L06-大数定律和中心极限定理的模拟

郑盼盼

2024-10-25

目录

6.1 大数定律

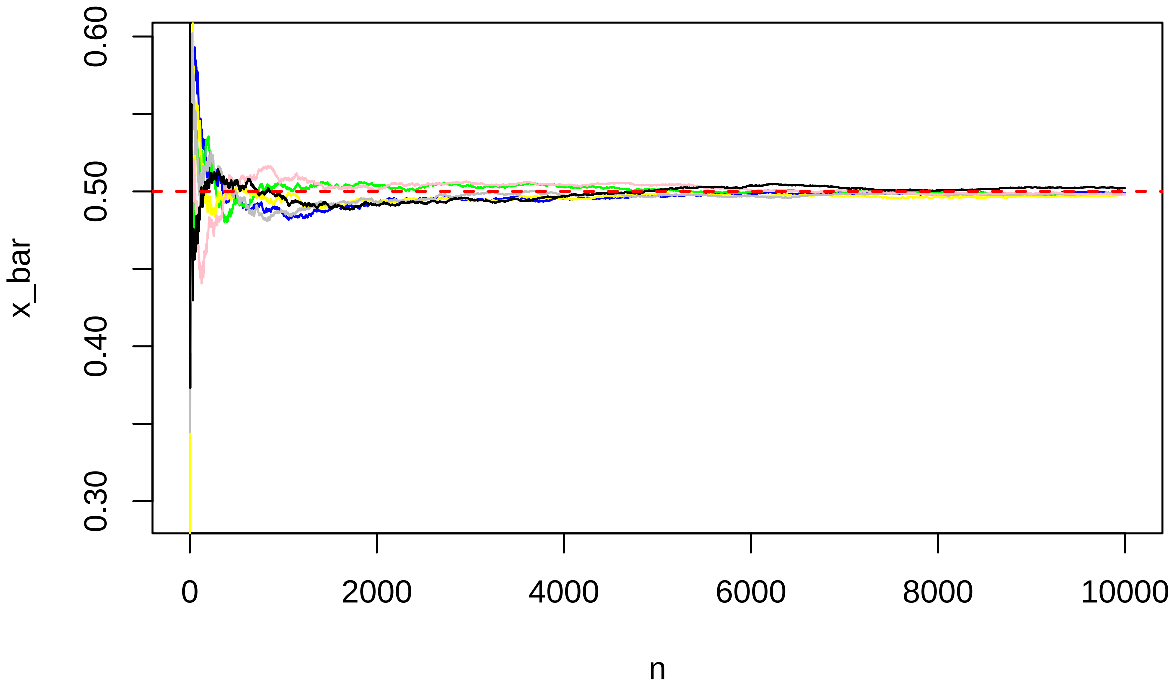
6.1.1 大数定律的模拟

考察均匀分布随机变量 $\xi \sim U(0,1)$ 重复观测的算术平均值随着重复观测的变化情况。

```r
# m 为我们希望生成的随机数的个数（重复观测次数）
m <- 10000
my_cols = c("blue", "green", "yellow", "grey", "pink", "black") # 由于我们希望生成 6
↪ 组随机数序列，我们定义六种颜色用于绘图

# 通过 for 循环生成 6 组随机数
for (i in 1:6){
  x <- runif(m, 0, 1) # 生成 m 个服从 U(0,1) 的随机数（重复观测 m 次）
  x_bar = c() # 定义一个空向量 x_bar 用于存储 x 中前 j 个元素算术平均数
```

```r
  # 内部的 for 循环用于计算 x 中前 j 个元素的算术平均数
  for (j in 1:m){
    x_bar <- c(x_bar, mean(x[1:j])) # 将计算得到前 j 个元素的算术平均数作为最后一个
↪   元素添加到向量 x_bar 中
  }
  # 当 i 为 1 时，通过 plot 新建一张图片
  if (i==1){
    plot(x_bar,
         type="l",
         lwd=1.,
         xlab="n",
         ylab="x_bar",
         col=my_cols[i])}
  else {
    # 若 i 不为 1，通过 lines 函数向图像中添加折线图
    lines(x_bar, lwd=1., col=my_cols[i])
  }
}
# 添加期望对应的线
abline(h=0.5,
       col="red",
       lwd=1.5,
       lty=2)
```



通过如上的代码，我们可以发现，随着观测数目的增加，随机变量 ξ 的观测值的算术平均值越来越接近其数学期望；设 ξ_i 为随机变量 ξ 第 i 次观测的结果，则有：

