Statistics and Application 统计学与应用, 2022, 11(1), 53-66
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111007

Hans汉斯

简单随机抽样中七个非常实用的R函数

白　宸[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：2022年1月9日；录用日期：2022年1月22日；发布日期：2022年2月9日

摘　要

简单随机抽样是抽样技术中最基本、最成熟、最简单的抽样设计方式。本文利用R软件对简单随机抽样中的总体均值和总体总值的点估计和区间估计问题，样本量的确定问题，以及子总体总值均值的估计问题进行了程序实现。针对简单随机抽样，本文自编了七个非常实用的R函数(程序)：compute_Y_bar_srs()、compute_Y_srs()、compute_P_N1_srs()、compute_n0_n_Y_bar_srs()、compute_n0_n_P_srs()、compute_Y_j_srs()及compute_Y_bar_j_srs()，它们将会为需要使用简单随机抽样进行实际问题分析的使用者提供极大的方便。

关键词

简单随机抽样，总体均值和总体总值，点估计和区间估计，样本量的确定，R函数

Seven Very Practical R Functions in Simple Random Sampling

Chen Bai[1*], Ying-Ying 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: Jan. 9th, 2022; accepted: Jan. 22nd, 2022; published: Feb. 9th, 2022

Abstract

Simple random sampling is the most basic, mature, and simple sampling design method in sampling technology. In this paper, R software is used to program the point estimation and interval estimation of population mean and total value, the determination of sample size, and the estima-

[*]共同第一作者。

文章引用: 白宸, 张应应. 简单随机抽样中七个非常实用的 R 函数[J]. 统计学与应用, 2022, 11(1): 53-66.
DOI: 10.12677/sa.2022.111007

tion of total and mean value of sub population in simple random sampling. For simple random sampling, we compile seven very practical R functions (programs): compute_Y_bar_srs(), compute_Y_srs(), compute_P_N1_srs(), compute_n0_n_Y_bar_srs(), compute_n0_n_P_srs(), compute_Y_j_srs(), and compute_Y_bar_j_srs(), which will provide great convenience for users who need to use simple random sampling to analyze practical problems.

Keywords

Simple Random Sampling, Population Mean and Total Value, Point Estimation and Interval Estimation, The Determination of Sample Size, R Function

1. 引言

简单随机抽样是抽样技术[1]-[7]中最基本、最成熟、最简单的抽样设计方式。同时 R 软件[8] [9]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。本文利用 R 软件对简单随机抽样中的总体均值和总体总值的点估计和区间估计问题，样本量的确定问题，以及子总体总值均值的估计问题进行了程序实现。针对简单随机抽样，本文自编了七个非常实用的 R 函数：compute_Y_bar_srs() (用于放回和不放回简单随机抽样下总体均值的点估计和区间估计)、compute_Y_srs() (用于放回和不放回简单随机抽样下总体总值的点估计和区间估计)、compute_P_N1_srs() (用于不放回简单随机抽样下总体比例及总体中具有某种属性单位的总个数的点估计和区间估计)、compute_n0_n_Y_bar_srs() (用于放回和不放回简单随机抽样下估计总体均值的样本量确定)、compute_n0_n_P_srs() (用于放回和不放回简单随机抽样下估计总体比例的样本量确定)、compute_Y_j_srs() (用于不放回简单随机抽样下子总体总值的估计)及compute_Y_bar_j_srs() (用于不放回简单随机抽样下子总体均值的估计)。我们在对这七个 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。我们相信，这七个 R 函数将会为需要使用简单随机抽样进行实际问题分析的使用者提供极大的方便。

2. 简单随机抽样中七个非常实用的 R 函数及应用举例

我们推荐简单随机抽样中七个非常实用的 R 函数。

R 函数 1：compute_Y_bar_srs()

对于放回和不放回简单随机抽样，给定样本数据 y_vector 及其它参数，得到计算总体均值的点估计和区间估计的 R 函数(程序)如下：

```
compute_Y_bar_srs = function(y_vector, n, N, alpha, replace = c(FALSE, TRUE)){
t = qnorm(1 - alpha / 2)
f = n / N
s2 = var(y_vector)
y_bar = mean(y_vector)
```

```
if (missing(replace)) replace = FALSE

if (replace == FALSE){
    v_y_bar = (1 - f) * s2 / n
}
else{
    v_y_bar = s2 / n
}

se_y_bar = sqrt(v_y_bar)
L = y_bar - t * se_y_bar
U = y_bar + t * se_y_bar
res = data.frame(t, f, s2, y_bar, v_y_bar, se_y_bar, L, U)
}
```

