worksheet03_group01

Antonio Alfaro de Prado, Hyunchang Oh, Tanya Toluay

2024-05-04

1 - Stratification

```
# Get the full data and re-run the analysis
lung_data <- read.csv("C:/Program Files/R/R-4.2.2/lung_data_all.csv")
kable(head(lung_data))
```

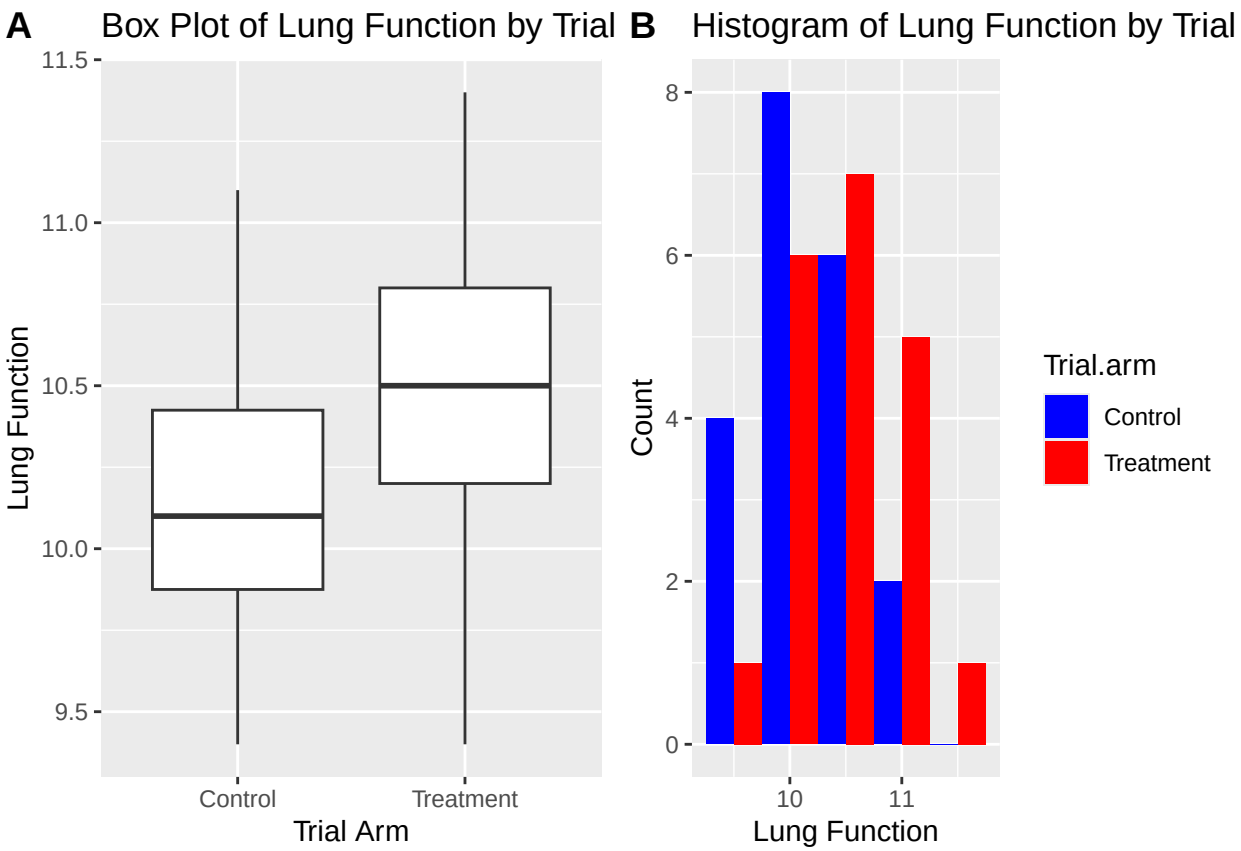
Subject.id	Lung.function	Trial.arm	Sex
1	11.0	Control	M
2	11.1	Control	M
3	9.5	Control	F
4	10.1	Control	F
5	9.7	Control	F
6	9.4	Control	F

```
lung_data$Trial.arm <- factor(lung_data$Trial.arm)
levels(lung_data$Trial.arm) <- c("Control", "Treatment")

# Box plot
box_plot <- ggplot(lung_data, aes(x = Trial.arm, y = Lung.function)) +
  geom_boxplot() +
  labs(title = "Box Plot of Lung Function by Trial Arm",
       x = "Trial Arm",
       y = "Lung Function")

# Histogram
histogram <- ggplot(lung_data, aes(x = Lung.function, fill = Trial.arm)) +
  geom_histogram(binwidth = 0.5, position = "dodge") +
  labs(title = "Histogram of Lung Function by Trial Arm",
       x = "Lung Function", y = "Count") +
  scale_fill_manual(values = c("Control" = "blue", "Treatment" = "red"))

combined_plots <- plot_grid(box_plot, histogram, labels = "AUTO", ncol = 2)
print(combined_plots)
```




```r
lung_function_control <- lung_data$Lung.function[lung_data$Trial.arm == "Control"]
lung_function_treatment <- lung_data$Lung.function[lung_data$Trial.arm == "Treatment"]

# t-test
t_test_result <- t.test(lung_function_control, lung_function_treatment)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  lung_function_control and lung_function_treatment
## t = -2.1569, df = 37.988, p-value = 0.0374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.63003538 -0.01996462
## sample estimates:
## mean of x mean of y
##    10.150    10.475
```

```r
# Run the analysis stratified by sex
lung_data_analysis <- lung_data %>%
  group_by(Sex) %>%
  summarize(mean_lung_function = mean(Lung.function))

summary_stats <- lung_data %>%
```

```r
  group_by(Sex) %>%
  summarise(
    mean_lung_function = mean(Lung.function),
    sd_lung_function = sd(Lung.function),
    min_lung_function = min(Lung.function),
    max_lung_function = max(Lung.function),
    median_lung_function = median(Lung.function)
  )

kable(summary_stats)
```

