Homework 1

张子乐,3230104237

记 $S_n = \sum_{i=1}^{n} X_i$

问题：

1. 样本容量 n 如何影响逼近速率？

2. 模拟次数如何影响逼近速率？

例如：

设随机变量 $X_i \sim \mathcal{U}(0,2)$ 或 $X_i \sim t(3)$，
取不同的样本容量:$n = 100$,$n = 500$,$n = 1000$

1. 样本容量 n 如何影响逼近速率？

首先写一套通用的函数

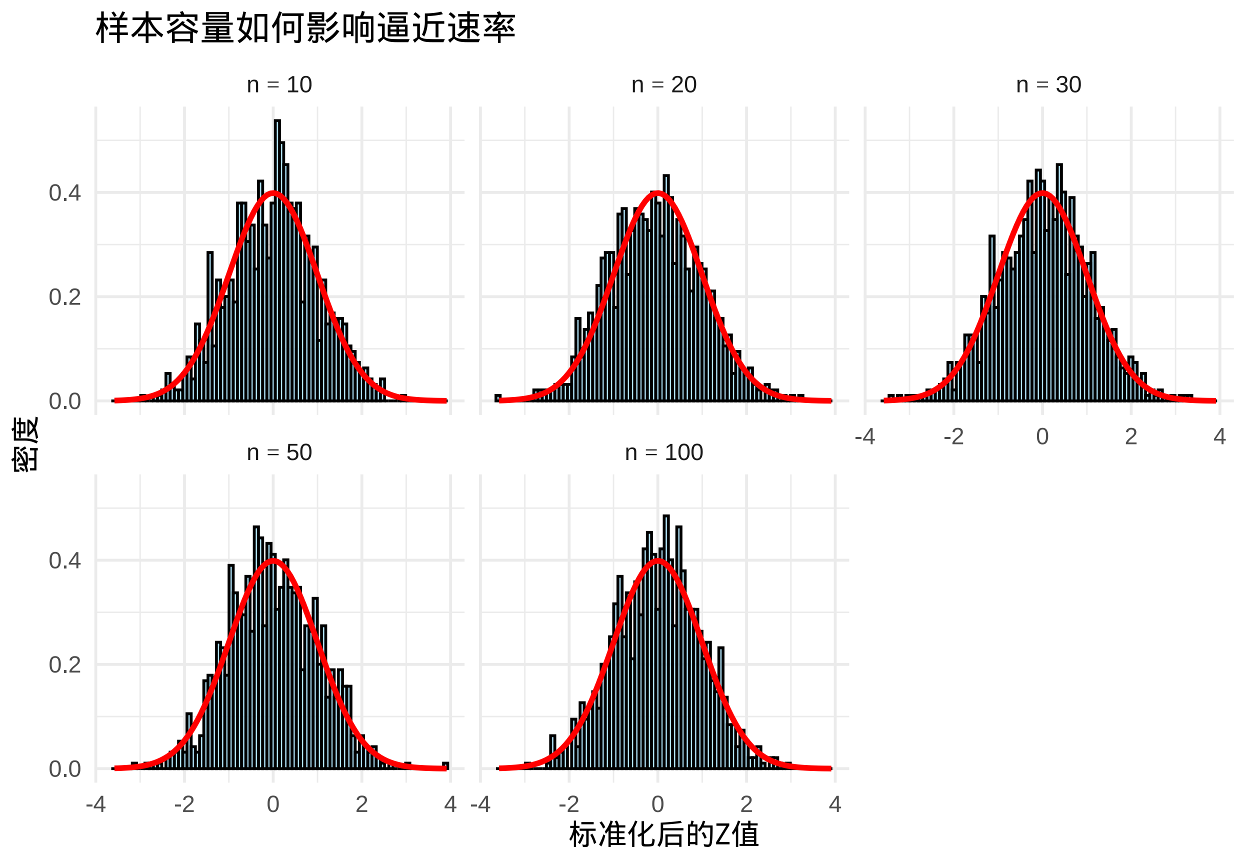
```r
test1 <- function(iter,mu,variance,sample_size,rand_func,...){
  # iter: 模拟次数
  # mu: 总体均值
  # variance: 总体方差
  # sample_size: 样本容量，这里分别取 10,20,30,50,100
  # rand_func: 随机变量服从的分布

  results <- list()

  # (Σxi-nu)/sqrt(n )
  standard <- function(nums){
    x = rand_func(nums,...)
    sn <- sum(x)
```

```r
    zn <- (sn-nums*mu)/sqrt(nums * variance)
    return(zn)
  }

  # 在每个样本容量下使用中心极限定理
  for(nums in sample_size){
    standard_sums <- replicate(iter,standard(nums))
     results[[as.character(nums)]] <- data.frame(
      Z = standard_sums,
      n = factor(nums)
    )
  }

  df <- do.call(rbind, results)

  # 绘图
  picture <- ggplot(df, aes(x = Z)) +
    geom_histogram(aes(y = after_stat(density)), bins = 80,
                   fill = "skyblue", color = "black",
                   alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0, sd = 1),
                  color = "red", linewidth = 1) +
    facet_wrap(~ n,
               labeller = label_bquote(n == .(as.character(n)))) +
    labs(
      title = " 样本容量如何影响逼近速率",
      x = " 标准化后的 Z 值",
      y = " 密度"
    ) +
    theme_minimal()

  print(picture)
}
```

