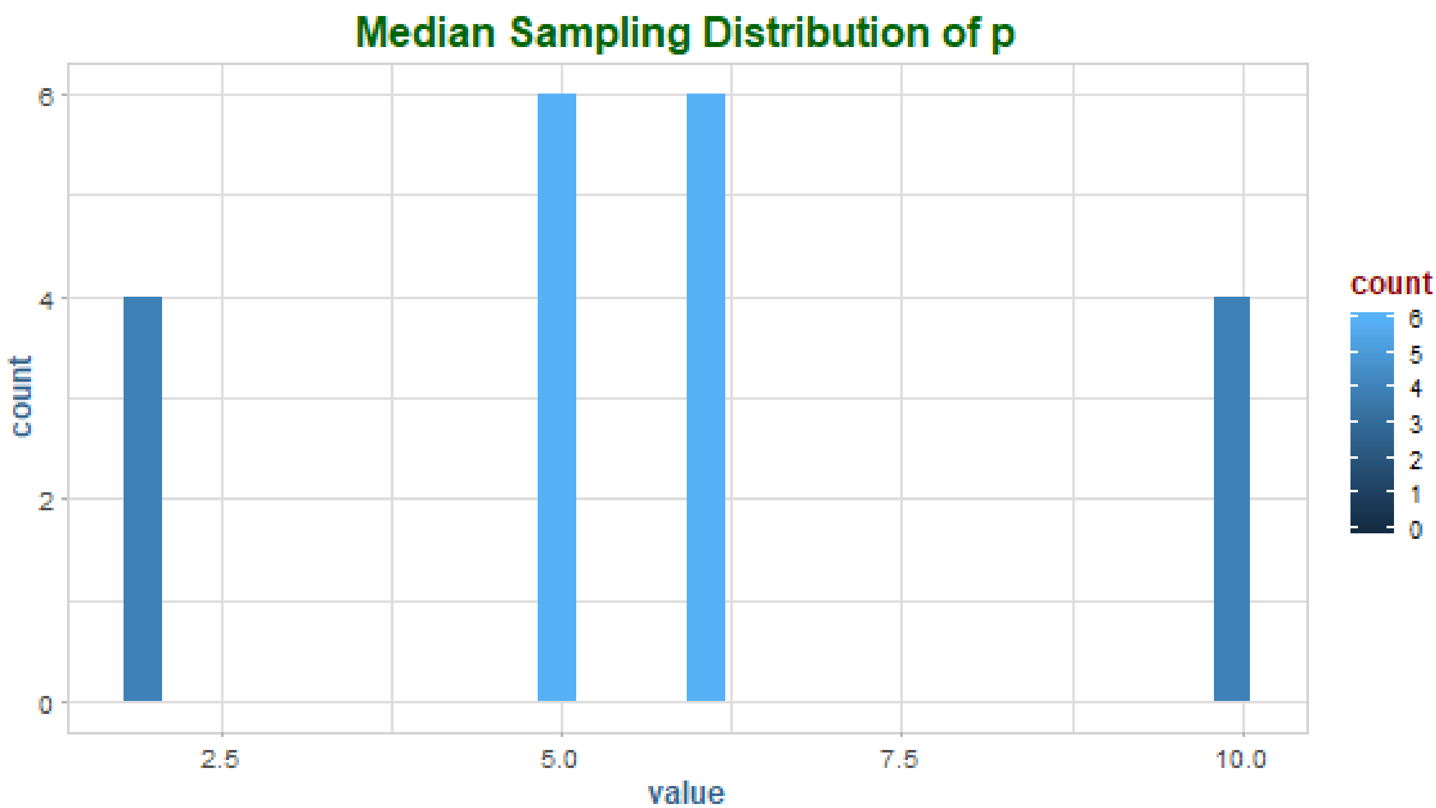
4.1

Consider the population {1, 2, 5, 6, 10, 12}.

Find (and plot) the sampling distribution of medians for samples of size 3 without replacement.

```
p <- c(1, 2, 5, 6, 10, 12)
c <- combinations(v = p, n = 6, r = 3)
t <- apply(c, 1, median)

ggplot(data.table(value = t), aes(value, fill = ..count..)) +
  geom_histogram(bins = 30) +
  labs(title = "Median Sampling Distribution of p")
```



Compare the median of the population to the mean of the medians.

