ACTIVITY2

Madeleine Schoderbek

2025-10-27

R Markdown

```
library(readxl)
Cpap <- read_excel("CPAPAdherence_Data_Clean.xlsx")
head(Cpap)
```

```
## # A tibble: 6 × 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>
```

#1

```
# Clean variable names to be consistent
Cpap <- clean_names(Cpap)

# Convert categorical variables to factors
Cpap <- Cpap %>%
  mutate(
    adherence = as.factor(adherence),
    sex = as.factor(sex),
    race = as.factor(race),
    ethnicity = as.factor(ethnicity),
    education = as.factor(education)
  )

# Check structure
str(Cpap)
```

```
## tibble [174 × 15] (S3: tbl_df/tbl/data.frame)
##  $ subject_id    : chr [1:174] "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 2 ...
##  $ education     : Factor w/ 2 levels "<= high school",..: 2 2 2 2 2 2 1 2
2 2 ...
##  $ race          : Factor w/ 4 levels "Black","NA","Other",..: 1 4 4 4 1 4
1 4 4 1 ...
##  $ age           : num [1:174] 62 71 75 62 55 70 67 75 75 63 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 1 1 2 2 2
...
##  $ 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     : Factor w/ 2 levels "Adherent","Non-adherent": 1 1 1 1 1
1 2 1 1 1 ...
##  $ 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 ...
```

```
render_cont <- function(x) {
  with(stats.apply.rounding(stats.default(x), digits = 2),
       c("", sprintf("%0.3f", p)))
}

render_cat <- function(x) {
  with(stats.apply.rounding(stats.default(x), digits = 2),
       c("", sprintf("%0.3f", p)))
}

my_render <- list(
  continuous = render_cont,
  categorical = render_cat
)
render_cont <- function(x) {
  with(stats.apply.rounding(stats.default(x), digits = 2),
       c("", sprintf("%0.3f", p)))
}

render_cat <- function(x) {
  with(stats.apply.rounding(stats.default(x), digits = 2),
       c("", sprintf("%0.3f", p)))
}

my_render <- list(
```

```
  continuous = render_cont,
  categorical = render_cat
)

Cpap <- as.data.frame(Cpap)


vars <- c("ethnicity", "education", "race", "age", "sex", "bmi",
          "ahi", "ess", "mmse", "odsi_bl", "odsi_6m", "adcs_12m", "avg_daily
_cpap")

catVars <- c("ethnicity", "education", "race", "sex")


table1 <- CreateTableOne(
  vars = vars,
  strata = "adherence",
  data = Cpap,
  factorVars = catVars,
  includeNA = TRUE,
  addOverall = TRUE
)

table1_text <- print(table1, showAllLevels = TRUE, quote = FALSE, noSpaces =
TRUE, test = TRUE)

##                              Stratified by adherence
##                               level                  Overall       Adherent
##   n                                                  174           128
##   ethnicity (%)               Hispanic or Latino     13 (7.5)      10 (7.8)
##                               Not Hispanic or Latino 161 (92.5)    118 (92.
2)
##   education (%)               <= high school         36 (20.7)     24 (18.8
)
##                               > high school          138 (79.3)    104 (81.
2)
##   race (%)                    Black                  37 (21.3)     18 (14.1
)
##                               NA                     1 (0.6)       0 (0.0)
##                               Other                  12 (6.9)      9 (7.0)
##                               White                  124 (71.3)    101 (78.
9)
##   age (mean (SD))                                    66.86 (7.52)  66.81 (7
.53)
##   sex (%)                     Female                 80 (46.0)     58 (45.3
)
##                               Male                   94 (54.0)     70 (54.7
)
##   bmi (mean (SD))                                    42.18 (7.21)  42.20 (7
```

```
.18)
##   ahi (mean (SD))                                  34.78 (20.82) 34.49 (2
1.20)
##   ess (mean (SD))                                  8.89 (4.96)   8.84 (5.
04)
##   mmse (mean (SD))                                 27.60 (1.78)  27.67 (1
.77)
##   odsi_bl (mean (SD))                              7.98 (6.08)   7.87 (6.
21)
##   odsi_6m (mean (SD))                              5.27 (4.93)   5.01 (4.
80)
##   adcs_12m (mean (SD))                             3.19 (1.47)   3.07 (1.
47)
##   avg_daily_cpap (mean (SD))                       5.15 (2.50)   6.42 (1.
32)
##                              Stratified by adherence
##                               Non-adherent  p      test
##   n                           46
##   ethnicity (%)               3 (6.5)       1.000
##                               43 (93.5)
##   education (%)               12 (26.1)     0.400
##                               34 (73.9)
##   race (%)                    19 (41.3)     <0.001
##                               1 (2.2)
##                               3 (6.5)
##                               23 (50.0)
##   age (mean (SD))             66.98 (7.57)  0.898
##   sex (%)                     22 (47.8)     0.904
##                               24 (52.2)
##   bmi (mean (SD))             42.15 (7.37)  0.966
##   ahi (mean (SD))             35.59 (19.91) 0.758
##   ess (mean (SD))             9.02 (4.79)   0.828
##   mmse (mean (SD))            27.39 (1.81)  0.361
##   odsi_bl (mean (SD))         8.30 (5.77)   0.677
##   odsi_6m (mean (SD))         6.18 (5.35)   0.224
##   adcs_12m (mean (SD))        3.69 (1.41)   0.053
##   avg_daily_cpap (mean (SD)) 1.61 (1.35)   <0.001
```

```r
table1_df <- as.data.frame(print(
  table1,
  showAllLevels = TRUE,
  quote = FALSE,
  noSpaces = TRUE,
  test = TRUE,
  smd = TRUE,
  printToggle = FALSE,
))

group_counts <- table(Cpap$adherence)
nonad_n <- group_counts["Non-adherent"]
```

```r
ad_n <- group_counts["Adherent"]

colnames(table1_df) <- c(
  "Level",
  "Overall",
  paste0("Non-adherent (n = ", nonad_n, ")"),
  paste0("Adherent (n = ", ad_n, ")"),
  "p-value",
  "test",
  "Effect Size (SMD)"
)

table1_df <- tibble::rownames_to_column(table1_df[, -6], var = "Characteristi
c")

table1_df$Characteristic <-
  c("N", "Ethnicity", "", "Education", "", "race", "", "","", "age.. mean.. S
D..", "sex", "", "bmi..mean..SD", "ahi..mean..SD", "ess..mean..SD", "mmse..me
an..SD", "odsi_bl..mean..SD", "odsi_6m..mean..SD", "adsc_12m..mean..SD..", "a
vg_daily_cpap..mean..SD")

# table1_df$Characteristic <- if_else(
#   grepl("\\.", table1_df$Characteristic) & !endsWith(table1_df$Characterist
ic, "."),
#   "",
#   table1_df$Characteristic
# )


library(flextable)
ft <- flextable(table1_df)
ft <- set_caption(ft, caption = "Table 1. Demographic and Clinical Characteri
stics (n = 174)")
ft <- fontsize(ft, size = 10)
ft <- align(ft, align = "center", part = "all")
ft <- align(ft, j = 1, align = "left", part = "all")
ft <- bold(ft, j = 1, part = "body")
ft <- theme_zebra(ft)
ft
```