此 R 函数(程序)的输入变量有：y_vector 是样本数据；n 是样本容量；N 是总体容量；alpha 是显著性水平；replace 是逻辑变量，取值为 FALSE(默认值，表示不放回抽样)和 TRUE(表示放回抽样)。

此 R 函数(程序)以数据框的形式作为输出变量：t 是抽样概率度；f 是抽样比；s2 是样本方差；y_bar 是样本均值；v_y_bar 是 y_bar 方差的无偏估计；se_y_bar 是 y_bar 标准误差的估计；L 和 U 分别是总体均值的 $1-\alpha$ 置信区间的左右端点。

下面我们举一个例子来说明 compute_Y_bar_srs()的使用方法。

例 1 ([6]中例 3.3)为调查某校大学生的电信消费水平，在全校 $N=15230$ 名学生中，用简单随机抽样的方法抽得一个 $n=36$ 的样本。对每个抽中的学生调查其上个月的电信支出金额 y_i，如表 1 所示。试以 95%的置信度估计该校大学生该月电信消费的平均支出额。

Table 1. Sample data of telecom consumption of 36 college students in a month
表 1. 36 名大学生某月电信消费的样本数据

样本序号	消费额/元	样本序号	消费额/元	样本序号	消费额/元
1	45	13	48	25	83
2	36	14	53	26	51
3	7	15	24	27	33
4	13	16	39	28	25
5	170	17	41	29	28
6	89	18	93	30	90
7	33	19	19	31	17
8	75	20	59	32	57
9	22	21	111	33	43
10	56	22	64	34	146
11	79	23	35	35	19
12	5	24	76	36	47

解：对于不放回简单随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ f = \frac{n}{N} \approx 0.002363756,\ s^2 \approx 1358.409$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 53.63889,\ v(\bar{y}) = \frac{1-f}{n}s^2 \approx 37.64438,\ se(\bar{y}) = \sqrt{v(\bar{y})} \approx 6.135502$$

$$L = \bar{y} - t \cdot se(\bar{y}) \approx 41.61353,\ U = \bar{y} + t \cdot se(\bar{y}) \approx 65.66425$$

代入数据计算如下：

```
> rm(list=ls(all=TRUE))
> source("subfunctions.R")
>
>y_vector =
+ c(45, 36, 7, 13, 170, 89, 33, 75, 22, 56, 79, 5,
+ 48, 53, 24, 39, 41, 93, 19, 59, 111, 64, 35, 76,
+ 83, 51, 33, 25, 28, 90, 17, 57, 43, 146, 19, 47)
> n = 36
> N = 15230
> alpha = 0.05
>
> #  不放回
> #  默认 replace = FALSE
>res_F = compute_Y_bar_srs(y_vector, n, N, alpha, replace = FALSE); res_F
          t           f       s2      y_barv_y_barse_y_bar
1 1.959964 0.002363756 1358.409 53.63889 37.64438 6.135502
          L         U
1 41.61353 65.66425
> res_F_1 = compute_Y_bar_srs(y_vector, n, N, alpha); res_F_1
          t           f       s2      y_barv_y_barse_y_bar
1 1.959964 0.002363756 1358.409 53.63889 37.64438 6.135502
          L         U
1 41.61353 65.66425
```

以上两种计算结果一致，说明对不放回简单随机抽样，参数 replace = FALSE 可以省略。

因此，对不放回简单随机抽样来说，该校大学生该月电信消费的平均支出额的估计为 53.64 元，可以以 95%的把握说该校大学生该月电信消费的人均支出额大约在 41.61~65.66 元之间。

对于放回简单随机抽样，由理论公式，可以计算：

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 53.63889,\ v(\bar{y}) = \frac{s^2}{n} \approx 37.73358,\ se(\bar{y}) = \sqrt{v(\bar{y})} \approx 6.142766$$

$$L = \bar{y} - t \cdot se(\bar{y}) \approx 41.59929,\ U = \bar{y} + t \cdot se(\bar{y}) \approx 65.67849$$

代入数据计算如下：

```
> # 放回
>res_T = compute_Y_bar_srs(y_vector, n, N, alpha, replace = TRUE); res_T
          t         f       s2   y_barv_y_barse_y_bar
1 1.959964 0.002363756 1358.409 53.63889 37.73358 6.142766
          L        U
1 41.59929 65.67849
```

因此，对放回简单随机抽样来说，该校大学生该月电信消费的平均支出额的估计为 53.64 元，可以以 95%的把握说该校大学生该月电信消费的人均支出额大约在 41.60~65.68 元之间。