Median of p = **5.5**. Mean of Medians of p = **5.7**

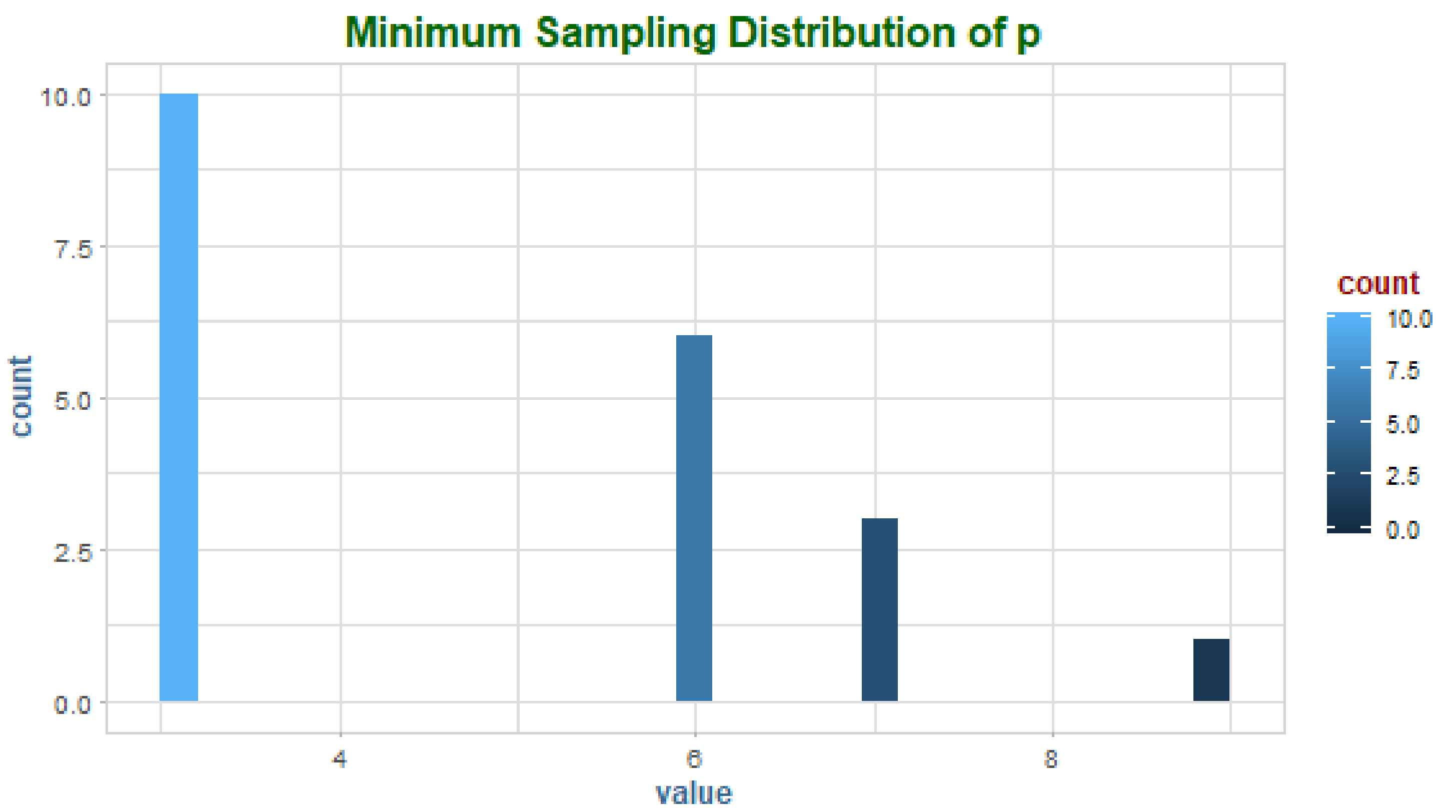
4.2

Consider the population {3, 6, 7, 9, 11, 14}.

For samples of size 3 without replacement, find (and plot) the sampling distribution for the minimum.

```
p <- c(3, 6, 7, 9, 11, 14)
c <- combinations(v = p, n = 6, r = 3)
t <- apply(c, 1, min)

ggplot(data.table(value = t), aes(value, fill = ..count..)) +
   geom_histogram(bins = 30) +
   labs(title = "Minimum Sampling Distribution of p")
```



What is the mean of the sampling distribution? **4.8**

The statistic is an estimate of some parameter - what is the value of that parameter?

This is an estimation of the minimum, which is: **3**

4.3

Let *A* denote the population {1, 3, 4, 5} and *B* the population {5, 7, 9}.

```
A <- c(1, 3, 4, 5)
B <- c(5, 7, 9)
```

Let *X* be a random value from *A*, and *Y* and random value from *B*.

a.) Find the sampling distribution of *X* + *Y*.

```
result = numeric(12)
index <- 1
for(j in 1:length(A))
{
   for(k in 1:length(B))
   {
      result[index] <- A[j] + B[k]
      index <- index + 1
   }
}

sort(result)
```

```
 [1]  6  8  8  9 10 10 10 11 12 12 13 14
```

b.) In this example, does the sampling distribution depend on whether you sample with or without replacement?

No.

Why or why not?

Because 5 in is both sets.

c.) Compute the mean of the values for each of *A* and *B* and the values in the sampling distribution of *X* + *Y*.

Mean of *A*: **3.25**. Mean of *B*: **7**.

Mean of *A* + *B*: **10.25**

How are the means related?

mean(A) + mean(B) = mean(A + B).

d.) Suppose you draw a random value from *A* and a random value from *B*.

```
prob <- sum(result >= 13) / length(result)
```