(1) 设随机变量 $X_i \sim \mathcal{U}(0,2)$

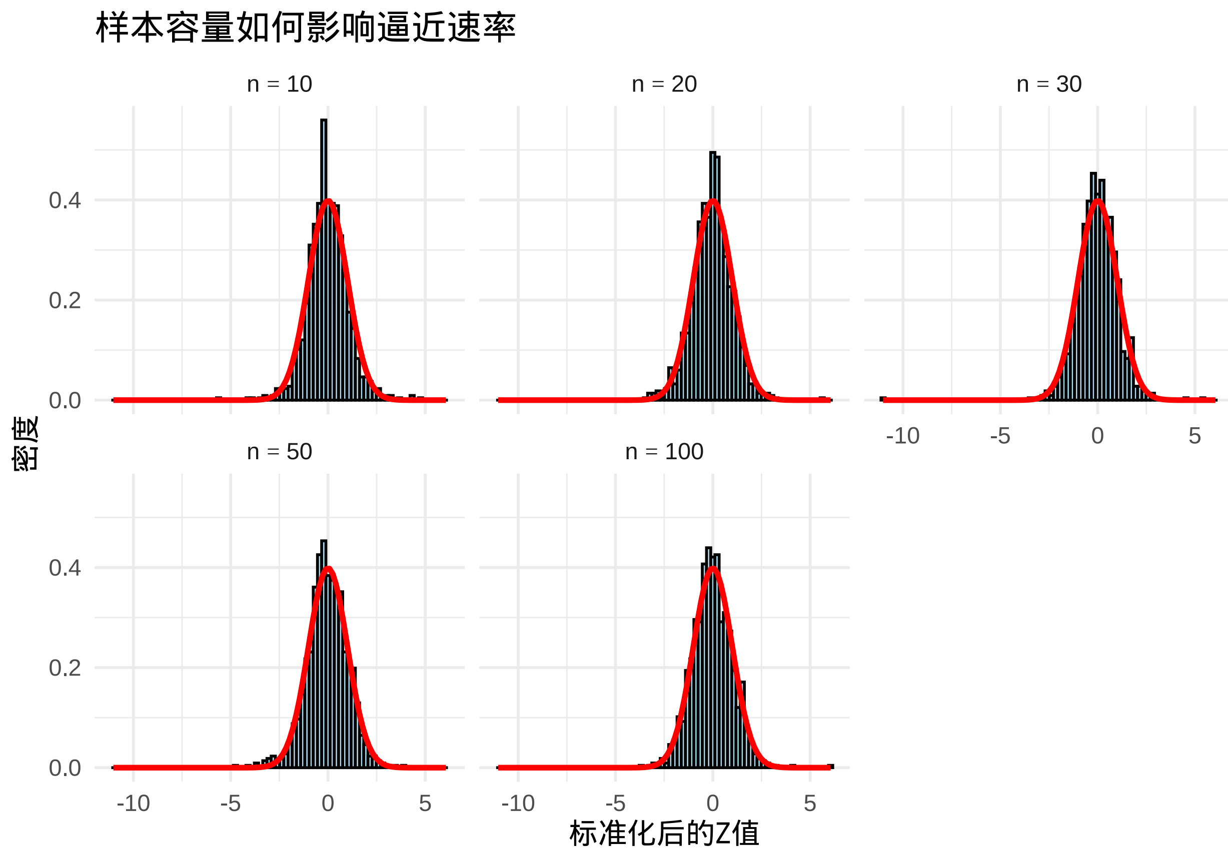
```r
iter <- 1000 # 模拟次数
mu <- 1 # 均值
```

```r
variance <- 1/3 # 方差
sample_size <- c(10,20,30,50,100) # 样本容量
test1(iter,mu,variance,sample_size,runif,min = 0, max = 2)
```



(2) 设随机变量 $X_i \sim t(3)$

```r
iter <- 1000 # 模拟次数
mu <- 0 # 均值
variance <- 3 # 方差
sample_size <- c(10,20,30,50,100) # 样本容量
test1(iter,mu,variance,sample_size,rt,df=3)
```



