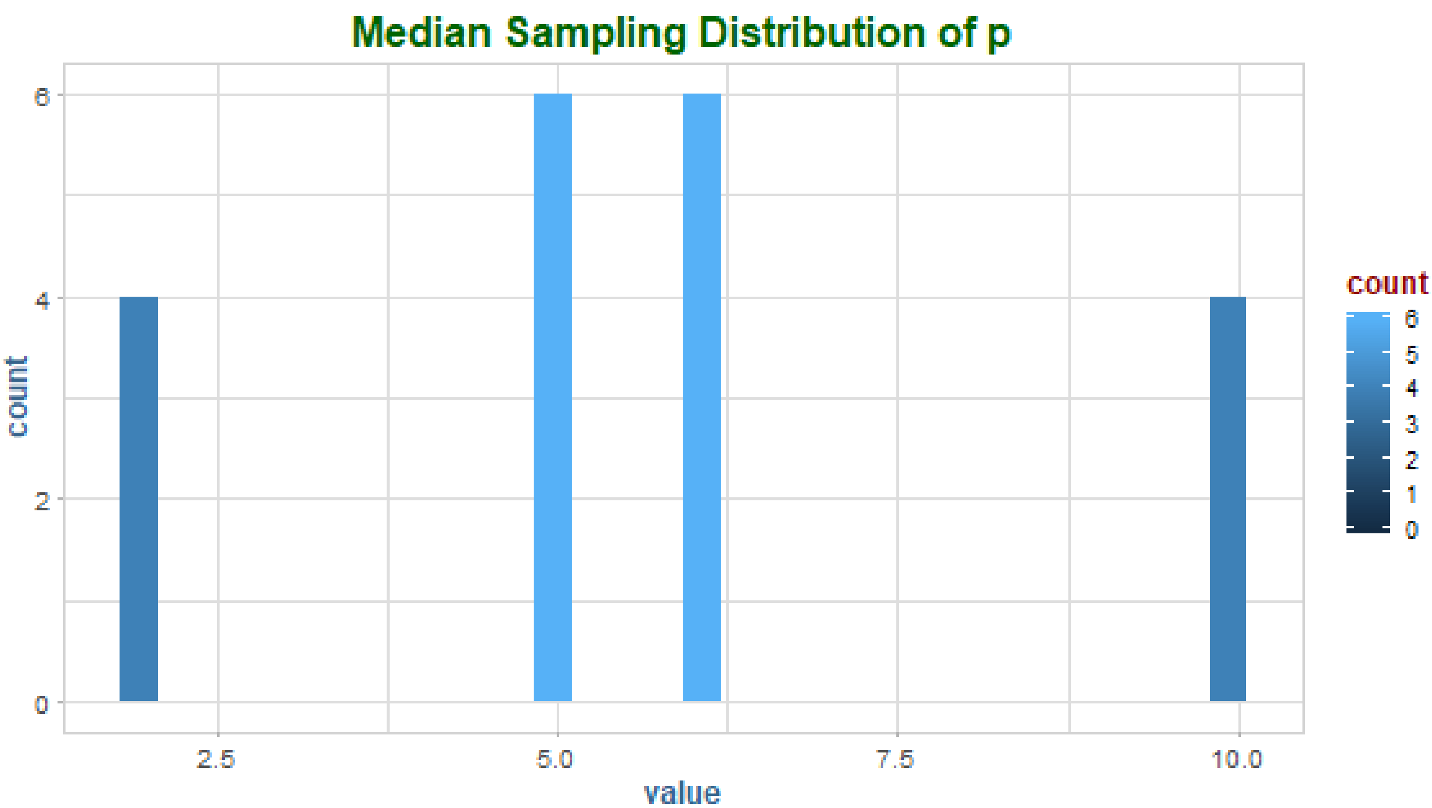
4.1

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.

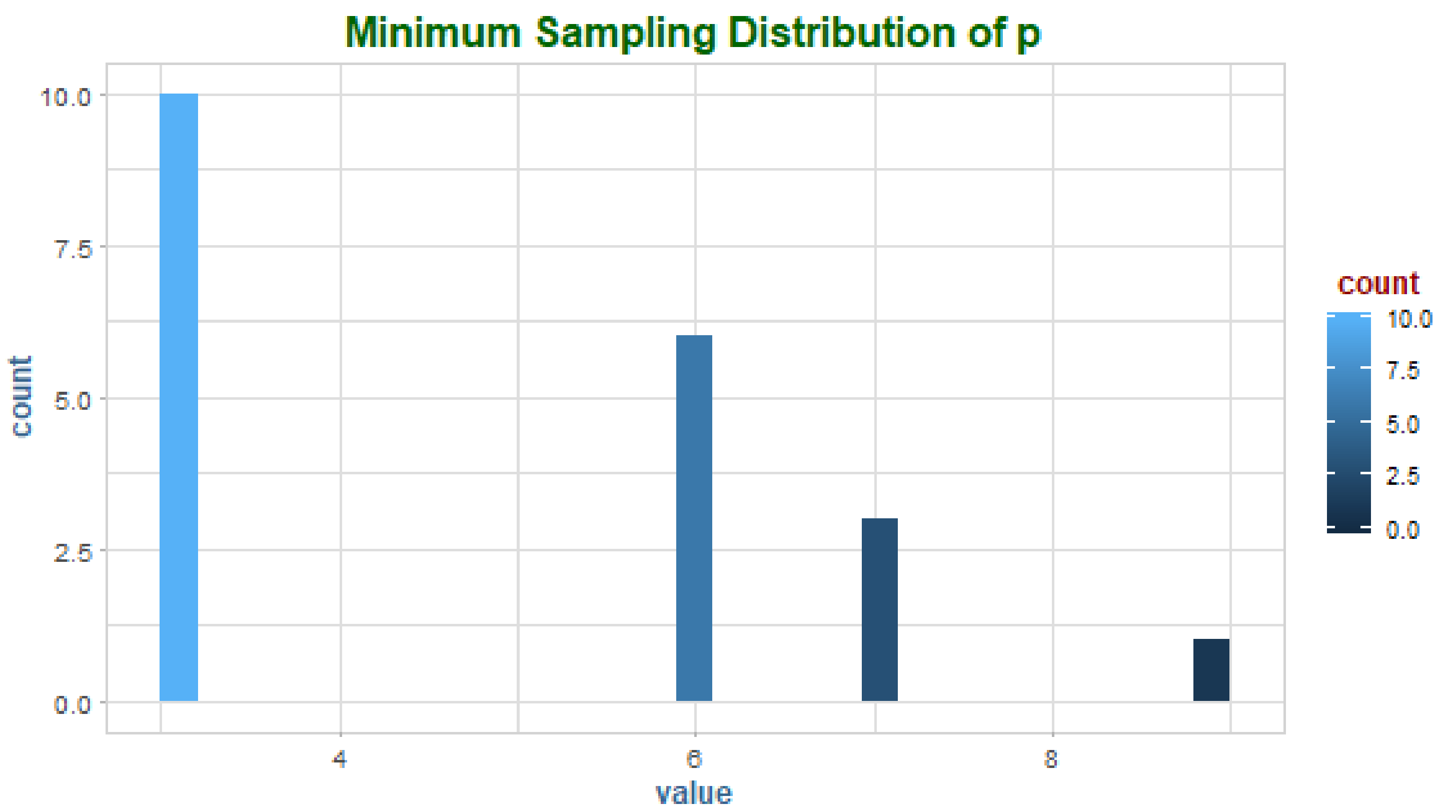
Median of p = **5.5**. Mean of Medians of p = **5.7**