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

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
N		174	128	46		

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
Ethnicity	Hispanic or Latino	13 (7.5)	10 (7.8)	3 (6.5)	1.000	0.050
	Not Hispanic or Latino	161 (92.5)	118 (92.2)	43 (93.5)		
Education	<= high school	36 (20.7)	24 (18.8)	12 (26.1)	0.400	0.177
	> high school	138 (79.3)	104 (81.2)	34 (73.9)		
race	Black	37 (21.3)	18 (14.1)	19 (41.3)	<0.001	0.705
	NA	1 (0.6)	0 (0.0)	1 (2.2)		
	Other	12 (6.9)	9 (7.0)	3 (6.5)		
	White	124 (71.3)	101 (78.9)	23 (50.0)		
age.. mean.. SD..		66.86 (7.52)	66.81 (7.53)	66.98 (7.57)	0.898	0.022
sex	Female	80 (46.0)	58 (45.3)	22 (47.8)	0.904	0.050
	Male	94 (54.0)	70 (54.7)	24 (52.2)		
bmi..mean..SD		42.18 (7.21)	42.20 (7.18)	42.15 (7.37)	0.966	0.007
ahi..mean..SD		34.78 (20.82)	34.49 (21.20)	35.59 (19.91)	0.758	0.054
ess..mean..SD		8.89 (4.96)	8.84 (5.04)	9.02 (4.79)	0.828	0.038
mmse..mean..SD		27.60 (1.78)	27.67 (1.77)	27.39 (1.81)	0.361	0.157
odsi_bl.. mean..SD		7.98 (6.08)	7.87 (6.21)	8.30 (5.77)	0.677	0.073
odsi_6m. .mean..SD		5.27 (4.93)	5.01 (4.80)	6.18 (5.35)	0.224	0.230

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
adsc_12m..mean..SD..		3.19 (1.47)	3.07 (1.47)	3.69 (1.41)	0.053	0.434
avg_daily_cpap..mean..SD		5.15 (2.50)	6.42 (1.32)	1.61 (1.35)	<0.001	3.613

```
ft <- add_footer_lines(ft,
  values = c(
    "Footnotes:",
    "Continuous variables were compared using t-tests or Wilcoxon rank-sum tests.",
    "Categorical variables were compared using Chi-squared or Fisher's exact tests.",
    "Effect sizes are reported as standardized mean differences (SMD).",
    "95% confidence intervals (95% CI) are shown where applicable.",
    "B. A significant P value illustrated in this data set was the difference in race between subjects who were adherent and non adherent to the CPAP therapy. The p value indicated was p < 0.001. A larger portion of the non adherent participants identified as black vs the adherent participants.",
    "C. A non significant P value illustrated in this data set was how there was not a large difference in age between the adherent and non adherent groups. The p value was 0.898. Therefore, age was not a controlling factor over CPAP adherence "
  )
)

ft
```

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

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
N		174	128	46		
Ethnicity	Hispanic or Latino	13 (7.5)	10 (7.8)	3 (6.5)	1.000	0.050
	Not Hispanic or Latino	161 (92.5)	118 (92.2)	43 (93.5)		

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
Education	<= high school	36 (20.7)	24 (18.8)	12 (26.1)	0.400	0.177
	> high school	138 (79.3)	104 (81.2)	34 (73.9)		
race	Black	37 (21.3)	18 (14.1)	19 (41.3)	<0.001	0.705
	NA	1 (0.6)	0 (0.0)	1 (2.2)		
	Other	12 (6.9)	9 (7.0)	3 (6.5)		
	White	124 (71.3)	101 (78.9)	23 (50.0)		
age.. mean.. SD..		66.86 (7.52)	66.81 (7.53)	66.98 (7.57)	0.898	0.022
sex	Female	80 (46.0)	58 (45.3)	22 (47.8)	0.904	0.050
	Male	94 (54.0)	70 (54.7)	24 (52.2)		
bmi..mean..SD		42.18 (7.21)	42.20 (7.18)	42.15 (7.37)	0.966	0.007
ahi..mean..SD		34.78 (20.82)	34.49 (21.20)	35.59 (19.91)	0.758	0.054
ess..mean..SD		8.89 (4.96)	8.84 (5.04)	9.02 (4.79)	0.828	0.038
mmse..mean..SD		27.60 (1.78)	27.67 (1.77)	27.39 (1.81)	0.361	0.157
odsi_bl.. mean..SD		7.98 (6.08)	7.87 (6.21)	8.30 (5.77)	0.677	0.073
odsi_6m. .mean..SD		5.27 (4.93)	5.01 (4.80)	6.18 (5.35)	0.224	0.230
adsc_12m..mean. .SD..		3.19 (1.47)	3.07 (1.47)	3.69 (1.41)	0.053	0.434
avg_daily_cpap..		5.15 (2.50)	6.42 (1.32)	1.61 (1.35)	<0.001	3.613

Characteristic	Level	Overall	Non-adherent (n = 46)	Adherent (n = 128)	p-value	Effect Size (SMD)
mean..SD						

Footnotes:

Continuous variables were compared using t-tests or Wilcoxon rank-sum tests.

Categorical variables were compared using Chi-squared or Fisher's exact tests.

Effect sizes are reported as standardized mean differences (SMD).

95% confidence intervals (95% CI) are shown where applicable.

