HW1

2024-04-08

Q0 A: My name is Ruyue Wang and my preferred name is Jasmine.

Q0 B: I have read and understood the entire syllabus.

Q0 C: I took 515, 514 at UW which are related to R coding. I'm familiar with using R studio, but some online sources are needed while trying to solve some problems.

Q0 D: I hope I can get more familiar and confident while using R in the future.

Q2 A: The function generate_data takes a single integer argument n, representing the number of data points to generate. It ensures that n is a positive integer. It then generates a vector sample_vec of length n, where each element is sampled from one of three distributions based on the values in the vector idx_vec, which is sampled from a uniform distribution over {1, 2, 3}. Specifically, elements with value 1 are sampled from a normal distribution with mean 10 and standard deviation 1, elements with value 2 are sampled from a gamma distribution with shape parameter 2 and scale parameter 2, and elements with value 3 are sampled from a chi-squared distribution with 3 degrees of freedom. The function returns the resulting vector sample_vec.

Q2 B:

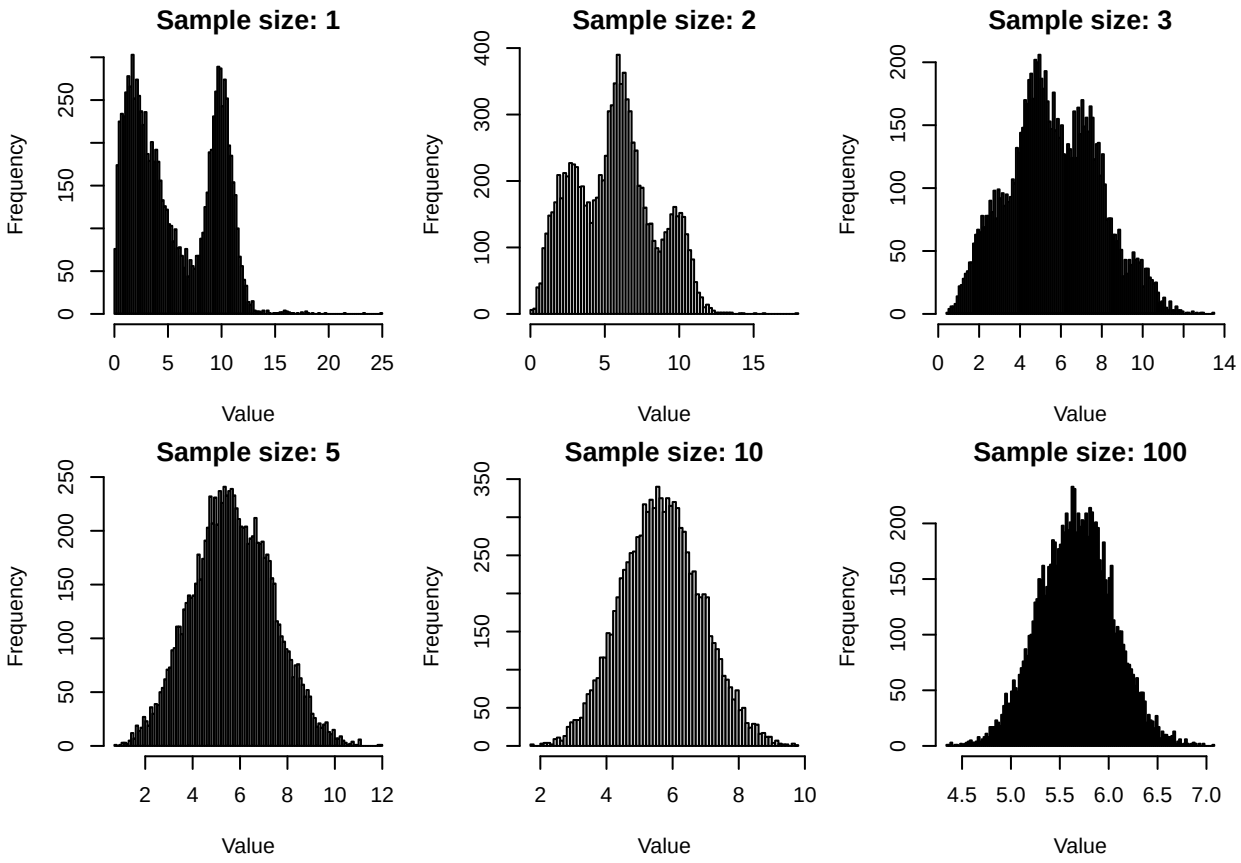
```r
# Load necessary library
library(ggplot2)

# Download and source the generate_data.R script
download.file("https://raw.githubusercontent.com/linnykos/561_s2024_public/main/homework1/generate_data
              destfile = "generate_data.R", method = "curl")
source("generate_data.R")

# Set up the plot parameters
par(mfrow=c(2, 3), mar=c(4, 4, 2, 1))

# Define the sample sizes
sample_sizes <- c(1, 2, 3, 5, 10, 100)

# Perform the simulations and plot the histograms
for (n in sample_sizes) {
  set.seed(123) # For reproducibility, remove if you want true randomness
  means <- replicate(10000, mean(generate_data(n)))
  hist(means, breaks=100, main=paste("Sample size:", n), xlab="Value", ylab="Frequency")
}
```



