$$\widehat{Var_1}(\hat{\tau}_{Yr})\approx N(N-n)\frac{s_y^2-2\hat{B}\hat{R}s_xs_y+\hat{B}^2s_x^2}{n}=N(N-n)\frac{s_e^2}{n}$$

再给出第二种总体均值、总体总数比例估计量抽样分布方差的估计：

$$\widehat{Var_2}(\hat{\mu}_{Yr})=\widehat{Var}(\hat{B}\mu_X)=\widehat{Var}(\hat{B})\mu_X^2\approx\left(1-\frac{n}{N}\right)\left(\frac{\mu_X}{\bar{x}}\right)^2\frac{s_e^2}{n}$$
$$\widehat{Var_2}(\hat{\tau}_{Yr})\approx N(N-n)\left(\frac{\mu_X}{\bar{x}}\right)^2\frac{s_e^2}{n}$$

下给出上述公式中所有等式的推导。

证明. 从模型的角度，根据 MSE 的估计来看（最后一行是使用了 SRS 均值的方差公式）：

$$
\begin{aligned}
Var(\hat{\mu}_{Yr}) \approx MSE(\hat{\mu}_{Yr}) &\approx E\left[(\bar{y} - B\bar{x})^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} y_i - B\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] \\
&= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - Bx_i)\right]^2\right\} \\
\varepsilon_i = y_i - Bx_i,\ \mu_\varepsilon &= E(\varepsilon_i) = 0 \\
Var(\hat{\mu}_{Yr}) &\approx E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - Bx_i)\right]^2\right\} \\
&= E\left[(\bar{\varepsilon})^2\right] \\
&= E\left[(\bar{\varepsilon} - 0)^2\right] \\
&= E\left[(\bar{\varepsilon} - \mu_\varepsilon)^2\right] \\
&= Var(\bar{\varepsilon}) \\
&= \left(1 - \frac{n}{N}\right)\frac{\sigma_\varepsilon^2}{n}
\end{aligned}
$$

对于 σ_ε^2：

$$
\begin{aligned}
\sigma_\varepsilon^2 &= \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 + B^2\sum_{i=1}^{N} x_i^2 - 2B\sum_{i=1}^{N} x_i yi\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - \mu_Y + \mu_Y)^2 + B^2\sum_{i=1}^{N}(x_i - \mu_X + \mu_X)^2 - 2B\sum_{i=1}^{N} x_i yi\right] \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i - \mu_Y)^2 + 2\sum_{i=1}^{N}(y_i - \mu_Y)\mu_Y + N\mu_Y^2\right. \\
&\qquad \left. + B^2\sum_{i=1}^{N}(x_i - \mu_X)^2 + 2B^2\sum_{i=1}^{N}(x_i - \mu_X)\mu_X + B^2 N\mu_X^2 - 2B\sum_{i=1}^{N} x_i y_i\right] \\
&= \sigma_Y^2 + B^2\sigma_X^2 + \frac{1}{N-1}\left[N\mu_Y^2 + B^2 N\mu_X^2 - 2B\sum_{i=1}^{N} x_i y_i\right] \\
&= \sigma_Y^2 + B^2\sigma_X^2 + \frac{1}{N-1}\left[N\mu_Y^2 + B^2 N\mu_X^2 - 2B\sum_{i=1}^{N}(x_i - \mu_X + \mu_X)(y_i - \mu_Y + \mu_Y)\right] \\
&= \sigma_Y^2 + B^2\sigma_X^2 + \frac{1}{N-1}\left[N\mu_Y^2 + B^2 N\mu_X^2 - 2B\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y) - 2BN\mu_X\mu_Y\right] \\
&= \sigma_Y^2 + B^2\sigma_X^2 - 2BCov(X,Y) + \frac{1}{N-1}\left[N\mu_Y^2 + B^2 N\mu_X^2 - 2BN\mu_X\mu_Y\right]
\end{aligned}
$$

因为：

$$
B = \frac{\mu_Y}{\mu_X}
$$

所以：

$$\begin{aligned} N\mu_Y^2 + B^2 N\mu_X^2 - 2BN\mu_X\mu_Y &= N\mu_Y^2 + \frac{\mu_Y^2}{\mu_X^2} N\mu_X^2 - 2\frac{\mu_Y}{\mu_X} N\mu_X\mu_Y \\ &= 2N\mu_Y^2 - 2N\mu_Y^2 \\ &= 0 \end{aligned}$$

也就有：

$$\sigma_\varepsilon^2 = \sigma_Y^2 - 2BR\sigma_X\sigma_Y + B^2\sigma_X^2$$

令 $e_i = y_i - \hat{B}x_i$，将它看作为 ε_i 的估计，则有：

$$s_e^2 = s_y^2 - 2\hat{B}\hat{R}s_x s_y + \hat{B}^2 s_x^2$$ □

3.6 置信区间

3.6.1 回归估计的置信区间

由于比例估计量抽样分布的方差公式中存在未知量，大样本情况下可得如下估计的置信区间：

$$\hat{\mu}_{Y_{reg}} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\mu}_{Y_{reg}})}$$

3.6.2 比例估计的置信区间

由于比例估计量抽样分布的方差公式中存在未知量，大样本情况下可得如下估计的置信区间：

$$\hat{B} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{B})}$$

$$\hat{\mu}_{Yr} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\mu}_{Yr})}$$

$$\hat{\tau}_{Yr} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\tau}_{Yr})}$$

