南開大學
Nankai University

统计与数据科学学院

《数据采集方法》课程报告

贵州茅台股票指标抽样与分析

小　　组：B+X9bo 组

姓　　名：蒋贵豪

年　　级：2021 级

专　　业：应用统计学

完成日期：2021 年 12 月 12 日

目录

1 数据介绍 **1**

2 简单随机抽样 **1**

2.1 理论部分与系数选取 2

2.2 结果展示与分析 3

2.2.1 结果展示 3

2.2.2 结果分析 4

3 分层随机抽样 **5**

3.1 分层抽样的概念 5

3.2 分层抽样的估计量及性质 5

3.3 分层随机抽样 6

3.4 总样本量及各层样本量的分配 7

3.4.1 总样本量的确定 7

3.4.2 各层样本量的分配 7

3.5 层的划分 7

3.6 实例数据实验及结果分析 8

4 等概率整群抽样和多阶段抽样 **11**

4.1 等概率整群抽样 11

4.1.1 基本概念 11

4.1.2 符号说明与公式 11

4.1.3 结果展示 11

4.2 等概率两阶段抽样 12

4.2.1 基本概念 12

4.2.2 符号说明与公式 12

4.2.3 结果展示 14

5 不等概抽样 **15**

5.1 不等概抽样的概念 15

5.2 在实验中使用不等概抽样的前提假设 15

5.3 不等概抽样的方法介绍 15
5.3.1 放回不等概抽样 (PPS) 15
5.3.2 不放回不等概抽样 (πPS) 16
5.4 金融数据实验 17

6 各类抽样方法的结果比较 **19**

7 小组分工 **19**

1 数据介绍

贵州茅台酒股份有限公司 (600519) 主要业务是茅台酒及系列酒的生产与销售。主导产品“贵州茅台酒”是世界三大蒸馏名酒之一，也是集国家地理标志产品、有机食品和国家非物质文化遗产于一身的白酒品牌。贵州茅台股价在 A 股市场“一骑绝尘”，因而其在 A 股时长有极高的关注度。

在我们本次的抽样与分析任务中，我们选取了贵州茅台贵州茅台 (600519) 近 3 年 (2018 年 11 月 23 日至 2021 年 11 月 23 日) 交易日的股票振幅、总手、金额、换手率和成交次数 5 项指标数据。我们直接从同花顺电脑端软件中直接导出了这些所需要的数据，共有 $N = 728$ 条。在我们的数据预处理时，因为我们所需要的数据从专业的证券软件中导出，我们的数据并无异常，并不需要进行数据清洗。我们所使用的原始数据表头展示见表1.1，详见“*600519.xlsx*”。

表 1.1: 原始数据表头展示

时间	振幅	总手	金额	换手率	成交次数
2018-11-23, 五	0.0184	2012621	1126652920	0.16	3240
2018-11-26, 一	0.0147	1912019	1063536980	0.15	3271
2018-11-27, 二	0.0154	2023382	1112920630	0.16	3368
2018-11-28, 三	0.0250	2773793	1535544400	0.22	3850
2018-11-29, 四	0.0373	4506945	2524752300	0.36	4008

2 简单随机抽样

选取振幅、总手、金额、换手率四个指标，对原始数据进行相关性分析，我们可以得到相关性图如图2.1所示。分析各个指标间的相关系数矩阵，发现换手和其他 3 个指标相关性较高。于是最终决定选用振幅、总手、金额 3 个指标作为主要分析对象，选择换手指标作为辅助变量，进行相关的比率估计和回归估计。

