# stat359 A2 wducharme

Wesley Ducharme

2024-10-07

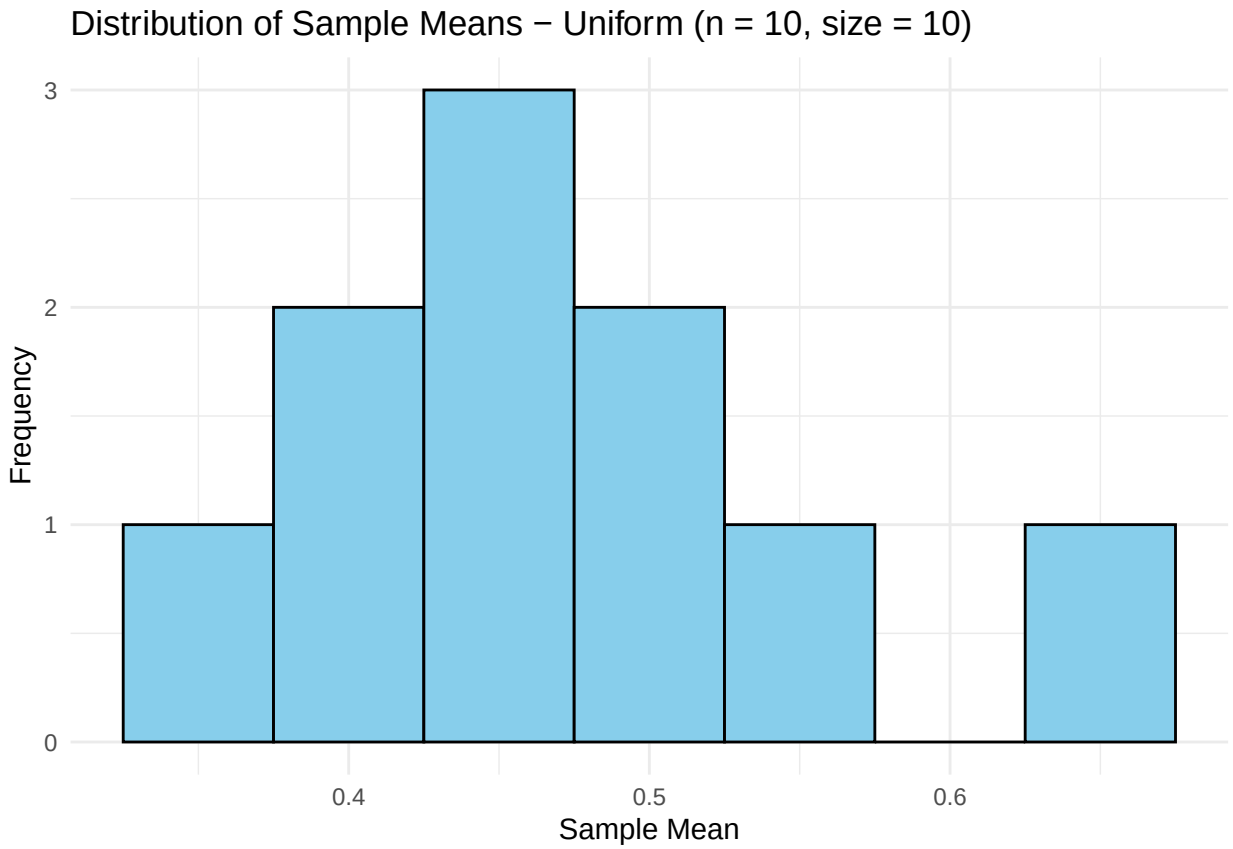
## Question 1:

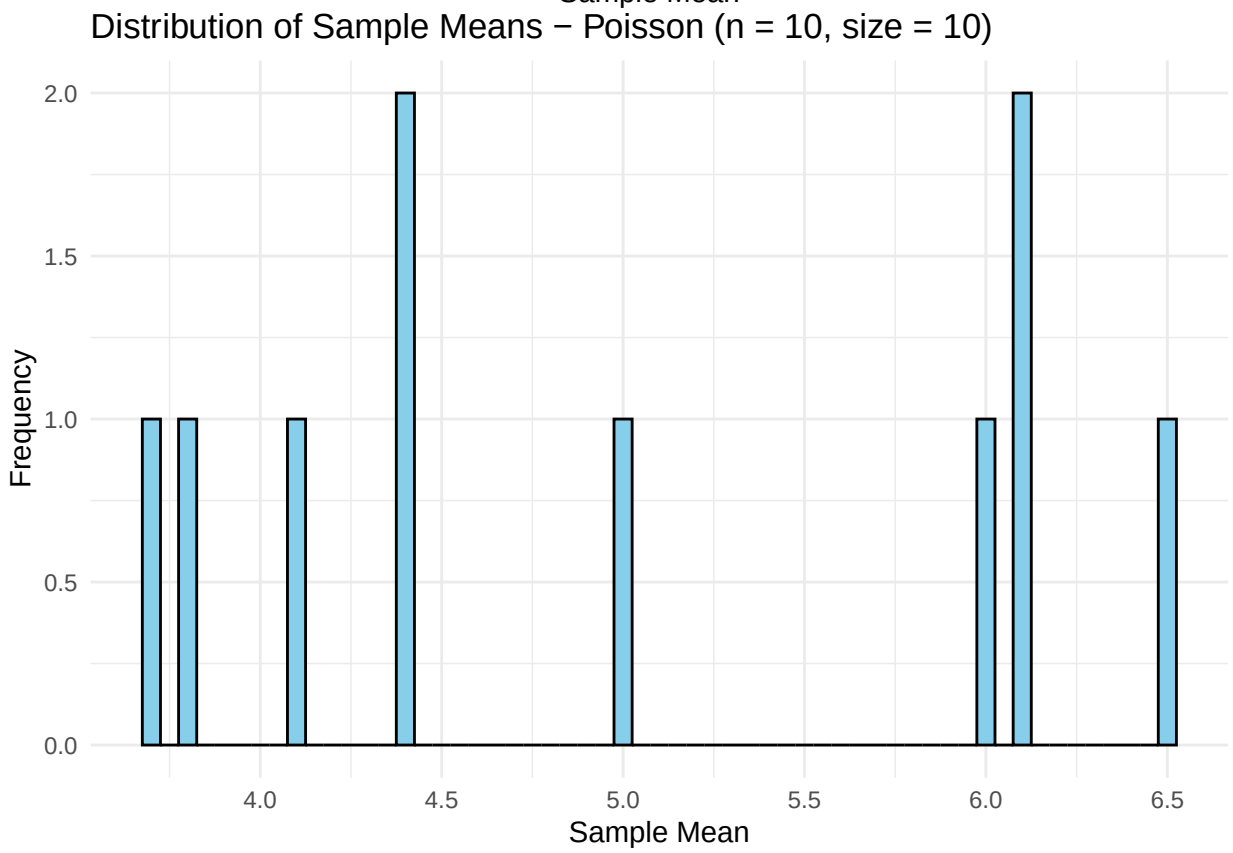
```
# Parameters for the distributions
a <- 0
b <- 1
lambda <- 5
p <- 0.20

# Sample sizes and number of samples
sample_sizes <- c(10, 20, 50, 100)
num_samples <- c(10, 100, 1000)

# Lists to store sample means
uniform_means <- list()
poisson_means <- list()
bernoulli_means <- list()

# Generates the sample means
for (n in num_samples) {
  uniform_means[[paste0("n_samples_", n)]] <- list()
  poisson_means[[paste0("n_samples_", n)]] <- list()
  bernoulli_means[[paste0("n_samples_", n)]] <- list()

  for (size in sample_sizes) {
    # Storeing the sample means for each distribution and sample size
    uniform_means[[paste0("n_samples_", n)]][[paste0("size_", size)]] <-
↪   apply(replicate(n, runif(size, min = a, max = b)), 2, mean)
    poisson_means[[paste0("n_samples_", n)]][[paste0("size_", size)]] <-
↪   apply(replicate(n, rpois(size, lambda = lambda)), 2, mean)
    bernoulli_means[[paste0("n_samples_", n)]][[paste0("size_", size)]] <-
↪   apply(replicate(n, rbinom(size, 1, prob = p)), 2, mean)
  }
}

library(ggplot2)
```

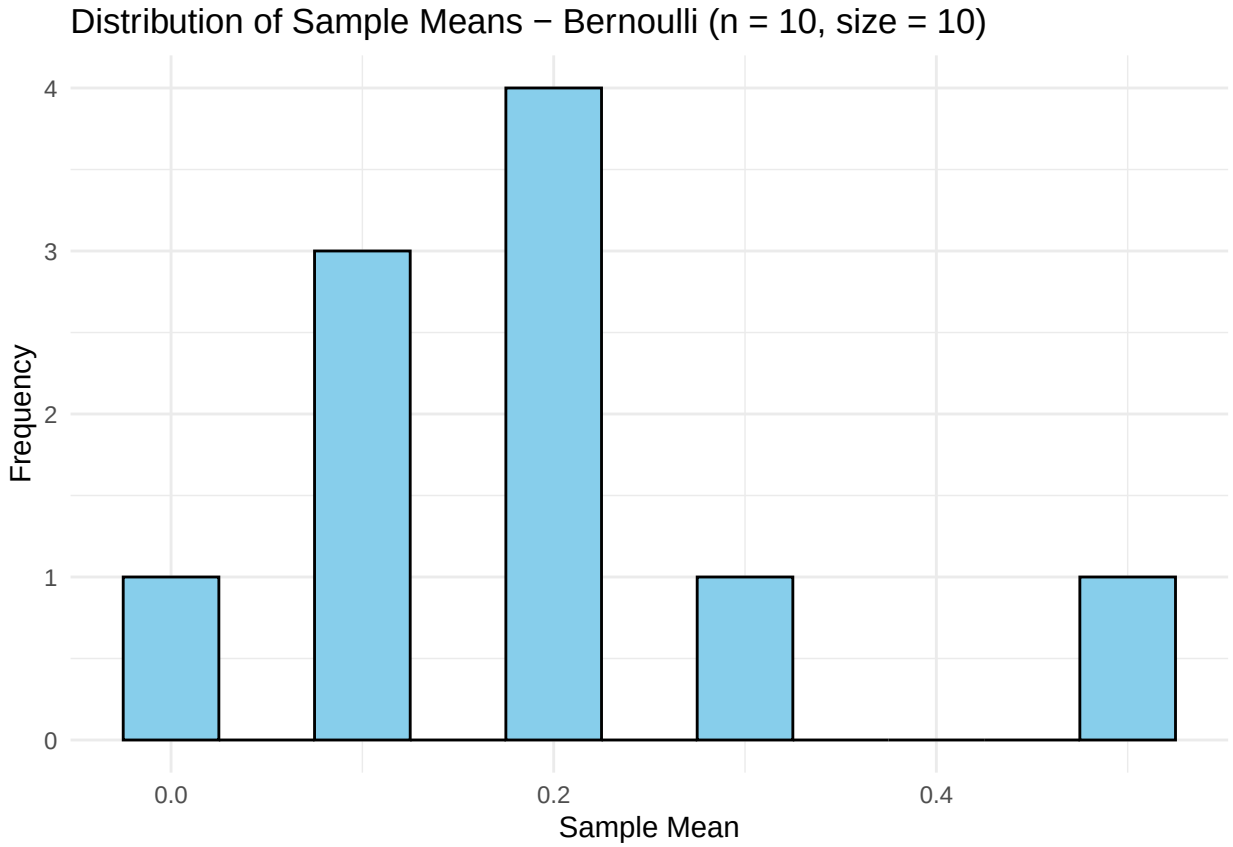
```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

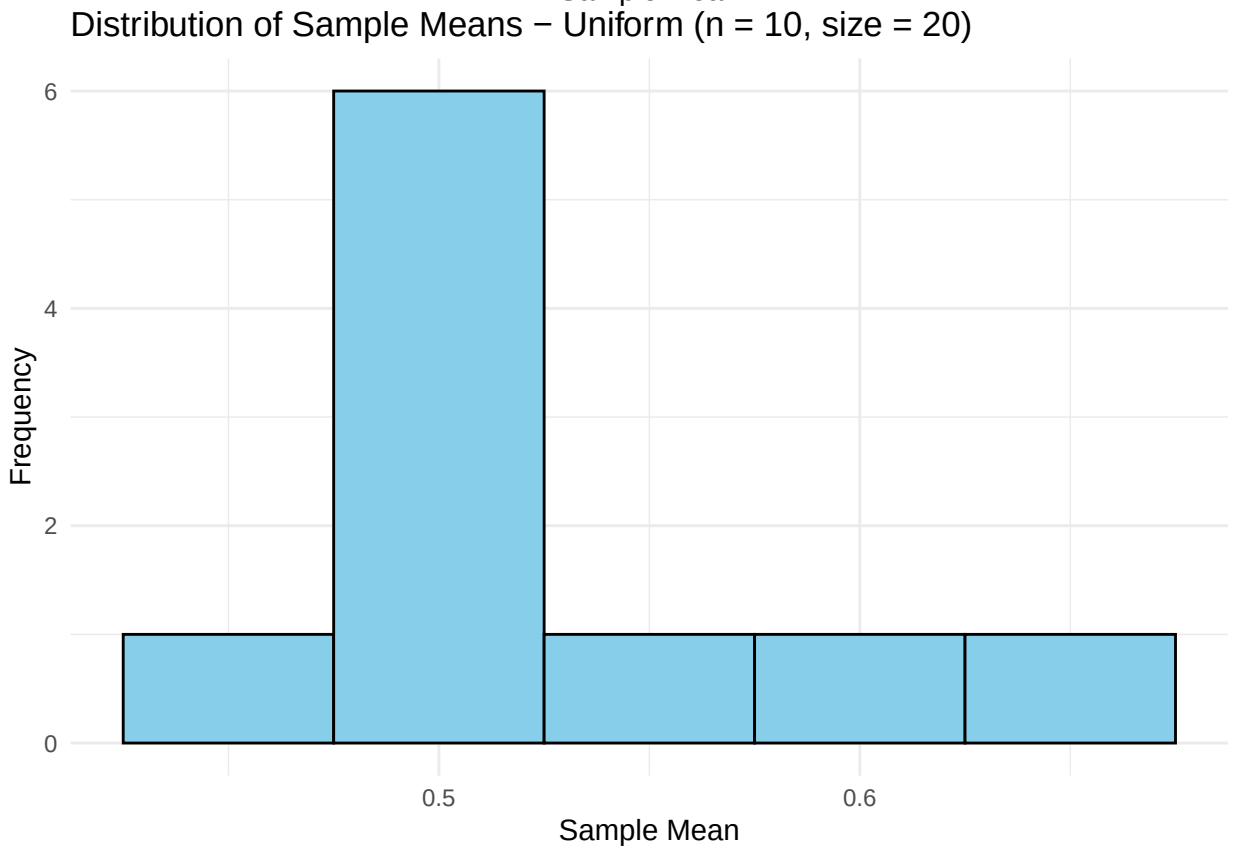
a)

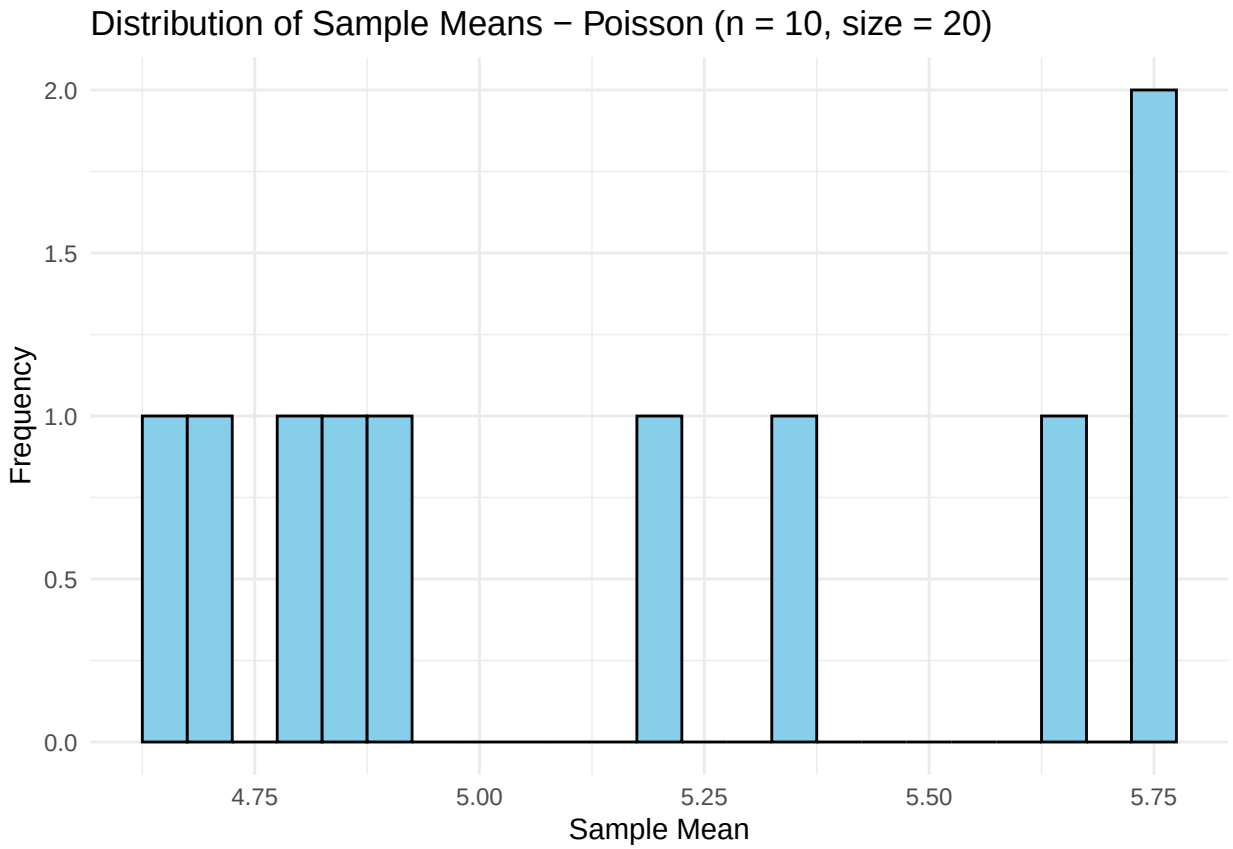
```r
# Function to create and save a histogram of sample means
plot_histogram <- function(sample_means, dist_name, n, size) {
  data <- data.frame(means = sample_means)
  p <- ggplot(data, aes(x = means)) +
    geom_histogram(binwidth = 0.05, color = "black", fill = "skyblue") +
    ggtitle(paste0("Distribution of Sample Means - ", dist_name,
                   " (n = ", n, ", size = ", size, ")")) +
    xlab("Sample Mean") +
    ylab("Frequency") +
    theme_minimal()

  # Display the plot
  print(p)
}

par(mfrow=c(6,6))

# Loop through all combinations of distributions, sample sizes, and number of samples
for (n in num_samples) {
  for (size in sample_sizes) {
    # Uniform distribution
    plot_histogram(uniform_means[[paste0("n_samples_", n)]][[paste0("size_", size)]],
                   "Uniform", n, size)

