Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 546-565
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111062

Hans汉斯

分层随机抽样中九个非常实用的R函数

崔　娅[1*]，张应应[1,2*#]

[1]重庆大学数学与统计学院统计与精算学系，重庆
[2]重庆大学分析数学与应用重庆市重点实验室，重庆

收稿日期：2021年12月26日；录用日期：2022年1月21日；发布日期：2022年1月28日

摘　要

分层抽样技术是在实际工作中应用得非常广泛的抽样技术之一。但在文献中，还没有方便地可以用于在分层随机抽样中仅给定基本的样本数据时就能解决总体均值和总体比例的点估计和区间估计问题，计算总体均值时样本量的确定及分配问题，计算总体比例时样本量的确定及分配问题，事后分层抽样下总体均值和总体比例的点估计和区间估计等问题通用的R函数。本文自编了九个通用的R函数：Compute_Y_bar_st()、Compute_Y_bar_prop_from_y_bar_h_s_h_st()、Compute_Y_bar_srs_pst()、Compute_P_st()、Compute_P_from_a_h_st()、Compute_P_srs_pst()、Compute_nh_given_n_Y_bar_st()、Compute_n_nh_Y_bar_st()及Compute_n_nh_P_st()，它们将会为需要使用分层抽样技术以提高估计精度进行实际问题分析的使用者提供极大的方便。

关键词

分层随机抽样，总体均值和总体比例，点估计和区间估计，样本量的确定及分配，R函数

Nine Very Practical R Functions in Stratified Random Sampling

Ya Cui[1*], Yingying Zhang[1,2*#]

[1]Department of Statistics and Actuarial Science, College of Mathematics and Statistics, Chongqing University, Chongqing
[2]Chongqing Key Laboratory of Analytic Mathematics and Applications, Chongqing University, Chongqing

Received: Dec. 26th, 2021; accepted: Jan. 21st, 2022; published: Jan. 28th, 2022

[*]第一作者。
[#]通讯作者。

文章引用: 崔娅, 张应应. 分层随机抽样中九个非常实用的 R 函数[J]. 应用数学进展, 2022, 11(1): 546-565.
DOI: 10.12677/aam.2022.111062

Abstract

Stratified sampling technique is one of the sampling techniques widely used in practical work. But in the literature, there are no convenient generic R functions to solve the problem of point estimation and interval estimation of population mean and population proportion, the problem of total sample size and each layer sample size when calculating population mean, the problem of total sample size and each layer sample size when calculating population proportion, and the problem of point estimation and interval estimation of population mean and population proportion in post-stratification sampling, in stratified random sampling when only basic sample data are given. We compile nine generic R functions: Compute_Y_bar_st(), Compute_Y_bar_prop_from_y_bar_h_s_h_st(), Compute_Y_bar_srs_pst(), Compute_P_st(), Compute_P_from_a_h_st(), Compute_P_srs_pst(), Compute_nh_given_n_Y_bar_st(), Compute_n_nh_Y_bar_st(), and Compute_n_nh_P_st(), which will provide great convenience for users who need to use stratified sampling technology to improve the estimation accuracy for practical problem analysis.

Keywords

Stratified Random Sampling, The Population Mean and Population Proportion, Point Estimation and Interval Estimation, Determination and Distribution of Sample Size, R Function

1. 引言

分层抽样技术[1]-[11]是在实际工作中应用得非常广泛的抽样技术[12]-[17]之一。同时 R 软件[18] [19]作为统计学的常用编程工具，它具有完全免费、简洁高效、运行方便等优点。本文选取 R 软件对分层随机抽样中的总体均值和总体比例的点估计和区间估计问题，计算总体均值时样本量的确定及分配问题，计算总体比例时样本量的确定及分配问题，事后分层抽样下总体均值和总体比例的点估计和区间估计等问题进行了程序实现。针对分层随机抽样，本文自编了九个非常实用的 R 函数。我们在对这九个 R 函数进行输入变量及输出变量的解释后给出了相应实际问题的 R 程序实现。这些内容构成了本文第一作者毕业论文的核心内容[20]。我们相信，这九个 R 函数将会为需要使用分层抽样技术以提高估计精度进行实际问题分析的使用者提供极大的方便。

2. 分层随机抽样中九个非常实用的 R 函数及应用举例

我们推荐分层随机抽样中九个非常实用的 R 函数。R 函数 1~3 用于解决给定各种样本信息时分层随机抽样下总体均值的点估计和区间估计问题。R 函数 4~6 用于解决给定各种样本信息时分层随机抽样下总体比例的点估计和区间估计问题。R 函数 7~9 用于解决给定各种样本信息时分层随机抽样下总体均值和总体比例的样本量的确定及分配问题。

R 函数 1：Compute_Y_bar_st()

对于分层随机抽样，给定样本单位的观察值组成的矩阵 y_matrix 等信息，得到计算总体均值的分层

随机抽样的点估计和区间估计的 R 函数(程序) Compute_Y_bar_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中(下载链接：https://pan.baidu.com/s/1y0UvE24vfVm8dTVAUYnskg，提取码：1234)。

下面我们举一个例子来说明该 R 函数的使用方法。

例 1 ([16]中例 4.1)为调查某地区住户的平均家庭成员数，将该地区分成城市和乡村 2 层，每层按简单随机抽样抽取 10 户，调查所获得的数据见表 1 所示。请估计该地区住户的平均家庭成员数及其 95%的置信区间。

Table 1. Household membership survey data
表 1. 家庭成员数调查数据

层(h)	居民户总数/户	家庭成员数(y_{hi})/人									
		1	2	3	4	5	6	7	8	9	10
城市	250	3	2	3	4	3	3	4	5	2	3
乡村	500	3	4	5	5	4	3	6	2	4	4

