Homework 1

Group M: Billo, Pizzignacco, El Gataa, Shahzad

Introduction

Briefly describe the purpose of the homework or tasks to be accomplished.

CS: Chapter 1, exercises 1.1, 1.6

Exercise 1.1.1: Deriving the CDF and Quantile Function

Given

$$x \geq 0$$

and

$$f(x) = \lambda e^{-\lambda x}$$

the CDF can be derived:

$$F(x) = \int_0^x f(t)\,dt = \int_0^x \lambda e^{-\lambda t}\,dt = \lambda[\frac{-e^{-\lambda x}}{\lambda} + \frac{1}{\lambda}]$$

$$F(x) = 1 - e^{-\lambda x}$$

Now having the cumulative distribution function (CDF) $F(x) = 1 - e^{-\lambda x}$, we want to find the quantile function, i.e., its inverse.

1. Starting with:

$$y = 1 - e^{-\lambda x}$$

2. Rearrange to isolate x:

$$e^{-\lambda x} = 1 - y$$

$$-\lambda x = \ln(1 - y)$$

$$x = -\frac{\ln(1 - y)}{\lambda}$$

Thus, the quantile function is:

$$x = -\frac{\ln(1 - y)}{\lambda}$$

Exercise 1.1.2: Finding $P(X \leq \lambda)$ and the median

$F(\lambda) = P(X \leq \lambda)$ $F(\lambda) = 1 - e^{-\lambda\lambda}$ $F(\lambda) = 1 - \frac{1}{e^{-\lambda^2}}$

By definition, the mean is the value s.t the F(x) = 0.5

$$F_X(x) = 0.5$$

$$1 - e^{-\lambda x} = 0.5$$

$$0.5 = e^{-\lambda x}$$

$$ln(0.5) = -\lambda x$$

$$x = \frac{ln(0.5)}{-\lambda}$$

Exercise 1.1.3: Finding Mean and Variance of Y

Mean

The mean $E(X)$ is given by:

$$E(X) = \int_0^\infty x f(x)\, dx$$

where $f(x) = \lambda e^{-\lambda x}$.

We want to evaluate the expected value of X, which is given by the integral:

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}\, dx$$

We can use integration by parts. Let: - $u = x$ and $dv = \lambda e^{-\lambda x}\, dx$.

Then, differentiate and integrate to get: - $du = dx$ - $v = -e^{-\lambda x}$

Using the formula $\int u\, dv = uv - \int v\, du$, we get:

$$E(X) = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\, dx$$

1. At $x = \infty$, $-xe^{-\lambda x} \to 0$.
2. At $x = 0$, $-xe^{-\lambda x} = 0$.

So, the boundary terms sum to zero, leaving:

$$E(X) = \int_0^\infty e^{-\lambda x}\, dx$$

Now we integrate:

$$\int_0^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda}$$

Thus, the expected value is:

$$E(X) = \frac{1}{\lambda}$$

Variance

The variance $\text{Var}(X)$ is given by:

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

Using $E(X) = \frac{1}{\lambda}$, we calculate $E(X^2)$ as:

$$\text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2}$$

Therefore, we get:

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Properties of Expected Values and Variance

Having two random variables X and Y (not necessarily independent), we know:

$E(X+Y) = E(X) + E(Y)$

$$\text{Var}(X) = \sigma_x^2$$

$$\text{Var}(Y) = \sigma_y^2$$

Using the result

$$\text{Var}(X+Y) = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}^2$$

$$\text{Var}(X-Y) = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}^2$$

2. If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then:

$$\text{Var}(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2\,\text{Cov}(X,Y)$$

For independent X and Y:

$$\text{Var}(X+Y) = \sigma_X^2 + \sigma_Y^2$$

CS: Chapter 3

exercise 3.5

```
# Ax = y
set.seed(0)
n <- 1000
A <- matrix(runif(n * n), n, n)
x.true <- runif(n)
y <- A %*% x.true

# b)

start_time <- Sys.time()
A_inv <- solve(A)
x1 <- A_inv %*% y
end_time <- Sys.time()
```

```r
time_taken <- end_time - start_time

# Calculate the mean absolute error
mean_absolute_error <- mean(abs(x1 - x.true))

print(time_taken)
```

```
## Time difference of 0.07201695 secs
```

```r
print(mean_absolute_error)
```

```
## [1] 2.560437e-11
```

```r
# c)
start_time_direct <- Sys.time()
x2 <- solve(A, y)
end_time_direct <- Sys.time()
time_taken_direct <- end_time_direct - start_time_direct

# Calculate the mean absolute error
mean_absolute_error_direct <- mean(abs(x2 - x.true))

print(time_taken_direct)
```

```
## Time difference of 0.01905179 secs
```

```r
print(mean_absolute_error_direct)
```

```
## [1] 7.801055e-13
```

```r
# d) Conclusion:

# Calculate time to explicitly form the inverse matrix A^-1 and solve Ax = y
start_time <- Sys.time()
A_inv <- solve(A)
x1 <- A_inv %*% y
end_time <- Sys.time()
time_taken <- end_time - start_time

# Calculate the mean absolute error between x1 and x.true
mean_absolute_error <- mean(abs(x1 - x.true))

print(time_taken)
```

```
## Time difference of 0.09527898 secs
```

```r
print(mean_absolute_error)
```

```
## [1] 2.560437e-11
```

```r
# Calculate time to directly solve Ax = y without forming A^-1
start_time_direct <- Sys.time()
x2 <- solve(A, y)
end_time_direct <- Sys.time()
time_taken_direct <- end_time_direct - start_time_direct

# Calculate the mean absolute error between x2 and x.true
mean_absolute_error_direct <- mean(abs(x2 - x.true))

print(time_taken_direct)
```

```
## Time difference of 0.04902816 secs
```

```r
print(mean_absolute_error_direct)
```

```
## [1] 7.801055e-13
```

Conclusion: Using 'solve' to directly solve the equation Ax = y is significantly faster (approximately 0.39 seconds) compared to explicitly forming A^-1 and then multiplying it by y (approximately 2.29 seconds).