3.7 比例估计样本容量的选择

$$n_r = \frac{Nu^2\sigma_\varepsilon^2}{u^2\sigma_\varepsilon^2 + Nd^2}$$

3.8 回归估计、比例估计与 HT 估计的比较

回归估计与比例估计是有偏的，但它们的方差比 HT 估计小很多，MSE 更小。

Chapter 4

标记重捕法

不对标记重捕法的具体操作进行介绍，高中都学过。

下给出标记重捕法 (tag recapture)的符号说明。

1. X：初始样本容量，即被标记数据。
2. y：被重捕的样本数。
3. x：重捕样本中被标记的数量。
4. t：总体总数。

4.1 假设

1. 种群是封闭的，种群数量在标记与重捕期间没有增减。
2. 每个样本都是来自种群的简单随机样本。
3. 两次样本独立。
4. 标记不能丢失。

即：重捕个体中已标记个体的比例与种群中已标记个体的比例相等。

4.2 总体总数 t 的估计

4.2.1 点估计

定理 4.1. 标记重捕法中总体总数 τ 的点估计如下：

$$\hat{t} = \frac{y}{x}X$$

它有如下性质：

$$Var(\hat{t}) = \frac{(yX)^2}{E^3(x)}\frac{(t-y)(t-X)}{t(t-1)}, \quad \widehat{Var}(\hat{t}) = \frac{Xy(X-x)(y-x)}{x^3}$$

证明. 在标记重捕法中，$x \sim H(y, X, t)$。因此可得：

$$E(x) = \frac{yX}{t},\ Var(x) = \frac{yX(t-y)(t-X)}{t^2(t-1)}$$

而：

$$\hat{t} = yX\frac{1}{x}$$

所以由定理 0.2：

$$\begin{aligned} Var(\hat{t}) &= Var\left(yX\frac{1}{x}\right) \\ &= (yX)^2 Var\left(\frac{1}{x}\right) \\ &\approx (yX)^2\left[\frac{-1}{E^2(x)}\right]^2 Var(x) \\ &= \frac{(yX)^2}{E^4(x)}\frac{yX}{t}\frac{(t-y)(t-X)}{t(t-1)} \\ &= \frac{(yX)^2}{E^3(x)}\frac{(t-y)(t-X)}{t(t-1)} \end{aligned}$$

用 x 替代 $E(x)$（相当于是一次简单随机抽样，利用无偏性），然后考虑 t 较大时的近似，最后还剩一个 t，用 $\hat{t}$ 带入进行计算，即可得到：

$$\begin{aligned} \widehat{Var}(\hat{t}) &= \frac{(yX)^2}{x^3}\frac{(t-y)(t-X)}{t(t-1)} \\ &\approx \frac{(yX)^2}{x^3}\frac{(t-y)(t-X)}{t^2} \\ &= \frac{(yX)^2}{x^3}\left(1-\frac{X}{t}\right)\left(1-\frac{y}{t}\right) \\ &\approx \frac{Xy(X-x)(y-x)}{x^3} \end{aligned}$$

□

极端情况下的修正

在极端情况下，x 可能为 0 或很小，那么 $\widehat{Var}(\hat{t})$ 就会无限大，此时作如下修正（即不带入 $\hat{t}$，而是代入 $\tilde{t}$）：

$$\begin{aligned} \tilde{t} &= \frac{(X+1)(y+1)}{x+1} - 1 \\ \widehat{Var}(\tilde{t}) &= \frac{(X+1)(y+1)(y-x)(X-x)}{(x+1)^2(x+2)} \end{aligned}$$

4.2.2 区间估计

正态近似求置信区间

由点估计方差公式，易得如下估计的总体总数地置信区间：

$$\hat{t} \pm u_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{t})},\quad \tilde{t} \pm u_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\tilde{t})}$$

正态近似置信区间可能存在的问题

正态近似置信区间可能会存在：置信区间左端点小于两次捕捉到的总数的现象，这显然是不合理的。

Pearsonχ^2 检验求置信区间

由标记重捕法使用条件，第一次被捕到和第二次被捕到这两件事情是独立的，由此可构建如下的列联表：

	第二次捕获: 是	第二次捕获: 否
第一次捕获：是	a	b
第一次捕获：否	c	d

表 4.1: 标记重捕法的列联表示意图

在这个表里，a, c, b 显然都是已知的，只有 d 是未知的。可以通过给 d 赋值的方式，去检验列联表行列变量之间是否独立（参考 section 0.3.1），选择合适的 d 值（即让独立性检验结果显著的 d 值）作为置信区间。

似然比检验

由样本可计算出 $\hat{t}$，然后可以构建如下假设：

$$H_0 : \theta = \hat{t} \quad H_1 : \theta = \theta_A$$

进行似然比检验（参考 section 0.4）。置信区间为拒绝零假设的 θ_A 构成的区间，即比 $\hat{t}$ 更适合作为模型参数的 θ_A 构成置信区间。

bootstrap 求置信区间

在第二个样本中进行 bootstrap，有放回的抽取 y 个样本，计算每个样本对总体总数的估计值 $\hat{t}$，重复 N 次。将 N 个 $\hat{t}$ 从小到大排序，在此基础上取分位点即产生置信区间，

代码

以上四种方法的代码如下：

```r
tag_recapture_CI <- function(a, b, c, d, alpha = 0.05,
    method = c("normal", "chisq", "fisher", "likelihood", "bootstrap"),
    x.correct = FALSE, N = 1000, seed = 42) {
  # Ensure the 'method' parameter is valid
  method <- match.arg(method)