Sex	mean_lung_function	sd_lung_function	min_lung_function	max_lung_function	median_lung_function
F	9.995454	0.3538594	9.4	10.6	10.1
M	10.700000	0.3547990	10.1	11.4	10.7

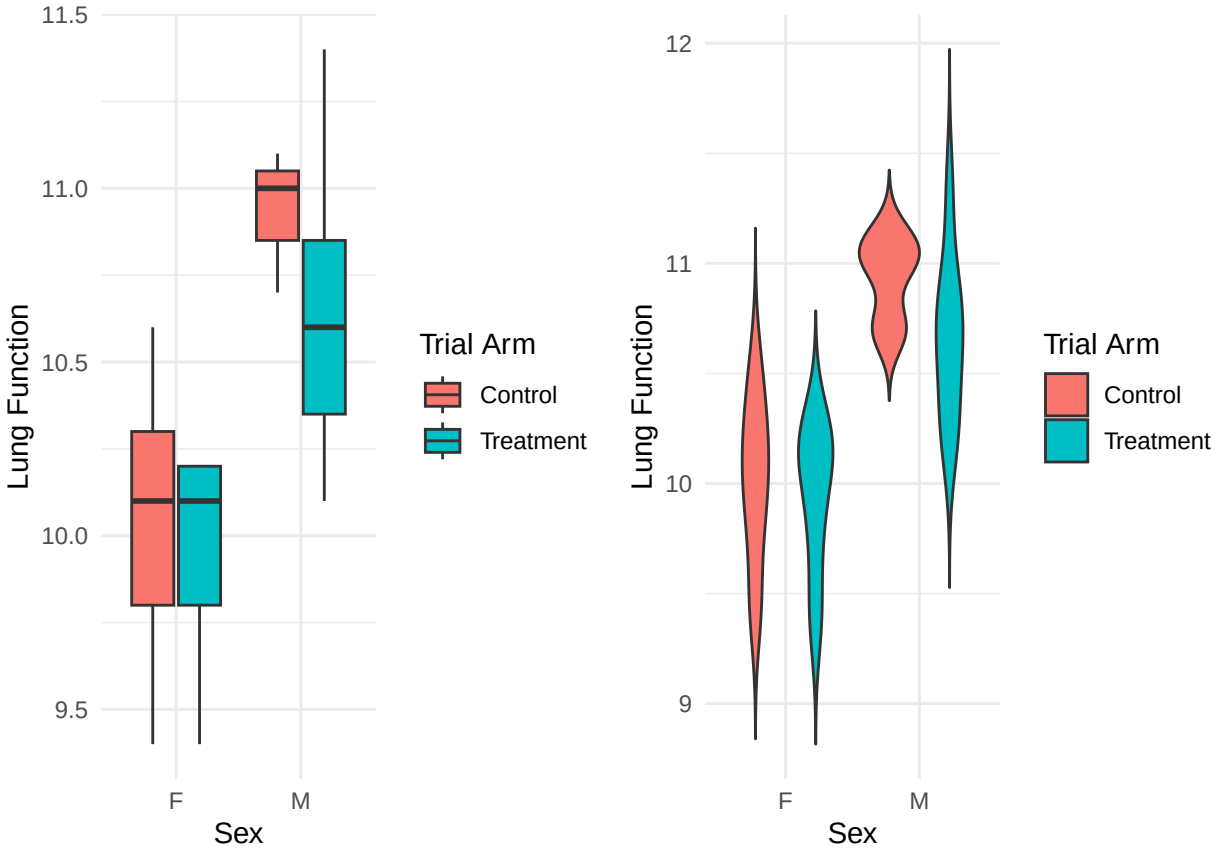
```r
# Anova
anova_result <- aov(Lung.function ~ Trial.arm * Sex, data = lung_data)
summary(anova_result)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Trial.arm      1  1.056   1.056    8.35  0.00649 **
## Sex            1  4.008   4.008   31.69 2.18e-06 ***
## Trial.arm:Sex  1  0.066   0.066    0.52  0.47536
## Residuals     36  4.554   0.126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Box plot
boxplot_plot <- ggplot(lung_data, aes(x = Sex, y = Lung.function, fill = Trial.arm)) +
  geom_boxplot() +
  labs(x = "Sex", y = "Lung Function", fill = "Trial Arm") +
  theme_minimal()

# Violin plot (to see density better)
violin_plot <- ggplot(lung_data, aes(x = Sex, y = Lung.function, fill = Trial.arm)) +
  geom_violin(trim = FALSE) +
  labs(x = "Sex", y = "Lung Function", fill = "Trial Arm") +
  theme_minimal()

combined_plots <- grid.arrange(boxplot_plot, violin_plot, ncol = 2)
```




```
print(combined_plots)
```

```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

Independent Samples t-test:

The t-test results indicate a statistically significant difference in lung function between the control and treatment groups (t = -2.1569, df = 37.988, p = 0.0374). The mean lung function in the treatment group (10.475) is higher than in the control group (10.150). The 95% confidence interval for the difference in means (-0.630 to -0.020) suggests that, on average, the treatment group has a higher lung function compared to the control group.

ANOVA with Sex as a Factor:

Trial Arm (Treatment vs. Control): The ANOVA results show a significant main effect of trial arm (F = 8.35, p = 0.00649), indicating that there are differences in lung function between the treatment and control groups. This supports the findings of the t-test.