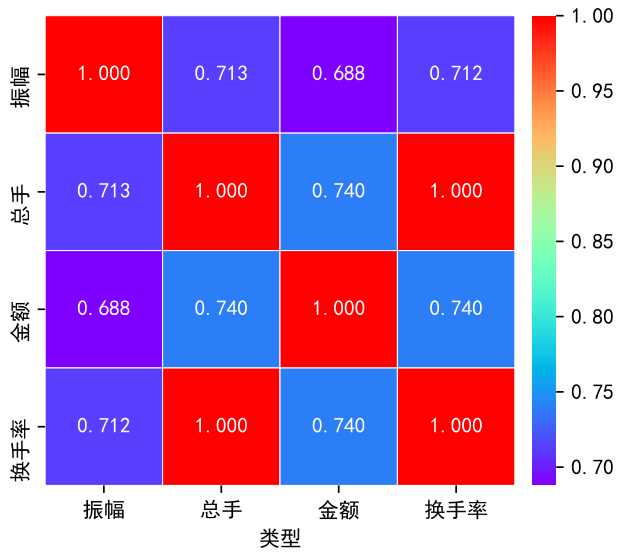

图 2.1: 指标相关系数热力图

2.1 理论部分与系数选取

采用简单随机抽样的方式在近三年的数据中抽取样本点，其中样本量的选取依据公式：

$$n_0 = 1/\left(\frac{1}{N} + \frac{d^2}{z_{\alpha/2}^2 S^2}\right) = \frac{N z_{\alpha/2}^2 S^2}{N d^2 + z_{\alpha/2}^2 S^2} \tag{2.1}$$

其中，N 为我们的总体数据量，S 为总体方差，d 为绝对误差。我们选定的置信度 $1-\alpha = 95\%$, 相对误差 $r = d/\bar{y} = 0.1$。通过该步骤，可以确定每个指标的样本量如表2.1所示：

表 2.1: 简单随机抽样样本量确定

振幅	总手	金额	换手率
81	73	107	73

每个指标分别确定抽样的样本量之后，可以计算简单随机抽样样本均值 $\bar{y}$ 和方差 s^2 以及 $\bar{y}$ 的方差。因而可以比较样本均值与总体的均值之间的差异。抽样均值的方差为：

$$\mathrm{Var}(\bar{y}) = \frac{1-f}{n} S^2 \tag{2.2}$$

其中，n 为表2.1中给出的样本量。f 为样本容量与总体规模的比例。

然后，分别进行比率估计和回归估计：

对总体均值的比率估计量为式2.3：

$$\hat{\bar{Y}}_R = \bar{y}_R = \bar{X}\frac{\bar{y}}{\bar{x}} = \frac{1}{N}X\hat{R} \tag{2.3}$$

其中，$\hat{R} = \bar{y}/\bar{x}$。回归估计均值的方差为式2.4：

$$\widehat{\mathrm{Var}}\left(\bar{y}_R\right) \approx \frac{1-f}{n}\left(s^2 - 2\hat{R}s_{yx} + \hat{R}^2 s_x^2\right) \tag{2.4}$$

其中，s_{xy} 是 x 和 y 的样本协方差，和 s_x^2 是 x 的样本方差。

回归估计的总体均值回归估计量为式2.5：

$$\bar{y}_{br} = \bar{y} + b(\bar{X} - \bar{x}) \tag{2.5}$$

其中 b 为：

$$b = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2.6}$$

回归估计量的方差为式2.7：

$$\mathrm{Var}\left(\bar{y}_{tr}\right) \approx \frac{1-f}{n}S^2\left(1-\rho^2\right) \tag{2.7}$$

其中 ρ 为 Y 与 X 的相关系数：

$$\rho = \frac{S_{xy}}{S_x S} \tag{2.8}$$

2.2 结果展示与分析

2.2.1 结果展示

下面的表2.2、表2.3、表2.4展示振幅、总手、金额的简单随机抽样、比率估计和回归估计得到的总体均值的估计，以及这个估计量的方差和置信区间的情况。

表 2.2: 振幅的估计

方法	总体均值估计	估计量方差	置信区间
总体	0.026	2.18×10^{-7}	$[0.001, 0.051]$
简单随机抽样	0.026	1.72×10^{-6}	$[0.024, 0.029]$
比率估计	0.026	1.07×10^{-6}	$[0.024, 0.028]$
回归估计	0.026	9.64×10^{-7}	$[0.024, 0.028]$

