《数理统计分析》

实验报告

|  |  |
| --- | --- |
| 实验名称 | 随机变量及其分布（中心极限定理证明） |
| 班 级 | 区块链工程221 |
| 学生姓名 | 高鑫宇 |
| 学生学号 | 2022131012 |
| 指导教师 | 高路路 |
| 成 绩 |  |

2024年11月27日

成都信息工程大学 人工智能学院

**一、实验目的：**

学习、理解基于R语言数据处理和分析的基本操作，和对该内容的具体实践。

**二、实验环境：**

Windows10操作系统、R、RStudio

**三、实验步骤：**

1. 完成以下题目：

1）. 设在10个产品中有2个不合格品，若以中随机取出4个，则其中不合格品数X是离散随机变量，它可取0，1，2三个值。

（a） X取这些值的概率为多少？

# 定义不放回抽样概率函数

P\_not\_replace <- function(k, total = 10, defective = 2, sample = 4) {

choose(defective, k) \* choose(total - defective, sample - k) / choose(total, sample)

}

# 定义 X 的取值范围

k\_values <- 0:2

# 计算每个 X 的概率

P\_X <- sapply(k\_values, P\_not\_replace)

# 输出结果

names(P\_X) <- paste0("P(X=", k\_values, ")")

P\_X

（b） 对于同样问题，若用放回抽样，则从10个产品（其中有2个不合格品）中随机取出4个，其中不合格品数Y是另一个随机变量，它可取0，1，2，3，4五个值，Y取这些值的概率为多少？

# 定义参数

n <- 4 # 抽样次数

p <- 0.2 # 抽到不合格品的概率

k\_values <- 0:4 # Y 的可能取值

# 计算每个 Y 的概率

P\_Y <- dbinom(k\_values, size = n, prob = p)

# 输出结果

names(P\_Y) <- paste0("P(Y=", k\_values, ")")

P\_Y

2）. 在一次制造过程中，不合格品率为 0.1，如今在成品中随机取出6个，记X为6个成品中的不合格品数，则X服从二项分布b(6, 0.1)，简记为X~b(6, 0.1）。现研究如下三个问题：

（a）恰有1个不合格品的概率是多少？

# 参数设置

n <- 6 # 试验次数

p <- 0.1 # 不合格品率

k <- 1 # 恰有1个不合格品

# 计算概率

P\_X\_1 <- dbinom(k, size = n, prob = p)

P\_X\_1

（b）不超过1个不合格品的概率为多少？

# 参数设置

n <- 6 # 试验次数

p <- 0.1 # 不合格品率

k <- 1 # 不超过1个不合格品

# 计算概率

P\_X\_leq\_1 <- pbinom(k, size = n, prob = p)

P\_X\_leq\_1

（c）二项分布b(6, 0.1）的均值、方差与标淮差分别为多少？

# 参数设置

n <- 6 # 试验次数

p <- 0.1 # 不合格品率

# 计算均值

mean\_X <- n \* p

# 计算方差

variance\_X <- n \* p \* (1 - p)

# 计算标准差

std\_dev\_X <- sqrt(variance\_X)

# 输出结果

mean\_X

variance\_X

std\_dev\_X

3）. 自动车床生产的零件长度X（毫米）服从N(30，0.752），者零件的长度在30±1.5毫米之间为合格品，求生产的零件是合格品的概率。

# 参数设置

mu <- 30 # 均值

sigma <- 0.75 # 标准差

lower\_bound <- 28.5 # 下限

upper\_bound <- 31.5 # 上限

# 计算合格品的概率

P\_qualified <- pnorm(upper\_bound, mean = mu, sd = sigma) - pnorm(lower\_bound, mean = mu, sd = sigma)

P\_qualified

4. 抽样调查表明，考生的外语成绩 （总分为100分）近似服从正态分布，平均成绩为72分，96分以上占总数的2.3%。试求考生外语成缆在60分至84 分之间的概率。

# 给定的参数

mu <- 72 # 平均成绩

x\_upper <- 96 # 上限

prob\_upper <- 0.023 # 上面2.3%对应的概率

# 计算标准差

Z\_upper <- qnorm(1 - prob\_upper) # 对应的Z值

sigma <- (x\_upper - mu) / Z\_upper # 计算标准差

# 计算在60分至84分之间的概率

x\_lower <- 60

x\_upper <- 84

# 计算对应的Z值

Z\_lower <- (x\_lower - mu) / sigma

Z\_upper <- (x\_upper - mu) / sigma

# 计算概率

P\_60\_to\_84 <- pnorm(Z\_upper) - pnorm(Z\_lower)