由上述计算结果可得，不放回简单随机抽样的置信区间比放回简单随机抽样的置信区间略小一点。因为总体容量较大而样本容量较小，所以在这个例子中两者之间相差很小。

R 函数 2：compute_Y_srs()

对于放回和不放回简单随机抽样，给定样本数据 y_vector 及其它参数，得到计算总体总值的点估计和区间估计的 R 函数(程序)如下：

```
compute_Y_srs = function(y_vector, n, N, alpha, replace = c(FALSE, TRUE)){
t = qnorm(1 - alpha / 2)
f = n / N
s2 = var(y_vector)
y_bar = mean(y_vector)
Y_hat = N * y_bar
if (missing(replace)) replace = FALSE

if (replace == FALSE){
    v_y_bar = (1 - f) * s2 / n
}
else{
    v_y_bar = s2 / n
}

v_Y_hat = N^2 * v_y_bar
se_Y_hat = sqrt(v_Y_hat)
L = Y_hat - t * se_Y_hat
U = Y_hat + t * se_Y_hat
res = data.frame(t, f, s2, y_bar, v_y_bar, Y_hat, v_Y_hat, se_Y_hat, L, U)
}
```

此 R 函数(程序)的输入变量有：y_vector 是样本数据；n 是样本容量；N 是总体容量；alpha 是显著性水平；replace 是逻辑变量，取值为 FALSE (默认值，表示不放回抽样)和 TRUE (表示放回抽样)。

此 R 函数(程序)以数据框的形式作为输出变量：t 是抽样概率度；f 是抽样比；s2 是样本方差；y_bar 是样本均值；v_y_bar 是 y_bar 方差的无偏估计；Y_hat 是样本总值；v_Y_hat 是 Y_hat 方差的无偏估计；se_Y_hat 是 Y_hat 标准误差的估计；L 和 U 分别是总体总值的$1-\alpha$置信区间的左右端点。

下面我们举一个例子来说明 compute_Y_srs()的使用方法。

例 2 ([6]中例 3.4)试以 95%的置信度估计例 1 中该校大学生该月电信消费的总支出额。

解：对于不放回简单随机抽样，由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ f = \frac{n}{N} \approx 0.002363756,\ s^2 \approx 1358.409$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 53.63889,\ v(\bar{y}) = \frac{1-f}{n} s^2 \approx 37.64438$$

$$\hat{Y} \approx 816920.3,\ v(\hat{Y}) = N^2 v(\bar{y}) \approx 8731723778,\ se(\hat{Y}) = \sqrt{v(\hat{Y})} \approx 93443.69$$

$$L = \hat{Y} - t \cdot se(\hat{Y}) \approx 633774,\ U = \hat{Y} + t \cdot se(\hat{Y}) \approx 1000067$$

代入数据计算如下：

```
> rm(list=ls(all=TRUE))
> source("subfunctions.R")
>
>y_vector =
+ c(45, 36, 7, 13, 170, 89, 33, 75, 22, 56, 79, 5,
+ 48, 53, 24, 39, 41, 93, 19, 59, 111, 64, 35, 76,
+ 83, 51, 33, 25, 28, 90, 17, 57, 43, 146, 19, 47)
> n = 36
> N = 15230
> alpha = 0.05
>
> #  不放回
> #  默认 replace = FALSE
>res_F = compute_Y_srs(y_vector, n, N, alpha, replace = FALSE); res_F
          t           f       s2  y_barv_y_barY_hat
1 1.959964 0.002363756 1358.409 53.63889 37.64438 816920.3
  v_Y_hatse_Y_hat      L       U
1 8731723778 93443.69 633774 1000067
> res_F_1 = compute_Y_srs(y_vector, n, N, alpha); res_F_1
          t           f       s2  y_barv_y_barY_hat
1 1.959964 0.002363756 1358.409 53.63889 37.64438 816920.3
  v_Y_hatse_Y_hat      L       U
1 8731723778 93443.69 633774 1000067
```

以上两种计算结果一致，说明对不放回简单随机抽样，参数 replace = FALSE 可以省略。

因此，对不放回简单随机抽样来说，该校大学生该月电信消费的总支出额的估计为 816,920.3 元，可以以 95%的把握说该校大学生该月电信消费的总支出额大约在 633,774~1,000,067 元之间。