表 2.3: 总手的估计

方法	总体均值估计	估计量方差	置信区间
总体	3890360.50	4330595718.31	$[410293.53, 7370427.47]$
简单随机抽样	4062302.25	51703742188.99	$[3616636.68, 4507967.83]$
比率估计	3887104.56	17040106.57	$[3879013.89, 3895195.23]$
回归估计	3887109.94	16933568.60	$[3879044.61, 3895175.28]$

表 2.4: 金额的估计

方法	总体均值估计	估计量方差	置信区间
总体	5471567360.00	1.3368×10^{16}	$[-642648729, 11585783451]$
简单随机抽样	5397944832.00	7.0001×10^{16}	$[4879384396, 5916505268]$
比率估计	5466988067.67	5.4772×10^{16}	$[5008289632, 5925686503]$
回归估计	5467341913.79	5.4312×10^{16}	$[5010574754, 5924109073]$

2.2.2 结果分析

从得到的结果来看，总体而言，3 种方法得到的均值估计，估计量的方差和置信区间都比较令人满意。相对而言，通过比较总体的情况，可以发现，由于这 3 个指标和换手率之间存在着较高的相关性，因此，利用比率估计和回归估计得到的结果要优于简单随机抽样的估计结果。在我们分析的数据中，可以发现回归估计最优，比率估计其次，二者不分伯仲，简单随机抽样相对精度要差一点。

3 分层随机抽样

3.1 分层抽样的概念

分层抽样 (stratified sampling) 又称为类型抽样或分类抽样，即在每一层中独立进行抽样，总的样本由各层样本组成，总体参数根据各层样本参数的汇总做出估计。表3.1和表3.2给出了本章所用到的相关符号说明：

表 3.1: 分层抽样相关符号说明表 1

符号	含义
N	总体单元数
n	样本单元数
L	总层数
h	下标，表示第 h 层
i	下标，表示层内第 i 个单元
N_h	第 h 层的总体单元数
n_h	第 h 层的样本单元数
Y_{hi}	第 h 层第 i 个总体单元的取值
y_{hi}	第 h 层第 i 个样本单元的取值

3.2 分层抽样的估计量及性质

对于分层抽样，各层抽样是独立进行的，即可用不同的抽样方法对各层指标独立进行推算。对于总体指标的估计，是各层相应指标估计值的加权和。

总体均值 $\bar{Y}$ 的估计为：

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{\bar{Y}}_h \tag{3.1}$$

总体均值 $\bar{Y}$ 的简单估计为：

$$\hat{\bar{y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h \tag{3.2}$$

表 3.2: 分层抽样相关符号说明表 2

符号	公式	含义
W_h	$\frac{N_h}{N}$	第 h 层的总体权值
w_h	$\frac{n_h}{n}$	第 h 层的样本权值
f_h	$\frac{n_h}{N_h}$	第 h 层的抽样比
$\bar{Y}_h$	$\frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi}$	第 h 层的总体均值
$\bar{y}_h$	$\frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$	第 h 层的样本均值
Y_h	$\sum_{i=1}^{N_h} Y_{hi} = N_h\bar{Y}_h$	第 h 层的总体总量
y_h	$\sum_{i=1}^{n_h} y_{hi} = n_h\bar{y}_h$	第 h 层的样本总量
S_h^2	$\frac{\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h)^2}{N_h-1}$	第 h 层的总体方差
s_h^2	$\frac{\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2}{n_h-1}$	第 h 层的样本方差