P\_60\_to\_84

5）. 从某厂生产的一批铆钉中随机抽取10个，测得其直径（单位：毫米）分别为：13.35， 13.38， 13.40，13.43，13.32，13.48，13.34，13.47，13.44，13.50。试求铆钉头部直径这一总体的均值与标准差 的估计。

# 样本数据

x <- c(13.35, 13.38, 13.40, 13.43, 13.32, 13.48, 13.34, 13.47, 13.44, 13.50)

# 计算样本均值

sample\_mean <- mean(x)

# 计算样本标准差

sample\_sd <- sd(x)

# 输出结果

sample\_mean

sample\_sd

6）. 针对原始电影数据：

（a） 请用 sample 函数随机抽取 100个数据并计算其基本统计量。

df1 <- read.csv("D:/douban\_movies\_cn\_clean.csv")

names(df1) <- c("id","title","year","cover","director","author","actor","genre","date","country","language","runtime","othername","imdb","avgRate","rateNum","five","four","three","two","one","tags","intro")

# 设置随机种子，以确保结果可复现

set.seed(123)

# 随机抽取 100 个数据

sample\_df <- df1[sample(nrow(df1), 100), ]

# 计算基本统计量（针对数值型列，如 avgRate, rateNum, runtime 等）

summary(sample\_df)

（b） 请对之前生成的数据（学号和电影编号的末尾两位数字相同的数据）计算其统计量。

summary(df)

（c） 请结合如下基于二项分布的中心极限定理验证函数，编写函数分别对（原始数据，9000+行）电影时长和评分变量进行多次抽样，观察统计量的变化（验证中心极限定理）。

文本

中度可信度描述已自动生成

sim.clt <- function(m = 100, n = 10, df1, var\_name) {

# 获取目标变量（runtime 或 avgRate）

var\_data <- df1[[var\_name]]

# 确保数据是数值型并去除 NA 值

var\_data <- na.omit(as.numeric(var\_data)) # 移除 NA 并确保为数值型

if (length(var\_data) < n) {

stop("数据的长度小于抽样大小，请检查数据。")

}

# 存储每次抽样的均值

sample\_means <- numeric(m)

# 设置随机种子，以确保结果可复现

set.seed(123)

# 抽样并计算均值

for (i in 1:m) {

# 随机抽取 n 个样本（放回抽样）

sample\_data <- sample(var\_data, n, replace = TRUE)

# 计算样本均值

sample\_means[i] <- mean(sample\_data)

}

# 绘制均值的分布，观察是否趋近于正态分布

hist(sample\_means, prob = TRUE, breaks = 20,

main = paste("CLT for", var\_name, "(n=", n, "m=", m, ")"),

col = "lightblue", border = "black")

# 添加标准正态分布曲线

curve(dnorm(x, mean = mean(sample\_means), sd = sd(sample\_means)),

add = TRUE, col = "red")

}

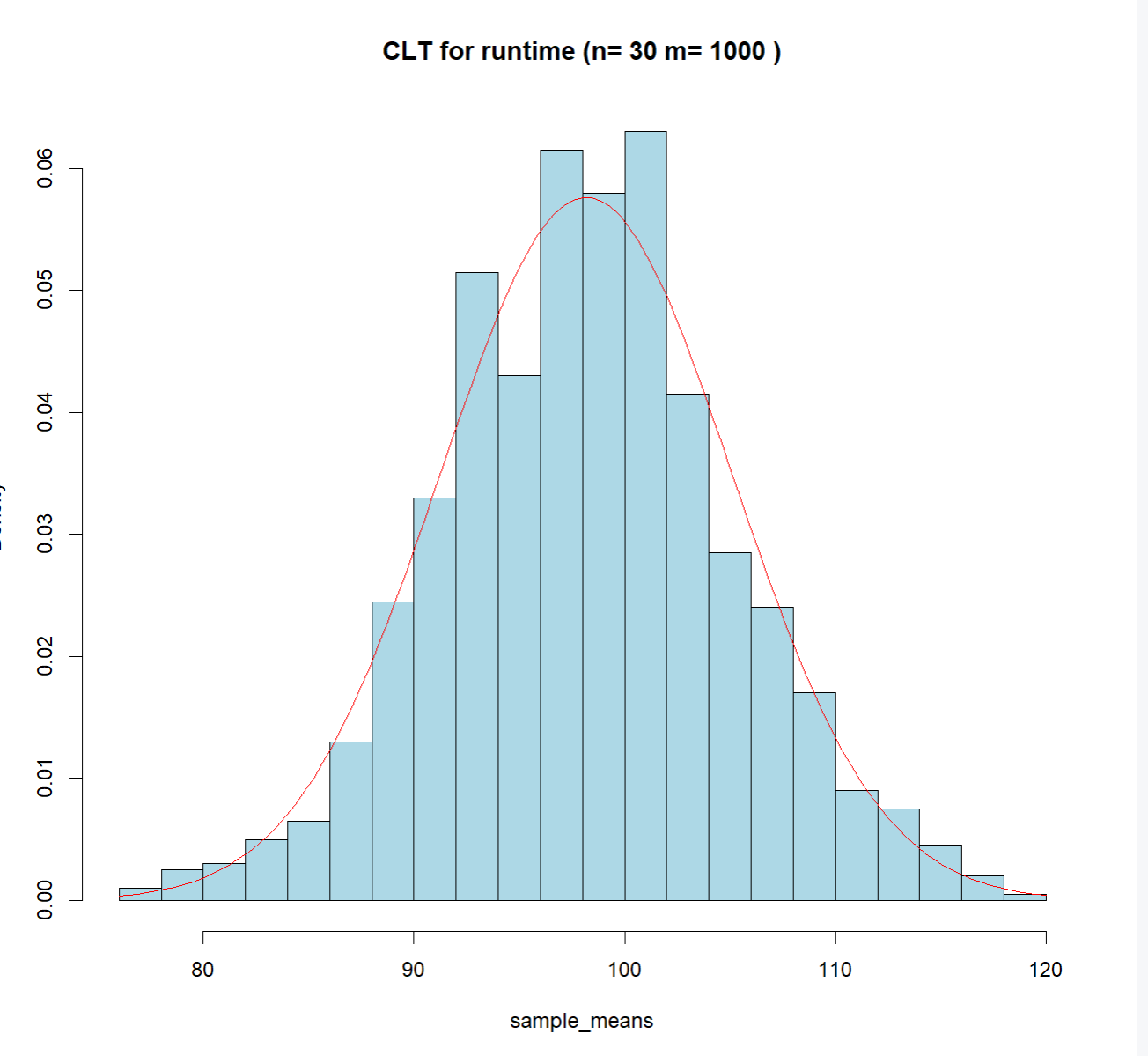
# 对9000+行数据进行抽样（movie runtime 和 avgRate）

