

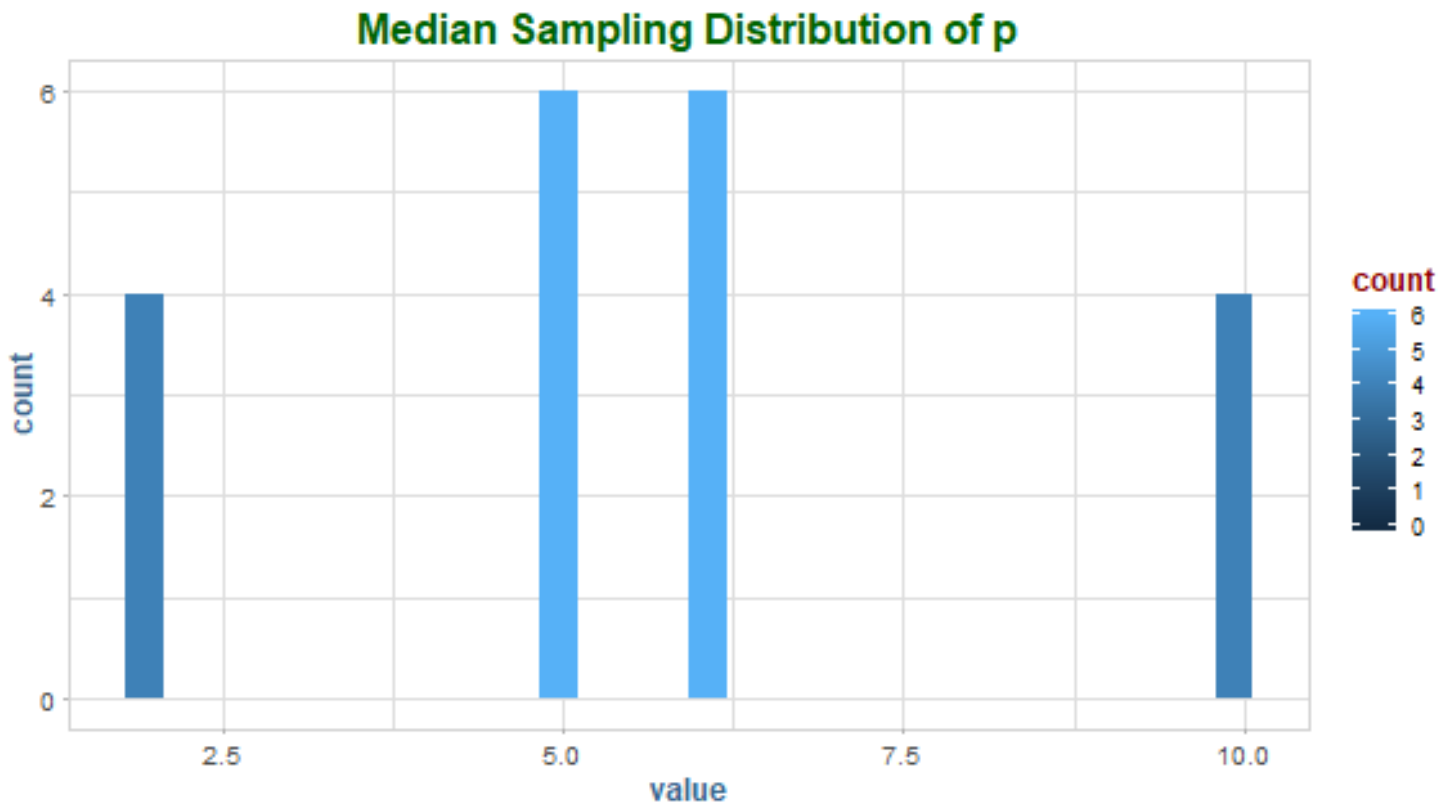
## 4.1

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  $p = 5.7$

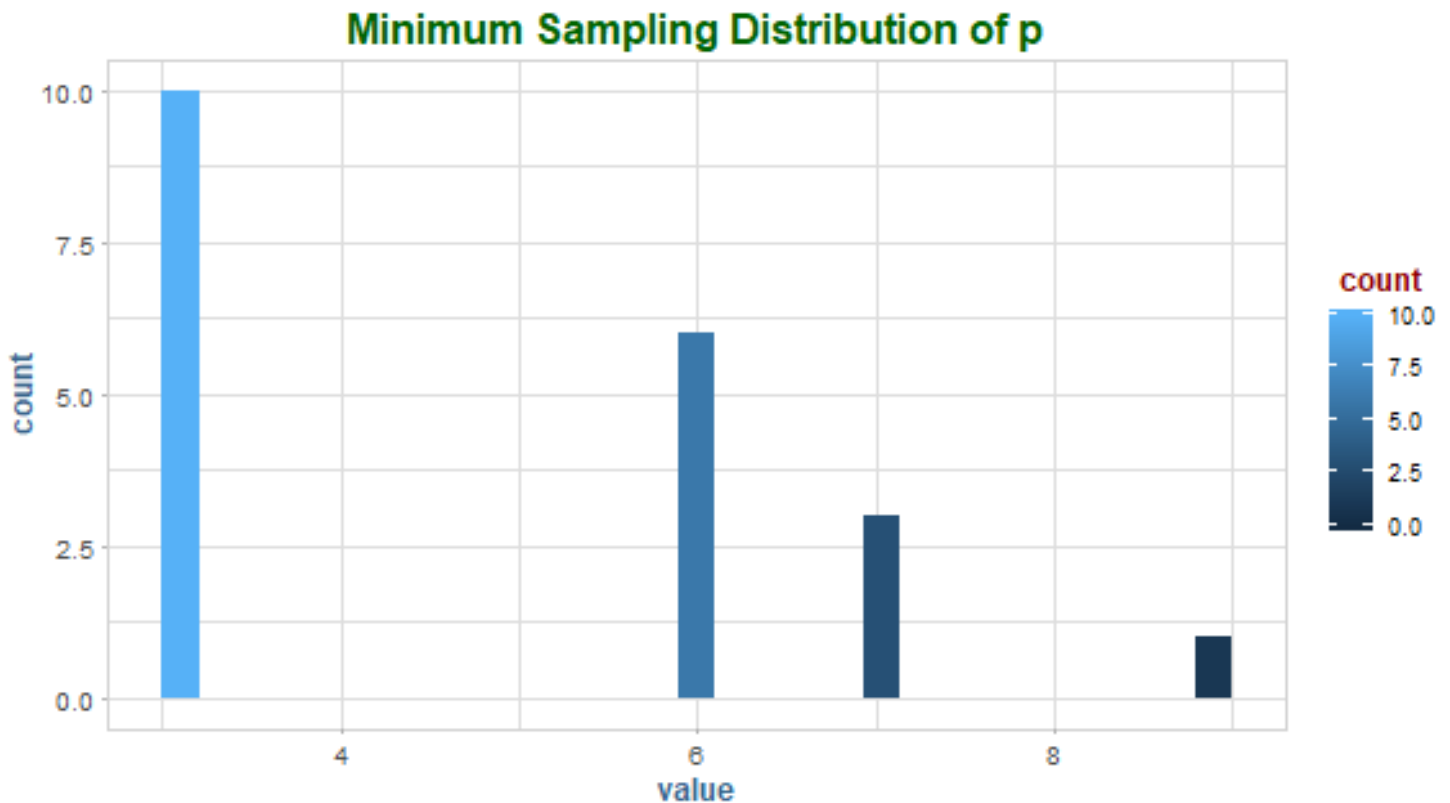
## 4.2

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 is in 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?

$\text{mean}(A) + \text{mean}(B) = \text{mean}(A + B)$ .

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

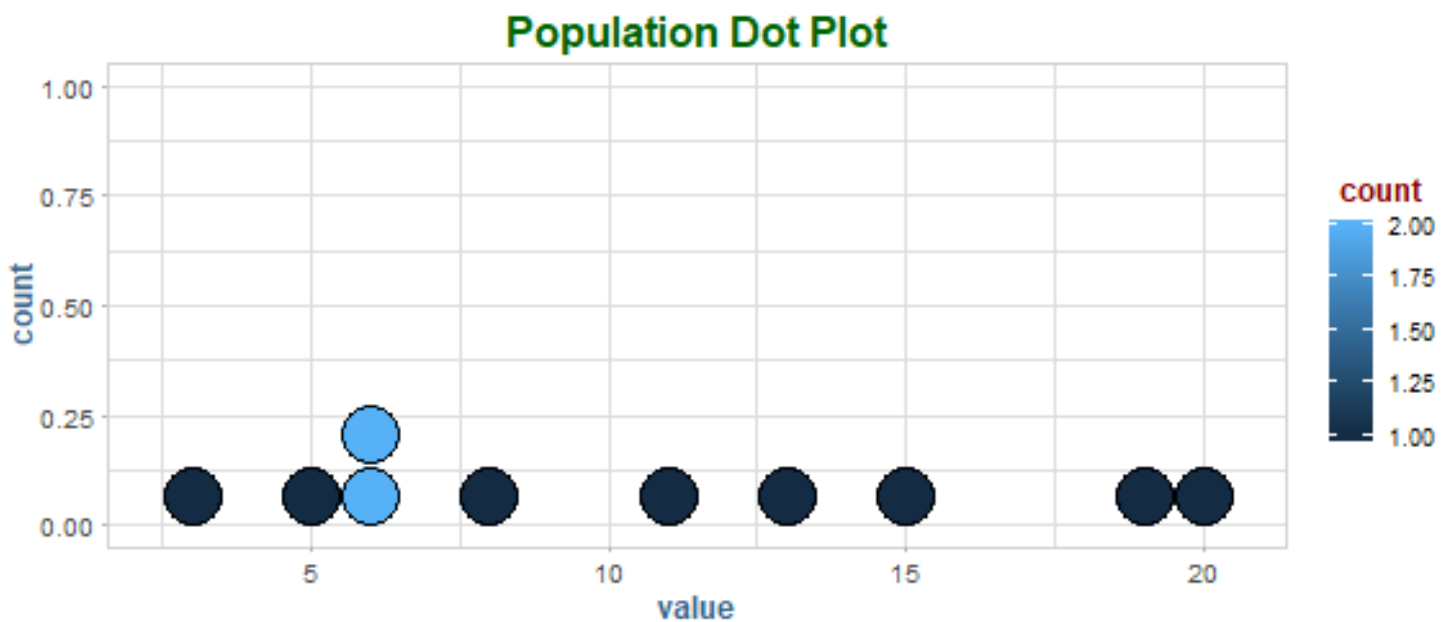
Consider the population  $\{3, 5, 6, 6, 8, 11, 13, 15, 19, 20\}$ .

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")
```



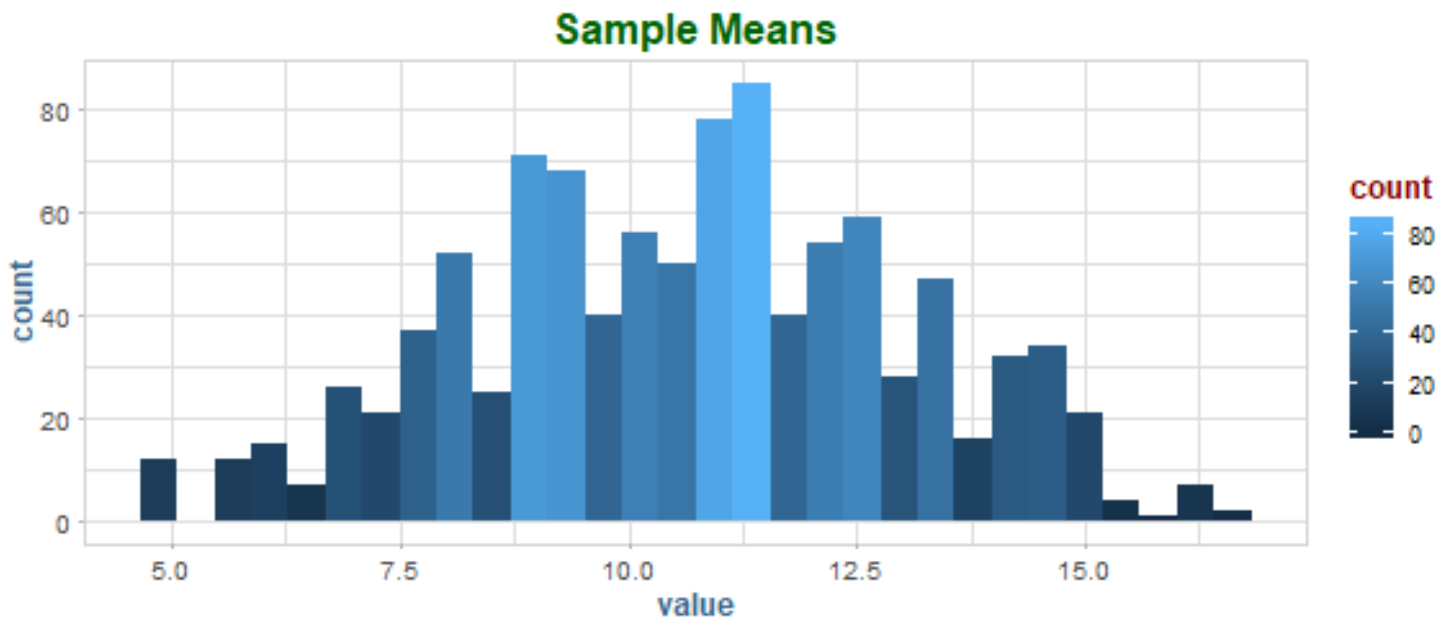
$$\mu = 10.6, \sigma = 5.9851668$$

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] )
}
```

