不等概抽样

课程： 数据采集方法
小组： B+X9bo 组
姓名： 蒋贵豪
时间： 2021 年 11 月

统计与数据科学学院

目录

不等概抽样

2021 年 11 月 8 日

0　前言

前面讨论的抽样方法多是等概率抽样，即每个总体单元都具有相同的入样概率。等概率抽样方法容易设计和解释，但有时不如不等概抽样有效率。尤其在抽样单元规模差异很大时，经常采用不等概抽样，即每个单元入样的概率不相等。本章介绍不等概抽样。

1　不等概抽样

1.1　不等概抽样的概念

在本章之前的方法都是等概率抽样，例如简单随机抽样、分层抽样和整群抽样。等概率抽样每个总体单元都有相同的入样概率。也就是说，对于总体中的每个单元，我们把他们的地位都看作是相同的。

但是，现实情况中，我们往往会遇到总体单元之间差异非常大的情况。例如，我们需要研究费城的养老院中人士对维持生命所治疗的偏好。我们的目标总体是在费城各养老院正式登记的人士。现在我们知道费城有 294 家养老院，37652 个床位。在抽样前，我们已知了床位数，不知道居住人士的数量。如果我们使用两阶段的整群抽样，首先是在所有养老院中做简单随机抽样，抽取部分养老院。然后再在抽中的养老院中做简单随机抽样，抽取部分人士。这样的话会导致一些问题，如大的养老院和小的养老院被抽到的概率一样大。如果对维持生命所治疗方式偏好与养老院床位成正比，无偏估计量可能就会有较大的方差。还有去不同养老院的调查员的工作量也有可能有较大差异，这是我们不希望看到的。

于是，当总体单元差异较大的时候，我们一般会放弃简单的等概抽样。一种方法是将总体的单元按大小分层，较大的单元层抽样比高，较小的单元层抽样比低。或者是采用不等概抽样来减少抽样的方差，而不是采用清晰的分层。也就是说，给每个单元与它的规模适配的入样概率。这种抽样方法下的总体总量方差很可能小于等概率整群抽样下的估计量方差。

1.2 不等概抽样适用情况

不等概抽样适用的情况如下：

(1) 抽样单元在总体中所占的地位不一致。

(2) 调查的总体单元与抽样总体的单元不一致。

(3) 需要改善估计量。

除了上述情况，不等概抽样还广泛用于整群抽样、多阶段抽样中初级单元规模相差较大的情况。不等概抽样的优点是提高估计的精度。但是现实的使用中，必须要有说明每个单元规模大小的辅助变量来确定每个单元的入样概率。不等概抽样编制抽样框的过程相对于简单随机抽样更为复杂。

1.3 不等概抽样的分类

Brewer 和 Hanif 在其著作《不等概抽样》中介绍了 50 多种不等概抽样方法，但是我们常用的只有大约 10 种。不等概抽样可以按多种原则分类，如按样本单元是否放回，可以分为放回不等概抽样和不放回不等概抽样。

1.3.1 放回不等概抽样

每次在总体中对每个单元按入样概率进行抽样，抽取出来的样本单元放回总体，然后再进行下一次抽样，这种不等概抽样方式为放回不等概抽样。此时，每次的抽样过程是从同一个总体中独立进行的。由于抽样是放回的，某个单元在样本中可能出现多次。在这种情况时，对于单元的调查只进行一次，但是在计算时需要按抽中几次算几次的原则。

放回不等概抽样中，最常用的时按照总体单元的规模大小来确定单元每次入样的概率。假设总体中第 i 个单元的大小为 M_i, 总体的总规模为 $M_0 = \sum_{i=1}^{N} M_i$, 每次抽样中的第 i 个单元被抽中的概率用 Z_i 表示，如果：

$$Z_i = \frac{M_i}{M_0} = \frac{M_i}{\sum_{i=1}^{N} M_i} \tag{1.1}$$

那么这种不等该抽样称作放回的与规模大小成比例的概率抽样 (probability proportional to size), 简称 PPS 抽样。

PPS 抽样的实施主要有两种方法：代码法和拉希里法：

(1) 代码法：在 PPS 抽样中，赋予每个单元与 M_i 相等的代码数，将代码累加得到 M_0，每次抽样都产生一个 $[1, M_0]$ 之间的随机数，设为 m, 则代码对应的单元被抽中。如此进行 n 次，就构成了 PPS 抽样的样本。如果 M_i 不是整数，那么乘以某一个倍数，使得 $M_0 Z_i$ 为整数，每个单元赋予与 $M_0 Z_i$ 相等的代码数后再进行代码法抽样。