```

```r
  if (method == "chisq") {
    # Pearson's Chi-squared test
    p.values <- sapply(d, function(d_i) {
      chisq.test(matrix(c(a, b, c, d_i), nrow = 2, byrow = TRUE),
                 correct = FALSE)$p.value
    })
    valid_d <- d[p.values > alpha]
    if (length(valid_d) == 0) {
      stop("No values satisfy the p-value > alpha condition.")
    }
    return(sum(c(a, b, c)) + range(valid_d))

  } else if (method == "fisher"){
    # Fisher's exact test
    p.values <- sapply(d, function(d_i) {
      fisher.test(matrix(c(a, b, c, d_i), nrow = 2, byrow =
      ↪  TRUE))$p.value
    })
    valid_d <- d[p.values > alpha]
    if (length(valid_d) == 0) {
      stop("No values satisfy the p-value > alpha condition.")
    }
    return(sum(c(a, b, c)) + range(valid_d))

  } else if (method == "normal") {
    # Normal approximation method
    X <- a + b
    y <- a + c
    x <- a
    if (x.correct) {
      t_hat <- (X + 1) * (y + 1) / (x + 1) - 1
      V_hat <- ((X + 1) * (y + 1) * (y - x) * (X - x)) / ((x + 1)^2 * (x
      ↪  + 2))
    } else {
      t_hat <- y * X / x
      V_hat <- y * X * (y - x) * (X - x) / x^3
    }
    CI_lower <- t_hat - qnorm(1 - alpha / 2) * sqrt(V_hat)
```

```r
    CI_upper <- t_hat + qnorm(1 - alpha / 2) * sqrt(V_hat)
    return(c(CI_lower, CI_upper))

  } else if (method == "bootstrap") {
    # Bootstrap method
    set.seed(seed)
    sample.frame <- c(rep(1, a), rep(0, c)) # Construct sample frame
    bootstrap_estimates <- replicate(N, {
      sampled <- sample(sample.frame, a + c, replace = TRUE)
      x_boot <- sum(sampled)
      (a + b) * (a + c) / x_boot
    })
    CI <- quantile(bootstrap_estimates, probs = c(alpha / 2, 1 - alpha /
    ↪  2))
    return(floor(CI))

  } else if (method == "likelihood") {
    # Likelihood method
    t_hat <- (a + b) * (a + c) / a
    ll_max <- dhyper(a, a + b, t_hat - (a + b), a + c, log = TRUE)
    ll <- sapply(d, function(d_i) {
      dhyper(a, a + b, d_i, a + c, log = TRUE)
    })
    valid_d <- d[2 * (ll_max - ll) < qchisq(1 - alpha, df = 1)]
    if (length(valid_d) == 0) {
      stop("No values satisfy the likelihood condition.")
    }
    return(a + b + range(valid_d))
  }
}
```

Chapter 5

分层抽样

分层抽样与分类型辅助变量息息相关。其核心为：

把目标群体分为 H 个亚群体 (stratum)。亚群体之间不重叠，它们构成整个群体（每个个体属于且只属于某个亚群体）。在每个亚群体中通过概率抽样的方法进行独立抽样，然后汇集信息进行群体估计。

当亚群体内的个体值趋于一致时，分层抽样有意义。

为什么要使用分层抽样

分层抽样相比于 SRS 具有如下优势：

1. 分层抽样可以避免因为不同类型的样本对研究结果会产生显著差异从而导致的严重样本选择偏倚。

2. 分层抽样过程有可能更易于管理，同时可以降低成本。

3. 分层抽样不仅可以估计群体的特征，还可以估计亚群体特征。

4. 样本数相同的情况下，分层抽样通常比 SRS 更加精确。当亚群体内的个体值趋于一致时，该结论尤其正确。

分层随机抽样基本设置

1. 必须知道每个 $N_h,\ h = 1, 2, \ldots, H$。

2. 在每一个层里面独立地使用 SRS。

公式	含义
$\tau_h = \sum_{j=1}^{N_h} Y_{hj}$	第 h 层的总量
$\tau = \sum_{h=1}^{H} \tau_{Y_h}$	总体总量
$\mu_h = \frac{\sum_{j=1}^{N_h} Y_{hj}}{N_h}$	第 h 层的均值
$\mu = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} Y_{hj}}{N}$	总体均值
$\sigma_h^2 = \frac{\sum_{j=1}^{N_h}\left(Y_{hj}-\mu_h\right)^2}{N_h-1}$	第 h 层的方差
$\sigma^2 = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h}\left(Y_{hj}-\mu\right)^2}{N-1}$	总体方差

表 5.1: 分层抽样部分计算公式

5.1 参数估计

5.1.1 亚群体特征的估计

因为分层随机抽样在亚群体中为简单随机抽样，由简单随机抽样的估计公式即有如下公式：

$$\hat{\mu}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj}$$

$$\hat{\tau}_h = \frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj} = N_h\hat{\mu}_h$$

$$\hat{\sigma_h^2} = s_h^2 = \frac{1}{n_h-1}\sum_{j=1}^{n_h}(y_{hj}-\hat{\mu}_h)^2$$

5.1.2 总体总量 $\hat{\tau}$ 的估计

计算公式

$$\hat{\tau}_{str} = \sum_{h=1}^{H}\hat{\tau}_h = \sum_{h=1}^{H} N_h\bar{y}_h$$

抽样权重形式

$$\hat{\tau}_{str} = \sum_{h=1}^{H} N_h\bar{y}_h = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj} = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\frac{N_h}{n_h}y_{hj} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}y_{hj}$$

5.1.3 总体均值 $\hat{\mu}$ 的估计

计算公式

$$\hat{\mu}_{str} = \frac{\hat{\tau}_{str}}{N} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h$$

抽样权重形式

只需注意到 $N = \sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}$：

$$\hat{\mu}_{str} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_{hj}}$$

5.1.4 估计的性质

无偏性

无偏性是显然的：在每一层里面使用的是简单随机抽样，而简单随机抽样的估计量是无偏的，总和也自然是无偏的。

方差

由各层样本之间的独立性以及每层中的抽样实际是 SRS，立即可得如下分层随机抽样估计量的方差公式：

$$Var(\hat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}$$

$$\widehat{Var}(\hat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

$$Var(\hat{\mu}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}$$

$$\widehat{Var}(\hat{\mu}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

其中：

$$\sigma_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (Y_{hj} - \mu_h)^2$$
$$s_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

置信区间

$$\hat{\mu}_{str} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\mu}_{str})},\ \hat{\tau}_{str} \pm u_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var})(\hat{\tau}_{str})}$$

自由度问题

当使用 t 置信区间的时候，如果各层之间方差是齐的，那么自由度即为 $n - H$。如果方差不齐，则使用 Satterwaithe approximation 来估计自由度：

$$Dof = \left(\sum_{h=1}^{H} a_h s_h^2 \right)^2 \div \sum_{h=1}^{H} \frac{(a_h s_h^2)^2}{(n_h - 1)}$$

其中：

$$a_h = \frac{N_h(N_h - n_h)}{n_h}$$

5.1.5 群体比例问题

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$
$$\widehat{Var}(\hat{p}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n - 1}$$

