**云南大学数学与统计学院**

**上机实验报告 （6）**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **实验课程名** | | | **抽样技术与应用实验** | | | | | | | | | **成绩** | |  | |
| **学号** |  | | | | **姓名** | | **枫叶** | **专业** | | **统计学** | **年级** | | **2021级** | | |
| **实验项目名称** | | | | **整群抽样** | | | | **日期** | |  | **实验时间** | | | |  |
| **指导教师** | | **潘东东老师** | | | | **实验地址（室）** | | | **数统学院格物楼3507** | | | | | | |
| **教师评语** | |  | | | | | | | | | | | | | |

1. **实验目的：**

1、掌握使用R软件编写抽样框的技巧；

2、在R中实现对初级单元进行PPS抽样的二阶抽样；

3、对抽出来的样本计算总体总量Y的估计。

1. **实验内容:**

P217 例9.4，将抽样框改为行列值为150\*60（即150个居委会），每个居委会所包含的户数在 [30,60] 之间（服从均匀分布），每户的值位于 [350, 6000] 之间（服从均匀分布），再假设指标值中包含7个异常值（随机挑选7个位置替换），分别为（0，0，0，12500，100500，15600，14500）。第一阶段用PPS抽样抽20个居委会，每个居委会抽6户，求 的估计及其置信度为90%的置信区间。

1. **使用环境:**

**R**

1. **实验步骤:**

library(sampling)

library(tidyverse)

m <- sample(30:60,size = 150,replace = T) #生成各居委会规模，户数为整数

M <- runif(sum(m),350,6000) #生成各户的指标值，金额可为小数

prob <- m/sum(m)

replace <- sample(1:150\*60,size = 7) #抽取替换为异常值的户

M[replace] <- c(0,0,0,12500,100500,15600,14500) #放入异常值

id <- c();idd <-c()

for (i in 1:150){

  id <- c(id,rep(i,m[i]))

  idd <- c(idd,1:m[i])

}

POP <- tibble(committee = id,

           family = idd,

           value = M)

POP[POP$committee %in% sample(1:150,60),] %>% #第一次抽样

  group\_by(committee) %>%

    sample\_n(6) %>% #第二次抽样

      group\_by(committee) %>%

      summarise(mean = mean(value)) -> ybar #各居委会均值

Ybar <- mean(ybar$mean) #总的均值

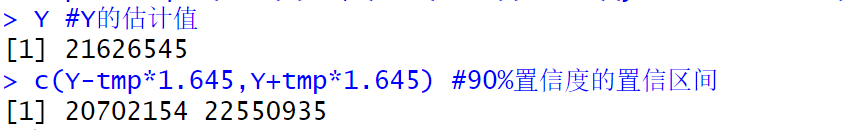
Y <- Ybar\*sum(m) #总体总值

tmp <- sqrt(sum(m)^2/(60\*(60-1))\*sum((ybar$mean-Ybar)^2)) #标准差

Y #Y的估计值

c(Y-tmp\*1.645,Y+tmp\*1.645) #90%置信度的置信区间

1. **结果**



1. **实验结果分析**