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\xi_i=\mathbb{E}(\xi)$$

这就是 P85 所说的**柯尔莫哥洛夫强大数定律**

6.1.2 经验分布

经验分布　设 $X_1, X_2, \cdots, X_n$ 为随机变量 X 的 n 次重复观测，称

$$F_n(x)=\frac{n(\{i:1\le i\le n, X_i\le x\})}{n}$$

为 X 的**经验分布**，此处：

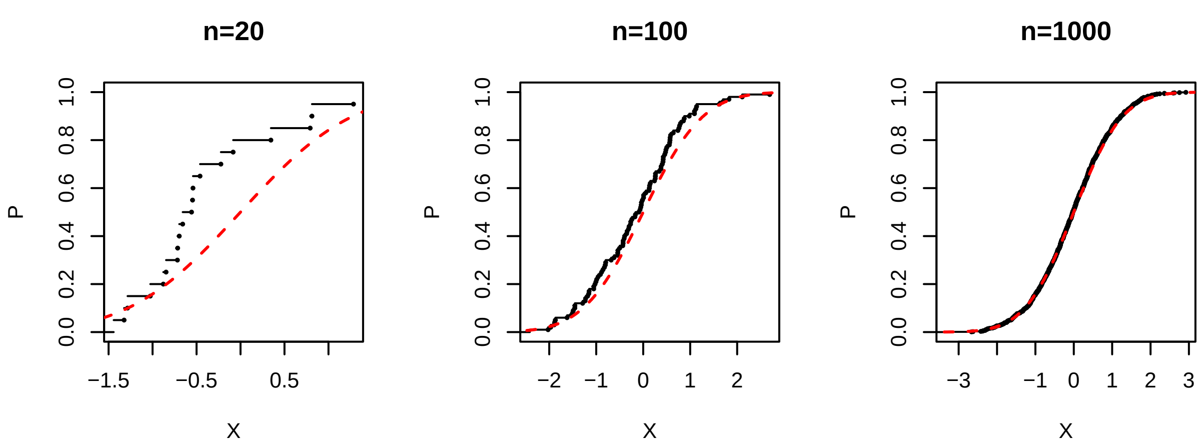
$$n(\{i:1\le i\le n, X_i\ \le x\})$$

表示 $X_1, X_2, \dots, X_n$ 中小于 x 的值的个数。

R 语言模拟　随机变量 $X\sim N(0,1)$，对其分别重复观测 $20, 100, 1000$ 次，绘制其经验分布图像，并和标准正态分布的分布函数曲线进行对比

```
# 设置绘图的布局为 1 行 3 列，且每张图片呈正方形
par(mfrow=c(1,3),pty = "s")
```

```r
# 迭代不同的观测数量，绘制不同的经验分布曲线
for (n in c(20,100,1000)){
  x <- seq(-4,4, by=0.001) # 定义横坐标 x 用于后面绘制标准正态分布分布函数曲线
  X <- rnorm(n)   # 对于随机变量 X 重复观测 n 次，得到观测值向量 X
  min_X = min(X)  # 找出最小的观测值
  max_X = max(X)  # 找出最大的观测值
  Z = unique(sort(X)) # 对于观测值 X 从小到大进行排序，并找出所有的不重复值，作为向
↪  量 Z
  fn = c() # 定义存储经验分布函数值的向量 fn
  # 依次遍历 Z 中的元素
  for (i in Z){
    fn <- c(fn, sum(X <= i)/n) # 统计小于 i 的观测值个数，并除以总观测值数目（即经验
↪  分布函数值），并将计算的结果作为最后一个元素添加进向量 fn
  }
  plot(c(min_X-1,min_X), c(0,0), type="l",
       xlim = c(min_X, max_X),
       ylim = c(0,1),
       xlab = "X",
       ylab = "P",
       main = paste("n=", n, sep=""))
  points(c(0), min_X, cex=0.3)
  for (i in 2:length(Z)){
    lines(c(Z[i-1],Z[i]), c(fn[i-1],fn[i-1]))
    points(c(Z[i]), fn[i-1],cex=0.3)
  }
  lines(x, pnorm(x), lwd=1.5, lty=2, col="red")
}
```



通过上图，我们可以发现，随着重复观测数目 n 的增加，经验分布愈发接近理论的分布函数值：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

6.1.3 大数定律与蒙特卡洛模拟

根据大数定律，对于随机变量 X 和其重复观测值 $X_1, X_2, \cdots, X_n$，我们有

$$\mathbb{E}(X) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$$

对于函数 $f(x)$ 在 $[a,b]$ 上的定积分：

$$\int_a^b f(x)\,dx$$

设随机变量 $X \sim U(a,b)$，我们有：

$$\mathbb{E}[f(X)] = \int_a^b \frac{f(x)}{b-a}\,dx = \frac{1}{b-a}\int_a^b f(x)\,dx \tag{6.1}$$