5.2 估计方法思考

在分层抽样中，如果使用如下公式估计 μ：

$$\tilde{\mu}_{str} = \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} y_{hj}}{n}$$

均值

讨论阳性率问题。

该方法的总体均值为：

$$
\begin{aligned}
E(\tilde{\mu}_{str}) &= E\left(\frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h} y_{hj}}{n}\right) \\
&= \frac{1}{n}\sum_{h=1}^{H} E\left(\sum_{j=1}^{N_h} Y_{hj} Z_{hj}\right) \\
&= \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{N_h} Y_{hj} E(Z_{hj}) \\
&= \frac{1}{n}\sum_{h=1}^{H} E(Z_{hj}) \sum_{j=1}^{N_h} Y_{hj} \\
&= \frac{1}{n}\sum_{h=1}^{H} \frac{n_h}{N_h} N_h p_h \\
&= \frac{1}{n}\sum_{h=1}^{H} n_h p_h
\end{aligned}
$$

而真实的总体均值为：

$$
\mu = \frac{t}{N} = \frac{1}{N}\sum_{h=1}^{H} N_h p_h
$$

如果想要无偏，则显然需要满足：

$$
\frac{n_h}{N_h} = \frac{n}{N}
$$

5.3 分配原则

分层随机抽样要考虑两个问题：

1. 如何定义层？
2. 每个层里面样本量是多少？

5.3.1 比例分配

比例分配 (proportional allocation)是指在分层抽样中令 $\pi_{hj} = \frac{n}{N} = \frac{n_h}{N_h},\ h = 1, 2, \ldots, H,\ j = 1, 2, \ldots, N_h$。这种分配方式不会出现极端情况，即样本几乎都出自某一层的现象。

比例分配与 SRS 的比较

可以注意到此时所有单元的入样概率都一样。

由总体均值、总体总量估计量的计算公式可知：比例分配与 SRS 对于总体均值、总体总量估计的期望是一样的，即估计量的期望是一样的。但是两种方式对于总体均值、总体总量估计的方差不一样。在 n 相同的情况下，$Var(\tilde{\mu}_{str})$, $Var(\tilde{\tau}_{str})$ 通常比 $Var(\hat{\mu}_{srs})$, $Var(\hat{\tau}_{str})$ 小。

证明. 从方差分析的角度去分析。因为两种估计方式中总体总量的估计都是总体均值估计的 N 倍，所以只需证明总体均值的情况即可。

回头补方差分析

由 $\frac{n_h}{N_h}=\frac{n}{N}$ 可得：

$$Var(\tilde{\tau}_{str})=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}=\sum_{h=1}^{H}N_h\left(1-\frac{n}{N}\right)\frac{n}{N}\sigma_h^2=\left(1-\frac{n}{N}\right)\frac{1}{N}\sum_{h=1}^{H}N_h\sigma_h^2$$

从理论角度看待 SSe，即计算总体而非样本的 SSe，可得到：

$$SSe=\sum_{h=1}^{H}\sum_{j=1}^{N_h}(Y_{hj}-\mu_h)^2=\sum_{h=1}^{H}(N_h-1)\sigma_h^2$$

所以：

$$Var(\tilde{\tau}_{str})=\left(1-\frac{n}{N}\right)\frac{1}{N}\sum_{h=1}^{H}N_h\sigma_h^2=\left(1-\frac{n}{N}\right)\frac{1}{N}\left(SSe+\sum_{h=1}^{H}\sigma_h^2\right)$$

而：

$$\begin{aligned}
Var(\hat{\tau}_{srs}) &= N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}\\
&=\frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{SST}{N-1}\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)(SSA+SSe)\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)SSA+\left(1-\frac{n}{N}\right)\frac{N}{n}\left(\frac{N}{N-1}SSe\right)\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)SSA+\left(1-\frac{n}{N}\right)\frac{N}{n}\left(SSe+\frac{1}{N-1}SSe\right)\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)SSA+\left(1-\frac{n}{N}\right)\frac{N}{n}\left(SSe+\sum_{h=1}^{H}\frac{N_h-1}{N-1}\sigma_h^2\right)\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)SSA+\left(1-\frac{n}{N}\right)\frac{N}{n}\left[SSe+\sum_{h=1}^{H}\left(\frac{N-1}{N-1}-\frac{N-N_h}{N-1}\right)\sigma_h^2\right]\\
&=\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)SSA+\left(1-\frac{n}{N}\right)\frac{N}{n}\left(SSe+\sum_{h=1}^{H}\sigma_h^2\right)+\left(1-\frac{n}{N}\right)\frac{N}{n}\left[\sum_{h=1}^{H}\left(-\frac{N-N_h}{N-1}\right)\sigma_h^2\right]\\
&=Var(\tilde{\tau}_{str})+\frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)\left[SSA-\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)\sigma_h^2\right]
\end{aligned}$$

由上式可以看出，如果比例分配估计量的方差比 SRS 估计量的方差大，则需要：

$$SSA < \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \sigma_h^2$$

而这种情况在实践中几乎见不到。 □

从以上推导中也可以看出，组间差异越大，即 SSA 越大，比例分配估计量的方差比 SRS 估计量的方差小得越多，也即精确得更多。而如果每一个层中的方差很大，即 σ_h^2 很大，有可能会使比例分配估计量的方差大于 SRS 估计量的方差，所以在选择分层的时候，要使层内差异小。综上，层间差异大、层内差异小时，比例分配下的分层随机抽样比 SRS 效果更好。

5.3.2 最优分配

在考虑分配方式的时候有如下三点主要因素：

1. 每层中的个体总数 N_h。
2. 层内差异 σ_h^2。
3. 在每个层内抽样的平均成本 c_h。

显然，比例分配没有考虑第二点和第三点。

成本一致最小化方差

当不同层之间抽样成本一致的时候，可以最小化估计量方差。可以将问题转化为：

$$\min f(\vec{n}) = Var(\hat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$
$$s.t. \quad \sum_{h=1}^{H} n_h = n$$