结论

- 样本容量 n 越大，标准化后样本和的分布越接近标准正态分布，逼近效果越好。
- 在 $n \geq 30$ 后，已经非常逼近标准正态分布。
- Berry-Esseen 定理指出逼近速率是 $O(\frac{1}{n})$，但常数项依赖于原始分布的三阶矩
- 从图像上看，对比均匀分布和 t 分布，两者的逼近速率有所差异。

2. 模拟次数如何影响逼近速率？

同样先写一个通用的函数

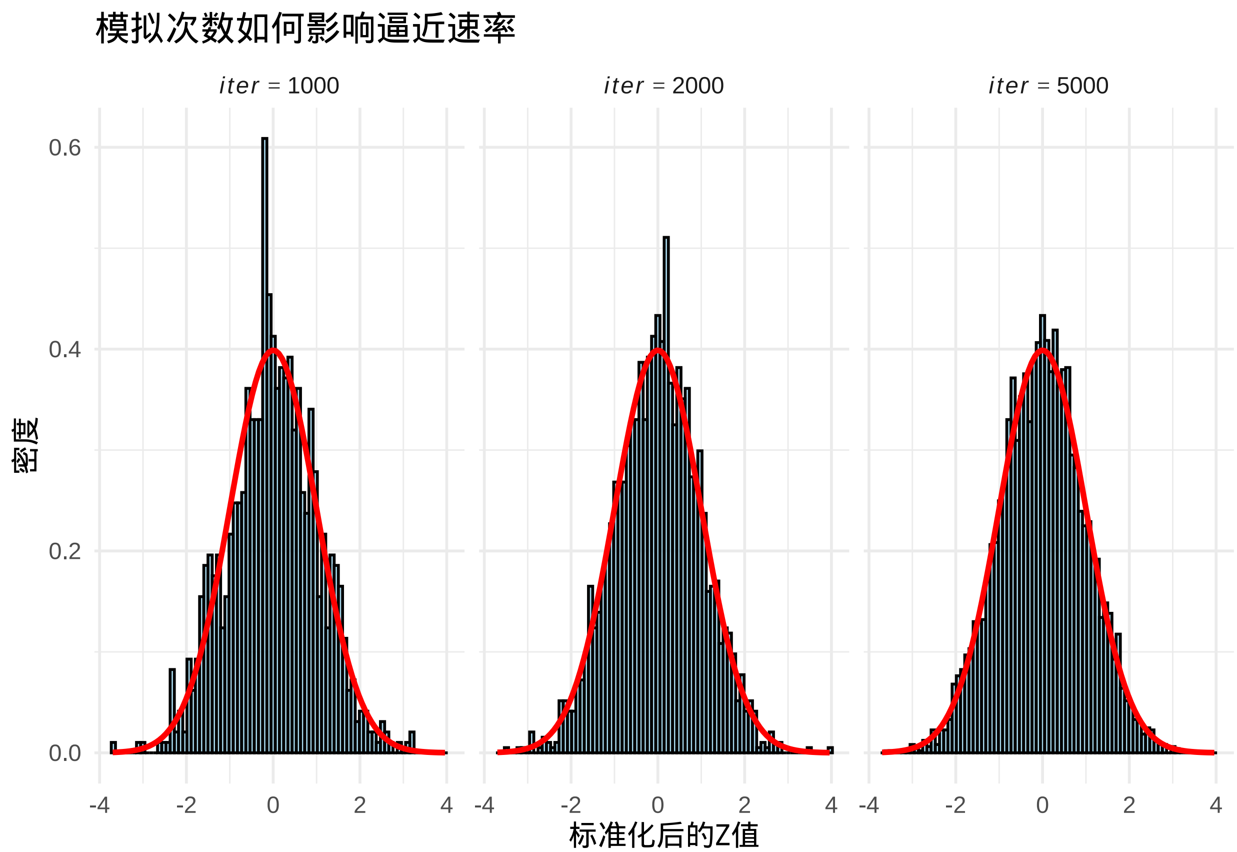
```
test2 <- function(iter,mu,variance,sample_size,rand_func,...){
  # iter: 模拟次数
  # mu: 总体均值
  # variance: 总体方差
  # sample_size: 样本容量，这里取 100
  # rand_func: 随机变量服从的分布
```

```r
results <- list()

# (Σxi-nu)/sqrt(n )
standard <- function(sample_size){
  x = rand_func(sample_size,...)
  sn <- sum(x)
  zn <- (sn-sample_size*mu)/sqrt(sample_size * variance)
  return(zn)
}

# 在每个样本容量下使用中心极限定理
for(it in iter){
  standard_sums <- replicate(it,standard(sample_size))
   results[[as.character(it)]] <- data.frame(
    Z = standard_sums,
    iter_count = factor(it)
  )
}

df <- do.call(rbind, results)

# 绘图
picture <- ggplot(df, aes(x = Z)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 80,
                 fill = "skyblue", color = "black",
                 alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = 0, sd = 1),
                color = "red", linewidth = 1) +
   facet_wrap(~ iter_count,
              labeller = label_bquote(italic(iter) == .(as.character(iter_count)))) +
  labs(
    title = " 模拟次数如何影响逼近速率",
    x = " 标准化后的 Z 值",
    y = " 密度"
  ) +
  theme_minimal()
```

```
  print(picture)
}
```