结合上述的大数定律：

$$\mathbb{E}[f(X)] = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{1}{n}f(X_i) \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i) \tag{6.2}$$

结合公式 (6.1), (6.2)，我们可以估算 $f(x)$ 在 $[a,b]$ 上的定积分

$$\int_a^b f(x)\,dx \approx (b-a)\frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

例 1 我们可以利用如上方法计算

$$\int_1^2 x^2\,dx$$

```r
n <- 10000          # 定义观测值数目 n
Y <- runif(n, 1, 2) # 生成 n 个服从 U(1,2) 的随机数，形成 n 维向量 X
estim_int <- (2-1) * mean(Y^2) # 使用蒙特卡洛模拟对于如上积分进行估计
acc_int <- integrate(function(x) x^2, 1,2) # 通过 integrate 函数直接计算如上积分
cat("蒙特卡洛模拟的结果：", estim_int, "\n实际结果：", acc_int$value)
```

```
## 蒙特卡洛模拟的结果： 2.339596
## 实际结果： 2.333333
```

例 2 设 $X \sim N(0,1)$，尝试写出用蒙特卡洛模拟方法近似计算概率 $\mathbb{P}(0.1 < X < 2)$ 的 R 程序代码。

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{x^2}{2}\right]$$

- **答**：根据

$$\mathbb{P}(0.1 < X < 2) = \int_{0.1}^{2} \varphi(x)\,dx \approx (2-0.1) \times \frac{1}{N}\sum_{i=1}^{N} \varphi(Y_i)$$

其中 $Y_1, \dots, Y_N$ 是 $Y \sim U(0.1, 2)$ 的 n 次重复观测值。

```
Y <- runif(10000,0.1,2) # 生成 10000 个服从 U(0.1,1) 的随机数，组成向量 Y
c <- 1/sqrt(2*pi)
phiY <- c * exp(-Y^2/2) # 将随机向量 Y 代入标准正态分布的密度函数，得到向量 phiY
estim_int <- (2-0.1) * mean(phiY) # 利用蒙特卡洛模拟估计如上定积分
acc_int <- pnorm(2,0,1) - pnorm(0.1,0,1)
cat(" 蒙特卡洛估计结果: ", estim_int,"\n实际结果: ", acc_int)
```

```
## 蒙特卡洛估计结果:  0.4395562
## 实际结果:  0.437422
```

当然我们也可以不通过积分，而直接通过经验分布的方式，近似计算

```
X <- rnorm(100000)
estim_P <- sum(X > 0.1 & X < 2)/length(X)
print(estim_P)
```

```
## [1] 0.4388
```

6.2 中心极限定理的模拟

通过大数定律我们可知，随着 $n \to \infty$，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X)$，那我们如何评价 $\bar{X}$ 对于 $\mathbb{E}(X)$ 的估计误差呢？**中心极限定理！**

中心极限定理 设随机变量 X 的方差为大于 0 的实数，若 $X_1, X_2, \dots, X_n$ 为 X 的 n 次重复观测，则

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{\bar{X} - \mathbb{E}(X)}{\sqrt{D(X)/n}} \le x\right\} = \Phi(x)$$

其中 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

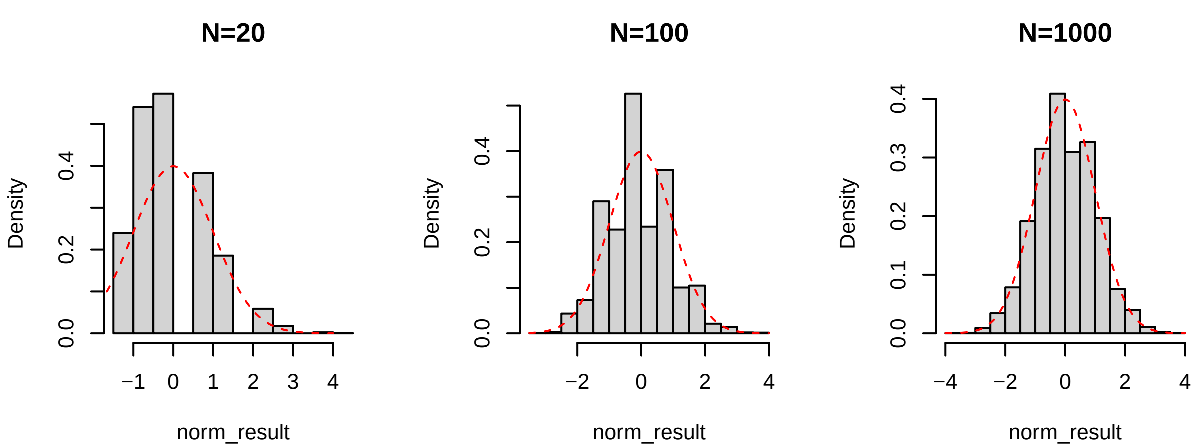
中心极限定理的模拟 我们可以通过如下代码对于模拟中心极限定理，对于 $X \sim B(1, 0.1)$，令：

