习题5.4

a. 说明题设中的抽样方法是整群抽样

本题以学术期刊为抽样框，从1285本学术期刊中抽取26本，然后调查抽中的学术期刊中的所有学术文章。所有，学术期刊是初级单元，学术文章是次级单元，抽样方法为单阶整群抽样。

b. 估计使用非概率抽样的学术文章的比例

记 t_i 为第 i 本学术期刊中使用非概率抽样的学术文章数量，M_i 为第 i 本学术期刊中使用抽样调查方法的学术文章数量，则使用非概率抽样的学术文章比例的估计量为

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{137}{148} = 0.9257$$

估计量的标准差为

$$\text{SE}(\hat{\bar{y}}_r) = \sqrt{\frac{1-f}{n\bar{M}^2}\frac{1}{n-1}\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2} = 0.0340$$

以下为计算的代码：

```
journal = read.csv("journal.csv")
n = 26; N = 1285; f = n/N; Mbar = mean(journal$numemp)
yr = sum(journal$nonprob) / sum(journal$numemp)
SE_yr = sqrt((1-f)/(n*Mbar^2) * 1/(n-1) * sum((journal$nonprob-yr*journal$numemp)^2))
```

c. 评述作者的观点：接受使用非概率抽样的学术文章

从上述估计量可以得出结论：使用非概率抽样的学术文章占学术文章的绝大多数（92.57%），因此作者的观点有一定的合理性。然而在计算上述估计量时，我们假定了作者抽取PSU时使用了SRS，但题目并未告诉我们作者使用的是SRS，所以上述的估计量可能存在一定偏差。

习题5.5

a. 估计明年想去西班牙语国家旅行的学生总数

记 y_{ij} 为第 i 个班级的第 j 位学生是否想去西班牙旅游（0-1变量），t_i 为第 i 个班级中想去西班牙国家旅游的学生数，M_i 为第 i 个班级的总人数，有 $t_i = \sum_{j=1}^{M_i} y_{ij}$ 。题目中未告诉72个班级的总人数 M_0 ，故使用无偏估计，总体总和的估计量为

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i\in\mathcal{S}} t_i = 453.6 \approx 454$$

估计量的标准误为

$$\text{SE}(\hat{t}_{\text{unb}}) = N\sqrt{\frac{1-f}{n}s_t^2} = 72 \times \sqrt{\left(\frac{1}{10} - \frac{1}{72}\right) \times 28.01} = 111.822$$

则明年想去西班牙语国家旅行的学生总数的95%置信区间为

$$\left[\hat{t}_{\text{unb}} - z_{\alpha/2}\text{SE}(\hat{t}_{\text{unb}}) \ , \ \hat{t}_{\text{unb}} + z_{\alpha/2}\text{SE}(\hat{t}_{\text{unb}})\right] = \Big[234.432, 672.767\Big]$$

以下为计算的代码：

```
library(dplyr)
spanish = read.csv("spanish.csv")
n = 10; N = 72; f = n/N
t_unb = N/n * sum(spanish$trip)
spanish_group = spanish %>% group_by(class) %>%
  summarise(M_i=n(), t_i=sum(trip), mean_score=mean(score))
s2_t = var(spanish_group$t_i)
SE_t = N * sqrt((1-f)/n * s2_t)
CI_t_lb = t_unb - qnorm(0.975) * SE_t; CI_t_ub = t_unb + qnorm(0.975) * SE_t
```

b. 估计学生西班牙语测试成绩的均分

记 t_i 为第 i 个班级的西班牙语测试总成绩，M_i 为第 i 个班级的总人数，则 t_i 与 M_i 有一定的正相关关系，使用比估计效果更好。西班牙语测试成绩均分的估计量为

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{13092}{196} = 66.80$$

估计量的标准误为

$$\text{SE}(\hat{\bar{y}}_r) = \sqrt{\frac{1-f}{n\bar{M}^2}\frac{1}{n-1}\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2} = 2.709$$

