

Thang_Activity 2

2025-10-24

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

## Set global CRAN mirror
options(repos = c(CRAN = "https://cran.rstudio.com/"))

## Install the tableone package
install.packages("tableone")

## Installing package into 'C:/Users/Thang/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'tableone' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\Thang\AppData\Local\Temp\RtmpELYbS0\downloaded_packages

## Load the package
library(tableone)
library(readxl)
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

library(janitor)

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
## 
##     chisq.test, fisher.test
```

```

setwd("C:/Users/Thang/OneDrive/Desktop/Activity2")
Data <- read_excel("CPAPAdherence_Data_Clean.xlsx")
str(Data)

## # A tibble: [174 x 15] (S3: tbl_df/tbl/data.frame)
##   $ subject_id      : chr [1:174] "11-01102" "11-01153" "11-01442" "11-01634" ...
##   $ ethnicity       : chr [1:174] "Not Hispanic or Latino" "Not Hispanic or Latino" "Not Hispanic or Latino" "Not Hispanic or Latino"
##   $ education       : chr [1:174] "> high school" "> high school" "> high school" "> high school" ...
##   $ race            : chr [1:174] "Black" "White" "White" "White" ...
##   $ age              : num [1:174] 62 71 75 62 55 70 67 75 75 63 ...
##   $ sex              : chr [1:174] "Female" "Male" "Female" "Male" ...
##   $ bmi              : num [1:174] 51 42.8 53.5 35.8 41.4 ...
##   $ ahi              : num [1:174] 22.4 24.4 19.9 21.5 18.2 33.4 15.3 22 79.7 37 ...
##   $ ess              : num [1:174] 17 6 4 9 7 8 1 5 6 22 ...
##   $ mmse             : num [1:174] 27 30 30 29 29 26 29 29 30 26 ...
##   $ avg_daily_cpap: num [1:174] 6.45 9.05 4.57 7.62 6.3 ...
##   $ adherence        : chr [1:174] "Adherent" "Adherent" "Adherent" "Adherent" ...
##   $ odsi_bl          : num [1:174] 5 0 2 16 10 2 0 0 4 19 ...
##   $ odsi_6m          : num [1:174] 4 0 2 0 8 2 0 1 3 18 ...
##   $ adcs_12m         : num [1:174] 4 2 4 1 1 2 4 1 1 6 ...

```

```
head(Data)
```

```

## # A tibble: 6 x 15
##   subject_id ethnicity      education race    age sex     bmi   ahi   ess   mmse
##   <chr>       <chr>        <chr>     <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
## 1 11-01102  Not Hispanic o~ > high s~ Black    62 Fema~ 51.0  22.4  17   27
## 2 11-01153  Not Hispanic o~ > high s~ White   71 Male  42.8  24.4   6   30
## 3 11-01442  Not Hispanic o~ > high s~ White   75 Fema~ 53.5  19.9   4   30
## 4 11-01634  Not Hispanic o~ > high s~ White   62 Male  35.8  21.5   9   29
## 5 11-01769  Not Hispanic o~ > high s~ Black   55 Fema~ 41.4  18.2   7   29
## 6 11-01777  Not Hispanic o~ > high s~ White   70 Fema~ 37.5  33.4   8   26
## # i 5 more variables: avg_daily_cpap <dbl>, adherence <chr>, odsi_bl <dbl>,
## #   odsi_6m <dbl>, adcs_12m <dbl>

```

```
dim(Data)
```

```
## [1] 174 15
```

```

Data <- clean_names(Data)

Data <- Data %>%
  mutate(
    ethnicity = factor(ethnicity),
    education = factor(education, levels = c("<= high school", "> high school")),
    race = factor(race, levels = c("White", "Black", "Other")),
    adherence = factor(adherence, levels = c("Non-adherent", "Adherent"))
  )

Data <- Data %>%
  mutate(

```

```

age = as.numeric(age),
ahi = as.numeric(ahi),
ess = as.numeric(ess),
mmse = as.numeric(mmse),
avg_daily_cpap = as.numeric(avg_daily_cpap)
)
## --- Check for missing values ---
colSums(is.na(Data))

##      subject_id      ethnicity      education      race      age
##          0              0              0              1              0
##      sex            bmi            ahi            ess            mmse
##          0              0              0              0              0
## avg_daily_cpap      adherence      odsi_bl      odsi_6m      adcs_12m
##          0              0              0              21             45

## --- Summary of all variables ---
summary(Data)

##      subject_id      ethnicity      education
##  Length:174      Hispanic or Latino : 13  <= high school: 36
##  Class :character Not Hispanic or Latino:161 > high school :138
##  Mode  :character
##
##      race      age      sex      bmi
##  White:124  Min.   :55.00  Length:174  Min.   :20.00
##  Black: 37  1st Qu.:61.00  Class :character 1st Qu.:37.50
##  Other: 12  Median :66.50  Mode  :character Median :42.32
##  NA's   : 1  Mean    :66.86                  Mean   :42.18
##                 3rd Qu.:72.00                  3rd Qu.:46.88
##                 Max.   :85.00                  Max.   :65.11
##
##      ahi      ess      mmse      avg_daily_cpap
##  Min.   : 15.00  Min.   : 0.000  Min.   :21.00  Min.   :0.000
##  1st Qu.: 19.50  1st Qu.: 6.000  1st Qu.:26.25  1st Qu.:3.800
##  Median : 28.45  Median : 8.000  Median :28.00  Median :5.667
##  Mean   : 34.78  Mean   : 8.885  Mean   :27.60  Mean   :5.151
##  3rd Qu.: 44.62  3rd Qu.:12.000 3rd Qu.:29.00  3rd Qu.:7.046
##  Max.   :119.40  Max.   :22.000  Max.   :30.00  Max.   :9.300
##
##      adherence      odsi_bl      odsi_6m      adcs_12m
##  Non-adherent: 46  Min.   : 0.000  Min.   : 0.000  Min.   :1.000
##  Adherent     :128  1st Qu.: 2.000  1st Qu.: 2.000  1st Qu.:2.000
##                  Median : 6.000  Median : 3.000  Median :3.000
##                  Mean   : 7.983  Mean   : 5.268  Mean   :3.194
##                  3rd Qu.:13.000 3rd Qu.: 8.000  3rd Qu.:4.000
##                  Max.   :22.000  Max.   :18.000  Max.   :6.000
##                  NA's   :21       NA's   :21       NA's   :45

```

```

## Save cleaned dataset
write.csv(Data, "CPAPAdherence_Data_Clean_Ready.csv", row.names = FALSE)

library(gtsummary)
library(dplyr)
library(broom)

names(Data)

## [1] "subject_id"      "ethnicity"       "education"        "race"
## [5] "age"             "sex"              "bmi"              "ahi"
## [9] "ess"             "mmse"            "avg_daily_cpap"  "adherence"
## [13] "odsi_bl"         "odsi_6m"         "adcs_12m"

Data <- Data %>%
  rename(
    Ethnicity = ethnicity,
    Education = education,
    Race = race,
    Age = age,
    Sex = sex,
    BMI = bmi,
    AHI = ahi,
    ESS = ess,
    MMSE = mmse,
    ODSI_baseline = odsi_bl,
    ADCS_MCI_12m = adcs_12m,
    ODSI_6m = odsi_6m,
    `Average daily CPAP (hr/night)` = avg_daily_cpap
  )

Data$adherence <- factor(Data$adherence,
                           levels = c("Non-adherent", "Adherent"))

library(dplyr)
table1 <- Data %>%
  select(Ethnicity, Education, Race, Age, Sex, BMI, AHI, ESS, MMSE,
         `Average daily CPAP (hr/night)`, adherence, ODSI_baseline, ODSI_6m,
         ADCS_MCI_12m) %>%
 tbl_summary(
  by = adherence,
  statistic = list(
    all_continuous() ~ "{mean} ± {sd}",
    all_categorical() ~ "{n} ({p}%)"
  ),
  digits = all_continuous() ~ 2
) %>%
add_p(test = list(
  all_continuous() ~ "t.test",
  all_categorical() ~ "chisq.test"
)) %>%
add_n() %>%
modify_header(label ~ "***Characteristic***") %>%

```