B. A significant P value illustrated in this data set was the difference in race between subjects who were adherent and non adherent to the CPAP therapy. The p value indicated was $p < 0.001$. A larger portion of the non adherent participants identified as black vs the adherent participants.

C. A non significant P value illustrated in this data set was how there was not a large difference in age between the adherent and non adherent groups. The p value was 0.898. Therefore, age was not a controlling factor over CPAP adherence

#2

Step 1: Is daytime sleepiness (ODSI score) different for adherent vs non-adherent #participants from baseline to 6 months?

H0: There is no difference in the mean change in ODSI scores between adherent and non adherent participants.

H1: There is a difference in the mean change in ODSI scores between adherent and non adherent participants.

Step 2: I used a Welch two- sample t test to compare the mean change in ODSI scores between the adherent and non adherent groups. a= 0.05.

Step 3:

```
# Calculate change in ODSI from baseline to 6 months
Cpap <- Cpap %>%
  mutate(odsi_change = odsi_6m - odsi_bl)

t.test(odsi_change ~ adherence, data = Cpap)

##
##  Welch Two Sample t-test
##
## data:  odsi_change by adherence
## t = -1.1984, df = 55.171, p-value = 0.2359
## alternative hypothesis: true difference in means between group Adherent an
d group Non-adherent is not equal to 0
## 95 percent confidence interval:
##  -3.7948357  0.9544996
## sample estimates:
##     mean in group Adherent mean in group Non-adherent
##                  -2.949580                  -1.529412
```

Step 4: Based on the results of the two-sample t-test : (t(55.17) = -1.20, p = 0.236), there was not a large statistical difference in change of ODSI scores between the adherent and non adherent subjects. The adherent subjects improved by 2.95 points, vs the non adherent group that improved by 1.53 points. Although the adherent subjects showed a higher decrease in sleep time, this difference compared to the non adherent group was not significant. Thus, both groups had a similar experience in improvements of sleep from baseline to 6 months, therefore we retain the null hypothesis.

#3

Step 1: H0: The probability of subjects having excessive daytime sleepiness with an ODSI score of more than 6 (>6) is the same at baseline and at 6 months.

H1: the probability of subjects having excessive daytime sleepiness with an ODSI score of more than 6 (>6) is different at baseline and at 6 months.

Step 2: I will use the McNemar's test because this problem has us comparing (2) paired proportions of the same subjects at two of the same time points and these 2 variables are both dichotomous. a= 0.05

#Step. 3:

```
Cpap <- Cpap %>%
mutate(
odsi_high_bl = ifelse(odsi_bl >= 6, 1, 0),
odsi_high_6m = ifelse(odsi_6m >= 6, 1, 0)
)

table_sleepiness <- table(Cpap$odsi_high_bl, Cpap$odsi_high_6m)
mcnemar.test(table_sleepiness)

##
##  McNemar's Chi-squared test with continuity correction
##
## data:  table_sleepiness
## McNemar's chi-squared = 6.6852, df = 1, p-value = 0.009722
```

#Step 4: If we use a significant level of a= 0.05, and p<0.05 we will reject the null hypothesis because p= 0.009722.

#Step 5: There was a significant change in the variation of subjects who experienced excessive daytime sleepiness from baseline to 6 months being that p= 0.0097. This means that the odds of subjects having excessive daytime sleepiness decreased after 6 months. This makes sense and suggests that the CPAP treatment could have helped reduce daytime sleepiness over the course from baseline to 6 months.

#4

#A.

Step 1: H0: The population mean of MMSE is equal to 23, $\mu = 23$ (not cognitvely impaired)

H1: The population mean of MMSE is less than 23, μ < 23 (cogntively impaired)

#Step 2: I will use the z test to calculate this, p=0.05

#Step 3:

```
sigma <- 2


xbar <- mean(Cpap$mmse)
n <- length(Cpap$mmse)
mu0 <- 23  # hypothesized mean


z <- (xbar - mu0) / (sigma / sqrt(n))


p_value <- pnorm(z)


z; p_value
```

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

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

Step 4: If we set the criterion that p= 0.05, being that p > 0.05 in this z test (p= 1). We would fail to reject the null hypothesis.

Step 5: Since our z test statistic was well above 23 (z = 30), and our p value was 1, we can conclude that there is no proof that the average MMSE score is below 23 and that there is no coginitive impairment. The average is much higher, thus, this group has strong cognitive abiltiies.

#B.

Step 1: H0: The population mean of MMSE is equal to 23, μ = 23 (not cognitvely impaired)

H1: The population mean of MMSE is less than 23, μ < 23 (cogntively impaired).

Step 2: I will use the one sample t-test because we do not know the SD. a = 0.05

Step 3:

```
t_test_result <- t.test(Cpap$mmse, mu = 23, alternative = "less")
t_test_result
```

```
##
##  One Sample t-test
##
## data:  Cpap$mmse
## t = 34.081, df = 173, p-value = 1
## alternative hypothesis: true mean is less than 23
## 95 percent confidence interval:
##     -Inf 27.8208
## sample estimates:
## mean of x
##   27.5977
```

```
t_stat <- t_test_result$statistic
p_val_t <- t_test_result$p.value
xbar <- mean(Cpap$mmse)
s <- sd(Cpap$mmse)
n <- length(Cpap$mmse)

cat("One-sample t-test (H1: mean < 23)\n")
```

```
## One-sample t-test (H1: mean < 23)
```

```
cat("t = ", round(t_stat, 4), ", df = ", t_test_result$parameter, ", p = ", s
ignif(p_val_t, 3), "\n")
```

```
## t =  34.081 , df =  173 , p =  1
```

```
cat("Sample mean =", round(xbar, 3), ", sample SD =", round(s, 3), ", n =", n
, "\n")
```

```
## Sample mean = 27.598 , sample SD = 1.78 , n = 174
```

Step 4: If we set the criterion that p = 0.05, we would fail to reject the null hypothesis because p > 0.05 in these results; p = 1.

Step 5: The subjects scored a 27.6 on the MMSE on average, which is above the cutoff of 23 that points towards cognitive impairment. This test also found that there was no proof that the groups score on average was below the cutoff of p = 1. Thus, the participants are mentally healthy.