What is the probability that the sum is 13 or larger? **16.6666667%**

4.4

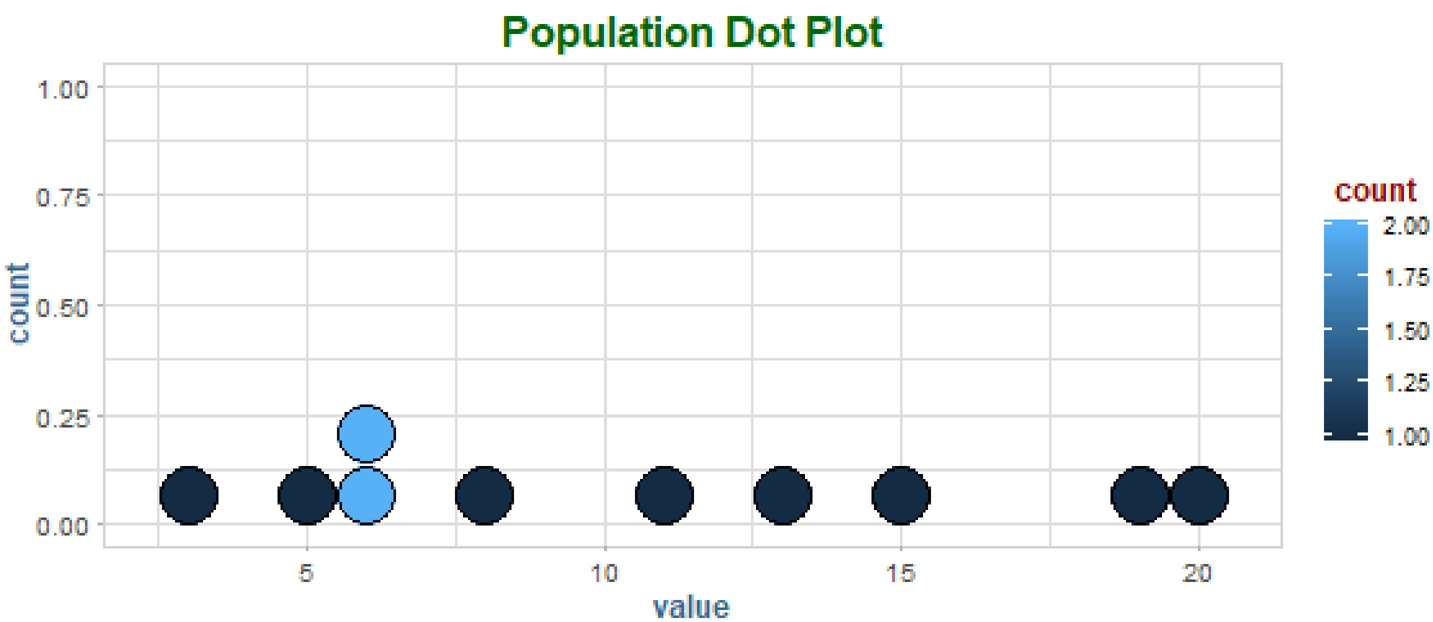
Consider the population {3, 5, 6, 6, 8, 11, 13, 15, 19, 20}.

a.) Compute the mean and standard deviation and create a dot plot of its distribution.

```
p <- c(3, 5, 6, 6, 8, 11, 13, 15, 19, 20)

mu <- mean(p)
sigma <- sd(p)

ggplot(data.table(value = p)) +
   geom_dotplot(aes(value, fill = ..count..), binwidth = 1) +
   labs(title = "Population Dot Plot")
```



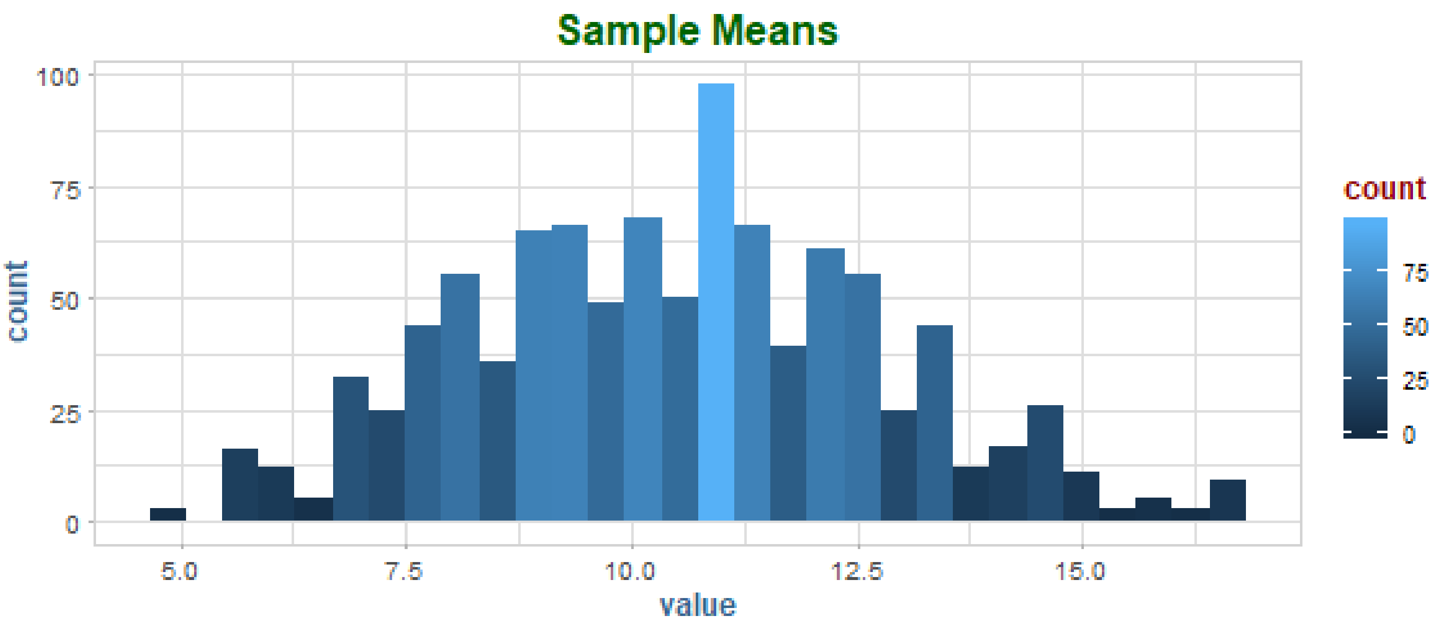
$\mu = 10.6, \sigma = 5.9851668$

b.) Simulate the sampling distribution of $\bar{X}$ by taking random samples of size 4 and plot your results.

```
N <- 10e2
results <- numeric(N)

for( i in 1:N)
{
   index <- sample(length(p), size = 4, replace = F)
   results[i] <- mean( p[index] )
}
```

```r
ggplot(data.table(value = results)) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Sample Means")
```




```r
xbar <- mean(results)
se <- sd(results) / sqrt(N)
```