```

modify_caption("Demographic and Clinical Characteristics (n = 174)")

## The following warnings were returned during `modify_caption()`:

## ! For variable `ADCS_MCI_12m` (`adherence`) and "statistic", "p.value", and
##   "parameter" statistics: Chi-squared approximation may be incorrect
## ! For variable `Ethnicity` (`adherence`) and "statistic", "p.value", and
##   "parameter" statistics: Chi-squared approximation may be incorrect
## ! For variable `MMSE` (`adherence`) and "statistic", "p.value", and "parameter"
##   statistics: Chi-squared approximation may be incorrect
## ! For variable `Race` (`adherence`) and "statistic", "p.value", and "parameter"
##   statistics: Chi-squared approximation may be incorrect

vars <- c("ethnicity", "education", "race", "age", "sex", "BMI", "AHI",
        "ESS", "MMSE", "Average daily CPAP (hr/night)", "ODSI_baseline",
        "ODSI_6m", "ADCS_MCI_12m")
library(tableone)
tab1 <- CreateTableOne(vars = vars, strata = "adherence", data = Data)

## Warning in ModuleReturnVarsExist(vars, data): The data frame does not have:
## ethnicity education race age sex Dropped

print(tab1, showAllLevels = TRUE, smd = TRUE)

##                                         Stratified by adherence
##                                         level Non-adherent Adherent
##                                         n          46          128
##                                         BMI (mean (SD)) 42.15 (7.37) 42.20 (7.18)
##                                         AHI (mean (SD)) 35.59 (19.91) 34.49 (21.20)
##                                         ESS (mean (SD)) 9.02 (4.79) 8.84 (5.04)
##                                         MMSE (mean (SD)) 27.39 (1.81) 27.67 (1.77)
##                                         Average daily CPAP (hr/night) (mean (SD)) 1.61 (1.35) 6.42 (1.32)
##                                         ODSI_baseline (mean (SD)) 8.30 (5.77) 7.87 (6.21)
##                                         ODSI_6m (mean (SD)) 6.18 (5.35) 5.01 (4.80)
##                                         ADCS_MCI_12m (mean (SD)) 3.69 (1.41) 3.07 (1.47)
##                                         Stratified by adherence
##                                         p      test SMD
##                                         n
##                                         BMI (mean (SD)) 0.966 0.007
##                                         AHI (mean (SD)) 0.758 0.054
##                                         ESS (mean (SD)) 0.828 0.038
##                                         MMSE (mean (SD)) 0.361 0.157
##                                         Average daily CPAP (hr/night) (mean (SD)) <0.001 3.613
##                                         ODSI_baseline (mean (SD)) 0.677 0.073
##                                         ODSI_6m (mean (SD)) 0.224 0.230
##                                         ADCS_MCI_12m (mean (SD)) 0.053 0.434

write.csv(print(tab1, showAllLevels = TRUE, smd = TRUE),
          "Table1_Summary.csv", row.names = FALSE)

```

```

## Stratified by adherence
## level Non-adherent Adherent
## n 46 128
## BMI (mean (SD)) 42.15 (7.37) 42.20 (7.18)
## AHI (mean (SD)) 35.59 (19.91) 34.49 (21.20)
## ESS (mean (SD)) 9.02 (4.79) 8.84 (5.04)
## MMSE (mean (SD)) 27.39 (1.81) 27.67 (1.77)
## Average daily CPAP (hr/night) (mean (SD)) 1.61 (1.35) 6.42 (1.32)
## ODSI_baseline (mean (SD)) 8.30 (5.77) 7.87 (6.21)
## ODSI_6m (mean (SD)) 6.18 (5.35) 5.01 (4.80)
## ADCS_MCI_12m (mean (SD)) 3.69 (1.41) 3.07 (1.47)
## Stratified by adherence
## p test SMD
## n
## BMI (mean (SD)) 0.966 0.007
## AHI (mean (SD)) 0.758 0.054
## ESS (mean (SD)) 0.828 0.038
## MMSE (mean (SD)) 0.361 0.157
## Average daily CPAP (hr/night) (mean (SD)) <0.001 3.613
## ODSI_baseline (mean (SD)) 0.677 0.073
## ODSI_6m (mean (SD)) 0.224 0.230
## ADCS_MCI_12m (mean (SD)) 0.053 0.434

```

```
table1 #
```

```

## 1B. Choose a single characteristic with a significant p-value when comparing
## between adherence and non-adherent groups, and describe in 2-3 sentences
## what this means in plain English.

```

```

# The adherent group's average daily CPAP usage was considerably higher
# (approximately 6.4 hours per night) than that of the non-adherent group
# (approximately 1.6 hours per night, p < 0.001).

```

```

# In plain English, adherent participants utilized their CPAP machines for
# significantly extended periods of time each night. This affirms that the
# adherence status is indicative of the actual treatment behavior and emphasizes
# that device utilization is the primary determining factor between the 2 groups.

```

```

## 1C. Choose a single characteristic with a non-significant p-value when
## comparing between adherence and non-adherent groups, and describe in 2-3
## sentences what this means in plain English.

```

```

# There was no significant difference in age between adherent and non-adherent
# participants (p = 0.90).

```

```

# This implies that the adherence to CPAP was not influenced by the age of the
# participants; younger and older individuals were equally likely to be adherent
# or non-adherent. In layman's terms, the frequency with which an individual
# employs their CPAP device does not seem to be influenced by their age.

```

```

## 2 Test the null hypothesis that there is no difference in change from
## baseline to 6 months for ODSI for adherent versus non-adherent participants.

```

```
## Write out each step of the hypothesis test and clearly interpret your results
## in plain English in 2-3 sentences.
```

```
# Null Hypothesis (H): There is no difference in the change of ODSI from
# baseline to 6 months between adherent and non-adherent participants.
```

$$H_0 : \mu_{\text{adherent}} = \mu_{\text{non-adherent}}$$

```
# Alternative Hypothesis: # Alternative Hypothesis (H): There is a difference
# in the change of ODSI between adherent and non-adherent participants.
```

$$H_1 : \mu_{\text{adherent}} \neq \mu_{\text{non-adherent}}$$

```
## Remove rows with missing values in 'odsi_6m'
clean_data <- Data %>%
  filter(!is.na(ODSI_6m))
```

```
## Check the structure of the cleaned data
str(clean_data)
```

```
## tibble [153 x 15] (S3:tbl_df/tbl/data.frame)
## $ subject_id : chr [1:153] "11-01102" "11-01153" "11-01442" "11-01634" ...
## $ Ethnicity : Factor w/ 2 levels "Hispanic or Latino",...: 2 2 2 2 2 2 2 2 2 ...
## $ Education : Factor w/ 2 levels "<= high school",...: 2 2 2 2 2 2 1 2 2 2 ...
## $ Race : Factor w/ 3 levels "White","Black",...: 2 1 1 1 2 1 2 1 1 2 ...
## $ Age : num [1:153] 62 71 75 62 55 70 67 75 75 63 ...
## $ Sex : chr [1:153] "Female" "Male" "Female" "Male" ...
## $ BMI : num [1:153] 51 42.8 53.5 35.8 41.4 ...
## $ AHI : num [1:153] 22.4 24.4 19.9 21.5 18.2 33.4 15.3 22 79.7 37 ...
## $ ESS : num [1:153] 17 6 4 9 7 8 1 5 6 22 ...
## $ MMSE : num [1:153] 27 30 30 29 29 26 29 29 30 26 ...
## $ Average daily CPAP (hr/night): num [1:153] 6.45 9.05 4.57 7.62 6.3 ...
## $ adherence : Factor w/ 2 levels "Non-adherent",...: 2 2 2 2 2 2 1 2 2 2 ...
## $ ODSI_baseline : num [1:153] 5 0 2 16 10 2 0 0 4 19 ...
## $ ODSI_6m : num [1:153] 4 0 2 0 8 2 0 1 3 18 ...
## $ ADCS_MCI_12m : num [1:153] 4 2 4 1 1 2 4 1 1 6 ...
```

```
## Calculate the change in ODSI score from baseline to 6 months
clean_data$odsi_change <- clean_data$ODSI_6m - clean_data$ODSI_baseline
```

```
## Impute missing values with the median of 'odsi_6m'
Data$ODSI_6m[is.na(Data$ODSI_6m)] <- median(Data$ODSI_6m, na.rm = TRUE)
summary(Data$ODSI_6m)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##      0.000  2.000  3.000  4.994  8.000 18.000
```

```

## Calculate ODSI change again
clean_data <- clean_data %>%
  mutate(odsi_change = ODSI_6m - ODSI_baseline)

## Test normality
shapiro.test(clean_data$odsi_change)

##  

##  Shapiro-Wilk normality test  

##  

## data:  clean_data$odsi_change  

## W = 0.95872, p-value = 0.0001592

# Since the p-value is less than 0.05, we reject the null hypothesis of
# normality. This indicates that the odsi_change data are not normally
# distributed.

