习题4.9

a. 估计总体中农场数少于600个的县的农用土地总面积（92年）

假定总体中农场数少于600个的县数未知，即子总体的 N_d 未知，对于子总体总和的估计量为

$$\hat{t}_u = N\bar{u} = N\frac{n_d}{n}\bar{y}_d = \frac{N}{n}\sum_{i\in\mathcal{S}_d} y_i$$

这一估计量的标准误为

$$\mathrm{SE}(\hat{t}_u) = N\cdot\mathrm{SE}(\bar{u}) = N\cdot\sqrt{\frac{1-f}{n}s_u^2}$$

其中

$$s_u^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(u_i-\bar{u})^2 = \frac{1}{n-1}\left[\sum_{i\in\mathcal{S}_d} y_i^2 - \frac{1}{n}(n_d\bar{y}_d)^2\right]$$

计算的代码和结果如下：

```
agsrs = read.csv("agsrs.csv")
N = 3078; n = 300
acres92_less = agsrs[agsrs$farms92<600, "acres92"]
nd_ybar_less = sum(acres92_less)
s2_u_less = (sum(acres92_less^2)-nd_ybar_less^2/n) / (n-1)
that_less = nd_ybar_less * N / n
SE_less = N * sqrt((1/n-1/N) * s2_u_less)
```

得 $n_d\bar{y}_d = 48532145$, $s_u^2 = 109710284064$, 则子总体总和的估计量为 $\hat{t}_u = 497939808$, 标准误为 $\mathrm{SE}(\hat{t}_u) = 55919525$.

b. 估计总体中农场数不少于600个的县的农用土地总面积（92年）

假定总体中农场数不少于600个的县数为1338个，即子总体的 $N_d = 1338$ ，对于子总体总和的估计量为

$$\hat{t}_{yd} = N_d\bar{y}_d = \frac{N_d}{n_d}\sum_{i\in\mathcal{S}_d} y_i$$

这一估计量的标准误为

$$\mathrm{SE}(\hat{t}_{yd}) = N_d\cdot\mathrm{SE}(\bar{y}_d) = N_d\cdot\sqrt{\frac{1-f}{n}\frac{n^2}{n_d^2}\frac{1}{n-1}\sum_{i\in\mathcal{S}_d}(y_i-\bar{y}_d)^2} = N_d\cdot\sqrt{\frac{1-f}{n_d}\frac{n}{n-1}\frac{n_d-1}{n_d}s_{yd}^2}$$

计算的代码和结果如下：

```
acres92_more = agsrs[agsrs$farms92>=600, "acres92"]
nd_ybar_more = sum(acres92_more)
N_more = 1338; n_more = length(acres92_more)
that_more = nd_ybar_more / n_more * N_more
SE_more = N_more * sqrt((1-n/N) * n/(n-1) * (n_more-1)/(n_more^2) * var(acres92_more))
```

得 $n_d = 129$, $n_d\bar{y}_d = 40836969$, 子总体总和的估计量为 $\hat{t}_{yd} = 423564841$, 标准误为 $\mathrm{SE}(\hat{t}_{yd}) = 28838198$.

习题4.16

对所抽取的SRS按照例3.2的信息进行事后分层，则92年农用土地面积总体均值的估计量为

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

其方差的估计量为

$$\hat{V}(\bar{y}_{\text{post}}) \approx \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} s_h^2$$

用事后分层方法估计总体均值，其近似的95%置信区间为

$$[\bar{y}_{\text{post}} - z_{\alpha/2}\mathrm{SE}(\bar{y}_{\text{post}}) \text{ , } \bar{y}_{\text{post}} + z_{\alpha/2}\mathrm{SE}(\bar{y}_{\text{post}})]$$

计算的代码和结果如下：

```
agpop = read.csv("agpop.csv"); selectrs = read.csv("selectrs.csv", header=F)
agsrs2 = agpop[selectrs$V5,]; N = 3078; n = 300
strat = c("NC","NE","S","W"); N_h = c(1054,220,1382,422); W_h = N_h/N
ybar_vec = c(); s2_vec = c()
for (region in strat){
  acres92_strat = agsrs2[agsrs2$region==region, "acres92"]
  ybar_vec = append(ybar_vec, mean(acres92_strat))
  s2_vec = append(s2_vec, var(acres92_strat))
}
ybar_post = sum(W_h*ybar_vec); Vhat = (1/n-1/N) * sum(W_h*s2_vec); SE_post = sqrt(Vhat)
post_CI_lb = ybar_post-qnorm(0.975)*SE_post; post_CI_ub = ybar_post+qnorm(0.975)*SE_post
```