这里“st”是“$stratified$”一词的缩写，意为用分层抽样的方法得到的估计。对于一般的分层抽样，如果 $\hat{\bar{Y}}_h$ 是 $\bar{Y}_h$ 的无偏估计 $(h = 1, 2, ..., L)$，则 $\hat{\bar{Y}}_{st}$ 是 $\bar{Y}$ 的无偏估计，也就是说，只要对各层估计无偏，则总体估计也无偏。

$\hat{\bar{Y}}_{st}$ 的方差为：

$$\mathrm{Var}(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{L} W_h^2 \mathrm{Var}(\hat{\bar{Y}}_h) \tag{3.3}$$

3.3 分层随机抽样

如果各层的抽样都是简单随机抽样，那么这种分层抽样为分层随机抽样。对于分层随机抽样，总体均值 $\bar{Y}$ 的简单估计 $\bar{y}_{st}$ 是 $\bar{Y}$ 的无偏估计，$\bar{y}_{st}$ 的方差为：

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \mathrm{Var}(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} S_h^2 \tag{3.4}$$

$\mathrm{Var}(\bar{y}_{st})$ 的无偏估计为：

$$\mathrm{var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \mathrm{var}(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} s_h^2 \tag{3.5}$$

3.4 总样本量及各层样本量的分配

3.4.1 总样本量的确定

设总的样本量为 n，n 的确定与所需要的绝对精度 d 或相对精度 r 有关，式3.6给出总样本量 n 的值：

$$n = \frac{\sum_{h=1}^{L} W_h^2 S_h^2 / w_h}{(\frac{d}{u})^2 + \sum_{h=1}^{L} W_h S_h^2 / N} = \frac{\sum_{h=1}^{L} W_h^2 S_h^2 / w_h}{(\frac{r\bar{Y}}{u})^2 + \sum_{h=1}^{L} W_h S_h^2 / N} \tag{3.6}$$

这里 $u = z_{\alpha/2}$，代表标准正态分布的右分位点。

3.4.2 各层样本量的分配

设第 h 层样本量为 n_h，那么样本量在各层的分配满足：

$$n_h = n w_h \tag{3.7}$$

下面是三种分配方法：

a. 平均分配：$n_h = \frac{n}{L}$ 即 $w_h = \frac{1}{L}$。

b. 按比例分配：$w_h = \frac{n_h}{n} = \frac{N_h}{N} = W_h$，此时 $f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$

c. Neyman 最优分配：$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$

3.5 层的划分

对于分层抽样，我们的分层原则是：总体中的每一个单元一定只属于某一个层，并且尽可能使层内单元的指标值相近，层间单元差异大，以提高精度；同时还要便于组织抽样实施。

3.6 实例数据实验及结果分析

我们首先确定总样本量、层的划分及各层样本量的分配。

观察各个变量的内容，可以看出第一个变量（时间）是可以看作按月份或者按星期的定性变量，后五个变量（振幅、总手、金额、换手、成交量）都是定量变量，因此我们将依据时间作为分层抽样的依据。

首先提取时间中的星期和月份信息，我们以振幅为例子，创建两个变量 **day** 和 **month**，可以观察到，在这一年内，星期一至星期五的数据量大致是均匀分配的，如表3.3所示：

表 3.3: 按星期划分的数据数量

星期	一	二	三	四	五
总量	144	143	143	143	147

而每月的数据量却有一定的差别，2、5、10 三个月份分别对应春节、五一劳动节与十一国庆节，因此数据量较少，如表3.4所示：

表 3.4: 按月份划分的数据数量

月份	1	2	3	4	5	6	7	8	9	10	11	12
总量	58	50	66	63	56	60	68	65	62	50	65	65

然后分别根据这两个变量进行分层。考虑各层样本量的分配，分别使用平均分配、按比例分配和 Neyman 最优分配。选定相对精度为 $r = 0.1$，确定的总样本量以及各层样本量如表3.5和表3.6所示：