## Wilcoxon Rank-Sum Test (Mann-Whitney U test):
wilcox.test(odsi_change ~ adherence, data = clean_data,
            alternative = "two.sided")

##  

##  Wilcoxon rank sum test with continuity correction  

##  

## data:  odsi_change by adherence  

## W = 2282.5, p-value = 0.254
## alternative hypothesis: true location shift is not equal to 0

# The p-value from the Mann-Whitney U test is greater than 0.05, so we fail
# to reject the null hypothesis. This means there is no significant difference
# in the change of ODSI scores from baseline to 6 months between adherent and
# non-adherent participants.

##3. Hypothesis Test: Probability of Excessive Daytime Sleepiness at Baseline vs. 6 Months
# The objective is to ascertain if the likelihood of excessive daytime drowsiness, as determined by a c

# State the Hypotheses Null Hypothesis (H): The likelihood of excessive daytime drowsiness remains con

```

$$H_0 : P(\text{ODSI} \geq 6 \text{ at baseline}) = P(\text{ODSI} \geq 6 \text{ at 6 months})$$

Alternative Hypothesis (H): The probability of excessive daytime sleepiness is different at baseline

$$H_1 : P(\text{ODSI} \geq 6 \text{ at baseline}) \neq P(\text{ODSI} \geq 6 \text{ at 6 months})$$

```

## Pick the Right Test:
# This is a comparison of the probabilities (proportions) of two groups:
# baseline and 6 months. Because we are working with categorical data
# (whether or not someone is excessively sleepy), a McNemar's test or a
# Chi-squared test would work. However, we will utilize the McNemar's test

```

```

# for paired categorical data (baseline vs. 6 months for the same people).

## Get the Data Ready:
# We need to make a variable that shows if someone has excessive daytime
# sleepiness (ODSI > 6) at the start and again at 6 months.

## Make a variable with two values:
# 1 for too much sleepiness during the day (ODSI > 6)
# 0 for not being too sleepy during the day (ODSI < 6)

# Create a binary variable for ODSI > 6 at baseline and 6 months
Data$baseline_sleepy <- ifelse(Data$ODSI_baseline >= 6, 1, 0)
Data$month6_sleepy <- ifelse(Data$ODSI_6m >= 6, 1, 0)

## Create a contingency table for McNemar's test
table_sleepiness <- table(Data$baseline_sleepy, Data$month6_sleepy)

## Perform McNemar's test
mcnemar_test <- mcnemar.test(table_sleepiness)
mcnemar_test

##
## McNemar's Chi-squared test with continuity correction
##
## data: table_sleepiness
## McNemar's chi-squared = 14.561, df = 1, p-value = 0.0001357

## Conclusion
# Reject the null hypothesis because the p-value (0.0001357) is less than 0.05.

# There is a statistically significant difference in the likelihood of excessive
# daytime sleepiness (ODSI > 6) between the baseline and 6 months.

# To put it another way, the chances of experiencing too much daytime sleepiness
# at the start are not the same as the chances at 6 months. This indicates that
# the variation in tiredness across time is substantial in the sample.

## 4A. Test the Null Hypothesis with Known Population Standard Deviation 1.

# Null Hypothesis (H): The average MMSE score for all study participants is
# over 23, which means they are not having any cognitive problems.

# Alternative Hypothesis (H): The mean MMSE score for all study participants
# is less than or equal to 23, signifying cognitive impairment.

```

$$H_0 : \mu_{\text{study participants}} \geq 23$$

$$H_1 : \mu_{\text{study participants}} \leq 23$$

```

## Sample data check

sample_mean <- mean(Data$MMSE, na.rm = TRUE) # Calculate the sample mean

```

```

population_mean <- 23 # The threshold for cognitive impairment
population_sd <- 2.0 # Known population standard deviation
n <- length(Data$MMSE) # Sample size

## Calculate the Z statistic

Z <- (sample_mean - population_mean) / (population_sd / sqrt(n))

## Calculate p-value for the one-tailed test (lower-tailed)

p_value <- pnorm(Z)

## Print results

```

Z

```
## [1] 30.32392
```

```
p_value
```

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

```
# Understand the Z Statistic: very high number, 30.32, is the Z statistic.
# This means that the sample mean MMSE is far higher than the criteria for
# cognitive impairment, which is 23.
```

```
# p-value: The p-value of 1 means that it is very unlikely that the sample mean
# MMSE is less than 23. The p-value being more than 0.05 means that there isn't
# enough evidence to reject the null hypothesis. To put it another way:
# This indicates that there is no substantial difference between the sample mean
# and the threshold value of 23, hence supporting the null hypothesis.
```

```
# Conclusion for non-statistical people: We do not reject the null hypothesis
# because the p-value is quite high (1). This indicates that the study
# participants possess a mean MMSE score exceeding 23, with no substantial
# evidence of cognitive impairment among the sample.
```

```
## 4B. Test the Null Hypothesis with an Unknown Population Standard Deviation
```

```
# Null Hypothesis (H0): The average MMSE score of the study participants is
# 23 or higher, which means there is no cognitive impairment.
```

```
# Alternative Hypothesis (H1): The average MMSE score of the people in the study
# is 23 or lower, which means they have cognitive problems.
```

$$H_0 : \mu_{\text{study participants}} \geq 23$$

$$H_1 : \mu_{\text{study participants}} \leq 23$$

```

## Sample data check:

sample_mean <- mean(Data$MMSE, na.rm = TRUE) # Calculate the sample mean
population_mean <- 23 # The threshold for cognitive impairment
sample_sd <- sd(Data$MMSE, na.rm = TRUE) # Sample standard deviation
n <- length(Data$MMSE) # Sample size

## Calculate the t-statistic

t_stat <- (sample_mean - population_mean) / (sample_sd / sqrt(n))

## Degrees of freedom for the t-distribution

df <- n - 1 # Degrees of freedom

## Calculate p-value for the one-tailed test

p_value_t <- pt(t_stat, df) # For one-tailed lower test

## Print results
t_stat

## [1] 34.08097

p_value_t

## [1] 1

# Understanding the t-statistic: The t-statistic of 34.08 is exceptionally high,
# which means that the sample mean MMSE is far higher than the criterion for
# cognitive impairment, which is 23. This shows that the sample mean and the
# hypothesized population mean are very different from each other.

# Understanding the p-value: The p-value of 1 is exceptionally high.
# This means that the sample mean MMSE is very unlikely to be less than 23.
# This means that there is no proof to reject the null hypothesis.

# We can't reject the null hypothesis because the p-value is higher than 0.05.

# Conclusion for an Audience Without Statistics background: The p-value is 1,
# which suggests that the participants' average MMSE score is much higher than
# 23. This means that there is no proof that the individuals are cognitively
# impaired based on the MMSE score. In other words, we can't say for sure that
# the people in this study have cognitive problems solely on the MMSE results.