对于放回简单随机抽样，由理论公式，可以计算：

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i\approx 53.63889,\ v(\bar{y})=\frac{s^2}{n}\approx 37.73358$$

$$\hat{Y}\approx 816920.3,\ v(\hat{Y})=N^2v(\bar{y})\approx 8752412343,\ se(\hat{Y})=\sqrt{v(\hat{Y})}\approx 93554.33$$

$$L=\hat{Y}-t\cdot se(\hat{Y})\approx 633557.2,\ U=\hat{Y}+t\cdot se(\hat{Y})\approx 1000283$$

代入数据计算如下：

```
># 放回
>res_T = compute_Y_srs(y_vector, n, N, alpha, replace = TRUE); res_T
          t           f       s2      y_barv_y_barY_hat
1 1.959964 0.002363756 1358.409 53.63889 37.73358 816920.3
v_Y_hatse_Y_hat        L       U
1 8752412343 93554.33 633557.2 1000283
```

因此，对放回简单随机抽样来说，该校大学生该月电信消费的总支出额的估计为 816,920.3 元，可以以 95%的把握说该校大学生该月电信消费的总支出额大约在 633,557~1,000,283 元之间。

值得一提的是，我们的计算结果与[6]中例 3.4 的计算结果稍有不同，原因是我们的中间结果使用的是变量(比如 $\bar{y}$)的精确值，而[6]中例 3.4 的中间结果使用的是变量的四舍五入之后的值。以后的类似情况不再赘述。

R 函数 3：compute_P_N1_srs()

对于不放回简单随机抽样，得到计算总体比例及总体中具有某种属性单位的总个数的点估计和区间估计的 R 函数(程序)如下：

```
compute_P_N1_srs = function(n1, n, N, correct, alpha){
t = qnorm(1 - alpha / 2)
p = n1 / n
f = n / N
v_p = (1 - f) / (n - 1) * p * (1 - p)
se_p = sqrt(v_p)
N1_hat = N * p
v_N1_hat = N * (N - n) / (n - 1) * p * (1 - p)
se_N1_hat = sqrt(v_N1_hat)
if (correct == TRUE){
    L_P = p - (t * se_p + 1 / (2 * n))
    U_P = p + (t * se_p + 1 / (2 * n))
    L_N1 = N * L_P
    U_N1 = N * U_P
    res = data.frame(p, v_p, se_p, N1_hat, v_N1_hat, se_N1_hat, L_P, U_P, L_N1, U_N1)
}
else{
    L_P = p - (t * se_p)
    U_P = p + (t * se_p)
```

```
    L_N1 = N * L_P
    U_N1 = N * U_P
    res = data.frame(p, v_p, se_p, N1_hat, v_N1_hat, se_N1_hat, L_P, U_P, L_N1, U_N1)
  }
  }
```

此 R 函数(程序)的输入变量有：n1 是样本中具有某种属性的个数；n 是样本容量；N 是总体容量；correct 是逻辑变量，用于判断是否对近似置信区间进行连续性修正，取值为 FALSE (表示不进行连续性修正)和 TRUE (表示进行连续性修正)；alpha 是显著性水平。

此 R 函数(程序)以数据框的形式作为输出变量：p 是样本比例；v_p 是 p 方差的无偏估计；se_p 是 p 标准误差的估计；N1_hat 是总体中具有某种属性单位的总个数的简单估计量 $\widehat{N}_1$；v_N1_hat 是 $\widehat{N}_1$ 方差的无偏估计；se_N1_hat 是 $\widehat{N}_1$ 标准误差的估计；L_P 和 U_P 是总体比例的 $1-\alpha$ 置信区间的左右端点；L_N1 和 U_N1 是总体中具有某种属性单位的总个数的 $1-\alpha$ 置信区间的左右端点。

下面我们举一个例子来说明 compute_P_N1_srs()的使用方法。

例 3 ([6]中例 3.5)试以 95%的置信度估计例 1 中该校大学生该月电信消费支出超出 80 元的人数及其比例。

解：对于不放回简单随机抽样，若进行连续性修正，由理论公式，可以计算：

$$p=\frac{n_1}{n}\approx 0.1944444,\ v(p)=\frac{1-f}{n-1}pq\approx 0.00446473,\ se(p)=\sqrt{v(p)}\approx 0.06681864$$

$$\widehat{N}_1=Np\approx 2961.389,\ v\left(\widehat{N}_1\right)=\frac{N(N-n)}{n-1}pq\approx 1035607,\ se\left(\widehat{N}_1\right)=\sqrt{v\left(\widehat{N}_1\right)}\approx 1017.648$$