表 3.5: 按星期分层各分配方法样本量确定

星期	一	二	三	四	五	总量
平均分配	16	16	16	16	16	80
比例分配	16	16	16	16	17	81
Neyman 最优分配	17	18	16	16	14	81

表 3.6: 按月份划分的数据数量

月份	1	2	3	4	5	6	7	8	9	10	11	12	总量
平均分配	7	7	7	7	7	7	7	7	7	7	7	7	84
比例分配	6	5	7	7	6	6	7	7	6	5	7	7	76
Neyman 最优分配	5	7	8	6	6	5	9	7	7	4	5	5	74

我们使用分层抽样的 **sampling** 包 **strata** 函数进行抽样，并计算总体均值和方差的估计值，所得到的结果如表3.7、表3.8和表3.9所示：

表 3.7: 振幅的分层抽样估计

方法	均值估计	估计量方差	所需样本量
总体	2.589×10^{-2}	1.588×10^{-4}	/
按星期平均分配	2.529×10^{-2}	1.711×10^{-4}	80
按月平均分配	2.562×10^{-2}	1.438×10^{-4}	84
按星期比例分配	2.368×10^{-2}	8.083×10^{-5}	81
按月比例分配	2.553×10^{-2}	1.643×10^{-4}	76
按星期 Neyman 分配	2.361×10^{-2}	1.345×10^{-4}	81
按月 Neyman 分配	2.479×10^{-2}	1.305×10^{-4}	74

表 3.8: 总手的分层抽样估计

方法	均值估计	估计量方差	所需样本量
总体	3.890×10^{6}	3.153×10^{12}	/
按星期平均分配	3.991×10^{6}	2.999×10^{12}	72
按月平均分配	4.170×10^{6}	4.293×10^{12}	65
按星期比例分配	3.772×10^{6}	1.678×10^{12}	72
按月比例分配	3.743×10^{6}	1.774×10^{12}	65
按星期 Neyman 分配	3.748×10^{6}	2.013×10^{12}	71
按月 Neyman 分配	3.773×10^{6}	2.423×10^{12}	62

表 3.9: 金额的分层抽样估计

方法	均值估计	估计量方差	所需样本量
总体	5.472×10^9	9.732×10^{18}	/
按星期平均分配	5.161×10^9	8.429×10^{18}	107
按月平均分配	5.630×10^9	1.157×10^{19}	101
按星期比例分配	4.883×10^9	8.280×10^{18}	107
按月比例分配	5.761×10^9	1.238×10^{19}	101
按星期 Neyman 分配	5.380×10^9	9.031×10^{18}	106
按月 Neyman 分配	5.143×10^9	7.462×10^{18}	94

4 等概率整群抽样和多阶段抽样

4.1 等概率整群抽样

4.1.1 基本概念

整群抽样是将总体分为若干群，抽取某一个群中的所有个体。群体的划分可以按照行政或地域，也可以人为进行划分。而分群的原则与分层抽样恰恰相反：群内方差尽可能大，群间方差尽可能小。当群规模较大时精度不高，但便于操作；群规模较小时虽然精度较高，但费用也可能会比较高。当规模过大时需要多阶段抽样。

4.1.2 符号说明与公式

N 和 n 分别为总体群数、样本群数。因为我们抽的是群，这里群就是基本单元。M_i 为第 i 个群的个体数量，这里假定各群规模相等：

$$M_1 = M_2 = \ldots\ldots = M_N = M \tag{4.1}$$

M_0 为总体的个体数量：

$$M_0 = \sum_{i=1}^{N} M_i \tag{4.2}$$

对总体均值的估计为:

$$\hat{\bar{Y}} = \bar{\bar{y}} = \sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij}/nM = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i \tag{4.3}$$

总体方差为:

$$S^2 = \frac{1}{M_0 - 1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2 \tag{4.4}$$

总体均值估计量的方差为：

$$\mathrm{Var}(\bar{\bar{y}}) = \frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 \tag{4.5}$$