(2) 拉希里法：令 $M^* = \max\limits_{1\le i\le N}\{M_i\}$. 每次抽样分别产生一个 $[1, N]$ 之间的随机数 i 和 $[1, M^*]$ 之间的随机数 m. 如果 $M_i \ge m$，则第 i 个样本被抽中；否则，重抽一组 (i, m). 重复上述操作，直到抽满 n 个样本。

1.3.2 不放回不等概抽样

对于放回抽样，其对总体参数估计及方差的估计较为简单，但是样本单元可能被抽中很多次，抽样中得到了重复的信息。于是，我们可以采用不放回的不等概抽样，也就是每次在总体中对每个单元按照入样概率进行抽样，抽取出来的样本不再放回总体。虽然不放回抽样比放回抽样的效率要高，但是样本的不独立性加大了抽样实施及参数估计和精度计算的难度。

对于不放回的不等概抽样，样本的抽取有如下几种方法：

(1) 逐个抽取法。每次从总体未被抽中的单元中以一定概率抽取一个单元样本。

(2) 重抽法。以一定的概率逐个进行放回抽样，如果抽到重复单元，则放弃所有抽到的单元。直至抽到规定的样本量且所有样本单元不重复为止。

(3) 全样本抽取法。对总体每个单元分别按一定的概率决定其是否入样。

(4) 系统抽样法。将总体单元按某种方式进行排序，将规定的入样概率汇总，根据样本量确定抽样间距 k, 在 $1 \sim k$ 产生一个随机数，并确定相应的初始单元，以后在总体中每隔 k 个单元抽出一个作为样本单元。

2　一阶段放回不等概抽样

在每一次抽样时，抽到每个单元的概率均相等，记抽到第 i 个单元的概率为 φ_i. 我们以调查某学校学生一周用于学习初等统计学的时长为例，如下表 2.1 所示：

表 2.1: 某学校学生一周用于学习初等统计学的时长

Class Number	M_i	φ_i
1	44	0.068006
2	33	0.051005
3	26	0.040185
4	22	0.034003
5	76	0.117465
6	63	0.097372
7	20	0.030912
8	44	0.068006
9	54	0.083462
10	34	0.052550
11	46	0.071097
12	44	0.037094
13	46	0.071097
14	100	0.154560
15	15	0.023184
Total	647	1

说明：

1. 本次调查共 15 个班级，每班人数为 M_i, 总人数为 647。

2. 按与班级人数成正比的概率抽取班级，对抽中的班级内的全体同学做调查。

我们可以用以下两种方法实施调查：

2.1　一阶段放回不等概抽样的方法

2.1.1　累计规模法

1. 计算每个班级累计规模范围。

2. 在 [1,647] 间生成 n 个随机整数。

3. 若某随机数落入一个班级的累积规模区间，则此班级为抽中样本单元。

表 2.2 给了累计规模区间：

表 2.2: 初等统计学班级累计规模区间

Class Number	M_i	φ_i	Cumulative M_i Range
1	44	0.068006	1 - 44
2	33	0.051005	45 - 77
3	26	0.040185	78 - 103
4	22	0.034003	104 - 125
5	76	0.117465	126 - 201
6	63	0.097372	202 - 264
7	20	0.030912	265 - 284
8	44	0.068006	285 - 328
9	54	0.083462	329 - 382
10	34	0.052550	383 - 416
11	46	0.071097	417 - 462
12	44	0.037094	463 - 486
13	46	0.071097	487 - 532
14	100	0.154560	533 - 632
15	15	0.023184	633 - 647
Total	647	1	

若产生 5 个随机数 $\{487, 369, 221, 326, 282\}$，则由表 2.2 可知，抽中的 5 个班级号为 $\{13, 9, 6, 8, 7\}$.