Compute the mean and standard error, and compare to the population mean and standard deviation.

mean: 10.47325, standard error: 0.0728957

c.) Use the simulation to find $P(\bar{X} < 11)$.

```r
prob <- mean(results < 11)
```

$P(\bar{X} < 11) = 56.9\%$

4.5

Consider two populations A = {3, 5, 7, 9, 10, 16}, B = {8, 10, 11, 15, 18, 25, 28}.

```
A <- c(3, 5, 7, 9, 10, 16)
B <- c(8, 10, 11, 15, 18, 25, 28)
```

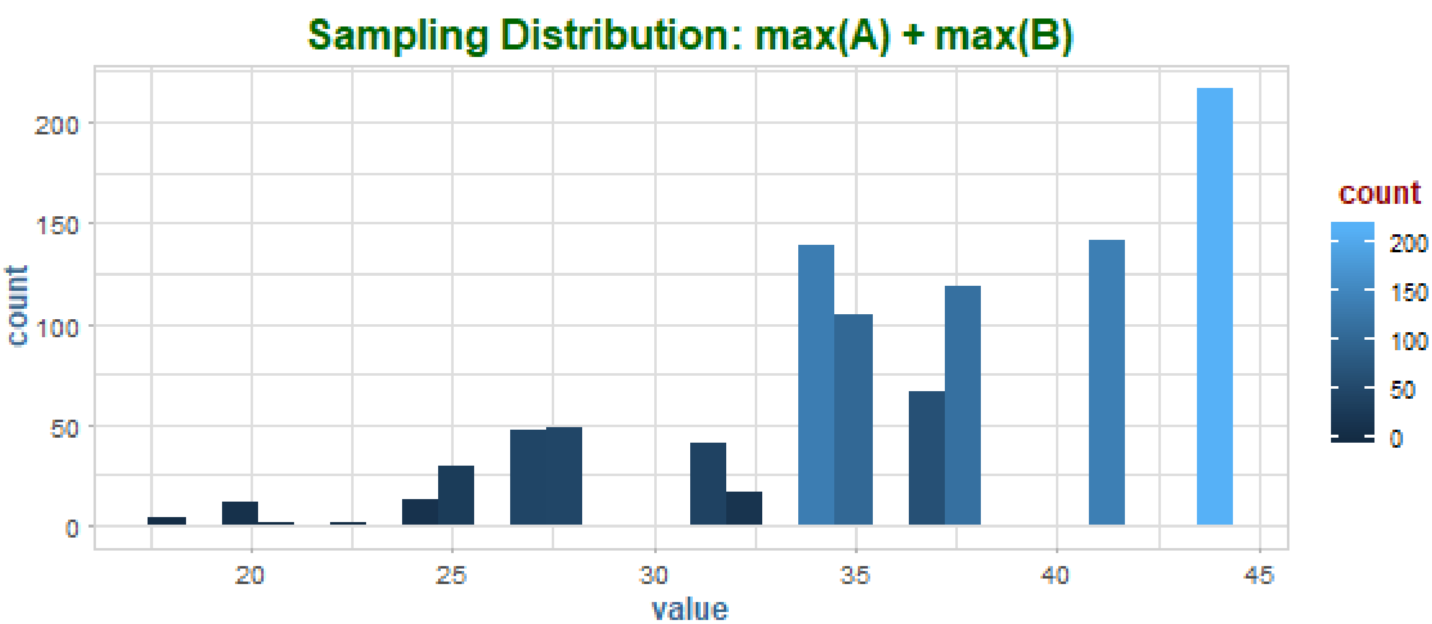
a.) Using R, draw random samples (without replacement) of size 3 from each population, and simulate the sampling distribution of the sum of their maximums.

```
N <- 10e2
results <- numeric(N)

for(i in 1:N)
{
   samp.a <- sample(A, 3, replace = F)
   samp.b <- sample(B, 3, replace = F)

   results[i] <- max(samp.a) + max(samp.b)
}

ggplot(data.table(value = results)[, index := .I]) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Sampling Distribution: max(A) + max(B)")
```



b.) Use your simulation to estimate the probability that the sum of the maximums is less than 20.

```
prob <- mean(results < 20)
```

Probability: 0.4%

c.) Draw random samples of size 3 from each population, and find the maximum of the union of these two sets.