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



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

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

mean: 10.64125, standard error: 0.0756742

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

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

$P(\bar{X} < 11) = 52.5\%$

## 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)
```

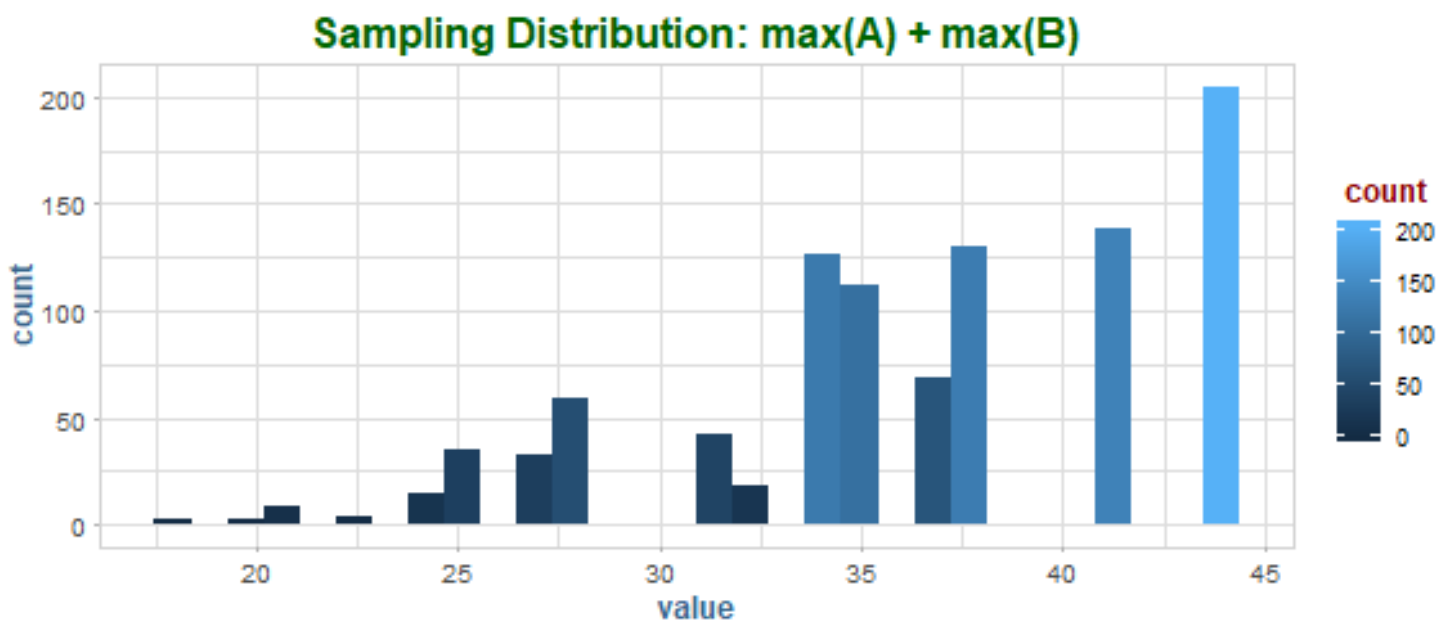
a.) Using R, draw random samples (without replacement) of size 3 from each population, and simulate the sampling distribution of the sum of their maximums.

```
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: 0.3%

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

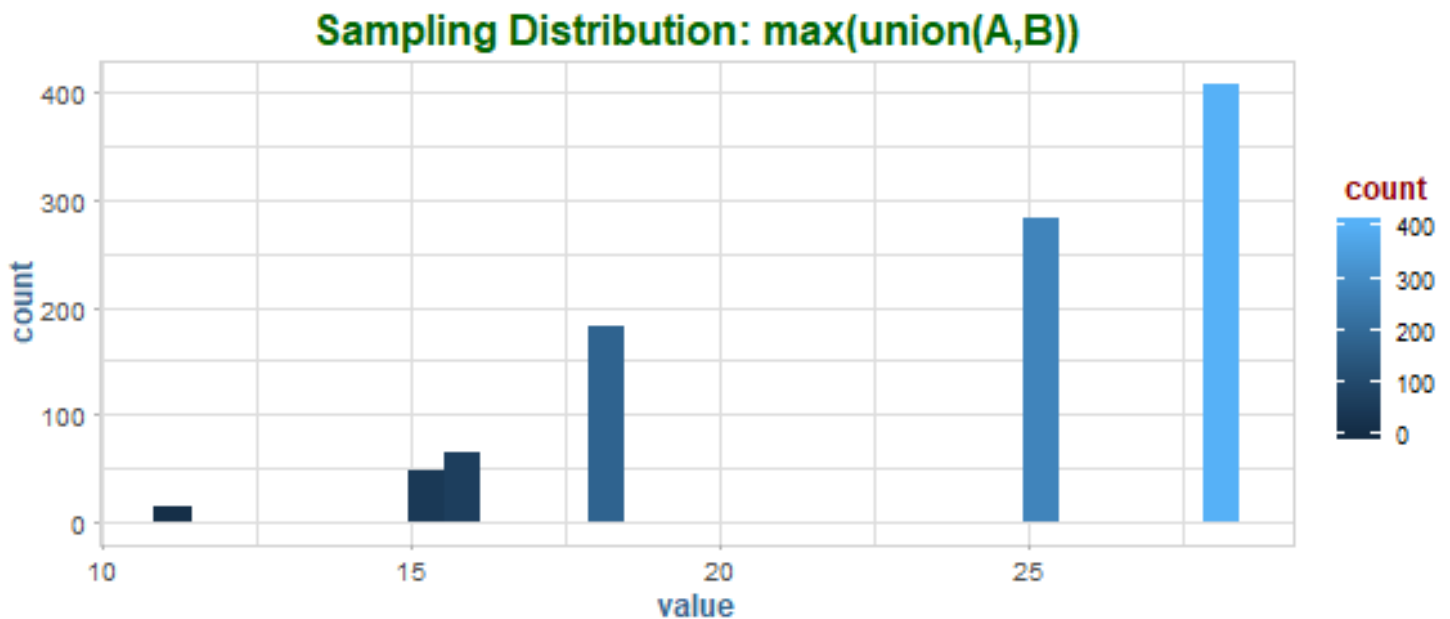
Simulate the sampling distribution of the maximums of this union.

```
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%

## 4.6

The data set *Recidivism* contains the population 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 prison.

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

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

```
[1] 0.3164141
```

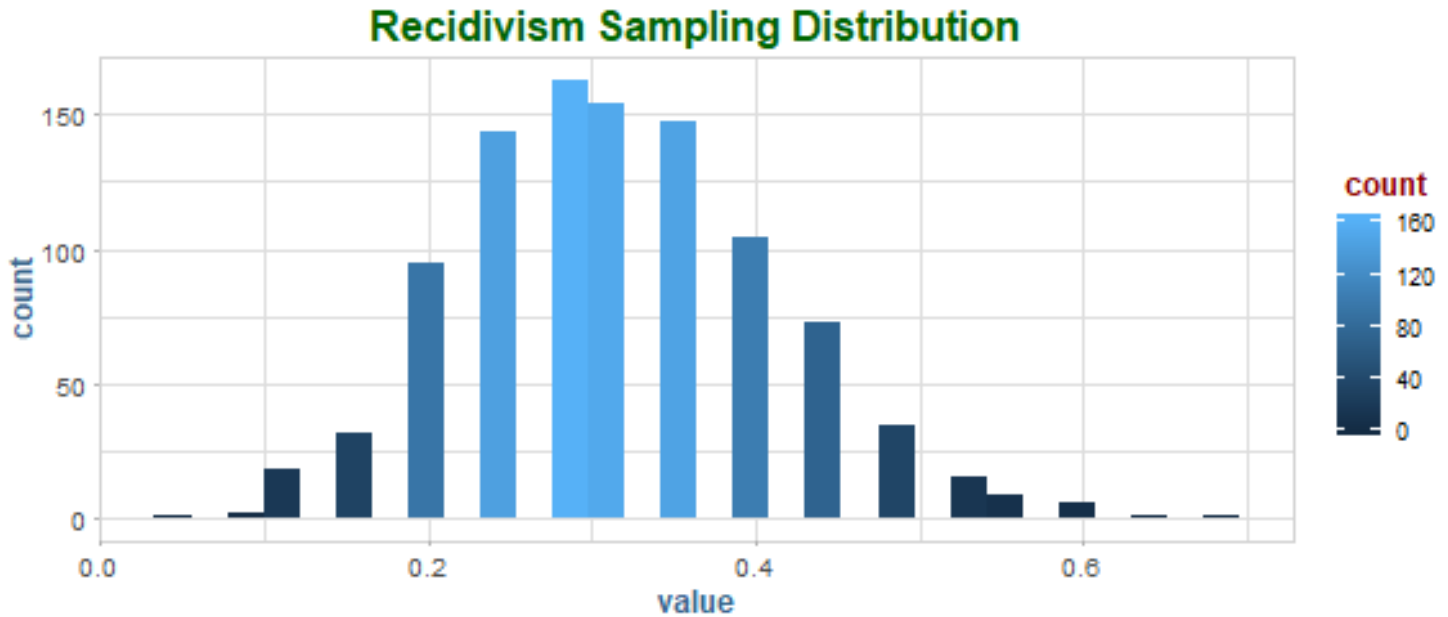
```
N <- 10e2
results <- numeric(N)

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)
```

$$\mu = 0.31716, \sigma = 0.0189239$$

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

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

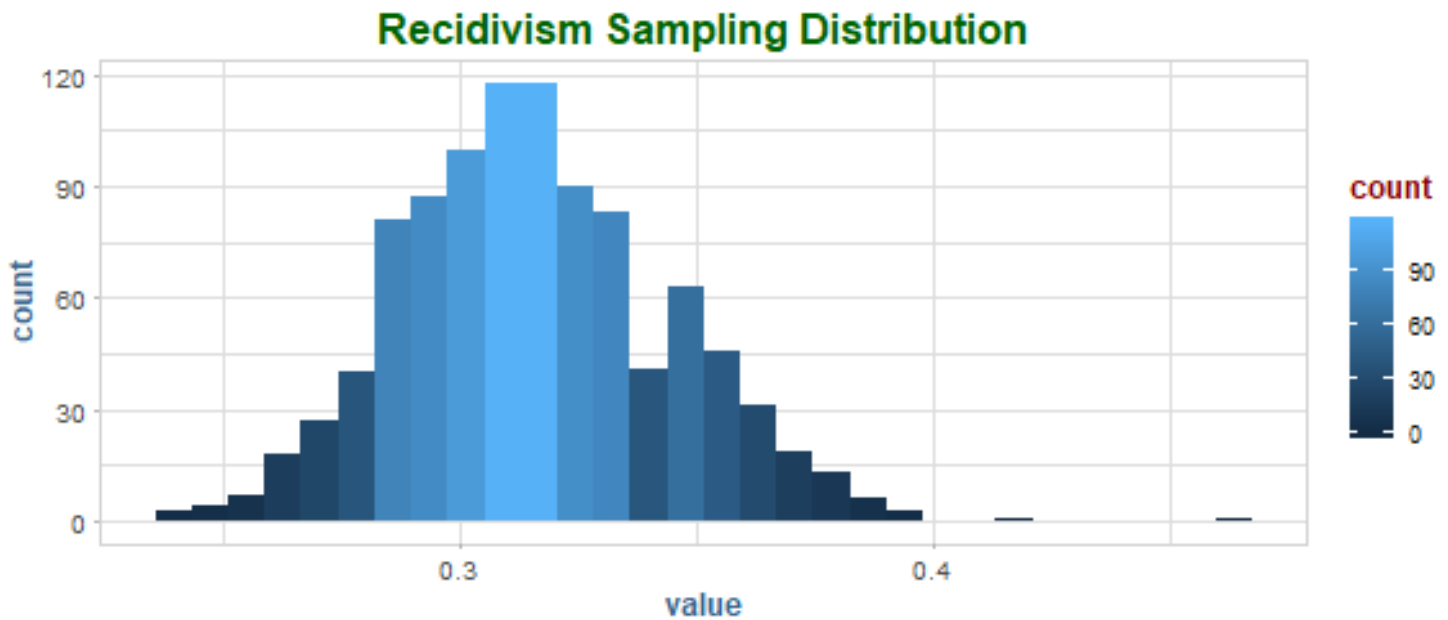
Theoretical: 0.0433139

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")
```



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

$$\mu = 0.31624, \sigma = 0.001795$$

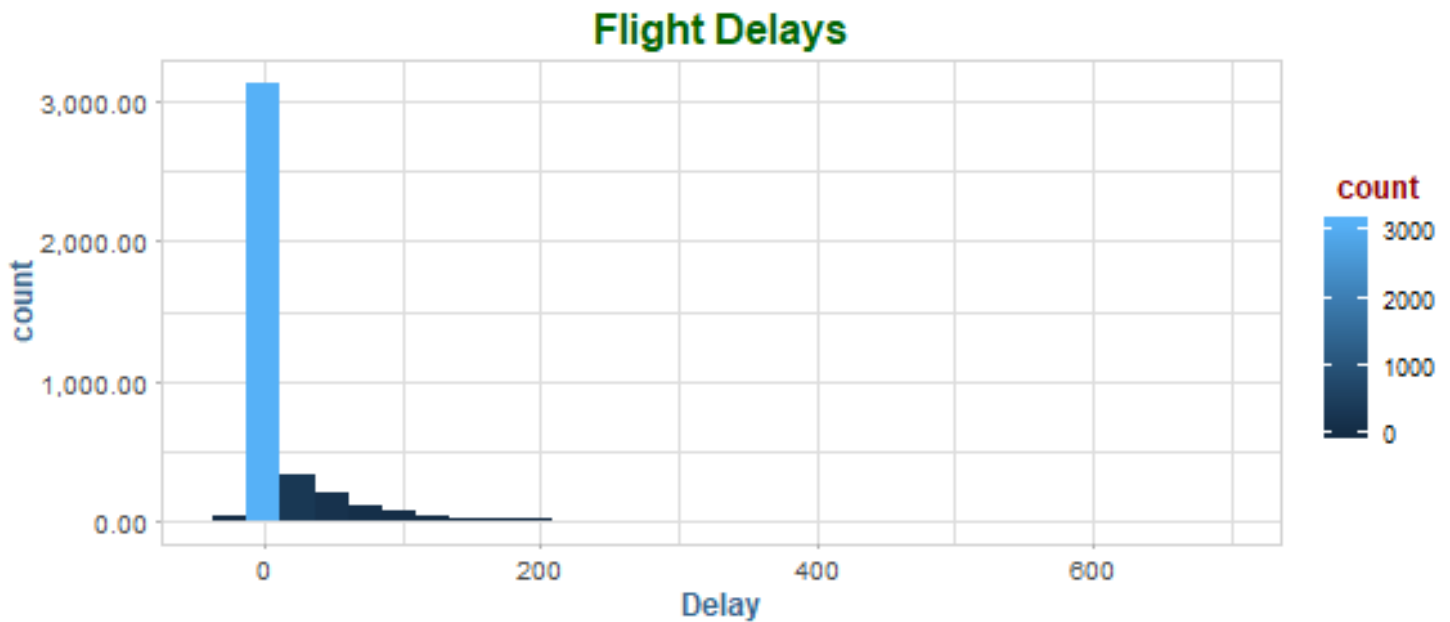
## 4.7

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).