2.1.2 Lahiri 方法

步骤:

1. 抽取随机数 $i \in \{1, \cdots, N\}$。

2. 抽取随机数 $M \in \{1, \cdots, Max\{M_i\}\}$。如果 $M \leq M_i$，则抽中第 i 个单元进入样本。否则，回到第一步。

3. 重复上述过程，直到抽取 5 个班级样本。

表 2.3: Lahiri 方法过程

First Random Number(psu i)	Second Random Number	M_i	Action
12	6	24	$6 < 24$; include psu 12 in sample
14	24	100	Include sample
1	65	44	$65 > 44$; discard pair of numbers and try again
7	84	20	$84 > 20$;try again
10	49	34	Try again
14	47	100	Include
15	43	15	Try again
5	24	76	Include
11	87	46	Try again
1	36	55	Include

由表 2.3 可知，抽中的班级序号为：$\{12, 14, 14, 5, 1\}$。

2.2 一阶段放回不等概抽样的估计

我们用总体总量 Hansen-Hurwitz 估计，其表达式如下：

1. 总量无偏估计：

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\varphi_i} = \frac{M_0}{n}\sum_{i=1}^{n}\frac{y_i}{M_i} \tag{2.1}$$

这里 M_i 表示样本中第 i 个 PSU 的规模；

2. 方差无偏估计：

$$\mathrm{Var}(\hat{Y}_{HH}) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{\varphi_i} - \hat{Y}_{HH}\right)^2 = \frac{M_0^2}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{M_i} - \frac{\hat{Y}_{HH}}{M_0}\right)^2 \tag{2.2}$$

接上表 2.3，抽中班级为：$\{12, 14, 14, 5, 1\}$，绘制下表 2.4，其中 γ_i 是第 i 个班的所有同学上个周用于学习初等统计学的用时。

表 2.4: 不等概抽样样本统计

Class	φ_i	γ_i	$\frac{\gamma_i}{\varphi_i}$
12	$\frac{24}{647}$	75	2021.875
14	$\frac{100}{647}$	203	1313.410
14	$\frac{100}{647}$	203	1313.410
5	$\frac{76}{647}$	191	1626.013
1	$\frac{44}{647}$	168	2470.364

记该年级所有学生一周用于学习初等统计学的时长的估计为 $\hat{Y}_{HH}$, 代入公式 (2.1) 计算可得：

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\varphi_i} = \frac{(2021.875 + 1313.410 + 1313.410 + 1626.013 + 2470.364)}{5} = 1749.014$$

该估计量的方差为：

$$\begin{aligned}\mathrm{Var}(\widehat{Y}_{HH}) &= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{\varphi_i} - \widehat{Y}_{HH}\right)^2 \\ &= \frac{(2021.875 - 1749.014)^2 + \ldots + (2470.364 - 1749.014)^2}{5 \times 4} = (222.42)^2\end{aligned}$$

记该年级平均每位学生一周用于学习初等统计学的时长的估计为 $\bar{y}$, 则 $\bar{y}$ 的值以及标准差为：

$$M_0 = 647, \bar{y} = \frac{\hat{Y}_{HH}}{M_0} = \frac{1749.014}{647} = 2.7$$

$$\mathrm{SE}(\bar{y}) = \frac{\sqrt{\mathrm{Var}(\hat{Y}_{HH})}}{M_0} = \frac{222.42}{647} = 0.34$$

3 多阶段放回不等概抽样

3.1 两阶段放回不等概抽样

我们知道，在多阶段抽样中如果初级单元规模大小不等时，需要对初级单元进行放回不等概抽样。对初级单元进行抽样时，需事先规定每个初级单元被抽中的概率 Z_i。对被抽中的每个初级单元，再抽取 m_i 个二级单元。如果某个初级单元被抽中多次，则应该对其二级单元抽取多个独立样本。例如，某个初级单元被重复抽中两次，则对其二级单元抽取一个大小为 m_i 的样本后，将这 m_i 个二级单元放回，再重新抽取一个大小为 m_i 的样本。当然，这两个样本中的二级单元可能会有重复，应记录下这些样本的情况。实际调查时，对重复的二级单元只调查一次，但计算的时候，应该按照被抽中的次数进行重复计算。