则西班牙语测试成绩均分的95%置信区间为

$$\left[\hat{\bar{y}}_r - z_{\alpha/2}\text{SE}(\hat{\bar{y}}_r) \ , \ \hat{\bar{y}}_r + z_{\alpha/2}\text{SE}(\hat{\bar{y}}_r)\right] = [61.486, 72.106]$$

以下为计算的代码：

```
yr = sum(spanish$score) / sum(spanish_group$M_i)
spanish_group$sum_score = spanish_group$M_i * spanish_group$mean_score
Mbar = mean(spanish_group$M_i)
SE_yr = sqrt((1-f)/(n*Mbar^2) * 1/(n-1) *
               sum((spanish_group$sum_score-yr*spanish_group$M_i)^2))
CI_y_lb = yr - qnorm(0.975) * SE_yr; CI_y_ub = yr + qnorm(0.975) * SE_yr
```

习题5.36

a. 以team为PSU进行单阶整群抽样

数据集 `baseball` 中共有30支队伍，每支队伍中约有25人，若以队伍为PSU进行单阶整群抽样，抽取容量约为150的样本，则要采用SRS从30支队伍中随机抽取6支队伍，并且以每支队伍中的所有数据为分析对象。以下为单阶整群抽样的代码：

```
library(SDAResources); data("baseball")
baseball_teamPop = baseball %>% group_by(team) %>%
  summarise(M_i=n(), mean_logsal=mean(log(salary)), sum_pitcher=sum(pos=="P"))
set.seed(1234)
baseball_teamSpl = baseball_teamPop[sample(1:30,6),]
```

由此抽取的6支队伍为 TBA, MIL, SFN, PHI, BOS, FLO, 样本容量为 $\sum_{i\in\mathcal{S}} M_i = 157$.

b. 画样本中 logsal 的分组箱线图

以 team 为组别，样本中 logsal 分组箱线图的作图代码和结果如下所示：

```
library(ggplot2)
baseball_spl = baseball[baseball$team %in% baseball_teamSpl$team,]
baseball_spl$logsal = log(baseball_spl$salary)
ggplot(baseball_spl, aes(x=team,y=logsal,color=team)) + geom_boxplot()
```