$$L_P=p-\left(t\cdot se(p)+\frac{1}{2n}\right)\approx 0.04959344,\ U_P=p+\left(t\cdot se(p)+\frac{1}{2n}\right)\approx 0.3392955$$

$$L_{N_1}=NL_P\approx 755.308,\ U_{N_1}=NU_P\approx 5167.47$$

代入数据计算如下：

```
> rm(list=ls(all=TRUE))
> source("subfunctions.R")
>
> # 修正后
> res = compute_P_N1_srs(n1 = 7, n = 36, N = 15230, correct = TRUE, alpha = 0.05); res
                p              v_pse_p      N1_hat v_N1_hat se_N1_hat
1 0.1944444 0.00446473 0.06681864 2961.389    1035607    1017.648
              L_P             U_P       L_N1       U_N1
1 0.04959344 0.3392955 755.308 5167.47
```

若不进行连续性修正，由理论公式，可以计算：

$$L_P=p-t\cdot se(p)\approx 0.06348232,\ U_P=p+t\cdot se(p)\approx 0.3254066$$

$$L_{N_1}=NL_P\approx 966.8358,\ U_{N_1}=NU_P\approx 4955.942$$

代入数据计算如下：

```
> #修正前
```

```
> res = compute_P_N1_srs(n1 = 7, n = 36, N = 15230, correct = FALSE, alpha = 0.05); res
          p          v_pse_p     N1_hat v_N1_hat se_N1_hat
1 0.1944444 0.00446473 0.06681864 2961.389    1035607    1017.648
          L_P          U_P        L_N1        U_N1
1 0.06348232 0.3254066 966.8358 4955.942
```

因此，对该校大学生该月电信消费支出超出 80 元的人数的估计为 2961 人，修正前可以以 95%的把握说该校大学生该月电信消费支出超出 80 元的人数大约为 967~4956 人，修正后可以以 95%的把握说该校大学生该月电信消费支出超出 80 元的人数大约为 755~5167 人。对该校大学生该月电信消费支出超出 80 元的人数比例的估计为 19.44%，修正前可以以 95%的把握说该校大学生该月电信消费支出超出 80 元的人数比例大约为 6.35%~32.54%，修正后可以以 95%的把握说该校大学生该月电信消费支出超出 80 元的人数比例大约为 4.96%~33.93%。

R 函数 4：compute_n0_n_Y_bar_srs()

对于总体均值，可以从给定方差 V 出发，得到计算样本量的 R 函数(程序)如下：

```
compute_n0_n_Y_bar_srs_from_V = function(V, S2, N){
n0 = S2 / V
n = n0 / (1 + n0 / N)
res = data.frame(n0, n)
}
```

此 R 函数(程序)的输入变量有：V 是方差；S2 是总体方差的估计；N 是总体容量。

此 R 函数(程序)以数据框的形式作为输出变量：n0 是无限总体或放回简单随机抽样情形下所需要的样本量；n 是不放回简单随机抽样所需要的样本量。

基于

$$V=\left(\frac{\Delta}{t}\right)^2=\left(\frac{\gamma\bar{Y}}{t}\right)^2=\left(CV\cdot\bar{Y}\right)^2=SE^2$$

和 compute_n0_n_Y_bar_srs_from_V()，利用 switch 语句，对于总体均值，得到计算样本量的 R 综合函数(程序)如下：

```
compute_n0_n_Y_bar_srs = function(
Given = c("V", "Delta", "gamma", "CV", "SE"),
input, alpha, Y_bar, S2, N){
t = qnorm(1 - alpha/2)

V = switch(Given,
V = input,
Delta = (input / t)^2,
gamma = (input * Y_bar / t)^2,
CV = (input * Y_bar)^2,
SE = input^2
)
```

res = compute_n0_n_Y_bar_srs_from_V(V, S2, N)

}

此R函数(程序)的输入变量有：Given是一个取值为字符串的表达式，可以是“V”,“Delta”,“gamma”，“CV”，“SE”中的一个；input是给定的精度，Given = “input”，比如Given = “V”，input = V；alpha是显著性水平；Y_bar是总体均值，当其未知时可由其估计值来代替；S2是总体方差的估计；N是总体容量。

此R函数(程序)以数据框的形式作为输出变量：n0是无限总体或放回简单随机抽样情形下所需要的样本量；n是不放回简单随机抽样所需要的样本量。

下面我们举一个例子来说明compute_n0_n_Y_bar_srs()的使用方法。

例4 ([6]中例3.6)在例1中，如果要求以95%的置信度估计该校大学生该月人均电信消费支出的绝对允许误差不超过5元，样本量应确定为多少？

解：由理论公式，可以计算：

$$n_0 = \frac{t^2 S^2}{\Delta^2} \approx 208.731,\ n = \frac{n_0}{1+\frac{n_0}{N}} \approx 205.909$$