解：对于分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ N = \sum_{h=1}^{2} N_h = 750$$

$$\bar{y}_{st} = \sum_{h=1}^{2} W_h \bar{y}_h \approx 3.733333,\ v(\bar{y}_{st}) = \sum_{h=1}^{2} W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \approx 0.06708148$$

$$se(\bar{y}_{st}) = \sqrt{v(\bar{y}_{st})} \approx 0.2590009,\ \Delta = t \cdot se(\bar{y}_{st}) \approx 0.5076325,\ \gamma = \frac{\Delta}{\bar{y}_{st}} \approx 0.135973$$

$$L_{\bar{Y}} = \bar{y}_{st} - \Delta \approx 3.225701,\ U_{\bar{Y}} = \bar{y}_{st} + \Delta \approx 4.240966$$

代入数据，调用 R 函数 Compute_Y_bar_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.1。

因此，估计得该地区住户的平均家庭成员数为 3.733 人，抽样标准误为 0.259 人，平均家庭成员数的 95%置信区间为(3.226，4.241)人。

R 函数 2：Compute_Y_bar_prop_from_y_bar_h_s_h_st()

对于按比例分配的分层随机抽样，给定各层的样本均值 y_bar_h 和各层的样本标准差 s_h 等信息，得到计算总体均值的分层随机抽样的点估计和区间估计的 R 函数(程序) Compute_Y_bar_prop_from_y_bar_h_s_h_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 2 ([16]中练习 4.5)某开发区利用电话调查对区内居民消费冷冻食品情况进行调查，将电话号码(6 位数字)的前 2 位作为一部分，后 4 位作为一部分，前 2 位代表局号，局号及每个局号中拥有的电话数可以找到，按局号分层，按每个局号(去掉商户后)拥有的电话数比例分配样本量(各层抽样比可以忽略)。调查后各层样本户购买冷冻食品支出的中间结果见表 2 所示。试估计该开发区居民户购买冷冻食品的平均支出及估计的 95%置信区间。

Table 2. Sample households purchase frozen food expenditure
表 2. 样本户购买冷冻食品支出

序号	层权 W_h /%	样本量 n_h	样本平均 $\bar{y}_h$ /元	样本标准差 s_h
1	8.2	16	89	105
2	6.5	13	56	74
3	13.7	27	102	186
4	5.6	11	76	97
5	11.8	24	97	106
6	11.6	23	79	89
7	17.0	34	83	112
8	9.8	20	52	73
9	8.8	18	36	44
10	7.0	14	52	65

解：对于按比例分配的分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ n = \sum_{h=1}^{10} n_h = 200,\ \bar{y}_{st} = \sum_{h=1}^{10} W_h \bar{y}_h \approx 75.792$$

$$v(\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^{10} W_h s_h^2 \approx 59.46035,\ se(\bar{y}_{st}) = \sqrt{v(\bar{y}_{st})} \approx 7.711054$$

$$L_{\bar{Y}} = \bar{y}_{st} - t \cdot se(\bar{y}_{st}) \approx 60.67861,\ U_{\bar{Y}} = \bar{y}_{st} + t \cdot se(\bar{y}_{st}) \approx 90.90539$$

代入数据，调用 R 函数 Compute_Y_bar_prop_from_y_bar_h_s_h_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.2。

故该开发区居民户购买冷冻食品的平均支出为 75.792 元，标准误差为 7.711 元，其 95%置信区间为(60.679，90.905)元。

R 函数 3：Compute_Y_bar_srs_pst()

对于事后分层抽样，给定各层的样本均值 y_bar_h 和各层的样本标准差 s_h 等信息，得到计算总体均值的事后分层抽样和简单随机抽样的点估计和区间估计的 R 函数(程序) Compute_Y_bar_srs_pst()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 3 ([16]中例 4.6)某高校欲了解在校学生用于课外进修(如各种考证辅导班、外语辅导班等)的开支，在全校 8000 名学生中抽出了一个 200 人的简单随机样本。根据学生科的统计，本科生人数为全校学生的 70%，调查最近一个学期课外进修支出的结果见表 3 所示，试估计全校学生用于课外进修的平均开支。

Table 3. Results of the survey on expenditure on after-school education of current students
表 3. 在校学生课外进修开支调查结果

层(h)	层权(W_h)	样本量(n_h)/人	样本均值($\bar{y}_h$)/元	样本标准差(s_h)/元
本科生	0.7	120	253.4	231.00
研究生	0.3	80	329.4	367.00
合计	1	$n = 200$	$\bar{y} = 283.8$	$s \approx 294.57$

解：由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ n = \sum_{h=1}^{2} n_h = 200,\ f = \frac{n}{N} = 0.025,\ \bar{y} = \frac{1}{n}\sum_{h=1}^{2} n_h \bar{y}_h = 283.8$$

$$s^2 = \sum_{h=1}^{2} \frac{n_h - 1}{n-1} s_h^2 + \sum_{h=1}^{2} \frac{n_h}{n-1}\left(\bar{y}_h - \bar{y}\right)^2 \approx 86772.05,\ s = \sqrt{s^2} \approx 294.571$$

采用事后分层估计，则

$$v\left(\bar{y}_{pst}\right) = \frac{1-f}{n}\sum_{h=1}^{2} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{2}\left(1 - W_h\right)s_h^2 \approx 381.8343$$

$$\bar{y}_{pst} = \sum_{h=1}^{2} W_h \bar{y}_h = 276.2,\ se\left(\bar{y}_{pst}\right) = \sqrt{v\left(\bar{y}_{pst}\right)} \approx 19.54058$$

$$L_{\bar{Y}_{pst}} = \bar{y}_{pst} - t \cdot se\left(\bar{y}_{pst}\right) \approx 237.9012,\ U_{\bar{Y}_{pst}} = \bar{y}_{pst} + t \cdot se\left(\bar{y}_{pst}\right) \approx 314.4988$$

采用简单随机估计，则

$$v\left(\bar{y}\right) = \frac{1-f}{n} s^2 \approx 423.0137,\ se\left(\bar{y}\right) = \sqrt{v\left(\bar{y}\right)} \approx 20.5673$$

$$L_{\bar{Y}_{srs}} = \bar{y} - t \cdot se\left(\bar{y}\right) \approx 243.4888,\ U_{\bar{Y}_{srs}} = \bar{y} + t \cdot se\left(\bar{y}\right) \approx 324.1112$$

