习题5.21

a. 总体直方图及描述性统计量

臭氧观测数据的总体直方图及描述性统计量如下：

```
library(knitr); library(kableExtra)
ozone = read.csv('ozone.csv'); ozone_vec = as.vector(t(ozone[,2:25]))
hist(ozone[,2:25][!is.na(ozone[,2:25])], main='Histogram of Hourly Ozone Readings',
     xlab='Hourly Ozone Readings (ppb)')
```

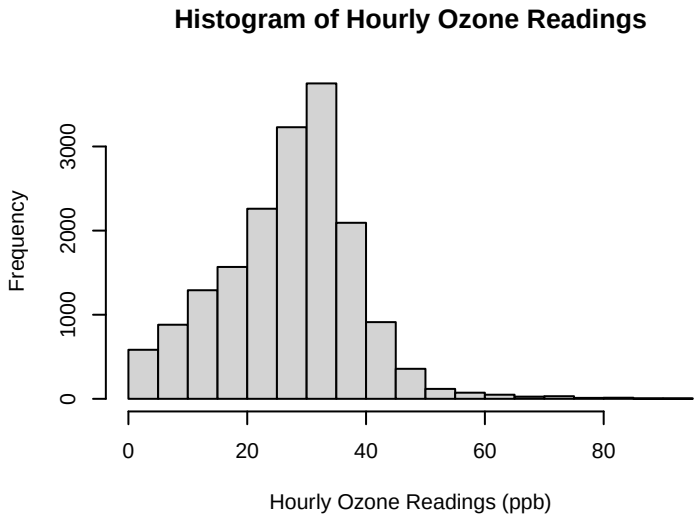

图 1: 臭氧观测数据的总体直方图

```
ozone_stats = data.frame(Mean=mean(ozone_vec,na.rm=T), Std.=sd(ozone_vec,na.rm=T),
                         Median=median(ozone_vec,na.rm=T))
kable(ozone_stats, caption='臭氧观测数据总体的描述性统计量', format='latex',
      digits=3, booktabs=TRUE) %>% kable_styling(latex_options = 'HOLD_position')
```

表 1: 臭氧观测数据总体的描述性统计量

Mean	Std.	Median
27.61	11.424	29

b. 对臭氧观测数据进行周期为24的系统抽样

从1到24中生成随机数 k , 则臭氧观测数据的系统抽样样本为GMT k 的所有观测。臭氧观测数据样本的直方图如下：

```
set.seed(1122)
k = sample(1:24, 1); ozone_sys24 = ozone[,k+1]
```

```
hist(ozone_sys24, main=paste0('Histogram of Ozone Readings at GMT',k),
     xlab=paste0('Ozone Readings at GMT',k))
```

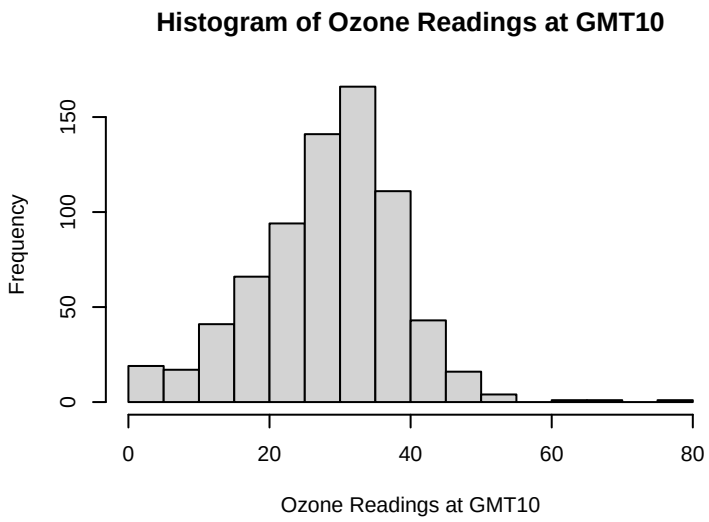

图 2: 臭氧观测数据样本的直方图

c. 臭氧观测数据系统抽样样本的描述性统计量、总体均值的区间估计

将 (b) 问中得到的系统抽样样本当作SRS，这个样本的均值、标准差和中位数如下：

```
ozone_sampleStats = data.frame(Mean=mean(ozone_sys24,na.rm=T),Std.=sd(ozone_sys24,na.rm=T),
                               Median=median(ozone_sys24,na.rm=T))
kable(ozone_sampleStats,caption='臭氧观测数据样本的描述性统计量',format='latex',
      digits=3, booktabs=TRUE) %>% kable_styling(latex_options = 'HOLD_position')
```