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

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

```
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)
```

$$\mu = 11.7379002, \sigma = 41.6304951$$

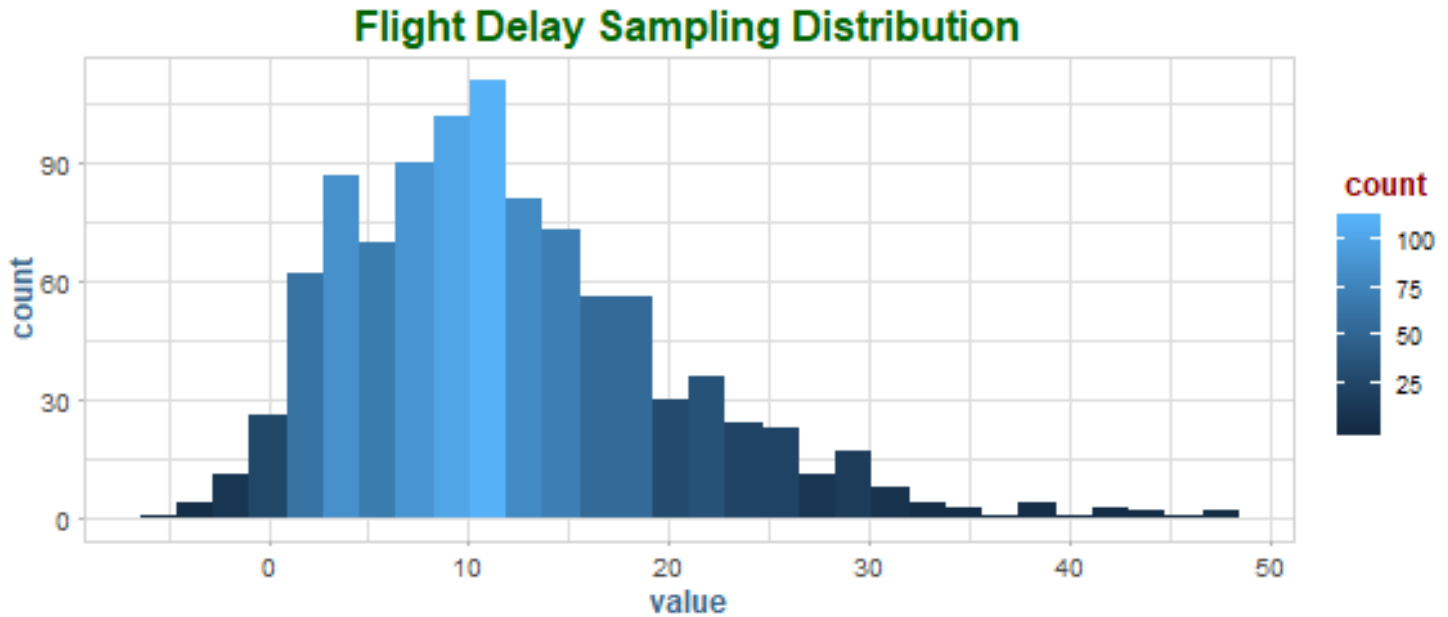
b.) Simulate the sampling distribution of  $\bar{x}$ , the sample mean of the length of the flight delays (*Delay*), for sample size 25.

```
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)
```

$$\mu = 12.03824, \Sigma = 1.655914$$

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

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

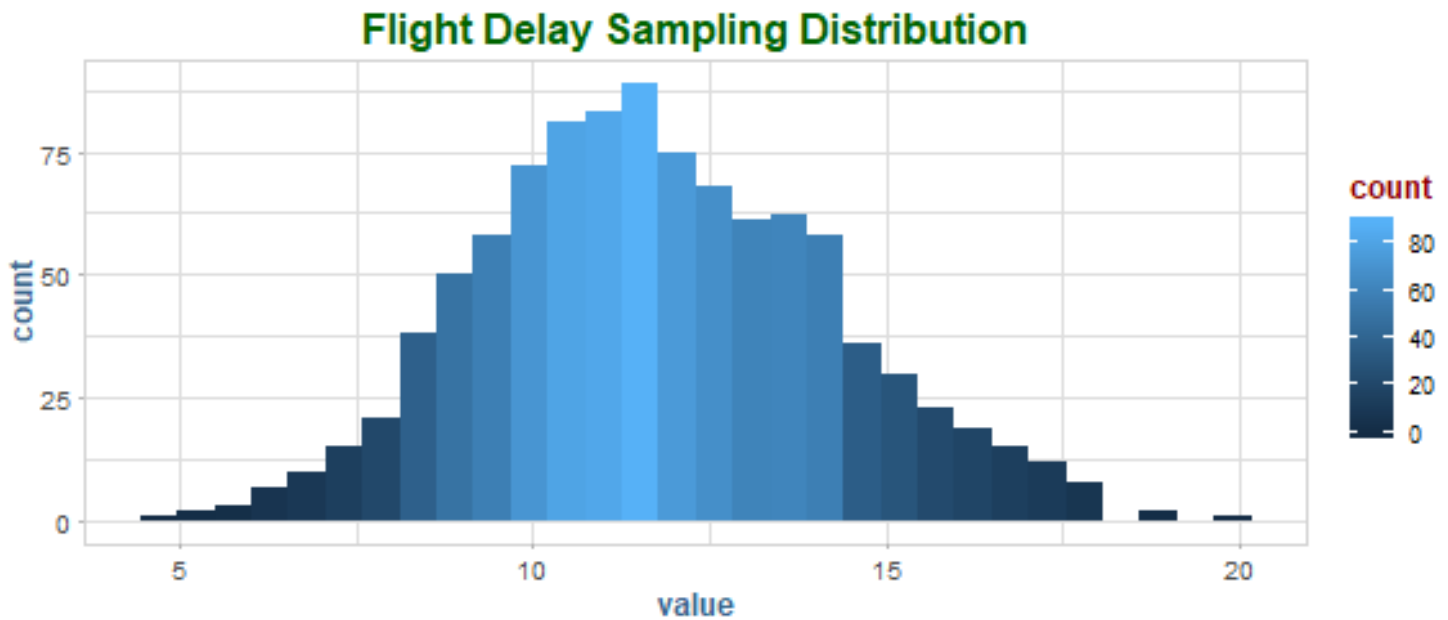
Theoretical: 2.7420512

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
```

$$\mu = 11.771008, \Sigma = 0.1573543$$

Theoretical: 0.0247604

## 4.8

Let  $X_1, X_2, \dots, X_{25}$  be a random sample from some distribution and  $W = T(X_1, X_2, \dots, X_n)$  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  $P(w < 1.5)$

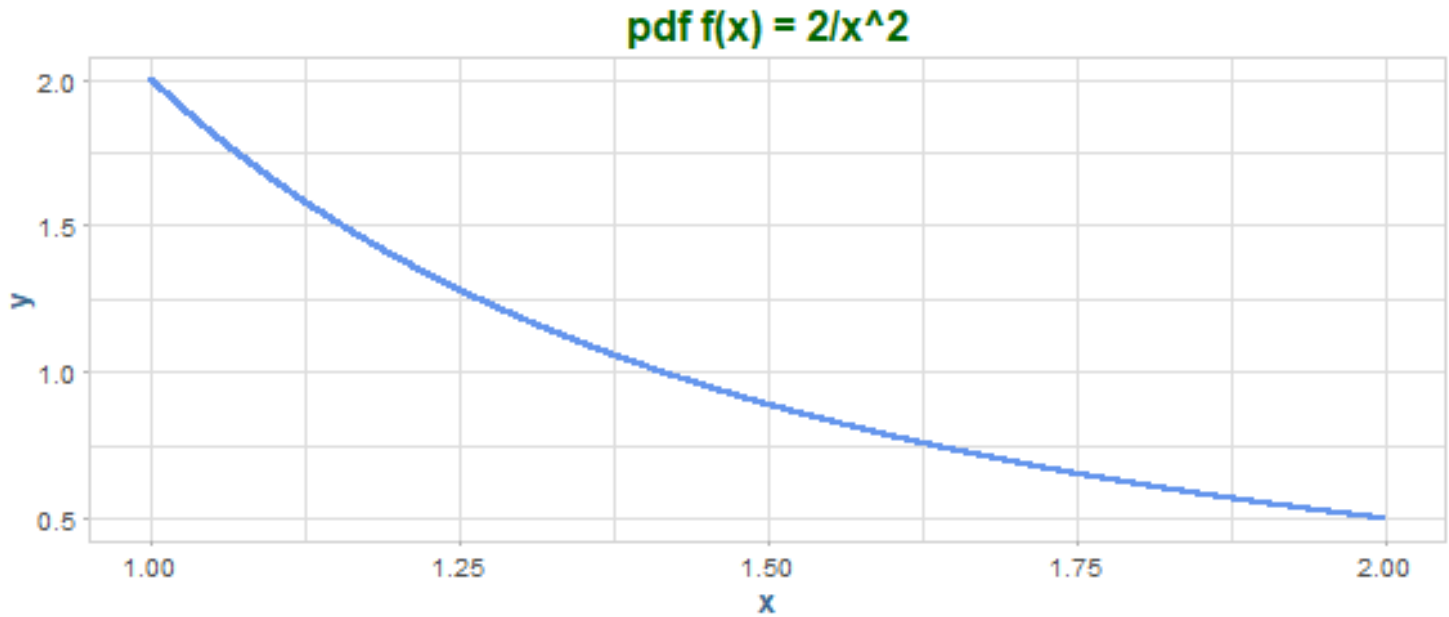
**Solution:**

```
f <- function(x) 2 / x^2

x <- seq( from = 1.0001, to = 1.999, by = 0.0001)

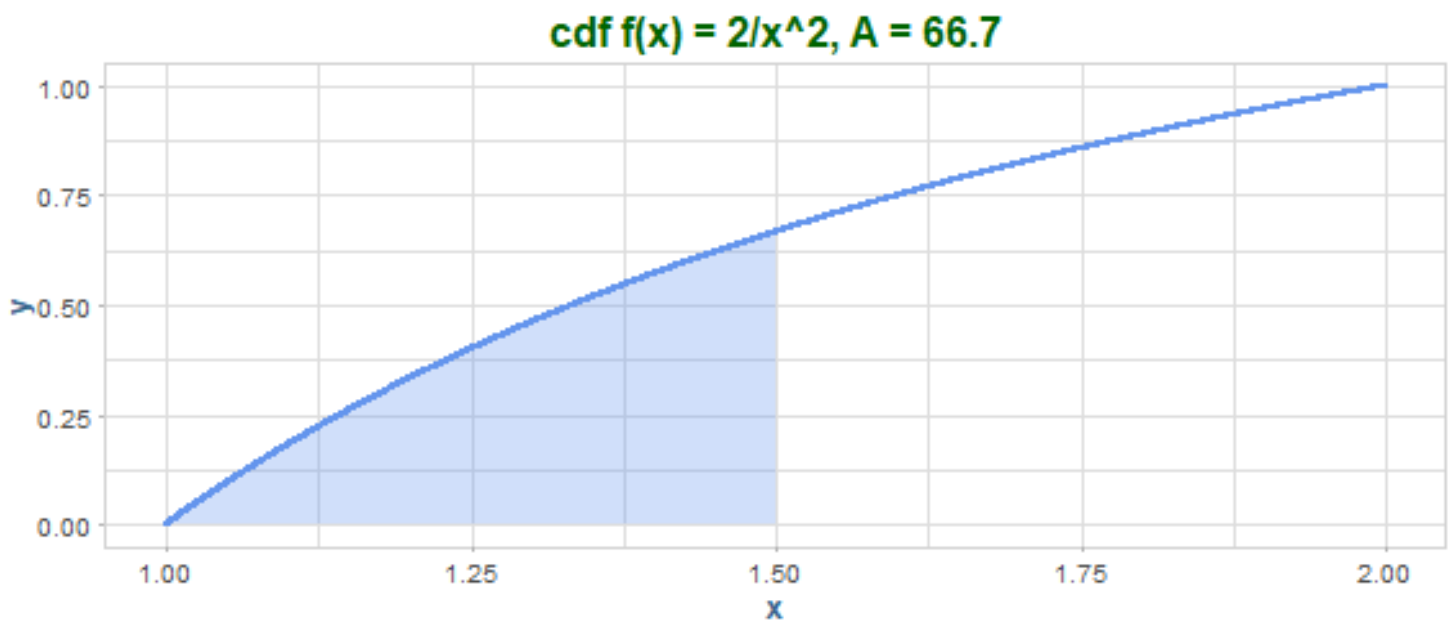
y <- f(x)

ggplot(data.table(x, y)) +
  geom_point(aes(x, y), color = "cornflowerblue", size = .6) +
  labs(title = "pdf f(x) = 2/x^2")
```