代入数据计算如下：

> ## 给定绝对允许误差 Delta

> Delta = 5

> alpha = 0.05

> t = qnorm(1 - alpha / 2); t

[1] 1.959964

> V = (Delta / t)^2; V

[1] 6.507944

>

>res_Delta = compute_n0_n_Y_bar_srs_from_V(V, S2 = 1358.41, N = 15230); res_Delta

n0 n

1 208.731 205.909

此外，我们也可以用综合函数compute_n0_n_Y_bar_srs()来计算：

> Delta = 5

> res_Delta_1 = compute_n0_n_Y_bar_srs(Given = "Delta", input = Delta, alpha = 0.05, Y_bar = 53.64, S2 = 1358.41, N = 15230); res_Delta_1

n0 n

1 208.731 205.909

因此，以95%的置信度估计该校大学生该月人均电信消费支出的绝对允许误差不超过5元，在无限总体或放回简单随机抽样情形下，所需要的样本量为209人；在不放回简单随机抽样情形下，所需要的样本量为206人。

R函数5：compute_n0_n_P_srs()

对于总体比例，可以从给定方差V出发，得到计算样本量的R函数(程序)如下：

compute_n0_n_P_srs_from_V = function(V, P, N){

```
n0 = P * (1 - P) / V
n = n0 / (1 + (n0 - 1) / N)
res = data.frame(n0, n)
}
```

此 R 函数(程序)的输入变量有：V 是方差；P 是总体比例；N 是总体容量。

此 R 函数(程序)以数据框的形式作为输出变量：n0 是无限总体或放回简单随机抽样情形下所需要的样本量；n 是不放回简单随机抽样所需要的样本量。

基于

$$V=\left(\frac{\Delta}{t}\right)^2=\left(\frac{\gamma P}{t}\right)^2=(CV\cdot P)^2=SE^2$$

和 compute_n0_n_P_srs_from_V()，利用 switch 语句，对于总体比例，得到计算样本量的 R 综合函数(程序)如下：

```
compute_n0_n_P_srs = function(
Given = c("V", "Delta", "gamma", "CV", "SE"),
input, alpha, P, N){
t = qnorm(1 - alpha/2)

V = switch(Given,
V = input,
Delta = (input / t)^2,
gamma = (input * P / t)^2,
CV = (input * P)^2,
SE = input^2
)
res = compute_n0_n_P_srs_from_V(V, P, N)
}
```

此 R 函数(程序)的输入变量有：Given 是一个取值为字符串的表达式，可以是“V”,“Delta”,“gamma”,“CV”，“SE”中的一个；input 是给定的精度，Given = “input”，比如 Given = “V”，input = V；alpha 是显著性水平；P 是总体比例，当其未知时可由其估计值来代替；N 是总体容量。

此 R 函数(程序)以数据框的形式作为输出变量：n0 是无限总体或放回简单随机抽样情形下所需要的样本量；n 是不放回简单随机抽样所需要的样本量。

下面我们举一个例子来说明 compute_n0_n_P_srs()的使用方法。

例 5 ([6]中例 3.7)在例 1 中，如果要求以 95%的置信度估计该校大学生该月电信消费支出超出 80 元的人数比例的相对允许误差不超过 10%，样本量至少应为多少？

解：由理论公式，可以计算：

$$n_0=\frac{t^2q}{\gamma^2p}\approx 1591.913,\ n=\frac{n_0}{1+\frac{n_0-1}{N}}\approx 1441.351$$

代入数据计算如下：

```
> gamma = 0.1
> p = 0.1944
> alpha = 0.05
> t = qnorm(1 - alpha / 2); t
[1] 1.959964
> V = (gamma * p / t)^2; V
[1] 9.837763e-05
>
>res_gamma = compute_n0_n_P_srs_from_V(V, P = p, N = 15230); res_gamma
          n0        n
1 1591.913   1441.351
```

此外，我们也可以用综合函数 compute_n0_n_P_srs()来计算：

```
> gamma = 0.1
> p = 0.1944
> res_gamma_1 = compute_n0_n_P_srs(Given = "gamma", input = gamma, alpha = 0.05, P = p, N = 15230); res_gamma_1
n0        n
1 1591.913   1441.351
```