对总体总值的估计通常是先构造初级单元总值 Y_i 的无偏估计 $\hat{Y}_i$ 然后用汉森-赫维茨估计量对总体总值 Y 进行估计：

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{Z_i} \tag{3.1}$$

由于 $\hat{Y}_i$ 是 Y_i 的无偏估计，$\hat{Y}_{HH}$ 是 Y 的无偏估计，且 $\hat{Y}_{HH}$ 的方差为：

$$\mathrm{Var}\left(\hat{Y}_{HH}\right) = \frac{1}{n}\left[\sum_{i=1}^{N} Z_i\left(\frac{Y_i}{Z_i} - Y\right)^2 + \sum_{i=1}^{N}\frac{\mathrm{Var}_2\left(\hat{Y}_i\right)}{Z_i}\right] \tag{3.2}$$

$\mathrm{Var}(\hat{Y}_{HH})$ 的一个无偏估计为：

$$\mathrm{var}\left(\hat{Y}_{HH}\right) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{Z_i} - \hat{Y}_{HH}\right)^2 \tag{3.3}$$

注意上述对第二阶抽样并没有做出特别的规定，而且估计量的方差估计式与第二阶段抽样的方式无关。

如果希望 $\hat{Y}_{HH}$ 是自加权的，由

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{Z_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{Z_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{Z_i m_i}\sum_{j=1}^{m_i} y_{ij}, \tag{3.4}$$

则要求：

$$\frac{M_i}{nZ_i m_i} = \text{常数}K \equiv \frac{1}{f_0} \tag{3.5}$$

这里，f_0 为总体中任意一个二级单元被抽中的概率。如果 f_0 事先确定，则

$$f_{2i} = \frac{m_i}{M_i} = \frac{f_0}{nZ_i} \tag{3.6}$$

记总体所有的二级单元数为 M_0，如果抽样时每个初级单元被抽中的概率与其拥有的二级单元数成比例，即初级单元被抽中概率为 $Z_i = M_i/M_0$。第二阶段对二级单元进行简单随机抽样，则 $m_i = m$ 时，样本是自加权的。这时，对总体总值的估计为：

$$\hat{Y}_{PPS} = M_0\bar{\bar{y}} = \frac{M_0}{n}\sum_{i=1}^{n}\bar{y}_i = \frac{M_0}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij} \tag{3.7}$$

估计量方差的样本估计为：

$$\text{var}\left(\hat{Y}_{PPS}\right) = \frac{M_0^2}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2 \tag{3.8}$$

实际工作中，如果初级单元大小不相等，人们通常喜欢在第一阶段抽样时按放回的与二级单元数成比例的 PPS 抽样，第二阶段抽样则进行简单随机抽样，且每个初级单元内的二级单元样本量都相同，这样得到的样本是自加权的，估计量的形式非常简单。

3.2 多阶段放回不等概抽样

一般情况下，各级单元的大小不相等。类似对初级单元大小不等的两阶段抽样的讨论，通常这时每一阶段的抽样采用与单元大小成比例的不等概抽样，而且通常抽样是放回的，即 PPS 抽样。

以三阶段抽样为例，记：

总体拥有 N 个初级单元, 每个初级单元拥有 M_i 个二级单元, 每个二级单元又拥有 k_{ij} 个三级单元;

各阶样本量分别为 n, m, k（注意 m, k 不随单元变化），即抽取 n 个初级单元，在每个样本初级单元中，抽取 m 个二级单元，在每个样本二级单元中，抽取 k 个三级单元;

每一阶单元的抽取概率为 Z_i, Z_{ij}, Z_{iju}, 它们满足:

$$\sum_{i=1}^{N}Z_i = 1, \sum_{j=1}^{M_i}Z_{ij} = 1, \sum_{u=1}^{K_{ij}}Z_{iju} = 1 \tag{3.9}$$

. 这时对总体总值 $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{u=1}^{K_{ij}}Y_{iju}$ 的无偏估计为：

$$\hat{Y} = \frac{1}{nmk}\sum_{i=1}^{n}\frac{1}{Z_i}\sum_{j=1}^{m}\frac{1}{Z_{ij}}\sum_{u=1}^{k}\frac{y_{iju}}{Z_{iju}} \tag{3.10}$$