```
a <- cumsum(y) / sum(y)
p <- round( a[x == 1.5], 4 ) * 100

d <- data.table(x, y = a)
ggplot(d) +
  geom_point(aes(x, y), color = "cornflowerblue", size = .6) +
  geom_area(aes(x, y), data = d[x < 1.5], fill = "cornflowerblue", alpha = .3) +
  labs(title = paste("cdf f(x) = 2/x^2, A =", p ))
```



Numerical solution: 66.7%

Analytical Solution:  $\int_1^{1.5} \frac{2}{x^2} = \frac{2}{3}$

## 4.9

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

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

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

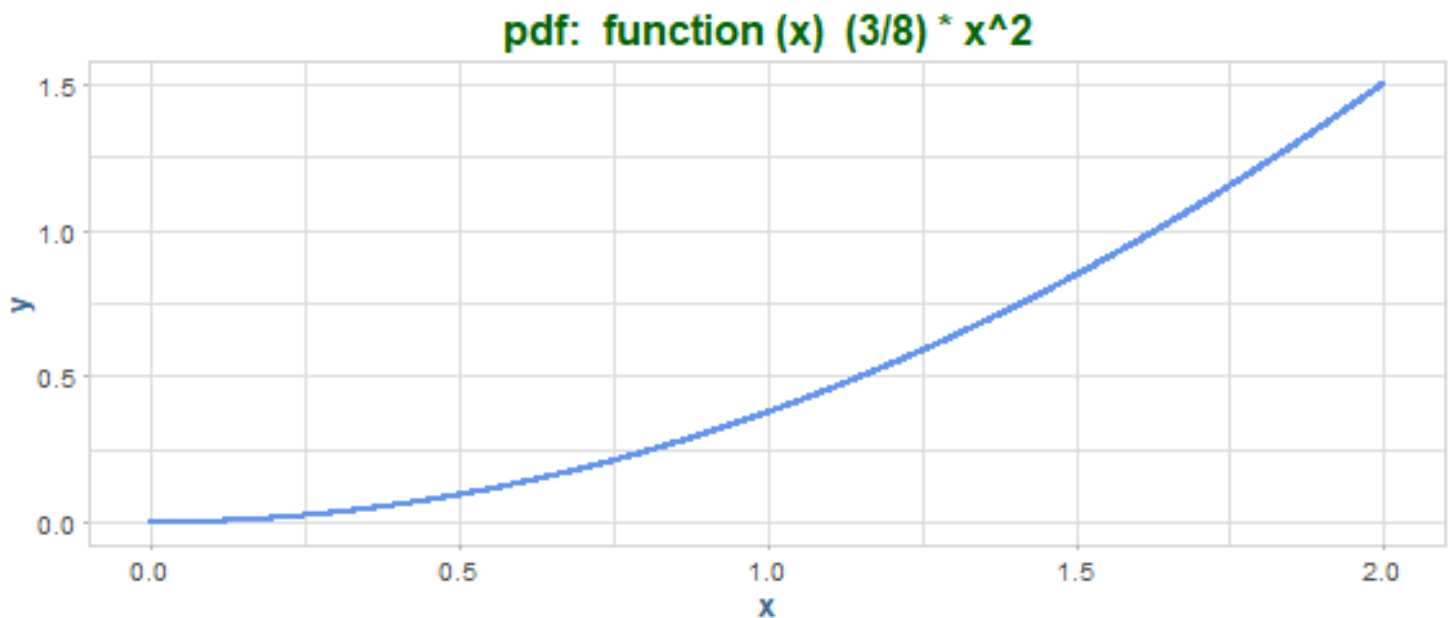
**Solution:**

```
f <- function(x) (3/8)*x**2

x <- seq( from = 0, to = 2, by = 0.001)

y <- f(x)

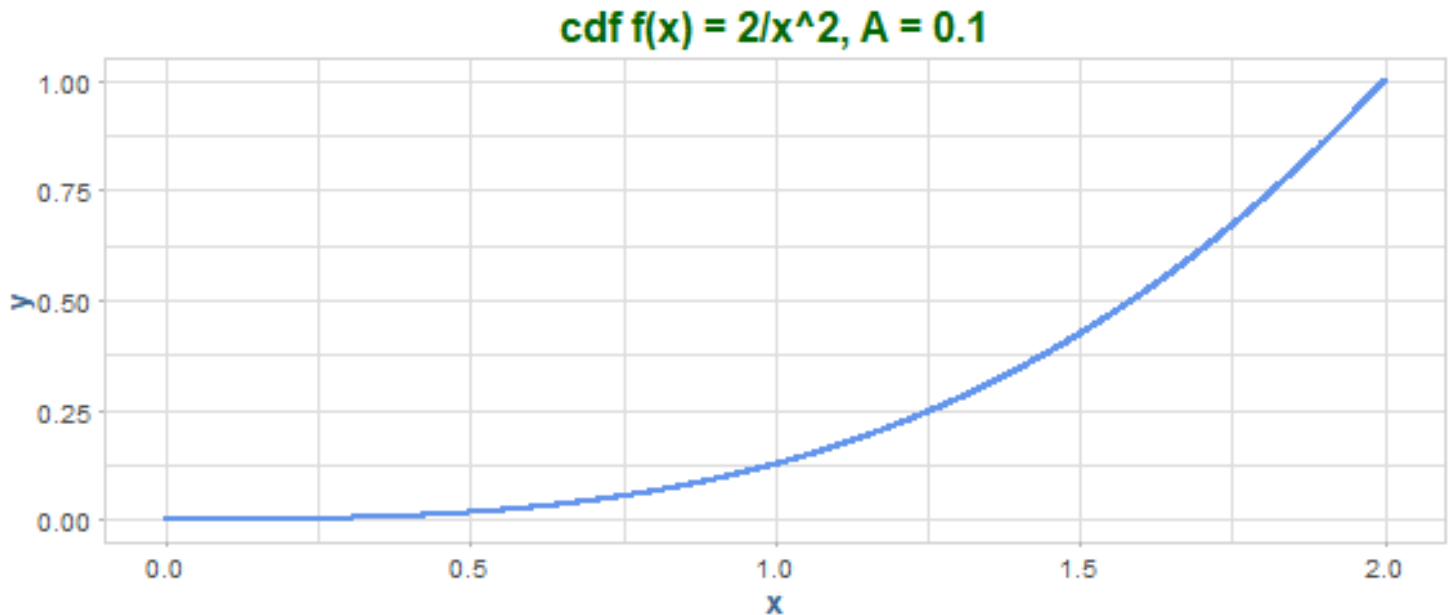
ggplot(data.table(x, y)) +
  geom_point(aes(x, y), color = "cornflowerblue", size = .6) +
  labs(title = paste("pdf: ", paste0(deparse(f), collapse = " ")))
```



```
a <- cumsum(y) / sum(y)
p <- round( a[x == 1/5], 4 ) * 100

d <- data.table(x, y = a)
```

```
ggplot(d) +
  geom_point(aes(x, y), color = "cornflowerblue", size = .6) +
  geom_area(aes(x, y), data = d[x < 1/5], fill = "cornflowerblue", alpha = .3) +
  labs(title = paste("cdf f(x) = 2/x^2, A =", p ))
```



Numerical Solution: 0.1%

Analytical Solution:  $\int_0^{\frac{1}{5}} \frac{x^3}{8} = \frac{.008}{8} = .001 = .1 \%$

#### 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.

```
z <- (51 - 48) / (9^2 / sqrt(30))
p <- pnorm(z, lower.tail = F)
```

Probability: **41.96%**

#### 4.11

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

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



```
z <- ( .5 - .55 ) / sqrt(.55 * (1 - .55) / 36)
p <- pnorm(z, lower.tail = T)
```

Probability: 27.32%

## 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  $P(\bar{X} \leq 4.6)$ .

```
z <- ( 4.6 - 6 ) / ( 10 * sqrt(20) )
p <- pnorm(z, lower.tail = T)
```

Probability: 48.75%

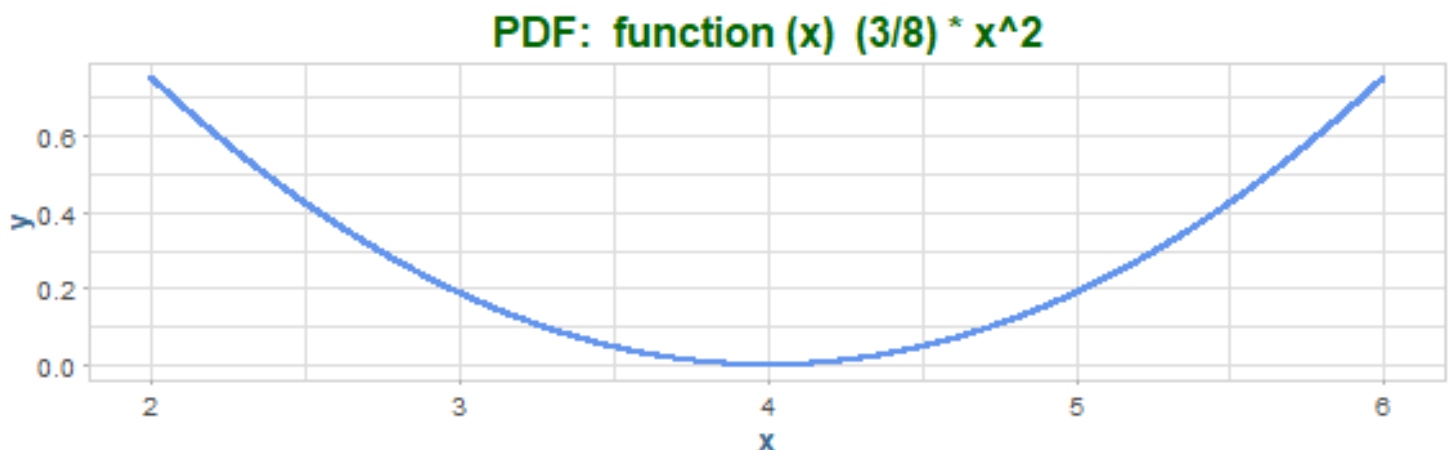
## 4.13

A random sample of size  $n = 244$  is drawn from a distribution with pdf  $f(x) = (3/16)(x - 4)^2, 2 \leq x \leq 6$ .

Use the CLT approximation to estimate  $P(X \geq 4.2)$ .

```
pdf <- function(x) (3/16)*(x - 4)^2
x <- seq(from = 2, to = 6, by = 0.001)
y <- pdf(x)

ggplot(data.table(x,y)) +
  geom_point(aes(x, y), col = "cornflowerblue", lwd = .8) +
  labs(title = paste("PDF: ", paste0(deparse(f), collapse = " ")))
```



```

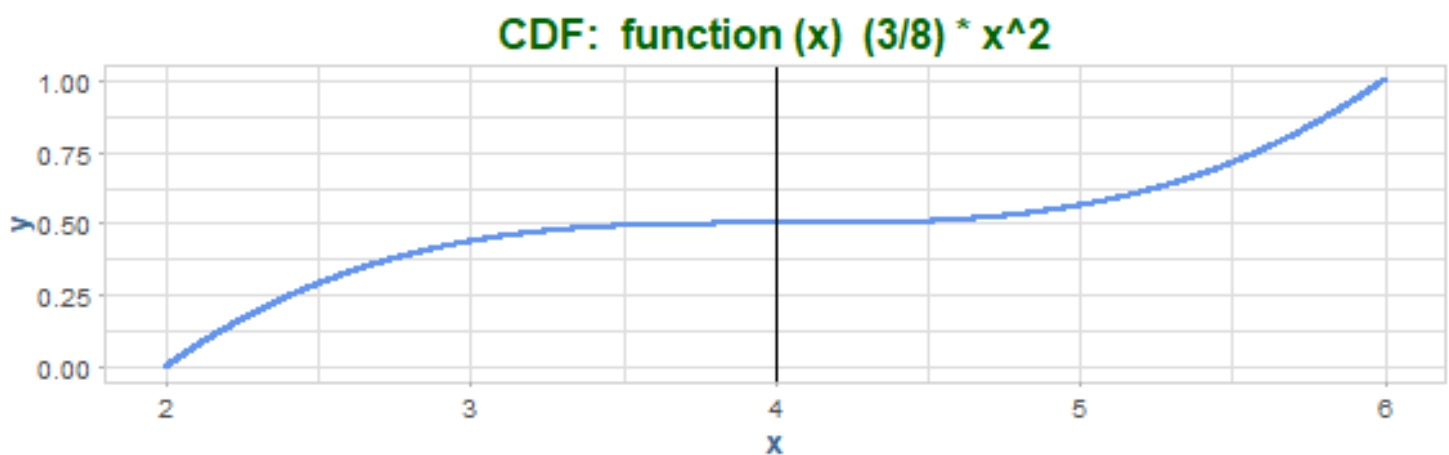
cdf <- function(x) (3/8)*(x - 4)

y <- cumsum(y) / sum(y)

ev <- x[min(which(y > .5))]

ggplot(data.table(x,y)) +
  geom_point(aes(x, y), col = "cornflowerblue", lwd = .8) +
  geom_vline(xintercept = ev) +
  labs(title = paste("CDF: ", paste0(deparse(f), collapse = " ")))

```



```

z <- ( 4.2 - ev ) / sqrt(244)
pnorm(z, lower.tail = F)

```

```
[1] 0.4949177
```

#### 4.14

According to the 2000 census, 28.6% of the US adult population recieved a high school diploma.

In a random sample of 800 US adults, what is the probability that between 220 and 230 (inclusive) people have a high school deploma?

Use the CLT approximation with continuity correction, and compare with the exact probability.

#### Solution:

The sampling distribution of  $\hat{p}$  is approximately normal with:

```

n <- 800
mu <- .286

```

```
ev <- 800 * mu
sigma <- sqrt(n*mu*(1-mu))
```

$$\mathbb{E}[X] = 228.8 \text{ and } \sigma = \sqrt{800(.286)(1 - .286)} = 12.7814$$