可得最佳分配 (optimal allocation)方案为：

$$n_k = \frac{nN_k\sigma_k}{\sum_{h=1}^{H} N_h\sigma_h}, \quad k = 1, 2, \ldots, H$$

证明. 使用 Lagrange 乘子法求解。引入 Lagrange 乘子 λ 即有：

$$f(\vec{n}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$
$$g(\vec{n}) = \sum_{h=1}^{H} n_h - n$$
$$h(\vec{n}) = f(\vec{n}) + \lambda g(\vec{n})$$

对 f 求偏导可得：

$$\begin{aligned}\frac{\partial f}{\partial n_h} &= -\frac{N_h^2\sigma_h^2}{N_h n_h} - \left(1-\frac{n_h}{N_h}\right)\frac{N_k^2\sigma_h^2}{n_k^2} \\ &= -\frac{N_h\sigma_h^2}{n_h} - \frac{N_h^2\sigma_h^2}{n_h^2} + \frac{N_h\sigma_h^2}{n_h} \\ &= -\frac{N_h^2\sigma_h^2}{n_h^2}\end{aligned}$$

所以：

$$\nabla f(\overrightarrow{n}) = (-\frac{N_1^2\sigma_1^2}{n_1^2}, -\frac{N_2^2\sigma_2^2}{n_2^2}, \ldots, -\frac{N_H^2\sigma_H^2}{n_H^2})$$
$$\nabla g(\overrightarrow{n}) = (1, 1, \ldots, 1)$$

当 f 取最小值时有：

$$\frac{\partial h(\overrightarrow{n})}{\partial \overrightarrow{n}} = (-\frac{N_1^2\sigma_1^2}{n_1^2} + \lambda, -\frac{N_2^2\sigma_2^2}{n_2^2} + \lambda, \ldots, -\frac{N_H^2\sigma_H^2}{n_H^2} + \lambda) = (0, 0, \ldots, 0)$$
$$\sum_{h=1}^{H} n_h = n$$

解得：

$$n_h = \frac{N_h\sigma_h}{\sqrt{\lambda}}, \quad h = 1, 2, \ldots, H$$

此时：

$$n = \sum_{h=1}^{H} n_h = \frac{\sum\limits_{h=1}^{H} N_h\sigma_h}{\sqrt{\lambda}}$$

于是：

$$\sqrt{\lambda} = \frac{\sum\limits_{h=1}^{H} N_h\sigma_h}{n}$$

所以：

$$n_k = \frac{nN_k\sigma_k}{\sum\limits_{h=1}^{H} N_h\sigma_h}, \quad k = 1, 2, \ldots, H$$ □

成本不一致最小化方差

如果不同层之间抽样成本不一致，且总抽样成本为：

$$c = c_0 + \sum_{h=1}^{H} c_h n_h$$

可以将问题转化为：

$$\min f(\overrightarrow{n}) = Var(\hat{\tau}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$
$$s.t. \quad c - c_0 - \sum_{h=1}^{H} c_h n_h = 0$$

则最佳分配方案为：

$$n_k = \frac{(c - c_0) N_k \sigma_k}{\sum_{h=1}^{H} N_h \sigma_h \sqrt{c_h}} \frac{1}{\sqrt{c_k}}, \quad k = 1, 2, \ldots, H$$

从上式中可看出，需要在个数多或方差大得层中分配更多的个体（方差大需要更多的个体来获得有代表性的样本），在成本高的层中分配较少的个体。

固定方差最小化成本

如果不同层之间抽样成本不一致，且总抽样成本为：

$$c = c_0 + \sum_{h=1}^{H} c_h n_h$$

此时如果固定方差最小化成本，则最佳分配方案为：

$$n_k = n \frac{\frac{N_k \sigma_k}{\sqrt{c_k}}}{\sum_{h=1}^{H} \frac{N_h \sigma_h}{\sqrt{c_h}}}, \quad k = 1, 2, \ldots, H$$

若计算出来 $n_k > N_k$，则令 $n_k = N_k$。

5.4 后分层

当使用了 SRS 后发现获取的样本比较极端时（比如研究人群体重，获得的样本中 90% 都是男性），此时使用后分层 (poststratification)，即先进行 SRS，然后将获得的样本分层。

5.4.1 参数估计

总体均值 μ 的估计

$$\hat{\mu}_{poststr} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

$$Var(\hat{\mu}_{poststr}) = E[Var(\hat{\mu}_{str}|\overrightarrow{n})] \approx \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 + \left(\frac{N-n}{N-1}\right) \frac{1}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \sigma_h^2$$

下证明：(1) 后分层估计量 $\hat{\mu}_{poststr}$ 是无偏估计；(2) 如上后分层方差公式正确。

证明. (1) 后分层估计量的计算公式本质和分层随机抽样一样，因此是无偏的。

(2) 由方差分解公式：

$$Var(\hat{\mu}_{poststr}) = Var(E[\hat{\mu}_{str}|\overrightarrow{n}]) + E[Var(\hat{\mu}_{str}|\overrightarrow{n})]$$

而：

$$\begin{aligned} E[\hat{\mu}_{str}|\overrightarrow{n}] &= E\left(\sum_{h=1}^{H}\sum_{j=1}^{N_h}\frac{Y_{hj}Z_{hj}}{n}\right) \\ &= \sum_{h=1}^{H}\sum_{j=1}^{N_h}\frac{Y_{hj}}{n}E(Z_{hj}) \\ &= \sum_{h=1}^{H}\sum_{j=1}^{N_h}\frac{Y_{hj}n_h}{nN_h} \\ &= \sum_{h=1}^{H}\frac{n_h}{n}\sum_{j=1}^{N_h}\frac{Y_{hj}}{N_h} \\ &= \sum_{h=1}^{H}\frac{n_h}{n}\mu_h \end{aligned}$$

可以看出上式是一个定值，所以：

$$Var(E[\hat{\mu}_{str}|\overrightarrow{n}]) = 0$$

于是：

$$\begin{aligned} Var(\hat{\mu}_{poststr}) &= E[Var(\hat{\mu}_{str}|\overrightarrow{n})] \\ &= E\left[\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}\right] \\ &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\sigma_h^2\left[E\left(\frac{1}{n_h}\right)-\frac{1}{N_h}\right] \end{aligned}$$