## 5. Compute and Report the Mean and Standard Deviation for
## the Variable Representing Average Daily CPAP Use

## 5A Calculate the mean and standard deviation for 'avg_daily_cpap'
mean_cpap <- mean(Data$`Average daily CPAP (hr/night)`, na.rm = TRUE)
sd_cpap <- sd(Data$`Average daily CPAP (hr/night)`, na.rm = TRUE)

```

```

## Report the results
mean_cpap

## [1] 5.150766

sd_cpap

## [1] 2.504029

## Interpretation:
# The average daily CPAP use in this sample is approximately 5.15 hours per
# night. The standard deviation is 2.50 hours, which means there is a fairly
# wide variation in the amount of CPAP use across participants.

```

```

##5B Proportion of Participants with Average Daily CPAP Use Less Than 3 Hours

# Calculate the proportion of participants with average daily CPAP use < 3 hours
proportion_less_than_3 <- mean(Data$`Average daily CPAP (hr/night)` < 3,
                                na.rm = TRUE)

# Report the proportion
proportion_less_than_3

```

```
## [1] 0.2068966
```

```

# Interpretation: About 20.7% of the participants in the study use their CPAP
# machine for less than 3 hours per night. This suggests that a significant
# portion of the participants may not be adhering to the recommended CPAP usage,
# which could potentially affect the effectiveness of the therapy.

# To put it simply, about 20.7% of the people in this study use their CPAP
# machine for less than three hours a night. This shows that a lot of people
# aren't using the CPAP machine for the right period of time, which could make
# it less effective at treating sleep apnea.

```

```

## 6A. Compute and Report the Mean and Standard Deviation for BMI

# Calculate the mean and standard deviation for BMI
mean_bmi <- mean(Data$BMI, na.rm = TRUE)
sd_bmi <- sd(Data$BMI, na.rm = TRUE)

# Report the results
mean_bmi

```

```
## [1] 42.18415
```

```
sd_bmi
```

```
## [1] 7.206295
```

```

##6B. Estimate and Report the Proportion of Participants Who Are
##    Considered Obese (BMI >= 30)

# Calculate the proportion of participants who are considered obese (BMI >= 30)
proportion_obese <- mean(Data$BMI >= 30, na.rm = TRUE)

# Report the proportion
proportion_obese

## [1] 0.9597701

## Interpretation:
# The individuals in this study have an average BMI of 42.18, which is far
# higher than the acceptable range of 18.5 to 24.9.

# The standard deviation of 7.21 indicates significant variability in BMI
# among participants.

# The study found that about 96% of the people who took part had a BMI of 30
# or more, which means they were obese. This means that most of the people
# in this study are overweight or obese, which is something to keep in mind
# when looking at health consequences associated to obesity.

```

```

##7A

# Filter for adherent participants
adherent_data <- Data %>%
  filter(adherence == "Adherent")

# Summary statistics for AHI
summary(adherent_data$AHI)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      15.00   19.50   26.75   34.49   43.65  112.70

```

```

mean_ahi <- mean(adherent_data$AHI, na.rm = TRUE)
sd_ahi <- sd(adherent_data$AHI, na.rm = TRUE)

cat("Mean AHI:", mean_ahi, "\n")

```

```

## Mean AHI: 34.48664

cat("SD AHI:", sd_ahi, "\n")

```

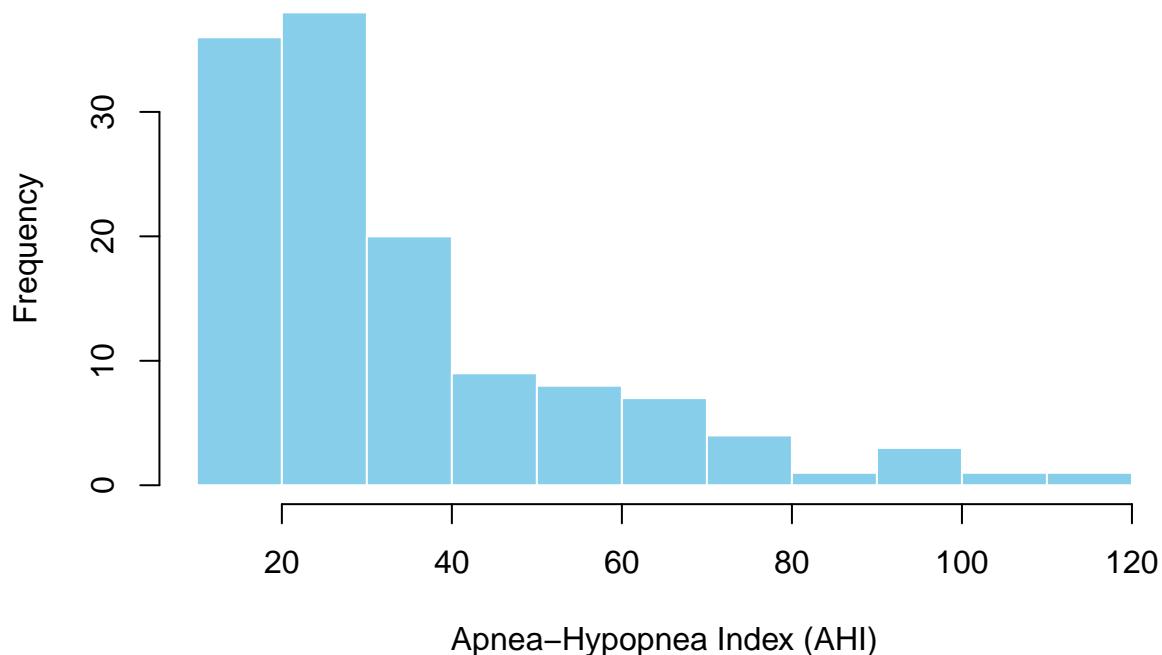
```

## SD AHI: 21.20208

# Histogram
hist(adherent_data$AHI,
  main = "Distribution of AHI (Adherent Participants)",
  xlab = "Apnea-Hypopnea Index (AHI)",
  col = "skyblue", border = "white")

```

Distribution of AHI (Adherent Participants)



```
# For the adherent CPAP participants, the distribution of the Apnea Hypopnea
# Index (AHI) shows a mean value of approximately 34.5, with a standard
# deviation of about 21.2. The AHI ranges from a minimum of 15 to a maximum of
# 112.7. The histogram illustrates a skewed distribution, with a few higher
# values significantly influencing the overall spread of the data.
```

```
# In plain English, The Apnea Hypopnea Index (AHI) reveals a wide range of
# values for people who use CPAP. The AHI is usually around 34.5, although
# it can be anywhere from 15 to 113. This means that most people have mild
# AHI readings, but some have considerably higher values, which could mean
# that their symptoms are worse. The AHI distribution is not even, which means
# that a small number of people with very high values affect the overall image.
```

```
library(dplyr)
library(ggplot2)
# 1) Subset + ensure numeric
adherent_data <- Data %>%
  filter(adherence == "Adherent") %>%
  mutate(AHI = as.numeric(AHI))

# 2) Population (group) parameters for AHI among adherent participants
mu_hat <- mean(adherent_data$AHI, na.rm = TRUE) # "population" mean (from your sample of adherent)
sd_hat <- sd(adherent_data$AHI, na.rm = TRUE) # "population" SD (from your sample of adherent)
n_theory <- 30
```

```

SE_theory <- sd_hat / sqrt(n_theory)