```
l <- pnorm((ev - 219.5) / sigma)
h <- pnorm((ev - 230.5) / sigma)

p <- l - h
```

Probability: 0.3195

## 4.15

If  $X_1, \dots, X_n$  are i.i.d. from  $\text{Unif}[0, 1]$ , how large should  $n$  be so that  $P(\bar{X} - \frac{1}{2} < 0.05) \geq 0.90$ , that is, is there at least a 90% chance that the sample mean is within 0.05 of  $\frac{1}{2}$ ? Use the CLT approximation.

## 4.16

Maria claims that she has drawn a random sample of size 30 from the exponential distribution with  $\lambda = 1/10$ .

The mean of her sample is 12.

a.) What is the expected value of a sample mean?

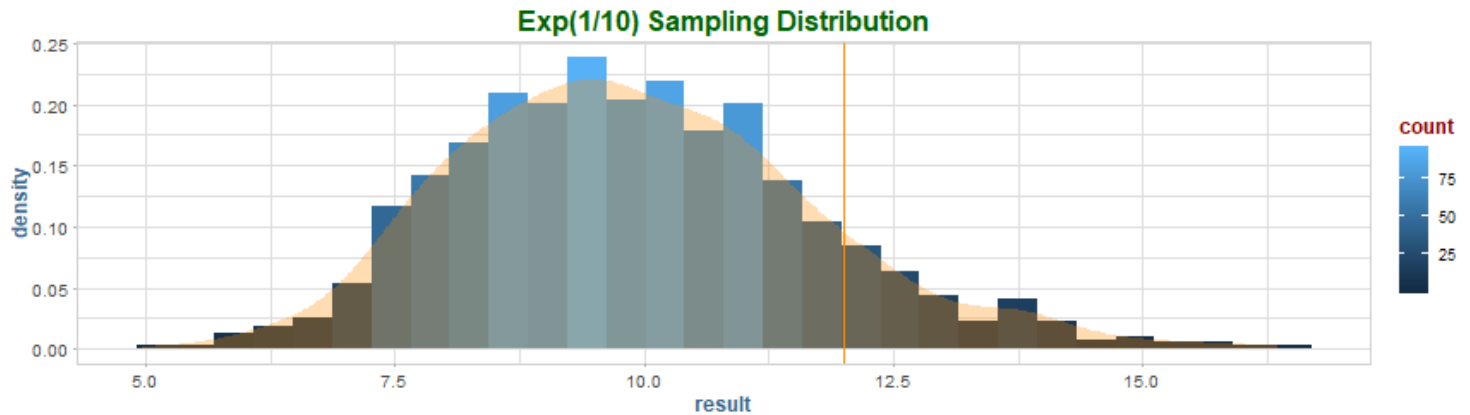
$$X \sim \text{Exp}(\frac{1}{10}), \mathbb{E}(x) = 10$$

b.) Run a simulation by drawing 1000 random samples, each of size 30, from  $\text{Exp}(1/10)$ , and compute the mean for each sample.

```
N <- 1000
result <- numeric(N)

for( i in 1:N)
{
  samp <- rexp( n = 30, rate = 1/10)
  result[i] <- mean(samp)
}

ggplot(data.table(result), aes(result)) +
  geom_histogram(aes(y = ..density.., fill = ..count..), bins = 30) +
  geom_vline(xintercept = 12, col = "darkorange") +
  stat_density( kernel = "gaussian", fill = "darkorange", alpha = .3) +
  labs(title = "Exp(1/10) Sampling Distribution")
```



```
p <- mean(result > 12)
```

What proportion of the sample means is as large or larger than 12? **12.1%**

c.) Is a mean of 12 unusual for a sample of size 30 from  $\text{Exp}(1/10)$ ?

**Yes**, only ~13% of the sample means have a value of 12 or higher.

## 4.17

Let  $X \sim N(15, 3^2)$  and  $Y \sim N(4, 2^2)$  be independent random variables.

a.) What is the exact sampling distribution of  $W = X - 2Y$ ?

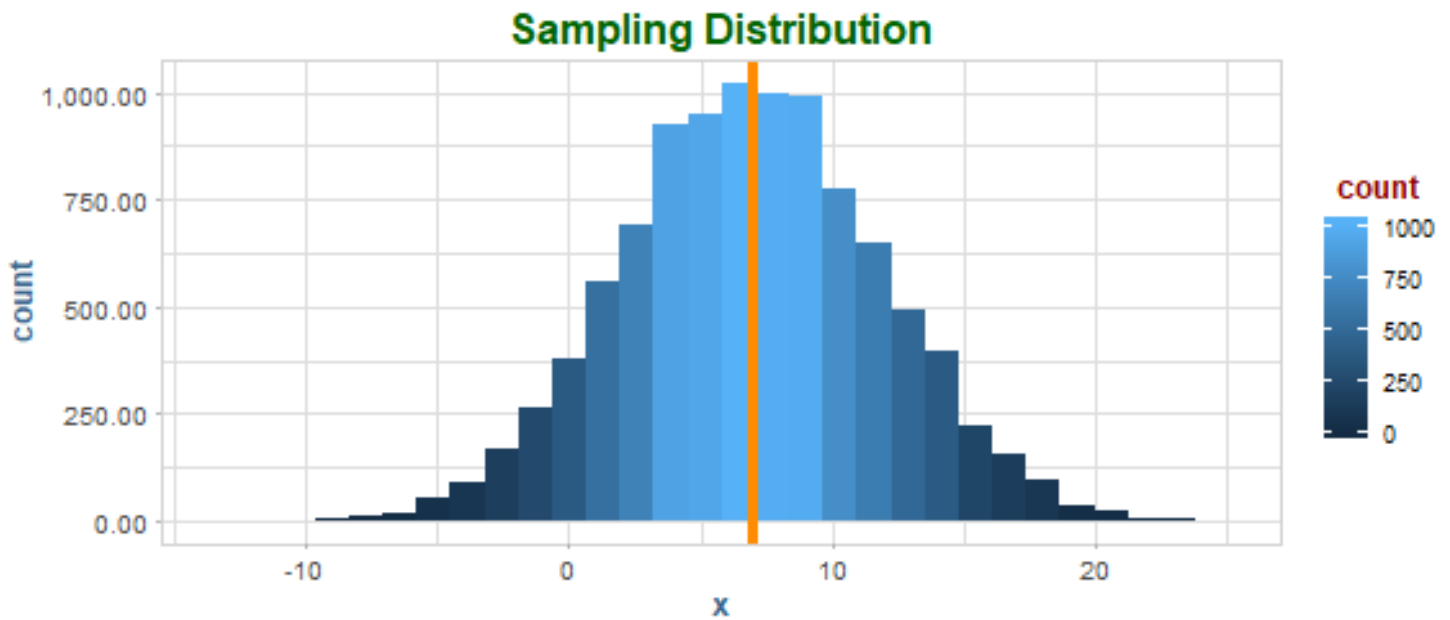
$$W \sim N(7, 5^2)$$

b.) Use R to simulate the sampling distribution of  $W$  and plot your results.

```
X <- rnorm(10e3, 15, 3)
Y <- rnorm(10e3, 4, 2)

W <- X - 2*Y

ggplot(data.table(x = W)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  geom_vline(xintercept = 7, col = "darkorange", lwd = 1.5) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



Check that the simulated mean and standard error are close to the theoretical mean and standard error.

```
mu <- mean(W)
sigma <- sd(W)
```

$$\mu = 6.9422036, \sigma = 4.9366139$$

c.) Use the simulated sampling to estimate  $P(W \leq 10)$ , and then check your estimate with an exact calculation.

```
phat <- mean(W <= 10)
p <- pnorm(10, mean = 7, sd = 5)
```

$$\hat{p} = 73.61\%$$

$$P(W \leq 10) = 72.57\%$$

## 4.18

Let  $X \sim \text{Pois}(4)$ ,  $Y \sim \text{Pois}(12)$ ,  $U \sim \text{Pois}(3)$  be independent random variables.

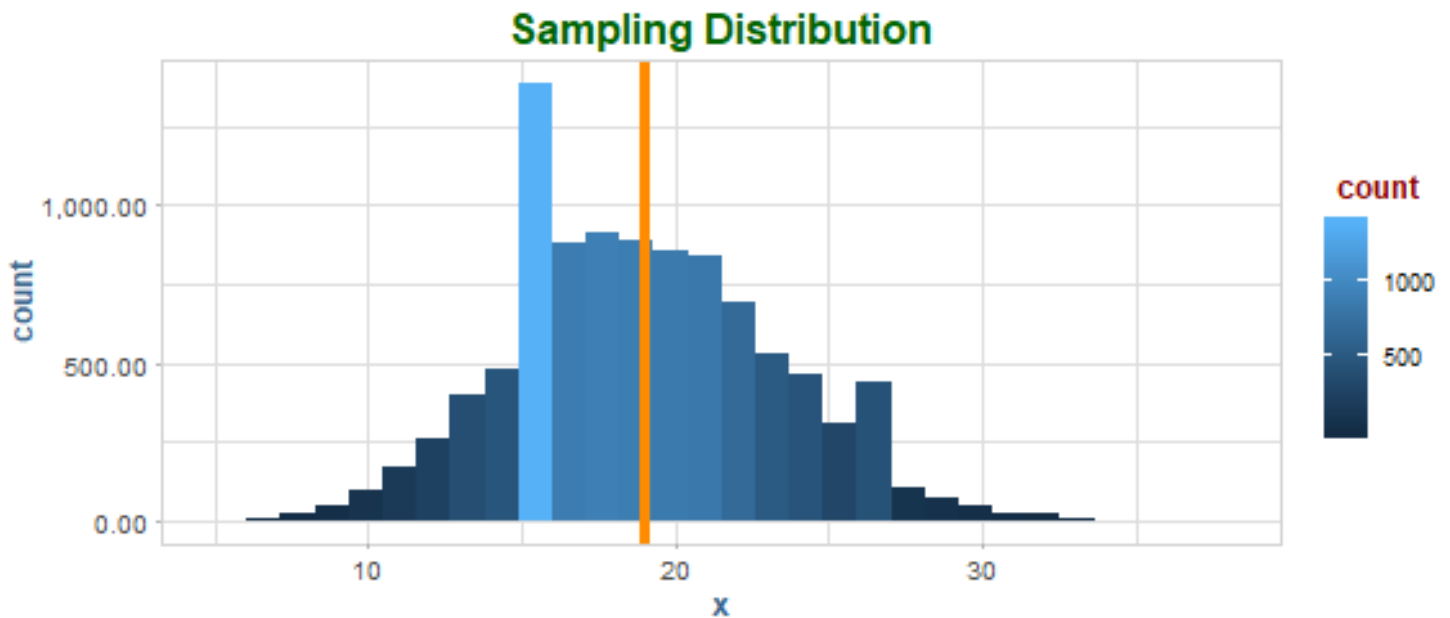
a.) What is the exact sampling distribution of  $W = X + Y + U$ ?

$$W \sim \text{Pois}(19)$$

b.) Use R to simulate the sampling distribution of  $W$  and plot your results.

```
W <- rpois(10e3, lambda = 19)

ggplot(data.table(x = W)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  geom_vline(xintercept = 19, col = "darkorange", lwd = 1.5) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



Check that the simulated mean and standard error are close to the theoretical mean and standard error.

```
mu <- mean(W)
sigma <- sd(W)
```

$$\mu = 19.0046, \sigma = 4.3554651$$

c.) Use the simulated sampling distribution to estimate  $P(W \leq 14)$  and then check your estimate with an exact calculation.

```
phat <- mean(W <= 14)
p <- ppois(14, lambda = 19)
```

$$\hat{p} = 14.99\%$$

$$P(W \leq 14) = 14.97\%$$

## 4.19

Let  $X_1, X_2, \dots, X_{10} \sim^{i.i.d} N(20, 8^2)$  and  $Y_1, Y_2, \dots, Y_{15} \sim^{i.i.d} N(16, 7^2)$ .

Let  $W = \bar{X} + \bar{Y}$

a.) Give the exact sampling distribution of  $W$ .

$$\sigma = (10 + 15) / \sqrt{10 + 15 - 1} = 3.1$$

$$W \sim N(36, 3.1^2)$$

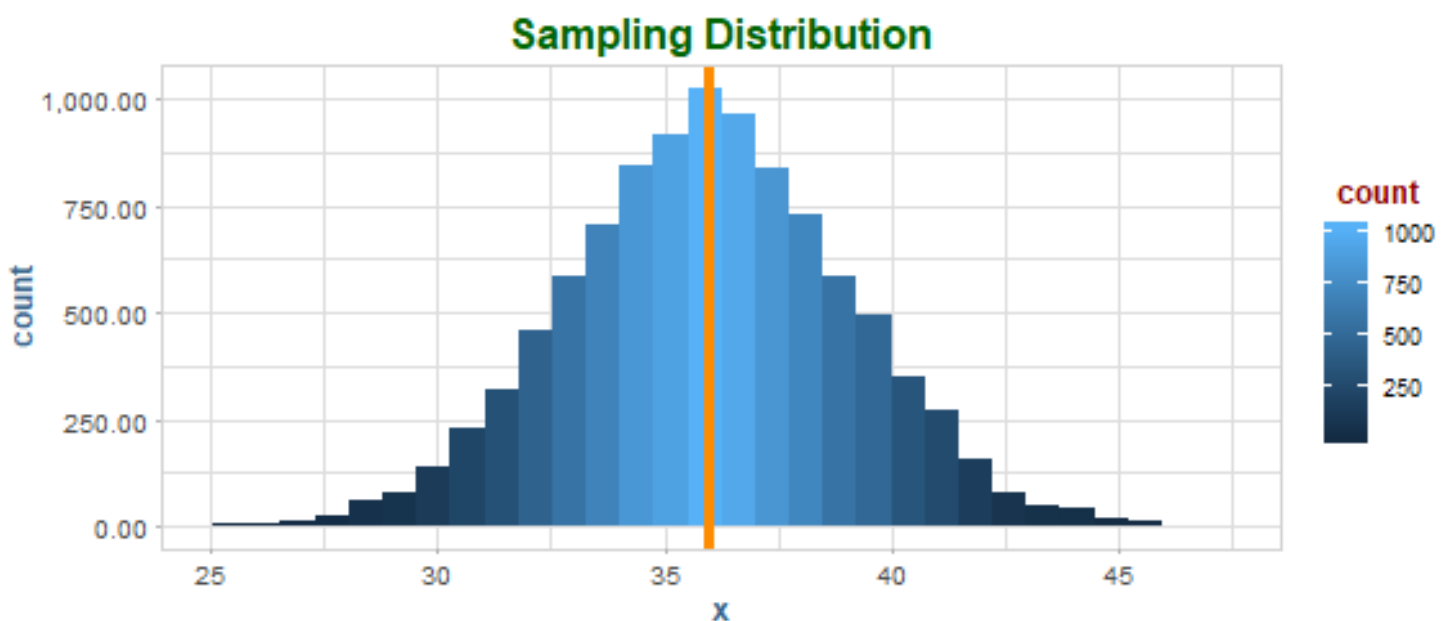
b.) Simulate the sampling distribution in R and plot your results.