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)
```

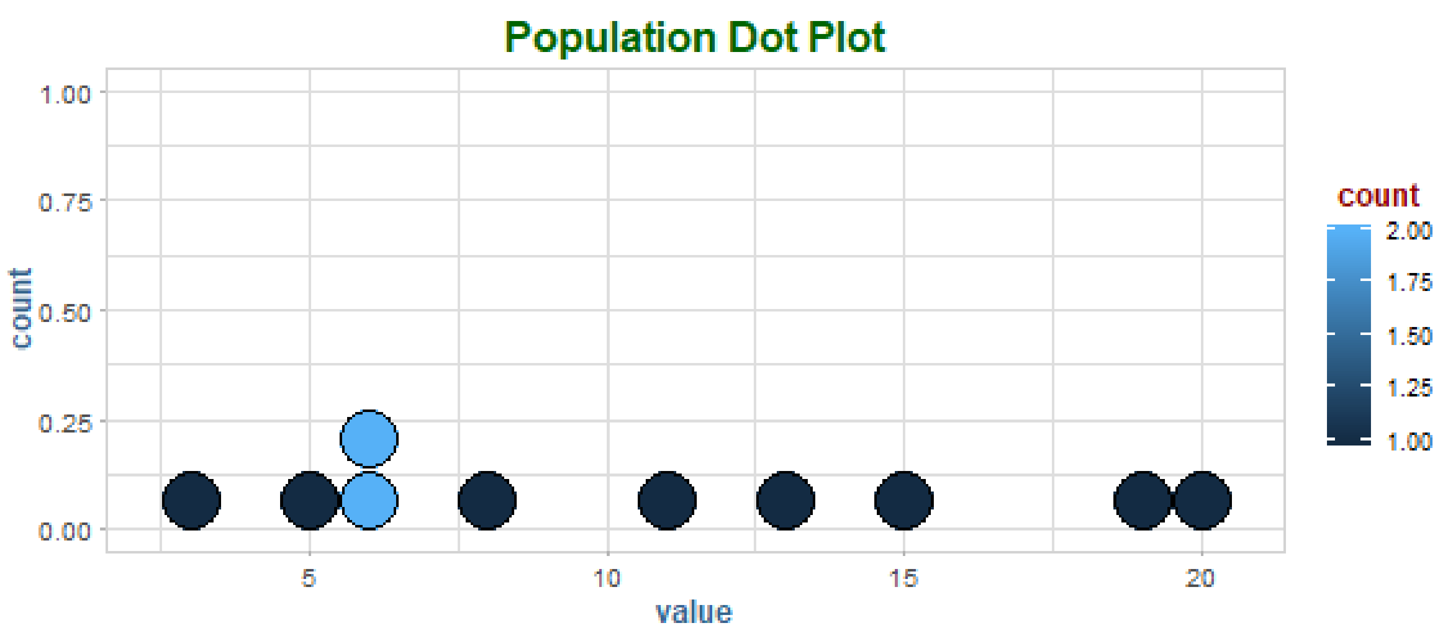
What is the probability that the sum is 13 or larger? **16.6666667%**

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")
```

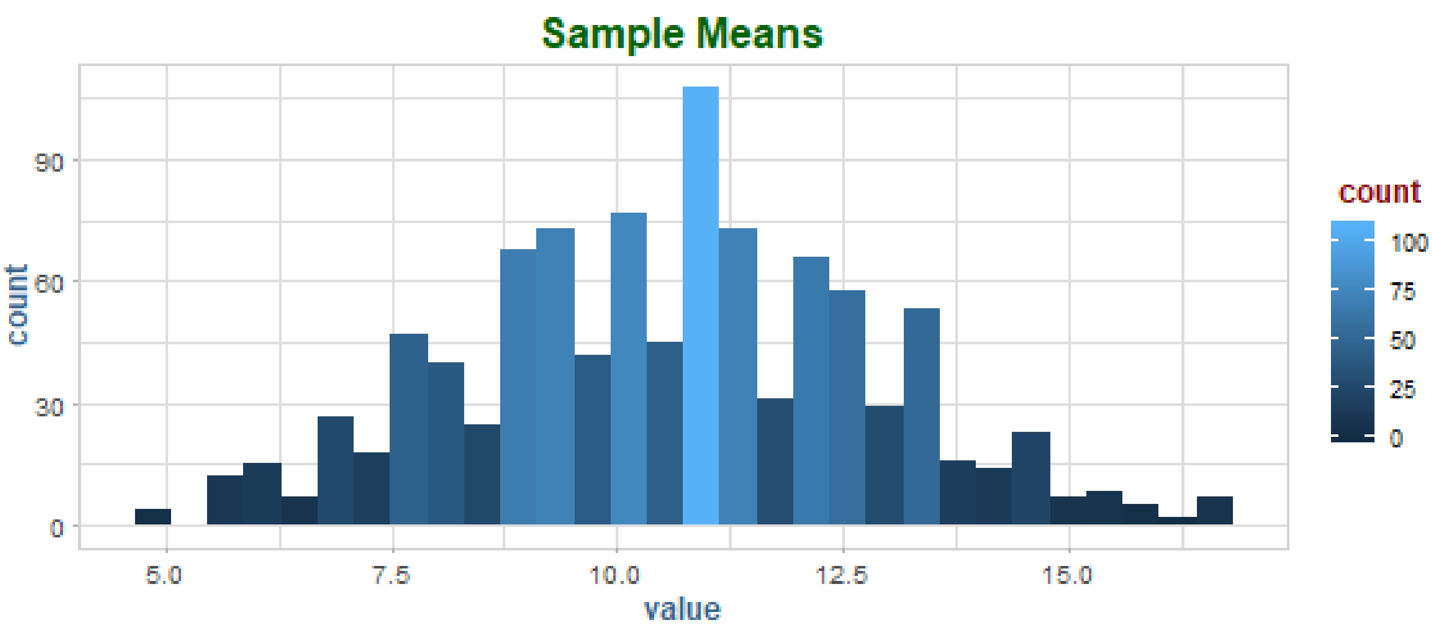


$\mu = 10.6, \sigma = 5.9851668$

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] )
}
```

```
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: 10.57325, standard error: 0.0715197

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

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

$P(\bar{X} < 11) = 55.2\%$

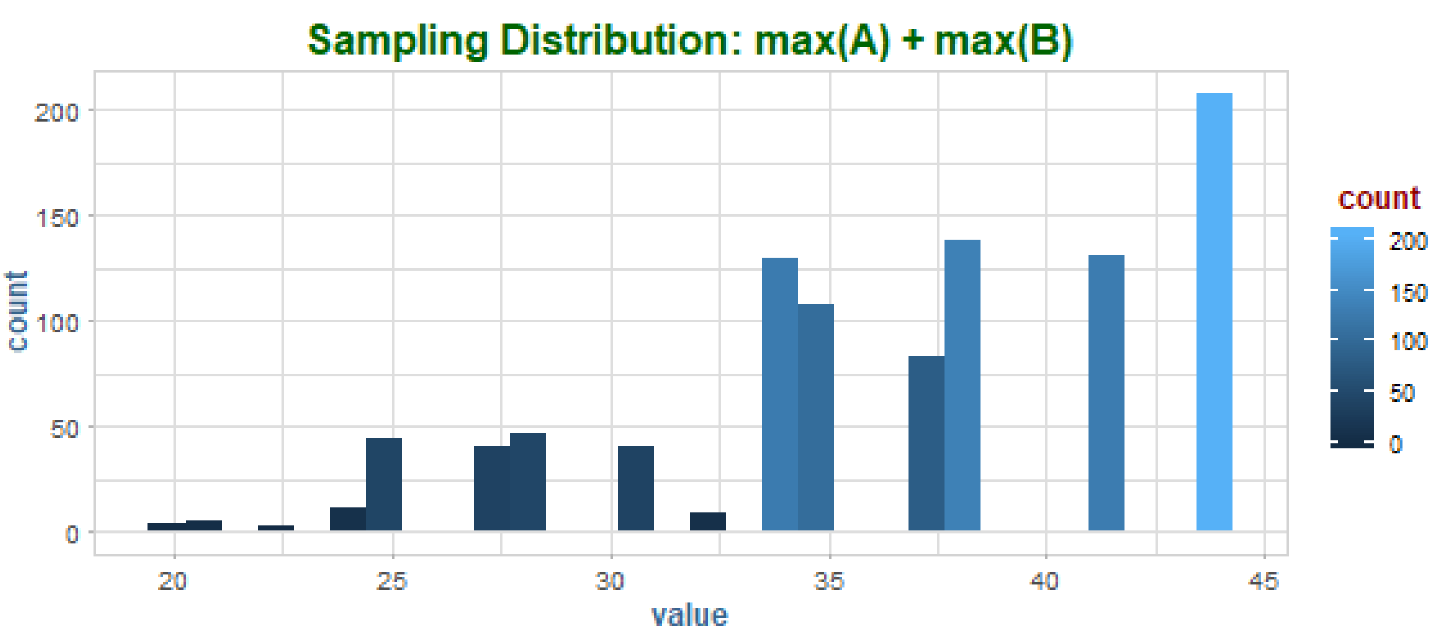
4.5

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.