由定理 0.2，利用泰勒展开，令 $E(n_h)=\mu$，即有：

$$\begin{aligned} E\left(\frac{1}{n_h}\right) &\approx E\left[\frac{1}{\mu}-\frac{1}{\mu^2}(n_h-\mu)+\frac{2}{\mu^3}\frac{(n_h-\mu)^2}{2!}\right] \\ &= \frac{1}{\mu}+\frac{1}{\mu^3}Var(n_h) \end{aligned}$$

由后分层原理，显然可以得到：

$$n_h \sim \text{Hyper}(n, N_h, N)$$

$$E(n_h)=\frac{nN_h}{N},\ Var(n_h)=\frac{nN_h(N-n)(N-N_h)}{N^2(N-1)}$$

于是：

$$E\left(\frac{1}{n_h}\right)=\frac{N}{nN_h}+\left(\frac{N}{nN_h}\right)^2\left(1-\frac{N_h}{N}\right)\frac{N-n}{N-1}$$

将上式代入 $Var(\hat{\mu}_{poststr})$ 即可得到：

$$Var(\hat{\mu}_{poststr})\approx\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2+\left(\frac{N-n}{N-1}\right)\frac{1}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)\sigma_h^2 \qquad \square$$

5.5 样本容量

忽略 FPC

当 $\frac{n_h}{N_h},\ h=1,2,\ldots,H$ 小的时候，忽略层内的 FPC，令：

$$Var(\hat{\mu}_{str})=\frac{1}{n}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{n}{n_h}\sigma_h^2=\frac{v}{n}$$

则 MOE 为 $u\sqrt{\frac{v}{n}}$，可得：

$$n=\frac{u^2v}{d^2}$$

不忽略 FPC

如果 $\frac{n_h}{N_h},\ h=1,2,\ldots,H$ 较大，则考虑层内的 FPC 或使用 Monte Carlo 算法。

Chapter 6

整群抽样

整群抽样 (cluster sampling)将所有个体划分为 N 个群 (cluster)，称群为初级抽样单元 (primary sampling units, psus)或抽样单位，然后通过 SRS 对群进行抽样，在抽出的每个群中使用独立的抽样方法得到最终的样本，称个体为二级抽样单元 (secondary sampling units, ssus)或观察单位。在整群抽样中，ssu 只有在它属于的 psu 被选中时才会有可能被包含在样本中。

整群抽样与 SRS、分层抽样的比较

1. 整群抽样中抽出的样本不如 SRS 得到的样本有代表性。因为群的划分往往是依据于地理信息等进行的，群内的差异往往较小，群间的差异可能较大。此时因为仅从抽中的群中抽样而未在别的群中抽样，样本就可能缺少代表性，即 SRS 样本单元比整群抽样样本单元提供的信息更多。

2. 整群抽样不同于分层抽样的地方是它并不选取辅助变量，虽然层的概念类似于群的概念，但群不由辅助变量产生，也就导致了样本代表性较低的结果。分层随机抽样中，$Var(\hat{\mu}_Y)$ 在群内差异小群间差异大的情况下较小，整群抽样中，$Var(\hat{\mu}_Y)$ 在群内差异大群间差异小的情况下较小（因为此时样本单元提供的信息更多），在实践中只能通过扩大群的个数来提高精确度。

3. 整群抽样会使抽样更加便捷，单位价格上信息更多。

整群抽样的分类

1. 一阶整群抽样 (one-stage cluster sampling)：一旦某个 psu 被选中，该 psu 中的 ssu 全部被选中。

2. 二阶整群抽样 (two-stage cluster sampling)：某个 psu 被选中后，还需对其中的所有 ssu 进行一次抽样，若此次抽中则包含在样本中。

什么情况下使用整群抽样

1. 当构建包含所有个体的抽样框架十分困难或根本做不到（此时就无法做到直接对个体进行抽样），但若把所有个体分为若干个群，构建包含所有群的抽样框架并对群进行抽样并不困难时，适合使用整群抽样。

2. 当目标群体在个体角度来讲分布广泛，调查成本太高，但可以将样本分群使一个群内的样本分布集中，从而可以降低抽样成本时，适合使用整群抽样。

6.1 一阶整群抽样

由一阶整群抽样的定义，选中某个 psu 后该 psu 中的所有 ssu 进入样本。

符号说明

符号	说明
N	总群数
n	抽样群数
M_i	第 i 个 psu 中的 ssu 个数
$M=\sum_{i=1}^{N} M_i$	ssu 的总数
y_{ij}	第 i 个 psu 中第 j 个样本单元的数值
$\mu_i=\sum_{j=1}^{M_i} \frac{y_{ij}}{M_i}$	第 i 个 psu 中的均值
$\mu=\sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{y_{ij}}{M}$	总体均值
$\tau_i=\sum_{j=1}^{M_i} y_{ij}$	第 i 个 psu 中的总量
$t_i=\sum_{j=1}^{M_i} y_{ij}$	入样的第 i 个 psu 中的总量
$\tau=\sum_{i=1}^{N} \tau_i$	总体总量
$\sigma_i^2=\sum_{j=1}^{M_i} \frac{(y_{ij}-\mu_{Y_i})^2}{M_i-1}$	第 i 个 psu 中的方差
$\sigma_{psu}^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i-\frac{\tau}{N}\right)^2$	psu 间的方差
$\sigma_M^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(M_i-\frac{M}{N}\right)^2$	psu 间 ssu 个数的方差
$\sigma^2=\sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{(y_{ij}-\mu)^2}{M-1}$	总体方差
$R=\frac{\sum_{i=1}^{N}(M_i-\frac{M}{N})(\tau_i-\frac{\tau}{N})}{(N-1)\sigma_M\sigma_{psu}}$	M_i 与 τ_i 的回归系数