Sex: There's a highly significant main effect of sex (F = 31.69, p < 0.001), suggesting that lung function differs between sexes. This could imply that biological differences between males and females contribute to variations in lung function.

Interaction Effect (Trial Arm by Sex): The interaction effect between trial arm and sex is not statistically significant (F = 0.52, p = 0.47536), indicating that the difference in lung function between the treatment and control groups is not moderated by sex. In other words, the treatment effect on lung function does not vary significantly between males and females.

Discussion:

Treatment Effect: The treatment appears to have a significant positive effect on lung function compared to the control group. This suggests that the treatment might be beneficial in improving lung function.

Sex Differences: The significant effect of sex on lung function highlights the importance of considering biological differences between males and females in lung function studies. It could be indicative of physiological disparities or potentially lifestyle factors that influence lung health differently between sexes.

2 - Confounders

w is generated from a normal distribution with mean 1 and some variance (given by the standard normal distribution). So, it follows a normal distribution with mean as 1 and variance as sigma^2

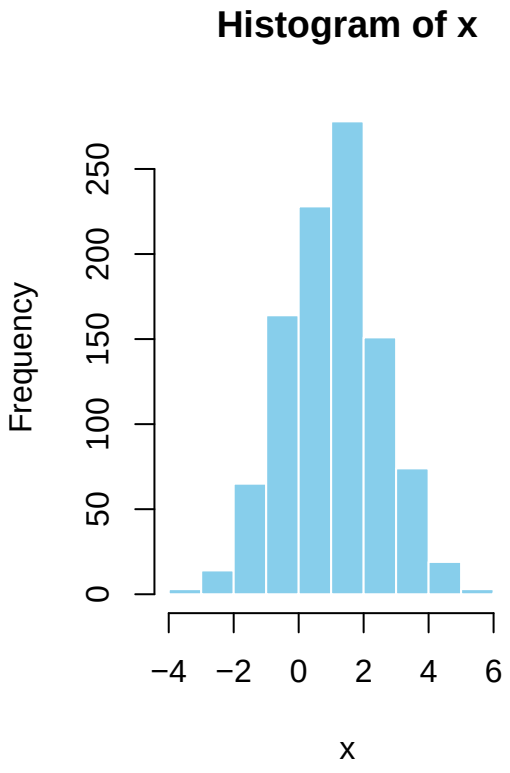
Both x and y are calculated by adding or subtracting w from x_0 and y_0, respectively. Since x_0 and y_0 are generated from standard normal distributions, the addition and subtraction of a normal variable result in another normal variable.

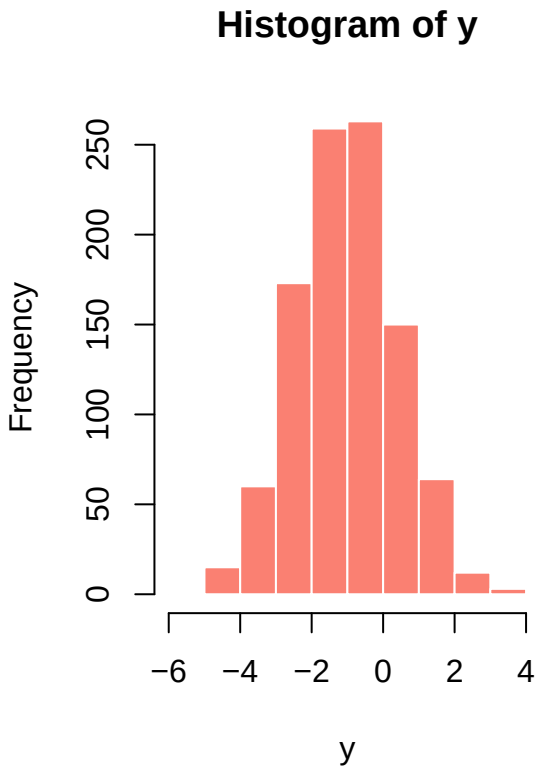
So, x and y both follow normal distributions.