    # Poisson distribution
    plot_histogram(poisson_means[[paste0("n_samples_", n)]][[paste0("size_", size)]],
                   "Poisson", n, size)

    # Bernoulli distribution
    plot_histogram(bernoulli_means[[paste0("n_samples_", n)]][[paste0("size_", size)]],
                   "Bernoulli", n, size)
  }
}
```

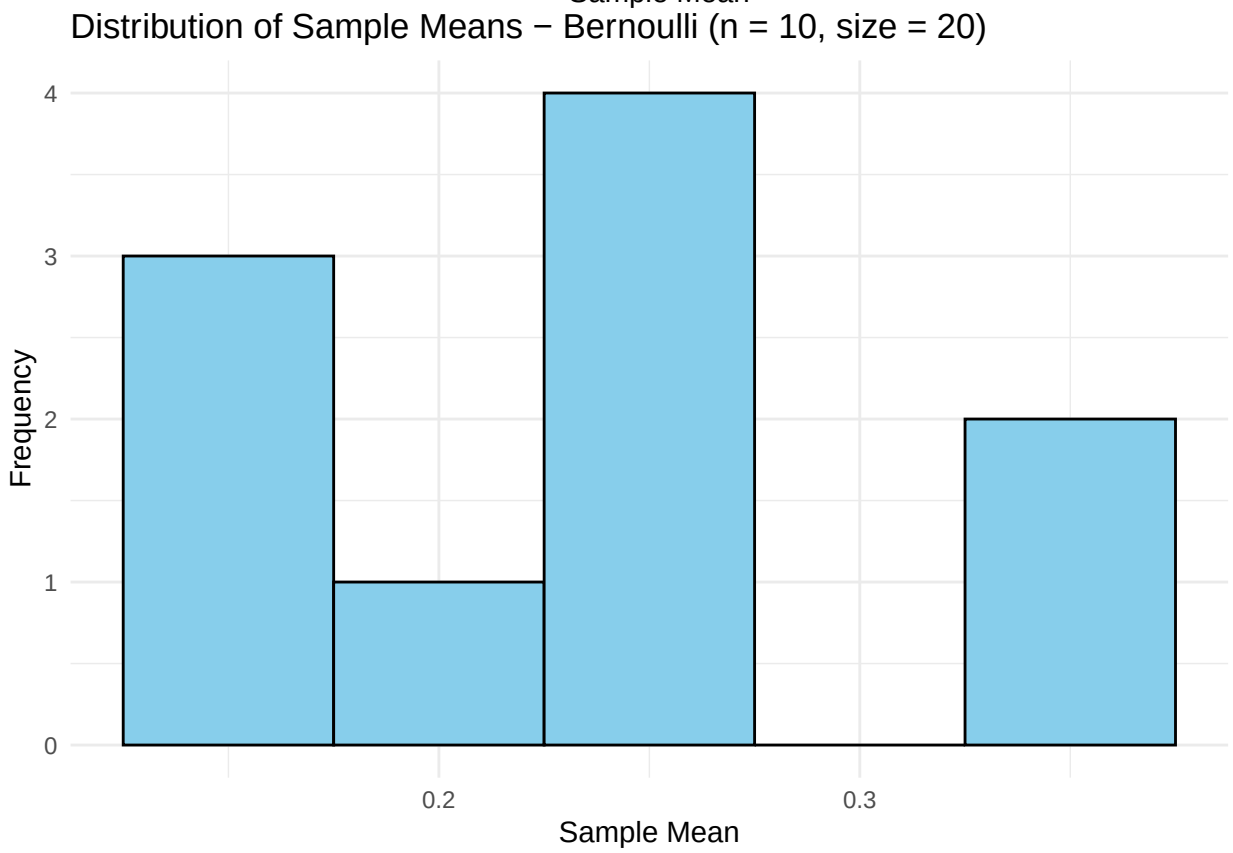

Distribution of Sample Means − Uniform (n = 10, size = 10)
Frequency
Sample Mean
0
1
2
3
0.4
0.5
0.6

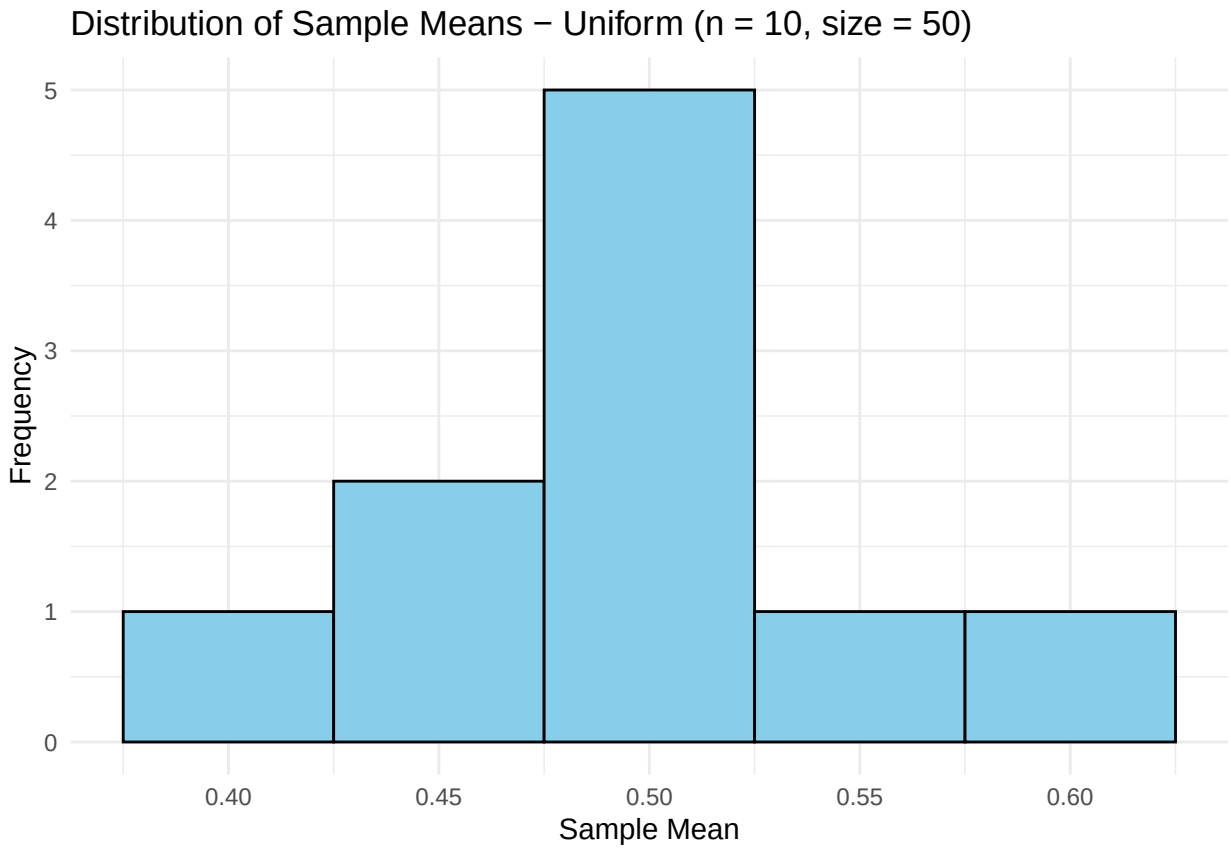

Distribution of Sample Means − Poisson (n = 10, size = 10)
Frequency
Sample Mean
0.0
0.5
1.0
1.5
2.0
4.0
4.5
5.0
5.5
6.0
6.5

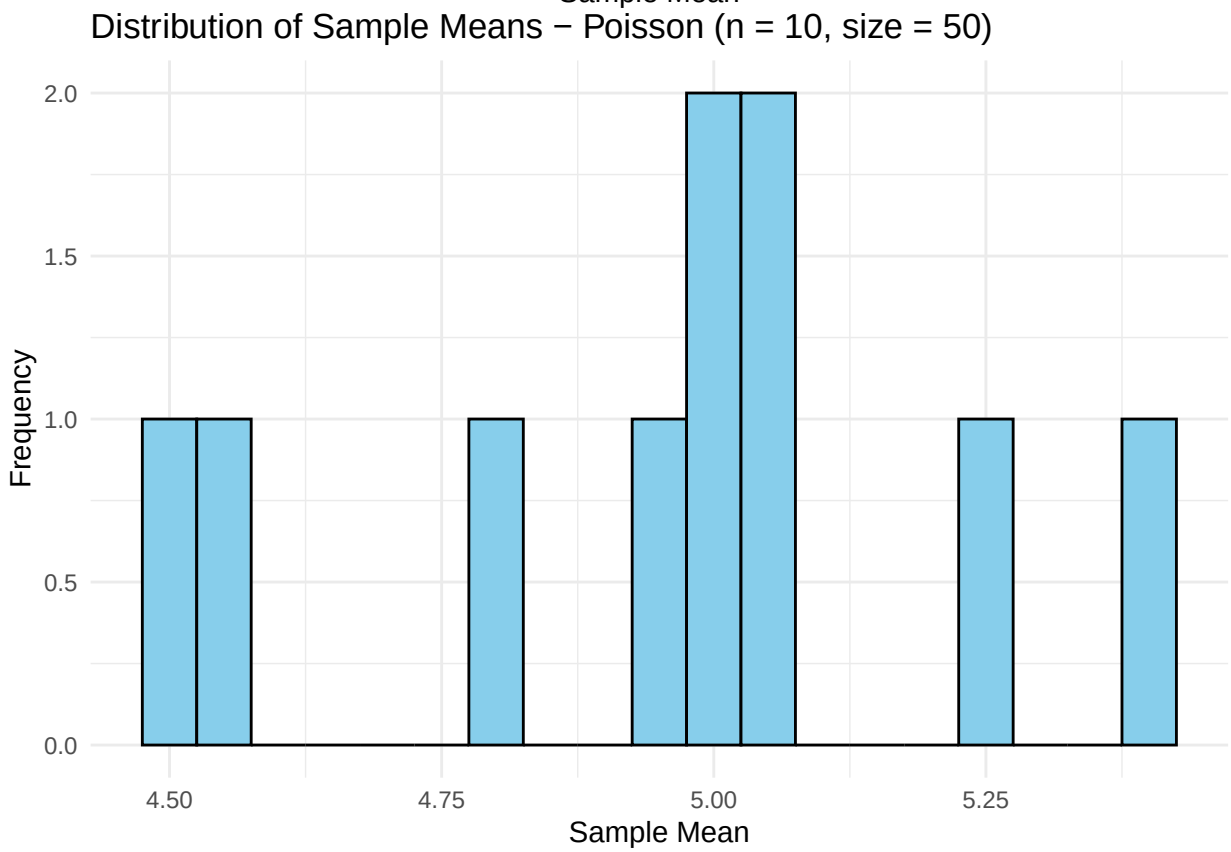

Distribution of Sample Means − Bernoulli (n = 10, size = 10)
Frequency
0
1
2
3
4
0.0
0.2
0.4
Sample Mean

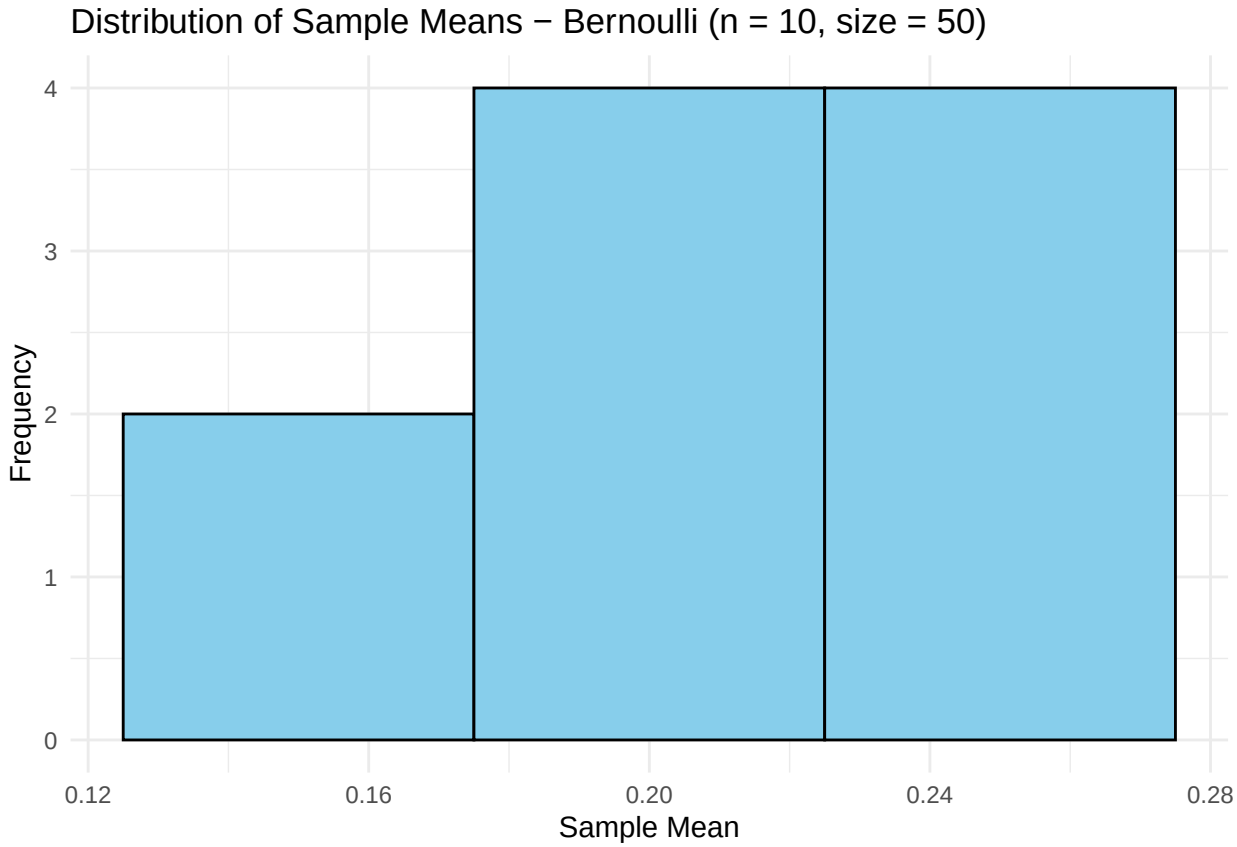

Distribution of Sample Means − Uniform (n = 10, size = 20)
Frequency
0
2
4
6
0.5
0.6
Sample Mean

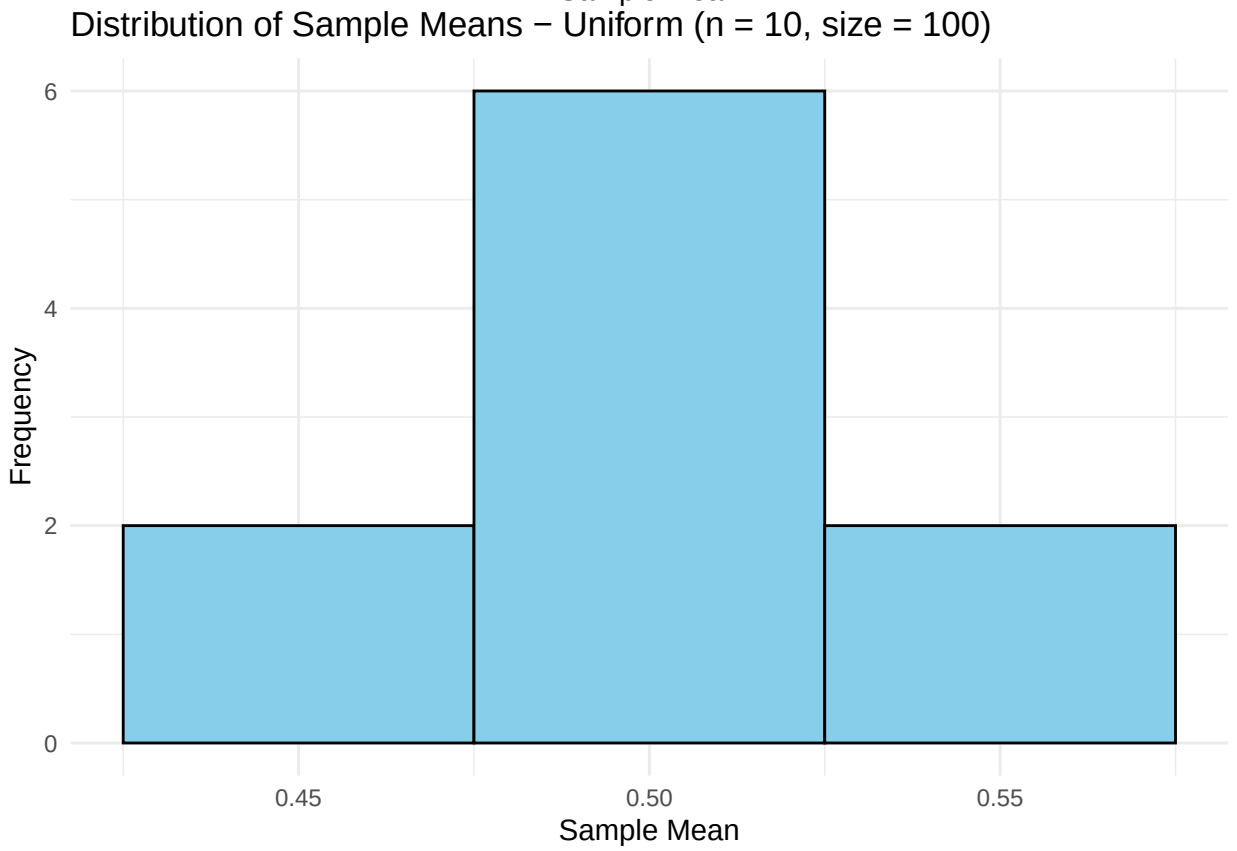

Distribution of Sample Means − Poisson (n = 10, size = 20)
Frequency
2.0
1.5
1.0
0.5
0.0
4.75
5.00
5.25
5.50
5.75
Sample Mean

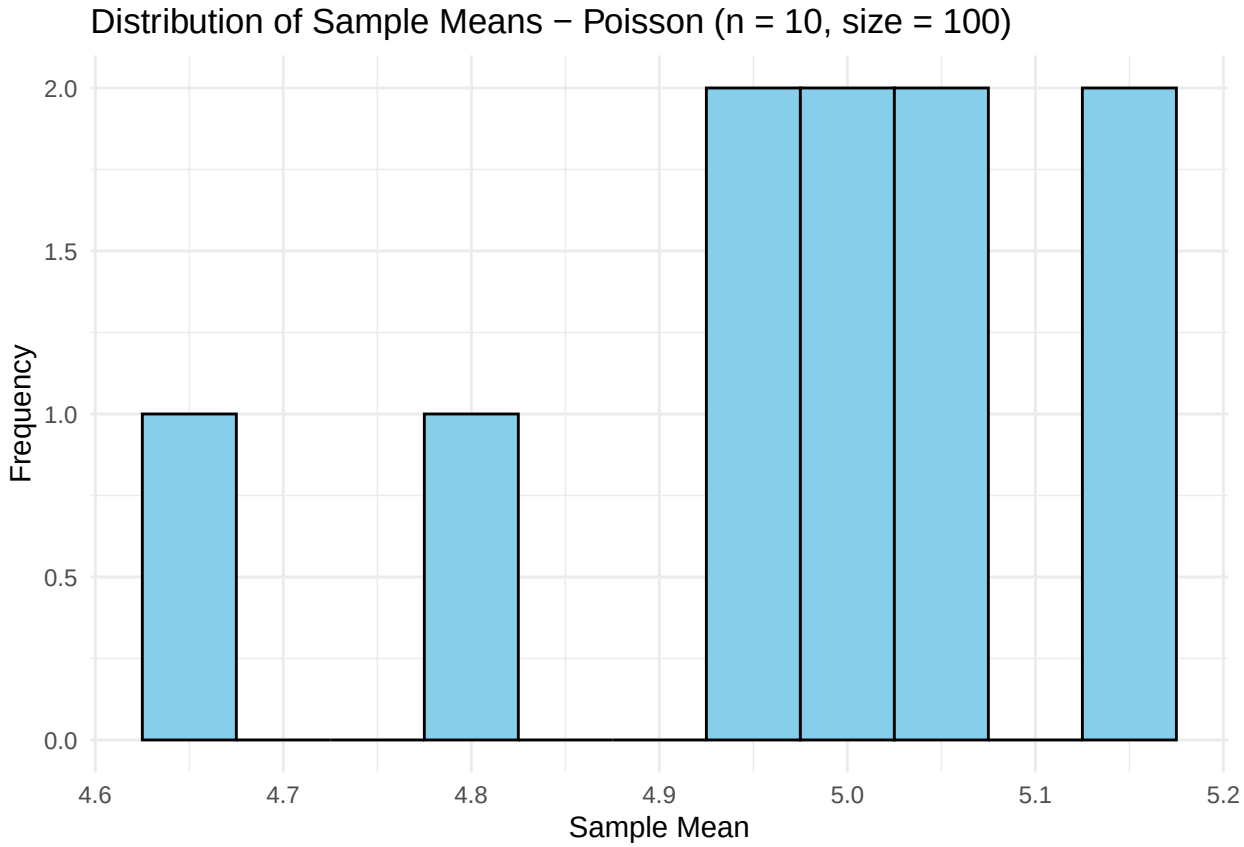

Distribution of Sample Means − Bernoulli (n = 10, size = 20)
Frequency
4
3
2
1
0
0.2
0.3
Sample Mean

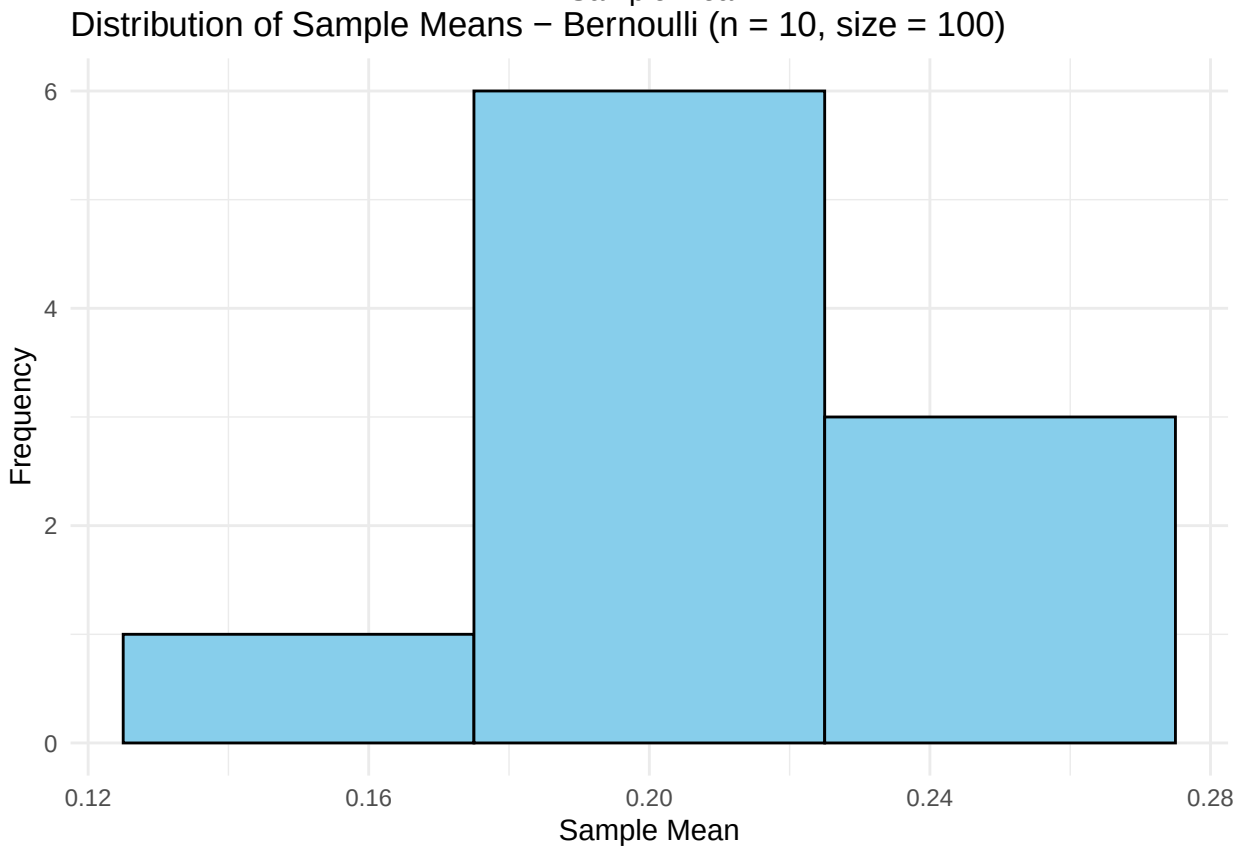

