Stat 361 - Recitation 9

MC Methods for Statistical Inference

Orçun Oltulu

23-24 / 04 / 2020

Monte Carlo Methods for Inferential Statistics

Monte Carlo simulations in statistics are computer experiments involving random sampling from probablility distributions to study properties of statistical methods. We use Monte Carlo (MC) simulation for the following purposes:

- Inference when the distribution of the test statistic is not known analytically.
- Comparing or assessing performances of inferential methods when parametric assumptions are violated.
- Testing the null and alternative hypotheses under various conditions.
- Evaluating the performance of inferential methods (e.g. power calculations).
- Comparing the quality of estimators.

Basic Monte Carlo:

- We estimate the distribution of the statistic by randomly sampling from the same population of interest and recording the value of statistic for each sample.
- The observed values of the statistic for these samples are used to estimate the distribution of statistic.
- By this way, we do not need to use some assumptions to make an inference about the statistic. (eg. In hypothesis testing for mean, we do not need to make an assumption that sample mean is approximately distributed normal. Since, in Monte Carlo method, we use emprical distribution of statistic, which is coming from random samples.)

Steps:

1. Determine the pseudo-population or model that represents the true population of interest.
2. Obtain a random sample from the pseudo-population.
3. Calculate a value for the test statistic of interest and store it.
4. Repeat steps 2 and 3 for trials.
5. Use the values found in step 4 to study the distribution of the statistic.

Question 1:

Suppose that X_1, X_2, X_3, X_4 are i.i.d. from a standard normal distribution. Estimate the sum of their squares $Y = X_1^2 + X_2^2 + X_3^2 + X_4^2$, E(Y).

```
M <- 1000
y <- numeric(M)
for(i in 1:M){
  x <- rnorm(4,0,1)
  y[i] <- sum(x^2)
}

estimate <- mean(y)
estimate
```

```
## [1] 4.121819
```

Also we know that

$$Y \sim \chi_4^2 \quad E(Y) = 4$$

Question 2:

Suppose that you are tossing a biased coin ($P_H = 0.6$) 15 times. What is the probability to obtain more than 10 Heads?

Binomial Probability:

Let random variable X is the number of Heads obtained and $X \sim Binomial(n, p)$

$$P(X > 10) = P(X = 11) + ... + P(X = 15)$$

Monte Carlo Simulation:

1. Draw 15 random variables x_i from Uniform(0,1), i = 1,…,15

2. Define "Head" if $x_i \leq 0.6$,"Tail" otherwise
3. Count the number of Heads, if the number of Heads are more than 10, increse the count k.
4. Repeat step 2 and 3 N times
5. The desired probability is approximately k/N

```
exact_prob <- 0
for(j in 11:15){
  exact_prob <- exact_prob + dbinom(j,15,0.6)
}
cat("Exact probability is",round(exact_prob,5),"\n","\n")
```

```
## Exact probability is 0.21728
##
```

```
cat("Monte Carlo Simulation","\n")
```

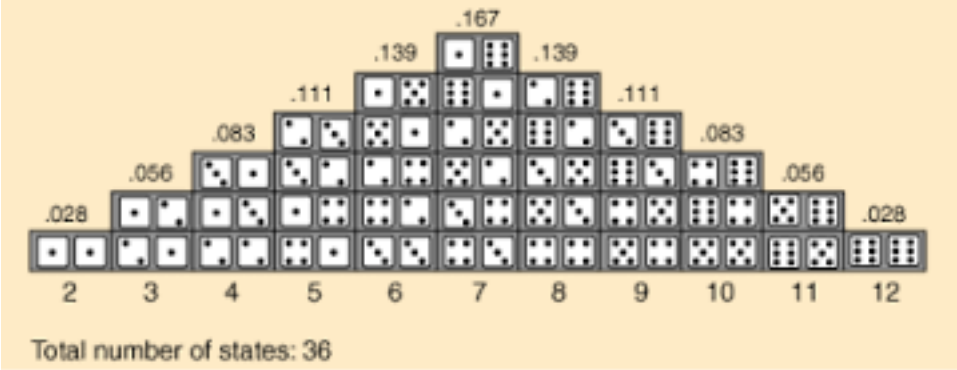
```
## Monte Carlo Simulation
```

```
N <- 10^seq(1,5)

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

  k <- 0 # counter

  for(i in 1:N[j]){
    x <- runif(15)
    num_heads <- sum(x<=0.6)

    # Alternatively
    # x <- sample(c("H","T"), size = 15, replace = T)
    # num_heads <- sum(x == "H")

    if(num_heads > 10){
      k <- k + 1
    }
  }
  est_prob <- k/N[j]
  cat("When N is",N[j],"estimated probability is",est_prob,"\n")
}
```

```
## When N is 10 estimated probability is 0.4
## When N is 100 estimated probability is 0.21
## When N is 1000 estimated probability is 0.231
## When N is 10000 estimated probability is 0.2257
## When N is 1e+05 estimated probability is 0.2172
```