```
N <- 10e3
result <- numeric(N)

for( i in 1:N)
{
  X <- rnorm(10, 20, 8)
  Y <- rnorm(15, 16, 7)

  result[i] <- mean(X) + mean(Y)
}

ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  geom_vline(xintercept = 36, col = "darkorange", lwd = 1.5) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



Check that the simulated mean and standard error are close to the exact mean and standard error.

```
mu <- mean(result)
sigma <- sd(result)
```

$$\mu = 35.9841727, \sigma = 3.1019113$$

c.) Use your simulation to find  $P(W < 40)$ . Calculate an exact answer and compare.

```
phat <- mean(result <= 40)

p <- pnorm(40, 36, 3)
```

$$\hat{p} = 90.08\%$$

$$P(W < 40) = 90.88\%$$

## 4.20

Let  $X_1, X_2, \dots, X_9 \sim^{i.i.d.} N(7, 3^2)$ , and  $Y_1, Y_2, \dots, Y_{12} \sim^{i.i.d.} N(10, 5^2)$ .

Let  $W = \bar{X} - \bar{Y}$ .

a.) Give the sampling distribution of  $W$ .

$$\sigma = (3 + 5)/\sqrt{9 + 12 - 1} = 1.79$$

$$W = N(-3, 1.79^2)$$

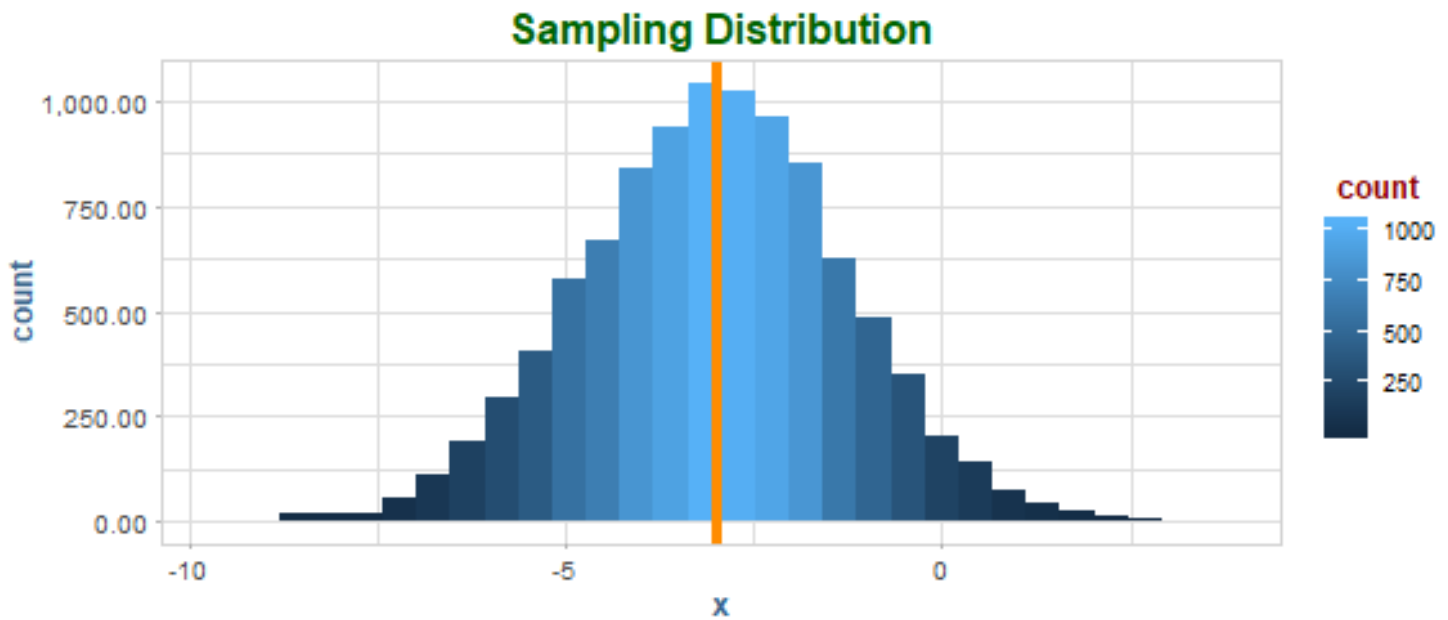
b.) Simulate the sampling distribution of  $W$  in R, and plot your results.

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  X <- rnorm(9, 7, 3)
  Y <- rnorm(12, 10, 5)
  result[i] <- mean(X) - mean(Y)
}

ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  geom_vline(xintercept = -3, col = "darkorange", lwd = 1.5) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```





Check that the simulated mean and standard error are close to the theoretical mean and standard error.

```
mu <- mean(result)
sigma <- sd(result)
```

$$\mu = -3.0310171, \sigma = 1.7456471$$

c.) Use your simulation to find  $P(W < -1.5)$ .

```
phat <- mean(result <= -1.5)
p <- pnorm(-1.5, -3, 1.79)
```

$$\hat{p} = 81.32\%$$

Calculate an exact answer and compare.

$$P(W < 1.5) = 79.9\%$$

## 4.21

Let  $X_1, X_2, \dots, X_N$  be a random sample from  $N(0, 1)$ . Let  $W = X_1^2 + X_2^2 + \dots + X_n^2$

What is the mean and variance of the sampling distribution of  $W$ ?

$$\mu = 0, \sigma = 1$$

Repeat using  $N = 4, N = 5$ .

$$N = 4, \sigma = 4/\sqrt{4-1} = 2.3$$

$$N = 5, \sigma = 5/\sqrt{5-1} = 2.5$$

What observations or conjectures do you have for general  $n$ ?

## 4.22

Let  $X$  be a uniform random variable on the interval  $[40, 60]$  and  $Y$  be a uniform random variable on  $[45, 80]$ .

Assume that  $X$  and  $Y$  are independent.

a.) Compute the expected value and variance of  $X + Y$ .

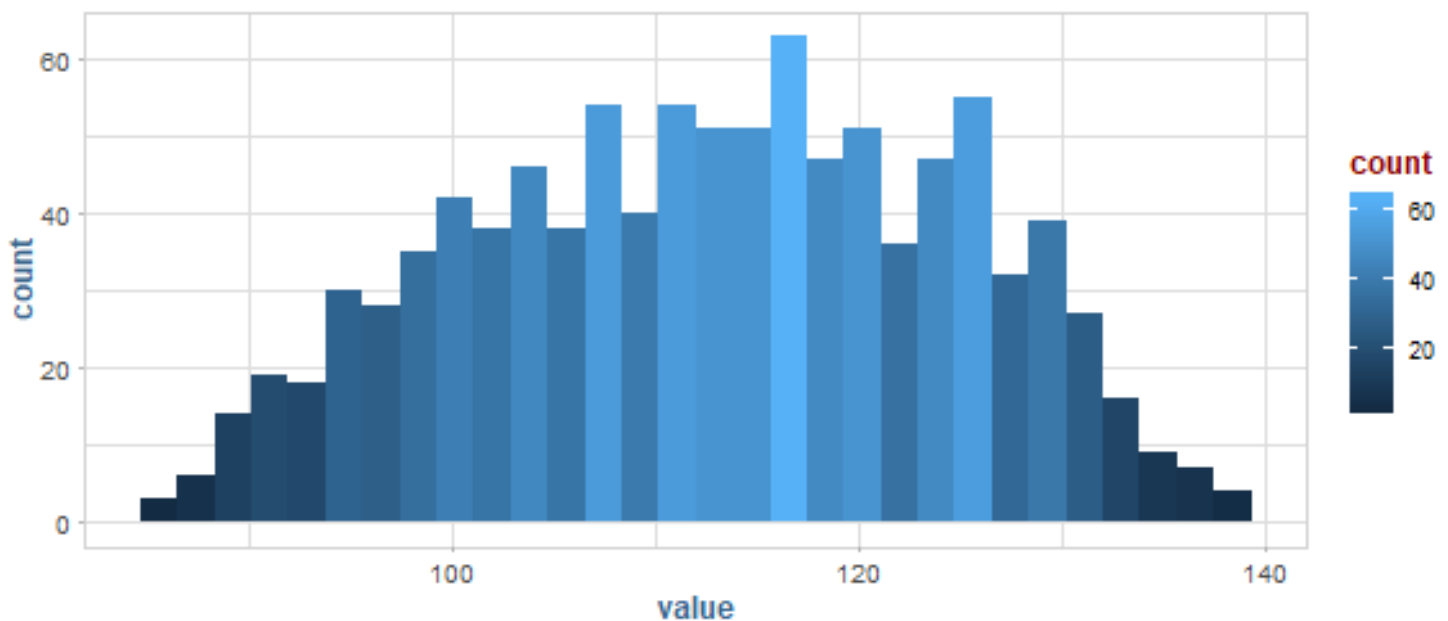
$$\mu = 112.5, \text{Var} = 1/24 * (140 - 85)^2 = 126.04$$

b.) Simulate a sampling distribution of  $X + Y$ .

```
X <- runif(1000, 40, 60)
Y <- runif(1000, 45, 80)

total <- X + Y

ggplot(data.table(value = total)) +
  geom_histogram(aes(value, fill = ..count..), bins = 30) +
  labs("X + Y Sampling Distribution")
```



Describe the sampling distribution of  $X + Y$ . **Approximately Normal**

Compute the mean and variance of the sampling distribution and compare this with the theoretical mean and variance.

```
mu <- mean(total)
var <- var(total)
```

$$\mu = 112.732947, Var = 140.5295884$$

c.) Suppose the time (in minutes) Andy takes to complete his statistics homework is  $Unif[40, 60]$  and the time Adam takes is  $Unif[45, 80]$ .

Assume they work independently.

One day they announce that their total time to finish an assignment was less than 90 minutes.

How likely is this?

```
p <- punif(90, 85, 140)
```

Probability: **9.09%**

## 4.23

Let  $X_1, X_2, \dots, X_{20} \sim^{i.i.d.} Exp(2)$ . Let  $X = \sum_{i=1}^{20} X_i$ .

a.) Simulate the sampling distribution of  $X$  in R.

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  samp <- rexp(n = 20, rate = 2)
  result[i] <- sum(samp)
}
```

b.) From your simulation, find  $\mathbb{E}[X]$  and  $Var[X]$ .

```
mu <- mean(result)
var <- var(result)
```

$$\mu = 10.0187404, Var = 5.0626136$$

c.) From your simulation, find  $P(X \leq 10)$ .

```
phat <- mean(result <= 10)
```

Probability: **52.95%**

## 4.24

Let  $X_1, X_2, \dots, X_{30} \sim^{i.i.d.} \text{Exp}(1/3)$  and let  $\bar{X}$  denote the sample mean.

a.) Simulate the sampling distribution of  $\bar{X}$  in R.

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  samp <- rexp(n = 30, rate = 1/3)
  result[i] <- mean(samp)
}
```

b.) Find the mean and standard error of the sampling distribution, and compare with the theoretical results.

```
mu <- mean(result)
sigma <- sd(result)
```

Sample:  $\mu = 3.010494$ ,  $\sigma = 0.5523455$

Theoretical:  $\mu = \frac{\lambda}{1} = 3$ ,  $\sigma = 3/\sqrt{30 - 1} = .56$

c.) From your simulation, find  $P(\bar{X} \leq 3.5)$ .

```
phat <- mean(result <= 3.5)
```

$\hat{p} = 0.8188$

d.) Estimate  $P(\bar{X} \leq 3.5)$  by assuming the CLT approximation holds.