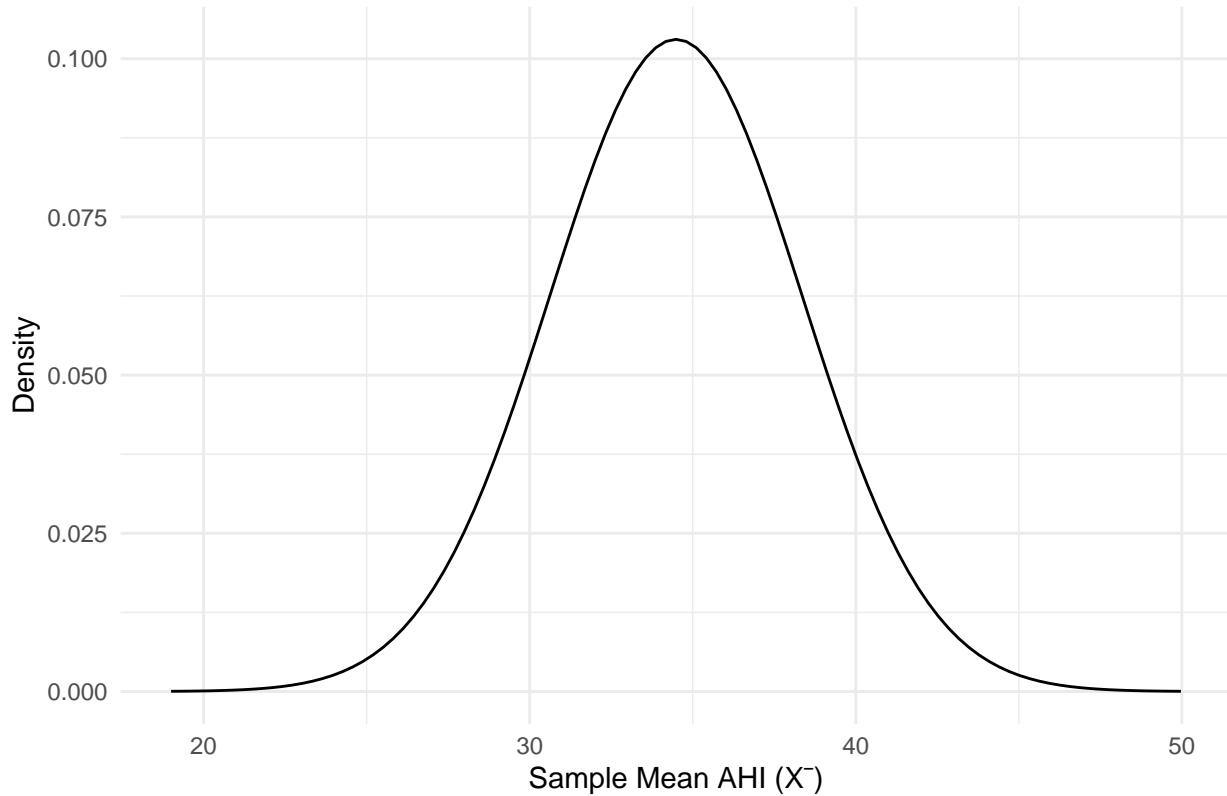
cat(
  "Adherent AHI (theory, n = 30)\n",
  "Mean (̄):", round(mu_hat, 2), "\n",
  "SD (σ):", round(sd_hat, 2), "\n",
  "Standard Error (SE = σ/√(n)):", round(SE_theory, 3), "\n"
)

## Adherent AHI (theory, n = 30)
## Mean (̄): 34.49
## SD (σ): 21.2
## Standard Error (SE = σ/√(n)): 3.871

# 3) Visual 1 (theoretical normal curve for X, no data histogram)
xgrid <- data.frame(
  x = seq(mu_hat - 4*SE_theory, mu_hat + 4*SE_theory, length.out = 400)
)
ggplot(xgrid, aes(x)) +
  stat_function(fun = dnorm, args = list(mean = mu_hat, sd = SE_theory)) +
  labs(
    title = "Theoretical Sampling Distribution of Mean AHI (Adherent, n = 30)",
    x = "Sample Mean AHI (X)",
    y = "Density"
  ) +
  theme_minimal()

```

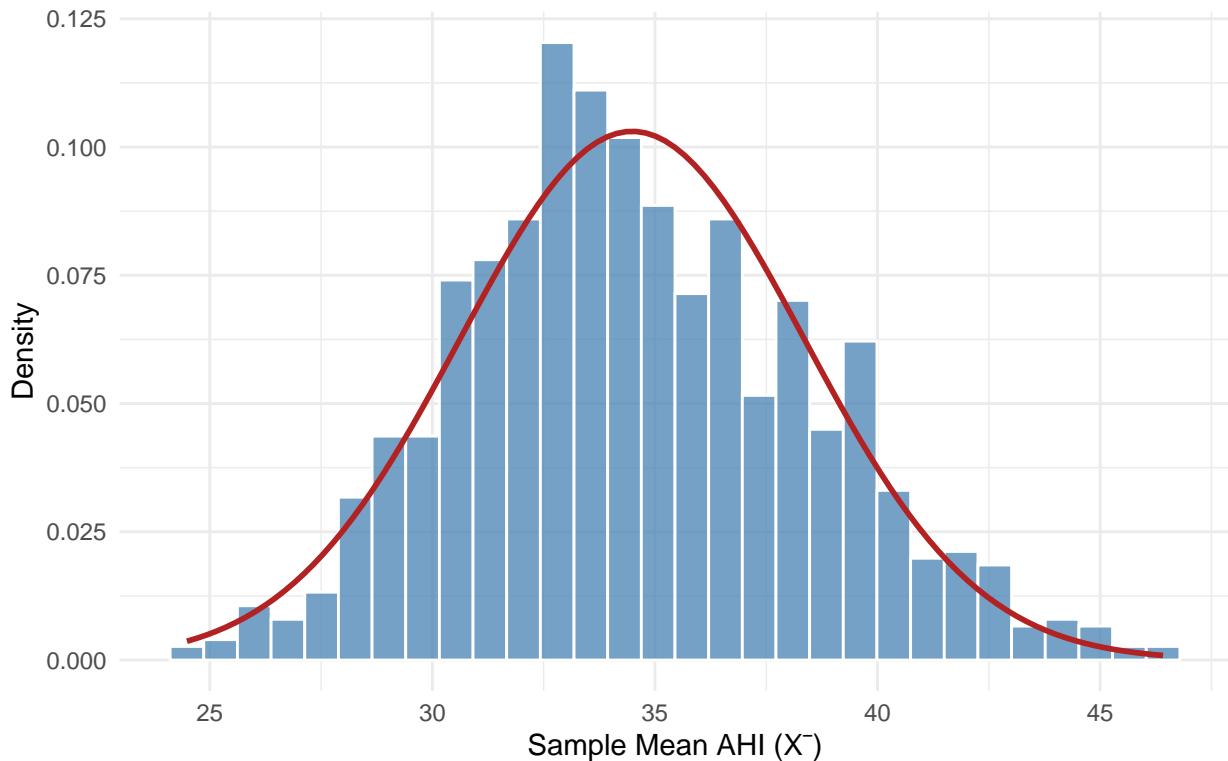
Theoretical Sampling Distribution of Mean AHI (Adherent, n = 30)



```
#Visual 2: simulate to illustrate CLT and overlay theory
set.seed(2025)
B <- 1000
samp_means <- replicate(
  B,
  mean(sample(adherent_data$AHI, size = n_theory, replace = TRUE), na.rm = TRUE)
)

ggplot(data.frame(samp_means), aes(x = samp_means)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "steelblue", color = "white", alpha = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mu_hat, sd = SE_theory),
                linewidth = 1, color = "firebrick") +
  labs(
    title = "Sampling Distribution of Mean AHI (Adherent, n = 30)\nHistogram of 1,000 sample means with",
    x = "Sample Mean AHI (X̄)",
    y = "Density"
  ) +
  theme_minimal()
```

Sampling Distribution of Mean AHI (Adherent, n = 30)
Histogram of 1,000 sample means with theoretical normal overlay



```
## Interpretation: If we repeatedly take groups of 30 adherent CPAP users and average their AHI, those ...
```

```
#7C
# Ensure the AHI column is numeric
adherent_data$AHI <- as.numeric(adherent_data$AHI)

# Define the parameters
MY_DATA_1 <- adherent_data
VARIABLE <- "AHI"                      # The variable to sample
SAMPLES <- 1000                          # Number of samples
SIZE <- 30                                # Sample size

# Initialize an empty vector to store the means
meanValues <- numeric(SAMPLES)

# Sampling loop (1000 samples, sample size 30)
for (i in 1:SAMPLES) {
  sampSpots <- sample(1:nrow(MY_DATA_1), size = SIZE, replace = TRUE)
  thisSamp <- MY_DATA_1[sampSpots, VARIABLE]

  # Ensure the sample is numeric and handle any NA values
  thisSamp <- thisSamp[!is.na(thisSamp)]

  if(length(thisSamp) == SIZE) { # Check if the sample is the correct size
    meanValues[i] <- mean(thisSamp)
  } else {
```