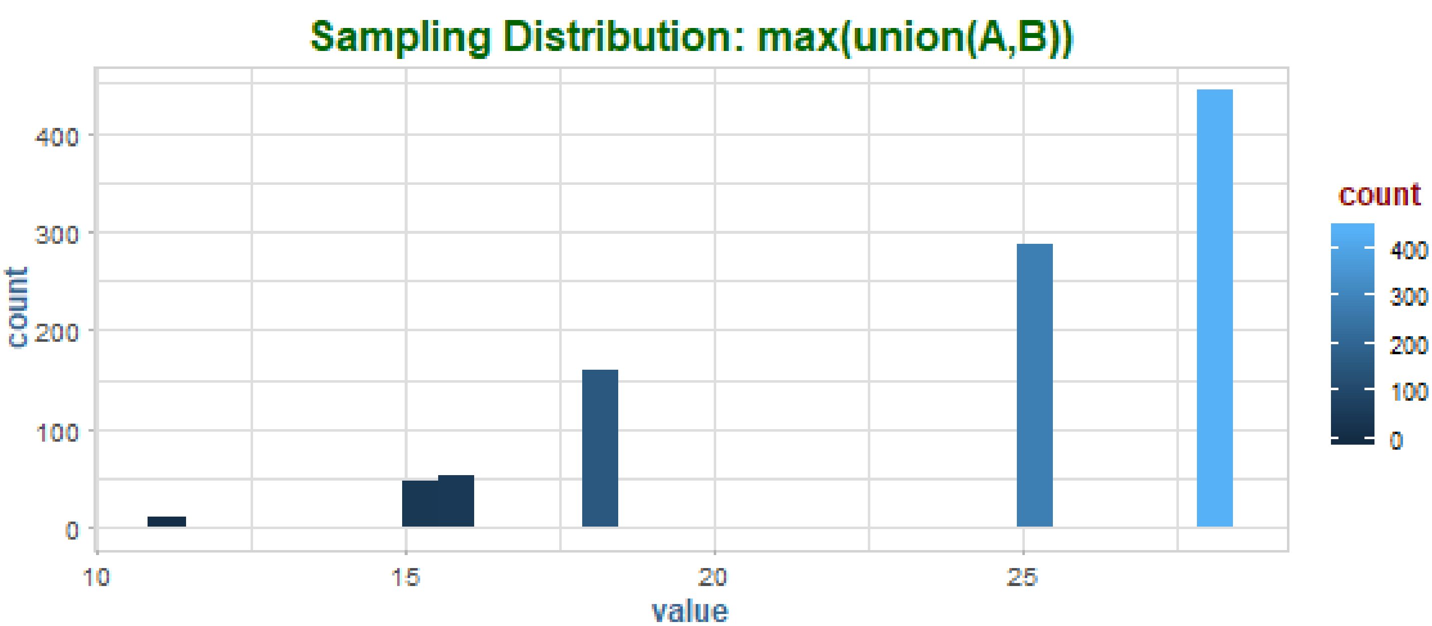
```
prob <- mean(results < 20)
```

Probability: 0%

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: 26.9%

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.

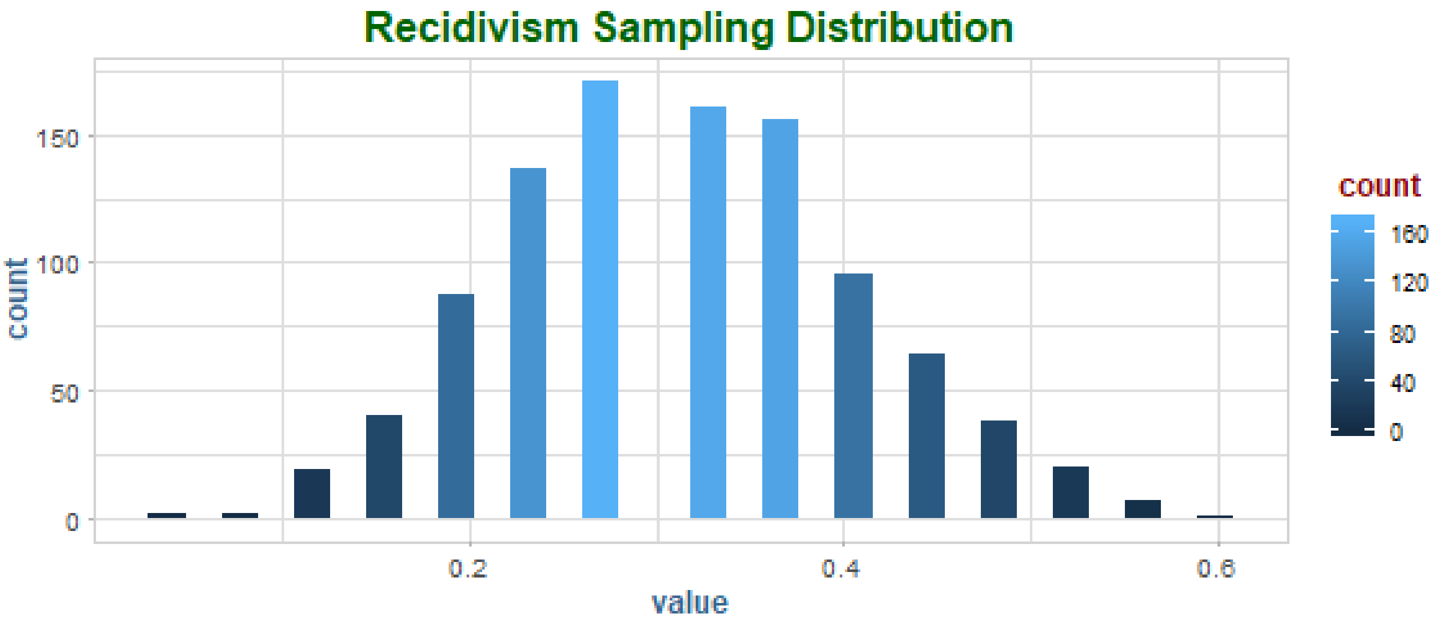
```
mean(Recidivism$Recid == "Yes")
```

```
[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)
```