Question 3:

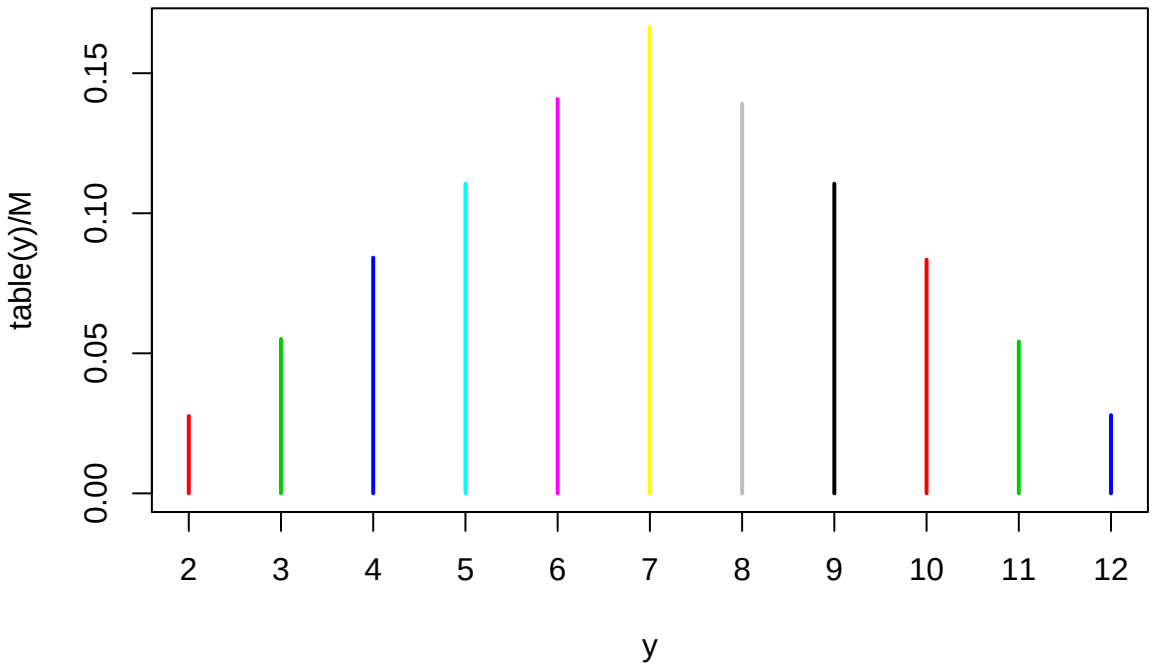
Consider, we want to calculate the probability of a particular sum of the throw of two dice (with each die having values one through six). In this particular case, there are 36 combinations of dice rolls:




```
M <- 10^5
y <- numeric()

for(i in 1:M){
  x <- sample(c(1:6),2,replace=TRUE)
  y[i] <- sum(x)
}

table(y)/M
```

```
## y
##       2       3       4       5       6       7       8       9      10      11
## 0.02763 0.05511 0.08415 0.11057 0.14079 0.16651 0.13912 0.11055 0.08346 0.05420
##      12
## 0.02791
```

```
plot(table(y)/M, col = c(2:12))
```



Question 4:

Consider, we want to calculate the bias of the estimators of location and scale parameters of normal distribution ($\mu = 3,\ \sigma^2 = 5$) when we have sample with sizes 5, 30 and 100.