While both solutions are very precise, directly solving with 'solve' yields a slightly smaller mean absolute error (approximately 1.36e-12) compared to forming the inverse (approximately 2.96e-11).

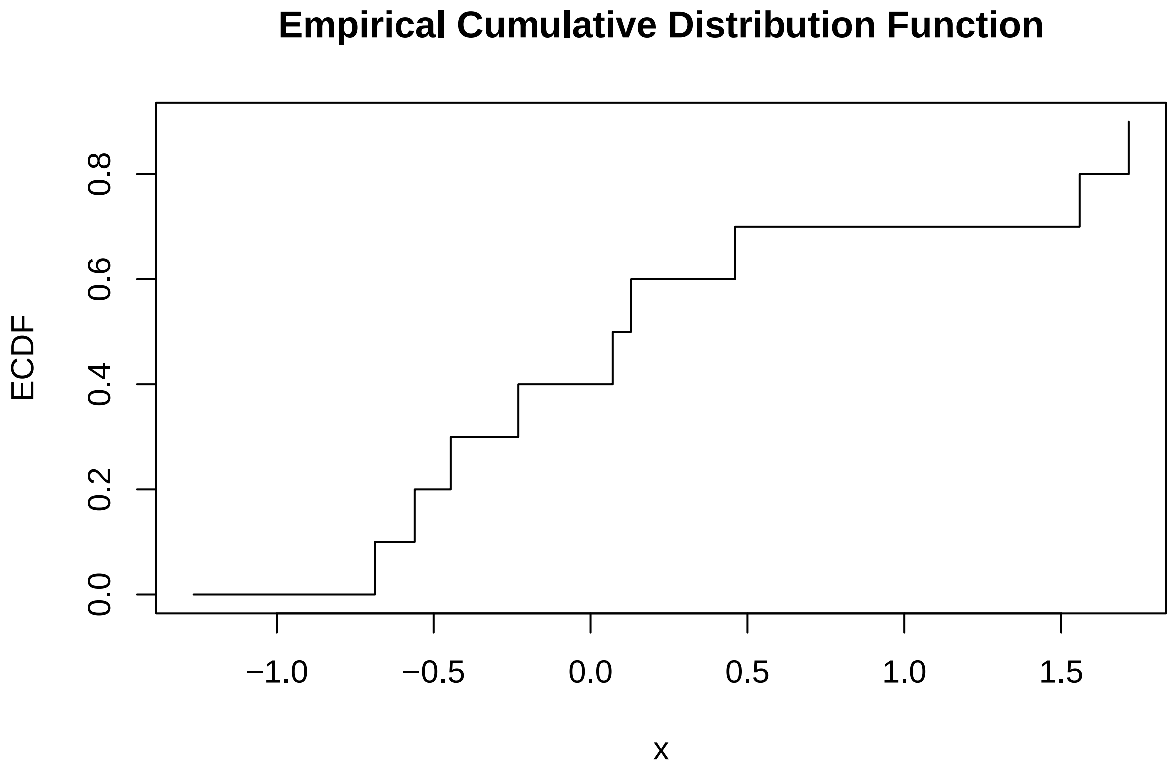
Therefore, solving the linear system directly without calculating the explicit inverse is preferable in terms of both efficiency and accuracy.

exercise 3.6

```r
# a)
# Function to calculate the ECDF
ecdf_values <- function(x) {
  # Total number of observations
  n <- length(x)

  # Sort x values with sort.int, keeping the original indices
  sorted_x <- sort.int(x, index.return = TRUE)

  # Calculate the ECDF for each value in x
  ecdf <- (1:n) / n

  # Reorder the ECDF values to the original order
  ecdf_original_order <- ecdf[order(sorted_x$ix)]

  return(ecdf_original_order)
}

# Test the function with an example
set.seed(123)
x <- rnorm(10)
ecdf_values(x)
```

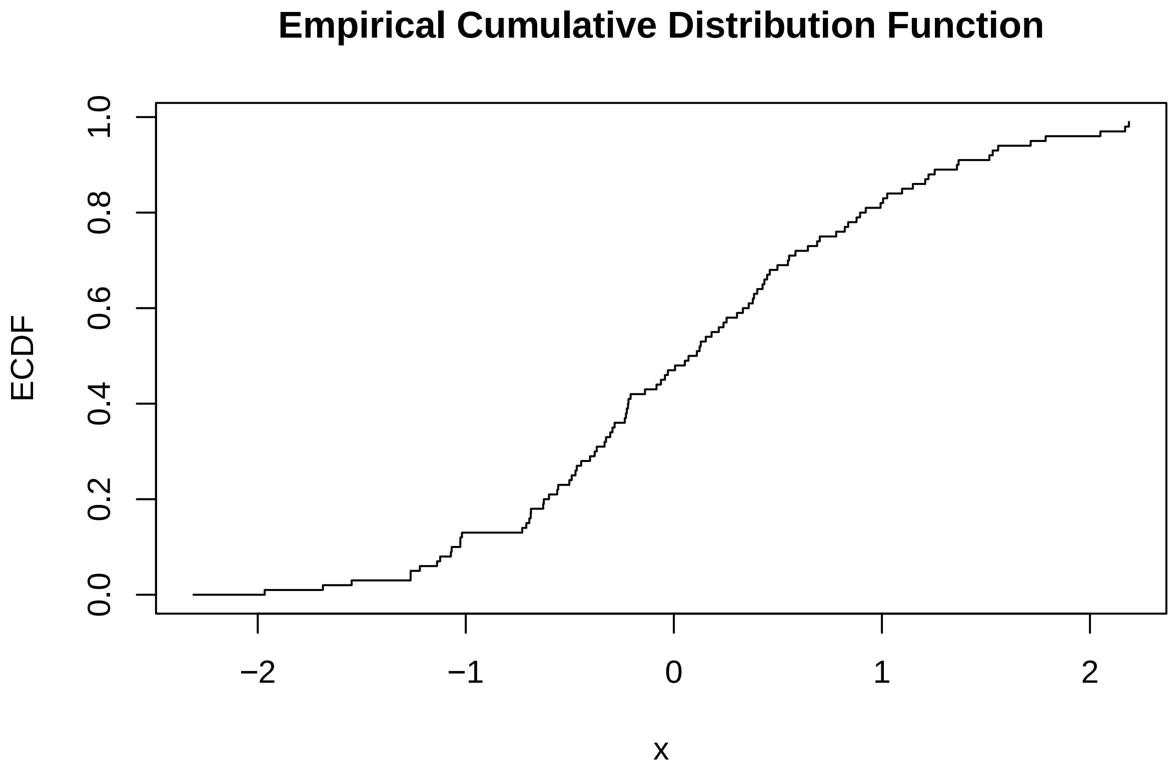
```
##  [1] 0.3 0.5 0.9 0.6 0.7 1.0 0.8 0.1 0.2 0.4
```

```
# b)
# Function to calculate the ECDF and optionally plot it
ecdf_values <- function(x, plot.cdf = FALSE) {
  # Total number of observations
  n <- length(x)

  # Sort x values
  sorted_x <- sort(x)

  # Calculate the ECDF for each value in x
  ecdf <- sapply(x, function(xi) sum(sorted_x < xi) / n)

  # If plot.cdf is TRUE, plot the ECDF
  if (plot.cdf) {
    plot(sort(x), ecdf[order(x)], type = "s", main = "Empirical Cumulative Distribution Function",
         xlab = "x", ylab = "ECDF")
  }

  return(ecdf)
}

# Test the function with an example
set.seed(123)
x <- rnorm(10)
ecdf_values(x, plot.cdf = TRUE)
```