Q2 C: The histograms visually demonstrate the Central Limit Theorem by showing that as the sample size increases, the distribution of the sample means approaches a normal distribution, regardless of the underlying distribution of the data, which is evidenced by the increasingly bell-shaped curves as the sample size grows from 1 to 100.

Q3 A: The head(df) function output displays the first six rows of the dataset, giving a snapshot of the data including donor ID, age at death, sex, APOE4 status, cognitive status, last CASI score, and Braak stage. The summary(df) function output provides a statistical summary for each variable, including data type, number of missing values, and for numerical data, measures of central tendency and dispersion such as minimum, maximum, quartiles, and mean.

Q3 B: dim() function can be used to print out the dimensionality of df. The class of df is shown as.

```
df <- read.csv("https://raw.githubusercontent.com/linnykos/561_s2024_public/main/homework1/sea-ad.csv")
class(df)
```

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

which shows as the data.frame

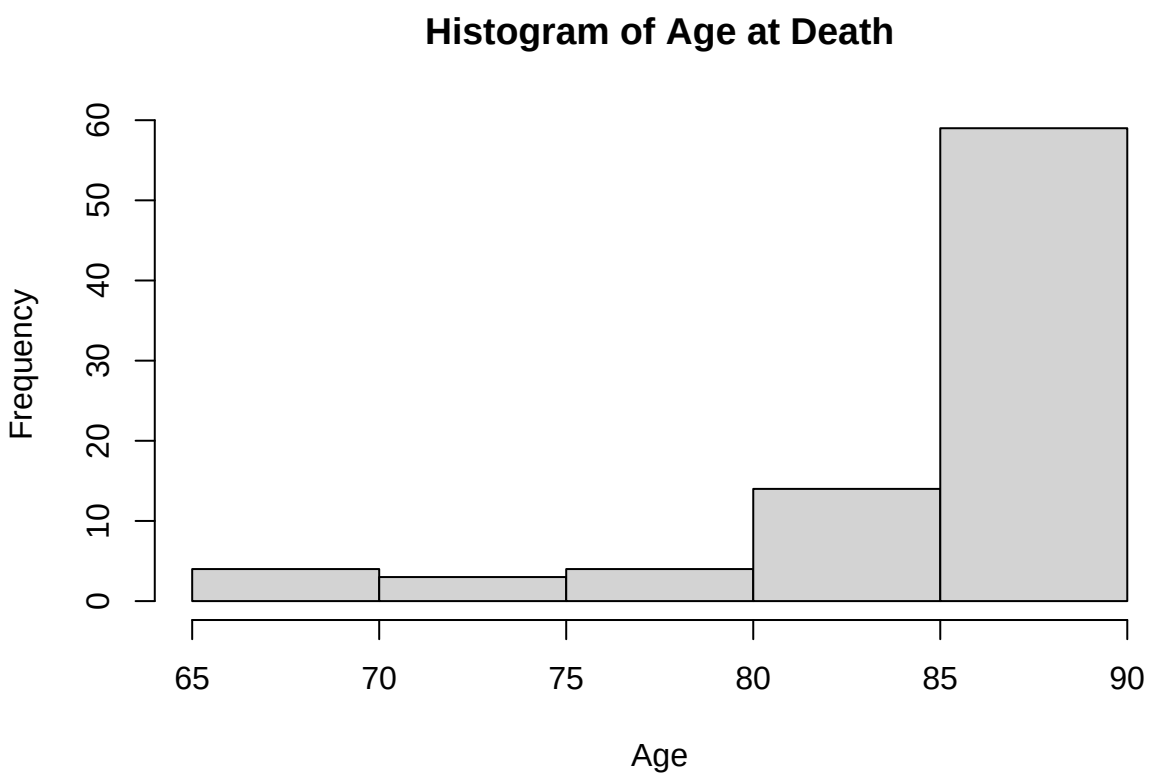
Q3 C: ## Setup Data Frame

Ensure the data frame is correctly loaded and displayed.

```
# Example data frame definition
df <- read.csv("https://raw.githubusercontent.com/linnykos/561_s2024_public/main/homework1/sea-ad.csv")
# Replace "90+" with "90"
df$Age.at.Death <- gsub("90\\+", "90", df$Age.at.Death)
```

```r
# Convert Age.at.Death to numeric
df$Age.at.Death <- as.numeric(df$Age.at.Death)

# Plot a histogram
hist(df$Age.at.Death, main="Histogram of Age at Death", xlab="Age", ylab="Frequency")
```