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}; \quad -\infty < x < \infty$$

We know that bias of estimator T is given by

$$\text{Bias}(\hat{T}) = E(\hat{T}) - T$$

Empirically, we can find the estimated bias by using

$$\hat{\text{Bias}}(\hat{T}) = \frac{\sum_{j=1}^{M} \hat{T}_j}{M} - T$$

Where $\hat{T}_j$ is the jth estimator coming from jth sample (we have M different sample mean to calculate the empirical distribution of the $\hat{T}$)

We can obtain the M different samples by using the true distribution of the $(x_1, x_2, ..., x_n)$ which is the normal distribution with mean μ and variance σ^2

Let

$$T_1 = \mu \quad \rightarrow \quad \hat{T_1} = \bar{x} \quad \text{for location parameter}$$

$$T_2 = \sigma^2 \quad \rightarrow \quad \hat{T_2} = S^2 \quad \text{for scale parameter}$$

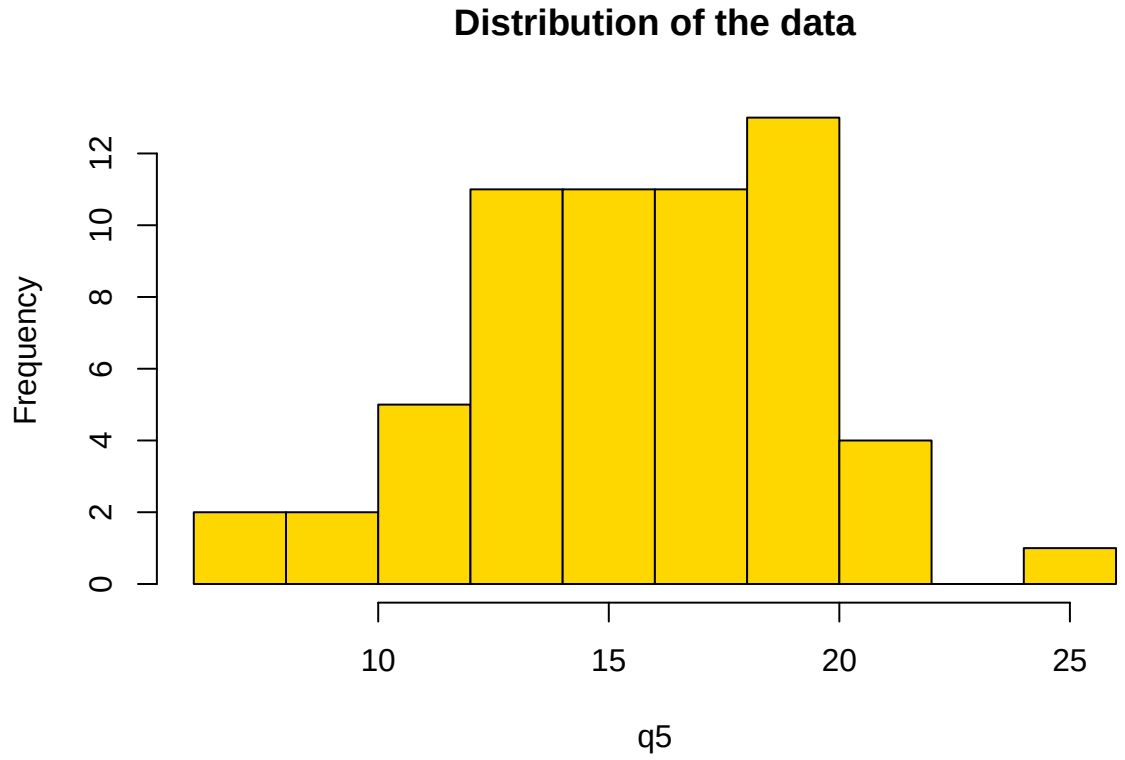
```
T1 <- 10; T2 <- 5 # set parameters
n <- c(5,30,100) # sample sizes
M <- 10^5

# creating a matrix to store bias values for both t1&t2
bias <- matrix(0,ncol = length(n), nrow = 2,
               dimnames = list(c("T1 bias","T2 bias"),
                               c(paste(rep("n",length(n)),n,sep=""))))

m <- numeric(M)
v <- numeric(M)
for(j in 1:length(n)){
  for(i in 1:M){
    x <- rnorm(n[j],T1,sqrt(T2))

    m[i] <- mean(x)
    v[i] <- var(x)
  }

  bias[1,j] <- mean(m) - T1
  bias[2,j] <- mean(v) - T2
}

round(bias,4)
```

```
##              n5     n30   n100
## T1 bias -0.0010 -0.0001 -2e-04
## T2 bias -0.0116 -0.0025  4e-04
```