$\mu = 0.31408, \sigma = 0.0184024$

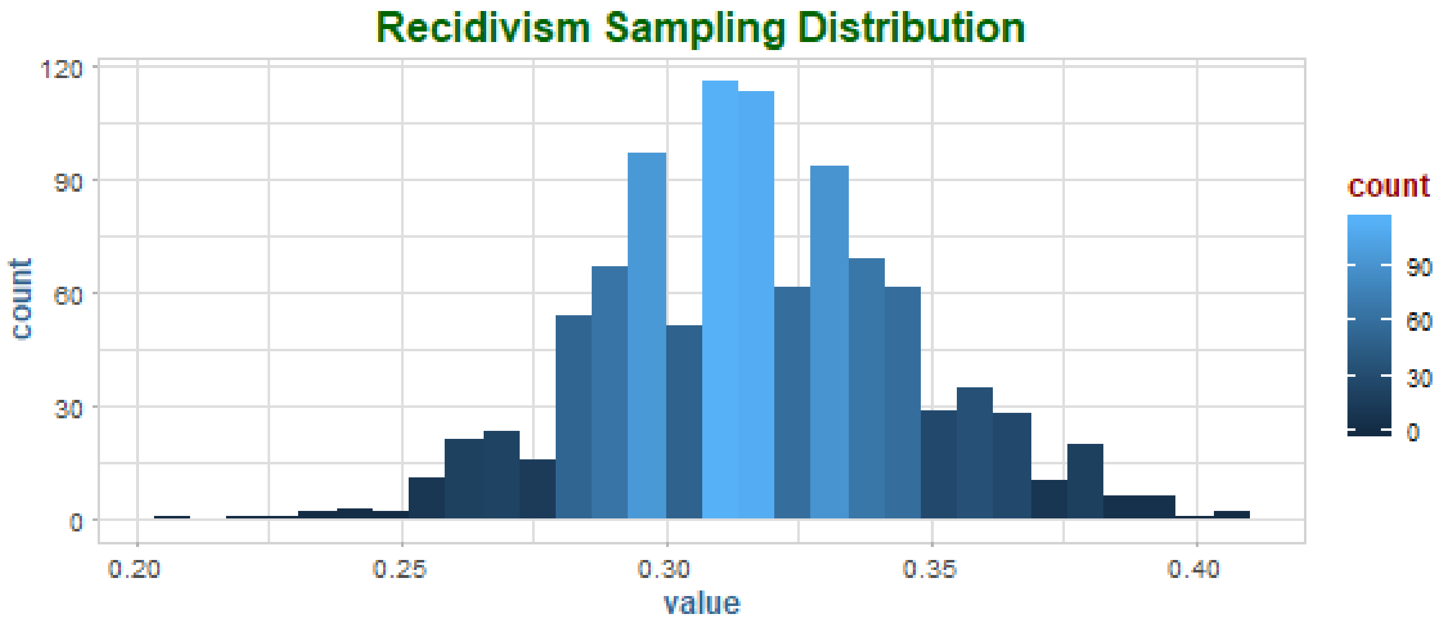
b.) Compare your estimate of the standard error with the theoretical standard error (*Corollary 4.3.2*).

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

Theoretical: 0.0430868

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)
```

$\mu = 0.317232, \sigma = 0.0018679$

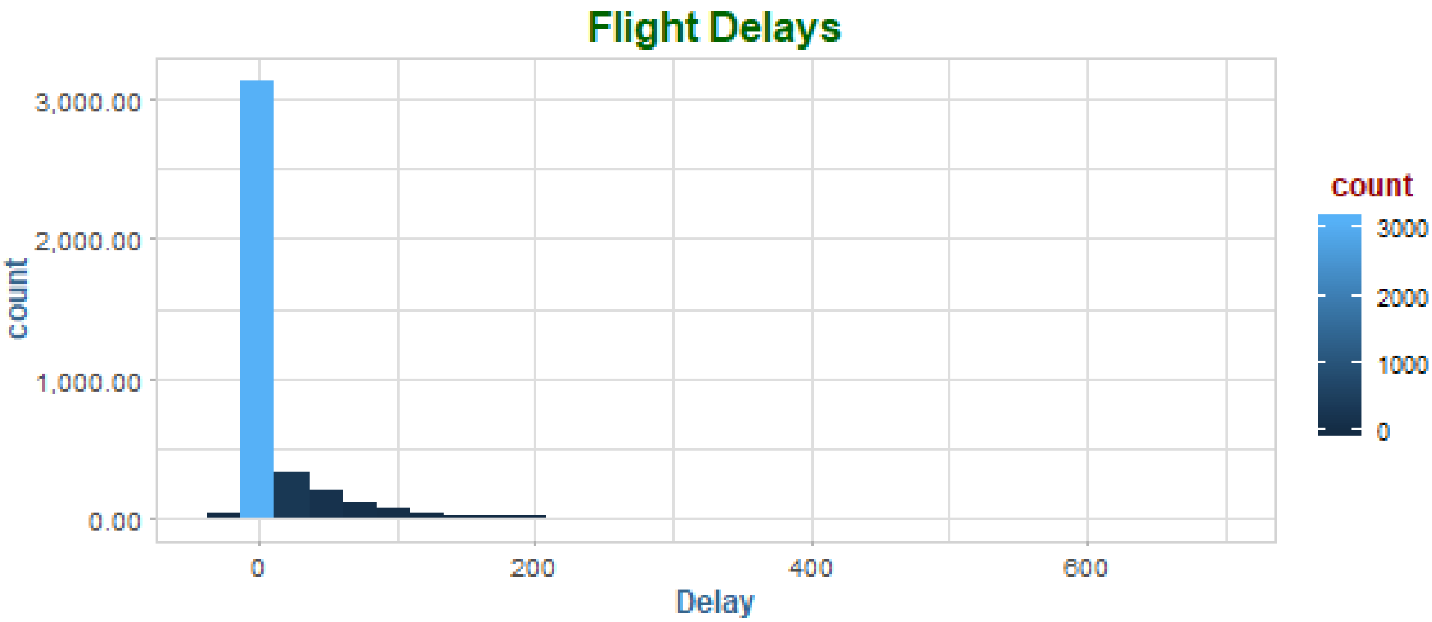
4.7

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)
```