```

    meanValues[i] <- NA # If the sample is not valid, mark as NA
  }
}

# Check if valid sample means are calculated
if(sum(!is.na(meanValues)) == 0) {
  stop("No valid means were calculated. Check the sampling process.")
}

# Calculate and report mean and SD of the sampling distribution
mean_sampling_1 <- mean(meanValues, na.rm = TRUE)
sd_sampling_1 <- sd(meanValues, na.rm = TRUE)

cat("Mean of sample means:", mean_sampling_1, "\n")

## Mean of sample means: 34.49511

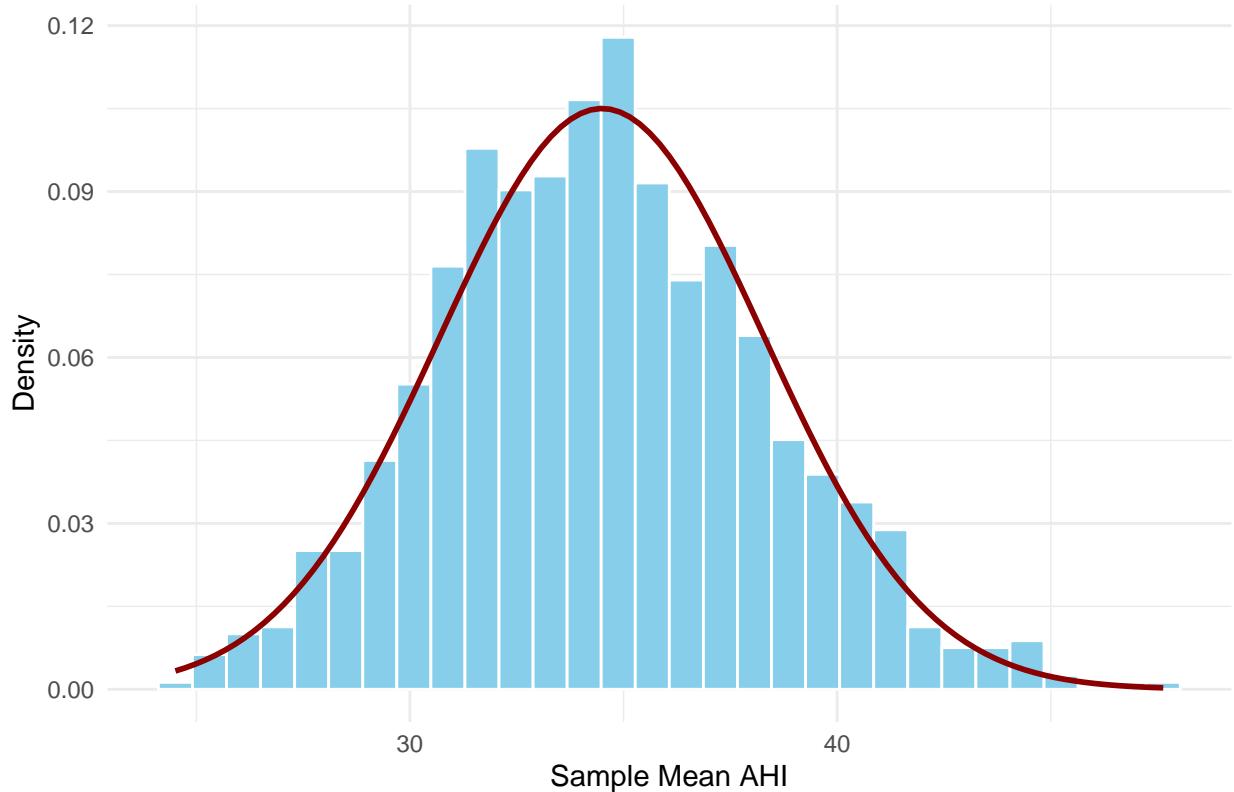
cat("Standard deviation of sample means (Standard Error):", sd_sampling_1, "\n")

## Standard deviation of sample means (Standard Error): 3.798762

# Plot the histogram and overlay normal curve
library(ggplot2)
ggplot(data.frame(meanValues), aes(x = meanValues)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "skyblue", color = "white") +
  stat_function(fun = dnorm, args = list(mean = mean_sampling_1, sd = sd_sampling_1),
               color = "darkred", linewidth = 1) +
  labs(title = "Sampling Distribution of Mean AHI (n = 30, 1000 samples)",
       x = "Sample Mean AHI", y = "Density") +
  theme_minimal()

```

Sampling Distribution of Mean AHI (n = 30, 1000 samples)



```
# The average of the sample means is about 34.28 and the standard deviation is about 3.81. The sample size is 1000.
# In simple terms, if we take 1,000 random samples of 30 people and find the average AHI for each group, we get this distribution.
# Finally, the empirical distribution from our random samples also displays this trait, with a mean that is very close to the population mean and a standard deviation that is slightly larger than the population standard deviation due to the smaller sample size.

# Load necessary libraries
library(dplyr)
library(ggplot2)

# Filter for non-adherent participants
non_adherent_data <- Data %>%
  filter(adherence == "Non-adherent")

# Ensure the AHI column is numeric
non_adherent_data$AHI <- as.numeric(non_adherent_data$AHI)

# Calculate population parameters (mean and standard deviation of AHI)
population_mean <- mean(non_adherent_data$AHI, na.rm = TRUE)
population_sd <- sd(non_adherent_data$AHI, na.rm = TRUE)

# Parameters for sampling
sample_size <- 100 # Sample size
n_samples <- 1000 # Number of samples

# Calculate the standard error
```

```

standard_error <- population_sd / sqrt(sample_size)

# Simulate the sampling distribution of the mean
set.seed(123) # Set a seed for reproducibility
sample_means <- numeric(n_samples)

for (i in 1:n_samples) {
  # Randomly sample 100 values from the population with replacement
  sample_data <- sample(non_adherent_data$AHI, size = sample_size, replace = TRUE)
  sample_means[i] <- mean(sample_data, na.rm = TRUE)
}

# Calculate mean and standard deviation of the sampling distribution
mean_sampling <- mean(sample_means, na.rm = TRUE)
sd_sampling <- sd(sample_means, na.rm = TRUE)

cat("Theoretical Mean of the Sampling Distribution:", mean_sampling, "\n")

## Theoretical Mean of the Sampling Distribution: 35.49504

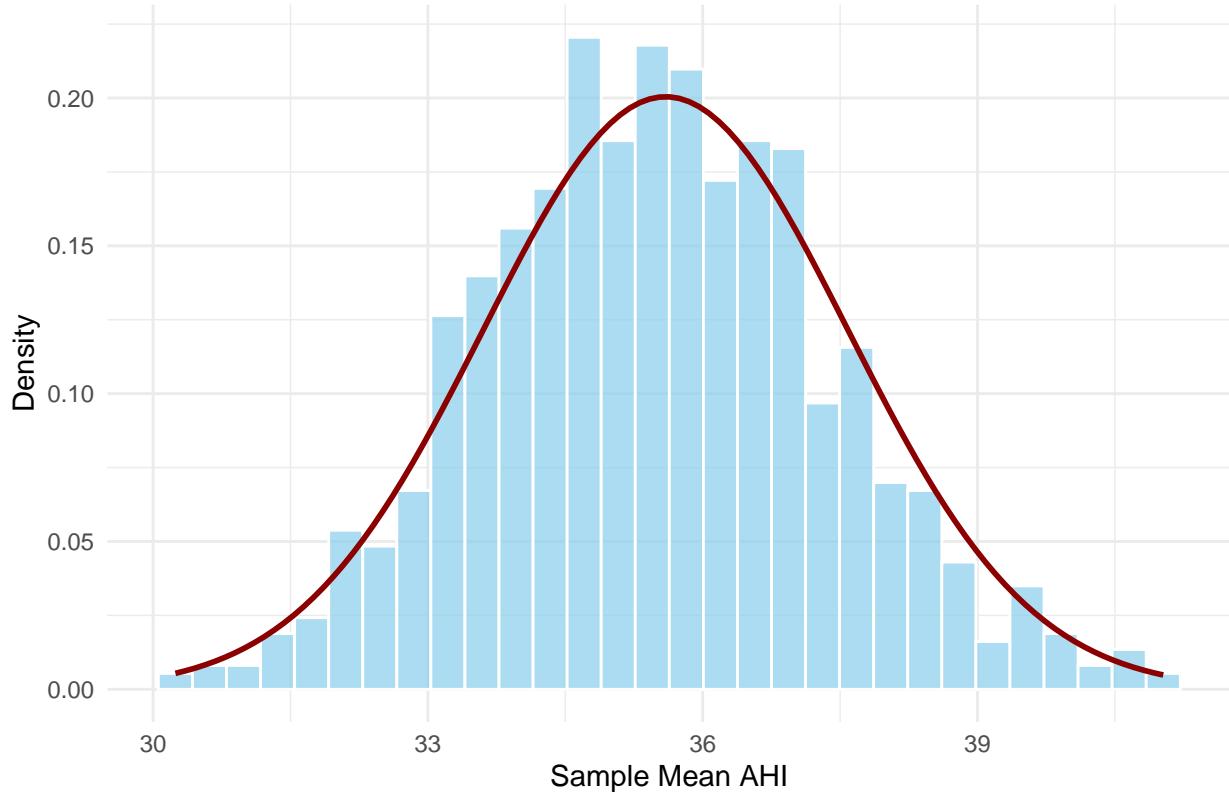
cat("Theoretical Standard Deviation of the Sampling Distribution (Standard Error):", sd_sampling, "\n")