代入数据，调用 R 函数 Compute_Y_bar_srs_pst()进行计算，详细的 R 程序输入及输出结果请见附录 A.3。

因此，采用事后分层估计的平均开支为 276.2 元，其抽样标准误为 19.541 元，总体均值的事后分层抽样的 95%置信区间为(237.901，314.499)元。而采用简单随机估计的平均开支为 283.8 元，其抽样标准误为 20.567 元，总体均值的简单随机抽样的 95%置信区间为(243.489，324.111)元。事后分层估计的抽样标准误比简单随机估计的抽样标准误要小。

R 函数 4：Compute_P_st()

对于分层随机抽样，给定样本单位的观察值组成的矩阵 y_matrix 等信息，得到计算总体比例的分层随机抽样的点估计和区间估计的 R 函数(程序) Compute_P_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 4 ([16]中例 4.2)对某地区的居民拥有家庭电脑的情况进行调查，以居民户为抽样单位，根据收入水平将居民户划分为 4 层，每层按简单随机抽样抽取 10 户，调查获得数据见表 4 所示。估计该地区居民拥有家庭电脑的比例及抽样标准误。

Table 4. The sample households have a home computer
表 4. 样本户拥有家庭电脑情况

层	居民户总数/户	样本户拥有家庭电脑情况/台									
		1	2	3	4	5	6	7	8	9	10
1	200	0	0	0	1	0	0	0	1	0	0
2	400	0	1	0	0	0	0	0	0	1	0
3	750	1	1	0	0	0	0	1	0	1	0
4	1500	1	0	0	0	0	0	0	0	0	0

解：对于分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ N = \sum_{h=1}^{4} N_h = 2850,\ p_{st} = \sum_{h=1}^{4} W_h p_h = 0.2$$

$$v(p_{st}) = \sum_{h=1}^{4} W_h^2 v(p_h) \approx 0.004998324,\ se(p_{st}) = \sqrt{v(p_{st})} \approx 0.07069883$$

$$\Delta = t \cdot se(p_{st}) \approx 0.1385672,\ \gamma = \frac{\Delta}{p_{st}} \approx 0.6928358$$

$$L_P = p_{st} - \Delta \approx 0.06143284\ ,\ U_P = p_{st} + \Delta \approx 0.3385672$$

代入数据，调用 R 函数 Compute_P_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.4。

因此，估计得该地区居民拥有家庭电脑的比例为 0.2，抽样标准误为 0.071，总体比例的分层估计的 95%置信区间为(0.061，0.339)。

R 函数 5：Compute_P_from_a_h_st()

对于分层随机抽样，给定各层样本中具有所考虑特征的单位数 a_h 等信息，得到计算总体比例的分层随机抽样的点估计和区间估计的 R 函数(程序)Compute_P_from_a_h_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 5 ([16]中练习 4.4)随着经济发展，某市居民正在悄悄改变过年的习惯，虽然仍有大多数居民除夕夜在家吃年夜饭、看电视节目，但也有些家庭到饭店吃年夜饭，或逛夜市，或利用过年的假期到外地旅游。为研究这种现象，某研究机构以市中心 165 万居民户作为研究对象，将居民户按 6 个行政区分层，在每个行政区随机抽出 30 户进行了调查(各层抽样比可以忽略)。每个行政区的情况以及在家吃年夜饭、看电视节目的居民户比例见表 5。试估计该市居民在家吃年夜饭的比例，并给出抽样标准误。

Table 5. Proportion of households and number of households
表 5. 居民户比例及在家居民户数

行政区(h)	居民户比例(W_h)	在家居民户数(a_h)/户
1	0.18	27
2	0.21	28
3	0.14	27
4	0.09	26
5	0.16	28
6	0.22	29

解：对于分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ p_{st} = \sum_{h=1}^{6} W_h p_h \approx 0.924$$

$$v(p_{st}) = \sum_{h=1}^{6} W_h^2 (1 - f_h) \frac{p_h q_h}{n_h - 1} \approx 0.0003969808,\ se(p_{st}) = \sqrt{v(p_{st})} \approx 0.01992438$$

$$L_P = p_{st} - t \cdot se(p_{st}) \approx 0.8849489,\ U_P = p_{st} + t \cdot se(p_{st}) \approx 0.9630511$$

$$p_h=\frac{a_h}{n_h},\ \text{即}\ p_1=0.900,\ p_2\approx0.933,\ p_3=0.900,\ p_4\approx0.867,\ p_5\approx0.933,\ p_6\approx0.967$$

$$q_h=1-p_h,\ \text{即}\ q_1=0.100,\ q_2\approx0.067,\ q_3=0.100,\ q_4\approx0.133,\ q_5\approx0.067,\ q_6\approx0.033$$

$$W_1p_1\approx0.162,\ W_2p_2\approx0.196,\ W_3p_3\approx0.126,\ W_4p_4\approx0.078,\ W_5p_5\approx0.149,\ W_6p_6\approx0.213$$

代入数据，调用 R 函数 Compute_P_from_a_h_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.5。

故该市居民在家吃年夜饭的比例为 0.924，抽样标准误为 0.020，95%置信区间为(0.885，0.963)。

R 函数 6：Compute_P_srs_pst()

对于事后分层抽样，给定各层样本中具有所考虑特征的单位数 a_h 等信息，得到计算总体比例的简单随机抽样和事后分层抽样的点估计和区间估计的 R 函数(程序) Compute_P_srs_pst()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 6 ([16]中练习 4.9)某公司进行财务审计，需要对原始凭证进行审核，该公司先后有两名出纳，由 A 出纳登记的原始凭证占 70%，B 出纳登记的原始凭证占 30%。审计人员从原始凭证中随机抽出 100 份，结果发现，由 A、B 出纳登记的原始凭证分别为 43 份和 57 份，差错分别为 1 份和 2 份。