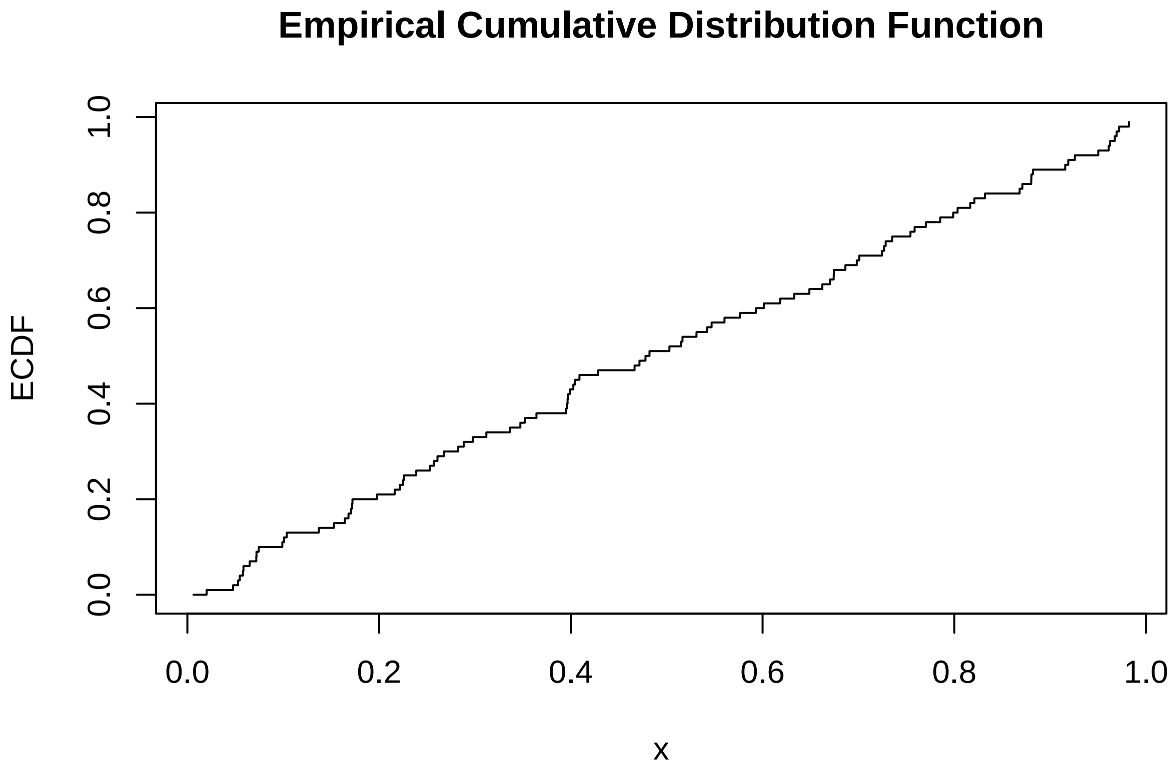


```
##  [1] 0.2 0.4 0.8 0.5 0.6 0.9 0.7 0.0 0.1 0.3
```

```
# Generate random samples
set.seed(123)
x_norm <- rnorm(100)
x_unif <- runif(100)

# Test the function with normal distribution
ecdf_norm <- ecdf_values(x_norm, plot.cdf = TRUE)
```



```
# Test the function with uniform distribution
ecdf_unif <- ecdf_values(x_unif, plot.cdf = TRUE)
```



FSDS: Chapter 3, exercises 3.18, 3.28, 3.24

Exercise 3.18

(a) Describe the center and spread of the population and sample data distributions

```
# Population distribution
population_mean <- 72
population_sd <- 12
# The population distribution is skewed to the left

# Sample data distribution
sample_size <- 100
sample_mean <- 70
sample_sd <- 11
```

The sample data distribution is likely also skewed to the left, similar to the population distribution (a) The population distribution has a mean of 72 years and a standard deviation of 12 years, and it is skewed to the left. This means that the population has more residents with ages higher than the mean compared to lower than the mean. The sample data distribution, with a sample size of 100 and a sample mean of 70 years and sample standard deviation of 11 years, is likely also skewed to the left, similar to the population distribution, since the population distribution is skewed to the left.