表 6.1: 符号说明表

一阶整群抽样的相关参数有两种估计方式，分别为无偏估计与比例估计。虽然无偏估计具有无偏性，但经过模拟研究，当群内总体总量与群内个体数量成正比时，比例估计的方差会比无偏估计小很多，此时应选择比例估计量对参数进行估计。

6.1.1 参数的无偏估计

总体总量的估计

可给出如下关于总体总量的估计：

$$\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} t_i = \sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij}y_{ij}$$

$$Var(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{\sigma_{psu}^2}{n} = N(N-n)\frac{\sigma_{psu}^2}{n}$$

$$\widehat{Var}(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{s_{psu}^2}{n} = N(N-n)\frac{s_{psu}^2}{n}$$

$$s_{psu}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(t_i - \frac{\hat{\tau}}{N}\right)^2$$

总体均值的估计

可给出如下关于总体均值的估计：

$$\hat{\mu} = \frac{\hat{\tau}}{M}$$

$$Var(\hat{\mu}) = Var\left(\frac{\hat{\tau}}{M^2}\right) = \frac{N^2}{M^2}\left(1-\frac{n}{N}\right)\frac{\sigma_{psu}^2}{n}$$

$$\widehat{Var}(\hat{\mu}) = \widehat{Var}\left(\frac{\hat{\tau}}{M^2}\right) = \frac{N^2}{M^2}\left(1-\frac{n}{N}\right)\frac{s_{psu}^2}{n}$$

6.1.2 参数的比例估计

由：

$$\mu = \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \frac{\sum_{i=1}^{N}\tau_i}{\sum_{i=1}^{N} M_i} = \frac{\tau}{M}$$

可设：

$$\mu = \frac{\tau}{M} = B$$

将 τ_i 看作样本单元值，将 M_i 看作辅助变量，可得到如下关于参数的比例估计：

$$\hat{B} = \hat{\mu}_r = \frac{\hat{\tau}}{\hat{M}} = \frac{\frac{N}{n}\sum_{i=1}^{n} t_i}{\frac{N}{n}\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} M_i}$$

$$Var(\hat{B}) = \frac{1}{M^2} Var(\hat{\tau}_r) \approx \left(1 - \frac{n}{N}\right) \frac{\sigma_{psu}^2 - 2BR\sigma_M\sigma_{psu} + B^2\sigma_M^2}{n(\frac{M}{N})^2}$$

$$\widehat{Var}(\hat{B}) \approx \left(1 - \frac{n}{N}\right) \frac{s_{psu}^2 - 2\hat{B}\hat{R}s_{psu}s_M + \hat{B}^2 s_M^2}{n(\frac{\sum_{i=1}^{n} M_i}{n})^2}$$

$$\widehat{Var}_1(\hat{\mu}_r) = \frac{1}{M^2} N\,(N-n)\,\frac{s_e^2}{n}$$

$$\widehat{Var}_1(\hat{\tau}_r) = N\,(N-n)\,\frac{s_e^2}{n}$$

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(t_i - \hat{B}M_i\right)^2$$

6.2 二阶抽样

整群抽样抽到群则群中所有单元进入样本，二阶抽样需要继续对每个群进行第二次抽样，抽中则进入样本，即我们只在选中的 psu 中选择一部分 ssu。

符号说明

符号	说明
N	总群数
n	抽样群数
J	每个群中分层的层数
M_{ij}	抽出的第 i 个群第 j 个层个体的总数
m_{ij}	从抽出的第 i 个群第 j 个层抽出来的样本单元总数
M_i	抽出的第 i 个群个体的总数
M	个体总数
y_{ijk}	第 i 个群第 j 个层的第 k 个个体是否阳性
$\bar{y}_{ij}$	第 i 个群第 j 个层样本的均值
μ_{ij}	第 i 个群第 j 个层的均值
τ_i	第 i 个群的总体总数
t_i	入样的第 i 个群的总体总数
τ	总体总量
σ_{ij}^2	第 i 个群第 j 个层的方差
p_{ij}	第 i 个群第 j 个层的流行率
p_i	第 i 个群的流行率
p	总流行率
Z_i	表示第 i 个群是否被抽中的示性变量

表 6.2: 符号说明表

6.2.1 总量的估计

可得到如下关于总量的估计：

$$\hat{t}_i = \sum_{j=1}^{J} \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk}$$

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} \hat{t}_i = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk}$$

$$E(\hat{t}_i) = t_i,\ E(\hat{\tau}) = \tau$$

$$Var(\hat{\tau}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i - \frac{\tau}{N}\right)^2 + \frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}}\left(1-\frac{m_{ij}}{M_{ij}}\right)\left(\frac{1}{M_{ij}-1}\right)\sum_{k=1}^{M_{ij}}(y_{ijk}-\mu_{ij})^2$$

$$\widehat{Var}(\hat{\tau}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{t}_i - \frac{\hat{\tau}}{N}\right)^2 + \frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}}\left(1-\frac{m_{ij}}{M_{ij}}\right)\left(\frac{1}{m_{ij}-1}\right)\sum_{k=1}^{m_{ij}}(y_{ijk}-\bar{y}_{ij})^2$$

无偏性的证明

由分层随机抽样总体总量估计量的无偏性，可以得到此时 $\hat{\tau_i}$ 的无偏性，进而可以证明 $\hat{\tau}$ 是无偏的：

$$E(\hat{\tau}) = E[E(\hat{\tau}|\overrightarrow{Z})] = E\left[E\left(\frac{N}{n}\sum_{i=1}^{n}\hat{t_i}|\overrightarrow{Z}\right)\right] = E\left[E\left(\frac{N}{n}\sum_{i=1}^{N}Z_i\hat{\tau_i}\right)\right]$$
$$= E\left[\sum_{i=1}^{N}\frac{N}{n}Z_iE(\hat{\tau_i})\right] = E\left[\sum_{i=1}^{N}\frac{N}{n}Z_i\tau_i\right] = \sum_{i=1}^{N}\frac{N}{n}\tau_iE(Z_i) = \sum_{i=1}^{N}\frac{N}{n}\tau_i\frac{n}{N} = \sum_{i=1}^{N}\tau_i = \tau$$