#5

#A. Mean and SD for CPAP use

```
# Replace 'avg_daily_use' with the actual column name if different
mean_use <- mean(Cpap$avg_daily_cpap, na.rm = TRUE)
sd_use   <- sd(Cpap$avg_daily_cpap, na.rm = TRUE)
n_use    <- sum(!is.na(Cpap$avg_daily_cpap))

mean_use; sd_use; n_use
```

```
## [1] 5.150766
```

```
## [1] 2.504029
```

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

```
cat("Mean average-daily CPAP use =", round(mean_use, 3),
    "hours; SD =", round(sd_use, 3), "hours; n =", n_use, "\n")
```

```
## Mean average-daily CPAP use = 5.151 hours; SD = 2.504 hours; n = 174
```

#B. CPAP use < 3 hours

```
# Parametric estimate (normal assumption, parameters = sample mean & sd)
prop_lt3_parametric <- pnorm(3, mean = mean_use, sd = sd_use)  # P(X < 3)

# Also compute the empirical proportion from the sample for comparison
prop_lt3_empirical <- mean(Cpap$avg_daily_cpap < 3, na.rm = TRUE)

prop_lt3_parametric; prop_lt3_empirical
```

```
## [1] 0.1951917
```

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

```
cat("Estimated proportion (parametric, Normal) with < 3 hours =",
    signif(prop_lt3_parametric, 3), "\n")
```

```
## Estimated proportion (parametric, Normal) with < 3 hours = 0.195
```

```
cat("Observed proportion in sample with < 3 hours =",
    signif(prop_lt3_empirical, 3), "\n")
```

```
## Observed proportion in sample with < 3 hours = 0.207
```

#C. Interpretation.

Based on the test results, participants used their CPAP machines for an average of just over 5 hours per night. The estimated proportion of individuals using the device for less than 3 hours nightly was 19.5%, closely matching the observed rate

of 20.6%. This means roughly 1 in 5 subjects had minimal usage, suggesting that a notable portion may not be fully benefiting from CPAP therapy. In contrast, the majority of participants appear to be using the treatment consistently and as recommended.

#6

#A. BMI mean and SD

```
mean_bmi <- mean(Cpap$bmi, na.rm = TRUE)
sd_bmi   <- sd(Cpap$bmi, na.rm = TRUE)
n_bmi    <- sum(!is.na(Cpap$bmi))

mean_bmi; sd_bmi; n_bmi
```

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

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

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

```
cat("Mean BMI =", round(mean_bmi, 2), "SD =", round(sd_bmi, 2), "n =", n_bmi,
"\n")
```

```
## Mean BMI = 42.18 SD = 7.21 n = 174
```

#B Obese subjects; BMI > 30

```
prop_obese_param <- 1 - pnorm(30, mean = mean_bmi, sd = sd_bmi)


prop_obese_emp <- mean(Cpap$bmi >= 30, na.rm = TRUE)

prop_obese_param; prop_obese_emp
```

```
## [1] 0.9545592
```

```
## [1] 0.9597701
```

```
cat("Estimated proportion (parametric, Normal) obese =", round(prop_obese_par
am, 3), "\n")
```

```
## Estimated proportion (parametric, Normal) obese = 0.955
```

```
cat("Observed proportion in sample obese =", round(prop_obese_emp, 3), "\n")
```

```
## Observed proportion in sample obese = 0.96
```

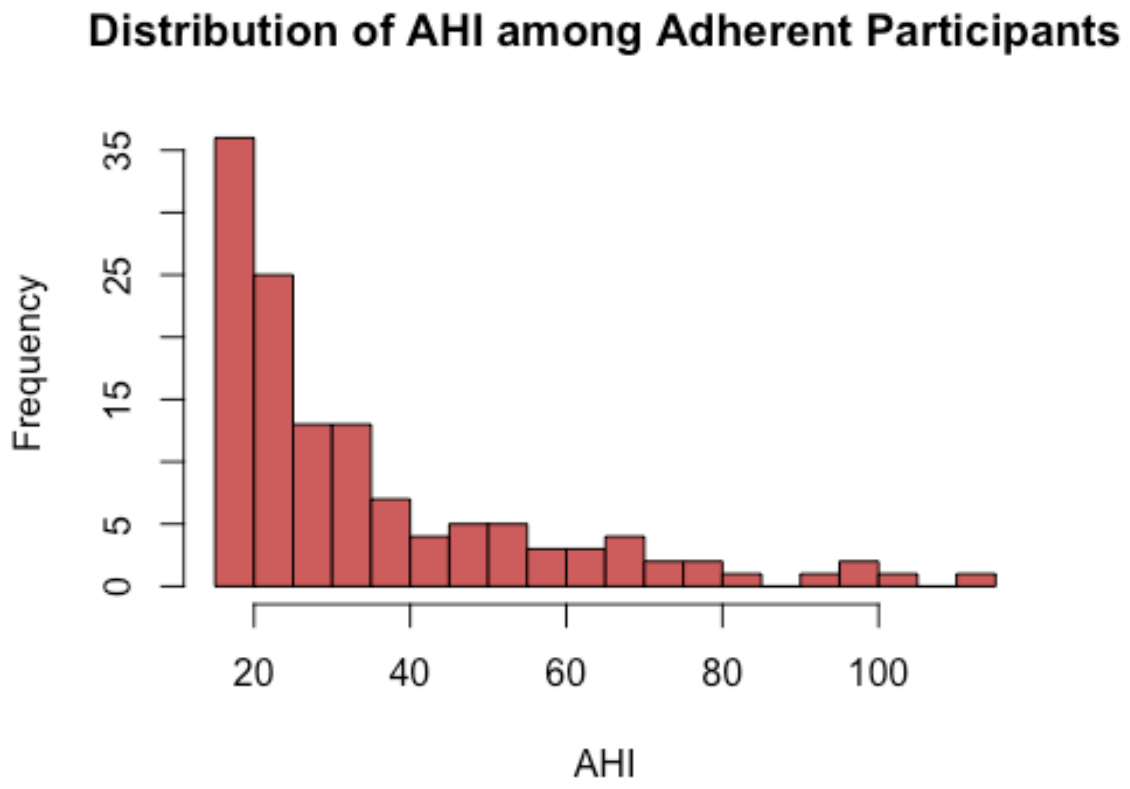
#C Interpretation.