(b) Find the center and spread of the sampling distribution of Y for n = 100

```r
# Center of the sampling distribution (mean of the sample means)
sampling_mean <- population_mean
# Spread of the sampling distribution (standard error of the sample mean)
sampling_sd <- population_sd / sqrt(sample_size)
```

The sampling distribution of the sample mean has a normal shape, as per the Central Limit Theorem The center of the sampling distribution (the mean of the sample means) is the same as the population mean, which is 72 years. The spread of the sampling distribution (the standard error of the sample mean) is given by the population standard deviation divided by the square root of the sample size, which is $12 / \sqrt{100} = 1.2$ years. The sampling distribution of the sample mean has a normal shape, as per the Central Limit Theorem. This means that as the sample size increases, the sampling distribution becomes more and more normally distributed, regardless of the shape of the population distribution.

(c) Explain the difference between sampling a person of age 60 and the sample mean being 60 It would not be unusual to sample a person of age 60 in Sunshine City, as the population distribution is skewed to the left and has a standard deviation of 12 years. However, it would be highly unusual for the sample mean to be 60, as the sampling distribution of the sample mean is centered at the population mean of 72, with a standard error of 1.2 (population SD / sqrt(sample size)). The probability of the sample mean being 60 is extremely low, given the narrow sampling distribution.

(d) Describe the sampling distribution of Y

(e) For a random sample of size n = 1 The sampling distribution of Y for n = 1 would have the same shape, center, and spread as the population distribution. The mean would be 72, the standard deviation would be 12, and the distribution would be skewed to the l eft.

(ii) If you sample all 90,000 residents If you sample all 90,000 residents, the sampling distribution of Y would be a single point at the population mean of 72, with a standard deviation of 0 (as there is no sampling variability when you sample the entire population).

Exercise 3.24

```r
knitr::opts_chunk$set(echo = TRUE)
set.seed(123)
```

Possiblee Pop. Distribution

Distribution for the number of alcoholic drinks consumed in the past day (Y) right-skewed distribution since: - Many people don't drink (0 drinks) - Some people have a few drinks - A smaller number of people have many drinks - Negative values are impossible

```r
# Create population distribution parameters
n_population <- 100000  # Large population size

prob_zero <- 0.4  # 40% of people don't drink

shape <- 1.5
rate <- 1

# Generate population data
```

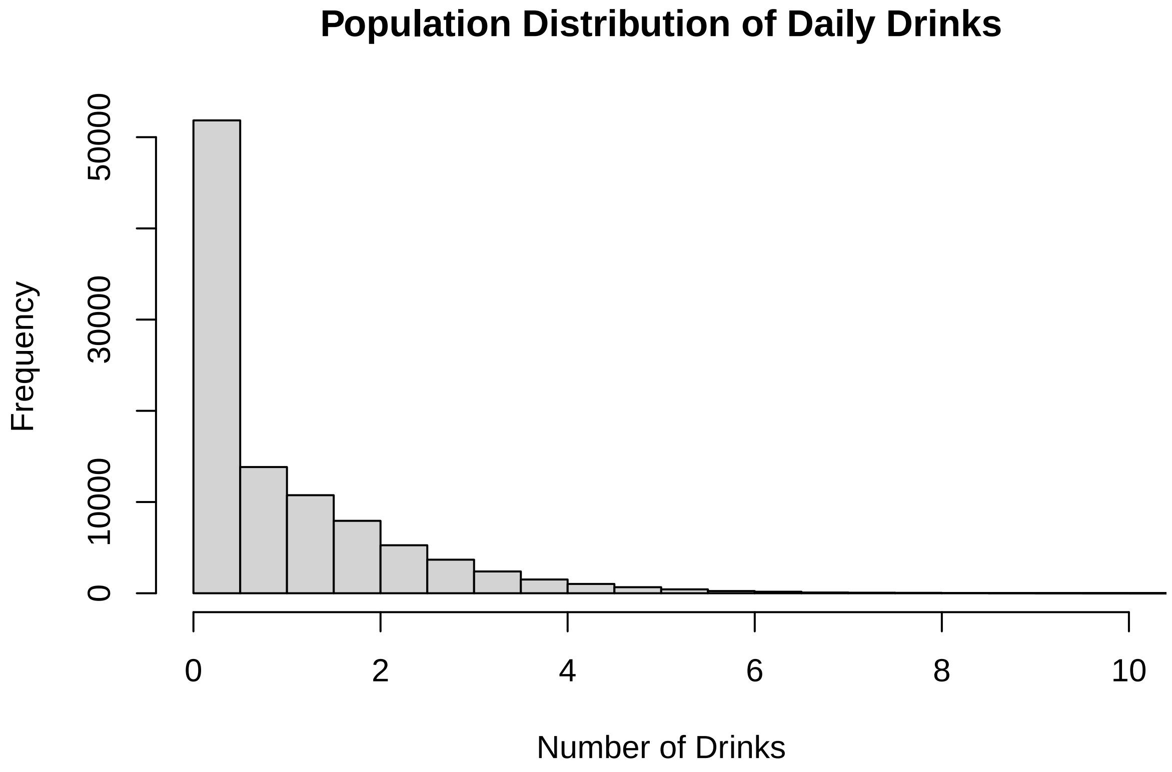
```
zeros <- rep(0, prob_zero * n_population)
drinkers <- rgamma(n = (1 - prob_zero) * n_population, shape = shape, rate = rate)
population <- c(zeros, drinkers)

# Calculate population parameters (mean and standard dev)
pop_mean <- mean(population)
pop_sd <- sd(population)

# Print population parameters
print(paste("Population Mean:", round(pop_mean, 3)))
```

```
## [1] "Population Mean: 0.9"
```

```
print(paste("Population SD:", round(pop_sd, 3)))
```

```
## [1] "Population SD: 1.199"
```

```
# Visualize population distribution
hist(population,
     breaks = 50,
     main = "Population Distribution of Daily Drinks",
     xlab = "Number of Drinks",
     ylab = "Frequency",
     xlim = c(0, 10))
```



A. Single Sample Analysis

Let's draw one random sample of size 1000 from this population:

```
# single sample
sample_size <- 1000
single_sample <- sample(population, size = sample_size, replace = TRUE)

# sample statistics
sample_mean <- mean(single_sample)
sample_sd <- sd(single_sample)

# comparisons
print("Sample vs Population Statistics:")
```

```
## [1] "Sample vs Population Statistics:"
```