它的方差为：

$$\begin{aligned}\mathrm{Var}(\hat{Y}) =& \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{Z_i} - Y^2\right) + \frac{1}{nm}\sum_{i=1}^{N}\frac{1}{Z_i}\left(\sum_{j=1}^{M_i}\frac{Y_{ij}^2}{Z_{ij}} - Y_i^2\right) \\ &+ \frac{1}{nmk}\sum_{i=1}^{N}\frac{1}{Z_i}\left[\sum_{j=1}^{M_i}\frac{1}{Z_{ij}}\left(\sum_{u=1}^{K_{ij}}\frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2\right)\right]\end{aligned} \tag{3.11}$$

式中：

$$Y_{ij} = \sum_{u=1}^{K_{ij}} Y_{iju}, Y_i = \sum_{j=1}^{M_i}\sum_{u=1}^{K_{ij}} Y_{iju} = \sum_{j=1}^{M_i} Y_{ij} \tag{3.12}$$

$\mathrm{Var}(\hat{Y})$ 的一个无偏估计为：

$$\mathrm{Var}(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{Y}_i - \hat{Y}\right)^2 \tag{3.13}$$

式中：

$$\hat{Y}_i = \frac{1}{Z_i m}\sum_{j=1}^{m}\frac{1}{Z_{ij}}\left(\frac{1}{k}\sum_{u=1}^{k}\frac{y_{iju}}{Z_{iju}}\right) \tag{3.14}$$

实际工作中，通常的做法是前两阶段抽样采用 PPS 抽样，即对初级单元和二级单元的抽样按放回的，与其单元大小成比例的概率抽样；最后一阶段抽样按等概率抽选。如果每一阶段抽样单元的样本量都相同（即 $m_i = m$，$k_j = k$），则样本是自加权的。这时

$$Z_i = \frac{\sum_{j=1}^{M_i} K_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i} K_{ij}} = \frac{\sum_{j=1}^{M_i} K_{ij}}{M_0}, Z_{ij} = \frac{K_{ij}}{\sum_{j=1}^{M_i} K_{ij}}, Z_{iju} = \frac{1}{K_{ij}} \tag{3.15}$$

注意：这时第三阶段抽样也是放回的，各阶段单元的大小是以最小单元数计算的。将其代入前面的式子，则总体总值的估计为：

$$\hat{Y} = \frac{M_0}{nmk}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{u=1}^{k} y_{iju} = M_0\bar{\bar{\bar{y}}} \tag{3.16}$$

$\bar{\bar{\bar{y}}}$ 是以三级单元计算的样本简单平均数。$\hat{y}$ 的表达式正好说明这时估计量是自加权的。$\hat{y}$ 方差的样本估计为：

$$\mathrm{var}(\hat{Y}) = \frac{M_0^2}{n(n-1)}\sum_{i=1}^{n}\left(\bar{\bar{y}}_i - \bar{\bar{\bar{y}}}\right)^2 \tag{3.17}$$

式中：

$$\bar{\bar{y}}_i = \frac{1}{mk}\sum_{j=1}^{m}\sum_{u=1}^{k} y_{iju} \tag{3.18}$$

如果对三级单元的抽样采用不放回简单随机抽样，上述公式仍然成立，只是估计量的理论方差比放回的情形小一些。

类似地，对于更多阶段的情形，除了最后一阶段采用等概率抽样（放回的或不放回的均可），前几阶段均采用 PPS 抽样，并且自第二阶段开始，每一阶段各样本单元的样本量相同（即 $m_i = m$，$k_j = k$,...），则样本是自加权的，其估计量的形式非常简单。

4 不放回不等概抽样

4.1 不放回不等概抽样概念

不放回的不等概抽样就是在可放回抽样方法的基础上，每一轮抽样都筛除已抽出样本，并进行抽样概率的重新计算。执行不放回的不等概抽样是一件困难的事情，方差估计也十分复杂。在不放回抽样的过程中，每一轮抽样都是不独立的，我们需要定义样本的入样概率，以方便后续计算。

在不放回不等概抽样中，每个单元入样的概率 π_i 及任意两个单元同时入样的概率 π_{ij} 统称为包含概率。

针对固定的 n，包含概率满足以下条件：

$$\sum_{i=1}^{N}\pi_i = n \tag{4.1}$$

$$\sum_{i\neq j}^{N}\pi_{ij} = (n-1)\pi_i \tag{4.2}$$

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij} = \frac{1}{2}n(n-1) \tag{4.3}$$

严格的 πPS 抽样定义：如果每个单元入样概率与其大小或规模的度量 M_i 严格成比例，记 $Z_i = \dfrac{M_i}{M_0}$, 则对于固定的 n，有 $\pi_i = nZ_i$.