The mean BMI among subjects was 42.18, with a standard deviation of 7.21, indicating considerable variability in body weight. Since a BMI of 30 or higher is classified as obese, the average suggests that the majority of participants fell within the obesity range. Both the observed and estimated proportions of obese individuals were approximately 95%, confirming that nearly all subjects were considered obese based on their BMI values.

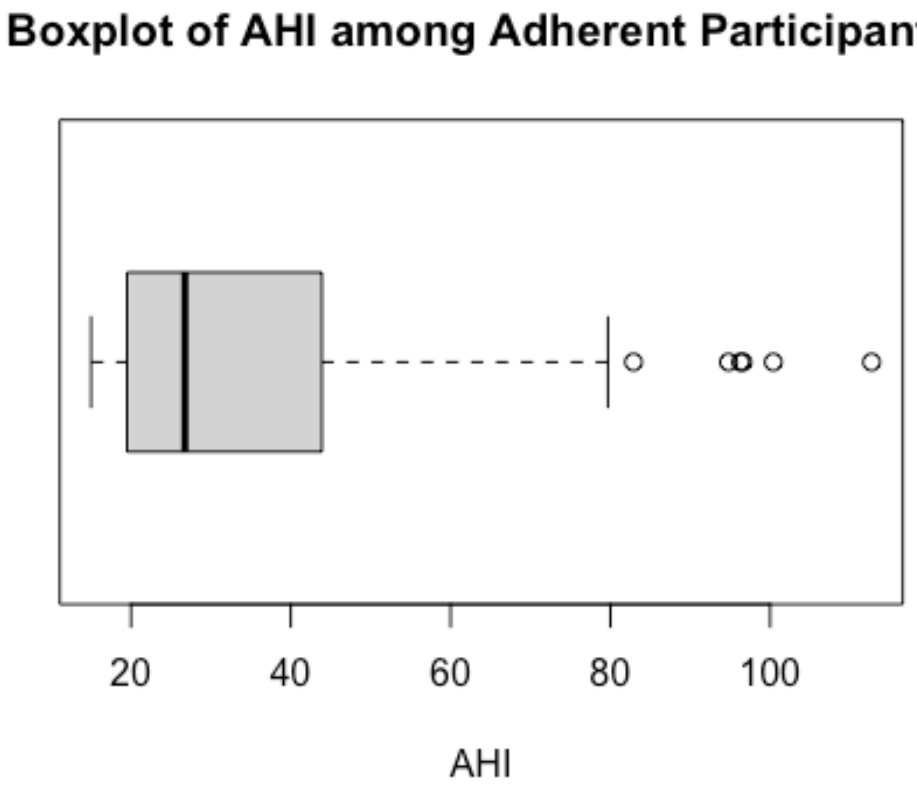
#7.

#A. AHI distribution in Adherent Subjects

```
ahi_adherent <- Cpap$ahi[Cpap$adherence == "Adherent"]
length(ahi_adherent)  # confirm n = 128

## [1] 128

mean_ahi <- mean(ahi_adherent, na.rm = TRUE)
sd_ahi   <- sd(ahi_adherent, na.rm = TRUE)
median_ahi <- median(ahi_adherent, na.rm = TRUE)
range_ahi <- range(ahi_adherent, na.rm = TRUE)
summary(ahi_adherent)

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

# Histogram
hist(ahi_adherent, breaks = 15, col = "indianred",
     main = "Distribution of AHI among Adherent Participants",
     xlab = "AHI", ylab = "Frequency")
```




```r
# Boxplot
boxplot(ahi_adherent, horizontal = TRUE,
        main = "Boxplot of AHI among Adherent Participants",
        xlab = "AHI")
```