表 2: 臭氧观测数据样本的描述性统计量

Mean	Std.	Median
28.768	10.175	30

用这个样本构造总体均值的95%置信区间估计，为

$$\left[\bar{y} - z_{\alpha/2}\mathrm{SE}(\bar{y}) \ , \ \bar{y} + z_{\alpha/2}\mathrm{SE}(\bar{y})\right]$$

其中

$$\mathrm{SE}(\bar{y}) = \sqrt{\frac{1-f}{n}s_y^2}$$

这一区间估计的计算代码和结果如下：

```
n = dim(ozone)[1]
ybar = mean(ozone_sys24,na.rm=T); s2_y = var(ozone_sys24,na.rm=T)
SE_ybar = sqrt((1-1/24)/n * s2_y)
```

```
ybar_CI_lb = ybar - qnorm(0.975)*SE_ybar; ybar_CI_ub = ybar + qnorm(0.975)*SE_ybar
print(paste0('总体均值的95%置信区间为 [',round(ybar_CI_lb,3),' , ',
             round(ybar_CI_ub,3),']'), quote=F)
```

```
## [1] 总体均值的95%置信区间为 [28.046 , 29.491]
```

而 (a) 问中得到的总体均值真实值为27.61，则这一置信区间不包含真实的总体均值。

d. 四个周期为96的系统抽样、用整群抽样的方法估计总体均值

臭氧观测数据的总体共17520个样本。对臭氧观测数据进行周期为96的系统抽样，即总体被分成96个PSU，前48个PSU中每个PSU含183个样本，后48个PSU中每个PSU含182个样本。现抽取4个独立的系统抽样样本，即从96个PSU中抽取4个PSU，并对所抽取的每个PSU进行全面调查，这相当于单阶整群抽样。

进行系统抽样的代码如下：

```
ozone_sys96 = matrix(c(ozone_vec,rep(NA,48)), nrow=183, ncol=96, byrow=T)
set.seed(12345)
PSU_k = sample(1:96, 4, replace=T)
ozone_sys96_PSU4 = ozone_sys96[,PSU_k]
print(paste0('所抽取的4个PSU为 ', paste(PSU_k,collapse=',')), quote=F)
```

```
## [1] 所抽取的4个PSU为 14,51,80,90
```

考虑每个PSU都有183个样本，最后不足的数据用NA值填充，则问题转化为群规模相同的单阶整群抽样。有 $N=96, n=4, M=183$.

总体均值的估计为

$$\hat{\bar{y}} = \frac{\hat{t}}{NM} = \frac{1}{n}\sum_{i\in\mathcal{S}} \bar{y}_{iU} = 28.615$$

估计量的标准误为

$$\mathrm{SE}(\hat{\bar{y}}) = \frac{1}{M}\sqrt{\frac{1-f}{n}s_t^2} = \sqrt{\frac{1-f}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\bar{y}_{iU}-\hat{\bar{y}})^2} = 2.195$$

则总体均值的95%置信区间为

$$\left[\hat{\bar{y}} - z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}) \ , \ \hat{\bar{y}} + z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}})\right] = [24.313 \ , \ 32.917]$$

以下是相关计算代码：

```
N = 96; n = 4; M = 183; f = n/N
y_iU = colMeans(ozone_sys96_PSU4, na.rm=T)
y_barhat = mean(y_iU)
SE_y_barhat = sqrt((1-f)/n * var(y_iU))
y_barhat_CI_lb = y_barhat - qnorm(0.975) * SE_y_barhat
y_barhat_CI_ub = y_barhat + qnorm(0.975) * SE_y_barhat
```

补充题

1. 总体均值估计量的计算及性质

采用二阶整群抽样法对总体进行抽样，有 $n=2, m=2, N=4, M=3$ ，则 $f=n/N=0.5$ ，$f_{2i}\equiv f_2=m/M=2/3$. 以下是抽样和估计量计算的代码：

```
supple = read.csv('supple.csv')
n=2; m=2; N=4; M=3; f=n/N; f2=m/M
set.seed(1111); PSU=sample(1:N, n)
SSU=matrix(rep(0,4),nrow=2,ncol=2,dimnames=list(paste0('PSU',PSU),c('SSU_k1','SSU_k2')))
set.seed(2222)
for (i in PSU){
  set.seed(sample(1:1000,1))
  SSU_num = sample(1:M, m)
  SSU[paste0('PSU',i), ] = as.vector(t(supple[i, SSU_num]))
}
y_barhat = sum(SSU) / (n*m)
kable(SSU, caption='采用二阶整群抽样法抽取的样本', format='latex', booktabs=TRUE) %>%
  kable_styling(latex_options = 'HOLD_position')
```