因为严格的 πPS 抽样中的包含概率 π_{ij} 不易求得，所以其估计量方差的估计也相当困难。只有在 $n = 2$ 时，我们才有一些简单的方法进行严格的 πPS 抽样，对于 $n > 2$ 的情形，严格的 πPS 抽样则相对复杂。实际工作中，我们可以通过对总体分层，在每一层中进行严格的 πPS 抽样。

下例中计算了当 $n = 2$ 时的不放回抽样的包含概率 π_{ij}：

例 4.1 超市销售额

一乡镇有 4 个超市，估计上月总的销售额 (Y_i)，$\psi(i)$ 为抽样概率（即超市面积的比率），总体单元情况如表 4.1：

P(第一次抽样时抽到超市 A)= $\dfrac{1}{16}$

P(第二次抽样时抽中 B| 第一次抽样时抽中 A)= $\dfrac{\frac{2}{16}}{1-\frac{1}{16}}$

P(第二次抽样时抽中 A| 第一次抽样时抽中 B)= $\dfrac{\frac{1}{16}}{1-\frac{2}{16}}$

表 4.1: 超市销售额抽样信息表

Store	Size(m^2)	ψ_i	Y_i (in Thousands)
A	100	$\frac{1}{16}$	11
B	200	$\frac{2}{16}$	20
C	300	$\frac{3}{16}$	24
D	1000	$\frac{10}{16}$	245
Total	1600	1	300

P(第一次抽样时抽中 A, 第二次抽样时抽中 B)≠P(第一次抽样时抽中 B, 第二次抽样时抽中 A)

当采样数量 $n = 2$ 时，P(样本中包含单元 i, k)=$\pi_{ik} = \psi_i \frac{\psi_k}{1-\psi_i} + \psi_k \frac{\psi_i}{1-\psi_k}$

表 4.2: $n = 2$ 时 π_{ik} 的计算结果

		Store k				
		A	B	C	D	π_i
Store i	A	-	0.0173	0.0269	0.1458	0.1900
	B	0.0173	-	0.0556	0.2976	0.3705
	C	0.0269	0.0556	-	0.4567	0.5393
	D	0.1458	0.2976	0.4567	-	0.9002
	π_k	0.1900	0.3705	0.5393	0.9002	2.0000

4.2 不放回不等概抽样的霍维茨-汤普森估计量

在不放回不等概抽样中，对总体总值 Y 的估计采用霍维茨-汤普森估计量：

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \alpha_i \frac{y_i}{\pi_i} \tag{4.4}$$

式中，$\alpha_i = 1$ 表示第 i 个单元入样，$\alpha_i = 0$ 表示第 i 个单元没有入样。$p(\alpha_i = 1) = \pi_i$,π_i 为第 i 个单元的包含概率。

如果 $\pi_i > 0$ $(i = 1, 2, ..., N)$, $\hat{Y}_{HT}$ 是 Y 的无偏估计:

$$\mathrm{E}(\hat{Y}_{HT}) = \sum_{i=1}^{N} \mathrm{E}(\alpha_i)\frac{y_i}{\pi_i} = \sum_{i=1}^{N} y_i = Y \tag{4.5}$$

$\hat{Y}_{HT}$ 的方差为：

$$\mathrm{Var}(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} {Y_i}^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} Y_iY_j \tag{4.6}$$

当 n 固定时，则有

$$\mathrm{Var}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j>i}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \tag{4.7}$$

如果 $\pi_i > 0$, $\pi_{ij} > 0$ $(i, j = 1, 2, ..., N; i \neq j)$，则 $\mathrm{Var}(\hat{Y}_{HT})$ 的无偏估计为：

$$\mathrm{var}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} y_iy_j \tag{4.8}$$

当 n 固定时，则 $\mathrm{Var}(\hat{Y}_{HT})$ 可以用耶茨、格伦迪和森提出的估计，也是无偏估计：

$$\mathrm{var}_{YGS}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{4.9}$$

在实际工作中，这两个估计式都不太理想，因为他们都有可能是负值。但当 $n = 2$ 时，$\mathrm{var}_{YGS}\left(\hat{Y}_{HT}\right)$ 总是大于 0。