同样，我们选用振幅、总手和金额 3 个指标作为主要分析对象，进行等概率整群分析。

4.1.3 结果展示

我们将原有的 728 个数据等规模分成 91 个群，设定样本群数为 5，也就是我们从中抽取 5 个群进行整群抽样。估计的目标变量是振幅、总手和金额的均值与总值，以及其相应的标准误差，结果如表4.1所示。

表 4.1: 等概率整群抽样指标估计结果

指标	总量估计	均值估计	均值估计的方差
振幅	19.889	0.02732	1.528739×10^{-5}
总手	3381328915	4644683	815248336635
金额	4.0434×10^{12}	5554174425	1.258927×10^{18}

4.2 等概率两阶段抽样

4.2.1 基本概念

多阶段抽样是整群抽样的一般化。如对于两阶段抽样，我们先从 N 个群中抽取 n 个群，第一阶段得到了：

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M_1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{iM_i} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nM_n} \end{bmatrix} \tag{4.6}$$

此时如果是整群抽样，那么抽样已经结束。但两阶段抽样还需要再从第 i 个群中抽取 m_i 个，而不是全部抽出。如果是多阶段抽样，那么这 $\sum_{i=1}^{n} m_i$ 个单元还不是最终的单元，需要再分别从各个单元中抽取部分个体。

其中，第一阶段抽到的称为初级单元 PSU；第二阶段抽到的称为二级单元 SSU；第三阶段抽到的称为三级单元 TSU；最终调查的单元称为最终单元 USU。

4.2.2 符号说明与公式

下面考虑每个 PSU 的规模相等，表4.2给出关于总量、均值的符号说明，表4.3给出关于方差的符号说明。

和整群抽样一样，用 PSU 平均的平均作为 SSU 均值的估计，式4.7第二个等号只在规模相等时成立。

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} \tag{4.7}$$

表 4.2: 关于均值符号的说明

符号	说明
$Y_i = \sum_{j=1}^{M} Y_{ij}$	总体第 i 个 PSU 的总量
$y_i = \sum_{j=1}^{m} y_{ij}$	样本第 i 个 PSU 的总量
$\bar{Y}_i = \frac{Y_i}{M}$	总体第 i 个 PSU 的均值
$\bar{y}_i = \frac{y_i}{m}$	样本第 i 个 PSU 的均值
$\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_i$	总体 SSU 的均值
$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$	样本 SSU 的均值

表 4.3: 关于方差符号的说明

符号	说明
$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$	总体 PSU 的方差
$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2$	样本 PSU 的方差
$S_{2i}^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2$	总体第 i 个 PSU 内 SSU 的方差
$s_{2i}^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2$	样本第 i 个 PSU 内 SSU 的方差
$S_2^2 = \frac{1}{N}\sum_{i=1}^{N} S_{2i}^2$	总体群内方差: 总体每个 PSU 内 SSU 的方差的平均
$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} s_{2i}^2$	样本群内方差: 样本每个 PSU 内 SSU 的方差的平均

方差的无偏估计量为：

$$\mathrm{var}(\bar{\bar{y}}) = \frac{1-f_1}{n}s_1^2 + \frac{f_1(1-f_2)}{nm}s_2^2 \tag{4.8}$$

同样，我们选用振幅、总手和金额 3 个指标作为主要分析对象，进行等概率两阶段抽样。

4.2.3 结果展示

将 728 个数据等规模分成 91 个群，第一阶段先从中抽取 25 个群，第二阶段从这 25 个群中每个群内抽取五分之一的样本单位，最终目标变量是振幅、总手和金额的均值与总值，以及均值的标准误差进行估计，结果如表4.4所示。

表 4.4: 等概率两阶段抽样指标估计结果

指标	总量估计	均值估计	均值估计的方差
振幅	17.29334	0.023722	0.00014605
总手	2874691215	3943335	4.4611×10^{12}
金额	4.112971×10^{12}	5649685972	1.3384×10^{19}