因此，以 95%的置信度估计该校大学生该月电信消费支出超出 80 元的人数比例的相对允许误差不超过 10%，在无限总体或放回简单随机抽样情形下，所需要的样本量为 1592；在不放回简单随机抽样情形下，所需要的样本量为 1442。

R 函数 6：compute_Y_j_srs()

对于简单随机抽样的子总体，当属于第 j 个子总体的单位数 N_j 未知时，得到计算子总体总值的 R 函数(程序)如下：

```
compute_Y_j_srs = function(N, n, n_j, y_bar_j, s_j, alpha){
t = qnorm(1 - alpha / 2)
p_j = n_j / n
q_j = 1 - p_j
Y_hat_j = N / n * n_j * y_bar_j
f = n / N
se_Y_hat_j = N * sqrt((1 - f) / n) * sqrt((n_j - 1) / (n - 1) * s_j^2 + n * p_j * q_j * y_bar_j^2 / (n - 1))
CV_hat_Y_hat_j = se_Y_hat_j / Y_hat_j
L_Y_j = Y_hat_j - t * se_Y_hat_j
U_Y_j = Y_hat_j + t * se_Y_hat_j
res = data.frame(Y_hat_j, se_Y_hat_j, CV_hat_Y_hat_j, L_Y_j, U_Y_j)
}
```

此 R 函数(程序)的输入变量有：N 是总体容量；n 是样本容量；n_j 是样本中属于第 j 个子总体的单位数；y_bar_j 是第 j 个子总体的样本均值；s_j 是第 j 个子总体的样本标准差；alpha 是显著性水平。

此 R 函数(程序)以数据框的形式作为输出变量：Y_hat_j 是第 j 个子总体总值的估计；se_Y_hat_j 是 Y_hat_j 标准误差的估计；CV_hat_Y_hat_j 是 Y_hat_j 变异系数的估计；L_Y_j 和 U_Y_j 分别是第 j 个子

总体总值的$1-\alpha$置信区间的左右端点。

下面我们举一个例子来说明 compute_Y_j_srs()的使用方法。

例 6 ([6]中例 3.8)某市地税局为估计餐饮业个体经营户的纳税情况，采用简单随机抽样的方法，以$N=15800$户个体经营户为总体，从中随机抽取了$n=800$户作为样本，其中$n_j=375$户属于餐饮行业，户均年纳税额$\bar{y}^{(j)}$为 4376 元，年纳税额标准差s_j为 755 元，试估计本市个体经营户中来自于餐饮业的全年纳税额$\hat{Y}^{(j)}$，并估计其标准误差$se\left(\hat{Y}^{(j)}\right)$和变异系数$\widehat{CV}\left(\hat{Y}^{(j)}\right)$。

解：由理论公式，可以计算：

$$se\left(\hat{Y}^{(j)}\right)=N\sqrt{\frac{1-f}{n}}\sqrt{\frac{n_j-1}{n-1}s_j^2+\frac{n}{n-1}p_jq_j\left[\bar{y}^{(j)}\right]^2}\approx 1222098$$

$$\hat{Y}^{(j)}=\frac{N}{n}n_j\bar{y}^{(j)}\approx 32409750,\ \widehat{CV}\left(\hat{Y}^{(j)}\right)=\frac{se\left(\hat{Y}^{(j)}\right)}{\hat{Y}^{(j)}}\approx 3.77\%$$

$$L_{Y^{(j)}}=\hat{Y}^{(j)}-t\cdot se\left(\hat{Y}^{(j)}\right)\approx 30014481,\ U_{Y^{(j)}}=\hat{Y}^{(j)}+t\cdot se\left(\hat{Y}^{(j)}\right)\approx 34805019$$

代入数据计算如下：

```
> res = compute_Y_j_srs(N = 15800, n = 800, n_j = 375, y_bar_j = 4376, s_j = 755, alpha = 0.05); res
Y_hat_jse_Y_hat_jCV_hat_Y_hat_jL_Y_jU_Y_j
1 32409750       1222098        0.03770774 30014481 34805019
```

因此，全部餐饮业个体经营户的全年纳税额为 32,409,750 元，其标准误差为 1,222,098 元，变异系数为 3.77%，并可以以 95%的把握说全部餐饮业个体经营户的全年纳税额大约为 30,014,481~34,805,019 元。

R 函数 7：compute_Y_bar_j_srs()