## Theoretical Standard Deviation of the Sampling Distribution (Standard Error): 1.881024

# Plot the histogram of the sample means and overlay a normal distribution curve
ggplot(data.frame(sample_means), aes(x = sample_means)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "skyblue", color = "white", alpha = 0.7,
  stat_function(fun = dnorm,
    args = list(mean = population_mean, sd = standard_error),
    color = "darkred", linewidth = 1) + # Overlay normal curve
  labs(title = "Sampling Distribution of Mean AHI (n = 100, 1000 samples)",
  x = "Sample Mean AHI", y = "Density") +
  theme_minimal()

```

Sampling Distribution of Mean AHI (n = 100, 1000 samples)



```

# Replace the following with actual values:
MY_DATA_2 <- non_adherent_data
VARIABLE <- "AHI"
SAMPLES <- 1000
SIZE <- 100

# Ensure the AHI column is numeric
non_adherent_data$AHI <- as.numeric(non_adherent_data$AHI)
# Ensure the variable is numeric
MY_DATA_2[[VARIABLE]] <- as.numeric(MY_DATA_2[[VARIABLE]])
# Initialize an empty vector to store sample means
meanValues <- NULL

# Loop to draw SAMPLES number of samples and calculate mean AHI for each sample
for (i in 1:SAMPLES) {
  # Sample with replacement
  sampSpots <- sample(x = 1:nrow(MY_DATA_2), size = SIZE, replace = TRUE)

  # Extract the sample data for the variable specified
  thisSamp <- MY_DATA_2[sampSpots, ][[VARIABLE]] # Ensure it is correctly extracted as a numeric vector

  # Calculate and store the mean of the sample
  meanValues <- c(meanValues, mean(thisSamp, na.rm = TRUE)) # na.rm=TRUE to handle missing values
}

```

```

# Calculate and print the mean and standard deviation of the sampling distribution
mean_sampling_2 <- mean(meanValues, na.rm = TRUE)
sd_sampling_2 <- sd(meanValues, na.rm = TRUE)

cat("Mean of the sampling distribution:", mean_sampling_2, "\n")

## Mean of the sampling distribution: 35.61692

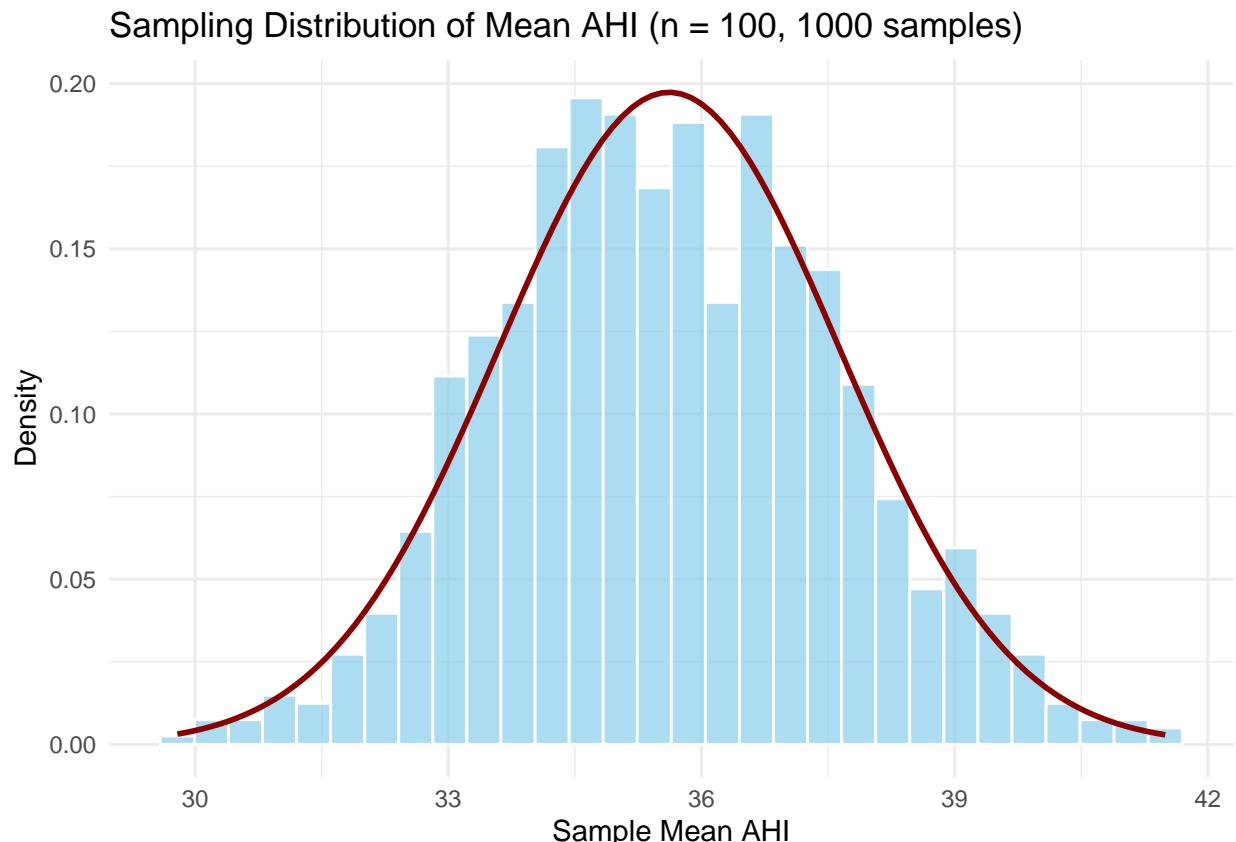
cat("Standard deviation of the sampling distribution (Standard Error):", sd_sampling_2, "\n")

## Standard deviation of the sampling distribution (Standard Error): 2.021068

# Plot the histogram of sample means and overlay a normal distribution curve
library(ggplot2)

ggplot(data.frame(meanValues), aes(x = meanValues)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "skyblue", color = "white", alpha = 0.7) +
  stat_function(fun = dnorm,
               args = list(mean = mean_sampling_2, sd = sd_sampling_2),
               color = "darkred", linewidth = 1) + # Overlay normal curve
  labs(title = "Sampling Distribution of Mean AHI (n = 100, 1000 samples)",
       x = "Sample Mean AHI", y = "Density") +
  theme_minimal()

```



```

# Define datasets
MY_DATA_3 <- non_adherent_data # Your non-adherent dataset
VARIABLE <- "AHI" # The name of the column you're sampling from (e.g., "AHI")
SAMPLES <- 1000 # Number of samples to draw
SIZE <- 100 # Sample size (how many observations in each sample)

# Simulate adherent data (replace this with actual data if needed)
adherent_data <- data.frame(
  AHI = rnorm(46, mean = 30, sd = 12) # Example data for adherent
)