1) 用简单随机抽样的公式估计登记原始凭证的差错率，并估计抽样标准误；

2) 用事后分层的公式估计登记原始凭证的差错率，并估计抽样标准误(有限总体校正系数$1-f\approx1$)。

解：1) 对于简单随机抽样，由理论公式，可以计算：

$$t=Z_{\alpha/2}\approx1.959964,\ a=\sum_{h=1}^{2}a_h=3,\ n=\sum_{h=1}^{2}n_h=100$$

$$p=\frac{a}{n}=0.03,\ q=1-p=0.97$$

$$v(p)=\frac{1-f}{n-1}pq\approx\frac{pq}{n-1}\approx0.0002939394,\ se(p)=\sqrt{v(p)}\approx0.01714466$$

$$L_{P_{srs}}=p-t\cdot se(p)\approx-0.003602918,\ U_{P_{srs}}=p+t\cdot se(p)\approx0.06360292$$

2) 对于事后分层抽样，由理论公式，可以计算：

$$p_{pst}=\sum_h W_hp_h\approx0.02680539$$

$$v_1(p_{pst})\approx\frac{1-f}{n}\sum_h W_hs_h^2+\frac{1}{n^2}\sum_h(1-W_h)s_h^2\approx0.0002692841,\ se_1(p_{pst})=\sqrt{v_1(p_{pst})}\approx0.01640988$$

$$L_{P_{pst,1}}=p_{pst}-t\cdot se_1(p_{pst})\approx-0.005357385,\ U_{P_{pst,1}}=p_{pst}+t\cdot se_1(p_{pst})\approx0.05896816$$

$$v_2(p_{pst})=\sum_h W_h^2(1-f_h)\frac{p_hq_h}{n_h-1}\approx0.0003194205,\ se_2(p_{pst})=\sqrt{v_2(p_{pst})}\approx0.01787234$$

$$L_{P_{pst,2}}=p_{pst}-t\cdot se_2(p_{pst})\approx-0.008223753,\ U_{P_{pst,2}}=p_{pst}+t\cdot se_2(p_{pst})\approx0.06183452$$

代入数据，调用 R 函数 Compute_P_srs_pst()进行计算，详细的 R 程序输入及输出结果请见附录 A.6。

故用简单随机抽样估计的差错率为 0.030，抽样标准误为 0.0171，95%置信区间为(−0.004，0.064)。用事后分层估计的差错率为 0.027，抽样标准误的第一种估计为 0.0164 < 0.0171，95%置信区间的第一种估计为(−0.005，0.059)，抽样标准误的第二种估计为 0.0179 > 0.0171，95%置信区间的第二种估计为(−0.008，

0.062)。

R 函数 7：Compute_nh_given_n_Y_bar_st()

对于按比例分配和尼曼分配的分层随机抽样，给定样本单位的观察值组成的矩阵 y_matrix 和样本量 n 等信息，得到计算总体均值时所需的各层样本量的 R 函数(程序) Compute_nh_given_n_Y_bar_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 7 ([16]中例 4.3)对某地区的 2850 户居民豆制品年消费支出进行调查，以居民户为抽样单位，根据收入水平将居民户划分为 4 层，每层按简单随机抽样抽取 10 户，调查获得以下数据，见表 6 所示。样本量为 $n=40$，按比例分配和尼曼分配时，各层的样本量分别应为多少？

Table 6. Annual consumption expenditure table of sample household soybean products
表 6. 样本户豆制品年消费支出表

层	居民户总数/户	样本户豆制品年消费支出/元									
		1	2	3	4	5	6	7	8	9	10
1	200	10	40	40	110	15	10	40	80	90	0
2	400	50	130	130	80	100	55	160	85	160	170
3	750	180	260	260	0	140	60	200	180	300	220
4	1500	50	35	15	0	20	30	25	10	30	25

解：对于分层随机抽样，由理论公式，可以计算：

$$L=4,\ N=\sum_{h=1}^{4}N_h=2850,\ \bar{y}_{st}=\sum_{h=1}^{4}W_h\bar{y}_h\approx 78.77193,\ \sum_{h=1}^{4}W_hs_h\approx 40.61926$$

$$W_h=\frac{N_h}{N},\ \text{即}W_1\approx 0.07018,\ W_2\approx 0.14035,\ W_3\approx 0.26316,\ W_4\approx 0.52632$$

$$\bar{y}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi},\ \text{即}\bar{y}_1=43.5,\ \bar{y}_2=112.0,\ \bar{y}_3=180.0,\ \bar{y}_4=24.0$$

$$s_h^2=\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2,\ \text{即}s_1^2\approx 1433.61,\ s_2^2\approx 1956.67,\ s_3^2\approx 8622.22,\ s_4^2\approx 193.33$$

$$W_1s_1\approx 2.6571,\ W_2s_2\approx 6.2083,\ W_3s_3\approx 24.4358,\ W_4s_4\approx 7.3181$$

按比例分配的样本量为 $n_h=nW_h$，即

$$n_1\approx 2.81,\ n_2\approx 5.61,\ n_3\approx 10.53,\ n_4\approx 21.05$$

按尼曼分配的样本量为

$$n_h=n\frac{W_hs_h}{\sum_{h=1}^{4}W_hs_h}$$

即

$$n_1\approx 2.62,\ n_2\approx 6.11,\ n_3\approx 24.06,\ n_4\approx 7.21$$

代入数据，调用 R 函数 Compute_nh_given_n_Y_bar_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.7。

因此，按比例分配确定的各层样本量为 3，6，10，21；按尼曼分配确定的各层样本量为 3，6，24，7。

R 函数 8：Compute_n_nh_Y_bar_st()

对于按比例分配和尼曼分配的分层随机抽样，给定样本单位的观察值组成的矩阵 y_matrix 及相应的精度要求，得到计算总体均值时所需的总样本量及各层样本量的 R 函数(程序) Compute_n_nh_Y_bar_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 8 ([16]中例 4.4 和例 4.3)在例 7 中，如果要求在 95%置信度下，相对允许误差不超过 10%，则按比例分配和尼曼分配时，总样本量及各层样本量分别为多少？