方差公式的证明

由方差的分解，可以得到：

$$Var(\hat{\tau}) = Var[E(\hat{\tau}|\overrightarrow{Z})] + E[Var(\hat{\tau}|\overrightarrow{Z})]$$

由 SRS 的结论，可以得到：

$$Var[E(\hat{\tau}|\overrightarrow{Z})] = Var\left[E\left(\sum_{i=1}^{N}\frac{N}{n}Z_i\hat{\tau_i}|\overrightarrow{Z}\right)\right] = Var\left(\sum_{i=1}^{N}\frac{N}{n}Z_i\tau_i\right) = N^2\left(1-\frac{n}{N}\right)\frac{\sigma_\tau^2}{n}$$
$$\sigma_\tau^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i - \frac{\tau}{N}\right)^2$$

由方差的分解，可以得到：

$$E[Var(\hat{\tau}|\overrightarrow{Z})] = E[E[\hat{\tau}|\overrightarrow{Z}] - E[\hat{\tau}|\overrightarrow{Z}]^2]$$

而：

$$E[\hat{\tau}|\overrightarrow{Z}] - E[\hat{\tau}|\overrightarrow{Z}]^2 = E\left[\left(\sum_{i=1}^{N}\frac{N}{n}Z_i\hat{\tau_i}\right)^2\right] - \left(\sum_{i=1}^{N}\frac{N}{n}Z_i\tau_i\right)^2$$
$$= E\left(\sum_{i=1}^{N}\frac{N^2}{n^2}Z_i^2\hat{\tau_i}^2 + \sum_{i=1}^{N}\sum_{j\neq i}\frac{N^2}{n^2}Z_iZ_j\hat{\tau_i}\hat{\tau_j}\right)$$
$$- \left(\sum_{i=1}^{N}\frac{N^2}{n^2}Z_i^2\tau_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}\frac{N^2}{n^2}Z_iZ_j\tau_i\tau_j\right)$$
$$= \frac{N^2}{n^2}\left(\sum_{i=1}^{N}Z_i^2E(\hat{\tau_i}^2) + \sum_{i=1}^{N}\sum_{j\neq i}Z_iZ_jE(\hat{\tau_i}\hat{\tau_j}) - \sum_{i=1}^{N}Z_i^2\tau_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}Z_iZ_j\tau_i\tau_j\right)$$
$$= \frac{N^2}{n^2}\left[\sum_{i=1}^{N}Z_i^2E^2(\hat{\tau_i}) + \sum_{i=1}^{N}Z_i^2Var(\hat{\tau_i}) + \sum_{i=1}^{N}\sum_{j\neq i}Z_iZ_jE(\hat{\tau_i})E(\hat{\tau_j})\right.$$
$$\left. - \sum_{i=1}^{N}Z_i^2\tau_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}Z_iZ_j\tau_i\tau_j\right]$$
$$= \frac{N^2}{n^2}\sum_{i=1}^{N}Z_i^2Var(\hat{\tau_i})$$

由 Z_i 的性质可得：

$$E[Var(\hat{\tau}|\overrightarrow{Z})] = E\left(\frac{N^2}{n^2}\sum_{i=1}^{N} Z_i^2 Var(\hat{\tau}_i)\right) = \frac{N^2}{n^2}\sum_{i=1}^{N} E(Z_i^2)Var(\hat{\tau}_i) = \frac{N}{n}\sum_{i=1}^{N} Var(\hat{\tau}_i)$$

由分层随机抽样层内方差的公式：

$$E[Var(\hat{\tau}|\overrightarrow{Z})] = \frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}}\left(1-\frac{m_{ij}}{M_{ij}}\right)\sigma_{ij}^2$$

$$\sigma_{ij}^2 = \left(\frac{1}{M_{ij}-1}\right)\sum_{k=1}^{M_{ij}}(y_{ijk}-\mu_{ij})^2$$

所以：

$$Var(\hat{\tau}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i-\frac{\tau}{N}\right)^2+\frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}}\left(1-\frac{m_{ij}}{M_{ij}}\right)\left(\frac{1}{M_{ij}-1}\right)\sum_{k=1}^{M_{ij}}(y_{ijk}-\mu_{ij})^2$$

6.2.2 流行率问题

关于流行率问题，有如下结论：

$$\hat{p}_i = \frac{\hat{t}_i}{M_i}$$

$$\hat{p} = \frac{\hat{\tau}}{M} = \frac{N}{nM}\sum_{i=1}^{n}\sum_{j=1}^{J}\frac{M_{ij}}{m_{ij}}\sum_{k=1}^{m_{ij}} y_{ijk}$$

$$Var(\hat{\tau}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i-\frac{\tau}{N}\right)^2+\frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}}\left(1-\frac{m_{ij}}{M_{ij}}\right)\frac{M_{ij}p_{ij}(1-p_{ij})}{M_{ij}-1}$$

$$\widehat{Var}(\hat{\tau}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{t}_i-\frac{\hat{\tau}}{N}\right)^2+\frac{N}{n}\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{M_{ij}^2}{m_{ij}-1}\left(1-\frac{m_{ij}}{M_{ij}}\right)\hat{p}_{ij}(1-\hat{p}_{ij})$$

由 $\hat{\tau}_i$ 和 $\hat{\tau}$ 的无偏性，显然 $\hat{p}_i$ 和 $\hat{p}$ 也是无偏估计。由一般情况下 $Var(\hat{\tau})$ 的计算公式，也容易得到流行率问题下 $Var(\hat{\tau})$ 的计算公式。

6.3 中英术语表

符号	英文全称	中文含义	首次出现页码
FPC	finite population correction fraction	有限群体校正分数	12
MOE	margin of error	误差幅度	17
psus	primary sampling units	初级抽样单元	45
SRS	simple random sampling	不放回型简单随机抽样	11
SRSWR	simple random sampling with replacement	放回型简单随机抽样	11
ssus	secondary sampling units	二级抽样单元	45