Distribution of Sample Means − Uniform (n = 10, size = 50)
Frequency
0
1
2
3
4
5
0.40
0.45
0.50
0.55
0.60
Sample Mean

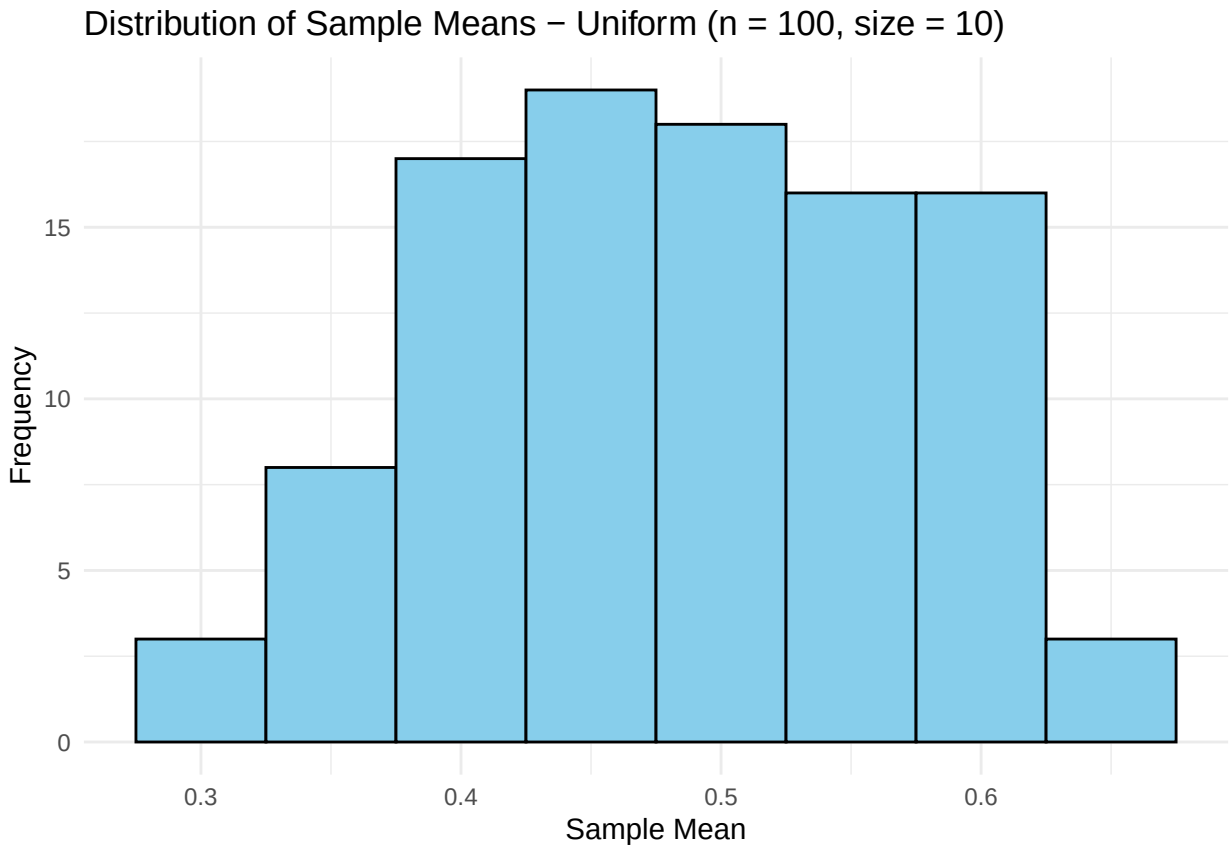

Distribution of Sample Means − Poisson (n = 10, size = 50)
Frequency
0.0
0.5
1.0
1.5
2.0
4.50
4.75
5.00
5.25
Sample Mean

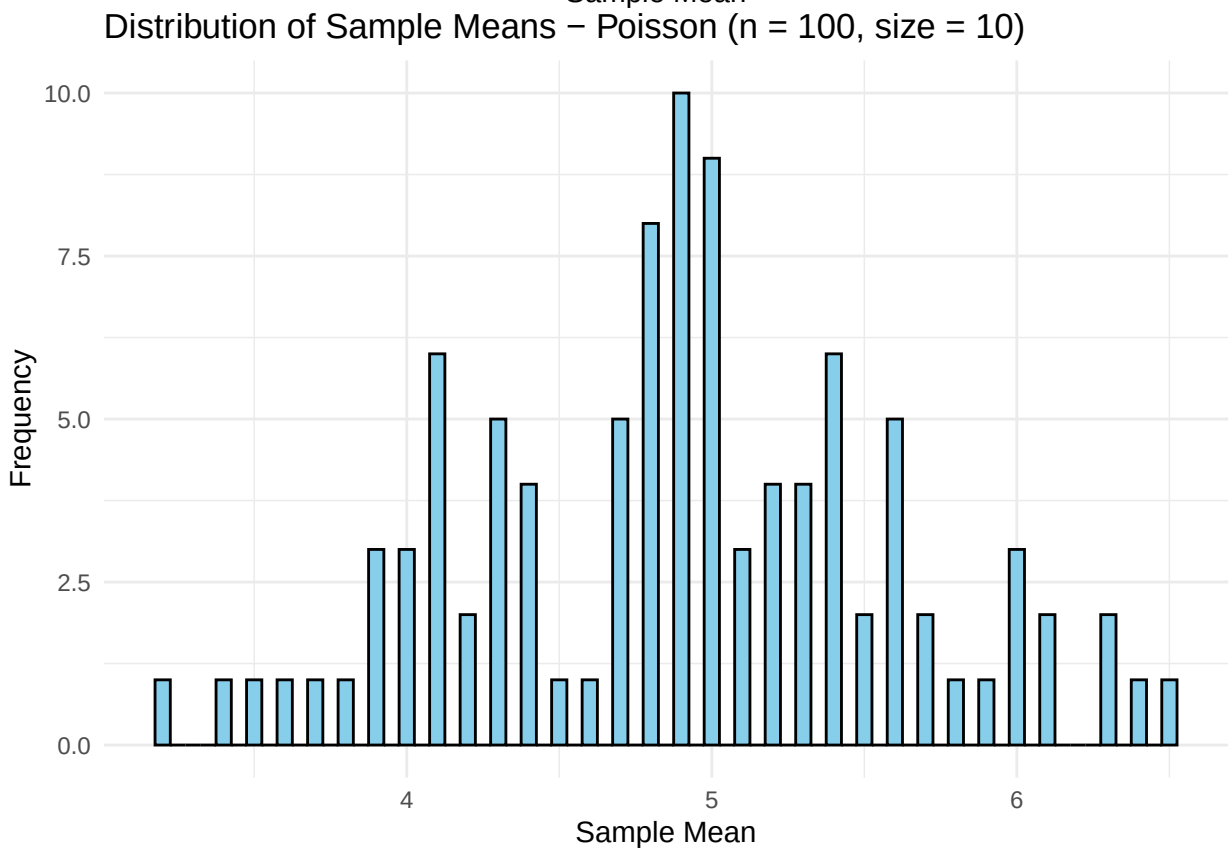

Distribution of Sample Means − Bernoulli (n = 10, size = 50)
Frequency
0
1
2
3
4
0.12
0.16
0.20
0.24
0.28
Sample Mean

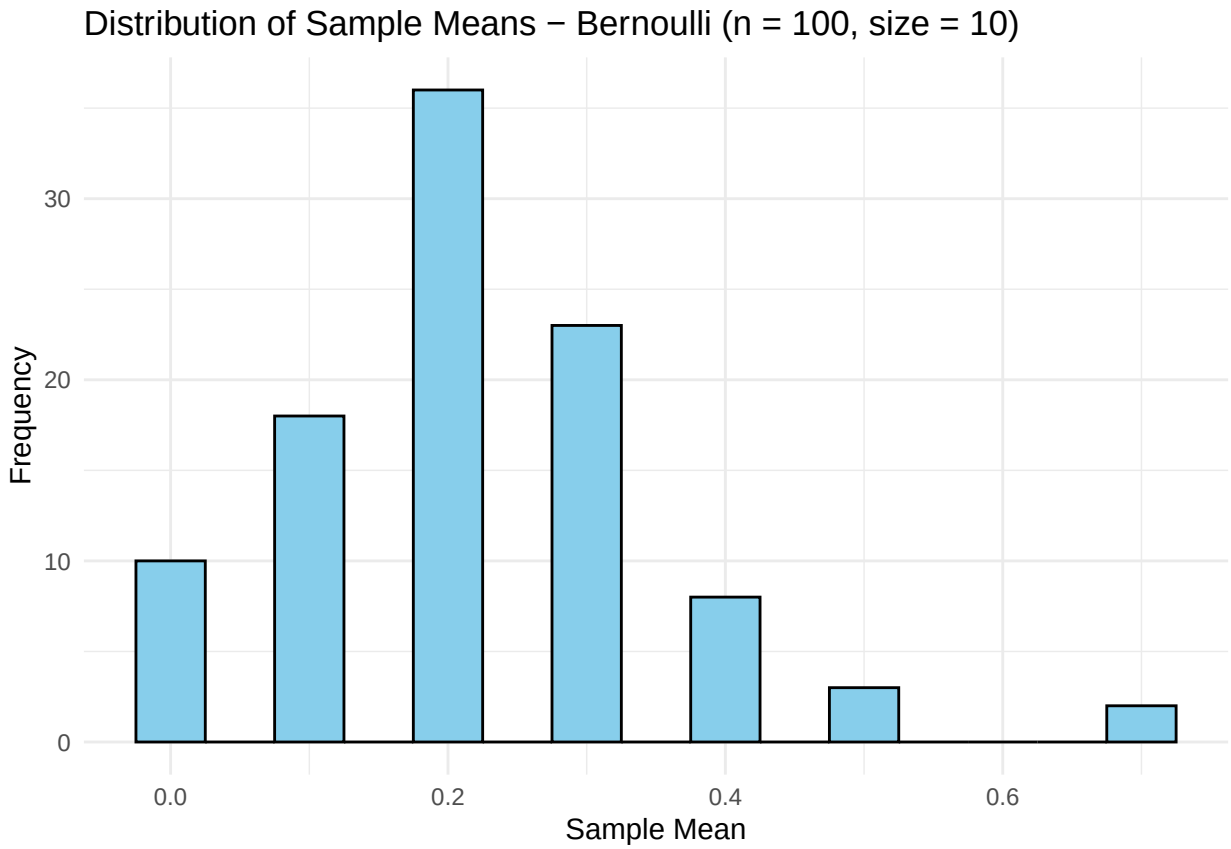

Distribution of Sample Means − Uniform (n = 10, size = 100)
Frequency
0
2
4
6
0.45
0.50
0.55
Sample Mean

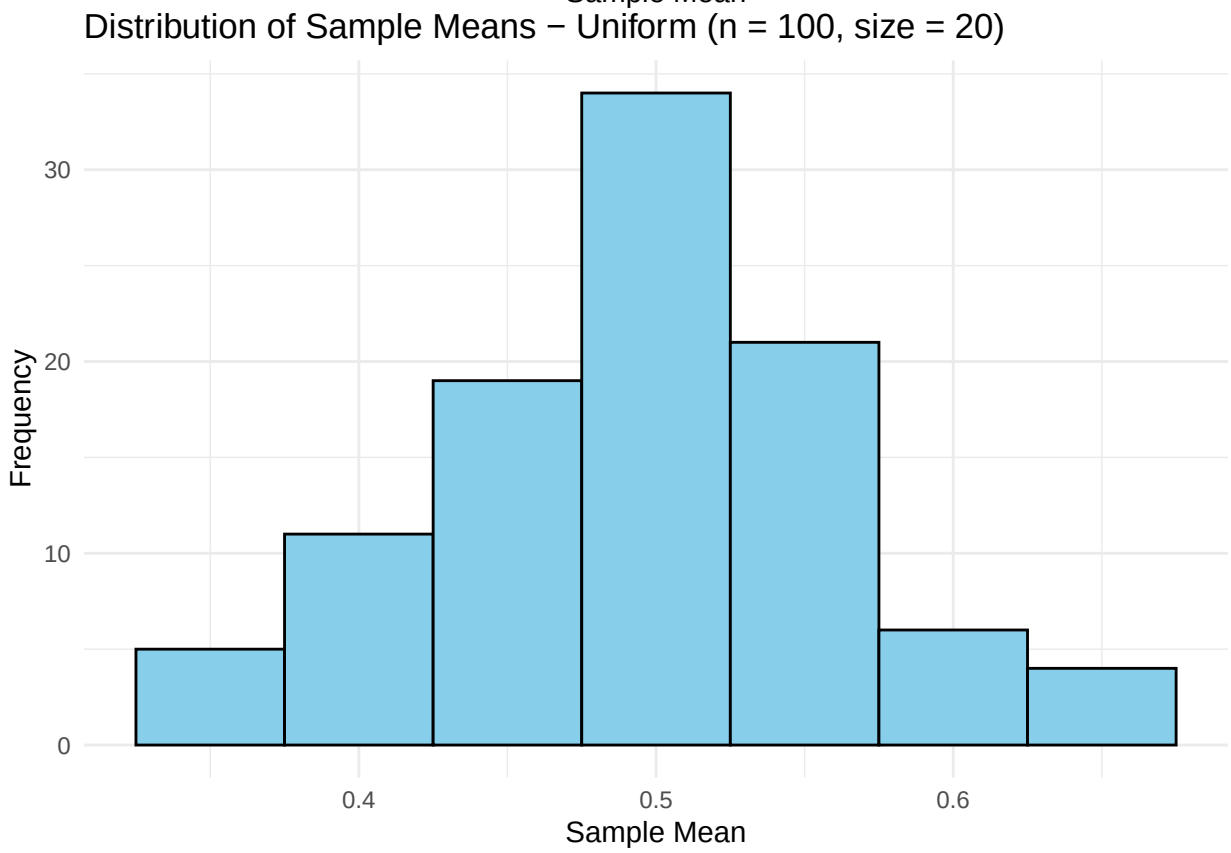

Distribution of Sample Means − Poisson (n = 10, size = 100)
Frequency
2.0
1.5
1.0
0.5
0.0
4.6
4.7
4.8
4.9
5.0
5.1
5.2
Sample Mean

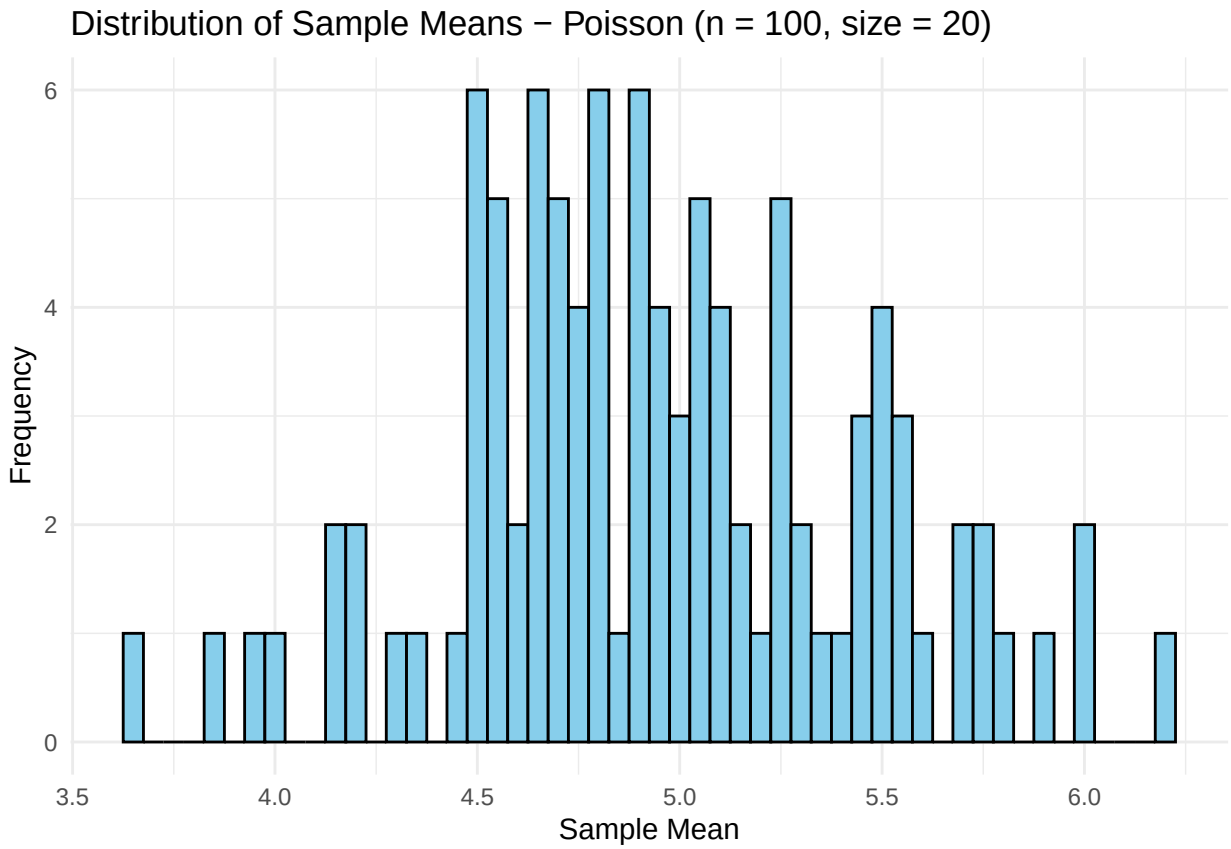

Distribution of Sample Means − Bernoulli (n = 10, size = 100)
Frequency
6
4
2
0
0.12
0.16
0.20
0.24
0.28
Sample Mean

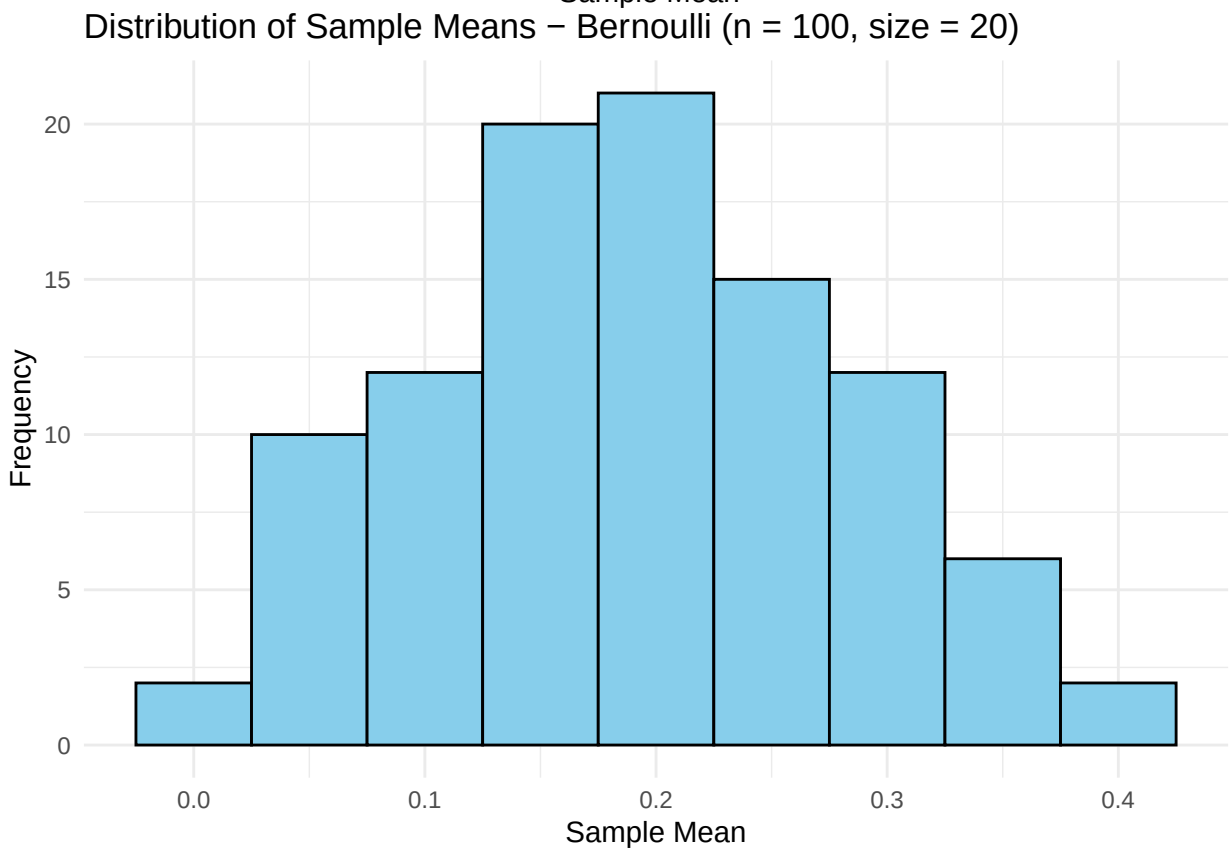

Distribution of Sample Means − Uniform (n = 100, size = 10)
Frequency
0
5
10
15
0.3
0.4
0.5
0.6
Sample Mean

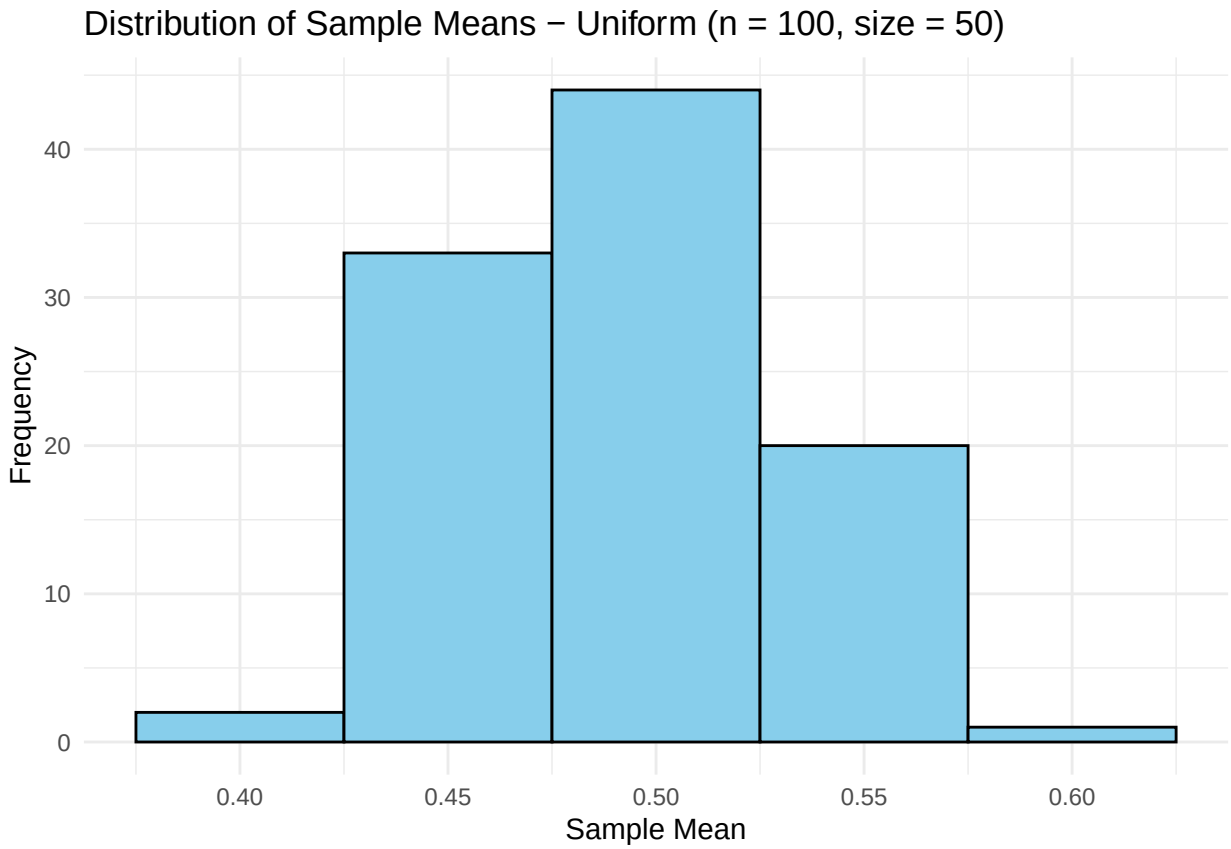

Distribution of Sample Means − Poisson (n = 100, size = 10)
Frequency
0.0
2.5
5.0
7.5
10.0
4
5
6
Sample Mean

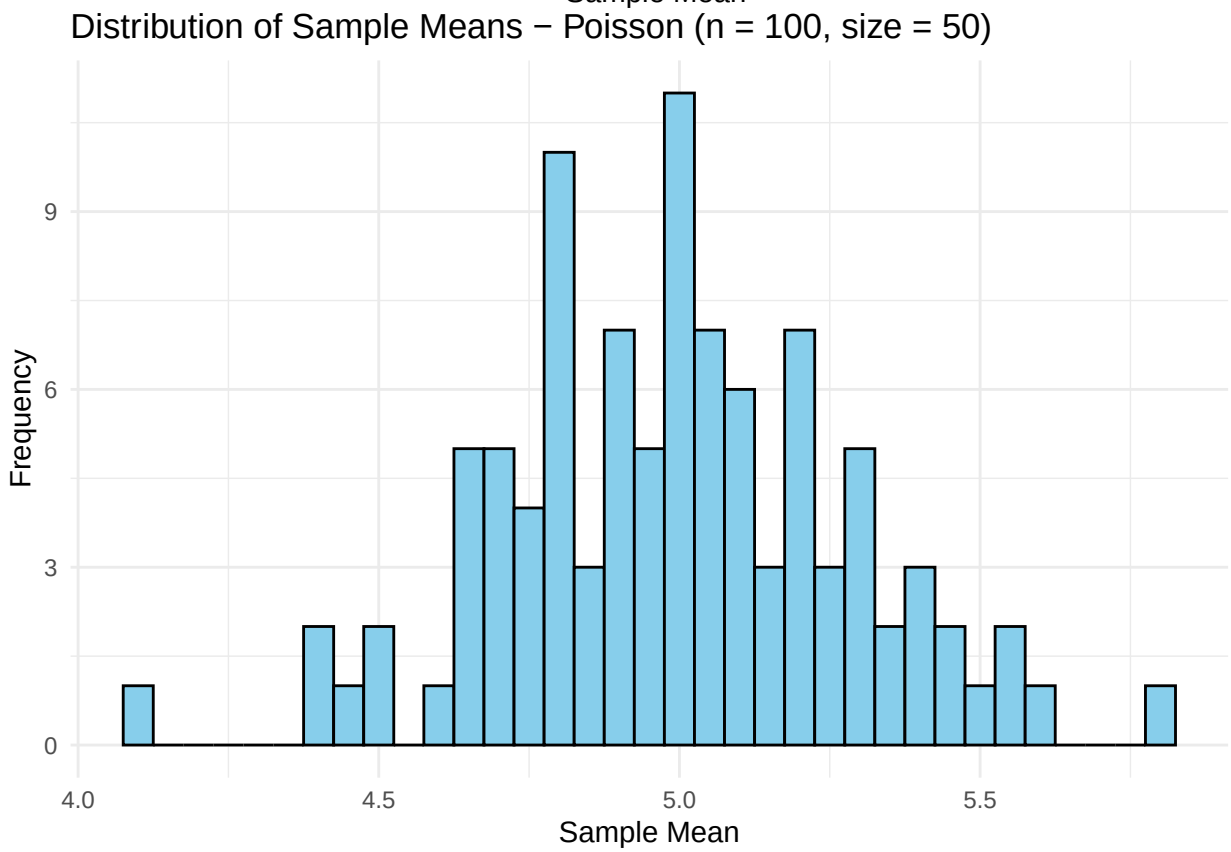