```
set.seed(123)
N <- 1000

w <- 1 + rnorm(N)
x_0 <- rnorm(N)
y_0 <- rnorm(N)
x <- x_0 + w
y <- y_0 - w

par(mfrow = c(1, 2))
hist(x, main = "Histogram of x", xlab = "x", col = "skyblue", border = "white")
hist(y, main = "Histogram of y", xlab = "y", col = "salmon", border = "white")
```






```r
t_test_result <- t.test(x, y)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 32.502, df = 1993.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.968431 2.221236
## sample estimates:
## mean of x mean of y
##  1.058593 -1.036240
```

Interpretation:

The extremely small p-value (< 2.2e-16) indicates strong evidence against the null hypothesis, suggesting that the means of x and y are significantly different.

The confidence interval (1.968, 2.221) indicates that we are 95% confident that the true difference in means between x and y falls between approximately 1.968 and 2.221.

The sample estimates show that the mean of x (1.058593) is notably higher than the mean of y (-1.036240).

In summary, based on these results, we can conclude that there is a significant difference between the means of x and y, with x having a higher mean than y.

```r
w <- 1 + rnorm(N)

w_p <- w * sample(c(-1, 1), N, replace = TRUE)

x_0 <- rnorm(N)
y_0 <- rnorm(N)

x_p <- x_0 + w_p
y_p <- y_0 - w_p

t_test_result <- t.test(x_p, y_p)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  x_p and y_p
## t = -0.83344, df = 1996.4, p-value = 0.4047
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.21614167  0.08722031
## sample estimates:
##   mean of x   mean of y
## -0.02752325  0.03693743
```

Interpretation:

The p-value (0.4047) is greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis.

This suggests that there is not enough evidence to conclude that there is a significant difference between the means of x_p and y_p.

The confidence interval (-0.216, 0.087) contains zero, further indicating that the true difference in means could plausibly be zero. The sample means of x_p and y_p are very close to each other, with y_p having a slightly higher mean than x_p, but this difference is not statistically significant.

In summary, based on these results, we cannot conclude that there is a significant difference between the means of x_p and y_p.

3 - Clustering

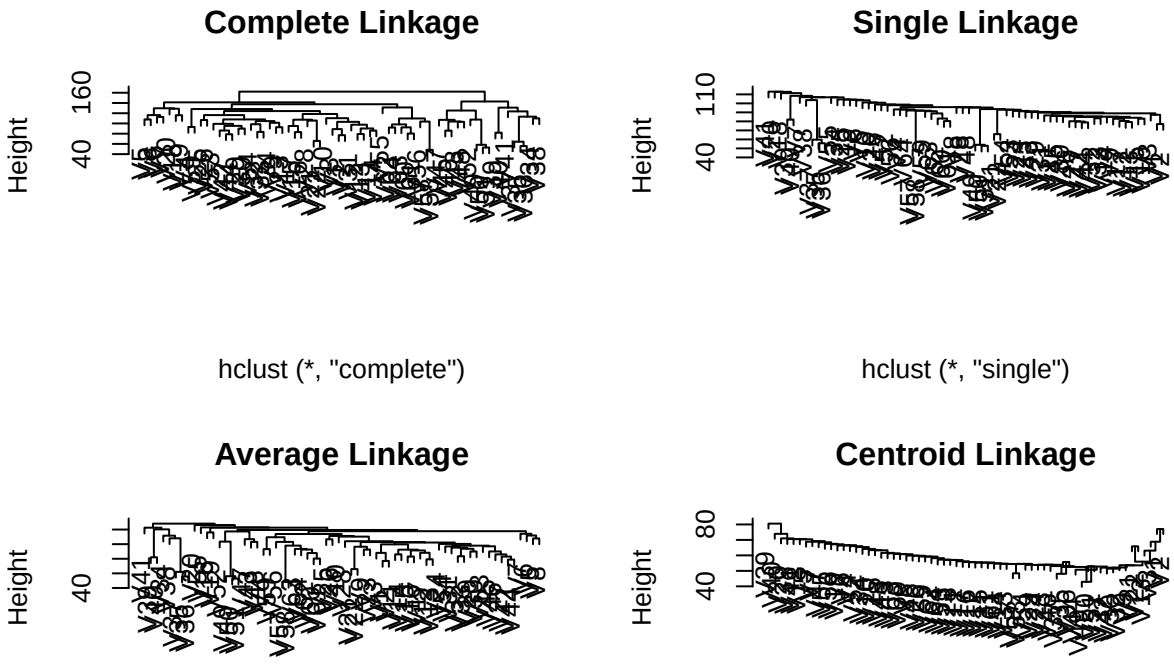
```r
library(ISLR2)
nci.labs <- NCI60$labs
nci.data <- NCI60$data