5 不等概抽样

5.1 不等概抽样的概念

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。而不等概率抽样的方法是认为总体中的每个单元的地位是不同的，这个不同由辅助变量得到。

不等概抽样分为放回不等概抽样 (PPS) 和不放回不等概抽样 (πPS)。其中 PPS 主要通过累计规模法 (代码法) 和拉希里方法进行采样，通过 Hansen-Hurwi 方法进行统计量估计；πPS 主要通过耶茨-格伦迪 (Yates-Grundy) 逐个抽取法进行非严格抽样，并使用 Raj 方法进行统计量估计。

5.2 在实验中使用不等概抽样的前提假设

在本实验中使用不等概抽样使用了以下假设：

- 在上述的指标相关性分析中得到，分析的总体为：振幅、总手、金额 3 个指标，辅助变量为：换手率。
- 以一日数据作为一个整体进行抽样。

5.3 不等概抽样的方法介绍

5.3.1 放回不等概抽样 (PPS)

放回不等概抽样中，最常用的时按照总体单元的规模大小 (辅助变量的大小) 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：

$$Z_i = \frac{M_i}{M_0} = \frac{M_i}{\sum_{i=1}^{N} M_i} \tag{5.1}$$

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 $M^* = \max_{1 \le i \le N} \{M_i\}$. 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。

总体总量 Hansen-Hurwitz 估计，其表达式如式5.2, 式5.3：

(a) 总量无偏估计：

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\varphi_i} = \frac{M_0}{n} \sum_{i=1}^{n} \frac{y_i}{M_i} \tag{5.2}$$

这里 M_i 表示样本中第 i 个抽样单元的规模；

(b) 方差无偏估计：

$$\mathrm{Var}(\hat{Y}_{HH}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i}{\varphi_i} - \hat{Y}_{HH} \right)^2 = \frac{M_0^2}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{M_i} - \frac{\hat{Y}_{HH}}{M_0} \right)^2 \tag{5.3}$$

5.3.2 不放回不等概抽样 (πPS)

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。

样本的抽取使用耶茨-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

$$\begin{cases} t_1 = \dfrac{y_1}{Z_1} \\ t_2 = y_1 + \dfrac{y_2}{Z_2}(1 - Z_1) \\ \cdots\cdots \\ t_n = y_1 + y_2 + \cdots + y_{n-1} + \dfrac{y_n}{Z_n}(1 - Z_1 - Z_2 - \cdots - Z_{n-1}) \end{cases} \tag{5.4}$$

则总体总值 Y 的 Raj 估计量如式5.5所示：

$$Y_{Raj} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{5.5}$$

式5.5是总体总值的无偏估计，其方差 $\mathrm{Var}\left(\hat{Y}_{Raj}\right)$ 的无偏估计式5.6所示：

$$\mathrm{var}\left(\hat{Y}_{Raj}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(t_i - \hat{Y}_{Raj}\right)^2 \tag{5.6}$$

5.4 金融数据实验

实验使用换手率作为辅助变量 (M_i), 估计振幅、总手、金额 3 个指标的总体总量和均值，以及统计量对应的方差。实验结果如表5.1、表5.2和表5.3所示。

从表5.1、表5.2和表5.3得到的结果中，我们使用的四种方法的抽样结果对总体重量和总体均值的估计都是令人满意的。振幅估计中，我们的均值相对误差率最高为 5.51%; 金额估计中，均值估计相对误差率最高 2.54%，而总手估计的误差率最高已经在 0.13%。我们也发现 Yates-Grundy 法得到均值估计的方差小于 PPS 法的方差。因此，使用不等概抽样的方法对贵州茅台近 3 年振幅、总手、金额指标有较好的估计。