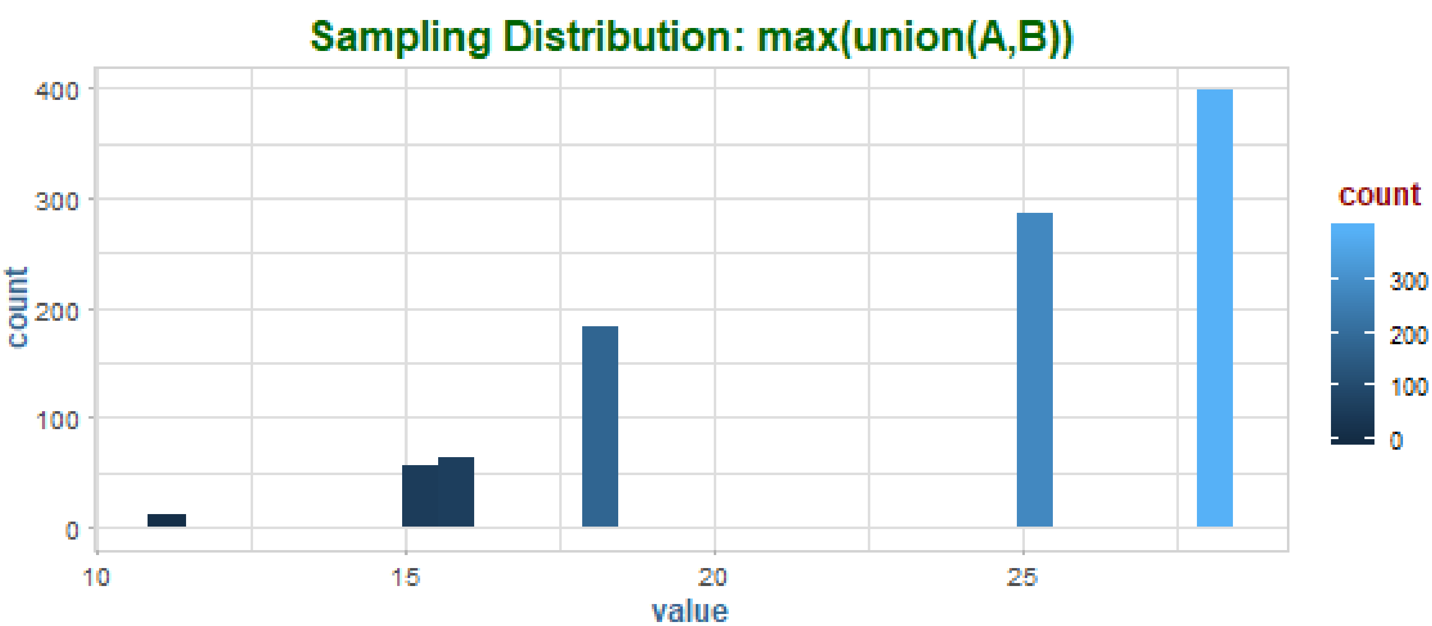
Simulate the sampling distribution of the maximums of this union.

```
results <- numeric(N)

for(i in 1:N)
{
   samp.a <- sample(A, 3, replace = F)
   samp.b <- sample(B, 3, replace = F)

   results[i] <- max(union(samp.a, samp.b))
}

ggplot(data.table(value = results)[, index := .I]) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Sampling Distribution: max(union(A,B))")
```



d.) Use simulation to find the probability that the maximum of the union is less than 20.

```
prob <- mean(results < 20)
```

Probability: 31.6%

4.6

The data set *Recidivism* contains the poopulation of all Iowa offenders convicted of either a felony or misdemeanor who were released in 2010 (case study in Section 1.4).

```
Recidivism <- data.table(read.csv(paste0(data.dir, "Recidivism.csv"),
                                  header = T))
```

Of these, 31.6% recidivated and were sent back to prision.

Simulate the sampling distribution of $\hat{p}$, the sample proportion of offeneders who recidivated, for random samples of size 25.

```
mean(Recidivism$Recid == "Yes")
```

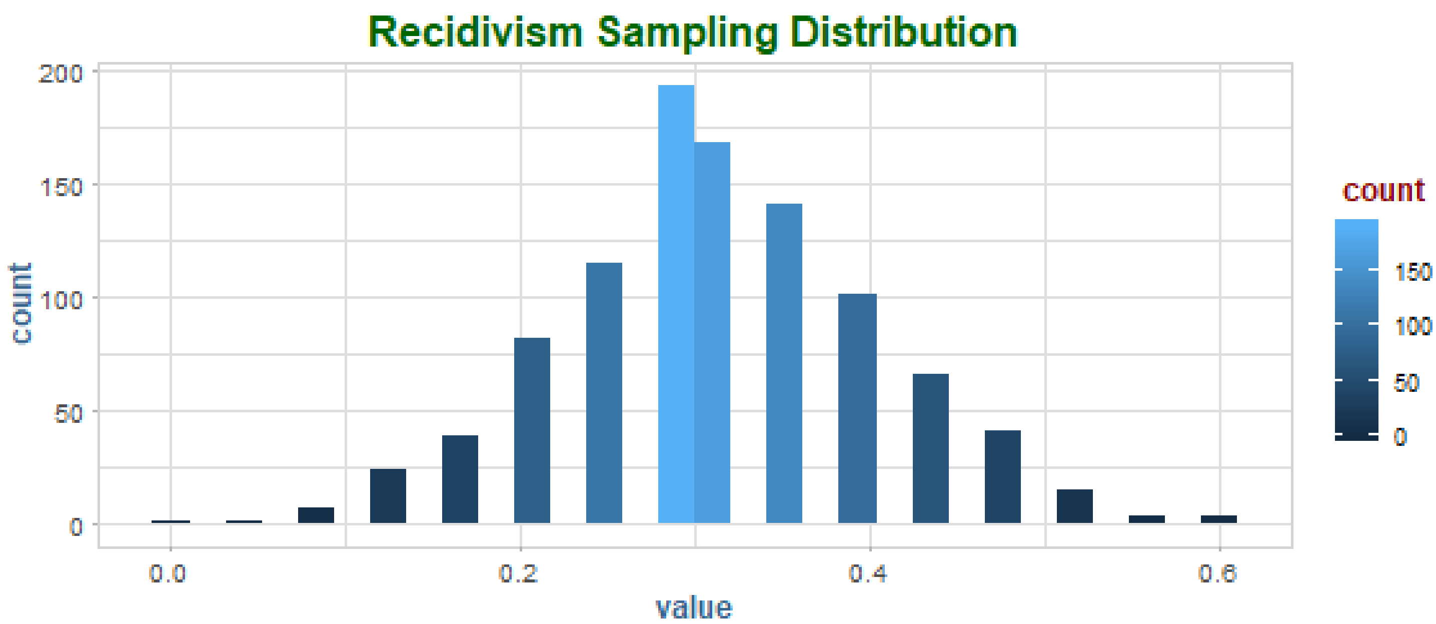
```
[1] 0.3164141
```

```
N <- 10e2
results <- numeric(N)

for(i in 1:N)
{
   samp <- sample(Recidivism$Recid, 25)
   results[i] <- mean(samp == "Yes")
}
```

a.) Create a histogram and describe the simulated sampling distribution of $\hat{p}$.

```
ggplot(data.table(value = results)) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Recidivism Sampling Distribution")
```



