习题 3.35

e. 最优配置是否值得考虑

读入第五周作业中抽取的分层样本数据（按比例配置），并计算每一层中变量 logsal 和 pitcher 的样本标准差，代码如下：

```
library(dplyr)
baseball = read.csv('baseball.csv', header=FALSE)
baseball_team = data.frame(team=baseball[[1]])
baseball_prop = read.csv('baseball_StS.csv')

by_team_raw = group_by(baseball_team, team)
by_team_sample = group_by(baseball_prop, team)
stats_baseball = by_team_raw %>% summarise(N_h=n())
stats_baseball$W_h = stats_baseball$N_h / sum(stats_baseball$N_h)

# logsal 和 pitcher 的层内样本标准差
sd_df = by_team_sample %>%
  summarise(s_h_logsal=sd(logsal), s_h_pitcher=sd(pitcher))
stats_baseball = merge(stats_baseball, sd_df, by='team')
```

将比例配置和最优配置的估计量方差进行比较，有公式

$$V_{\text{prop}}(\bar{y}_{str}) - V_{\text{Neyman}}(\bar{y}_{str}) = \frac{1}{n}\left[\sum_{h=1}^{H} W_h S_h^2 - \left(\sum_{h=1}^{H} W_h S_h\right)^2\right]$$

分别对变量 logsal 和 pitcher 采用上述公式计算两种配置方法的方差之差，代码如下：

```
attach(stats_baseball)
n = dim(baseball_prop)[1]
logsal_Vdiff = (sum(W_h*s_h_logsal^2) - (sum(W_h*s_h_logsal))^2) / n # logsal
pitcher_Vdiff = (sum(W_h*s_h_pitcher^2) - (sum(W_h*s_h_pitcher))^2) / n # pitcher
detach(stats_baseball)
```

计算结果如表 1 所示。由此可见，比例配置和 Neyman 配置的方差之差非常有限，所以没有必要采用 Neyman 配置，直接使用比例配置即可。

表 1：logsal 和 pitcher 两种配置方法的方差之差

变量名	$V_{\text{prop}}(\bar{y}_{str}) - V_{\text{Neyman}}(\bar{y}_{str})$
logsal	6.768×10^{-4}
pitcher	6.806×10^{-5}

f. 对样本采用最优配置

每一层的抽样成本相等，则最优配置的样本量计算公式为 $\frac{n_h}{n} = \frac{W_h S_h}{\sum_{l=1}^{H} W_l S_l}$；

分别对估算变量 `logsal` 和 `pitcher` 均值计算其最优配置的各层样本量，代码和结果如下：

```
stats_baseball$logsal_WS = stats_baseball$W_h * stats_baseball$s_h_logsal
stats_baseball$pitcher_WS = stats_baseball$W_h * stats_baseball$s_h_pitcher
stats_baseball$logsal_n_h = n * stats_baseball$logsal_WS / sum(stats_baseball$logsal_WS)
stats_baseball$pitcher_n_h = n * stats_baseball$pitcher_WS / sum(stats_baseball$pitcher_WS)
print(data.frame(team=stats_baseball$team, logsal_n_h=round(stats_baseball$logsal_n_h),
                 pitcher_n_h=round(stats_baseball$pitcher_n_h)))
```

```
   team logsal_n_h pitcher_n_h
1   ANA          6           5
2   ARI          7           6
3   ATL          5           6
4   BAL          5           4
5   BOS          5           5
6   CHA          7           4
7   CHN          5           6
8   CIN          4           6
9   CLE          4           6
10  COL          1           6
11  DET          6           5
12  FLO          5           4
13  HOU          2           0
14  KCA          5           6
15  LAN          2           5
16  MIL          4           5
17  MIN          6           5
18  MON          5           6
19  NYA          6           5
20  NYN          5           5
21  OAK          5           6
22  PHI          6           4
23  PIT          4           6
24  SDN          5           4
25  SEA          5           5
26  SFN          8           6
27  SLN          6           5
28  TBA          3           4
29  TEX          6           5
30  TOR          6           5
```

所以，变量 `logsal` 和 `pitcher` 的最优配置都和比例配置较为接近（比例配置每层的样本容量为 5）。