解：对于分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ L = 4,\ N = \sum_{h=1}^{4} N_h = 2850,\ \bar{y}_{st} = \sum_{h=1}^{4} W_h \bar{y}_h \approx 78.77193$$

$$\sum_{h=1}^{4} W_h s_h^2 \approx 2745.984,\ \sum_{h=1}^{4} W_h s_h \approx 40.61926,\ V = \left(\frac{\gamma \bar{y}_{st}}{t}\right)^2 \approx 16.15276$$

$$W_h = \frac{N_h}{N},\ \text{即} W_1 \approx 0.07018,\ W_2 \approx 0.14035,\ W_3 \approx 0.26316,\ W_4 \approx 0.52632$$

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi},\ \text{即} \bar{y}_1 = 43.5,\ \bar{y}_2 = 112.0,\ \bar{y}_3 = 180.0,\ \bar{y}_4 = 24.0$$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_h\right)^2,\ \text{即} s_1^2 \approx 1433.61,\ s_2^2 \approx 1956.67,\ s_3^2 \approx 8622.22,\ s_4^2 \approx 193.33$$

$$W_1 s_1^2 \approx 100.604,\ W_2 s_2^2 \approx 274.620,\ W_3 s_3^2 \approx 2269.006,\ W_4 s_4^2 \approx 101.754$$

$$W_1 s_1 \approx 2.657,\ W_2 s_2 \approx 6.208,\ W_3 s_3 \approx 24.436,\ W_4 s_4 \approx 7.318$$

按比例分配确定的

$$n_0 = \frac{\sum W_h s_h^2}{V} \approx 170.00$$

对 n_0 进行修正，得到修正后的样本量为

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \approx 160.43$$

且各层应分配的样本量为 $n_h = nW_h$，即

$$n_1 \approx 11.30,\ n_2 \approx 22.60,\ n_3 \approx 42.37,\ n_4 \approx 84.74$$

按尼曼分配确定的总样本量为

$$n = \frac{\left(\sum W_h s_h\right)^2}{V + \frac{\sum W_h s_h^2}{N}} \approx 96.40$$

且各层样本量为

$$n_h = n\frac{W_h s_h}{\sum_{h=1}^{4} W_h s_h}$$

即

$$n_1 \approx 6.35,\ n_2 \approx 14.83,\ n_3 \approx 58.35,\ n_4 \approx 17.48$$

代入数据，调用 R 函数 Compute_n_nh_Y_bar_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.8。

因此，按比例分配确定的总样本量为 161，各层样本量为 11，23，42，85；按尼曼分配确定的总样本量为 97，各层样本量为 6，15，58，18 (程序结果为 17，但由于总样本量需达到 97，故该层样本量需加 1 成 18)。

R 函数 9：Compute_n_nh_P_st()

对于按比例分配和尼曼分配的分层随机抽样，给定各层的总体单位数 N_h，各层的样本比例 p_h 及相应的精度要求，得到计算总体比例时所需的总样本量及各层样本量的 R 函数(程序) Compute_n_nh_P_st()。由于正文版面的限制，该 R 函数的内容及输入输出的解释放在了补充材料中。

下面我们举一个例子来说明该 R 函数的使用方法。

例 9 ([16]中例 4.5 和例 4.3)在例 4 中，如果要求在 95%的置信度下，绝对误差不超过 5%，则按比例分配和尼曼分配时，总样本量及各层样本量分别为多少？

解：对于分层随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ L = 4,\ N = \sum_{h=1}^{4} N_h = 2850,\ p_{st} = \sum_{h=1}^{4} W_h p_h = 0.2$$

$$V = \left(\frac{\Delta}{t}\right)^2 \approx 0.0006507944,\ \sum_{h=1}^{4} W_h p_h q_h \approx 0.1442105,\ \sum_{h=1}^{4} W_h \sqrt{p_h q_h} \approx 0.3710258$$

$$W_h = \frac{N_h}{N},\ \text{即} W_1 \approx 0.07018,\ W_2 \approx 0.14035,\ W_3 \approx 0.26316,\ W_4 \approx 0.52632$$

$$q_h = 1 - p_h,\ \text{即} q_1 = 0.8,\ q_2 = 0.8,\ q_3 = 0.6,\ q_4 = 0.9$$

$$W_1 p_1 \approx 0.014,\ W_2 p_2 \approx 0.028,\ W_3 p_3 \approx 0.105,\ W_4 p_4 \approx 0.053$$

$$W_1 p_1 q_1 \approx 0.011,\ W_2 p_2 q_2 \approx 0.022,\ W_3 p_3 q_3 \approx 0.063,\ W_4 p_4 q_4 \approx 0.047$$

$$W_1 \sqrt{p_1 q_1} \approx 0.028,\ W_2 \sqrt{p_2 q_2} \approx 0.056,\ W_3 \sqrt{p_3 q_3} \approx 0.129,\ W_4 \sqrt{p_4 q_4} \approx 0.158$$

按比例分配确定的

$$n_0 = \frac{\sum_{h=1}^{4} W_h p_h q_h}{V} \approx 221.59$$

对 n_0 进行修正，得到修正后的样本量为

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \approx 205.61$$

且各层样本量为 $n_h = nW_h$，即

$$n_1 \approx 14.46,\ n_2 \approx 28.91,\ n_3 \approx 54.21,\ n_4 \approx 108.42$$

按尼曼分配确定的总样本量为

$$n=\frac{\left(\sum_{h=1}^{4}W_h\sqrt{p_hq_h}\right)^2}{V+\frac{\sum_{h=1}^{4}W_hp_hq_h}{N}}\approx 196.27$$

且各层样本量为

$$n_h=n\frac{W_h\sqrt{p_hq_h}}{\sum_{h=1}^{4}W_h\sqrt{p_hq_h}}$$

即

$$n_1\approx 14.90,\ n_2\approx 29.81,\ n_3\approx 68.45,\ n_4\approx 83.84$$

代入数据，调用 R 函数 Compute_n_nh_P_st()进行计算，详细的 R 程序输入及输出结果请见附录 A.9。

因此，按比例分配所需的总样本量为 206，各层样本量为 15 (程序结果为 14，但由于总样本量需达到 206，故该层样本量需加 1 成 15)，29，54，108；按尼曼分配所需的总样本量为 197，各层样本量为 15，30，68，84。