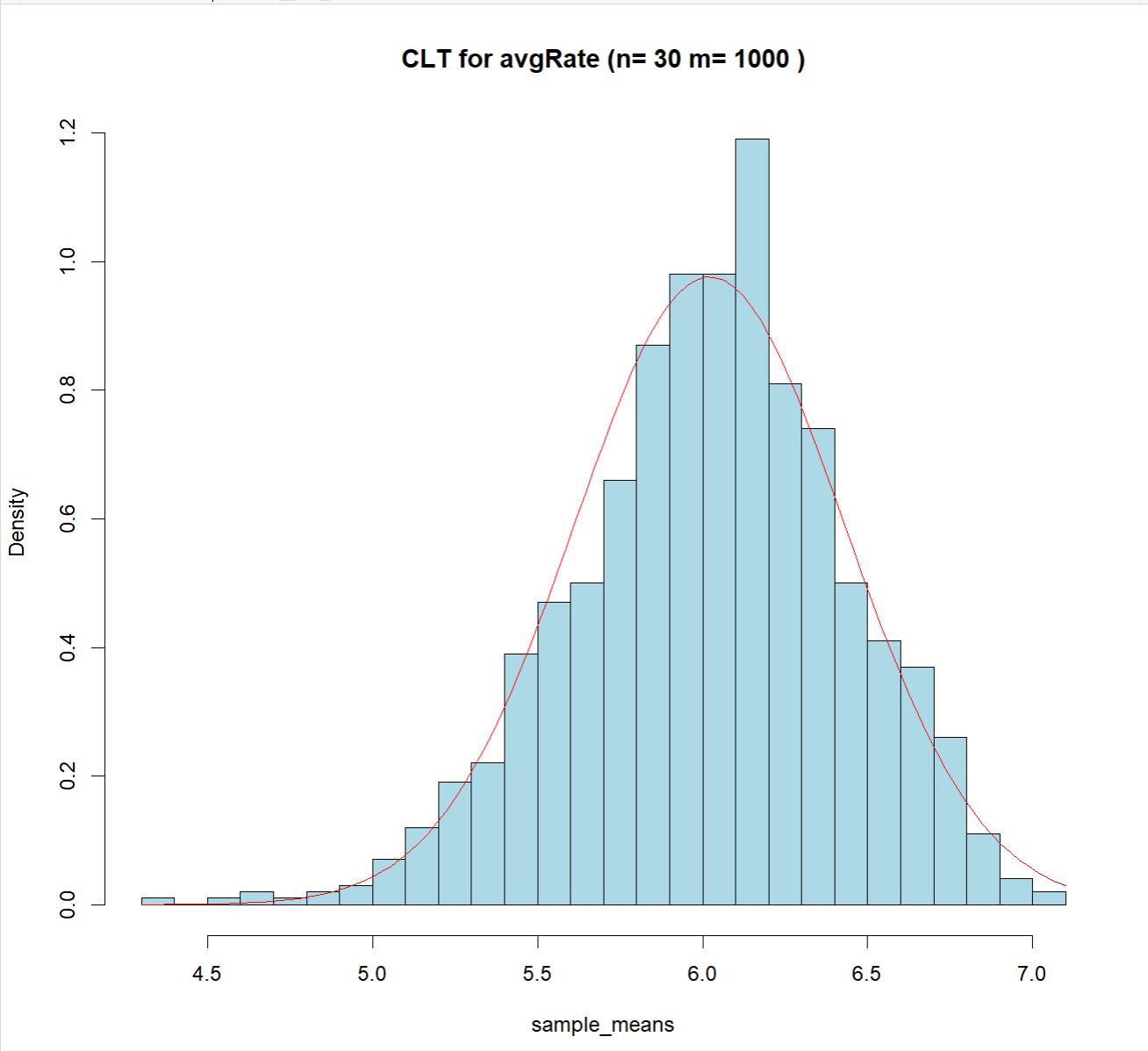
# 对电影时长（runtime）的抽样

sim.clt(m = 1000, n = 30, df1 = df1[9001:nrow(df1), ], var\_name = "runtime")

# 对电影评分（avgRate）的抽样

sim.clt(m = 1000, n = 30, df1 = df1[9001:nrow(df1), ], var\_name = "avgRate")





7）. 请结合如下标准正态分布概率表和t分布临界值表代码，构造 分布和F分布的时的临界值。

文本

描述已自动生成

# 卡方分布临界值表

alpha = 0.05 # 尾部概率

# 自由度范围

df = 1:30 # 假设自由度从 1 到 30

# 计算卡方分布的临界值

chi\_squared\_critical\_values = qchisq(1 - alpha, df)

# 打印卡方分布临界值表

data.frame("Degrees of Freedom" = df, "Critical Value (alpha=0.05)" = chi\_squared\_critical\_values)

# F 分布临界值表

alpha = 0.05 # 尾部概率

# 自由度1和自由度2范围

df1 = 1:10 # 自由度1范围c'le'rcler

df2 = 1:10 # 自由度2范围

# 计算 F 分布的临界值

f\_critical\_values = matrix(NA, nrow = length(df1), ncol = length(df2))

for (i in 1:length(df1)) {

for (j in 1:length(df2)) {

f\_critical\_values[i, j] = qf(1 - alpha, df1[i], df2[j])

}

}

# 打印 F 分布临界值表

f\_critical\_values\_df = data.frame(f\_critical\_values)