表 3: 采用二阶整群抽样法抽取的样本

	SSU_k1	SSU_k2
PSU4	2	3
PSU2	9	12

根据表3计算得

$$\hat{\bar{y}}=\frac{1}{nm}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}y_{ij}=6.5$$

根据课件p22，有

$$\hat{t}_{\text{unb}}=\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{N}{n}\frac{M}{m}y_{ij}=NM\cdot\hat{\bar{y}}$$

根据 $\hat{t}_{\text{unb}}$ 的三条性质及 $\hat{V}_{WR}(\hat{t}_{\text{unb}})$，可以推出 $\hat{\bar{y}}$ 的相应性质：

(1) $\hat{t}_{\text{unb}}$ 是 t 的无偏估计，有 $\hat{\bar{y}}=\dfrac{\hat{t}_{\text{unb}}}{NM}$, $\bar{y}_U=\dfrac{t}{NM}$, 则 $\hat{\bar{y}}$ 是 $\bar{y}_U$ 的无偏估计;

(2) $V(\hat{\bar{y}})=\dfrac{1}{N^2M^2}V(\hat{t}_{\text{unb}})=\dfrac{1-f}{M^2n}S_t^2+\dfrac{1-f_2}{nmN}\sum\limits_{i=1}^{N}S_i^2$;

(3) 方差的无偏估计：$\hat{V}(\hat{\bar{y}})=\dfrac{1}{N^2M^2}\hat{V}(\hat{t}_{\text{unb}})=\dfrac{1-f}{M^2n}s_t^2+\dfrac{1-f_2}{nmN}\sum\limits_{i\in\mathcal{S}}s_i^2$, $\text{SE}(\hat{\bar{y}})=\sqrt{\hat{V}(\hat{\bar{y}})}$;

(4) $\hat{V}_{WR}(\hat{\bar{y}})=\dfrac{1}{N^2M^2}\hat{V}_{WR}(\hat{t}_{\text{unb}})=\dfrac{s_t^2}{M^2n}$.

2. 总体参数及估计量方差的计算

根据以下的计算代码

```
supple_vec = as.vector(t(supple)); t_i = rowSums(supple)
ybar_U = mean(supple_vec); S2 = var(supple_vec); S2_t = var(t_i)
S2_i = apply(supple, 1, var)
V_y_barhat = (1-f)/(M^2 * n) * S2_t + (1-f2)/(n*m*N) * sum(S2_i)
```

得

$$\bar{y}_U = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = 7.5$$

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2 = 13$$

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_i-\bar{t}_U)^2 = 27.67$$

由 $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2$ ，得 $S_1^2 = 17.33$ ， $S_2^2 = 9$ ， $S_3^2 = 12.33$ ， $S_4^2 = 19$.

此外

$$V(\hat{\bar{y}}) = \frac{1-f}{M^2 n}S_t^2 + \frac{1-f_2}{nmN}\sum_{i=1}^{N}S_i^2 = 1.970$$