3. 总结

本文就分层随机抽样的 R 软件实现方面自编了九个非常实用的 R 函数，分别是 Compute_Y_bar_st() (用于分层随机抽样下总体均值的点估计和区间估计)、Compute_Y_bar_prop_from_y_bar_h_s_h_st() (用于给定各层的样本均值和各层的样本标准差等信息的按比例分配的分层随机抽样下总体均值的点估计和区间估计)、Compute_Y_bar_srs_pst() (用于事后分层抽样和简单随机抽样下总体均值的点估计和区间估计)、Compute_P_st() (用于分层随机抽样下总体比例的点估计和区间估计)、Compute_P_from_a_h_st() (用于给定各层样本中具有所考虑特征的单位数等信息的分层随机抽样下总体比例的点估计和区间估计)、Compute_P_srs_pst() (用于事后分层抽样和简单随机抽样下总体比例的点估计和区间估计)、Compute_nh_given_n_Y_bar_st() (用于给定总样本量等信息的按比例分配和尼曼分配的分层随机抽样下计算总体均值时所需的各层样本量)、Compute_n_nh_Y_bar_st() (用于按比例分配和尼曼分配的分层随机抽样下计算总体均值时所需的总样本量及各层样本量)及 Compute_n_nh_P_st() (用于按比例分配和尼曼分配的分层随机抽样下计算总体比例时所需的总样本量及各层样本量)。我们相信，这九个 R 函数一定可以给利用分层随机抽样以提高估计精度进行实际问题分析的使用者提供极大的方便。

基金项目

教育部人文社会科学研究西部和边疆地区项目：基于临床试验大数据的条件势的贝叶斯无效分析的基础研究(20XJC910001)，2020.1~2022.12。国家社科基金西部项目：基于贝叶斯的八种预测势在临床试验中用于节约新药研发成本的评价研究(21XTJ001)，2021.9~2024.12。国家自然科学基金面上项目：大数据驱动的中小微企业全息风险评估与介观调控机制研究(72071019)，2021.1~2024.12。

参考文献

[1] 金勇进, 石可. 极小信息量下分层抽样的样本分配的一个案例[J]. 统计研究, 2000, 17(2): 56-60.

[2] 闫在在, 马俊玲. 分层抽样下的分别乘积估计和联合乘积估计(英文) [J]. 工程数学学报, 2001(3): 133-135+109.

[3] 苟鹏凰. 关于分层抽样的一点思考[J]. 统计研究, 2005(11): 16-17.

[4] 黄莺, 李金昌. 双重分层抽样中的校正估计[J]. 统计研究, 2008(7): 66-69.

[5] 张宁. 分层抽样下的样本轮换理论研究[J]. 统计与信息论坛, 2008, 23(4): 33-36.

[6] 王克林. 存在测量误差时分层抽样层均值方差的估计[J]. 统计与信息论坛, 2011, 26(3): 16-20.

[7] 张建军, 乔松珊. 辅以排序集样本的分层抽样方法研究[J]. 统计与信息论坛, 2012, 27(5): 14-18.

[8] 陈兵, 吕恕. 有辅助信息可利用时的分层抽样下样本轮换研究[J]. 统计与决策, 2014(15): 13-15.

[9] 李林蔓. 分层抽样下样本量的分配方法研究[J]. 统计与决策, 2015(19): 18-20.

[10] 梁敏, 刘建平. 敏感问题的分层抽样方法探讨[J]. 统计与决策, 2017(5): 12-15.

[11] 李广丽, 朱涛, 袁天, 滑瑾, 张红斌. 混合分层抽样与协同过滤的旅游景点推荐模型研究[J]. 数据采集与处理, 2019, 34(3): 566-576.

[12] 金勇进, 蒋妍, 李序颖. 抽样技术[M]. 北京: 中国人民大学出版社, 2002.

[13] 孙山泽. 抽样调查[M]. 北京: 北京大学出版社, 2004.

[14] 杜子芳. 抽样技术及其应用[M]. 北京: 清华大学出版社, 2005.

[15] 杜智敏. 抽样调查与 MATLAB 和 SPSS 应用[M]. 北京: 电子工业出版社, 2010.

[16] 李金昌. 应用抽样技术[M]. 第三版. 北京: 科学出版社, 2015.

[17] 杨贵军, 尹剑, 孟杰, 王维真. 应用抽样技术[M]. 第二版. 北京: 中国统计出版社, 2020.

[18] R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org

[19] 薛毅, 陈丽萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.

[20] 崔娅. 分层抽样的 R 软件实现[D]: [学士学位论文]. 重庆: 重庆大学, 2021.

附录

本附录给出了九个例子的详细的 R 程序输入及输出的结果。

A.1. 例 1 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>n_h = c(10, 10)
>N_h = c(250, 500)
> alpha = 0.05
> y1 = c(3, 2, 3, 4, 3, 3, 4, 5, 2, 3); y1
 [1] 3 2 3 4 3 3 4 5 2 3
> y2 = c(3, 4, 5, 5, 4, 3, 6, 2, 4, 4); y2
 [1] 3 4 5 5 4 3 6 2 4 4
>y_matrix = rbind(y1, y2); y_matrix
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
y1    3    2    3    4    3    3    4    5    2     3
y2    3    4    5    5    4    3    6    2    4     4
>
> result = Compute_Y_bar_st(y_matrix, n_h, N_h, alpha); result
$df_0
         t    N  y_bar_st  v_y_bar_st  se_y_bar_st      Delta      gamma
1 1.959964  750  3.733333  0.06708148    0.2590009  0.5076325  0.135973
   L_Y_bar  U_Y_bar
1 3.225701 4.240966
$W_h
[1] 0.3333333 0.6666667
$f_h
[1] 0.04 0.02
$y_bar_h
[1] 3.2 4.0
$s_2_h
[1] 0.8444444 1.3333333
$W_h_y_bar_h
[1] 1.066667 2.666667
```

A.2. 例 2 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>W_h = c(0.082, 0.065, 0.137, 0.056, 0.118, 0.116, 0.17, 0.098, 0.088, 0.07)
>n_h = c(16, 13, 27, 11, 24, 23, 34, 20, 18, 14)
```

```
>y_bar_h = c(89, 56, 102, 76, 97, 79, 83, 52, 36, 52)
>s_h = c(105, 74, 186, 97, 106, 89, 112, 73, 44, 65)
> f = 0 # 各层抽样比可以忽略
> alpha = 0.05
>
> result = Compute_Y_bar_prop_from_y_bar_h_s_h_st(W_h, n_h, y_bar_h, s_h, f, alpha); result
         t   n y_bar_st v_y_bar_st se_y_bar_st  L_Y_bar  U_Y_bar
1 1.959964 200   75.792   59.46035    7.711054 60.67861 90.90539
```