colnames(f\_critical\_values\_df) = paste("df2 =", df2)

rownames(f\_critical\_values\_df) = paste("df1 =", df1)

f\_critical\_values\_df

**四、实验结果：**

1. 完成了什么实验？

离散随机变量的概率计算：通过超几何分布计算了不合格品数的概率

二项分布的概率与统计量：计算了二项分布下特定不合格品数量的概率，并求得了均值、方差与标准差。

正态分布下合格品概率的计算：通过正态分布计算了制造过程中合格品的概率。

正态分布应用于考试成绩：计算了考生外语成绩落在指定区间内的概率。

样本统计量的计算：计算了铆钉头部直径的样本均值和标准差。

数据抽样与统计量计算：通过 sample 函数对电影数据进行了随机抽样，并计算了基本统计量。

中心极限定理验证：通过编写抽样函数，验证了中心极限定理，并观察了统计量随抽样次数的变化。

分布临界值计算：通过标准正态分布、t分布、卡方分布和F分布的临界值表的构造，计算了相应的临界值。

2. 实验的结论

在离散随机变量和二项分布的实验中，能够通过概率公式准确计算各类概率。

正态分布和二项分布的均值与方差可有效描述概率分布的特征。

通过中心极限定理验证，可以观察到随着样本数增大，样本均值逐渐逼近正态分布。

F分布、卡方分布等分布的临界值计算为进行假设检验和统计推断提供了工具。

3. 遇到了什么困难，怎么解决的？

困难：在验证中心极限定理时，sample 函数中抽样参数的设定可能导致抽样大小不足或参数不匹配。

解决：通过检查抽样参数并确保每次抽样的大小不超过数据的总行数，调整了抽样代码并避免了这种错误。

4. 对实验有何认识？

概率与统计是理解和分析数据的重要工具。通过实验能够验证理论并深入理解各类分布的特性。

随着样本量增大，统计量会趋向稳定，证明了中心极限定理的实际应用。

使用R语言进行统计分析能够高效地计算和展示不同分布的特性，帮助做出更精确的统计推断。