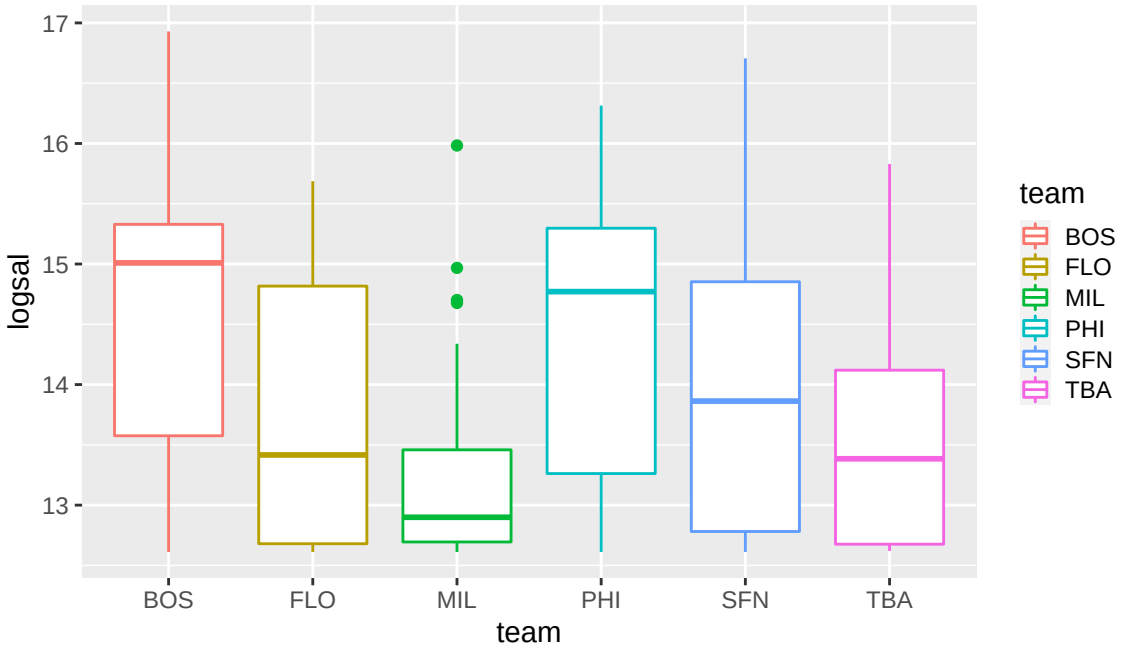

图 1: logsal的分组箱线图

c. 估计 logsal 的总体均值

采用比估计来估计 logsal 的总体均值，有

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{2193.035}{157} = 13.968$$

估计量的标准误为

$$\text{SE}(\hat{\bar{y}}_r) = \sqrt{\frac{1-f}{n\bar{M}^2}\frac{1}{n-1}\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2} = 0.196$$

则 logsal 总体均值的95%置信区间为

$$\left[\hat{\bar{y}}_r - z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_r)\ ,\ \hat{\bar{y}}_r + z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_r)\right] = [13.584, 14.353]$$

以下为计算的代码：

```
baseball_teamSpl$sum_logsal = baseball_teamSpl$M_i * baseball_teamSpl$mean_logsal
yr = sum(baseball_teamSpl$sum_logsal) / sum(baseball_teamSpl$M_i)
n = 6; N = 30; f = n/N; Mbar = mean(baseball_teamSpl$M_i)
SE_yr = sqrt((1-f)/(n*Mbar^2) * 1/(n-1) *
               sum((baseball_teamSpl$sum_logsal-yr*baseball_teamSpl$M_i)^2))
CI_y_lb = yr - qnorm(0.975) * SE_yr; CI_y_ub = yr + qnorm(0.975) * SE_yr
```

d. 估计 pitcher 的总体比例

仍采用比估计来估计 pitcher 的总体比例，有

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\mathrm{unb}}}{\hat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{74}{157} = 0.4713$$

估计量的标准误为

$$\mathrm{SE}(\hat{\bar{y}}_r) = \sqrt{\frac{1-f}{n\bar{M}^2}\frac{1}{n-1}\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2} = 0.01334$$

则 pitcher 总体比例的95%置信区间为

$$\left[\hat{\bar{y}}_r - z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_r)\ ,\ \hat{\bar{y}}_r + z_{\alpha/2}\mathrm{SE}(\hat{\bar{y}}_r)\right] = [0.4452, 0.4975]$$

以下为计算的代码：

```
yr2 = sum(baseball_teamSpl$sum_pitcher) / sum(baseball_teamSpl$M_i)
SE_yr2 = sqrt((1-f)/(n*Mbar^2) * 1/(n-1) *
               sum((baseball_teamSpl$sum_pitcher-yr2*baseball_teamSpl$M_i)^2))
CI_y2_lb = yr2 - qnorm(0.975) * SE_yr2; CI_y2_ub = yr2 + qnorm(0.975) * SE_yr2
```

e. 与使用SRS得到的估计量进行比较

将本题中采用单阶整群抽样和习题2.32中采用SRS得到的估计量进行比较，结果如表1所示：

表1：单阶整群抽样与SRS得到的估计量的比较

抽样方法	变量	点估计	95%置信下限	95%置信上限
单阶整群抽样	logsal总体均值	13.968	13.584	14.353
SRS	logsal总体均值	13.95	13.774	14.126
单阶整群抽样	pitcher总体比例	0.4713	0.4452	0.4975
SRS	pitcher总体比例	0.44	0.368	0.512

由表1可知，在估计 `logsal` 的总体均值时，采用SRS得到的置信区间更短，这是由于同一个PSU（team）中各球员的薪水更加接近，所以当抽取的样本容量相近时，SRS比单阶整群抽样得到的估计量更准确。在估计 `pitcher` 的总体比例时，采用单阶整群抽样得到的置信区间更短，这是由于 `pitcher` 的人数和队伍人数有一定的正相关关系，使用比估计能够得到更小的方差，即单阶整群抽样比SRS得到的估计量更准确。