(1) 设随机变量 $X_i \sim \mathcal{U}(0,2)$

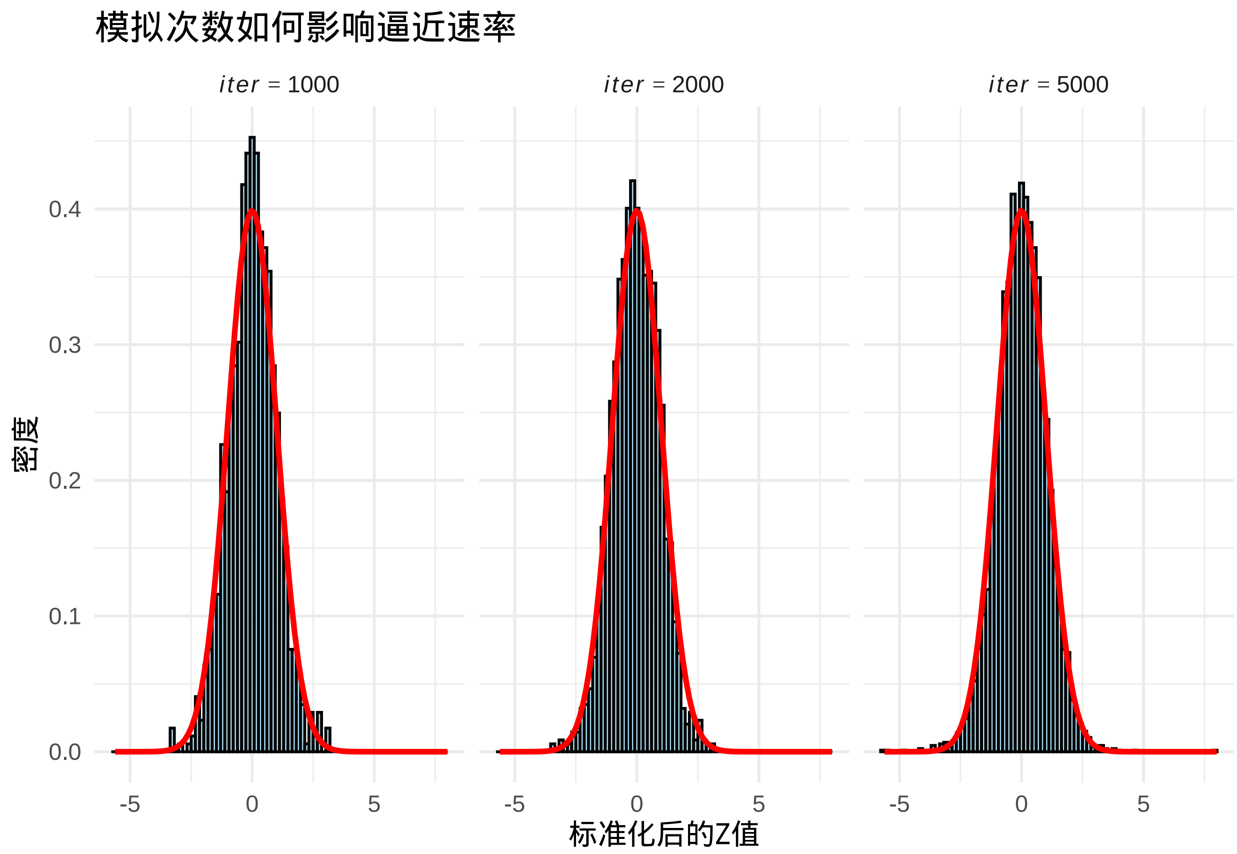
```
iter <- c(1000,2000,5000) # 模拟次数
mu <- 1 # 均值
variance <- 1/3 # 方差
sample_size <- 100 # 样本容量
test2(iter,mu,variance,sample_size,runif,min = 0, max = 2)
```



(2) 设随机变量 $X_i \sim t(3)$

```
iter <- c(1000,2000,5000) # 模拟次数
mu <- 0 # 均值
variance <- 3 # 方差
```

```r
sample_size <- 100 # 样本容量
test2(iter,mu,variance,sample_size,rt,df=3)
```



结论：

- 模拟次数不会影响逼近速率，但会影响模拟中心极限定理中的观测效果
- 模拟次数越多，经验分布越稳定、越接近真实分布