3. 所有二阶整群样本

二阶正群样本共有 $\binom{4}{2}\times\binom{3}{2}^2 = 6\times 3^2 = 54$ 个。以下是由代码生成的所有二阶整群样本及其统计量：

```
library(stringr)
PSU=c(); SSU=c(); Sample=c()
ybar_hat=c(); s2_t = c(); s2_1=c(); s2_2=c(); s2_3=c(); s2_4=c()
Vhat_ybar_hat=c(); Vhat_WR_y_barhat=c()

get_PSU_char = function(vec){
  return(paste0('[', vec[1], ',', vec[2], ']'))
}
get_SSU_char = function(vec){
  return(paste0('[(',vec[1],',',vec[2],'),(',vec[3],',',vec[4],')]'))
}
get_sample_char = function(vec){
  vec_char = sapply(vec, str_pad, 2, 'left')
  return(paste0('[(',vec_char[1],',',vec_char[2],'),(',vec_char[3],',',vec_char[4],')]'))
}
calc_s2_t = function(vec){
  t_i = c(mean(vec[1:2]) * 3, mean(vec[3:4]) * 3); s2_t = var(t_i)
  return(s2_t)
}
```

```
calc_Vhat_ybar = function(s2_t,s2_i_list){
  Vhat_ybar = (1-f)/(M^2*n) * s2_t + (1-f2)/(n*m*N) * sum(s2_i_list,na.rm=T)
  return(Vhat_ybar)
}

for (v1 in 1:dim(t(combn(N,n)))[1]){
  vec_i = t(combn(N,n))[v1,]
  PSU_sample = supple[vec_i,]
  for (v2 in 1:dim(t(combn(M,m)))[1]){
    vec_j1 = t(combn(M,m))[v2,]
    for (v3 in 1:dim(t(combn(M,m)))[1]){
      vec_j2 = t(combn(M,m))[v3,]; SSU_nums = c(vec_j1, vec_j2)
      SSU_sample_vec = c(t(PSU_sample[1,vec_j1]), t(PSU_sample[2,vec_j2]))
      PSU = c(PSU, get_PSU_char(vec_i)); SSU = c(SSU, get_SSU_char(SSU_nums))
      Sample = c(Sample, get_sample_char(SSU_sample_vec))

      ybar_hat = c(ybar_hat, mean(SSU_sample_vec))
      s2_t_value = calc_s2_t(SSU_sample_vec)
      s2_t = c(s2_t, s2_t_value)
      s2_i_list = rep(NA,4)
      s2_i_list[vec_i] = c(var(SSU_sample_vec[1:2]),var(SSU_sample_vec[3:4]))
      s2_1 = c(s2_1,s2_i_list[1]); s2_2 = c(s2_2,s2_i_list[2])
      s2_3 = c(s2_3,s2_i_list[3]); s2_4 = c(s2_4,s2_i_list[4])
      Vhat_ybar_hat = c(Vhat_ybar_hat, calc_Vhat_ybar(s2_t_value,s2_i_list))
      Vhat_WR_y_barhat = c(Vhat_WR_y_barhat, s2_t_value/(M^2*n))
    }
  }
}

SampleNum = 1:54

result = data.frame(cbind(SampleNum, PSU, SSU, Sample, ybar_hat, s2_t, s2_1, s2_2,
                          s2_3, s2_4, Vhat_ybar_hat, Vhat_WR_y_barhat))
res_print = data.frame(cbind(SampleNum, PSU, SSU, Sample, ybar_hat, s2_t, s2_1, s2_2, s2_3,
                             s2_4, round(Vhat_ybar_hat,2), round(Vhat_WR_y_barhat,2)))
colnames(res_print) = c("样本","PSU编号","SSU编号","样本数据",r"{$\hat{\bar{y}}$}",
                        "$s_t^2$","$s_1^2$","$s_2^2$","$s_3^2$","$s_4^2$",
                        r"{$\hat{V}(\hat{\bar{y}})$}",r"{$\hat{V}_{WR}(\hat{\bar{y}})$}")
opts = options(knitr.kable.NA = "/")
res_print %>%
  kable(format='latex', caption='所有二阶整群样本及其统计量', escape=F, booktabs=T,
        longtable=T) %>%
  kable_styling(latex_options='HOLD_position')
```