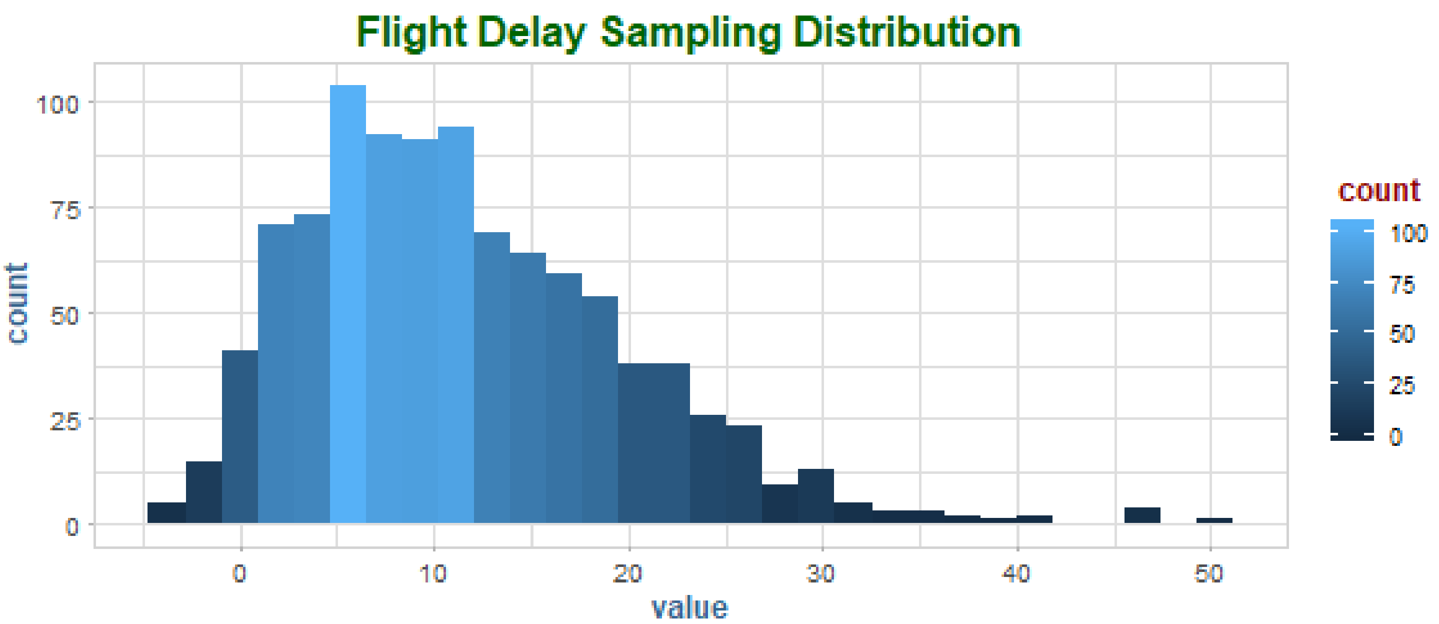
$\mu = 11.7379002, \sigma = 41.6304951$

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)
```

$\mu = 11.64688, \Sigma = 1.6759377$

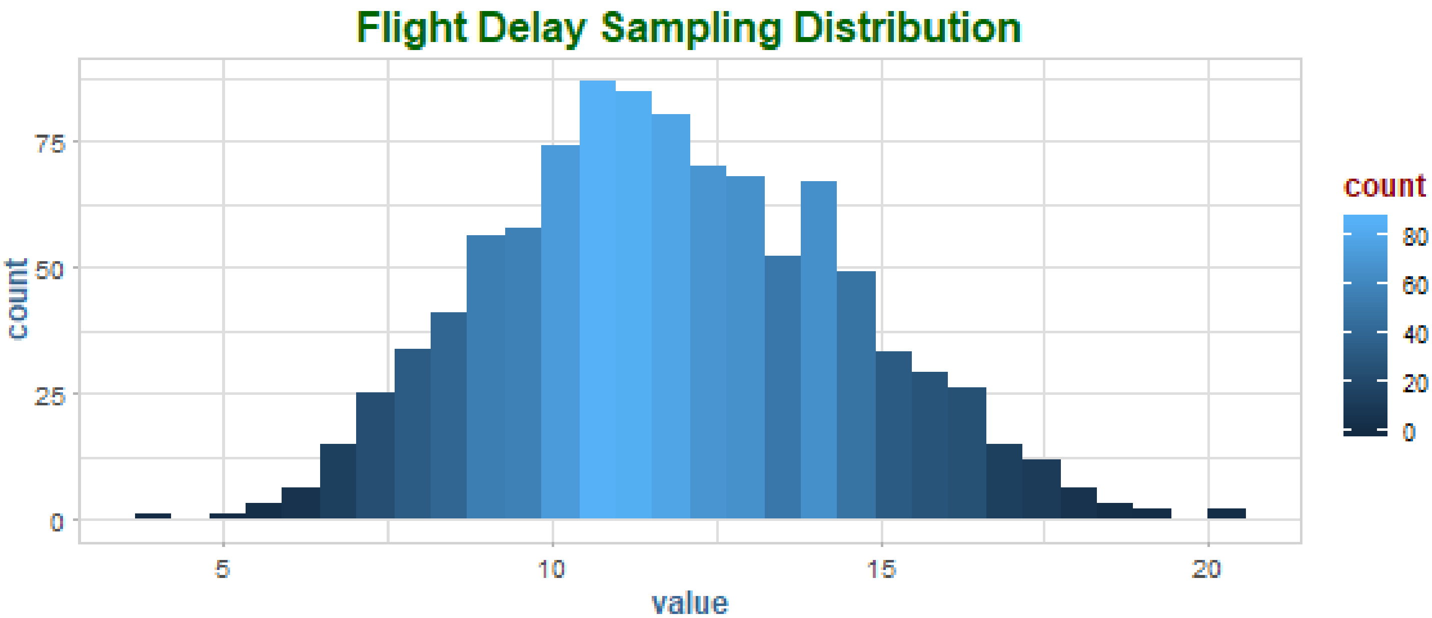
c.) Compare your estimate of the standard error with the theoretical standard error (*Corollary A.4.1*).

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

Theoretical: 2.8087672

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
```

$\mu = 11.80626, \Sigma = 0.1693495$

Theoretical: 0.0286793

4.8

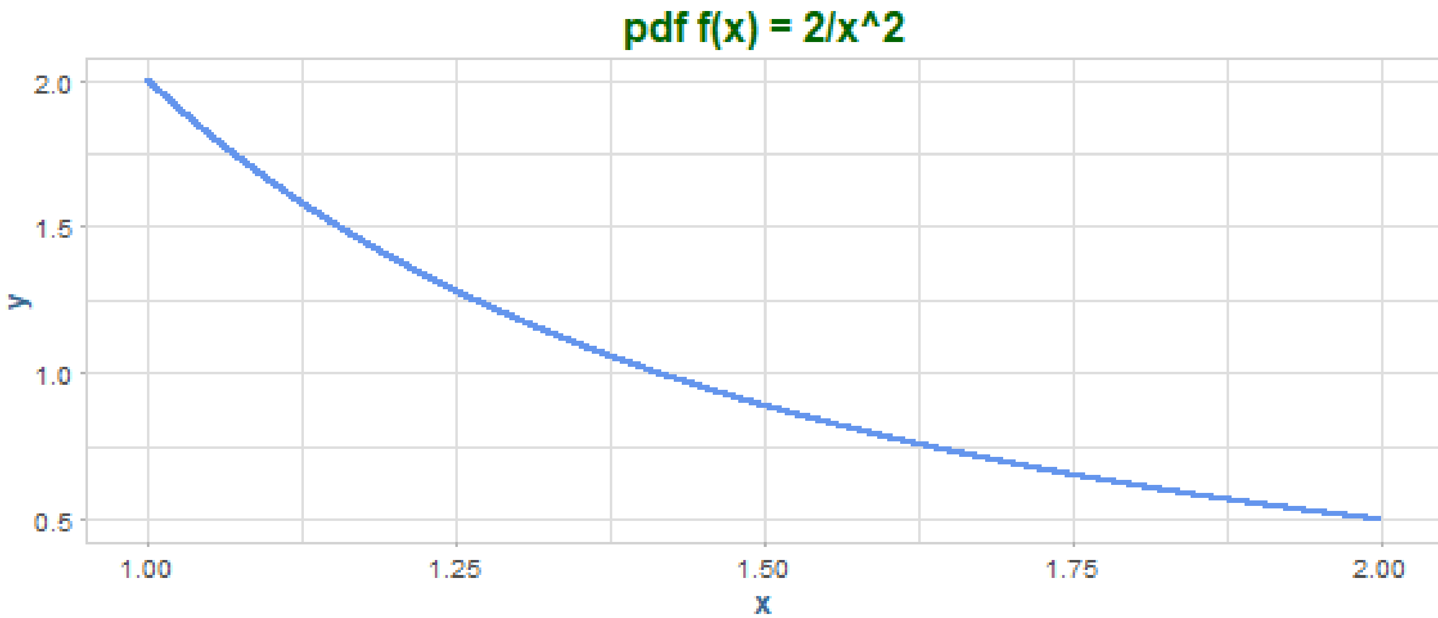
Let $X_1, X_2, \ldots, X_{25}$ be a random sample from some distribution and $W = T(X_1, X_2, \ldots, X_n)$ be a statistic.