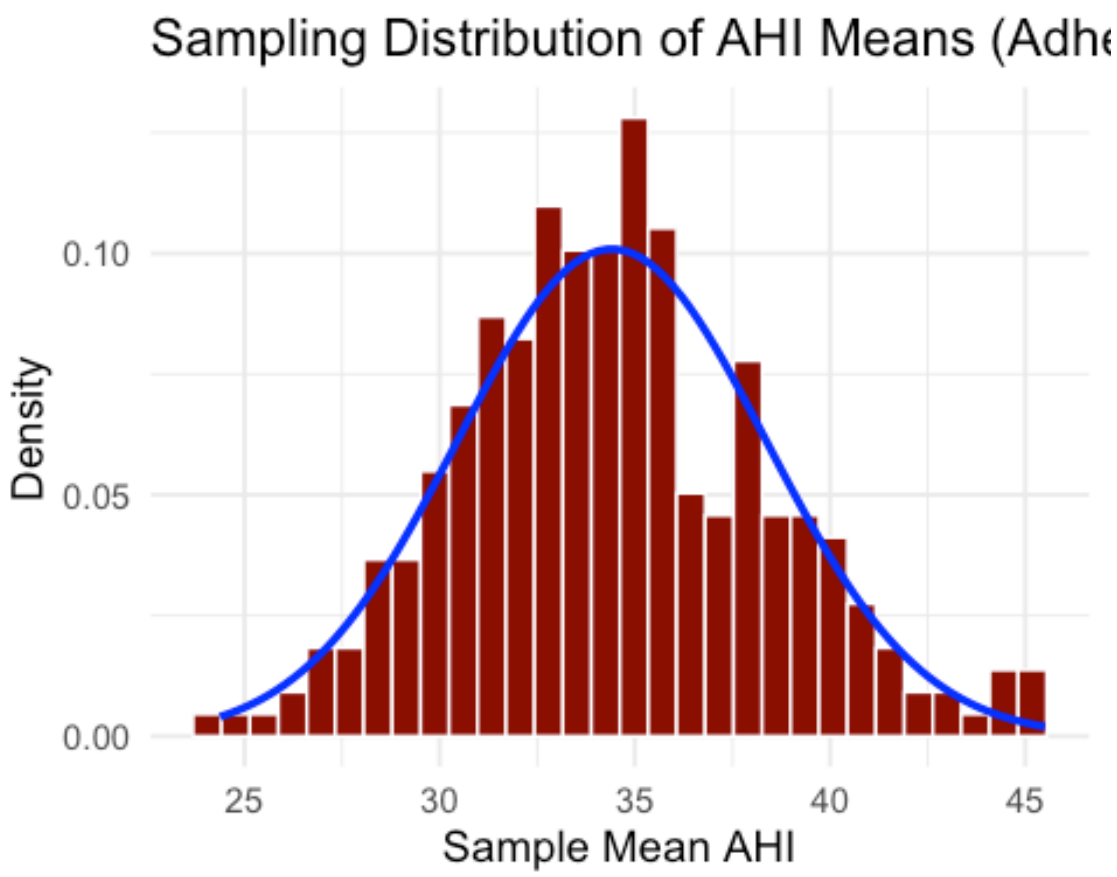


In the distribution of AHI scores among adherent participants, values ranged from 15 to 113, with a mean of approximately 34.5. Most participants had scores within a moderate range, though a few exhibited notably high values. This indicates substantial variability in baseline AHI measurements within the adherent group. The histogram reveals a right-skewed distribution, while the boxplot highlights several high scores and distinct outliers, further emphasizing the presence of extreme values in the data.

#B. Theoretical sampling distributions of AHI means (n = 30, Adherent)

```
MY_DATA <- subset(Cpap, adherence == "Adherent")
VARIABLE <- "ahi"
SAMPLES <- 300
SIZE <- 30

meanValues3 <- NULL

for (i in 1:SAMPLES) {
  sampSpots <- sample(x = 1:nrow(MY_DATA),
                      size = SIZE,
                      replace = TRUE)
```

```r
  thisSamp <- MY_DATA[sampSpots, names(MY_DATA) == VARIABLE]
  meanValues3 <- c(meanValues3, mean(thisSamp))
}

mean1 <- mean(meanValues3, na.rm = TRUE)
sd1 <- sd(meanValues3, na.rm = TRUE)

cat("Mean of sampling distribution:", mean1, "\n")

## Mean of sampling distribution: 34.40887

cat("SD of sampling distribution:", sd1, "\n")

## SD of sampling distribution: 3.953907

df_means1 <- data.frame(meanValues3)

ggplot(df_means1, aes(x = meanValues3)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "red4",
                 color = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean1, sd = sd1),
                color = "blue",
                size = 1.2) +
  labs(title = "Sampling Distribution of AHI Means (Adherent Participants)",
       x = "Sample Mean AHI",
       y = "Density") +
  theme_minimal(base_size = 14)
```

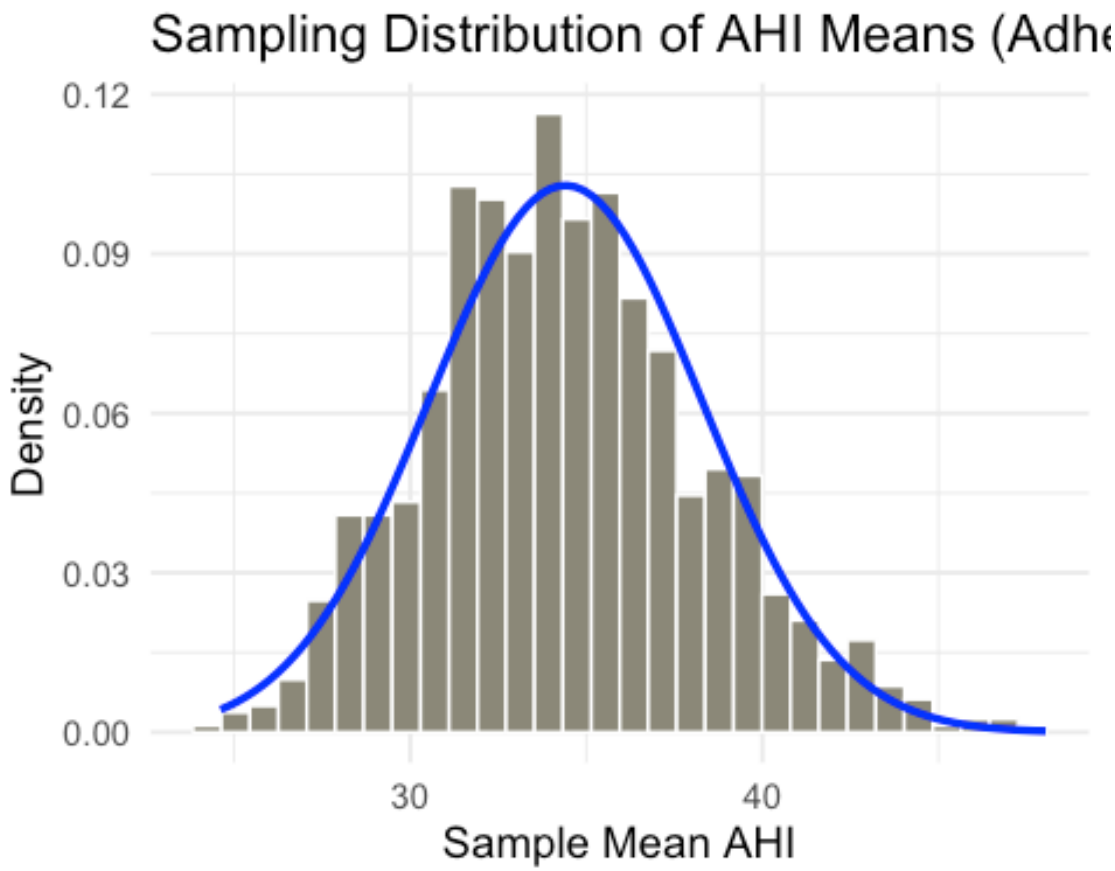


If we repeatedly drew samples of 30 adherent participants, the resulting sample means would cluster around 34.52, with a standard error of approximately 3.62. In this case, the sampling distribution would be roughly normal in shape and noticeably narrower than the distribution of individual AHI index scores. Being that the sample mean is based on multiple people, the spread of this distribution is smaller than the standard deviation of AHI index score data. Overall, these scores only differ slightly around the overall mean.