表 4: 所有二阶整群样本及其统计量

样本	PSU编号	SSU编号	样本数据	$\hat{\bar{y}}$	s_t^2	s_1^2	s_2^2	s_3^2	s_4^2	$\hat{V}(\hat{\bar{y}})$	$\hat{V}_{WR}(\hat{\bar{y}})$
1	[1,2]	[(1,2),(1,2)]	[(5,13),(6, 9)]	8.25	10.125	32	4.5	/	/	1.04	0.56
2	[1,2]	[(1,2),(1,3)]	[(5,13),(6,12)]	9	0	32	18	/	/	1.04	0
3	[1,2]	[(1,2),(2,3)]	[(5,13),(9,12)]	9.75	10.125	32	4.5	/	/	1.04	0.56
4	[1,2]	[(1,3),(1,2)]	[(5, 7),(6, 9)]	6.75	10.125	2	4.5	/	/	0.42	0.56
5	[1,2]	[(1,3),(1,3)]	[(5, 7),(6,12)]	7.5	40.5	2	18	/	/	1.54	2.25
6	[1,2]	[(1,3),(2,3)]	[(5, 7),(9,12)]	8.25	91.125	2	4.5	/	/	2.67	5.06
7	[1,2]	[(2,3),(1,2)]	[(13, 7),(6, 9)]	8.75	28.125	18	4.5	/	/	1.25	1.56
8	[1,2]	[(2,3),(1,3)]	[(13, 7),(6,12)]	9.5	4.5	18	18	/	/	0.88	0.25
9	[1,2]	[(2,3),(2,3)]	[(13, 7),(9,12)]	10.25	1.125	18	4.5	/	/	0.5	0.06
10	[1,3]	[(1,2),(1,2)]	[(5,13),(11, 8)]	9.25	1.125	32	/	4.5	/	0.79	0.06
11	[1,3]	[(1,2),(1,3)]	[(5,13),(11, 4)]	8.25	10.125	32	/	24.5	/	1.46	0.56
12	[1,3]	[(1,2),(2,3)]	[(5,13),(8, 4)]	7.5	40.5	32	/	8	/	1.96	2.25
13	[1,3]	[(1,3),(1,2)]	[(5, 7),(11, 8)]	7.75	55.125	2	/	4.5	/	1.67	3.06
14	[1,3]	[(1,3),(1,3)]	[(5, 7),(11, 4)]	6.75	10.125	2	/	24.5	/	0.83	0.56
15	[1,3]	[(1,3),(2,3)]	[(5, 7),(8, 4)]	6	0	2	/	8	/	0.21	0
16	[1,3]	[(2,3),(1,2)]	[(13, 7),(11, 8)]	9.75	1.125	18	/	4.5	/	0.5	0.06
17	[1,3]	[(2,3),(1,3)]	[(13, 7),(11, 4)]	8.75	28.125	18	/	24.5	/	1.67	1.56
18	[1,3]	[(2,3),(2,3)]	[(13, 7),(8, 4)]	8	72	18	/	8	/	2.54	4
19	[1,4]	[(1,2),(1,2)]	[(5,13),(3,10)]	7.75	28.125	32	/	/	24.5	1.96	1.56
20	[1,4]	[(1,2),(1,3)]	[(5,13),(3, 2)]	5.75	190.125	32	/	/	0.5	5.96	10.56
21	[1,4]	[(1,2),(2,3)]	[(5,13),(10, 2)]	7.5	40.5	32	/	/	32	2.46	2.25
22	[1,4]	[(1,3),(1,2)]	[(5, 7),(3,10)]	6.25	1.125	2	/	/	24.5	0.58	0.06
23	[1,4]	[(1,3),(1,3)]	[(5, 7),(3, 2)]	4.25	55.125	2	/	/	0.5	1.58	3.06
24	[1,4]	[(1,3),(2,3)]	[(5, 7),(10, 2)]	6	0	2	/	/	32	0.71	0
25	[1,4]	[(2,3),(1,2)]	[(13, 7),(3,10)]	8.25	55.125	18	/	/	24.5	2.42	3.06
26	[1,4]	[(2,3),(1,3)]	[(13, 7),(3, 2)]	6.25	253.125	18	/	/	0.5	7.42	14.06
27	[1,4]	[(2,3),(2,3)]	[(13, 7),(10, 2)]	8	72	18	/	/	32	3.04	4
28	[2,3]	[(1,2),(1,2)]	[(6, 9),(11, 8)]	8.5	18	/	4.5	4.5	/	0.69	1
29	[2,3]	[(1,2),(1,3)]	[(6, 9),(11, 4)]	7.5	0	/	4.5	24.5	/	0.6	0
30	[2,3]	[(1,2),(2,3)]	[(6, 9),(8, 4)]	6.75	10.125	/	4.5	8	/	0.54	0.56
31	[2,3]	[(1,3),(1,2)]	[(6,12),(11, 8)]	9.25	1.125	/	18	4.5	/	0.5	0.06
32	[2,3]	[(1,3),(1,3)]	[(6,12),(11, 4)]	8.25	10.125	/	18	24.5	/	1.17	0.56
33	[2,3]	[(1,3),(2,3)]	[(6,12),(8, 4)]	7.5	40.5	/	18	8	/	1.67	2.25
34	[2,3]	[(2,3),(1,2)]	[(9,12),(11, 8)]	10	4.5	/	4.5	4.5	/	0.31	0.25
35	[2,3]	[(2,3),(1,3)]	[(9,12),(11, 4)]	9	40.5	/	4.5	24.5	/	1.73	2.25
36	[2,3]	[(2,3),(2,3)]	[(9,12),(8, 4)]	8.25	91.125	/	4.5	8	/	2.79	5.06
37	[2,4]	[(1,2),(1,2)]	[(6, 9),(3,10)]	7	4.5	/	4.5	/	24.5	0.73	0.25
38	[2,4]	[(1,2),(1,3)]	[(6, 9),(3, 2)]	5	112.5	/	4.5	/	0.5	3.23	6.25
39	[2,4]	[(1,2),(2,3)]	[(6, 9),(10, 2)]	6.75	10.125	/	4.5	/	32	1.04	0.56