表1：SRS事后分层后的统计量

Stratum	N_h	n_h	$\bar{y}_h$	s_h^2
North Central	1054	107	350292.01	86857007373
Northeast	220	24	71970.83	4225007609
South	1382	130	206246.35	76969408653
West	422	39	598680.59	266418735738

表1为对所抽取的SRS进行事后分层后的统计量。计算得92年农用土地面积总体均值的估计量为 $\bar{y}_{\text{post}} = 299778.1$ ，其方差的估计量为 $\hat{V}(\bar{y}_{\text{post}}) \approx 304243347$ ，标准误为 $\text{SE}(\bar{y}_{\text{post}}) = 17442.57$ ，近似95%置信区间为 $[265591.3\ ,\ 333964.9]$ 。

例2.10得到的总体均值的近似95%置信区间为 $[260706\ ,\ 335088]$ ，与之相比，本题中采用事后分层的方法得到的近似95%置信区间更窄。

习题4.17

由例4.10, $N = 940$ ，$t_x = 9.407 \times 10^6$ ，各层的统计量如表2所示：

表2：各层的统计量

Stratum	N_h	n_h	$\bar{x}_h$	s_{xh}	$\bar{y}_h$	s_{yh}	r_h
1	102	70	59549.55	64047.95	38247.80	32470.78	0.62
2	838	101	5718.84	5982.34	3833.16	5169.72	0.77

根据各层的统计量计算得

$$\bar{y}_{\text{str}} = \sum_{h=1}^{H} W_h \bar{y}_h = 7567.515\ ,\quad \bar{x}_{\text{str}} = \sum_{h=1}^{H} W_h \bar{x}_h = 11560.045$$

联合回归估计的回归系数取

$$\hat{B}_1 = \frac{\sum_{h=1}^{H} W_h^2 (1-f_h) s_{yxh}/n_h}{\sum_{h=1}^{H} W_h^2 (1-f_h) s_{xh}^2/n_h} \approx 0.5017$$

其中 $s_{yxh} = s_{yh} s_{xh} r_h$ 。

则 y 的总体均值的联合回归估计为

$$\bar{y}_{\text{reg,c}} = \bar{y}_{\text{str}} + \hat{B}_1(\bar{x}_U - \bar{x}_{\text{str}}) \approx 6788.632$$

其标准误为

$$\text{SE}(\bar{y}_{\text{reg,c}}) = \sqrt{\sum_{h=1}^{H} \frac{W_h^2(1-f_h)}{n_h}\left(s_{yh}^2 - 2\hat{B}_1 s_{yxh} + \hat{B}_1^2 s_{xh}^2\right)} \approx 351.831$$

以下是本题的计算代码：

```
N = 940; t_x = 9407000; xbar_U = t_x/N
N_h = c(102,838); n_h = c(70,101); W_h = N_h/N; f_h = n_h/N_h
xbar_h = c(59549.55,5718.84); s_xh = c(64047.95,5982.34)
ybar_h = c(38247.80,3833.16); s_yh = c(32470.78,5169.72)
r_h = c(0.62,0.77); s_yxh = s_yh * s_xh * r_h

y_str = sum(W_h*ybar_h); x_str = sum(W_h*xbar_h)
B1_numer = sum(W_h^2 * (1-f_h) * s_yxh / n_h)
B1_denom = sum(W_h^2 * (1-f_h) * s_xh^2 / n_h)
```

```
B1_hat = B1_numer / B1_denom # 联合回归估计的回归系数
ybar_regc = y_str + B1_hat * (xbar_U-x_str) # y的总体均值的联合回归估计

SE_yreg = sqrt(
  sum(W_h^2 * (1-f_h) / n_h * (s_yh^2 - 2*B1_hat*s_yxh + B1_hat^2 * s_xh^2))
)
```