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

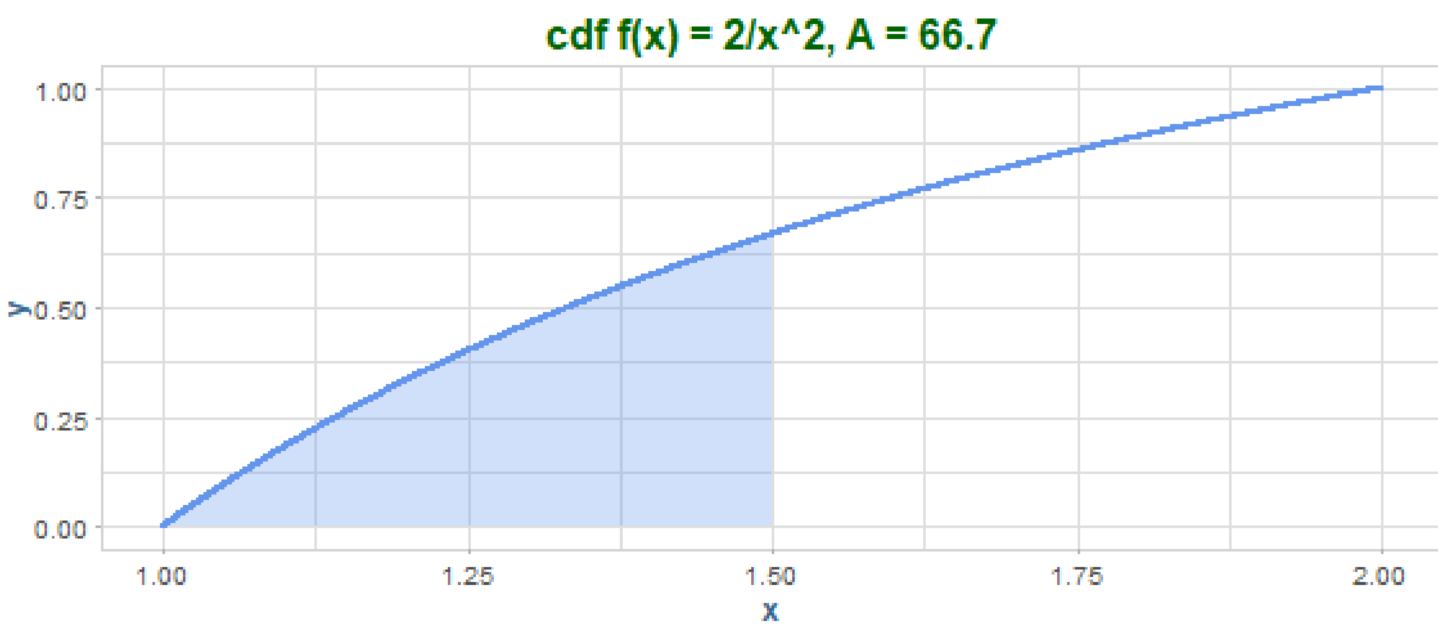
Find $P(w < 1.5)$

Solution:

```
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")
```




```r
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

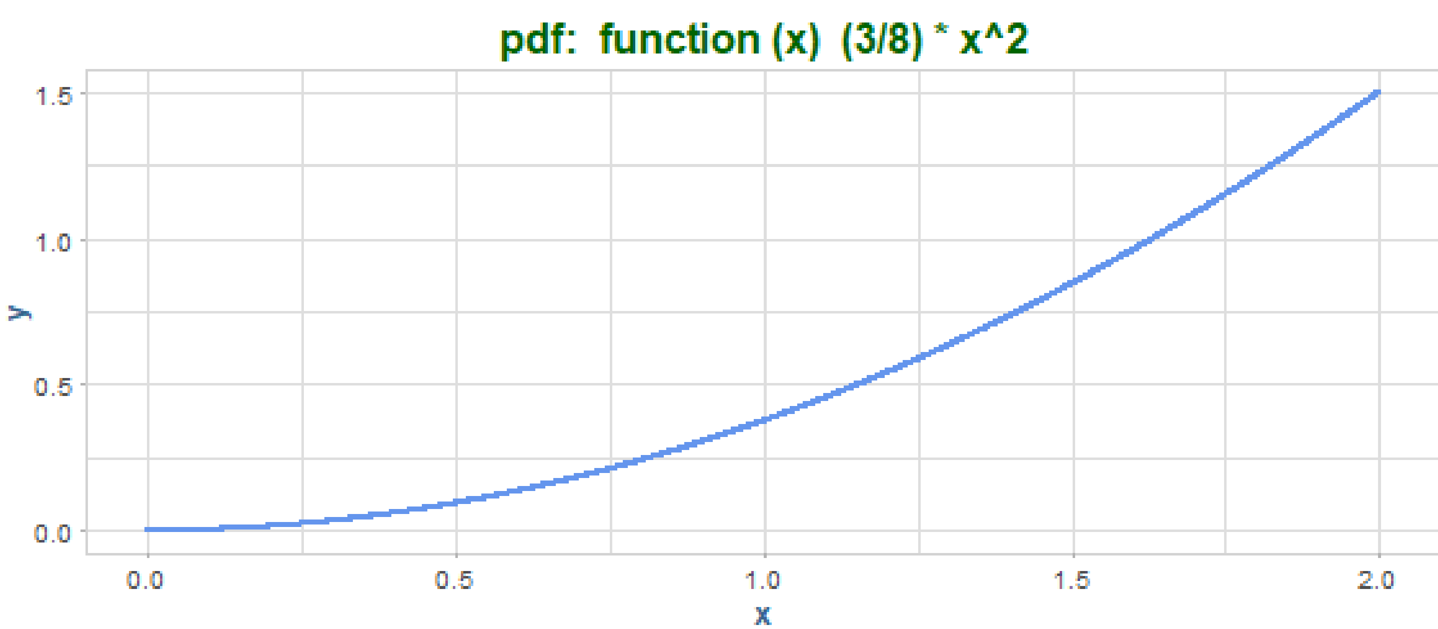
Let $X_1, X_2, \ldots, X_n$ be a random sample from some distribution and $Y = T(X_1, X_2, \ldots, X_n)$ be a statistic.

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

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

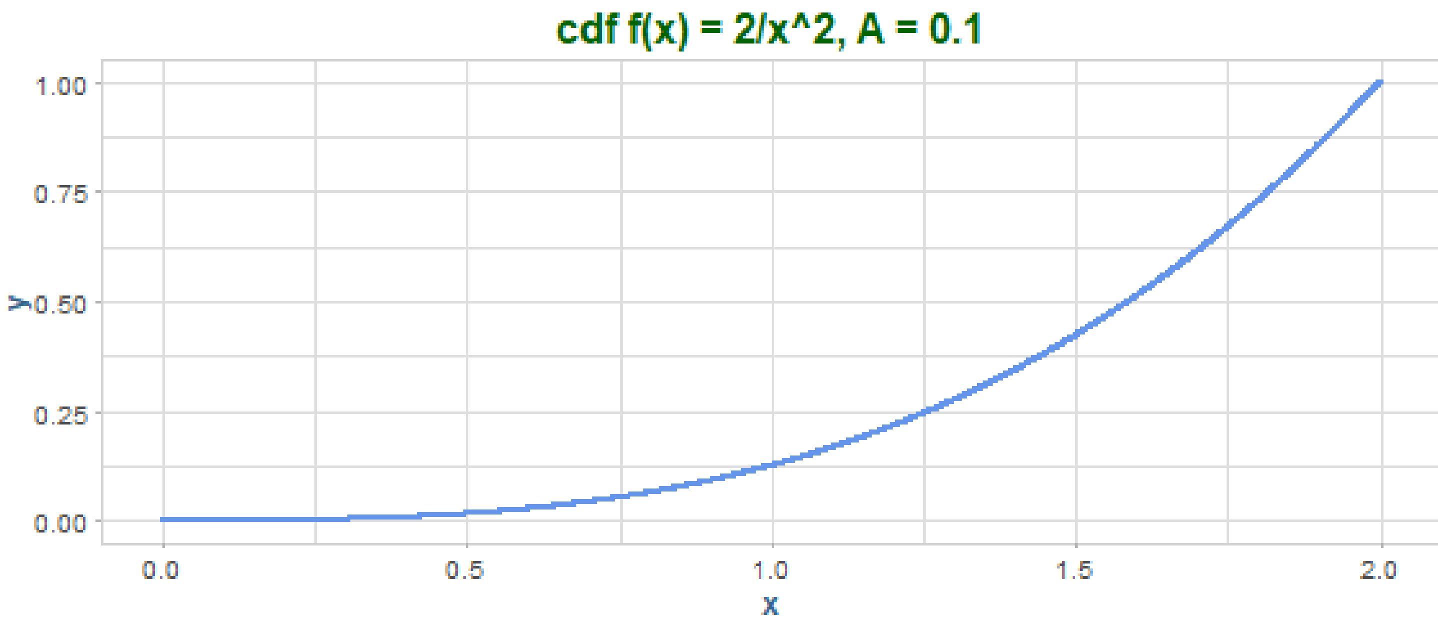
Solution:

```
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)
```

```r
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.