40	[2,4]	[(1,3),(1,2)]	[(6,12),(3,10)]	7.75	28.125	/	18	/	24.5	1.67	1.56
41	[2,4]	[(1,3),(1,3)]	[(6,12),(3, 2)]	5.75	190.125	/	18	/	0.5	5.67	10.56
42	[2,4]	[(1,3),(2,3)]	[(6,12),(10, 2)]	7.5	40.5	/	18	/	32	2.17	2.25
43	[2,4]	[(2,3),(1,2)]	[(9,12),(3,10)]	8.5	72	/	4.5	/	24.5	2.6	4
44	[2,4]	[(2,3),(1,3)]	[(9,12),(3, 2)]	6.5	288	/	4.5	/	0.5	8.1	16
45	[2,4]	[(2,3),(2,3)]	[(9,12),(10, 2)]	8.25	91.125	/	4.5	/	32	3.29	5.06
46	[3,4]	[(1,2),(1,2)]	[(11, 8),(3,10)]	8	40.5	/	/	4.5	24.5	1.73	2.25
47	[3,4]	[(1,2),(1,3)]	[(11, 8),(3, 2)]	6	220.5	/	/	4.5	0.5	6.23	12.25
48	[3,4]	[(1,2),(2,3)]	[(11, 8),(10, 2)]	7.75	55.125	/	/	4.5	32	2.29	3.06
49	[3,4]	[(1,3),(1,2)]	[(11, 4),(3,10)]	7	4.5	/	/	24.5	24.5	1.15	0.25
50	[3,4]	[(1,3),(1,3)]	[(11, 4),(3, 2)]	5	112.5	/	/	24.5	0.5	3.65	6.25
51	[3,4]	[(1,3),(2,3)]	[(11, 4),(10, 2)]	6.75	10.125	/	/	24.5	32	1.46	0.56
52	[3,4]	[(2,3),(1,2)]	[(8, 4),(3,10)]	6.25	1.125	/	/	8	24.5	0.71	0.06
53	[3,4]	[(2,3),(1,3)]	[(8, 4),(3, 2)]	4.25	55.125	/	/	8	0.5	1.71	3.06
54	[3,4]	[(2,3),(2,3)]	[(8, 4),(10, 2)]	6	0	/	/	8	32	0.83	0

以下根据上表中列出的所有二阶整群样本进行估计量性质的验证：

(1) $E(\hat{\bar{y}}) = \dfrac{1}{54}\sum_{k=1}^{54}\hat{\bar{y}}_k = 7.5$, 又 $\bar{y}_U = 7.5$, 则 $\hat{\bar{y}}$ 是 $\bar{y}_U$ 的无偏估计;

(2) $E\left[\hat{V}(\hat{\bar{y}})\right] = \dfrac{1}{54}\sum_{k=1}^{54}\hat{V}_k(\hat{\bar{y}}) = 1.970$, 又 $V(\hat{\bar{y}}) = 1.970$, 则 $\hat{V}(\hat{\bar{y}})$ 是 $V(\hat{\bar{y}})$ 的无偏估计;

(3) $\hat{V}(\hat{\bar{y}})$ 的方差为 3.086，并且是 $V(\hat{\bar{y}})$ 的无偏估计。而 $\hat{V}_{WR}(\hat{\bar{y}})$ 的方差为 13.608，期望为 2.738，说明它是 $V(\hat{\bar{y}})$ 的有偏估计。所以，用 $\hat{V}(\hat{\bar{y}})$ 来估计 $V(\hat{\bar{y}})$ 更好。

(4) $E(s_t^2) = \dfrac{1}{54}\sum_{k=1}^{54}s_{t,k}^2 = 49.292$, 又 $S_t^2 = 27.67$, 则 s_t^2 不是 S_t^2 的无偏估计.

下面是相关的计算代码：

```
E_ybar_hat = mean(ybar_hat)
E_Vhat_ybar_hat = mean(Vhat_ybar_hat)
var_Vhat = var(Vhat_ybar_hat)
E_Vhat_WR = mean(Vhat_WR_y_barhat); var_Vhat_WR = var(Vhat_WR_y_barhat)
E_s2_t = mean(s2_t)
```