Estimate the mean and standard error.

```
mu <- mean(results)
se <- sd(results) / sqrt(25)
```

$\mu = 0.31288, \sigma = 0.0186139$

b.) Compare your estimate of the standard error with the theoretical standard error (*Corollary 4.3.2*).

```
tse <- mu * ( 1 - mu ) / sqrt(25)
```

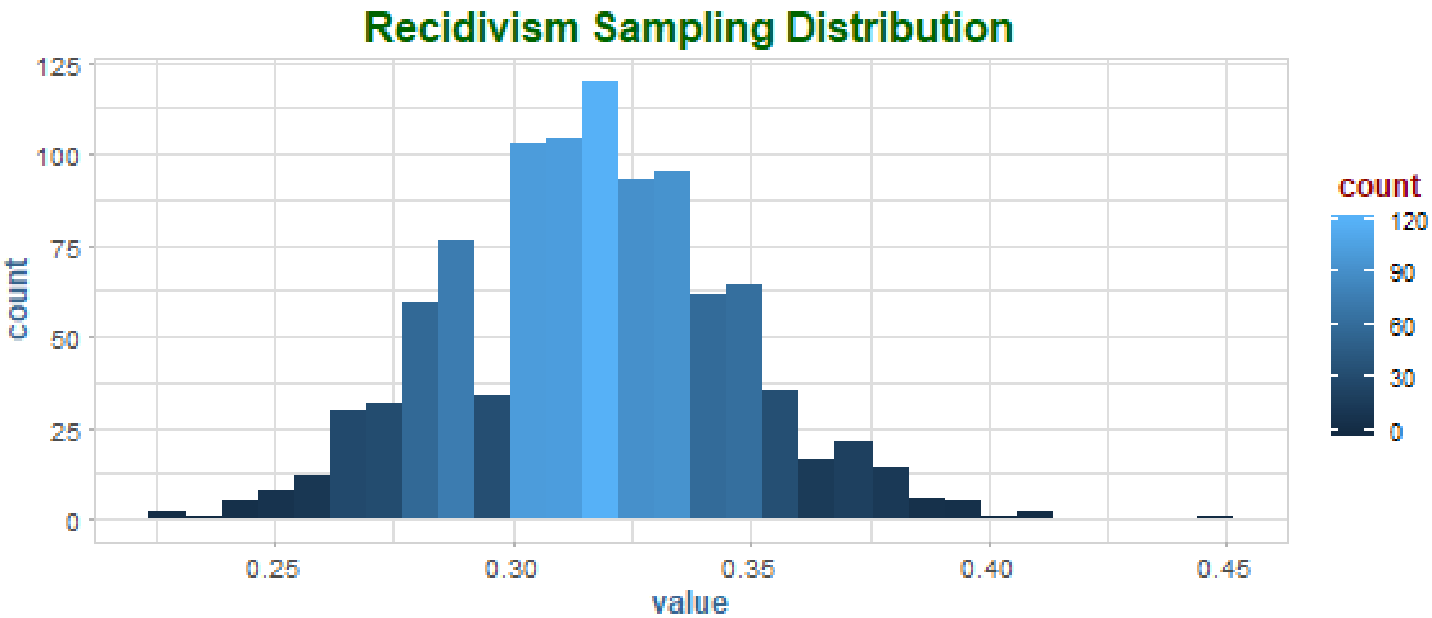
Theoretical: 0.0429972

c.) Repeat the above using samples of size 250, and compare with the $n = 25$ case.

```
N <- 10e2
results <- numeric(N)

for(i in 1:N)
{
   samp <- sample(Recidivism$Recid, 250)
   results[i] <- mean(samp == "Yes")
}

ggplot(data.table(value = results)) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Recidivism Sampling Distribution")
```




```r
mu <- mean(results)
se <- sd(results) / sqrt(250)
```