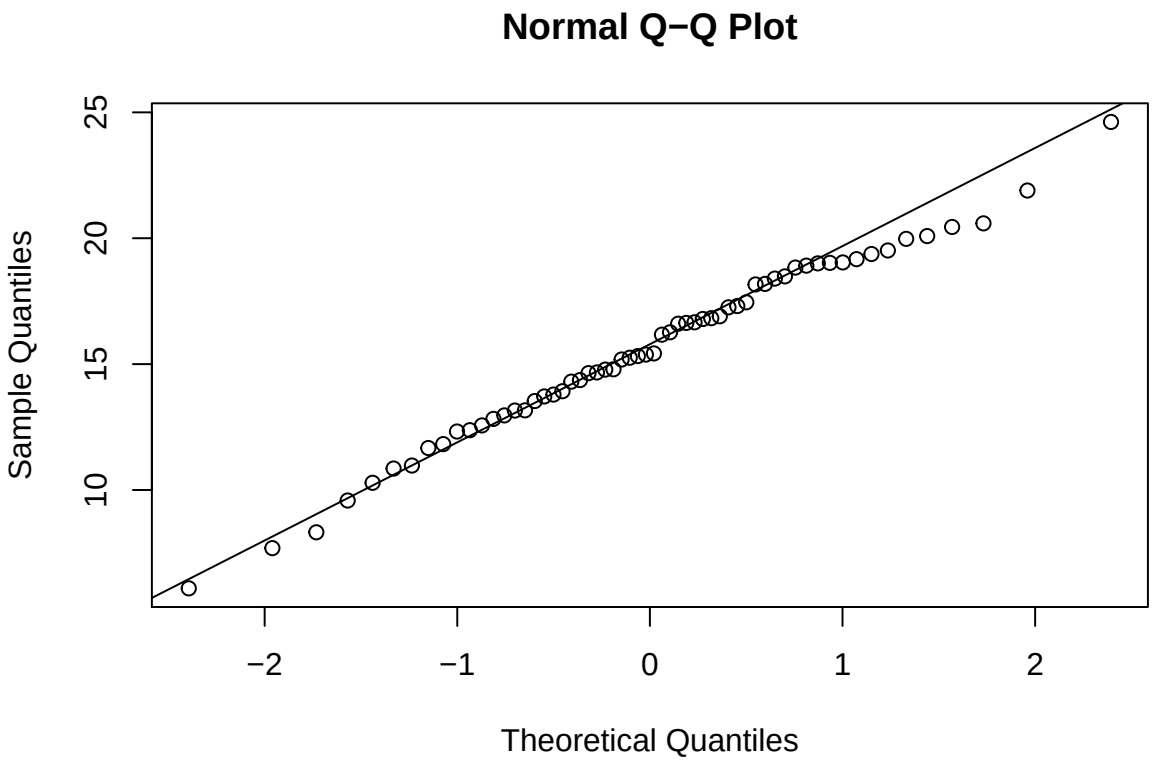
Question 5: Confidence Interval Estimation

Let's assume that you wanted to work on COVID-19 and wanted estimate a confidence interval for the mean recovery time (in days). Then you collect a random sample of size 60, you can find it in *q5.txt*. Write an R function to run a monte carlo simulation for estimating that confidence interval with 95% significance level.

```r
#read data into R.
q5 <- as.vector(read.table("q5.txt",header = T)$X)

#check its distributional properties
hist(q5, main = "Distribution of the data", col = "Gold")
```




```r
#it looks symmetric
qqnorm(q5);qqline(q5)
```




```r
shapiro.test(q5)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  q5
## W = 0.98914, p-value = 0.8719
```

```r
# We can conclude that the data comes from Normal
```

```r
recovery_time_CI <- function(data, M = 10^5, alpha = 0.05){
  # we are assuming that the population has from normal shape

  n <- length(data)
  mu <- mean(data);sig <- sd(data)

  storage <- matrix(0, nrow = M, ncol = 2,
                    dimnames = list(c(1:M),c("Lower","Upper")))

  for(i in 1:M){
    x <- rnorm(n,mu,sig)
```

```
    x_bar <- mean(x)
    x_sd <- sd(x)

    storage[i,1] <- x_bar - qnorm(1-alpha) * (x_sd/sqrt(n))
    storage[i,2] <- x_bar + qnorm(1-alpha) * (x_sd/sqrt(n))

  }
  out <- list(CI = apply(storage,2,mean),
              conf_level = 1-alpha)
  return(out)
}

recovery_time_CI(q5)
```

```
## $CI
##    Lower    Upper
## 14.80704 16.34155
##
## $conf_level
## [1] 0.95
```