习题4.42

a. 对于每一种业务类型估计卡车的2002年总里程数

首先，计算每一层的总单元数和总体的总单元数。数据集中变量 TABTRUCKS 为每一层的抽样权重，据此算出每一层的总单元数 N_h ，然后得总体单元数 $N=\sum_{h=1}^{H} N_h = 85174777$.

```
library(dplyr)
trucks = read.csv("vius.csv")
trucks_strat = trucks %>% group_by(STRATUM) %>% summarise(n_h=n())
trucks_strat$sampleWeight = unique(trucks$TABTRUCKS)
trucks_strat$N_h = round(trucks_strat$n_h * trucks_strat$sampleWeight)
N = sum(trucks_strat$N_h) # 总体单元数
trucks_strat$W_h = trucks_strat$N_h / N # 每一层的层权
```

本小题需要对14种业务类型（变量BUSINESS），估计每种业务类型中卡车在2002年的总里程数（变量MILES_ANNL）。删去BUSINESS中的缺失值，使用剩下的部分进行估计。然后，将所有样本按照变量BUSINESS分为14个子总体，每一个子总体都是一个分层抽样样本，按照分层抽样的方法估计每一个子总体中里程数的总体总和和95%置信区间。下面根据抽样权重 w_{dhj} (d 代表子总体，h 代表分层，j 代表样本单元) 构造各估计量的公式。

对于每个子总体，里程数总体总和和总体均值的估计量分别为

$$\hat{t}_{d,\text{str}} = \sum_{h=1}^{H} \sum_{j\in\mathcal{S}_{dh}} w_{dhj}\cdot y_j \ , \ \bar{y}_{d,\text{str}} = \frac{\hat{t}_{d,\text{str}}}{N_d} \ , \ \text{其中}\ N_d = \sum_{h=1}^{H} \sum_{j\in\mathcal{S}_{dh}} w_{dhj}$$

总体总和的方差估计量为

$$\hat{V}(\hat{t}_{d,\text{str}}) = \sum_{h=1}^{H} N_{dh}^2 \frac{1-f_{dh}}{n_{dh}} s_{dh}^2$$

则估计量的标准误为

$$\text{SE}(\hat{t}_{d,\text{str}}) = \sqrt{\hat{V}(\hat{t}_{d,\text{str}})} \quad , \quad \text{SE}(\bar{y}_{d,\text{str}}) = \frac{1}{N_d}\text{SE}(\hat{t}_{d,\text{str}})$$

由此可得每个子总体的总体均值和总体总和的95%置信区间。其中 $N_d = N\cdot n_d/n$, $N_{dh} = n_{dh}\cdot w_{dhj}$ ，这里 n, n_d, n_{dh} 都是去除缺失值后的值。此外，若出现 $n_{dh}=1$ ，即 s_{dh}^2 无法计算的情况，则用这一子总体中其余分层的样本方差的加权平均（权重为层权）去估计无法计算的 s_{dh}^2 .