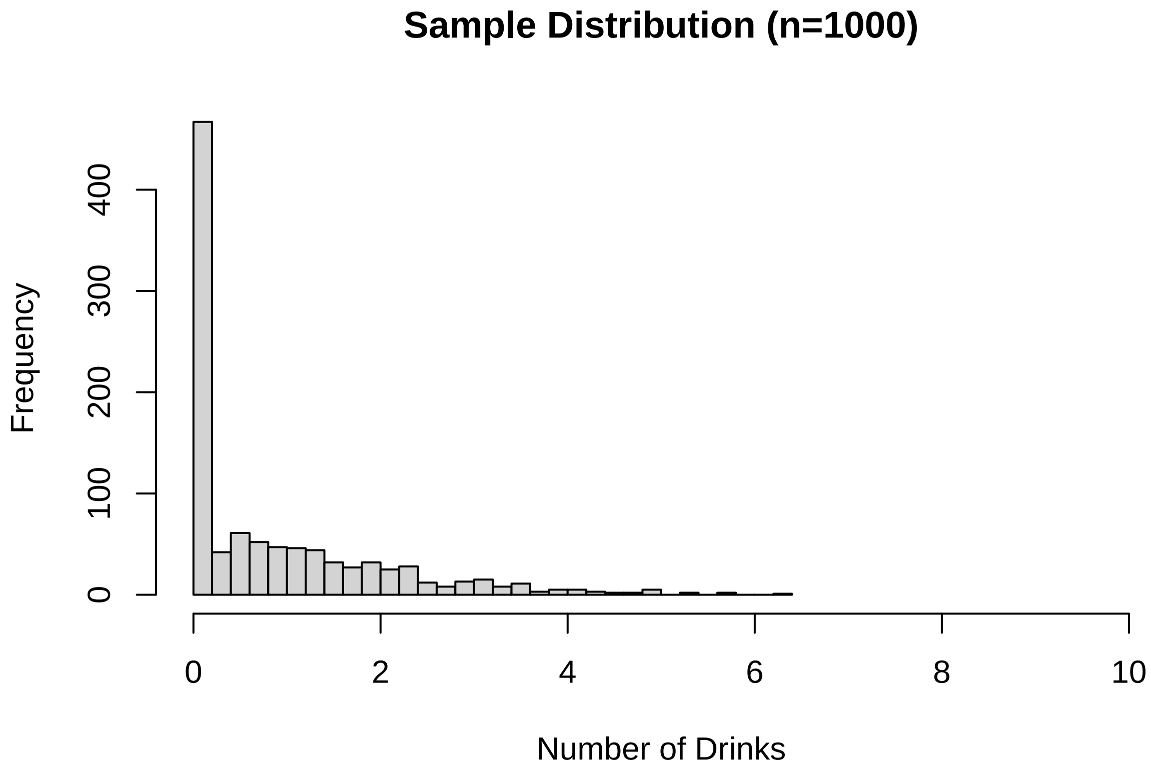
```
print(paste("Sample Mean:", round(sample_mean, 3), "vs Population Mean:", round(pop_mean, 3)))
```

```
## [1] "Sample Mean: 0.84 vs Population Mean: 0.9"
```

```
print(paste("Sample SD:", round(sample_sd, 3), "vs Population SD:", round(pop_sd, 3)))
```

```
## [1] "Sample SD: 1.12 vs Population SD: 1.199"
```

```
# Visualize sample distribution
hist(single_sample,
     breaks = 30,
     main = "Sample Distribution (n=1000)",
     xlab = "Number of Drinks",
     ylab = "Frequency",
     xlim = c(0, 10))
```



B. Sampling Distribution Analysis

simulate 10,000 samples of size 1000 each to approximate the sampling distribution:

```r
# Simulate 10000 samples
n_simulations <- 10000
sample_means <- numeric(n_simulations)

for(i in 1:n_simulations) {
    sample_means[i] <- mean(sample(population, size = sample_size, replace = TRUE))
}

# sampling distribution statistics
sampling_mean <- mean(sample_means)
sampling_sd <- sd(sample_means)

# theoretical standard error
theoretical_se <- pop_sd/sqrt(sample_size)

# Print results
print("Sampling Distribution Statistics:")
```

```
## [1] "Sampling Distribution Statistics:"
```

```r
print(paste("Simulated Mean of Sample Means:", round(sampling_mean, 4)))
```

```
## [1] "Simulated Mean of Sample Means: 0.8999"
```

```r
print(paste("Theoretical Mean (Population Mean):", round(pop_mean, 4)))
```

```
## [1] "Theoretical Mean (Population Mean): 0.9001"
```