# Scale the variables
scaled_nci_data <- scale(nci.data)

# Hierarchical clustering with different linkage methods
complete_hclust <- hclust(dist(scaled_nci_data), method = "complete")
single_hclust <- hclust(dist(scaled_nci_data), method = "single")
```

```r
average_hclust <- hclust(dist(scaled_nci_data), method = "average")
centroid_hclust <- hclust(dist(scaled_nci_data), method = "centroid")

# Plot dendrograms
par(mfrow = c(2, 2))
plot(complete_hclust, main = "Complete Linkage", xlab = "", ylab = "Height")
plot(single_hclust, main = "Single Linkage", xlab = "", ylab = "Height")
plot(average_hclust, main = "Average Linkage", xlab = "", ylab = "Height")
plot(centroid_hclust, main = "Centroid Linkage", xlab = "", ylab = "Height")
```



hclust (*, "complete")

hclust (*, "single")

hclust (*, "average")

hclust (*, "centroid")

```r
# Cut the dendrogram at a reasonable height
cut_height <- 1500
clusters <- cutree(complete_hclust, h = cut_height)

plot(complete_hclust, main = "Complete Linkage Dendrogram")
abline(h = cut_height, col = "red")

library(clValid)
```

```
## Loading required package: cluster
```

```r
cluster_counts <- seq(2, 10, by = 1)

km_validation <- list()
```

```r
# Evaluate K-means clustering for each cluster count

#In this line of code:

#nClust = k: This parameter specifies the number of clusters to be generated, which is determined by th
#clMethods = 'kmeans': This parameter indicates that the k-means clustering method is being used.
#validation = "internal": This parameter specifies that internal validation measures should be used to
#neighbSize = 3: This parameter specifies the neighborhood size for connectivity validation.

for (k in cluster_counts) {
    print(k)
    kmeans_clusters <- kmeans(scaled_nci_data, centers = k, nstart = 25)
    km_validation[[as.character(k)]] <- clValid(scaled_nci_data, nClust = k, clMethods = 'kmeans', valid
}
```

```
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
```

```r
# Print validation measures for K-means clustering
print("Validation measures for K-means clustering:")
```

```
## [1] "Validation measures for K-means clustering:"
```

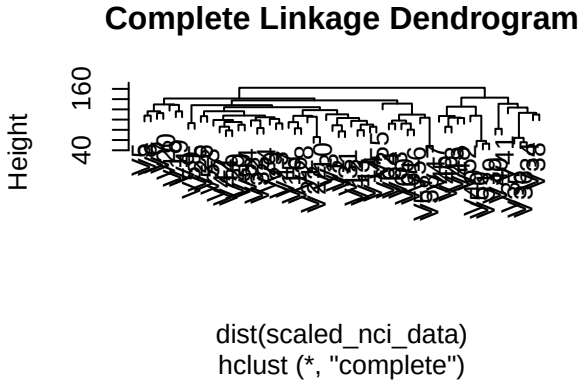
```r
print(km_validation)
```

```
## $`2`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  2
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`3`
##
## Call:
```

```
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  3
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`4`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  4
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`5`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  5
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`6`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
```

```
## Cluster sizes:
##  6
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`7`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  7
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`8`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  8
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
##
## $`9`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  9
##
## Validation measures:
##  Connectivity Dunn Silhouette
##
```