Compare this result with the one in part (c).

```
z <- (3.5 - 3) / sigma
p <- pnorm(z)
```

$p = 0.8173$

## 4.25

Consider the exponential distribution with density  $f(x) \frac{1}{20} e^{-x/20}$ , with mean and standard deviation of 20.

a.) Calculate the median of this distribution.

$$0.5 = \int_m^{\infty} f(x) dx = -e^{-x/20}$$

... =  $M = 20 * \log(2)$

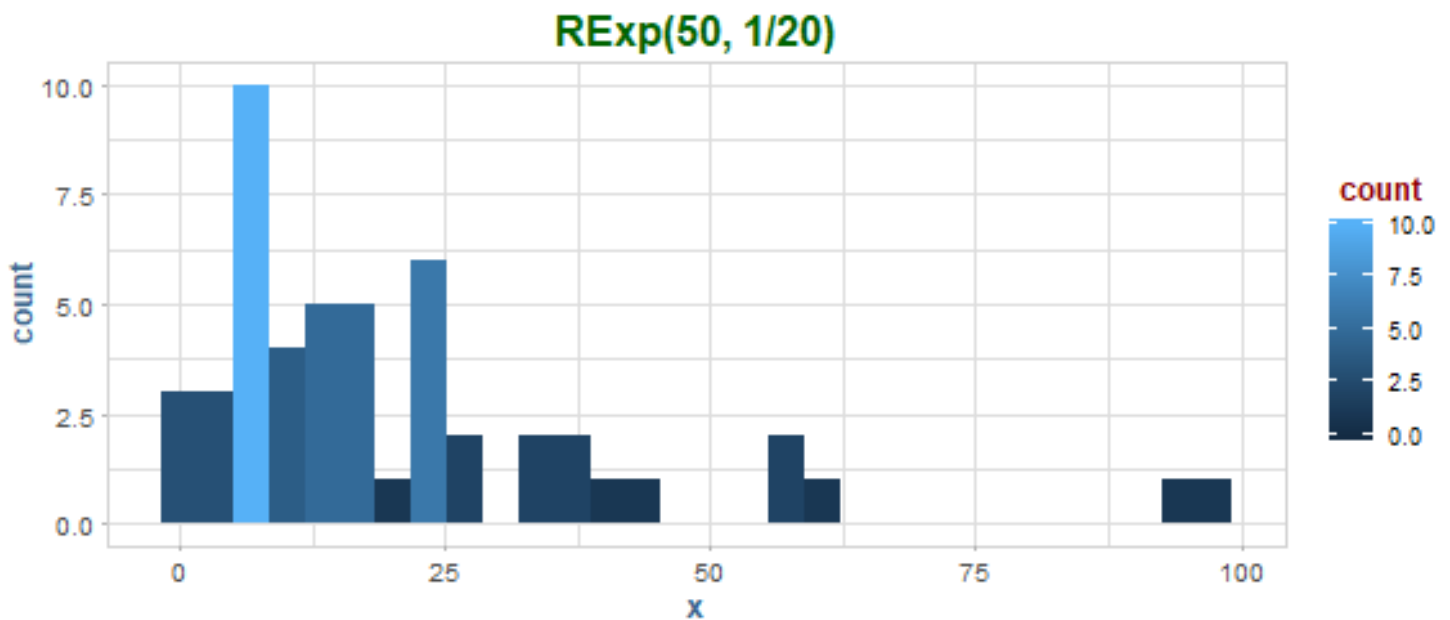
... = 13.8629436

b.) Using R, draw a random sample of size 50 and graph the histogram.

```
samp <- rexp( n = 50, rate = 1/20)

ggplot(data.table(x = samp)) +
  geom_histogram(aes(x, fill = ..count..)) +
  labs(title = "RExp(50, 1/20)")
```

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



What are the mean and standard deviation of your sample?

```
mu <- mean(samp)
sigma <- sd(samp)
```

$\mu = 21.5535161, \sigma = 21.3461975$

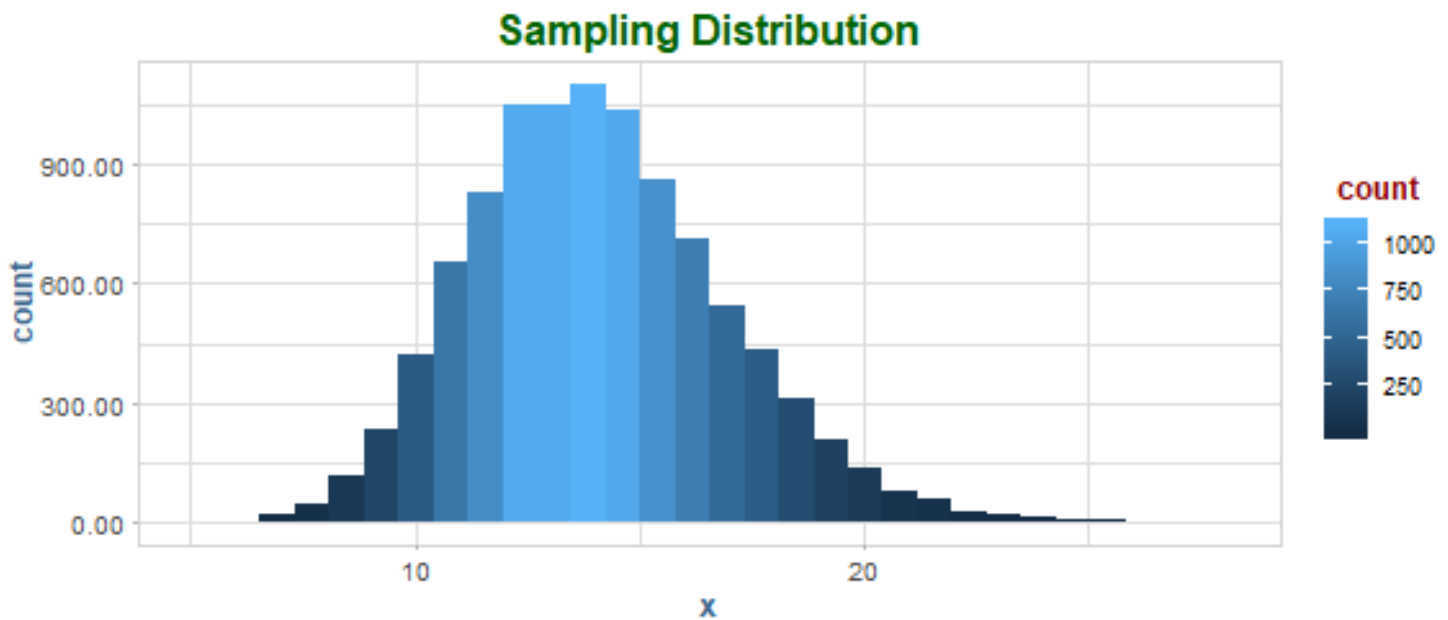
c.) Run a simulation to find the (approximate) sampling distribution for the median of sample size 50 from the exponential distribution and describe it.

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
```

```
{
  samp <- rexp( n = 50, rate = 1/20)
  result[i] <- median(samp)
}

ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



What is the mean and the standard error of this sampling distribution?

```
mu <- mean(result)
sigma <- sd(result) / sqrt(50)
```

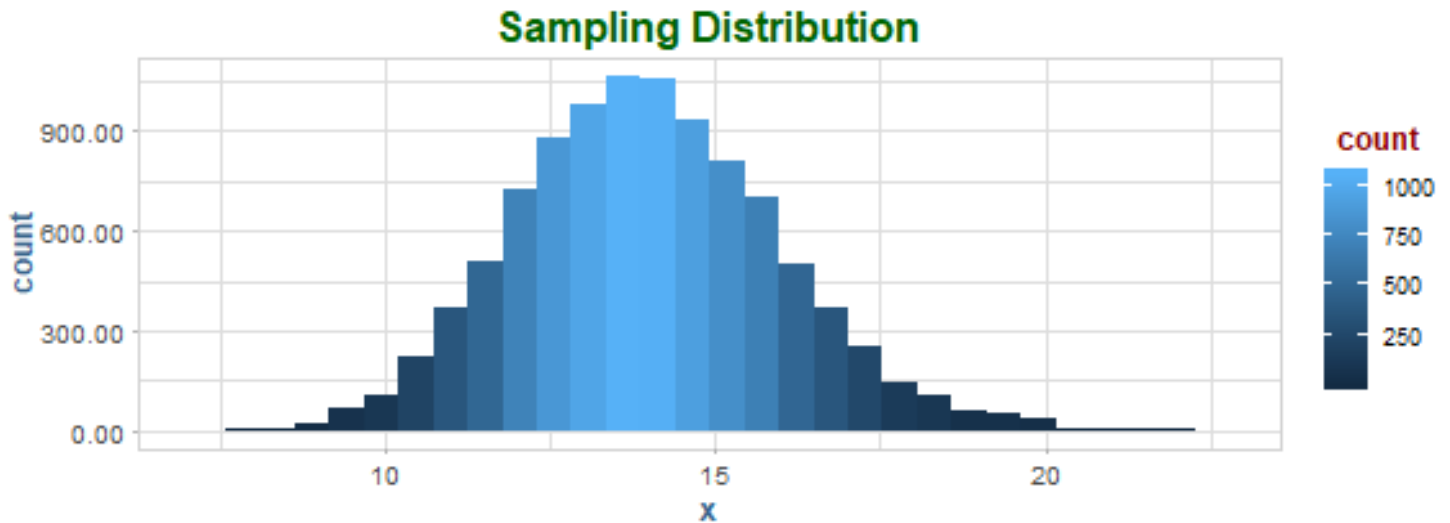
$\mu = 14.0874039, \sigma = 0.4016384$

d.) Repeat the above but use sample sizes  $n = 100, 500$  and  $1,000$ .

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  samp <- rexp( n = 100, rate = 1/20)
  result[i] <- median(samp)
}
```

```
ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



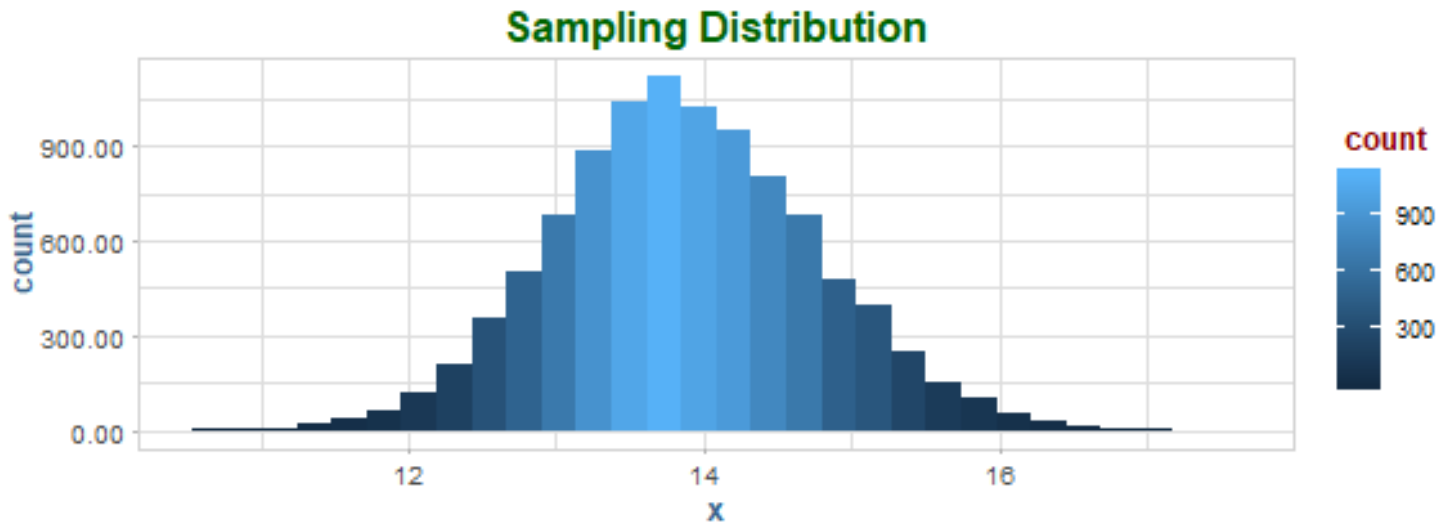
```
mu <- mean(result)
sigma <- sd(result) / sqrt(50)
```

$$\mu = 13.9680289, \sigma = 0.2806502$$

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  samp <- rexp( n = 500, rate = 1/20)
  result[i] <- median(samp)
}

ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```



```
mu <- mean(result)
sigma <- sd(result) / sqrt(50)
```

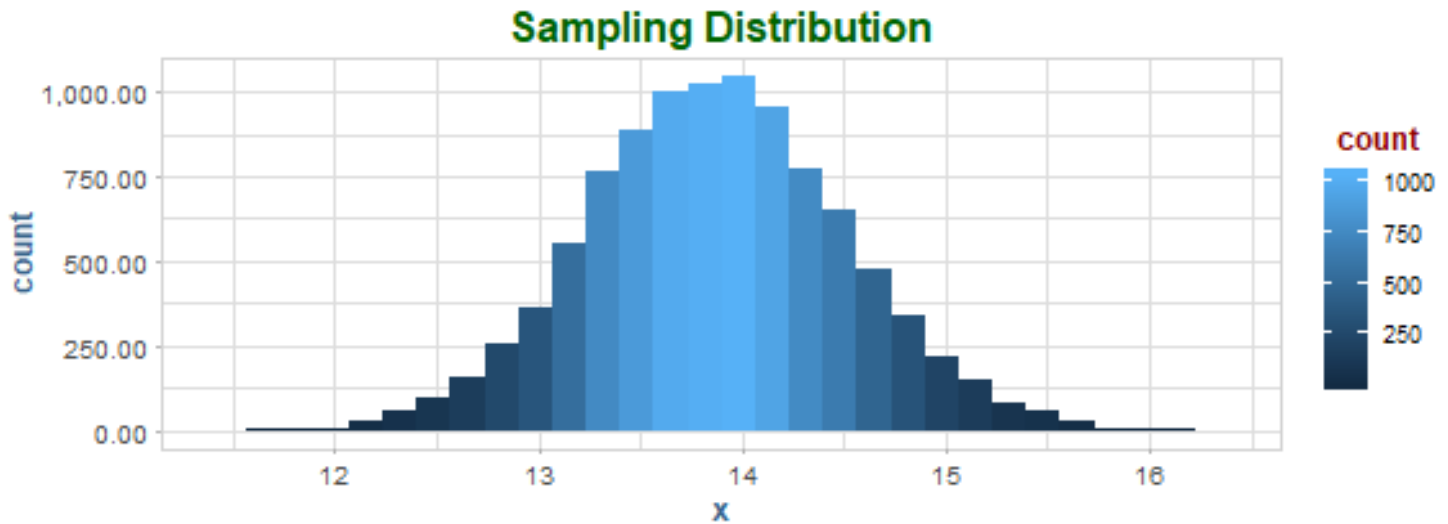
$$\mu = 13.8727789, \sigma = 0.125773$$

```
N <- 10e3
result <- numeric(N)

for(i in 1:N)
{
  samp <- rexp( n = 1000, rate = 1/20)
  result[i] <- median(samp)
}

ggplot(data.table(x = result)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  scale_y_continuous(labels = comma) +
  labs(title = "Sampling Distribution")
```





```
mu <- mean(result)
sigma <- sd(result) / sqrt(50)
```

$$\mu = 13.8752542, \sigma = 0.0897888$$

How does sample size affect the sampling distribution?

*The sample mean and standard error converge to the analytical solution with increased sample size.*

## 4.26

Prove theorem 4.2.1.