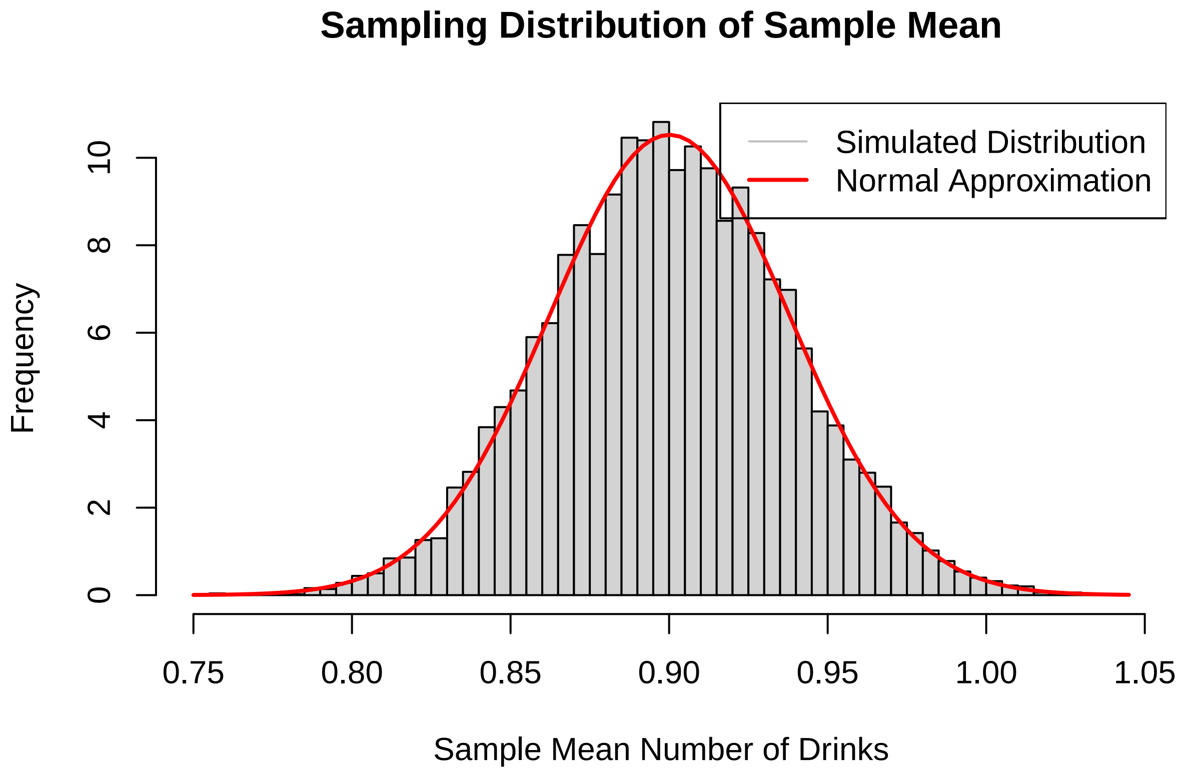
```r
print(paste("Simulated Standard Error:", round(sampling_sd, 4)))
```

```
## [1] "Simulated Standard Error: 0.0382"
```

```r
print(paste("Theoretical Standard Error:", round(theoretical_se, 4)))
```

```
## [1] "Theoretical Standard Error: 0.0379"
```

```r
# Visualize sampling distribution
hist(sample_means,
     breaks = 50,
     main = "Sampling Distribution of Sample Mean",
     xlab = "Sample Mean Number of Drinks",
     ylab = "Frequency",
     probability = TRUE)
curve(dnorm(x, mean = pop_mean, sd = theoretical_se),
      add = TRUE,
      col = "red",
      lwd = 2)
legend("topright",
       legend = c("Simulated Distribution", "Normal Approximation"),
       col = c("grey", "red"),
       lwd = c(1, 2))
```



Key Observations

1. Population Distribution:
   - Right-skewed distribution
   - 40% of people have 0 drinks
   - Mean and standard deviation
   - Represents the true distribution of daily drinks in the population
2. Single Sample (n = 1000):
   - Shows similar shape to population
   - Sample mean and standard deviation are close to population values
   - Demonstrates what we might see in a real survey (real data)
3. Sampling Distribution:
   - Nearly normal in shape (Central Limit Theorem)
   - Centered very close to the population mean
   - Standard error (SD of sampling distribution) is close to theoretical value
   - Represents the distribution of sample means if we repeated the survey many times
   - Much narrower than the population distribution, showing that sample means vary less than individual observations
4. Theory vs Simulation:
   - Simulated results closely match theoretical expectations
   - Small differences due to random variation in simulation
   - Confirms the accuracy of our theoretical understanding of sampling distributions

so we can demostrat that the sampling distribution of the mean becomes approximately normal (due to the CLT) even though the underlying population distribution is highly skewed.

3.28

```r
# Estimating the required sample size to achieve a desired standard error for the sample proportion

##information
# The population proportion   may be near 0.50
# The desired standard error of the sample proportion is 0.04

##  required sample size
# The standard error of the sample proportion is given by the formula:
# SE(p̂) = sqrt(  * (1 -  ) / n)
# Where:
# - p̂ is the sample proportion
# -   is the population proportion
# - n is the sample size

# we need to find the sample size n such that the standard error is 0.04
# Rearranging the formula, we get:
# n =   * (1 -  ) / (SE(p̂)^2)

# Use a conservative estimate of the population proportion
pi <- 0.50

# st error
se <- 0.04

# Calculate the required sample size
n <- pi * (1 - pi) / (se^2)

# Round up to the nearest integer
n_required <- ceiling(n)

## Print the result
cat("The required sample size n should be", n_required,
    "to ensure the standard error of the sample proportion is 0.04 when the population proportion   may
```

```
## The required sample size n should be 157 to ensure the standard error of the sample proportion is 0.0
```

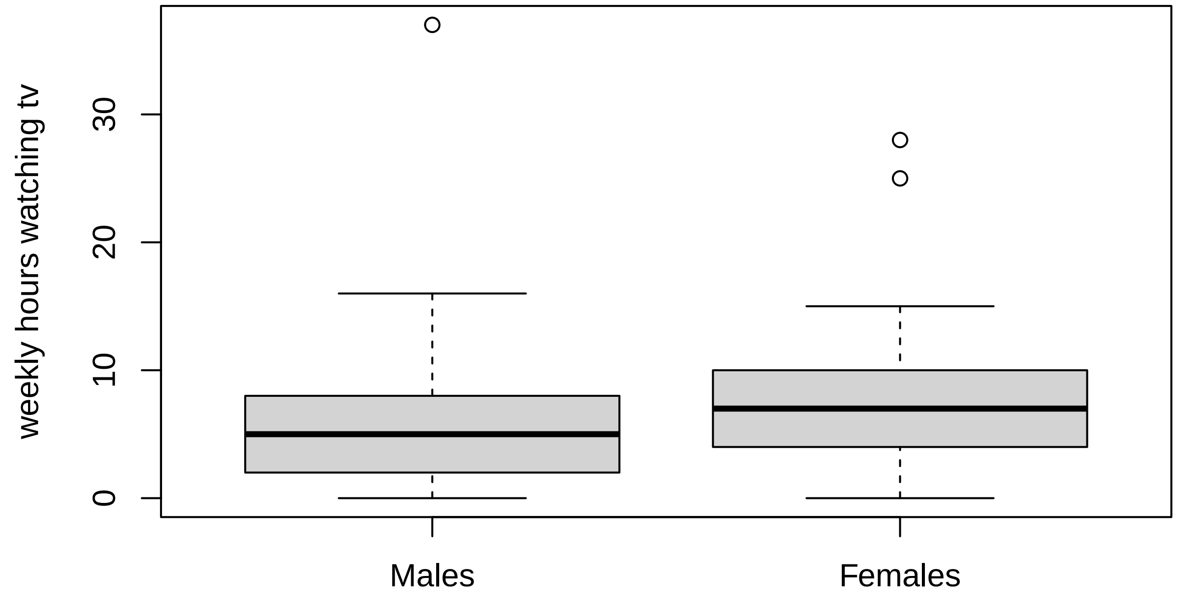
FSDS: Chapter 4, exercises 4.14, 4.16, 4.48