Distribution of Sample Means − Bernoulli (n = 100, size = 10)
Frequency
30
20
10
0
0.0
0.2
0.4
0.6
Sample Mean

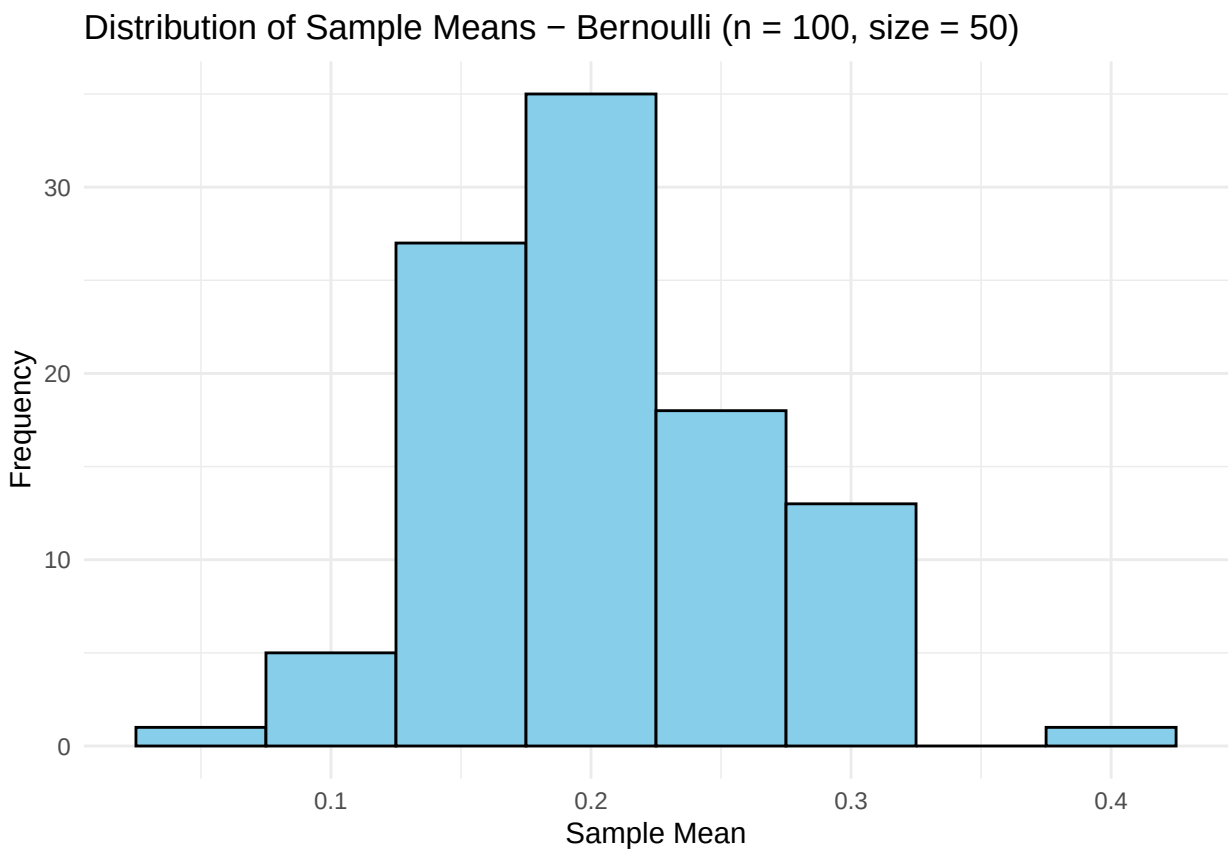

Distribution of Sample Means − Uniform (n = 100, size = 20)
Frequency
30
20
10
0
0.4
0.5
0.6
Sample Mean

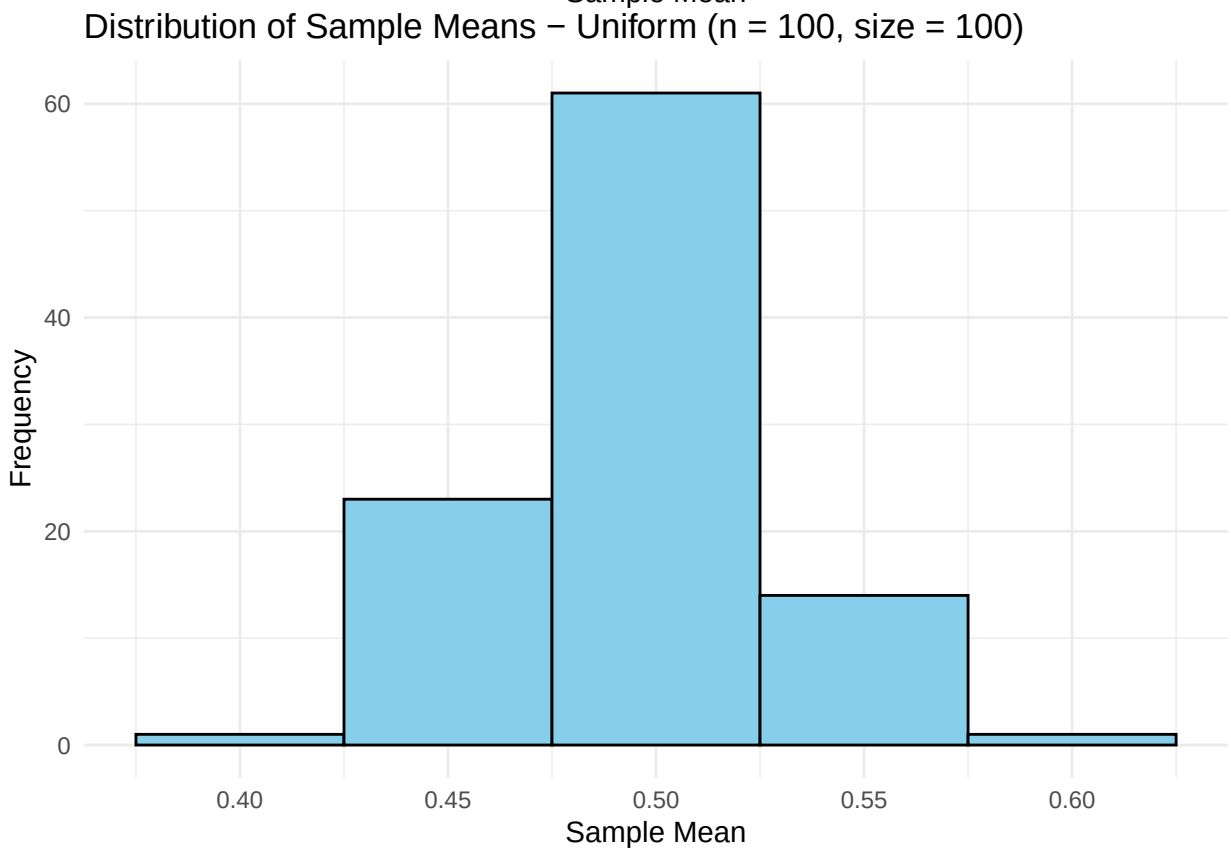

Distribution of Sample Means − Poisson (n = 100, size = 20)
Frequency
0
2
4
6
3.5
4.0
4.5
5.0
5.5
6.0
Sample Mean

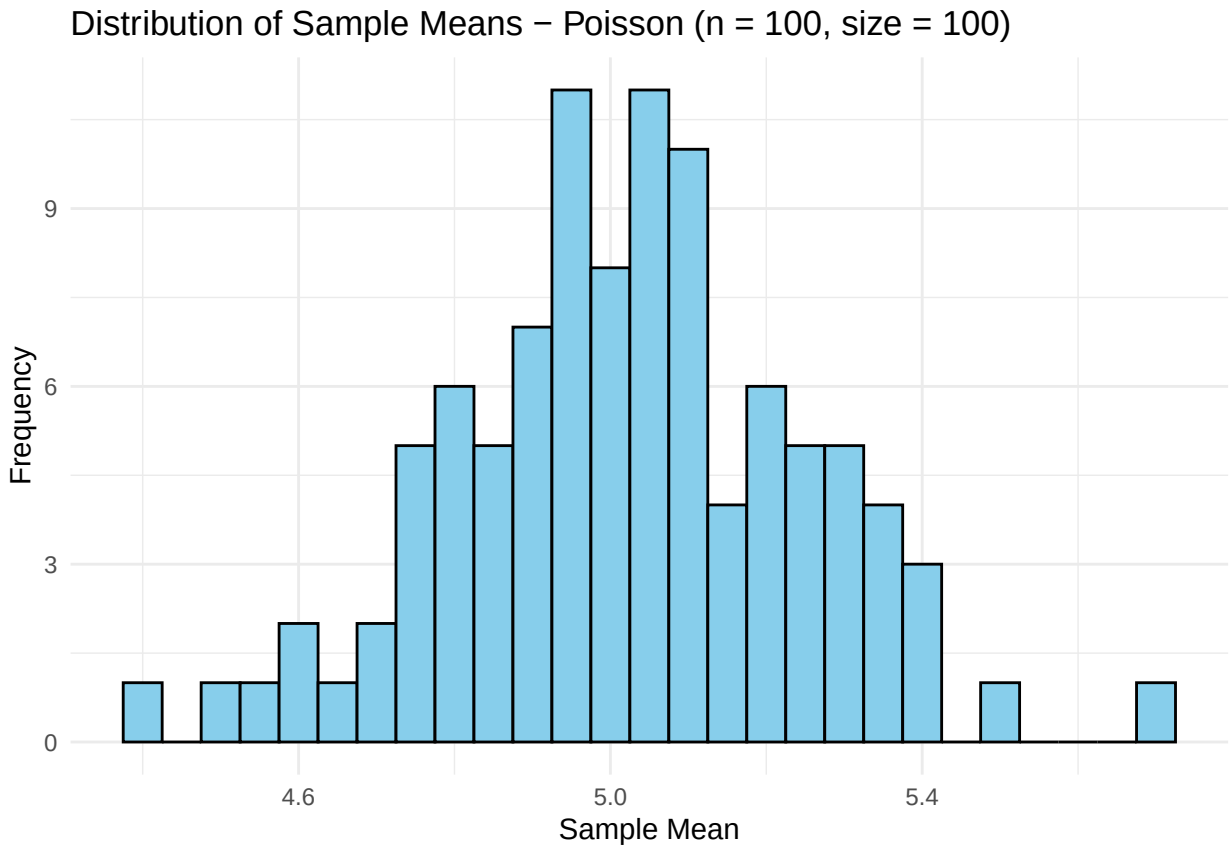

Distribution of Sample Means − Bernoulli (n = 100, size = 20)
Frequency
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
Sample Mean

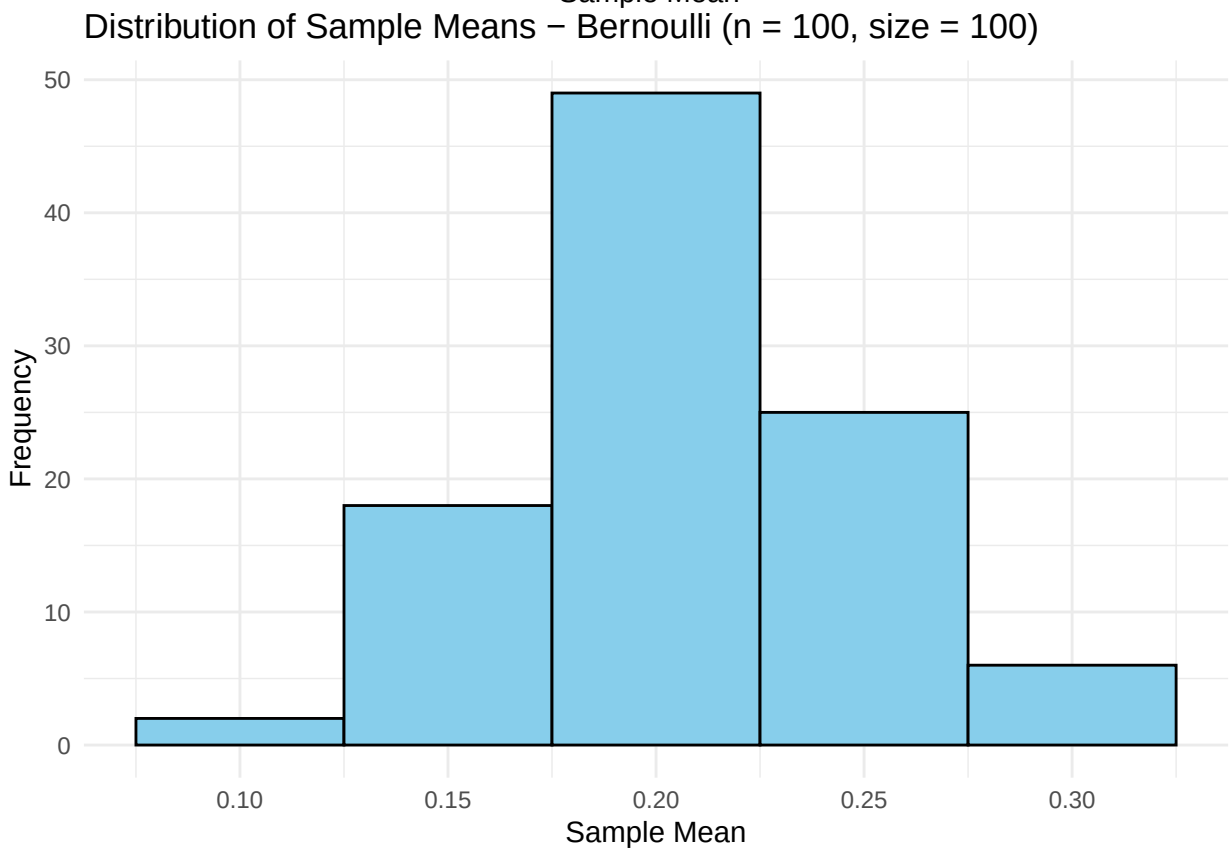

Distribution of Sample Means − Uniform (n = 100, size = 50)
Frequency
0
10
20
30
40
0.40
0.45
0.50
0.55
0.60
Sample Mean

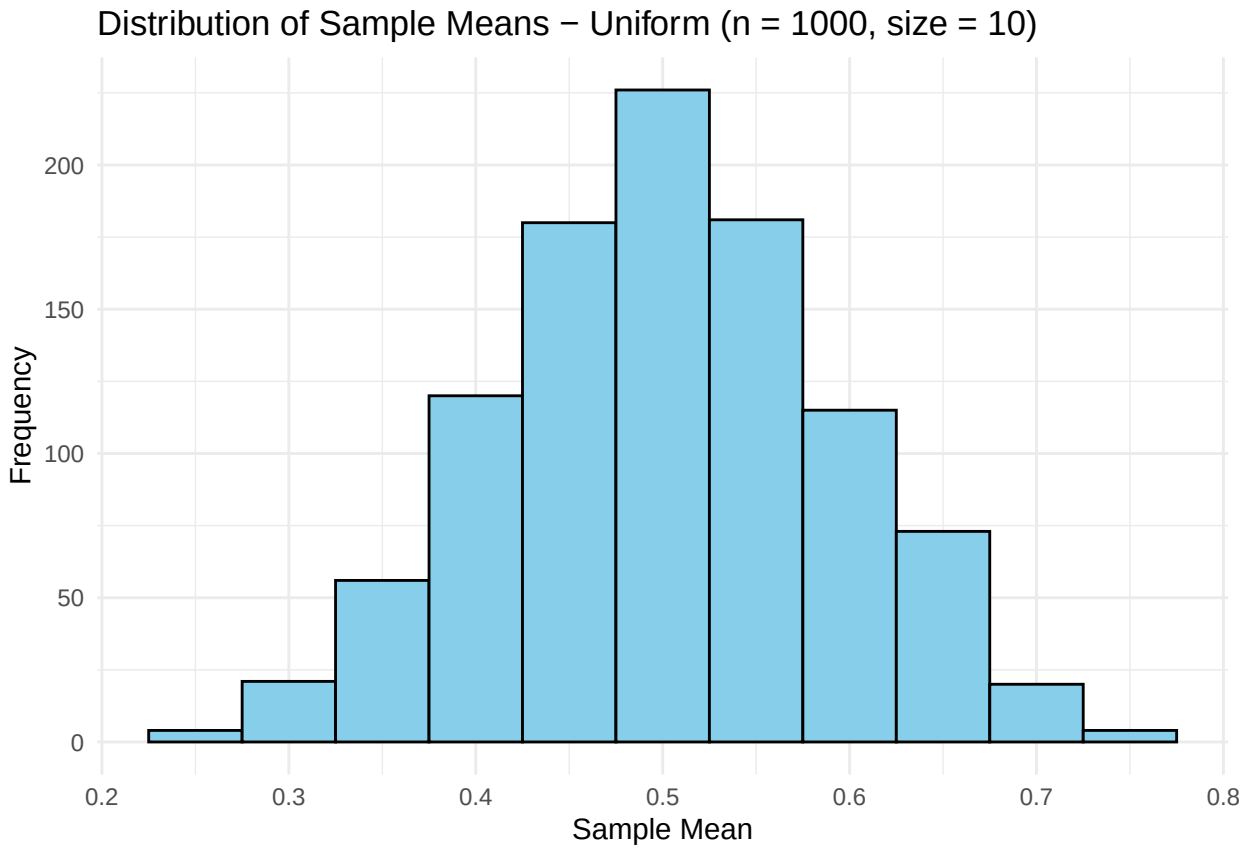

Distribution of Sample Means − Poisson (n = 100, size = 50)
Frequency
0
3
6
9
4.0
4.5
5.0
5.5
Sample Mean

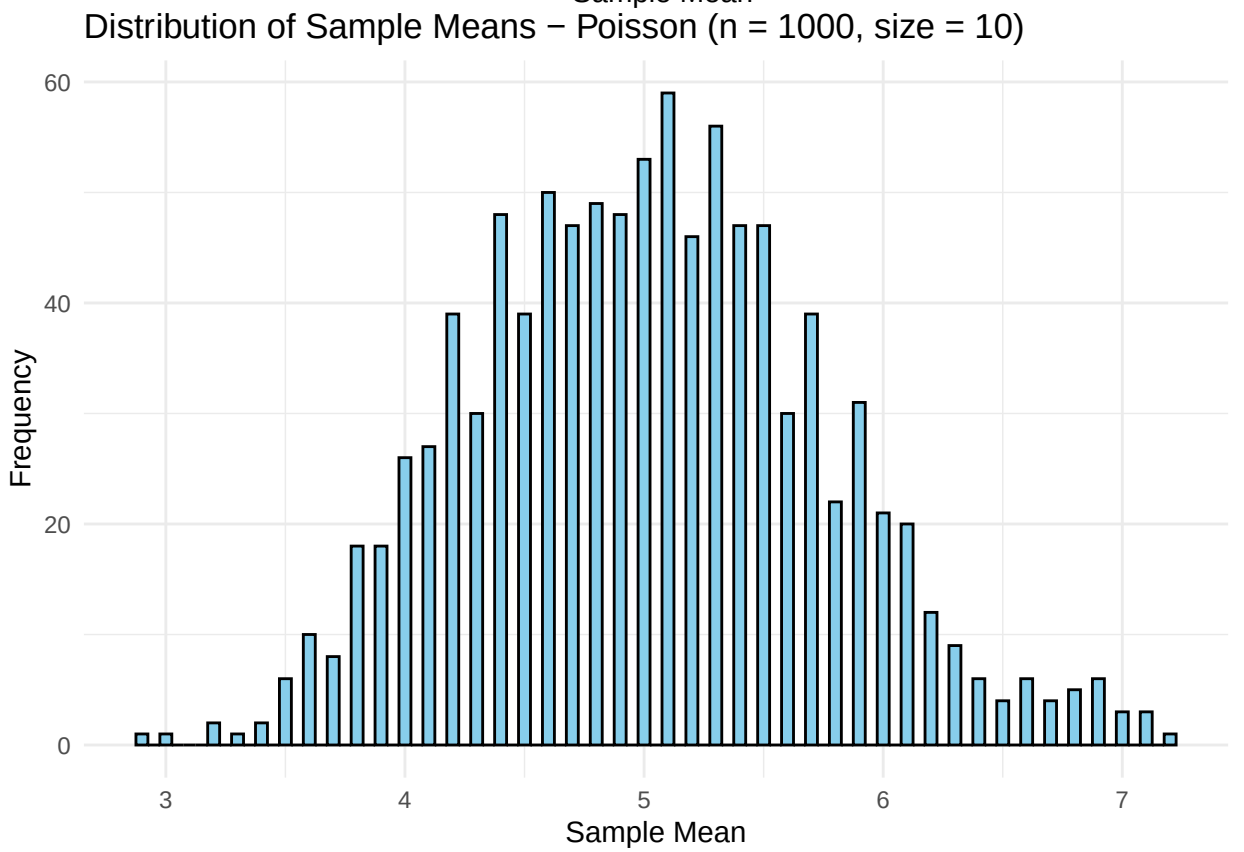

Distribution of Sample Means − Bernoulli (n = 100, size = 50)
Frequency
0
10
20
30
0.1
0.2
0.3
0.4
Sample Mean

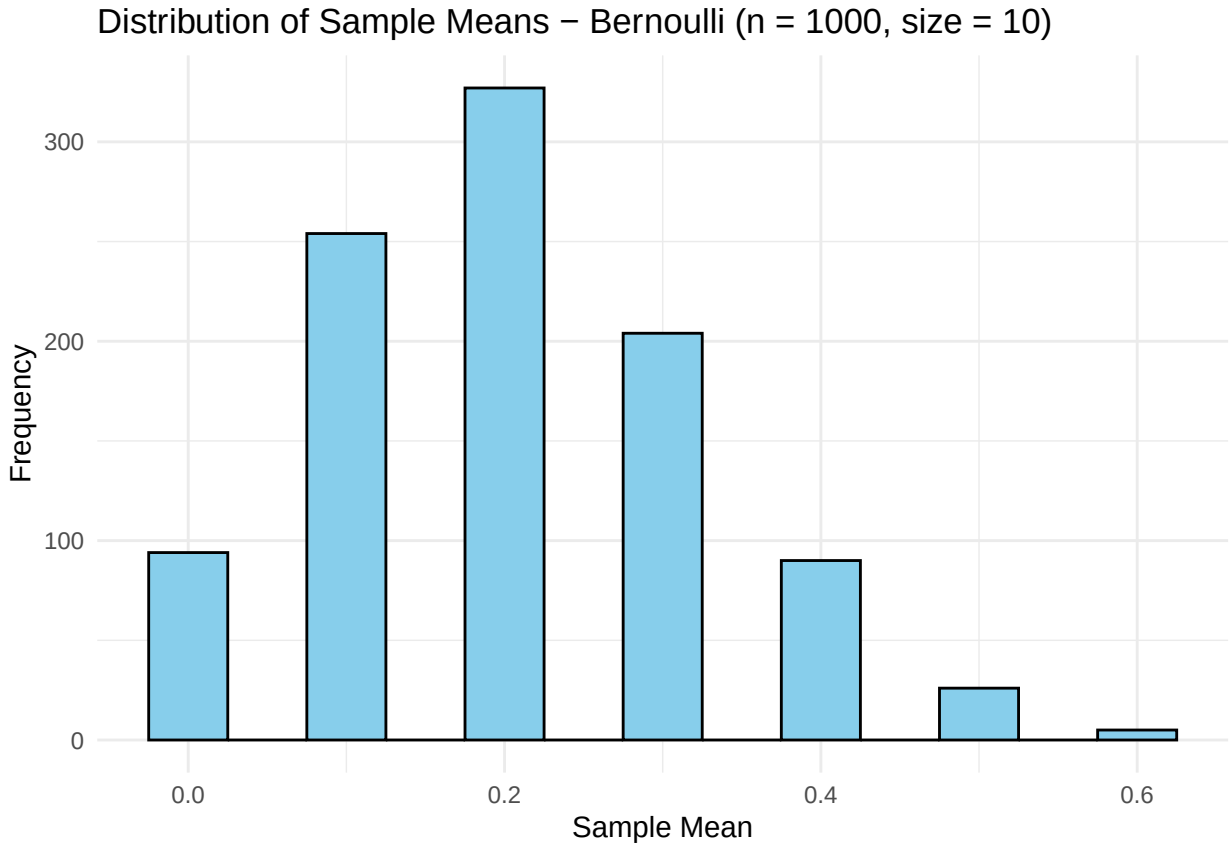

Distribution of Sample Means − Uniform (n = 100, size = 100)
Frequency
0
20
40
60
0.40
0.45
0.50
0.55
0.60
Sample Mean

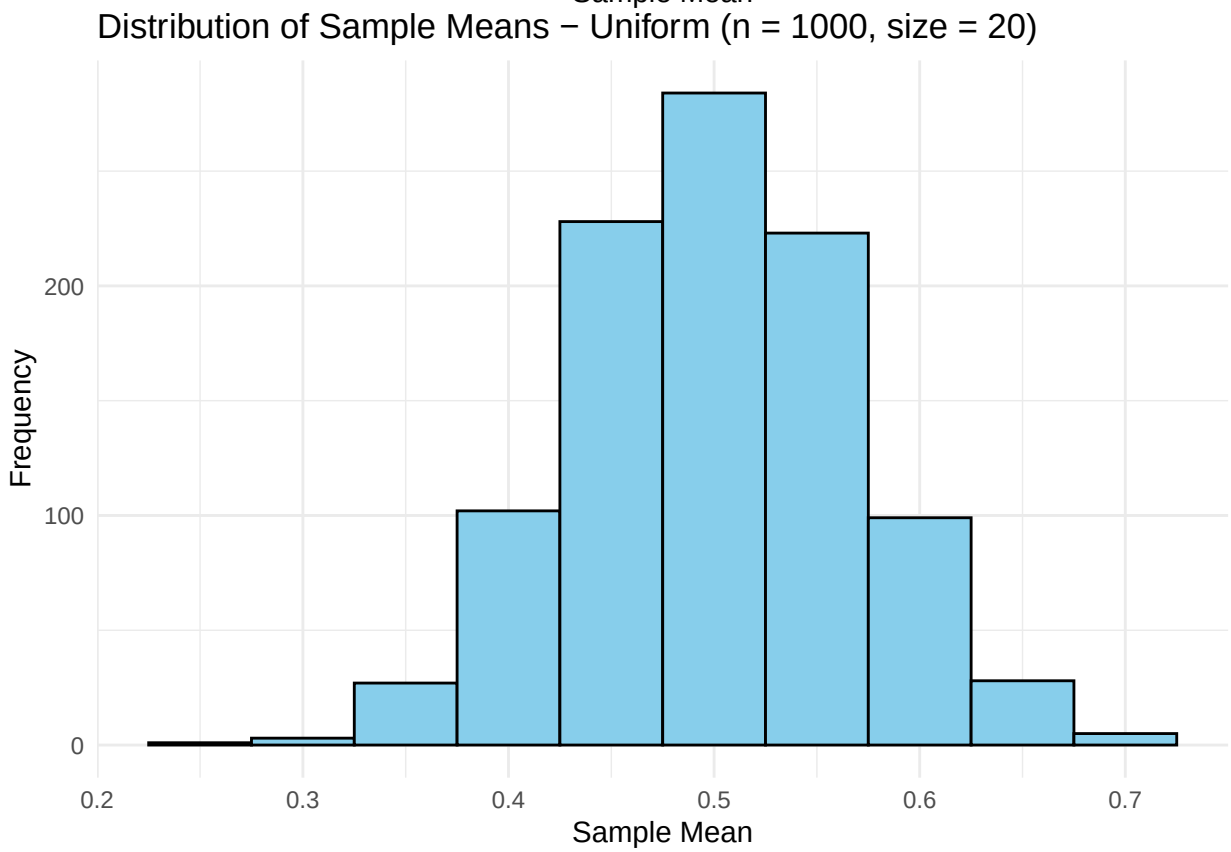