4.3 πPS 抽样的实现方法

4.3.1 严格的 πPS 抽样

当采样数量 $n = 2$ 时，通常采用逐个抽取样本的方法来保证抽样是不放回的，主要采用的方法是布鲁尔 (Brewer) 方法和 Duerbin 方法。

1. 布鲁尔方法：

(1). 假设对所有的 i, 有 $Z_i < \frac{1}{2}$;

(2). 按与 $\dfrac{Z_i(1-Z_i)}{1-2Z_i}$ 成正比的概率抽取第一个单元，记为 i;

(3). 剩下的 $N-1$ 个样本中，按与 $\dfrac{Z_j}{1-Z_i}(j \neq i)$ 成正比的概率抽取第二个。

2. Durbin 方法:

(1). 按照 Z_i 抽取第一个样本；

(2). 按与 $Z_j\left(\dfrac{1}{1-2Z_i}+\dfrac{1}{1-2Z_j}\right)$ 成正比的概率抽取第 2 个样本。

当采样数量 $n>2$ 时，也有几种采样方法。例如，可以采用分层抽样的思想将 $n=2$ 的布鲁尔方法和 Durbin 方法从 $n=2$ 拓展至 $n>2$，但是计算包含概率时十分复杂。下面介绍的水野法是一种比较简单的逐个抽取方法。

3. 水野法：

(1). 按照以下抽样概率抽取第一个样本单元：

$$Z_i^*=\frac{n(N-1)Z_i}{N-n}-\frac{n-1}{N-1},\quad i=1,2,\cdots,N \tag{4.10}$$

为了保证每个 $Z_i^*\geqslant 0$, 要求每个单元的大小满足：

$$M_i\geqslant\frac{(n-1)M_0}{n(N-1)} \tag{4.11}$$

(2). 在剩下的 $N-1$ 个样本中不放回、等概率地抽取出 $n-1$ 个样本单元。

备注：以上的抽样方法都是用于设定每一轮抽取样本单元的概率，在具体实施抽样的时候，主要采取代码法或拉希里法。

4.3.2 非严格的 πPS 抽样

非严格的 πPS 抽样是指在抽样过程中 n 不固定且随机确定的情况；也可以理解为不是严格的不放回抽样；或者说是包含概率 π_i 与单元大小并非严格成比例，即 $\pi_i=nZ_i$ 不严格成立。

针对非严格的 πPS 抽样，下面介绍一种抽样方法：耶茨-格伦迪 (Yates-Grundy) 逐个抽取法。

1. 耶茨-格伦迪 (Yates-Grundy) 逐个抽取法：

(1). 按照 Z_i 的概率抽取第一个样本，不妨记抽到的第一个样本单元为 1；

(2). 按照 $\dfrac{Z_i}{1-Z_1}$ 的概率抽取第二个样本，同样记抽到的第二个样本单元为 2；

(3). 按照 $\dfrac{Z_i}{1-Z_1-Z_2}$ 的概率抽取第三个样本，同样记抽到的第三个样本单元为 3；

以此类推直到抽取到 n 个样本截止抽样。

耶茨-格伦迪方法的 π_i 不容易计算，因而无法使用霍维茨-汤普森估计量。可以采用 Raj 估计量。

设 $y_1,y_2,\cdots,y_n$ 为按抽中顺序排列的样本单元的指标值，相应的 Z 值为 $Z_1,Z_2,\cdots,Z_n$，

令

$$\begin{cases} t_1 = \frac{y_1}{Z_1} \\ t_2 = y_1 + \frac{y_2}{Z_2}(1 - Z_1) \\ \cdots\cdots \\ t_n = y_1 + y_2 + \cdots + y_{n-1} + \frac{y_n}{Z_n}(1 - Z_1 - Z_2 - \cdots - Z_{n-1}) \end{cases} \tag{4.12}$$

则总体总值 Y 的 Raj 估计量为：

$$Y_{Raj} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{4.13}$$

它是总体总值的无偏估计，对其方差 $V\left(\hat{Y}_{Raj}\right)$ 的无偏估计为：

$$v\left(\hat{Y}_{Raj}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(t_i - \hat{Y}_{Raj}\right)^2 \tag{4.14}$$