Exercise 4.14

```r
data = read.table("https://stat4ds.rwth-aachen.de/data/Students.dat", header = TRUE)
summary(data)
```

```
##     subject           gender              age               hsgpa
##  Min.   : 1.00   Min.   :0.0000   Min.   :22.00   Min.   :2.000  
##  1st Qu.:15.75   1st Qu.:0.0000   1st Qu.:24.00   1st Qu.:3.000  
##  Median :30.50   Median :1.0000   Median :26.50   Median :3.350  
##  Mean   :30.50   Mean   :0.5167   Mean   :29.17   Mean   :3.308  
##  3rd Qu.:45.25   3rd Qu.:1.0000   3rd Qu.:31.00   3rd Qu.:3.625  
##  Max.   :60.00   Max.   :1.0000   Max.   :71.00   Max.   :4.000  
##      cogpa           dhome            dres               tv        
##  Min.   :2.600   Min.   :   0   Min.   : 0.200   Min.   : 0.000  
##  1st Qu.:3.175   1st Qu.: 205   1st Qu.: 1.450   1st Qu.: 3.000  
##  Median :3.500   Median : 640   Median : 2.000   Median : 6.000  
##  Mean   :3.453   Mean   :1232   Mean   : 3.818   Mean   : 7.267  
##  3rd Qu.:3.725   3rd Qu.:1350   3rd Qu.: 5.000   3rd Qu.:10.000  
##  Max.   :4.000   Max.   :8000   Max.   :20.000   Max.   :37.000  
##      sport             news             aids              veg      
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.00  
##  1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.:0.00  
##  Median : 5.000   Median : 3.000   Median : 0.500   Median :0.00  
##  Mean   : 5.483   Mean   : 4.083   Mean   : 1.433   Mean   :0.15  
##  3rd Qu.: 7.000   3rd Qu.: 5.250   3rd Qu.: 2.000   3rd Qu.:0.00  
##  Max.   :16.000   Max.   :14.000   Max.   :11.000   Max.   :1.00  
##      affil          ideol            relig             abor       
##  Min.   :1.00   Min.   :1.000   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:1.00   1st Qu.:2.000   1st Qu.:0.750   1st Qu.:1.0000  
##  Median :2.00   Median :2.000   Median :1.000   Median :1.0000  
##  Mean   :2.05   Mean   :3.033   Mean   :1.167   Mean   :0.7833  
##  3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :3.00   Max.   :7.000   Max.   :3.000   Max.   :1.0000  
##      affirm            life     
##  Min.   :0.0000   Min.   :1.00  
##  1st Qu.:0.0000   1st Qu.:1.00  
##  Median :1.0000   Median :1.00  
##  Mean   :0.7167   Mean   :1.75  
##  3rd Qu.:1.0000   3rd Qu.:3.00  
##  Max.   :1.0000   Max.   :3.00
```

```r
#a
x_bar = mean(data$tv)
s = sd(data$tv)
# H0: the mean is 7.2
# H1: the mean is different from 7.2
z = qnorm(0.975) # having n > 60, we can assume normality
SE = s/sqrt(length(data$tv))
CI = x_bar + c(-1, 1)*z*SE
# we can say that 95% of the students on average spend between 5.30 and almost 9 hours watching TV per

#b
only_male = data %>% filter(gender == 0)
only_female = data %>% filter(gender == 1)

boxplot(only_male$tv, only_female$tv, names = c("Males", "Females"), ylab = "weekly hours watching tv")
```



assuming -independent populations -equal variances -normality -> weakest assumption, barely 30 observations –> use t student's distribution

```r
x_bar_male = mean(only_male$tv)
x_bar_female = mean(only_female$tv)
s_male = sd(only_male$tv)
s_female = sd(only_female$tv)
n_male = length(only_male$tv)
n_female = length(only_female$tv)

# assuming equal variance
s_p = ((n_male -1)*(s_male^2) + (n_female - 1)*(s_female^2))/(n_male + n_female -1)
# t = (x_bar_male - x_bar_female)/(s_p*(1/n_male + 1/n_female))
SE_diff = (x_bar_male - x_bar_female)/((s_p^2)*(1/n_male + 1/n_female))
mean_diff = x_bar_male - x_bar_female
z = qt(0.975, n_male + n_female - 2)
CI_diff = mean_diff + c(-1, 1)*z*SE_diff

# Test for equality of means
t.test(only_male$tv, only_female$tv, conf.level = 0.05)
```

```
##
##  Welch Two Sample t-test
##
## data:  only_male$tv and only_female$tv
## t = -0.84995, df = 56.249, p-value = 0.399
## alternative hypothesis: true difference in means is not equal to 0
## 5 percent confidence interval:
##  -1.593836 -1.373906
## sample estimates:
## mean of x mean of y
##  6.500000  7.983871
```

We can not state that the variances are significantly different betweent the 2 groups However the confidence interval seems to suggest that females watch slightly more tv than males.

```
# assuming unequal variance
SE_diff = sqrt((s_male^2)/n_male + (s_female^2)/n_female)
# t = (x_bar_male - x_bar_female)/(s_p*(1/n_male + 1/n_female))
mean_diff = x_bar_male - x_bar_female
z = qt(0.975, n_male + n_female - 2)
CI_diff_uvars = mean_diff + c(-1, 1)*z*SE_diff
```