对于简单随机抽样的子总体，当属于第 j 个子总体的单位数 N_j 未知时，得到计算子总体均值的 R 函数(程序)如下：

```
compute_Y_bar_j_srs = function(N, n, n_j, y_bar_j, s_j, alpha){
t = qnorm(1 - alpha / 2)
f = n / N
v_y_bar_j = (1 - f) / n_j * s_j^2
se_y_bar_j = sqrt(v_y_bar_j)
L_Y_bar_j = y_bar_j - t * se_y_bar_j
U_Y_bar_j = y_bar_j + t * se_y_bar_j
res = data.frame(t, f, v_y_bar_j, se_y_bar_j, L_Y_bar_j, U_Y_bar_j)
}
```

此 R 函数(程序)的输入变量有：N 是总体容量；n 是样本容量；n_j 是样本中属于第 j 个子总体的单位数；y_bar_j 是第 j 个子总体的样本均值；s_j 是第 j 个子总体的样本标准差；alpha 是显著性水平。

此 R 函数(程序)以数据框的形式作为输出变量：t 是抽样概率度；f 是抽样比；v_y_bar_j 是 y_bar_j 方差的估计；se_y_bar_j 是 y_bar_j 标准误差的估计；L_Y_bar_j 和 U_Y_bar_j 分别是第 j 个子总体均值的 $1-\alpha$ 置信区间的左右端点。

下面我们举一个例子来说明 compute_Y_bar_j_srs()的使用方法。

例 7 (参考[6]中例 3.8)试以 95%的置信度估计例 6 中该市个体经营户中来自于餐饮业的平均纳税额。

解：由理论公式，可以计算：

$$t = Z_{\alpha/2} \approx 1.959964,\ f = \frac{n}{N} \approx 0.05063291$$

$$v\left(\bar{y}^{(j)}\right) = \frac{1-f}{n_j} s_j^2 \approx 1443.101,\ se\left(\bar{y}^{(j)}\right) = \sqrt{v\left(\bar{y}^{(j)}\right)} \approx 37.98817$$

$$L_{\bar{Y}^{(j)}} = \bar{y}^{(j)} - t \cdot se\left(\bar{y}^{(j)}\right) \approx 4301.545,\ U_{\bar{Y}^{(j)}} = \bar{y}^{(j)} + t \cdot se\left(\bar{y}^{(j)}\right) \approx 4450.455$$

代入数据计算如下：

```
> res = compute_Y_bar_j_srs(N = 15800, n = 800, n_j = 375, y_bar_j = 4376, s_j = 755, alpha = 0.05); res
        t          f v_y_bar_jse_y_bar_jL_Y_bar_jU_Y_bar_j
1 1.959964 0.05063291    1443.101     37.98817    4301.545    4450.455
```

因此，餐饮业个体经营户的平均纳税额为 4376 元，其方差的估计为 1443.10 元^2，标准误差的估计为37.99元，并可以以95%的把握说该市个体经营户中来自于餐饮业的平均纳税额大约为4301.55~4450.46元。

3. 总结

本文就简单随机抽样的 R 软件实现方面自编了七个非常实用的 R 函数，分别是 compute_Y_bar_srs() (用于放回和不放回简单随机抽样下总体均值的点估计和区间估计)、compute_Y_srs() (用于放回和不放回简单随机抽样下总体总值的点估计和区间估计)、compute_P_N1_srs() (用于不放回简单随机抽样下总体比例及总体中具有某种属性单位的总个数的点估计和区间估计)、compute_n0_n_Y_bar_srs() (用于放回和不放回简单随机抽样下估计总体均值的样本量确定)、compute_n0_n_P_srs() (用于放回和不放回简单随机抽样下估计总体比例的样本量确定)、compute_Y_j_srs() (用于不放回简单随机抽样下子总体总值的估计)及compute_Y_bar_j_srs() (用于不放回简单随机抽样下子总体均值的估计)。我们相信，这七个 R 函数一定可以给利用简单随机抽样进行实际问题分析的使用者提供极大的方便。

基金项目

本研究受教育部人文社会科学研究西部和边疆地区项目(20XJC910001)，国家社科基金西部项目(21XTJ001)和国家自然科学基金面上项目(72071019)支持。

参考文献

[1] 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.

[2] [美] G.卡尔顿, 著. 抽样调查导论[M]. 郝虹生, 译. 北京: 中国统计出版社, 2003.

[3] 孙山泽. 抽样调查[M]. 北京: 北京大学出版社, 2004.

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

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

[6] 李金昌. 应用抽样技术[M]. 第 3 版. 北京: 科学出版社, 2015.

[7] 杨贵军, 尹剑, 孟杰, 王维真. 应用抽样技术[M]. 第 2 版. 北京: 中国统计出版社, 2020.

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

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