#C Draw 1,000 samples of size 30 (with replacement) and calculate sample means.

```
# Load ggplot2
library(ggplot2)

# Subset adherent participants
MY_DATA <- subset(Cpap, adherence == "Adherent")
VARIABLE <- "ahi"
SAMPLES <- 1000
SIZE <- 30

# Generate sampling distribution of means
meanValues2 <- NULL
```

```r
for (i in 1:SAMPLES) {
  sampSpots <- sample(x = 1:nrow(MY_DATA),
                      size = SIZE,
                      replace = TRUE)
  thisSamp <- MY_DATA[sampSpots, names(MY_DATA) == VARIABLE]
  meanValues2 <- c(meanValues2, mean(thisSamp))
}

# Calculate mean and sd of the sampling distribution
mean_val <- mean(meanValues2, na.rm = TRUE)
sd_val <- sd(meanValues2, na.rm = TRUE)

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

## Mean of sampling distribution: 34.40605

cat("SD of sampling distribution:", sd_val, "\n")

## SD of sampling distribution: 3.878256

df_means <- data.frame(meanValues2)

ggplot(df_means, aes(x = meanValues2)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "cornsilk4",
                 color = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean_val, sd = sd_val),
                color = "blue",
                size = 1.2) +
  labs(title = "Sampling Distribution of AHI Means (Adherent Participants)",
       x = "Sample Mean AHI",
       y = "Density") +
  theme_minimal(base_size = 14)
```

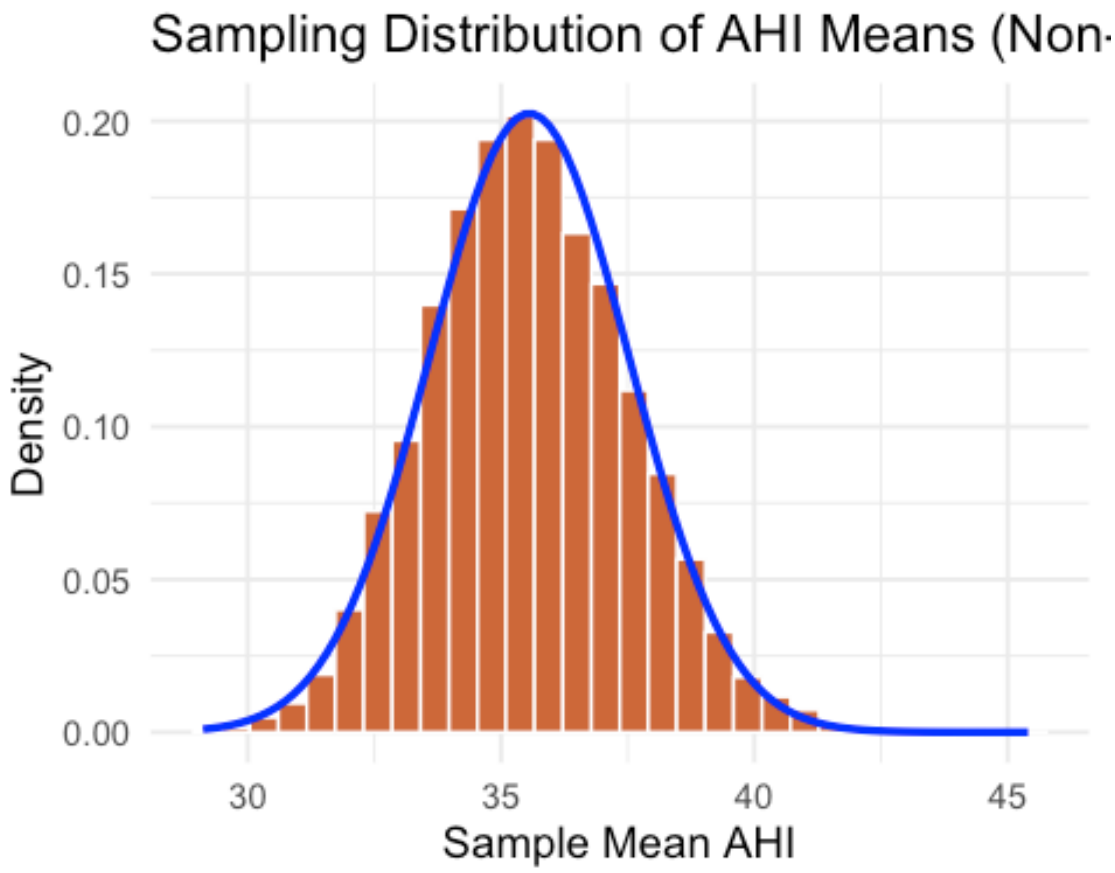


IV. This empirical sampling distribution closely resembles a normal curve, with a mean of 34.72 and a standard deviation of 3.87. Compared to a normal distribution with the same mean and standard deviation, it matches closely. The shape is bell-shaped and symmetric. While the original AHI scores are somewhat skewed, the sampling distribution of the means is approximately normal, as predicted by the Central Limit Theorem.

V. When comparing the empirical sampling distribution to the theoretical distribution from Problem B, both histograms appear very similar, with each centered around the same mean. Their standard deviations are slightly different (3.88 vs 3.62). With the SD of the adherent sampling distribution being larger, this tells us that the variability in the sample means larger. Overall, both problems indicate normal distributions and we can conclude that versus the individual AHI index scores, the sample means is more consistent and reliable.

#D. Theoretical sampling distribution of AHI means (n = 100, Non-Adherent)

```
MY_DATA <- subset(Cpap, adherence == "Non-adherent")
VARIABLE <- "ahi"
SAMPLES <- 10000
SIZE <- 100

meanValues4 <- NULL

for (i in 1:SAMPLES) {
  sampSpots <- sample(x = 1:nrow(MY_DATA),
                      size = SIZE,
                      replace = TRUE)
  thisSamp <- MY_DATA[sampSpots, names(MY_DATA) == VARIABLE]
  meanValues4 <- c(meanValues4, mean(thisSamp))
}
mean2 <- mean(meanValues4, na.rm = TRUE)
sd2 <- sd(meanValues4, na.rm = TRUE)

cat("Mean of sampling distribution:", mean2, "\n")

## Mean of sampling distribution: 35.55684

cat("SD of sampling distribution:", sd2, "\n")

## SD of sampling distribution: 1.96986

df_means1 <- data.frame(meanValues4)

ggplot(df_means1, aes(x = meanValues4)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "sienna3",
                 color = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean2, sd = sd2),
                color = "blue",
                size = 1.2) +
  labs(title = "Sampling Distribution of AHI Means (Non-Adherent Participants
)",
       x = "Sample Mean AHI",
       y = "Density") +
  theme_minimal(base_size = 14)
```