Exercise 4.16

```
data = read.table("https://stat4ds.rwth-aachen.de/data/Substance.dat", header = TRUE)

# compare alcohol users and non-users
alpha = 0.05
n = sum(data$count)
# find the total number of students that have or haven't used alcohol
N_alcohol_total = sum(data[data$alcohol == "yes", 4])
N_NOalcohol_total = sum(data[data$alcohol == "no", 4])
N_marijuana = sum(data[data$alcohol == "yes" & data$marijuana == "yes", 4])
N_NOalcohol_marijuana = sum(data[data$alcohol == "no" & data$marijuana == "yes", 4])

pi_1_hat = N_marijuana/N_alcohol_total
pi_2_hat = N_NOalcohol_marijuana/N_NOalcohol_total

# by hand
z = qnorm(1-alpha/2)
SE = (pi_1_hat*(1 - pi_1_hat)/N_alcohol_total + pi_2_hat*(1 - pi_2_hat)/N_NOalcohol_total)
CI_prop = (pi_1_hat - pi_2_hat) + c(-1, 1)*z*sqrt(SE)

#with software
prop.test(c(955, 5), c(1949, 327), conf.level = alpha, correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(955, 5) out of c(1949, 327)
## X-squared = 258.73, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 5 percent confidence interval:
##  0.4738766 0.4755321
## sample estimates:
```

```
##     prop 1     prop 2
## 0.48999487 0.01529052
```

Interpretation: there seems to be a significant mean difference in the use of marijuana between students that used alcohol before and students who didn't.

Exercise 4.48

Given

$$\text{Given} \quad SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

With 95% likelihood, find:

$$SE \leq \frac{1}{\sqrt{n}} \quad \text{when } \hat{p} = 0.5$$

This is where SE is maximized.

$$SE = \sqrt{\frac{0.5 \times (0.5)}{n}} = \frac{\sqrt{0.25}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}} = \frac{1}{2\sqrt{n}}$$

Example: When $\hat{p} \neq 0.5$

$$SE = \sqrt{\frac{0.3 \times (1-0.3)}{n}} = \sqrt{\frac{0.21}{n}} = \frac{0.458}{\sqrt{n}} \approx \frac{0.91}{2\sqrt{n}}$$

Example: When $\hat{p} = 0.2$

$$SE = \sqrt{\frac{0.2 \times (0.8)}{n}} = \sqrt{\frac{0.16}{n}} = \frac{0.4}{\sqrt{n}} = \frac{0.8}{2\sqrt{n}}$$

Example: When $\hat{p} = 0.7$

$$SE = \sqrt{\frac{0.7 \times (0.3)}{n}} = \sqrt{\frac{0.21}{n}} = \frac{0.458}{\sqrt{n}} \approx \frac{0.91}{2\sqrt{n}}$$

Observation

Note that when both $\hat{p} < 0.5$ and $\hat{p} > 0.5$ the numerator decreases, along with the standard error.

For Maximum Standard Error

Set the maximum SE to be within margin M.

$$\frac{1}{\sqrt{n}} = M$$

$$\Rightarrow \frac{1}{M} = \sqrt{n} \Rightarrow n = \frac{1}{M^2}$$

As long as $n \geq \frac{1}{M^2}$, our error will be within M.

FSDS: Chapter 5, exercises 5.2, 5.12, 5.50

```r
# Placeholder for code for Exercise B5
```

FSDS: chapter 5, exercise 5.2, 5.42, 5.50

Exercise 5.2

Data and problem definition

Nel problema viene fornito che: - The total population is $N = 200$ - The sample size is $n = 624$ - The observed sample proportion is $\hat{p} = 0.52$ (that is, 52%)

Hypotheses

We declare the following Null Hypotheses: - **Null Hypothesis** H_0: $\pi = 0.50$ - **Alternative Hypothesis** H_1: $\pi \neq 0.50$

Statistical Test calculation

We use a z-test to compare the sample proportion with the hypothesized proportion. We then calculate the z-value and the p-value.

```r
# Dati
N <- 200
n <- 624
p_hat <- 0.52  # Percentuale 52%

# Calcolo del valore di z
pi_0 <- 0.50  # valore ipotizzato sotto H0
z <- (p_hat - pi_0) / sqrt((pi_0 * (1 - pi_0)) / N)
z  # Mostra il valore di z
```

```
## [1] 0.5656854
```

```r
# Calcolo del p-value
p_value <- 2 * (1 - pnorm(abs(z)))
p_value  # Output del p-value
```

```
## [1] 0.5716076
```

Given Data for Menu A

$$\bar{X}_a = 22.30$$

$$\sigma_a = 6.88$$

$$n_a = 43$$

Given Data for Menu B

$$\bar{X}_b = 25.91$$

$$\sigma_b = 8.01$$

$$n_b = 50$$

Hypotheses

$$H_0 : \mu_a = \mu_b \quad \text{(the means are equal)}$$

$$H_a : \mu_a \neq \mu_b$$

t-test Calculation

The t-statistic is calculated as follows:

$$t = \frac{(\bar{X}_a - \bar{X}_b)}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}} = \frac{22.30 - 25.91}{\sqrt{\frac{6.88^2}{43} + \frac{8.01^2}{50}}} \approx -2.34$$

Results

Degrees of Freedom

$$df \approx 89$$

Standard Error

$$SE = 1.544$$

p-value

$$p\text{-value} = 0.021$$

EX 5.50

Given that the p-value is 0.057, and greater than 0.05, we cannot reject the null hypothesis.

1. The p-value of 0.057 means there is a 5.7% probability of obtaining a result like the one observed (or more extreme) if the null hypothesis is true.
2. The p-value of 0.057 indicates the probability of obtaining a test value as extreme as 120 or more extreme, given that the null hypothesis is true.
3. The p-value of 0.057 does not represent the probability of a Type I error. The probability of a Type I error is determined by the significance level α, not by the p-value.
4. We cannot accept H_0; we can only fail to reject it. Since the p-value of 0.057 is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis at the 5% significance level.