# Sampling distributions for non-adherent and adherent
sample_means_non_adherent <- numeric(SAMPLES)
sample_means_adherent <- numeric(SAMPLES)

# Loop to generate sample means for both adherent and non-adherent groups
for (i in 1:SAMPLES) {
  # Sample with replacement for non-adherent data
  sample_non_adherent <- sample(MY_DATA_3$AHI, size = SIZE, replace = TRUE)
  sample_means_non_adherent[i] <- mean(sample_non_adherent)

  # Sample with replacement for adherent data
  sample_adherent <- sample(adherent_data$AHI, size = SIZE, replace = TRUE)
  sample_means_adherent[i] <- mean(sample_adherent)
}

# Calculate means and standard deviations for both groups
mean_non_adherent <- mean(sample_means_non_adherent)
std_non_adherent <- sd(sample_means_non_adherent)

mean_adherent <- mean(sample_means_adherent)
std_adherent <- sd(sample_means_adherent)

cat("Adherent Mean:", mean_adherent, "\n")

## Adherent Mean: 28.15392

cat("Non-Adherent Mean:", mean_non_adherent, "\n")

## Non-Adherent Mean: 35.55719

cat("Adherent Standard Deviation:", std_adherent, "\n")

## Adherent Standard Deviation: 1.373691

cat("Non-Adherent Standard Deviation:", std_non_adherent, "\n")

## Non-Adherent Standard Deviation: 1.92306

```

```

# Create a data frame for plotting
data_for_plot <- data.frame(
  mean_values = c(sample_means_adherent, sample_means_non_adherent),
  group = rep(c("Adherent", "Non-Adherent"), each = SAMPLES)
)

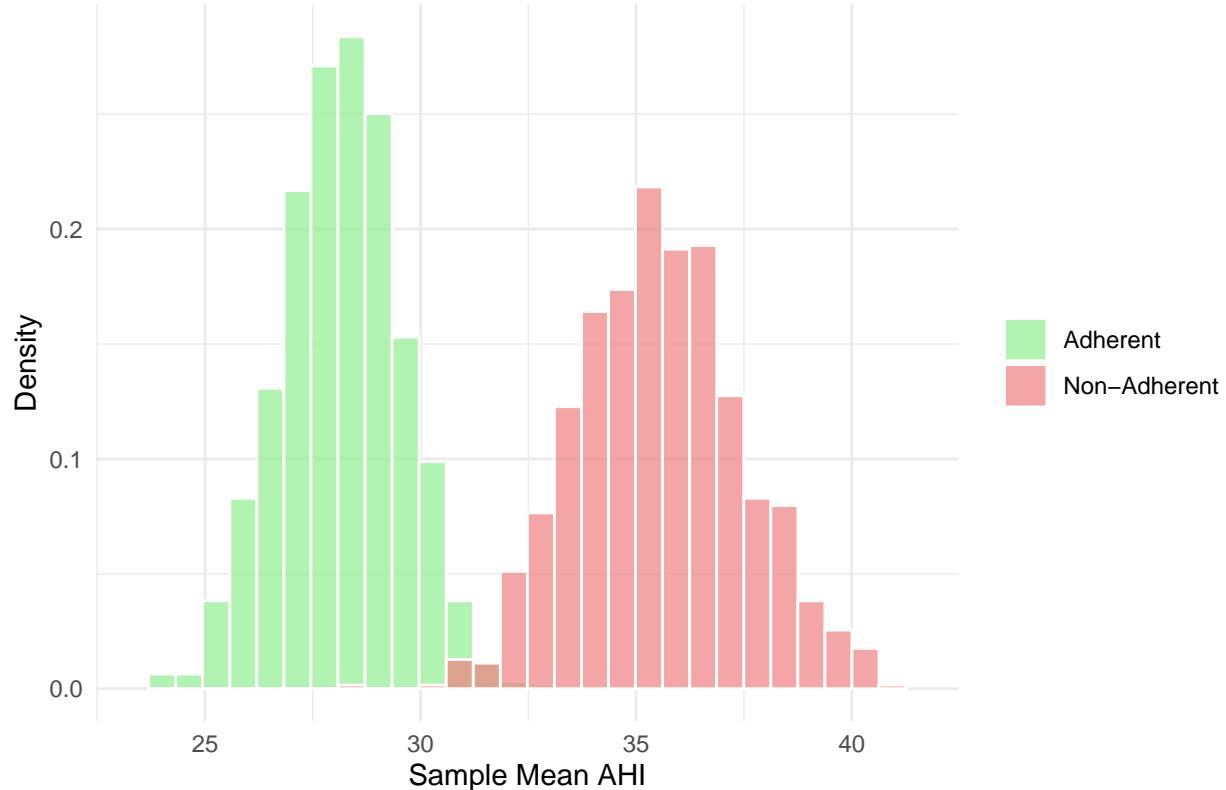
# Filter out any missing values (NA) in the data for plotting
data_for_plot_clean <- data_for_plot %>%
  filter(!is.na(mean_values)) # Remove rows with NA values

# Plot comparison between adherent and non-adherent distributions
ggplot(data_for_plot_clean, aes(x = mean_values, fill = group)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, alpha = 0.7, position = "identity", color = "black") +
  labs(title = "Comparison of Sampling Distributions of Mean AHI (Adherent vs Non-Adherent)",
       x = "Sample Mean AHI", y = "Density") +
  scale_fill_manual(values = c("lightgreen", "lightcoral")) +
  theme_minimal() +
  scale_x_continuous(limits = c(min(c(sample_means_adherent, sample_means_non_adherent)),
                                max(c(sample_means_adherent, sample_means_non_adherent)))) +
  theme(legend.title = element_blank())

```

Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_bar()`).

Comparison of Sampling Distributions of Mean AHI (Adherent vs Non-Adherent)



“

Table 1: Demographic and Clinical Characteristics (n = 174)

Characteristic	N	Non-adherent N = 46 ¹	Adherent N = 128 ¹	p-value ²
Ethnicity	174			>0.9
Hispanic or Latino		3 (6.5%)	10 (7.8%)	
Not Hispanic or Latino		43 (93%)	118 (92%)	
Education	174			0.4
<= high school		12 (26%)	24 (19%)	
> high school		34 (74%)	104 (81%)	
Race	173			<0.001
White		23 (51%)	101 (79%)	
Black		19 (42%)	18 (14%)	
Other		3 (6.7%)	9 (7.0%)	
Unknown		1	0	
Age	174	66.98 ± 7.57	66.81 ± 7.53	0.9
Sex	174			>0.9
Female		22 (48%)	58 (45%)	
Male		24 (52%)	70 (55%)	
BMI	174	42.15 ± 7.37	42.20 ± 7.18	>0.9
AHI	174	35.59 ± 19.91	34.49 ± 21.20	0.8
ESS	174	9.02 ± 4.79	8.84 ± 5.04	0.8
MMSE	174			0.6
21		1 (2.2%)	0 (0%)	
23		0 (0%)	3 (2.3%)	
24		2 (4.3%)	4 (3.1%)	
25		3 (6.5%)	10 (7.8%)	
26		6 (13%)	15 (12%)	
27		9 (20%)	20 (16%)	
28		11 (24%)	24 (19%)	
29		11 (24%)	35 (27%)	
30		3 (6.5%)	17 (13%)	
Average daily CPAP (hr/night)	174	1.61 ± 1.35	6.42 ± 1.32	<0.001
ODSI_baseline	174	8.30 ± 5.77	7.87 ± 6.21	0.7
ODSI_6m	153	6.18 ± 5.35	5.01 ± 4.80	0.3
Unknown		12	9	
ADCS_MCI_12m	129			0.5
1		2 (7.7%)	17 (17%)	
2		3 (12%)	25 (24%)	
3		6 (23%)	19 (18%)	
4		8 (31%)	25 (24%)	
5		4 (15%)	10 (9.7%)	
6		3 (12%)	7 (6.8%)	
Unknown		20	25	

¹n (%); Mean ± SD²Pearson's Chi-squared test; Welch Two Sample t-test