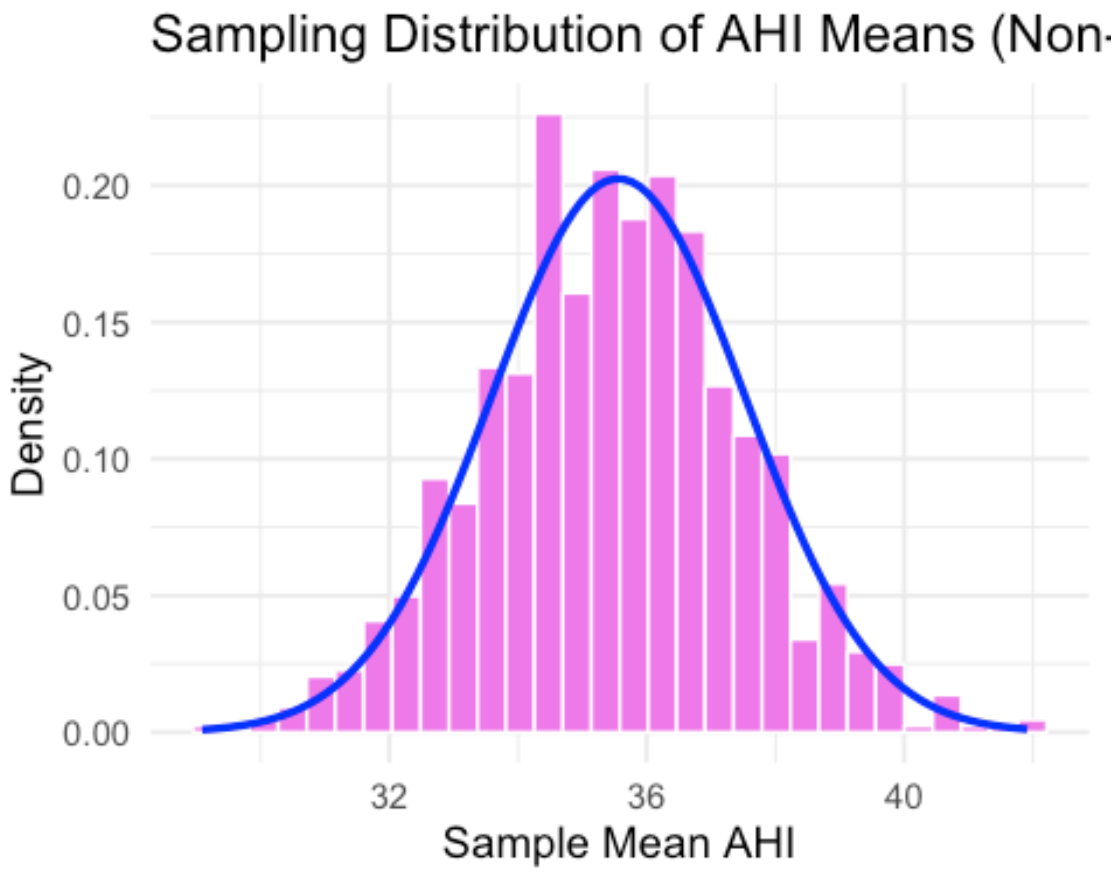


For non-adherent participants, the sampling distribution of the mean AHI for samples of size 100 is approximately normal, centered at a mean of 35.60 with a standard error of 1.97. The distribution is bell-shaped and balanced, indicating that most sample means cluster tightly around the population mean. Even if the original individual AHI scores are somewhat skewed, the averages of large samples usually tend to be more consistent and follow a normal pattern via the CLT.

E. Empirical Sampling Distribution for Non-Adherent subjects (n = 100)

```
MY_DATA <- subset(Cpap, adherence == "Non-adherent")
VARIABLE <- "ahi"
SAMPLES <- 1000
SIZE <- 100

meanValues5 <- NULL

for (i in 1:SAMPLES) {
  sampSpots <- sample(x = 1:nrow(MY_DATA),
                      size = SIZE,
                      replace = TRUE)
  thisSamp <- MY_DATA[sampSpots, names(MY_DATA) == VARIABLE]
```

```r
  meanValues5 <- c(meanValues5, mean(thisSamp))
}

mean3 <- mean(meanValues5, na.rm = TRUE)
sd3 <- sd(meanValues5, na.rm = TRUE)

cat("Mean of sampling distribution:", mean3, "\n")

## Mean of sampling distribution: 35.45734

cat("SD of sampling distribution:", sd3, "\n")

## SD of sampling distribution: 1.987399

df_means2 <- data.frame(meanValues5)

ggplot(df_means2, aes(x = meanValues5)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "orchid2",
                 color = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean2, sd = sd2),
                color = "blue",
                size = 1.2) +
  labs(title = "Sampling Distribution of AHI Means (Non-Adherent Participants
)",
       x = "Sample Mean AHI",
       y = "Density") +
  theme_minimal(base_size = 14)
```



IV. The empirical sampling distribution for the AHI index had a mean of 35.67 and a standard deviation of approximately 1.98. This sampling example closely matches that of a normal distibution. This sampling example illustrates a bell shape curve and a proportional alignment. Through drawing larger sample sizes, the sample means is more evenly balanced than the original AHI index scores.

V. The empirical mean and standard deviation closely mirror those from Part D, with both distributions centered from around 35.59-35.67 and showing a similar spread of approximately 1.97 to 1.98. Additionally, both distributions exhibit a roughly normal shape. This strong alignment reinforces the accuracy of the theoretical model in predicting how sample means behave when repeatedly drawing large samples from non-adherent participants.

#F. When comparing the sampling distributions of AHI between adherent and non-adherent participants, the overall mean was slightly higher for the non-adherent group (35.67) compared to the adherent group (34.73). The standard deviation of the sampling distribution was larger for adherent participants (SD = 3.88) and smaller for non-adherent participants (SD = 1.98), indicating greater consistency in

sample means among the non-adherent group. This difference in variability is primarily due to the disparity in sample sizes—100 for non-adherent versus 30 for adherent participants. Larger samples tend to produce more stable and less variable averages. Overall, both distributions approximate a normal shape, but the non-adherent group's distribution is narrower and more precise.