```
##
## $`10`
##
## Call:
## clValid(obj = scaled_nci_data, nClust = k, clMethods = "kmeans",
##     validation = "internal", neighbSize = 3)
##
## Clustering Methods:
##  kmeans
##
## Cluster sizes:
##  10
##
## Validation measures:
##  Connectivity Dunn Silhouette
```



Discussion

The k-means algorithm was applied, and internal validation measures, including Connectivity, Dunn index, and Silhouette index, were employed to gauge clustering quality.

Connectivity: This measure assesses the connectedness of clusters, indicating how well the data points within each cluster are connected to each other. Higher connectivity values suggest more cohesive clusters where data points are closely related to each other within the same cluster.

Dunn index: The Dunn index measures the compactness and separation of clusters. It compares the distance between clusters with the distance within clusters. Higher Dunn index values indicate better-defined clusters with smaller within-cluster distances and larger between-cluster distances.

Silhouette index: The Silhouette index evaluates the cohesion and separation of clusters by comparing the average distance of data points within clusters to the distance between clusters. A higher Silhouette index indicates that data points are well-clustered, with small within-cluster distances and large between-cluster distances.

4 - Review

(a) Is it a reasonable choice to use the chi square test to analyse your data? If not, what other test would you chose?

No, it's not a reasonable to use the chi-square test because it is used for categorical data analysis, while our data involves continuous variables (blood pressure measurements) and group comparisons. For analyzing continuous variables across different groups, particularly when comparing means between groups, in this case, since we have two groups (control vs. high alcohol consumption) and multiple factors (smoking status, obesity), an ANOVA with appropriate post-hoc tests would be suitable.

- **(b) You want to do a power analysis before you start the study. What quantities do you need to know or estimate to do this?**

We need to know i- how big of a difference we expect to see between the groups (effect size), ii- how sure we want to be that our results are real (significance level), iii- how likely we want to be to catch a real difference if it's there (power), and iv- how many people we'll need in our study (sample size).

(c) Your power analysis shows a rather weak power. What factor is the easiest to change to increase the power?

The easiest factor to change to increase power is typically the sample size. Increasing the sample size can directly enhance the power of our study. By including more participants, we increase the likelihood of detecting a true effect if it exists, thereby improving the study's power.

(d) If you put an equal amount of female/male participants in the high/low group this can be described as blocking. (T/F)

FALSE - Blocking typically involves separating participants into different groups based on these characteristics, not necessarily ensuring an equal number of participants from each gender in each group.

(e) Since you have many people in your study, it is necessary to correct for multiple testing. (T/F)

TRUE - Yes, if we deliberately allocate an equal number of male and female participants to each group (high alcohol consumption and low alcohol consumption), this can be described as blocking.

(f) In a good experimental design you only randomise background effects that you cannot block. (T/F)

TRUE - In a good experiment, we control what we can and randomize what we can't. Blocking helps control known differences, like gender, while randomization deals with things we can't control, like individual traits.

(g) What term describes the role of "smoking habit" in your first study?

In the context of the first study where "smoking habit" is suspected to influence blood pressure, it would be described as a potential confounding variable.

(h) Explain the contradictory results of your second study.

In the second study, the different results might be because we didn't consider all the factors that could affect blood pressure. In the first study, smoking was a problem, and in the second, obesity might have been the issue.

(i) It was not necessary to do the second study, because the first study already showed the effect of alcohol consumption on blood pressure (T/F)

FALSE - The second study was needed to clarify the effect of alcohol consumption without the influence of smoking.

(j) Your second study is proof that regular consumption of high amounts of alcohol increases blood pressure (T/F)

FALSE - The second study doesn't provide conclusive evidence that regular consumption of high amounts of alcohol increases blood pressure. It merely shows that when smoking, a potential confounding factor, is controlled for, the relationship between alcohol consumption and blood pressure changes. However, other factors like obesity were not fully addressed in the second study.