Distribution of Sample Means − Poisson (n = 100, size = 100)
Frequency
0
3
6
9
4.6
5.0
5.4
Sample Mean

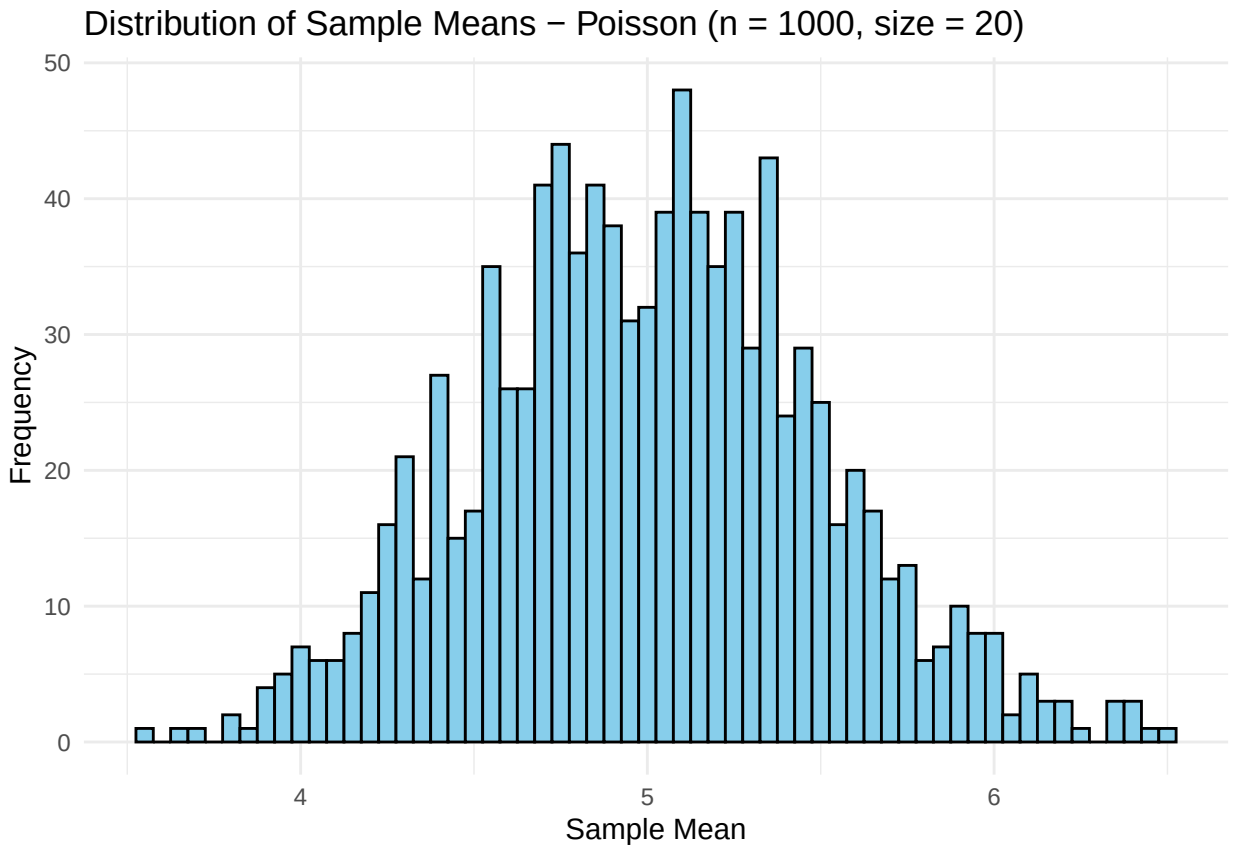

Distribution of Sample Means − Bernoulli (n = 100, size = 100)
Frequency
0
10
20
30
40
50
0.10
0.15
0.20
0.25
0.30
Sample Mean

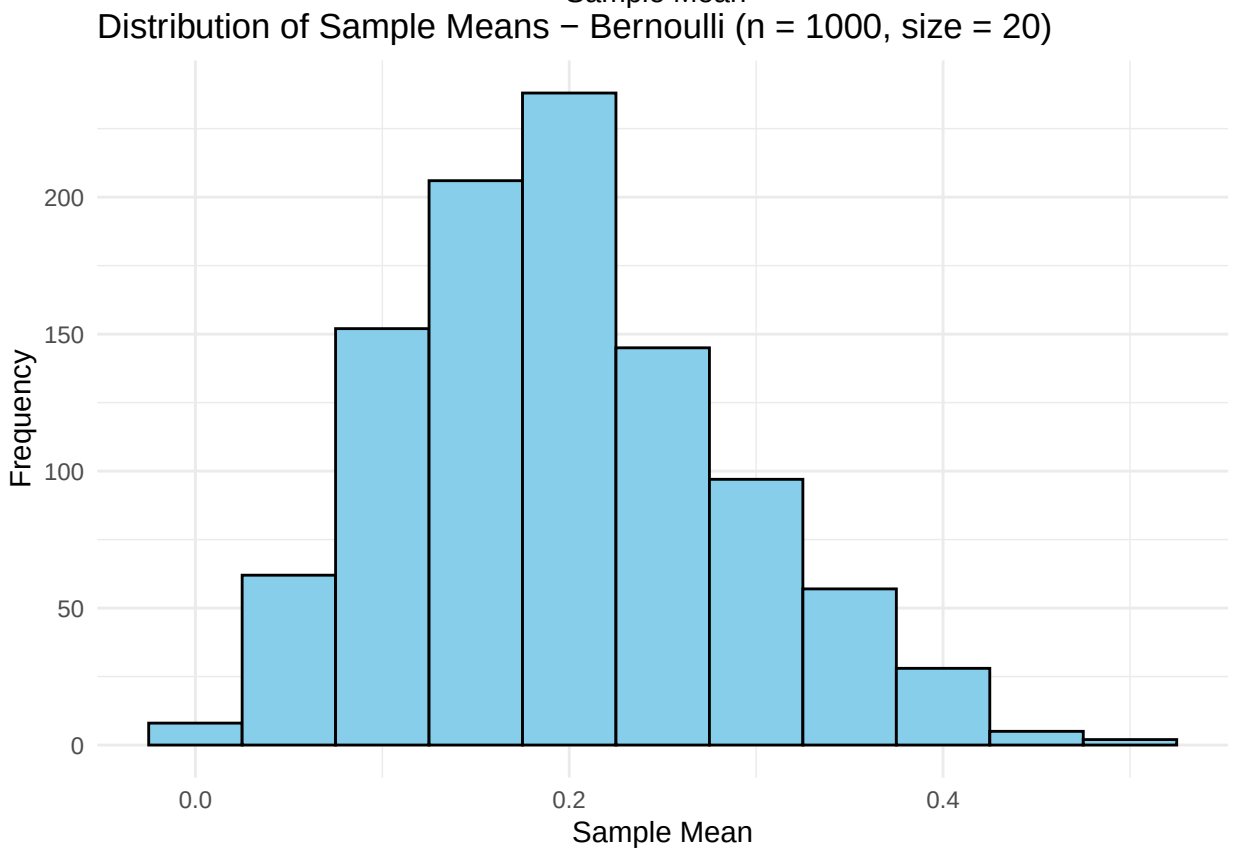

Distribution of Sample Means − Uniform (n = 1000, size = 10)
Frequency
0
50
100
150
200
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Sample Mean

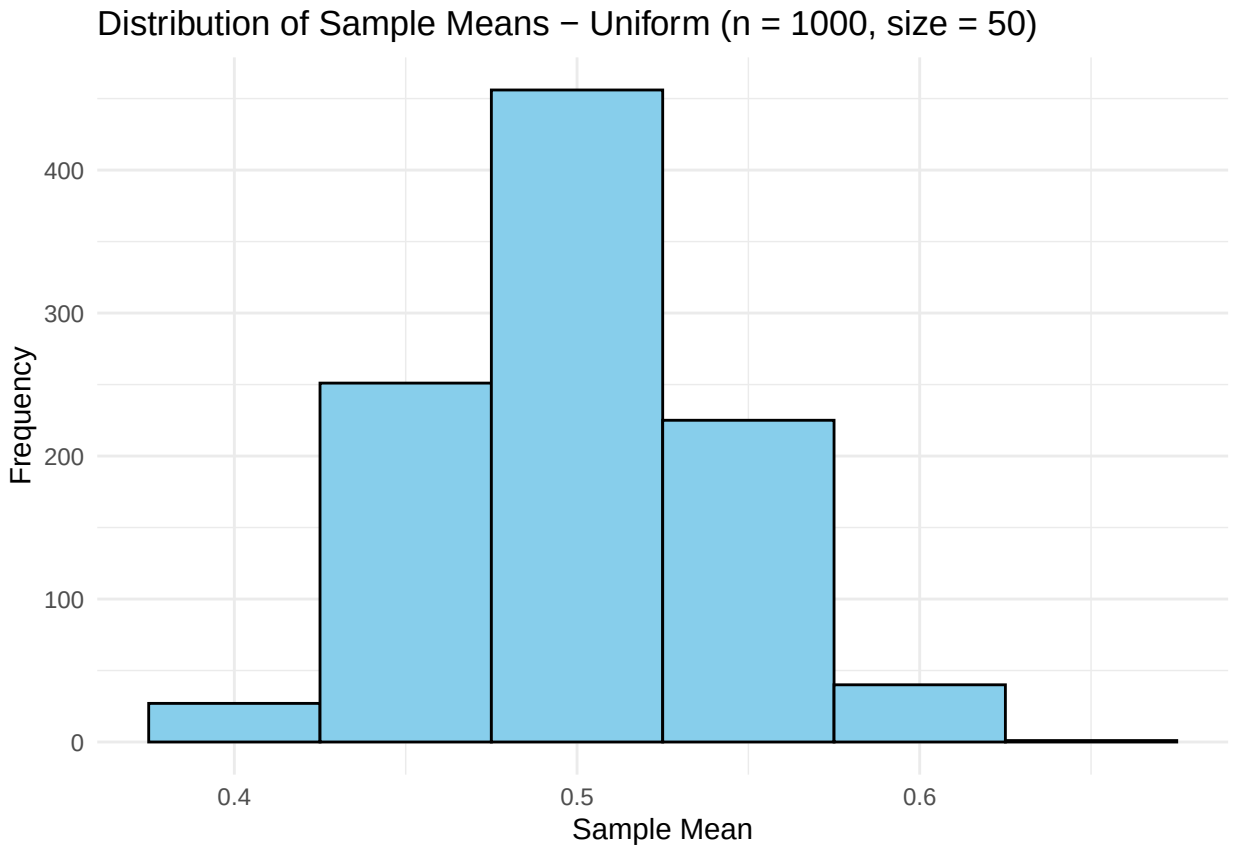

Distribution of Sample Means − Poisson (n = 1000, size = 10)
Frequency
0
20
40
60
3
4
5
6
7
Sample Mean

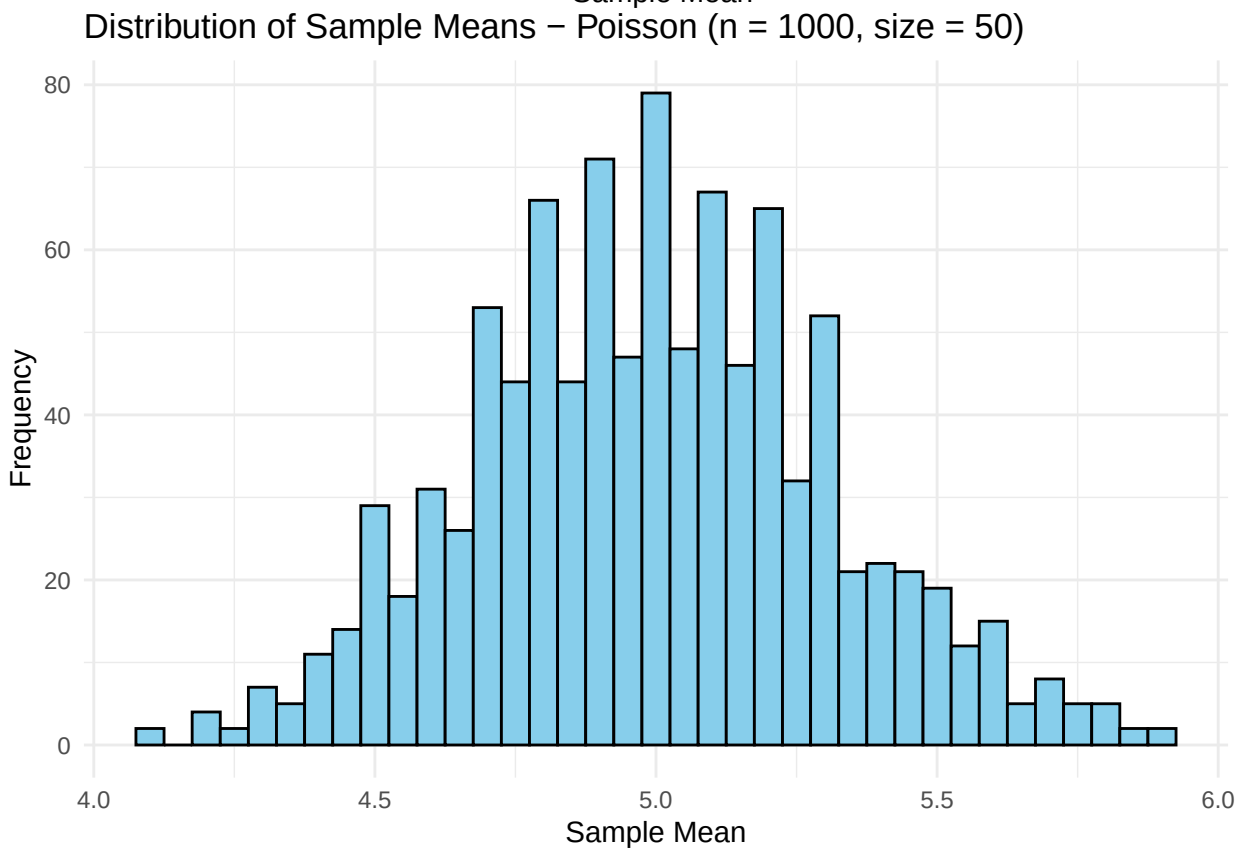

Distribution of Sample Means − Bernoulli (n = 1000, size = 10)
Frequency
300
200
100
0
0.0
0.2
0.4
0.6
Sample Mean

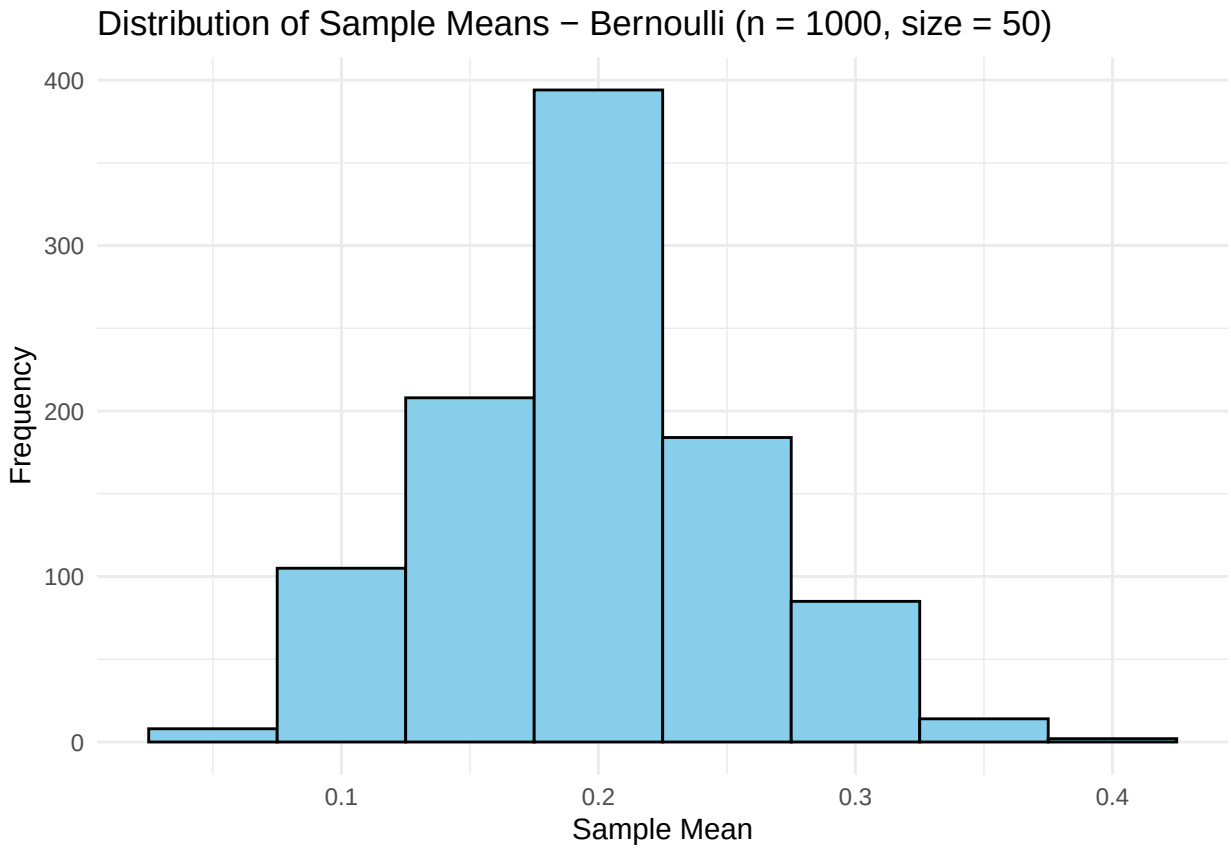

Distribution of Sample Means − Uniform (n = 1000, size = 20)
Frequency
200
100
0
0.2
0.3
0.4
0.5
0.6
0.7
Sample Mean

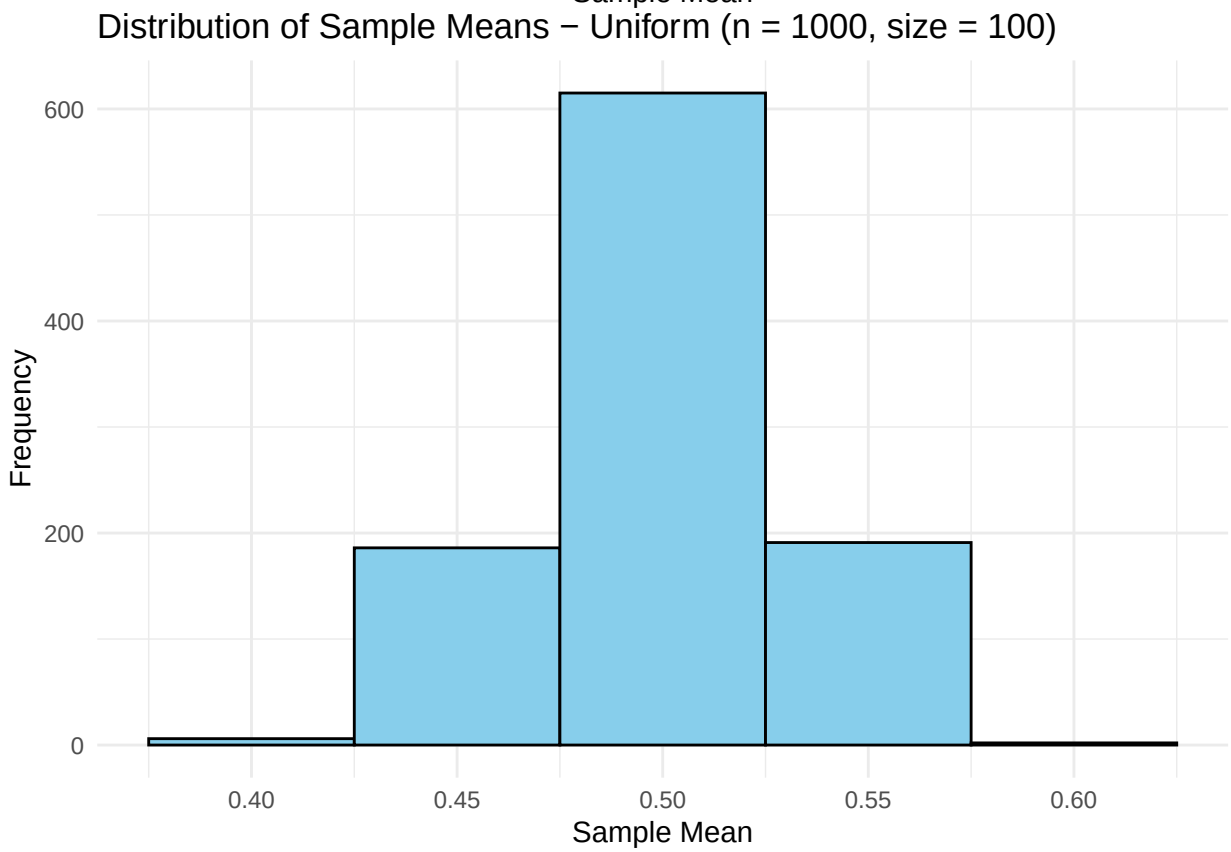

Distribution of Sample Means − Poisson (n = 1000, size = 20)
Frequency
Sample Mean
0
10
20
30
40
50
4
5
6

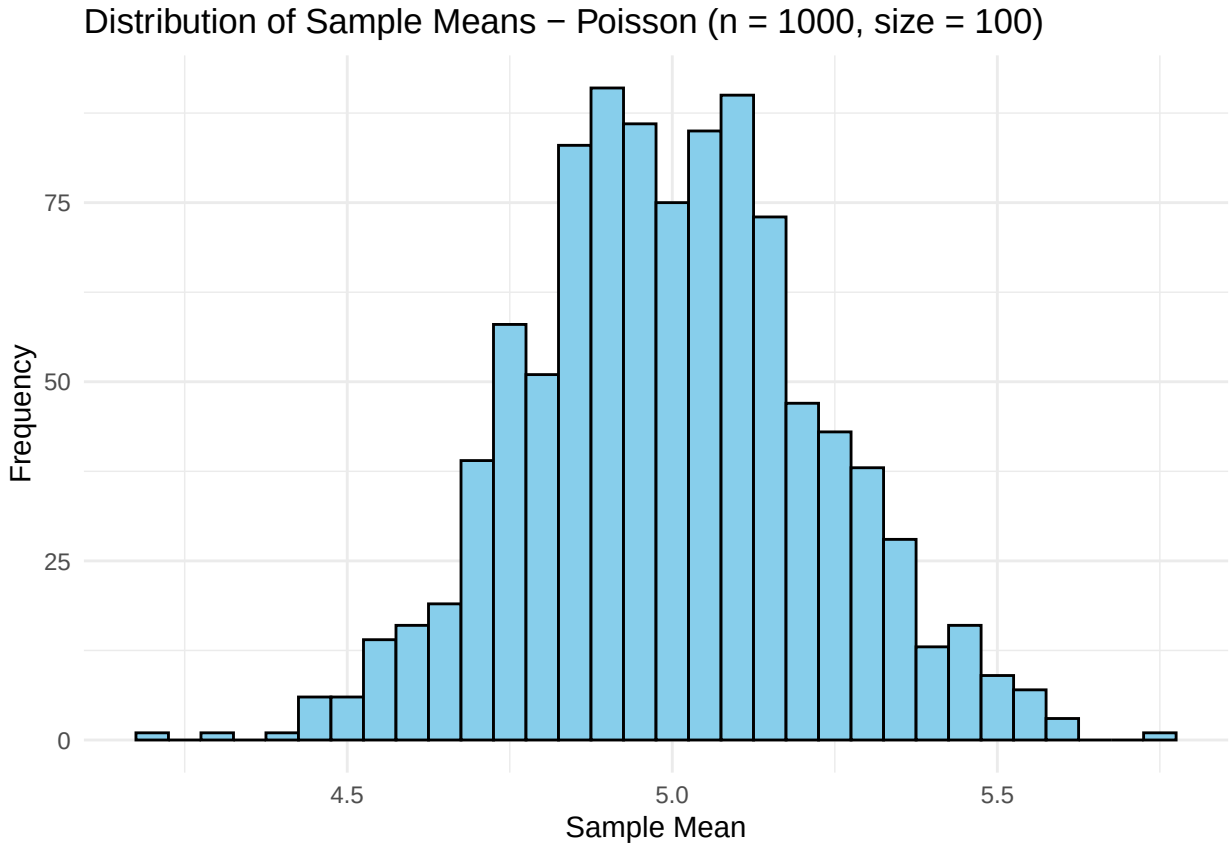

Distribution of Sample Means − Bernoulli (n = 1000, size = 20)
Frequency
Sample Mean
0
50
100
150
200
0.0
0.2
0.4

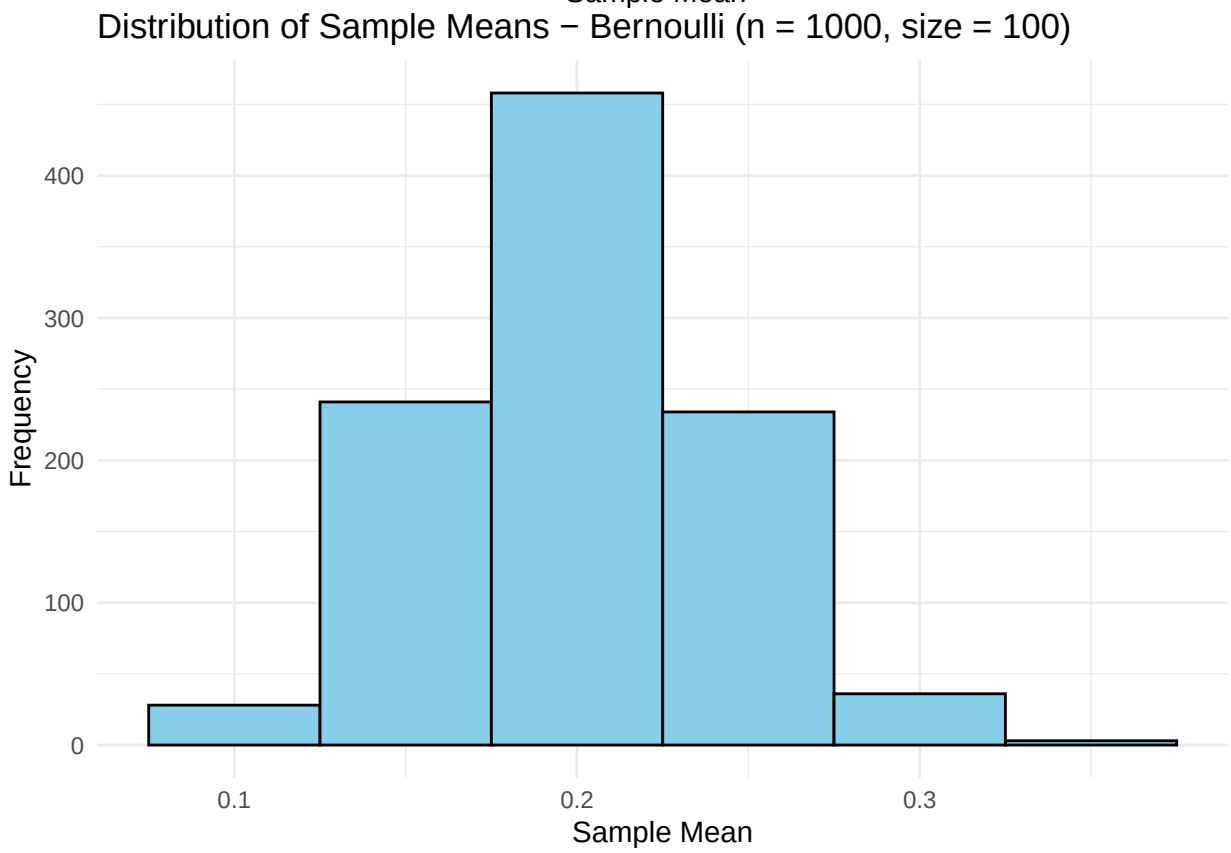

Distribution of Sample Means − Uniform (n = 1000, size = 50)
Frequency
0
100
200
300
400
0.4
0.5
0.6
Sample Mean


Distribution of Sample Means − Poisson (n = 1000, size = 50)
Frequency
0
20
40
60
80
4.0
4.5
5.0
5.5
6.0
Sample Mean


Distribution of Sample Means − Bernoulli (n = 1000, size = 50)
Frequency
400
300
200
100
0
0.1
0.2
0.3
0.4
Sample Mean


Distribution of Sample Means − Uniform (n = 1000, size = 100)
Frequency
600
400
200
0
0.40
0.45
0.50
0.55
0.60
Sample Mean




It can be observed in the above histograms that as the number of samples and the sample size increases

that the distributions approach a normal distribution.

b) As sample size increases we see a more consistent spread of data with less random dips. This does contribute towards the data becoming more normal but not as significantly as the number of samples does.

c) AS the the number of samples increases we see that it is the biggest contributor towards the distributions becoming more normal.

d) The Bernoulli and poisson distributions have a very clear normal distribution as the number of samples increases and the Uniform distribution also appears normal though it stands out to haev quite a steep peek which may be even too steep to be considered normal.

## Question 2:

```
salt_data <- read.table(file ='C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A2/data1/salt.txt', sep="",header=TRUE, na.strings="NA")

attach(salt_data)
```

a)

```
library(sn)
```

```
## Warning: package 'sn' was built under R version 4.3.3
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'sn'
```

```
## The following object is masked from 'package:stats':
##
##     sd
```

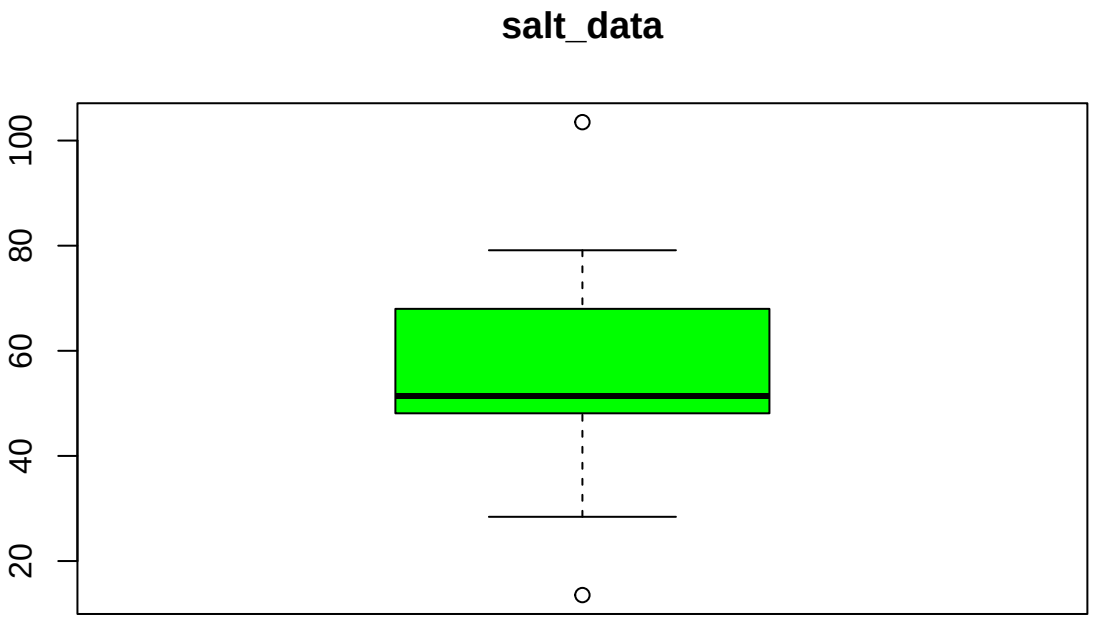
```
summary(salt_data)
```

```
##       salt
##  Min.   : 13.53
##  1st Qu.: 48.11
##  Median : 51.40
##  Mean   : 55.62
##  3rd Qu.: 67.98
##  Max.   :103.50
```

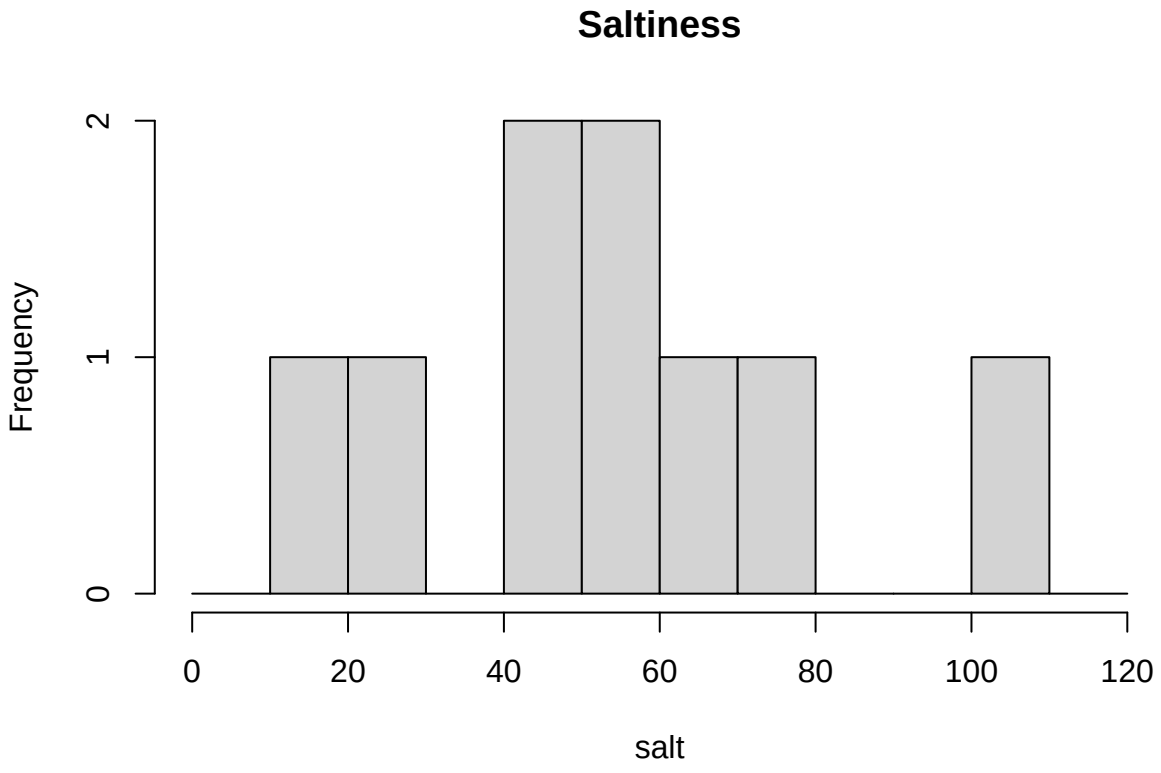
```
class(salt_data)
```