表 5.1: 振幅估计结果

振幅 (样本量 $n = 100$)	总量	总量估计值方差	均值	均值估计值方差	均值估计相对真实值误差
真实值	18.8514	-	0.0259	-	-
PPS 代码法估计	19.1826	0.3049	0.0263	5.75314×10^{-7}	0.0176
PPS 拉希里法估计	19.4051	0.3583	0.0267	6.76004×10^{-7}	0.0294
Yates-Grundy-代码法	19.1984	0.3294	0.0264	6.21525×10^{-7}	0.0184
Yates-Grundy-拉希里法	19.8894	0.2873	0.0273	5.42086×10^{-7}	0.0551

表 5.2: 总手估计结果

总手 (样本量 $n = 100$)	总量	总量估计值方差	均值	均值估计值方差	均值估计相对真实值误差
真实值	2832180989	-	3890358.501	-	-
PPS 代码法估计	2835503829	1.10123×10^{13}	3894922.842	2.08×10^{7}	0.001173244
PPS 拉希里法估计	2833509865	1.06228×10^{13}	3892183.881	2.00×10^{7}	0.000469206
Yates-Grundy-代码法	2832078133	7.18664×10^{12}	3890217.215	1.36×10^{7}	-3.63×10^{-5}
Yates-Grundy-拉希里法	2836035315	7.46144×10^{12}	3895652.905	1.41×10^{7}	0.001360904

表 5.3: 金额估计结果

金额 (样本量 $n = 100$)	总量	总量估计值方差	均值	均值估计值方差	均值估计相对真实值误差
真实值	3.983×10^{12}	-	5.472×10^{9}	-	-
PPS 代码法估计	4.033×10^{12}	1.921×10^{22}	5.540×10^{9}	3.624×10^{16}	1.248×10^{-2}
PPS 拉希里法估计	4.085×10^{12}	1.948×10^{22}	5.611×10^{9}	3.676×10^{16}	2.541×10^{-2}
Yates-Grundy-代码法	4.001×10^{12}	1.420×10^{22}	5.496×10^{9}	2.678×10^{16}	4.485×10^{-3}
Yates-Grundy-拉希里法	3.964×10^{12}	1.730×10^{22}	5.445×10^{9}	3.265×10^{16}	-4.862×10^{-3}

6　各类抽样方法的结果比较

在 2-5 章中，针对我们选取的贵州茅台近 3 年的指标数据，我们分别采取了简单随机抽样中的简单估计量、比率估计量和回归估计量；分层随机抽样中的按星期与月份分层，并使用了平均分配、按比例分配和 Neyman 最优分配 3 种方法确定样本量；等概率整群抽样和等概率两阶段抽样；不等概抽样中的 PPS 和 Yates-Grundy 法，并分别用代码法和拉希里法进行了实现。

从抽样的结果与真实值相比较来看，各种方法得到的总体均值的估计都是非常有效的。除了换手率指标因为数值较小，最大相对能达到 5%，其它指标的估计使用 4 种抽样方法的相对误差均在 1% 以下。

而从估计量的方差来看，因为简单随机抽样和不等概抽样的原理的相似性，它们两个得到的估计量方差较为接近。同样，由于分层随机抽样和等概率整群抽样的原理的相似性，它们得到的估计量的方差也较为接近。且分层随机抽样和等概率整群抽样得到的方差会比简单随机抽样和不等概抽样得到的方差大。

不过，由于金融指标的变换的原理非常复杂，我们在简单随机抽样中计算了置信区间。从置信区间中我们可以看到，其是一个非常大的范围，因此想十分准确的估计一个金融指标是非常困难的。我们的使用的抽样技术仅仅只是一种参考和指导的意义。在实际的应用中，我们应使用更为复杂的模型，并在具体的情况下，对模型进行相应的修改，以适合我们最后的估计。

7　小组分工

我们的抽样中用到的数据和代码可见：*https://github.com/zhehaoXu6/-Data-collection-methods_code2021*