$\mu = 0.31664, \sigma = 0.0018803$

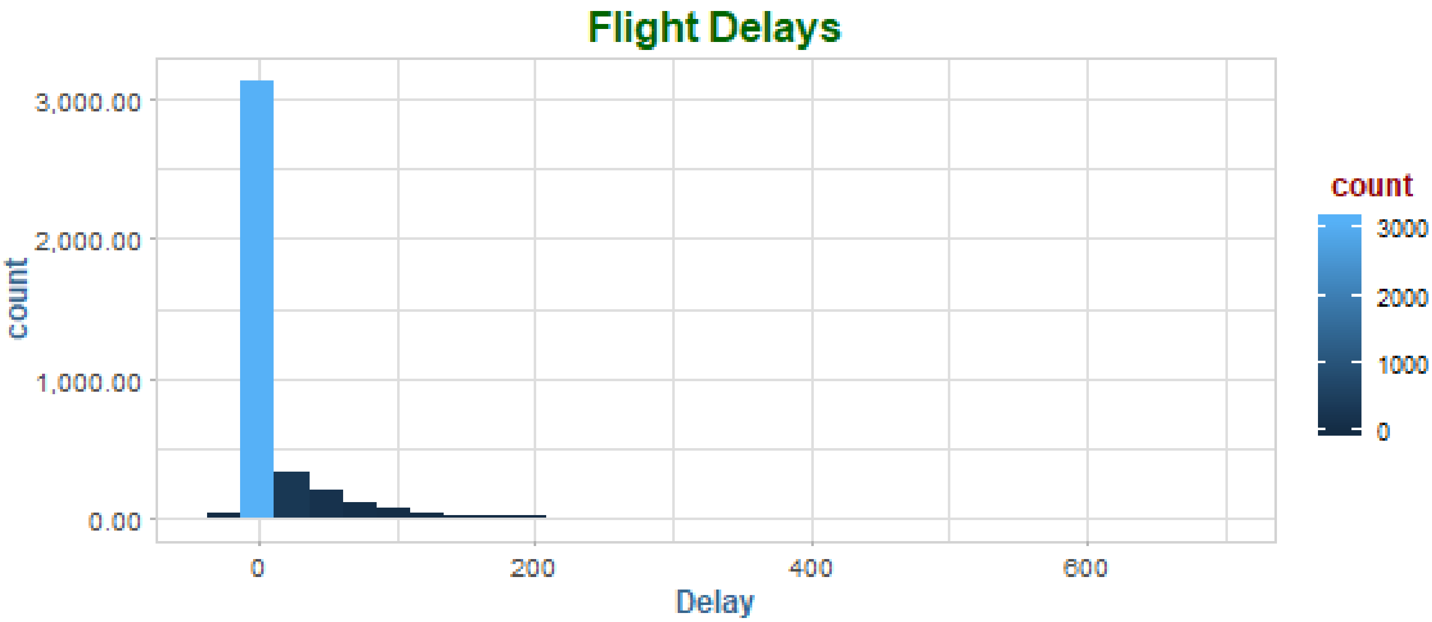
4.7

The data set *FlightDelays* contains the population of all flight departures by United Airlines and American Airlines out of LGA during May and June 2009 (case study in Section 1.1).

```r
Flights <- data.table(read.csv(paste0(data.dir, "FlightDelays.csv"),
                               header = T))
```

a.) Create a histogram of *Delay* and describe the distribution.

```r
ggplot(Flights, aes(Delay)) +
  geom_histogram(aes(fill = ..count..), bins = 30) +
  scale_y_continuous(labels = comma) +
  labs(title = "Flight Delays")
```



Compute the mean and standard deviation.

```
mu <- mean(Flights$Delay)
sigma <- sd(Flights$Delay)
```

$\mu = 11.7379002, \sigma = 41.6304951$

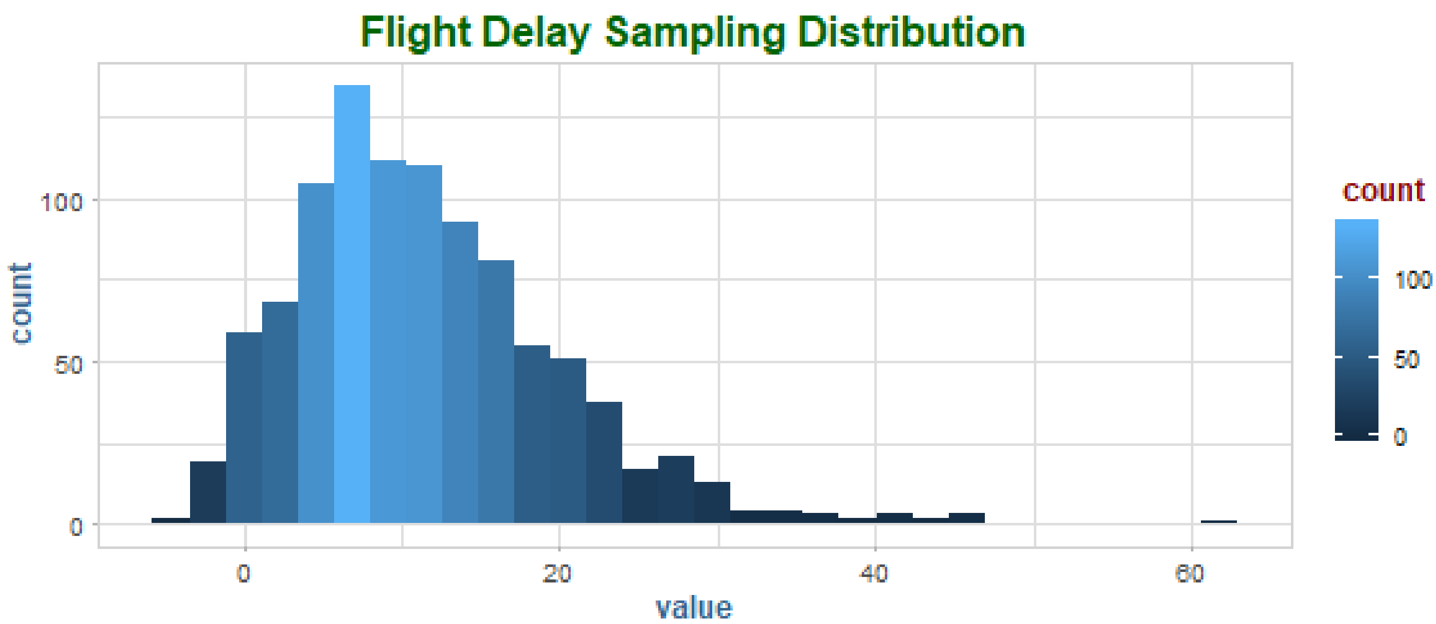
b.) Simulate the sampling distribution of $\bar{x}$, the sample mean of the length of the flight delays (*Delay*), for sample size 25.

```
N <- 10e2
results <- numeric(N)

for(i in 1:N)
{
   samp <- sample(Flights$Delay, 25, replace = F)
   results[i] <- mean(samp)
}
```