A.3. 例 3 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>W_h = c(0.7, 0.3)
>n_h = c(120, 80)
>y_bar_h = c(253.4, 329.4)
>s_h = c(231, 367)
> N = 8000
> alpha = 0.05
>
> result = Compute_Y_bar_srs_pst(W_h, n_h, y_bar_h, s_h, N, alpha); result
         t   n     f       s2       s y_bar_pst v_y_bar_pst
1 1.959964 200 0.025 86772.05 294.571     276.2    381.8343
  se_y_bar_pst L_Y_bar_pst U_Y_bar_pst y_bar  v_y_bar
1     19.54058    237.9012    314.4988 283.8 423.0137
  se_y_bar L_Y_bar_srs U_Y_bar_srs
1  20.5673    243.4888    324.1112
```

A.4. 例 4 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>n_h = c(10, 10, 10, 10)
>N_h = c(200, 400, 750, 1500)
> alpha = 0.05
> y1 = c(0, 0, 0, 1, 0, 0, 0, 1, 0, 0); y1
 [1] 0 0 0 1 0 0 0 1 0 0
> y2 = c(0, 1, 0, 0, 0, 0, 0, 0, 1, 0); y2
 [1] 0 1 0 0 0 0 0 0 1 0
> y3 = c(1, 1, 0, 0, 0, 0, 1, 0, 1, 0); y3
 [1] 1 1 0 0 0 0 1 0 1 0
> y4 = c(1, 0, 0, 0, 0, 0, 0, 0, 0, 0); y4
```

```
 [1] 1 0 0 0 0 0 0 0 0 0

>y_matrix = rbind(y1, y2, y3, y4); y_matrix
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
y1    0    0    0    1    0    0    0    1    0     0
y2    0    1    0    0    0    0    0    0    1     0
y3    1    1    0    0    0    0    1    0    1     0
y4    1    0    0    0    0    0    0    0    0     0
>
>result_P = Compute_P_st(y_matrix, n_h, N_h, alpha); result_P
$df_0
         t    N p_st      v_p_st    se_p_st     Delta
1 1.959964 2850  0.2 0.004998324 0.07069883 0.1385672
      gamma        L_P       U_P
1 0.6928358 0.06143284 0.3385672
$a_h
y1 y2 y3 y4
2  2  4  1
$W_h
[1] 0.07017544 0.14035088 0.26315789 0.52631579
$f_h
[1] 0.050000000 0.025000000 0.013333333 0.006666667
$p_h
 y1  y2  y3  y4
0.2 0.2 0.4 0.1
$q_h
 y1  y2  y3  y4
0.8 0.8 0.6 0.9
$v_p_h
         y1          y2          y3          y4
0.016888889 0.017333333 0.026311111 0.009933333
```

此外，我们也可以利用总体均值的分层随机抽样的点估计和区间估计的 R 函数（程序）Compute_Y_bar_st()来进行计算。

```
>result_Y_bar = Compute_Y_bar_st(y_matrix, n_h, N_h, alpha); result_Y_bar
$df_0
         t    N y_bar_st  v_y_bar_st se_y_bar_st     Delta
1 1.959964 2850       0.2 0.004998324  0.07069883 0.1385672
      gamma    L_Y_bar   U_Y_bar
1 0.6928358 0.06143284 0.3385672
$W_h
```

```
[1] 0.07017544 0.14035088 0.26315789 0.52631579
$f_h
[1] 0.050000000 0.025000000 0.013333333 0.006666667
$y_bar_h
[1] 0.2 0.2 0.4 0.1
$s2_h
[1] 0.1777778 0.1777778 0.2666667 0.1000000
$W_h_y_bar_h
[1] 0.01403509 0.02807018 0.10526316 0.05263158
```

A.5. 例 5 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>W_h = c(0.18, 0.21, 0.14, 0.09, 0.16, 0.22)
>n_h = c(30, 30, 30, 30, 30, 30)
>a_h = c(27, 28, 27, 26, 28, 29)
> N = 1650000
>N_h = N * W_h; N_h
[1] 297000 346500 231000 148500 264000 363000
>f_h = n_h / N_h; f_h
[1] 1.010101e-04 8.658009e-05 1.298701e-04 2.020202e-04
[5] 1.136364e-04 8.264463e-05
>f_h = c(0, 0, 0, 0, 0, 0); f_h # 各层抽样比可以忽略
[1] 0 0 0 0 0 0
> alpha = 0.05
> result = Compute_P_from_a_h_st(W_h, n_h, a_h, f_h, alpha); result
$df_0
        t  p_st       v_p_st    se_p_st       L_P       U_P
1 1.959964 0.924 0.0003969808 0.01992438 0.8849489 0.9630511
$p_h
[1] 0.9000000 0.9333333 0.9000000 0.8666667 0.9333333 0.9666667
$q_h
[1] 0.10000000 0.06666667 0.10000000 0.13333333 0.06666667
[6] 0.03333333
$W_h_p_h
[1] 0.1620000 0.1960000 0.1260000 0.0780000 0.1493333 0.2126667
```

A.6. 例 6 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
```

```
>W_h = c(0.7, 0.3)
>a_h = c(1, 2)
>n_h = c(43, 57)
>f_h = c(0, 0) # 各层抽样比可以忽略
> f = 0 # 抽样比可以忽略
> alpha = 0.05
>
> result = Compute_P_srs_pst(W_h, a_h, n_h, f_h, f, alpha); result
$df_srs
         t a   n    p    q          v_p       se_p      L_P_srs
1 1.959964 3 100 0.03 0.97 0.0002939394 0.01714466 -0.003602918
     U_P_srs
1 0.06360292
$df_pst
       p_pst    v_p_pst_1 se_p_pst_1    L_P_pst_1  U_P_pst_1
1 0.02680539 0.0002692841 0.01640988 -0.005357385 0.05896816
     v_p_pst_2 se_p_pst_2    L_P_pst_2  U_P_pst_2
1 0.0003194205 0.01787234 -0.008223753 0.06183452
$p_h
[1] 0.02325581 0.03508772
$q_h
[1] 0.9767442 0.9649123
$s2_h
[1] 0.02325581 0.03446115
```

