1. 读入数据

因为本题用到的数据仅为棒球选手的 ID 和 salary, 所以仅提取数据框中的 player 和 salary 两列, 读取的数据保存为变量名 baseball。

```
> baseball = read.csv('baseball.csv', header=FALSE)
> baseball = baseball[,c(3,4)]
> colnames(baseball) = c('player', 'salary')
> head(baseball)

    player salary
1 anderga0 6200000
2 colonba0 11000000
3 davanje0 375000
4 donnebr0 375000
5 eckstda0 2150000
6 erstada0 7750000
```

2. 抽取样本量为 150 的 SRS

采用以下方法实现 SRS, 步骤:

```
1) \diamondsuit \mathcal{S} = \emptyset, k = 0;
```

- 2) 由计算机独立地产生随机数 $x, x \sim U(0,1)$, 由此得到 R, R = [Nx] + 1, 则 $R \sim U(1,...,N)$;
- 3) if $R \in \mathcal{S}$, then go to 2) else $\mathcal{S} = \mathcal{S} \bigcup \{R\}, k = k + 1$;
- 4) if k < n, then go to 2) else end.

最终的抽样结果以变量 S 保存(S 中的每个元素表示抽样单元的序号)

```
[1] 575 698 607 707 364 133 260 406 580 789 28 122 587 1 312 369 310 321 [19] 143 759 362 261 770 564 514 311 557 434 181 387 633 5 150 544 295 289 [37] 693 721 493 107 624 343 740 617 207 257 48 35 44 499 769 660 252 170 [55] 584 398 582 65 348 189 631 786 604 781 175 757 120 479 755 549 403 298 [73] 268 39 494 767 523 407 694 411 7 16 116 244 659 401 641 49 645 63
```

[91] 570 410 574 598 77 318 235 492 777 416 720 509 689 201 172 486 306 602 [109] 303 634 722 785 469 8 256 446 72 545 566 638 753 149 223 618 27 536 [127] 648 529 556 148 78 644 553 173 503 118 765 790 396 733 297 635 281 167 [145] 626 345 675 480 591 475

3. 变量 salary 和 logsal 的直方图

```
> baseball_SRS = baseball[S,] # 用 baseball_SRS 储存抽样数据
> par(mfrow=c(1,2))
> salary = baseball_SRS$salary # 得到变量 salary
> logsal = log(baseball_SRS$salary) # 得到变量 logsal
> hist(salary) # salary 的直方图
> hist(logsal) # logsal 的直方图
```

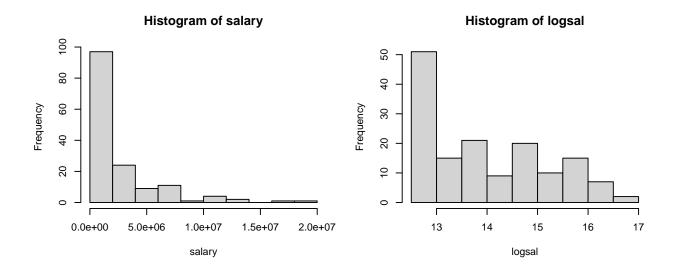


图 1: 左: salary 的直方图;右: logsal 的直方图

由上图可知, salary 的分布严重右偏,不是正态分布。经过对数变换后,logsal 的右偏性不及 salary,但仍有明显的右偏性,也不是正态分布。