Create a histogram and describe the simulated sampling distribution of $\bar{x}$.

```
ggplot(data.table(value = results)) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Flight Delay Sampling Distribution")
```



Estimate the mean and standard error.

```
mu <- mean(results)
se <- sd(results) / sqrt(25)
```

$\mu = 11.52556, \Sigma = 1.6739902$

c.) Compare your estimate of the standard error with the theoretical standard error (*Corollary A.4.1*).

```
tse <- var(results) / 25
```

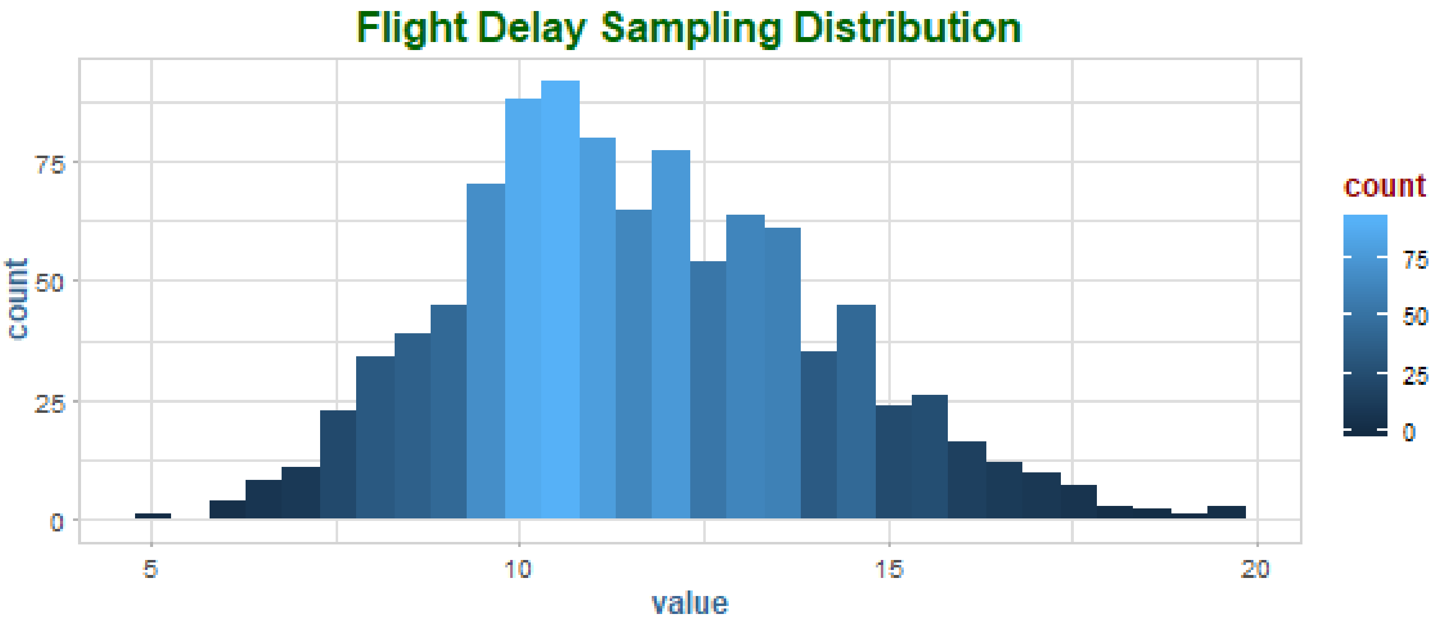
Theoretical: 2.8022431

d.) Repeat with sample size 250.

```
N <- 10e2
results <- numeric(N)

for(i in 1:N)
{
   samp <- sample(Flights$Delay, 250, replace = F)
   results[i] <- mean(samp)
}

ggplot(data.table(value = results)) +
   geom_histogram(aes(value, fill = ..count..), bins = 30) +
   labs(title = "Flight Delay Sampling Distribution")
```




```
mu <- mean(results)
se <- sd(results) / sqrt(250)
tse <- var(results) / 250
```

$\mu = 11.614224, \Sigma = 0.1563767$

Theoretical: 0.0244537

4.8

Let $X_1, X_2, \ldots, X_{25}$ be a random sample from some distribution and $W = T(X_1, X_2, \ldots, X_n)$ be a statistic.

Suppose the *sampling distribution* of W has a pdf given by $f(x) = \frac{2}{x^2}$, for 1 < x < 2.

Find $P(w < 1.5)$

Solution:

4.9

Let $X_1, X_2, \ldots, X_n$ be a random sample from some distribution and $Y = T(X_1, X_2, \ldots, X_n)$ be a statistic.

Suppose the *sampling distribution* of Y has pdf $f(y) = (3/8)y^2 \ for \ 0 \leq y \leq 2$.

Find $P(0 \leq Y \leq \frac{1}{5})$

Solution:

4.10

Suppose the heights of boys in a certain large city follow a distribution with mean 48 in. and variance 9^2.

Use the CLT approximation to estimate the probability that in a random sample of 30 boys, the mean height is more than 51 in.

4.11

Let $X_1, X_2, \ldots, X_{36} \sim Bern(.55)$ be independent, and let $\hat{p}$ denote the sample proportion.

Use the CLT approximation with continuity correction to find the probability that $\hat{p} \leq 0.5$.

4.12

A random sample of size $n = 20$ is drawn from a distribution with mean 6 and variance 10.

Use the CLT approximation to estimate $P(\bar{X} \leq 4.6)$.