计算代码和结果如下：

```
trucks_bus = trucks[!is.na(trucks$BUSINESS),]
n = dim(trucks_bus)[1]; busNum = 1:14
ybar_str_vec = c(); SE_y_vec = c(); ybar_CI_lower_vec = c(); ybar_CI_upper_vec = c()
t_str_vec = c(); SE_t_vec = c(); t_CI_lower_vec = c(); t_CI_upper_vec = c()

for (num in busNum){
  one_bus_df = trucks_bus[trucks_bus$BUSINESS==num,]
  one_bus_strat = one_bus_df %>% group_by(STRATUM, TABTRUCKS) %>%
    summarise(n_dh=n(), ybar_dh=mean(MILES_ANNL), s2_dh=var(MILES_ANNL))
  one_bus_strat$N_dh = one_bus_strat$n_dh * one_bus_strat$TABTRUCKS
  one_bus_strat$W_dh = one_bus_strat$N_dh / sum(one_bus_strat$N_dh)
  one_bus_strat[is.na(one_bus_strat$s2_dh),"s2_dh"]=sum(
    one_bus_strat$s2_dh*one_bus_strat$W_dh,na.rm=T)

  t_str = sum(one_bus_df$TABTRUCKS*one_bus_df$MILES_ANNL)
  ybar_str = t_str / sum(one_bus_df$TABTRUCKS)
  Vhat_t = sum(
    one_bus_strat$N_dh^2*(1/one_bus_strat$n_dh-1/one_bus_strat$N_dh)*one_bus_strat$s2_dh)
  SE_t = sqrt(Vhat_t)
  SE_ybar = SE_t / sum(one_bus_df$TABTRUCKS)
  t_CI_lower = t_str-qnorm(0.975)*SE_t; t_CI_upper = t_str+qnorm(0.975)*SE_t
  ybar_CI_lower=ybar_str-qnorm(0.975)*SE_ybar;ybar_CI_upper=ybar_str+qnorm(0.975)*SE_ybar

  ybar_str_vec = append(ybar_str_vec,round(ybar_str))
  SE_y_vec = append(SE_y_vec,round(SE_ybar))
  ybar_CI_lower_vec = append(ybar_CI_lower_vec,round(ybar_CI_lower))
  ybar_CI_upper_vec = append(ybar_CI_upper_vec,round(ybar_CI_upper))
  t_str_vec = append(t_str_vec,round(t_str))
  SE_t_vec = append(SE_t_vec,round(SE_t))
  t_CI_lower_vec = append(t_CI_lower_vec,round(t_CI_lower))
  t_CI_upper_vec = append(t_CI_upper_vec,round(t_CI_upper))
}

result_bus=data.frame(ybar_str=ybar_str_vec,SE_y=SE_y_vec,
                      y_CI_lb=ybar_CI_lower_vec,y_CI_ub=ybar_CI_upper_vec,
                      t_str=t_str_vec,SE_t=SE_t_vec,
                      t_CI_lb=t_CI_lower_vec,t_CI_ub=t_CI_upper_vec)
result_bus
```

```
##    ybar_str SE_y y_CI_lb y_CI_ub       t_str       SE_t     t_CI_lb     t_CI_ub
## 1     56452 1072   54352   58553 72272793289 1372143825 69583440810 74962145769
## 2     23306  607   22116   24496 20024589014  521506157 19002455729 21046722299
## 3     10768  443    9900   11637 24119946651  992819136 22174056901 26065836401
## 4     19210 1698   15883   22537  3411543277  301484263  2820644979  4002441575
```

```
## 5      15081  691    13726    16436 10244675655   469654524   9324169703 11165181608
## 6      16714  385    15960    17468 75906142636 1747735381 72480644235 79331641036
## 7      19650 1137    17421    21878 15384530602   890179172 13639811485 17129249719
## 8      23052  947    21195    24908 16963450921   696987138 15597381232 18329520609
## 9      17948  469    17029    18867 27470445448   717767085 26063647811 28877243084
## 10     14927 1414    12155    17700  5622014452   532694424  4577952567  6666076337
## 11     14410  779    12883    15937 10709275945   579076503  9574306855 11844245035
## 12      9537  908     7757    11317  1784083855   169891954  1451101744  2117065966
## 13     20461 1339    17837    23085  5816313888   380575301  5070400005  6562227771
## 14     16818  602    15637    17998 35776203775 1281302304 33264897407 38287510144
```

b. 对于每一种transmssn估计MPG的总体均值

本小题以4种transmssn作为子总体，每一个子总体仍为分层抽样样本，对每一个子总体估计变量MPG的总体均值。对于MPG中的缺失值，以每一层的样本均值补全；去除transmssn中的缺失值。