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

Probability: **41.96%**

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 $\hat{p} \leq 0.5$.

```
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)
```

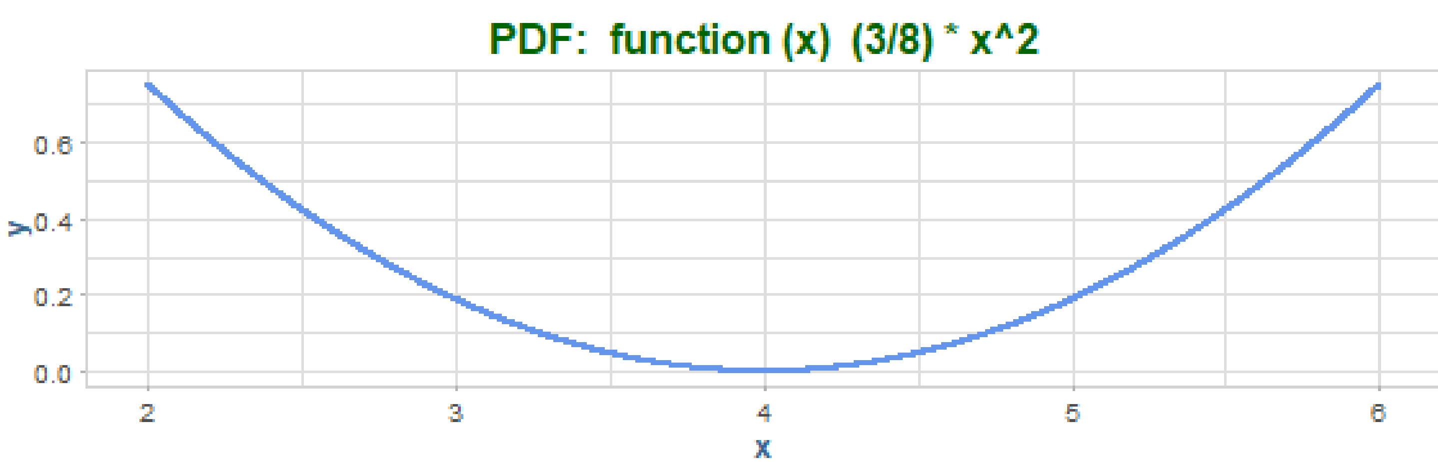
Probability: 48.75%

4.13

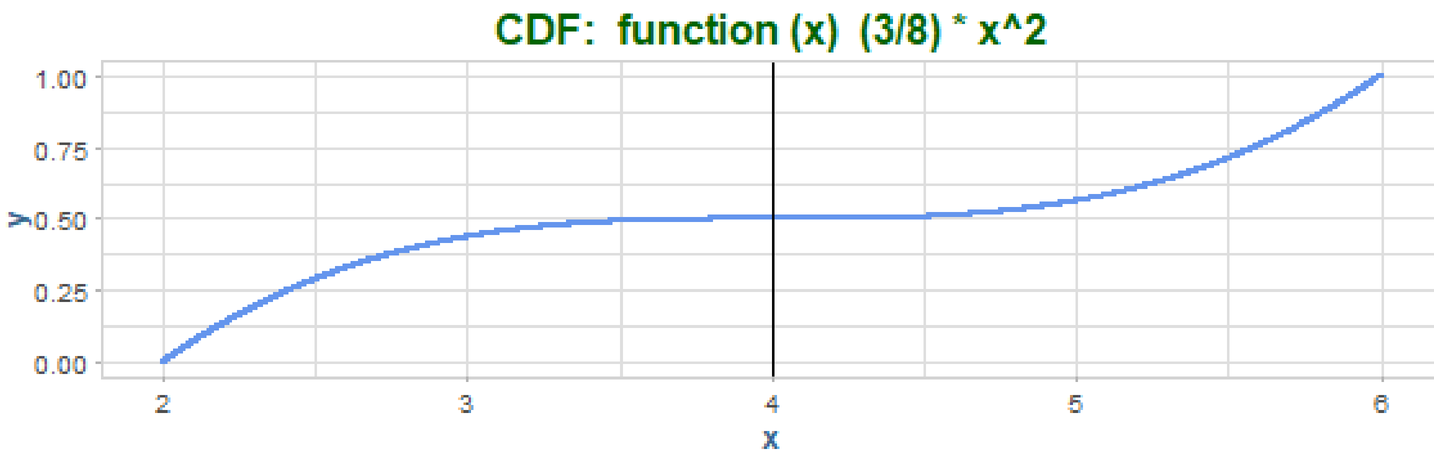
A random sample of size $n = 244$ is drawn from a distribution with pdf $f(x) = (3/16)(x - 4)^2, 2 \leq x \leq 6.$

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
```

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

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

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

p <- l - h
```

Probability: 0.3195

4.15

If $X_1, \ldots, X_n$ are i.i.d. from 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 $\lambda = 1/10$.

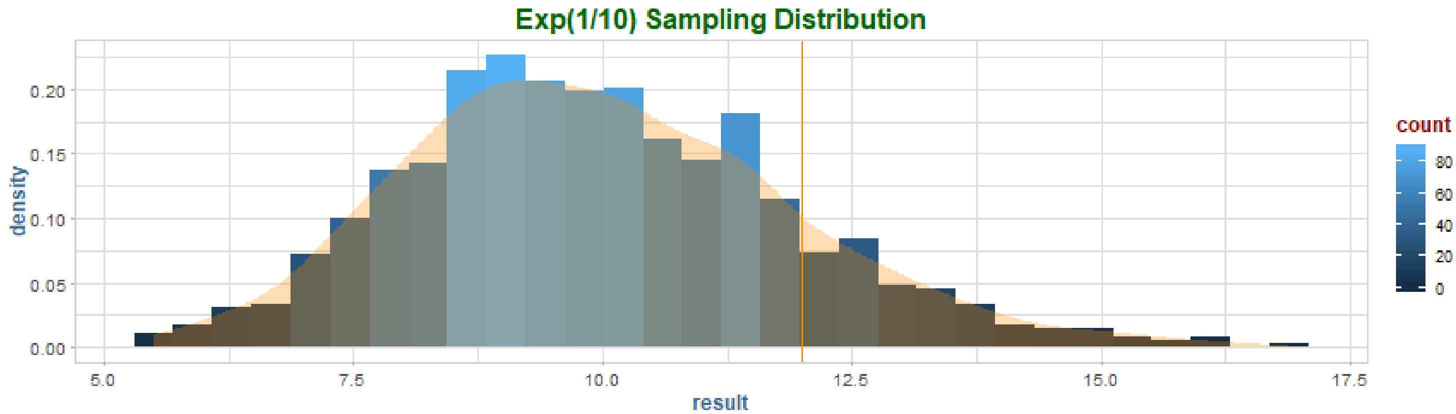
The mean of her sample is 12.

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

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

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

```r
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? **14%**

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

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

4.17

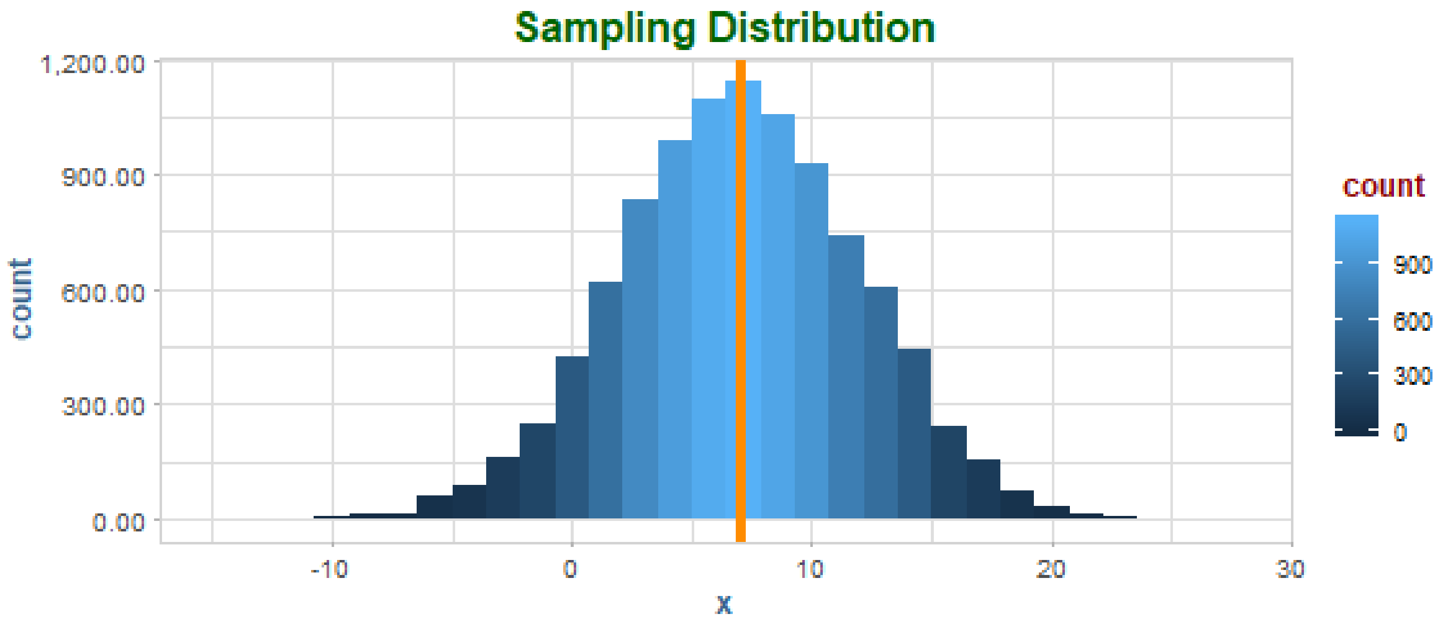
Let $X \sim N(15, 3^2)$ and $Y \sim N(4, 2^2)$ be independent random variables.

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)
```