```
## [1] "data.frame"
```

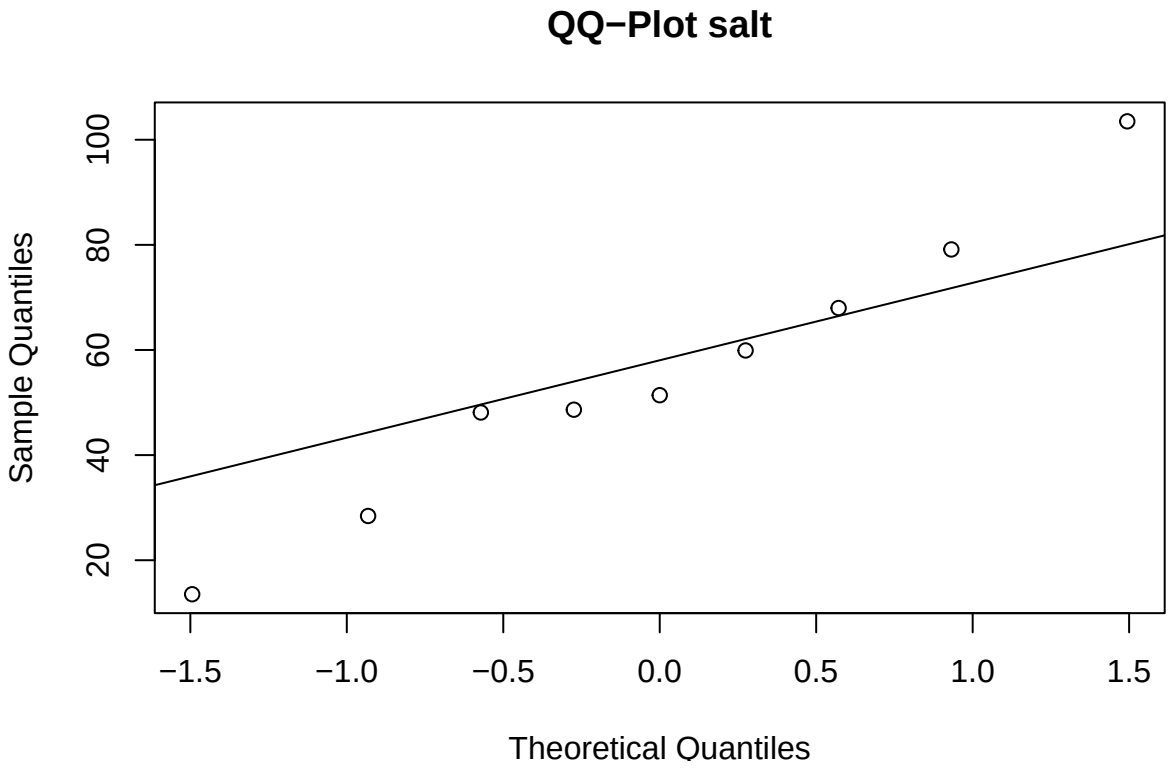
```r
boxplot(salt_data, col = "green")
title("salt_data")
```




```r
hist(salt, breaks=seq(0,120,by=10), main="Saltiness", xlab="salt")
```




```r
qqnorm(salt, main="QQ-Plot salt")
qqline(salt)
```




Based on what can be observed from the data visualization it appears that we could make some different interpretations. Based om the boxplot we would say it has a positive skew(right skew). Where from the histogram and QQ-plot we can see that it appears to be mostly symmetrical though it has a steep peak.

b)

```
#skew function
skew<-function(x){
m3<-sum((x-mean(x))^3)/length(x)
s3<-sqrt(var(x))^3
m3/s3  }

# getting skew_hat
skew_hat<-skew(salt)
skew_hat
```

```
## [1] 0.1723753
```

```
#Kurtosis function
kurtosis<-function(x) {
m4<-sum((x-mean(x))^4)/length(x)
s4<-var(x)^2
m4/s4 - 3  }

kurt_hat<-kurtosis(salt)
kurt_hat
```

```
## [1] -0.9342198
```

Estimation of the skew is 0.1724. Estimation of the kurtosis is -0.9342.

c)

```
salt ## data for bootstrapping
```

```
## [1]  13.53  28.42  48.11  48.64  51.40  59.91  67.98  79.13 103.50
```

```
#Bootstrap
B<-15000
salt_boot<-matrix(data=sample(salt,size=B*length(salt),replace=TRUE),nrow=length(salt),ncol=B)
skew_boot_sampled<-apply(salt_boot,2,skew)
boot_interval_skew<-quantile(skew_boot_sampled,probs=c(0.025,0.975))

boot_interval_skew
```

```
##      2.5%     97.5%
## -0.891026  1.109291
```

The 95% confidence interval for skewness includes 0. Which suggests that There is little or no skewness in the population data.

d)

```
salt ## data for bootstrapping
```

```
## [1]  13.53  28.42  48.11  48.64  51.40  59.91  67.98  79.13 103.50
```

```
#Bootstrap
B<-15000
salt_boot<-matrix(data=sample(salt,size=B*length(salt),replace=TRUE),nrow=length(salt),ncol=B)
kurt_boot_sampled<-apply(salt_boot,2,kurtosis)
boot_interval_kurt<-quantile(kurt_boot_sampled,probs=c(0.025,0.975))

boot_interval_kurt
```

```
##       2.5%      97.5%
## -1.8845268  0.5661748
```

The 95% confidence interval for kurtosis includes 0. Which suggests that there may be zero kurtosis.

e) For the salt data based on our 95% confidence intervals for skewness and kurtosis we can conclude that there is a zero or close to zero value for both skewness and kurtosis.

## Question 3:

```
fecundity_data <- read.table(file ='C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A2/data1/fecundity.txt', sep="",header=TRUE, na.strings="NA")

attach(fecundity_data)
```

Test if the variances are equal.

```
rs<-fecundity_data$RS[!is.na(fecundity_data$RS)]
ns<-fecundity_data$NS[!is.na(fecundity_data$NS)]
var.test(rs,ns)
```

```
##
##  F test to compare two variances
##
## data:  rs and ns
## F = 0.75551, num df = 24, denom df = 24, p-value = 0.4974
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3329286 1.7144557
## sample estimates:
## ratio of variances
##          0.7555074
```

P-value = 0.4974 There is little or no evidence to indicate that the variances are not equal. Based on the variance test. Conclude that the RS and NS do not significantly differ in population variance.

Therefore we can move forward with the assumption that the population variances are equal and perform a t-test to determine if the population means are equal. With H0: The popualtion means are equal.

```
t.test(rs,ns,alternative="two.sided",mu=0,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  rs and ns
## t = -3.4251, df = 48, p-value = 0.001268
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.880308  -3.351692
## sample estimates:
## mean of x mean of y
##    25.256    33.372
```

The p-value for the t-test is 0.001268. Therefore we have very strong evidence against the H0 and reject it. Conclude that RS and NS do differ in their population means.

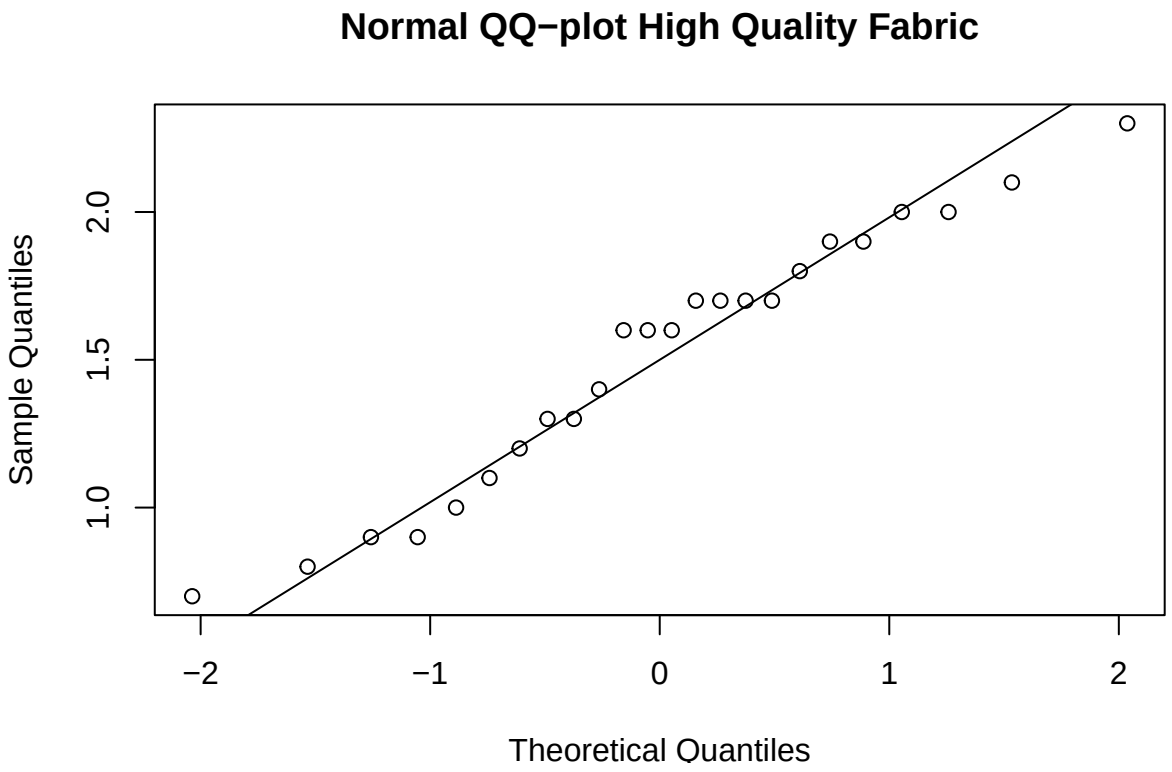
## Question 4

```
fabric_data <- read.table(file ='C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A2/data1/fabric.txt', sep="",header=TRUE, fill=TRUE) ##adjusted
↪  the content in the fabric.txt file so the read command would work.

attach(fabric_data)
fabric_data
```

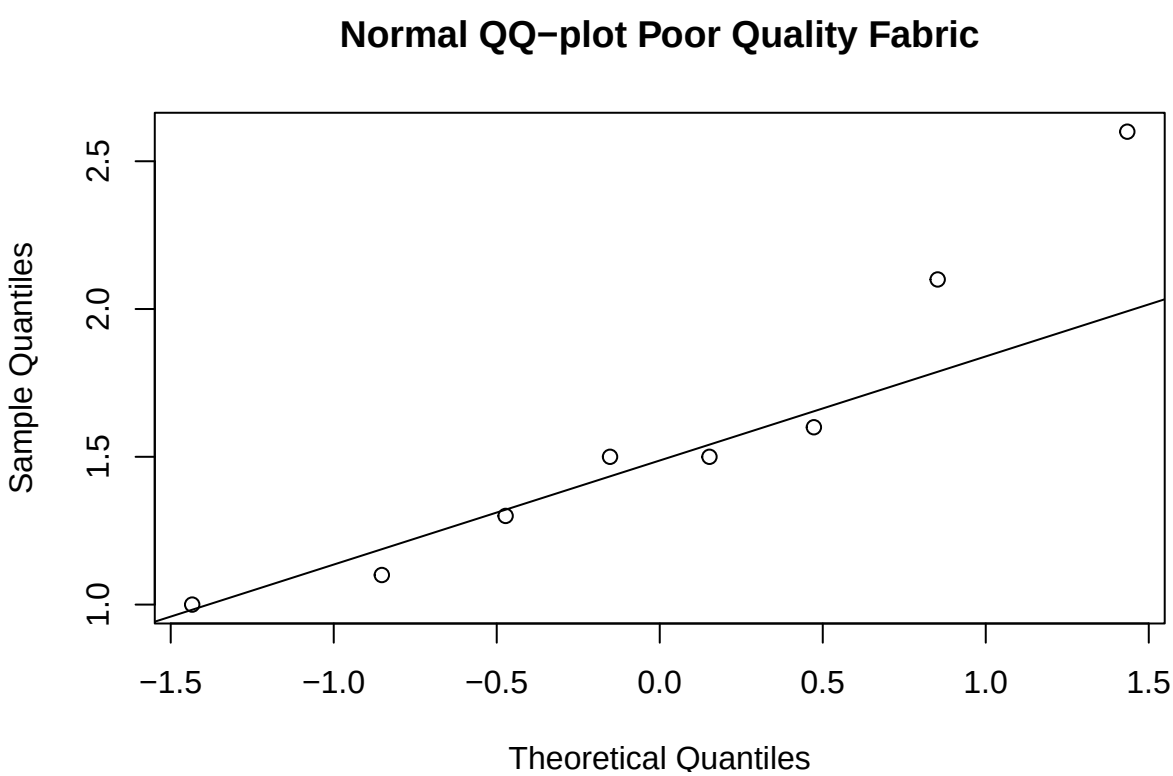
```
##      H   P
## 1  1.2 1.6
## 2  0.9 1.5
## 3  0.7 1.1
## 4  1.0 2.1
## 5  1.7 1.5
## 6  1.7 1.3
## 7  1.1 1.0
## 8  0.9 2.6
## 9  1.7  NA
## 10 1.9  NA
## 11 1.3  NA
## 12 2.1  NA
## 13 1.6  NA
## 14 1.8  NA
## 15 1.4  NA
## 16 1.3  NA
## 17 1.9  NA
## 18 1.6  NA
## 19 0.8  NA
## 20 2.0  NA
## 21 1.7  NA
## 22 1.6  NA
## 23 2.3  NA
## 24 2.0  NA
```

a) QQ-plot construction.

```
qqnorm(fabric_data$H, main="Normal QQ-plot High Quality Fabric")
qqline(fabric_data$H)
```



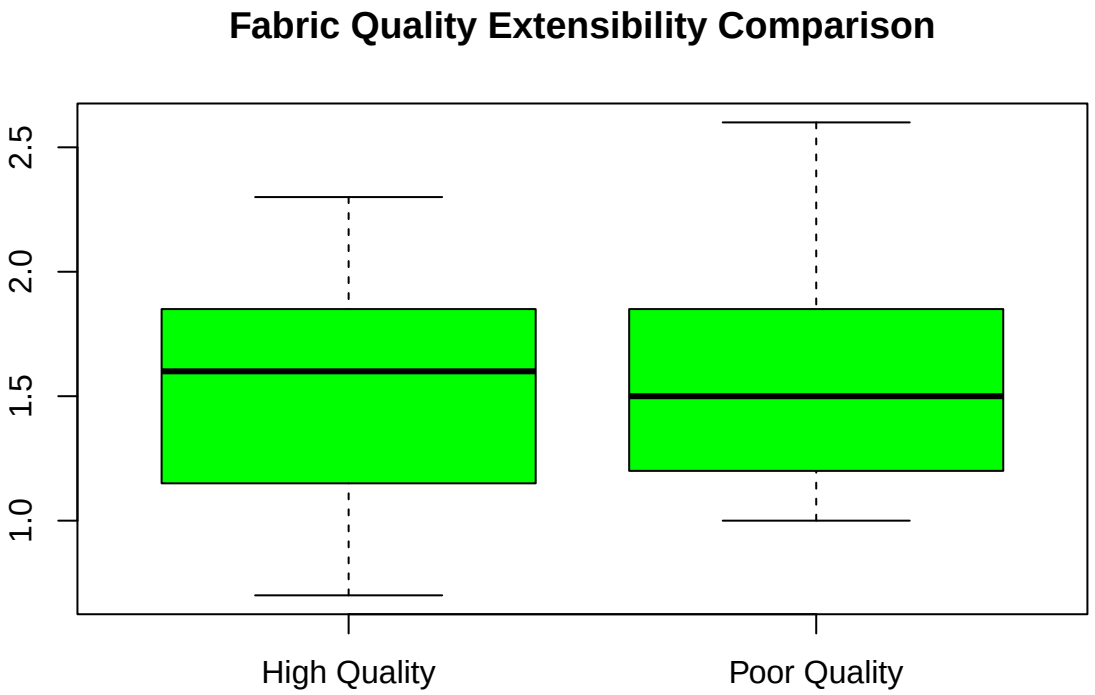

```r
qqnorm(fabric_data$P, main="Normal QQ-plot Poor Quality Fabric")
qqline(fabric_data$P)
```




H and P both look to be approximately normal judging by the QQ-plots.

b)

```r
# boxplot construction
boxplot(H,P, main="Fabric Quality Extensibility Comparison", col='green',
        names=c('High Quality','Poor Quality'))
```




There is a lot of overlap between the two box-plots but the box-plot for Poor Quality shows more extreme large value outliers. Which implies that its mean may be larger.

c)2-sample t-test to determine if there is a difference in their true means. First test to see if it is safe to assume the variances are different or equal.

```r
hq<-fabric_data$H[!is.na(fabric_data$H)]
pq<-fabric_data$P[!is.na(fabric_data$P)]
var.test(hq,pq)
```

```
##
##  F test to compare two variances
##
## data:  hq and pq
## F = 0.70158, num df = 23, denom df = 7, p-value = 0.4862
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1585015 2.0362234
## sample estimates:
## ratio of variances
##          0.7015781
```

P-value of the from the F-test is 0.4862. Therefore we have little or no evidence against H0. Thus it is safe to assume that there is no significant difference between the two samples variances.

Now the t-test.

```r
t.test(hq,pq,alternative="two.sided", mu=0, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  hq and pq
## t = -0.41638, df = 30, p-value = 0.6801
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4674695  0.3091362
## sample estimates:
## mean of x mean of y
##  1.508333  1.587500
```

P-value of the t-test is 0.6801. Therefore there is zero or little evidence against H0. Therefore we can conclude that the average extensibility does not differ for the two types of fabric.