Q3 D:

```r
# Assuming df is correctly defined and loaded with data
if(nrow(df) > 0) {
  df$Sex <- as.factor(df$Sex)
  df$APOE4.Status <- as.factor(df$APOE4.Status)
  df$Cognitive.Status <- as.factor(df$Cognitive.Status)
  df$Braak <- as.factor(df$Braak)
} else {
  message("Data frame is empty.")
}

# Display the structure of the dataframe to confirm changes
str(df)
```

```
## 'data.frame':    84 obs. of  7 variables:
##  $ Donor.ID        : chr  "H19.33.004" "H20.33.001" "H20.33.002" "H20.33.004" ...
##  $ Age.at.Death    : num  80 82 90 86 90 90 90 90 90 82 ...
##  $ Sex             : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 1 1 1 2 1 ...
##  $ APOE4.Status    : Factor w/ 2 levels "N","Y": 1 1 1 2 1 2 2 1 1 1 ...
##  $ Cognitive.Status: Factor w/ 2 levels "Dementia","No dementia": 2 2 2 1 2 2 1 2 2 2 ...
##  $ Last.CASI.Score : int  85 97 93 80 94 92 79 98 93 NA ...
##  $ Braak           : Factor w/ 6 levels "Braak 0","Braak II",..: 4 4 4 5 4 5 5 3 4 4 ...
```

```r
summary(df)
```

```
##    Donor.ID          Age.at.Death       Sex     APOE4.Status
##  Length:84          Min.   :65.00   Female:51   N:59
##  Class :character   1st Qu.:83.75   Male  :33   Y:25
##  Mode  :character   Median :90.00
##                     Mean   :86.26
##                     3rd Qu.:90.00
##                     Max.   :90.00
##
##     Cognitive.Status Last.CASI.Score       Braak
##  Dementia   :42      Min.   :66.00   Braak 0  : 2
##  No dementia:42      1st Qu.:80.00   Braak II : 4
##                      Median :89.00   Braak III: 6
##                      Mean   :87.32   Braak IV :23
##                      3rd Qu.:95.00   Braak V  :34
##                      Max.   :99.00   Braak VI :15
##                      NA's   :15
```

Q3 E: The summary in Question 3E is more informative because it provides specific counts and distributions for categorical variables and detailed statistics for numerical variables, allowing for a clearer understanding of the dataset's characteristics and any potential imbalances or outliers

Q3 F:

```r
# Display the relationship between Braak stages and Cognitive Status
table(df$Braak, df$Cognitive.Status)
```

```
##
##             Dementia No dementia
##   Braak 0          0           2
##   Braak II         2           2
##   Braak III        2           4
##   Braak IV         4          19
##   Braak V         20          14
##   Braak VI        14           1
```

Q3 G:

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

df <- read.csv("https://raw.githubusercontent.com/linnykos/561_s2024_public/main/homework1/sea-ad.csv")

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr)

# Calculate the quantiles for Last.CASI.Score, excluding NAs
score_quantiles <- quantile(df$Last.CASI.Score, probs = seq(0, 1, by = 0.25), na.rm = TRUE)

# Cut the Last.CASI.Score into quantile-based bins
df$Score.Quantiles <- cut(df$Last.CASI.Score, breaks = score_quantiles, include.lowest = TRUE,
                          labels = c("1st Quartile", "2nd Quartile", "3rd Quartile", "4th Quartile"))

# Create a table to show the relationship between the quantile bins and Cognitive Status
quantile_cognitive_table <- table(df$Score.Quantiles, df$Cognitive.Status)

# Print the table
quantile_cognitive_table
```

```
##
##                 Dementia No dementia
##   1st Quartile        17           1
##   2nd Quartile        10           8
##   3rd Quartile         2          15
##   4th Quartile         3          13
```

The table shows that higher Last.CASI.Scores, which fall into the 3rd and 4th quartiles, are predominantly associated with "No dementia," indicating that higher cognitive scores might be linked to lower instances of dementia. Conversely, the majority of "Dementia" cases are concentrated in the 1st quartile, suggesting that lower scores are strongly correlated with higher occurrences of dementia.