$\mu = 7.0152707, \sigma = 5.0203046$

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)
```

$\hat{p} = 72.43\%$

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

4.18

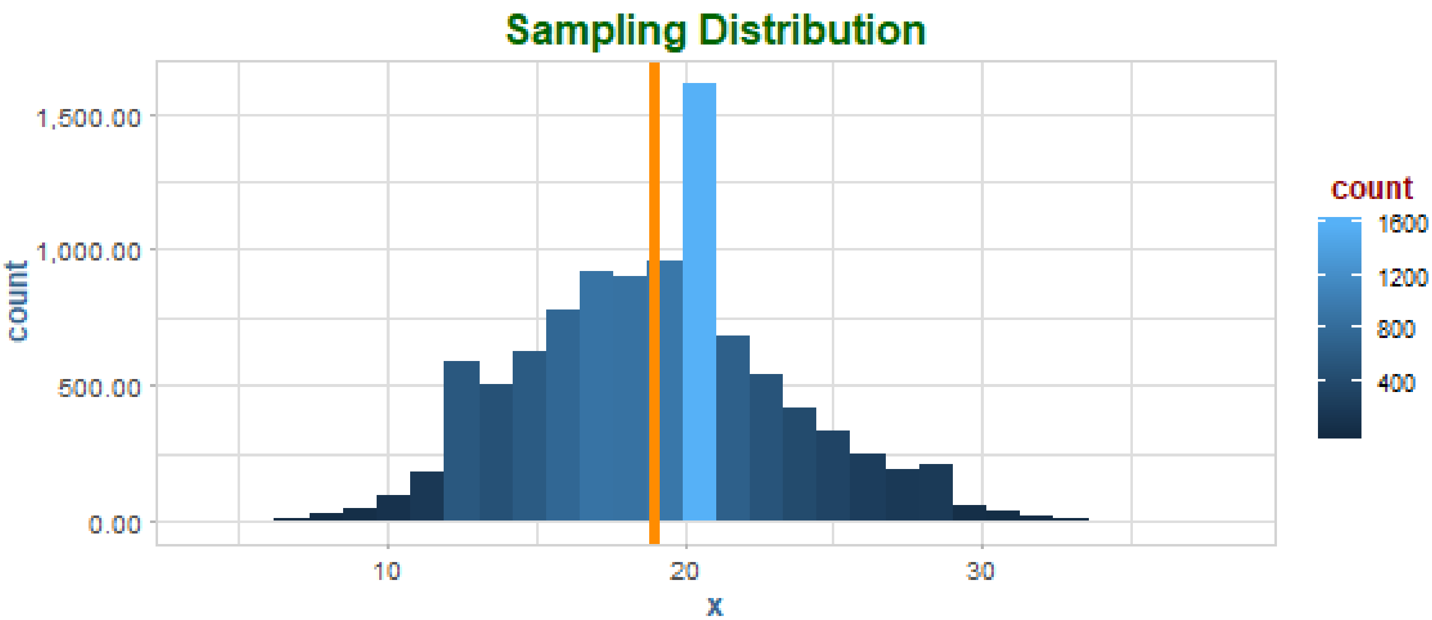
Let $X \sim Pois(4), Y \sim Pois(12), U \sim Pois(3)$ be independent random variables.

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

$W \sim Pois(19)$

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

```r
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.

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

$\mu = 19.0288, \sigma = 4.3990119$

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

```r
phat <- mean(W <= 14)

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

$\hat{p} = 14.69\%$

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

4.19

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

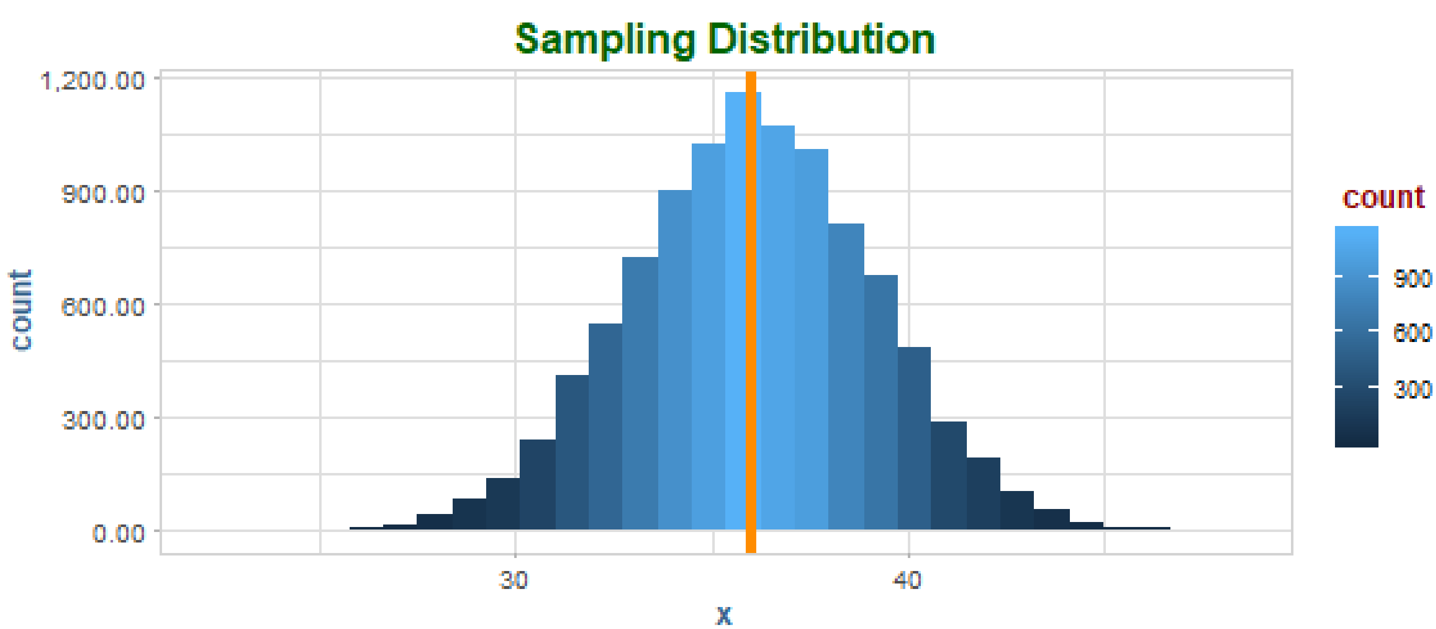
Let $W = \bar{X} + \bar{Y}$

a.) Give the exact sampling distribution of W.

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

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.

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

$\mu = 36.0265079, \sigma = 3.1146317$

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

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

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

$\hat{p} = 90.07$%

$P(W < 40) = 90.88$%