## 4.27

Let  $X_1, X_2 \sim^{i.i.d.} F$  with corresponding pdf  $f(x) = \frac{2}{x^2}, 1 \leq x \leq 2$ .

a.) Find the pdf of  $X_{max}$ .

$$F_{max}(x) = 8\left(\frac{1}{x^2} - \frac{1}{x^3}\right)$$

b.) Find the expected value of  $X_{max}$ .

$$\mathbb{E}[F_{max}] = 1.545$$

## 4.28

Let  $X_1, X_2, \dots, X_N \sim^{i.i.d.}$  with corresponding pdf  $f(x) = 3x^2, 0 \leq x \leq 1$ .

a.) Find the pdf for  $X_{min}$ .

b.) Find the pdf for  $X_{max}$ .

c.) If  $n = 10$ , find the probability that the largest value,  $X_{max}$ , is greater than 0.92.

**4.29**

Compute the pdf of the sampling distribution of the maximum samples of size 10 from a population with an exponential distribution with  $\lambda = 12$ .

**4.30**

Let  $X_1, X_2, \dots, X_N \sim^{i.i.d.} \text{Exp}(\lambda)$  with pdf  $f(x) = \lambda e^{-\lambda x}$ ,  $\lambda > 0, x > 0$ .

- Find the pdf  $f_{\min}(x)$  for the sample minimum  $X_{\min}$ . Recognize this as the pdf of a known distribution.
- Simulate in R the sampling distribution of  $X_{\min}$  of samples of size  $n = 25$  from the exponential distribution with  $\lambda = 7$ . Compare the theoretical expected value of  $X_{\min}$  with the simulated expected value.

**4.31**

Let  $X_1, X_2, \dots, X_n \sim^{i.i.d.} \text{Pois}(3)$ . Let  $X = \sum_{i=1}^n X_i$ .

Find the pdf for the sampling distribution of  $X$ .

**4.32**

Let  $X_1$  and  $X_2$  be independent exponential random variables, both with parameter  $\lambda > 0$ .

Find the cumulative distribution function for the sampling distribution of  $X = X_1 + X_2$ .

**4.33**

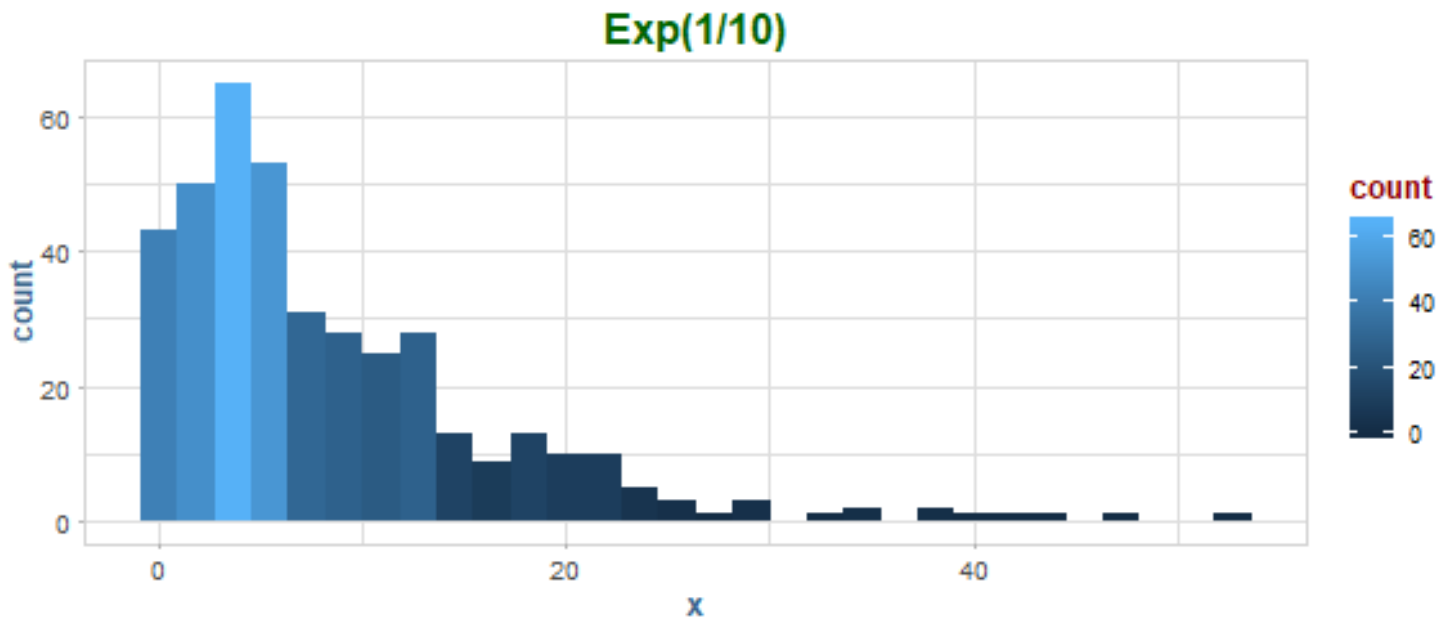
This simulation illustrates the CLT for a finite population.

```
N <- 400
n <- 5

finpop <- rexp(N, 1/10)

ggplot(data.table(x = finpop)) +
  geom_histogram(aes(x, fill = ..count..)) +
  labs(title = "Exp(1/10)")
```

``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



```
mean(finpop) #mean (mu) of your pop.
```

```
[1] 8.642582
```

```
sd(finpop) # stdev (sigma) of your pop.
```

```
[1] 8.266114
```

```
sd(finpop)/sqrt(n) # theoretical standard error of sampling distribution
```

```
[1] 3.696718
```

```
sd(finpop)/sqrt(n) * sqrt((N-n)/(N-1)) # without replacement
```

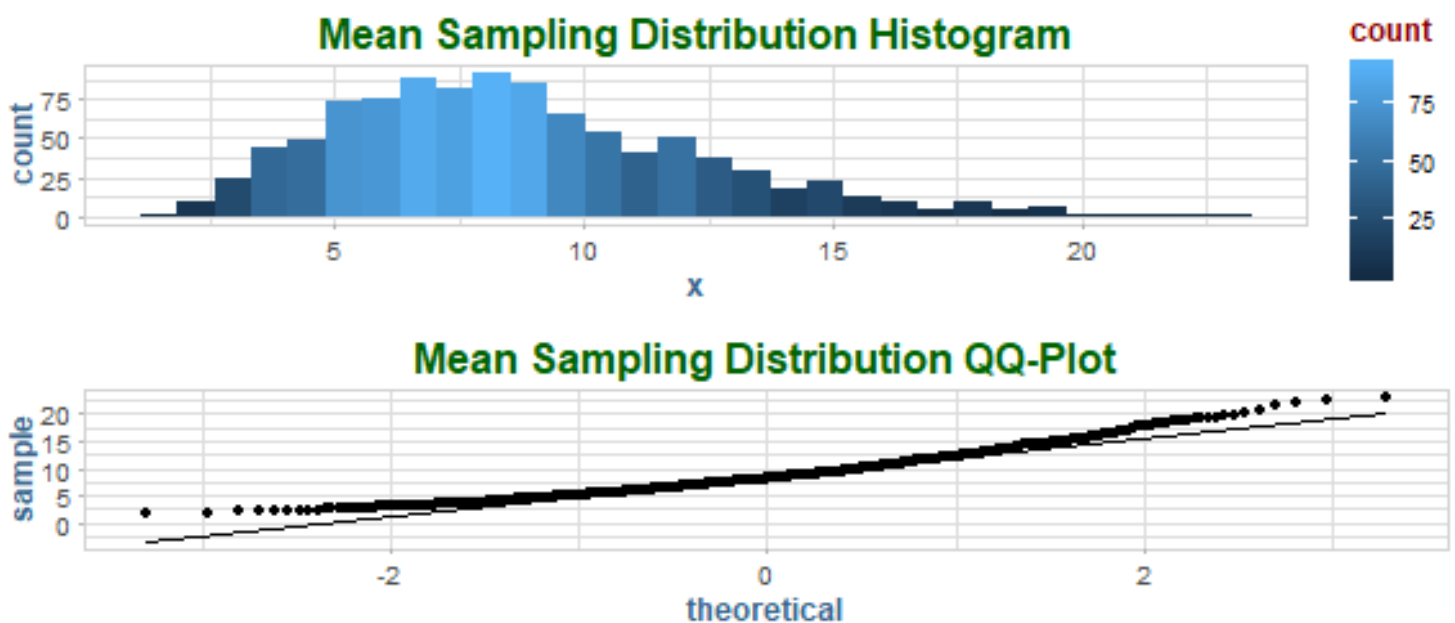
```
[1] 3.678142
```

```
Xbar <- numeric(1000)
for(i in 1:1000)
{
  x <- sample(finpop, n) # Random sample of size n
                           # (w/o replacement)
  Xbar[i] <- mean(x)
}
```

```
p1 <- ggplot(data.table(x = Xbar)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  labs(title = "Mean Sampling Distribution Histogram")

p2 <- ggplot(data.table(x = Xbar), aes(sample = x)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Mean Sampling Distribution QQ-Plot")

grid.arrange(p1, p2, nrow = 2)
```



```
mean(Xbar)
```

```
[1] 8.668039
```

```
sd(Xbar) # estimated standard error of sampling distribution
```

```
[1] 3.646921
```

- Does the sampling distribution of sample means appear approximately normal?
- Compare the mean and standard error of your simulated sampling distribution with the theoretical ones.
- Calculate  $(\sigma/\sqrt{n})(\sqrt{(N-n)/(N-1)})$ , where  $\sigma$  is the standard deviation of the finite population and compare with the (estimated) standard error of the sampling distribution.
- Repeat for larger  $n$ , say  $n = 20$  and  $n = 100$ .

### 4.34

Let  $X_1, X_2, \dots, X_n$  be independent random variables from  $N(\mu, \sigma)$ .

We are interested in the sampling distribution of the variance.

Run a simulation to draw random samples of size 20 from  $N(25, 7^2)$  and calculate the variance for each sample.

```
W <- numeric(1000)

for(i in 1:1000)
{
  x <- rnorm(20, 25, 7)
  W[i] <- var(x)
}

mean(W)
```

```
[1] 48.80045
```

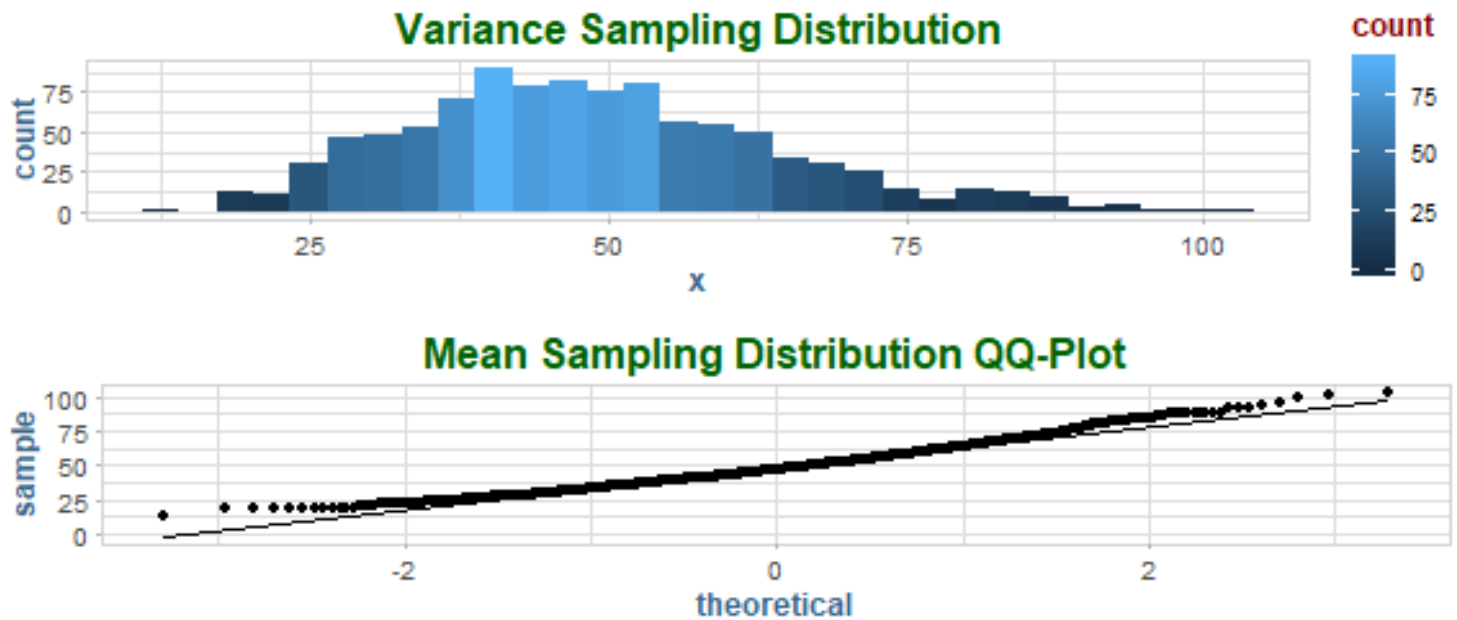
```
var(W)
```

```
[1] 240.5399
```

```
p1 <- ggplot(data.table(x = W)) +
  geom_histogram(aes(x, fill = ..count..), bins = 30) +
  labs(title = "Variance Sampling Distribution")

p2 <-
  ggplot(data.table(value = W), aes(sample = value)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Mean Sampling Distribution QQ-Plot")

grid.arrange(p1, p2, nrow = 2)
```



Does the sampling distribution appear to be normally distributed?

Repeat with  $n = 50$  and  $n = 200$ .

#### 4.35

A random sample of size  $n = 100$  is drawn from a distribution with pdf  $f(x) = 3(1 - x)^2, 0 \leq x \leq 1$ .

- Use the CLT approximation to estimate  $P(\bar{X} \leq 0.27)$ .
- Use the expanded CLT to estimate the same probability (dnorm).
- If  $X_1, X_2, X_3 \sim^{i.i.d.} Unif[0, 1]$ , then the minimum has density  $\mathbf{f}$  given above.

Use simulation to estimate the probability.