仍采用 (a) 问中的方法对每一个子总体估计MPG的总体均值及其95%置信区间。计算的代码和结果如下：

```
# MPG中的缺失值以每一层的样本均值补全
trucks$MPG = as.numeric(trucks$MPG)
trucks_strat = merge(trucks_strat, trucks %>% group_by(STRATUM) %>%
  summarise(mpgMean=mean(MPG,na.rm=T)), by="STRATUM")
for (strat in unique(trucks$STRATUM)){
  trucks[trucks$STRATUM==strat&is.na(trucks$MPG),"MPG"] =
    trucks_strat[trucks_strat$STRATUM==strat,"mpgMean"]
}

# 对每个子总体计算MPG的样本均值和95%置信区间
trucks_trans = trucks[!is.na(trucks$TRANSMSSN),]; n = dim(trucks_trans)[1]
transNum = 1:4
ybar_str_vec = c(); SE_y_vec = c(); ybar_CI_lower_vec = c(); ybar_CI_upper_vec = c()

for (num in transNum){
  one_trans_df = trucks_trans[trucks_trans$TRANSMSSN==num,]
  one_trans_strat = one_trans_df %>% group_by(STRATUM, TABTRUCKS) %>%
    summarise(n_dh=n(), ybar_dh=mean(MPG), s2_dh=var(MPG))
  one_trans_strat$N_dh = one_trans_strat$n_dh * one_trans_strat$TABTRUCKS
  one_trans_strat$W_dh = one_trans_strat$N_dh / sum(one_trans_strat$N_dh)
  one_trans_strat[is.na(one_trans_strat$s2_dh),"s2_dh"]=sum(
    one_trans_strat$s2_dh*one_trans_strat$W_dh,na.rm=T)

  ybar_str = sum(one_trans_df$TABTRUCKS*one_trans_df$MPG) / sum(one_trans_df$TABTRUCKS)
  Vhat_t = sum(
    one_trans_strat$N_dh^2*
      (1/one_trans_strat$n_dh-1/one_trans_strat$N_dh)*one_trans_strat$s2_dh)
```

```
  SE_t = sqrt(Vhat_t)
  SE_ybar = SE_t / sum(one_trans_df$TABTRUCKS)
  ybar_CI_lower=ybar_str-qnorm(0.975)*SE_ybar;ybar_CI_upper=ybar_str+qnorm(0.975)*SE_ybar

  ybar_str_vec = append(ybar_str_vec,round(ybar_str,4))
  SE_y_vec = append(SE_y_vec,round(SE_ybar,4))
  ybar_CI_lower_vec = append(ybar_CI_lower_vec,round(ybar_CI_lower,4))
  ybar_CI_upper_vec = append(ybar_CI_upper_vec,round(ybar_CI_upper,4))
}

result_trans=data.frame(ybar_str=ybar_str_vec,SE_y=SE_y_vec,
                        y_CI_lb=ybar_CI_lower_vec,y_CI_ub=ybar_CI_upper_vec)
result_trans
```

```
##   ybar_str   SE_y y_CI_lb y_CI_ub
## 1  16.7049 0.0368 16.6328 16.7770
## 2  15.8367 0.0756 15.6886 15.9848
## 3  14.7265 0.4773 13.7910 15.6620
## 4  16.7684 0.7304 15.3369 18.2000
```

c. 估计MILES_ANNL和MILES_LIFE之比

采用分层抽样的联合比估计，以 y 记变量MILES_ANNL，x 记变量MILES_LIFE，则比估计为

$$\hat{B} = \frac{\bar{y}_{\text{str}}}{\bar{x}_{\text{str}}}$$

估计量的标准误为

$$\text{SE}(\hat{B}) = \frac{1}{\bar{x}_{\text{str}}}\text{SE}(\bar{y}_{rc}) = \frac{1}{\bar{x}_{\text{str}}}\sqrt{\sum_{h=1}^{H}\frac{W_h^2(1-f_h)}{n_h}\left(s_{yh}^2 - 2\hat{B}s_{yxh} + \hat{B}^2 s_{xh}^2\right)}$$

其95%置信区间为

$$\left[\hat{B} - z_{\alpha/2}\text{SE}(\hat{B})\ ,\ \hat{B} + z_{\alpha/2}\text{SE}(\hat{B})\right]$$

计算的代码和结果如下：

```
trucks_strat$f_h = trucks_strat$n_h / trucks_strat$N_h
trucks_strat = merge(trucks_strat,
  trucks %>% group_by(STRATUM) %>%
    summarise(annlMean=mean(MILES_ANNL),lifeMean=mean(MILES_LIFE),s2_yh=var(MILES_ANNL),
              s2_xh=var(MILES_LIFE), s_yxh=cov(MILES_ANNL,MILES_LIFE)),
  by="STRATUM"
)
ybar_str=sum(trucks_strat$W_h*trucks_strat$annlMean)
xbar_str=sum(trucks_strat$W_h*trucks_strat$lifeMean)
```

```
Bhat = ybar_str / xbar_str
Vhat_ybar_rc = sum(trucks_strat$W_h^2 * (1-trucks_strat$f_h) / trucks_strat$n_h *
                       (trucks_strat$s2_yh-2*Bhat*trucks_strat$s_yxh+trucks_strat$s2_xh))
SE_Bhat = sqrt(Vhat_ybar_rc) / xbar_str
B_CI_lb = Bhat - qnorm(0.975)*SE_Bhat; B_CI_ub = Bhat + qnorm(0.975)*SE_Bhat
```

得 $\hat{B} = 0.1244$, $\mathrm{SE}(\hat{B}) = 0.00485$, 95%置信区间为 $[0.1149, 0.1339]$.