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

取样本量 $N = 20, 100, 1000$，分别重复抽样 10000 次，得到样本均值 $\bar{X}_N$。绘制标准化后的 $\bar{X}_N$ 的直方图，并与标准正态分布的概率密度曲线进行对比。

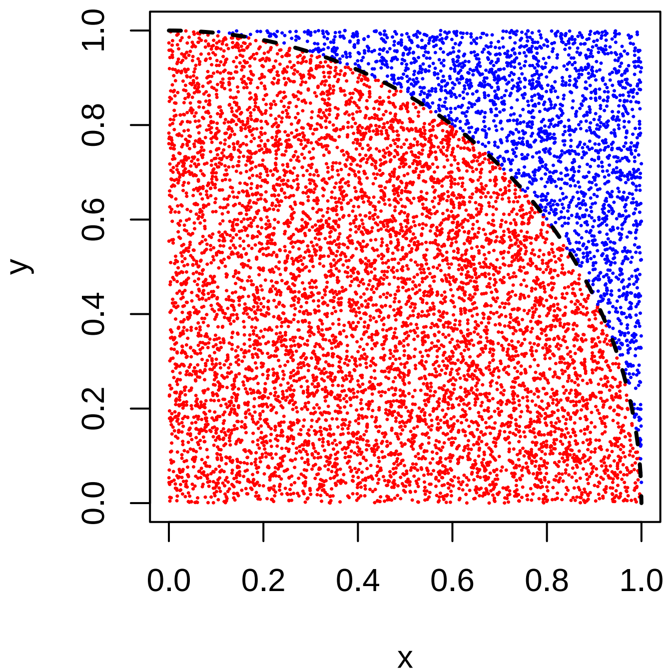
```
n <- 1
p <- 0.1
par(mfrow = c(1, 3), pty = "s")
N <- c(20, 100, 1000)
m <- 10000

for (i in N) {
  sim_result <- matrix(rbinom(m * i, n, p), m, i)
  mean_result = rowSums(sim_result) / i
  norm_result = (mean_result - n * p) / sqrt(n * p * (1 - p) / i)
  hist(norm_result,
       freq = F,
       main = paste("N=", i, sep = ""))
  lines(seq(-4, 4, by = 0.001),
        dnorm(seq(-4, 4, by = 0.001)),
        col = "red",
        lty = 2)
}
```



* 利用蒙特卡洛模拟估计 π

在一个边长为 1 的正方形内（坐标范围 [0,1]），画一个以原点为中心的半径为 1 的 1/4 圆弧（如下图黑色虚线所示），



现在往正方形区域内随机生成一点 (x, y)，根据几何概型，其落入圆弧内部（红色区域）的概率为

$$\frac{\pi/4}{1} = \frac{\pi}{4} \approx \frac{n(\{(x,y) : x^2 + y^2 \leq 1\})}{n}$$

于是，我们可以结合如上的原理，利用如下代码估算 π

```
n <- 10000
x <- runif(n,0,1)
y <- runif(n,0,1)
points <- matrix(c(x,y), n, 2)
distance <- rowSums(points ^ 2)
estim_pi <- sum(distance <= 1)/n * 4
cat("estimate pi:", estim_pi)
```

```
## estimate pi: 3.1172
```

Questions

1. 使用 R 语言代码模拟研究 $X \sim P(5)$ 的重复观测数据的算术平均值和观测次数之间的关系，总结规律。(hints：大数定律
2. 通过 1000 次模拟观测数据估计 $X \sim B(10, 0.8)$ 的数学期望 $\mathbb{E}(X)$，讨论估计的结果是否为随机变量，并判断估计误差的取值范围。(hints：中心极限定理
3. 已知数学考试的平均成绩（5 分制）为 4.10，标准差为 0.3，估算 1000 名学生的成绩之和小于 400 的概率 (hints: 中心极限定理，成绩之和与成绩平均值之间的关系，正态分布的变换
4. $X \sim B(10, 0.4)$，写出统计学方法估计 $\mathbb{E}(\sin(X))$ 的程序代码，并给出估计结果。
5. 写出用统计学方法估计 $s = \sum_{i=1}^{100} i^4$ 的程序代码，并给出估计结果和误差。
6. 写出用统计学方法估计 $\int_0^1 \cos(x)\, dx$ 的程序代码，并给出估计结果和估计误差。