A.7. 例 7 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
> n = 40
>N_h = c(200, 400, 750, 1500)
> y1 = c(10, 40, 40, 110, 15, 10, 40, 80, 90, 0); y1
 [1]  10  40  40 110  15  10  40  80  90   0
> y2 = c(50, 130, 130, 80, 100, 55, 160, 85, 160, 170); y2
 [1]  50 130 130  80 100  55 160  85 160 170
> y3 = c(180, 260, 260, 0, 140, 60, 200, 180, 300, 220); y3
 [1] 180 260 260   0 140  60 200 180 300 220
> y4 = c(50, 35, 15, 0, 20, 30, 25, 10, 30, 25); y4
 [1] 50 35 15  0 20 30 25 10 30 25
>y_matrix = rbind(y1, y2, y3, y4); y_matrix
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
```

```
y1   10   40   40  110   15   10   40   80   90    0
y2   50  130  130   80  100   55  160   85  160  170
y3  180  260  260    0  140   60  200  180  300  220
y4   50   35   15    0   20   30   25   10   30   25
>
> result = Compute_nh_given_n_Y_bar_st(y_matrix, n, N_h); result
$df_0
  L    N y_bar_st sum_W_h_s_h
1 4 2850 78.77193    40.61926
$n_h_prop
       n1        n2        n3        n4
2.807018  5.614035 10.526316 21.052632
$n_h_Neyman
       n1        n2        n3        n4
2.616548  6.113663 24.063233  7.206555
$W_h
[1] 0.07017544 0.14035088 0.26315789 0.52631579
$y_bar_h
   y1    y2    y3    y4
 43.5 112.0 180.0  24.0
$s2_h
       y1        y2        y3        y4
1433.6111 1956.6667 8622.2222  193.3333
$W_h_s_h
       y1        y2        y3        y4
2.657057  6.208312 24.435769  7.318124
```

A.8. 例 8 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
>N_h = c(200, 400, 750, 1500)
> alpha = 0.05
> gamma = 0.1
> y1 = c(10, 40, 40, 110, 15, 10, 40, 80, 90, 0); y1
 [1]  10  40  40 110  15  10  40  80  90   0
> y2 = c(50, 130, 130, 80, 100, 55, 160, 85, 160, 170); y2
 [1]  50 130 130  80 100  55 160  85 160 170
> y3 = c(180, 260, 260, 0, 140, 60, 200, 180, 300, 220); y3
 [1] 180 260 260   0 140  60 200 180 300 220
> y4 = c(50, 35, 15, 0, 20, 30, 25, 10, 30, 25); y4
```

```
 [1] 50 35 15  0 20 30 25 10 30 25
>y_matrix = rbind(y1, y2, y3, y4); y_matrix
   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
y1   10   40   40  110   15   10   40   80   90     0
y2   50  130  130   80  100   55  160   85  160   170
y3  180  260  260    0  140   60  200  180  300   220
y4   50   35   15    0   20   30   25   10   30    25
>
> result = Compute_n_nh_Y_bar_st(y_matrix, N_h, alpha, Given = "gamma", input = gamma);
result
$df_0
         t L    N y_bar_st sum_W_h_s2_h sum_W_h_s_h        V
1 1.959964 4 2850 78.77193     2745.984    40.61926 16.15276
   n0_prop   n_prop n_Neyman
1 170.0009 160.4313 96.39512
$n_h_prop
       n1        n2        n3        n4
11.29825 22.59649 42.36842 84.73684

$n_h_Neyman
       n1        n2        n3        n4
 6.34513 14.82563 58.35334 17.47590
$W_h
[1] 0.07017544 0.14035088 0.26315789 0.52631579
$y_bar_h
   y1    y2    y3    y4
 43.5 112.0 180.0  24.0
$s2_h
       y1         y2         y3         y4
1433.6111 1956.6667 8622.2222  193.3333
$W_h_s2_h
       y1         y2         y3         y4
100.6043  274.6199 2269.0058  101.7544
$W_h_s_h
       y1         y2         y3         y4
2.657057  6.208312 24.435769  7.318124
```

A.9. 例 9 的详细的 R 程序输入及输出的结果

```
>rm(list = ls(all = TRUE))
> source("subfunctions.R")
```

```
>n_h = c(10, 10, 10, 10)
>N_h = c(200, 400, 750, 1500)
>p_h = c(0.2, 0.2, 0.4, 0.1)
> alpha = 0.05
> Delta = 0.05
>
> result = Compute_n_nh_P_st(N_h, p_h, alpha, Given = "Delta", input = Delta); result
$df_0
         t L    N p_st             V sum_W_h_p_h_q_h
1 1.959964 4 2850  0.2 0.0006507944       0.1442105
  sum_W_h_sqrt_p_h_q_h  n0_prop   n_prop n_Neyman
1            0.3710258 221.5915 205.6054 196.2663
$n_h_prop
      n1        n2        n3        n4
14.45614  28.91228  54.21053 108.42105
$n_h_Neyman
      n1       n2       n3       n4
14.90415 29.80830 68.45169 83.83586
$W_h
[1] 0.07017544 0.14035088 0.26315789 0.52631579
$q_h
[1] 0.8 0.8 0.6 0.9
$W_h_p_h
[1] 0.01403509 0.02807018 0.10526316 0.05263158
$W_h_p_h_q_h
[1] 0.01122807 0.02245614 0.06315789 0.04736842
$W_h_sqrt_p_h_q_h
[1] 0.02807018 0.05614035 0.12892051 0.15789474
```