习题4.24

1. $E(\bar{y}_j)$, $V(\bar{y}_j)$, j=1,2

因为样本中来自子总体 $\mathcal{U}_j$ 的子样本可视为从该子总体中抽取的一个容量为 n_j 的SRS，且两个子样本相互独立（条件独立），则由SRS的性质知

$$E(\bar{y}_j) = \bar{y}_{U_j} \ , \ V(\bar{y}_j) = \frac{1-f_j}{n_j} S_j^2 \ , \ j = 1, 2$$

2. $V(\bar{y}_1 - \bar{y}_2)$

由于两个子样本相互独立，有

$$V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2) = \frac{1-f_1}{n_1} S_1^2 + \frac{1-f_2}{n_2} S_2^2$$

3. 给出 $V(\bar{y}_1 - \bar{y}_2)$ 的一个合理估计

给定 $n_j > 1$ 时，s_1^2, s_2^2 是 S_1^2, S_2^2 的无偏估计。当 N_1, N_2 已知时，取

$$\hat{V}(\bar{y}_1 - \bar{y}_2) = \frac{1 - n_1/N_1}{n_1} s_1^2 + \frac{1 - n_2/N_2}{n_2} s_2^2$$

当 N_1, N_2 未知时，可用 $N \cdot n_j / n$ 去估计 N_j , $j = 1, 2$ ，则

$$\hat{V}(\bar{y}_1 - \bar{y}_2) = \frac{1-f}{n_1} s_1^2 + \frac{1-f}{n_2} s_2^2$$

4. 构造 $\bar{y}_{1U} - \bar{y}_{2U}$ 的95%置信区间

在 $\bar{y}_1 - \bar{y}_2$ 渐近正态的假定下，$\bar{y}_{1U} - \bar{y}_{2U}$ 的95%置信区间为

$$\left[\bar{y}_1 - \bar{y}_2 - z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_1 - \bar{y}_2)} \ , \ \bar{y}_1 - \bar{y}_2 + z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_1 - \bar{y}_2)}\right]$$

5. 第9题中两个子总体均值之差的95%置信区间

当 N_1, N_2 未知时，有

```
agsrs = read.csv("agsrs.csv")
N = 3078; n = 300; f = n/N
acres92_less = agsrs[agsrs$farms92<600, "acres92"]; n1=length(acres92_less)
acres92_more = agsrs[agsrs$farms92>=600, "acres92"]; n2=length(acres92_more)
mean_diff = mean(acres92_more) - mean(acres92_less)
s2_1 = var(acres92_less); s2_2 = var(acres92_more)
Vhat_unknown = (1-f)/n1 * s2_1 + (1-f)/n2 * s2_2
unknown_CI_lb = mean_diff - qnorm(0.975) * sqrt(Vhat_unknown)
unknown_CI_ub = mean_diff + qnorm(0.975) * sqrt(Vhat_unknown)
```

得两子总体均值之差（农场数不少于600 - 农场数少于600）的点估计为 $\bar{y}_2 - \bar{y}_1 = 32751.94$，标准误为 $\mathrm{SE}(\bar{y}_2 - \bar{y}_1) = 36071.69$，95%置信区间为 $[-37947.28, 103451.16]$.

当 $N_1 = 1740, N_2 = 1338$ 时，有

```
N1 = 1740; N2 = 1338
Vhat_known = (1-n1/N1)/n1 * s2_1 + (1-n2/N2)/n2 * s2_2
known_CI_lb = mean_diff - qnorm(0.975) * sqrt(Vhat_known)
known_CI_ub = mean_diff + qnorm(0.975) * sqrt(Vhat_known)
```

得两子总体均值之差（农场数不少于600 - 农场数少于600）的点估计为 $\bar{y}_2 - \bar{y}_1 = 32751.94$，标准误为 $\mathrm{SE}(\bar{y}_2 - \bar{y}_1) = 36068.86$，95%置信区间为 $[-37941.73, 103445.60]$.