Homework1

Peyton Hall

01/22/2025

Load Necessary Libraries

```
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(ggplot2)
```

1. An experiment is designed to determine how speaker size affects loudness. The researcher measures loudness of 18 speakers randomly: 6 speakers have small diameters, 6 have medium diameters, and 6 have larger diameters. The researcher records the loudness (in decibels) of the speakers playing the same sound in the table below, and compared the three types of speakers.

   Small diameter speaker loudness, 15, 21, 16, 14, 20, 22 Medium diameter speaker loudness, 32, 41, 46, 39, 52, 43 Large diameter speaker loudness, 52, 50, 61, 69, 43, 58

a) Identify the independent variable (IV) IV = Speaker Size
b) Identify whether the independent variable is categorical or numeric. IV = Ordinal Categorical Variable (i.e. ordered)
c) Identify the dependent variable (DV) DV = loudness (in decibels)
d) Identify whether the DV is categorical or numeric DV = Discrete Numeric Variable (i.e. integers only)
e) What can be the confounding variables? (list one possible variable) One possible confounding variable could be the acoustics of the room.

Question 1 Parts f) and g)

```
# f) Enter the data in R as a data frame with one dependent and one independent
#    variable

# each = 6 ensures each speaker sizes' loudness is repeated six times.
```

```r
# This matches the six loudness data points to each size.
# rep() function replicates values in a vector
IV_Size <- rep(c("Small", "Medium", "Large"), each = 6)
DV_Loudness <- c(15, 21, 16, 14, 20, 22, 32, 41, 46, 39,
                 52, 43, 52, 50, 61, 69, 43, 58)

# create data frame using the vectors
speaker_data <- data.frame(IV_Size, DV_Loudness)
speaker_data
```

```
##    IV_Size DV_Loudness
## 1    Small          15
## 2    Small          21
## 3    Small          16
## 4    Small          14
## 5    Small          20
## 6    Small          22
## 7   Medium          32
## 8   Medium          41
## 9   Medium          46
## 10  Medium          39
## 11  Medium          52
## 12  Medium          43
## 13   Large          52
## 14   Large          50
## 15   Large          61
## 16   Large          69
## 17   Large          43
## 18   Large          58
```

```r
# g) Use pipeline to calculate the mean and standard deviation of loudness for
#    each type of speaker. List the mean and standard deviation for each speaker
#    here:

speaker_stats <- speaker_data %>%
  group_by(IV_Size) %>%
  summarize(Mean_Loudness = mean(DV_Loudness), SD_Loudness = sd(DV_Loudness))
speaker_stats
```

```
## # A tibble: 3 x 3
##   IV_Size Mean_Loudness SD_Loudness
##   <chr>           <dbl>       <dbl>
## 1 Large            55.5        9.14
## 2 Medium           42.2        6.74
## 3 Small            18          3.41
```

2. Dr. Optimist believes he has discovered a "smarts" pill will increase IQ. He first measures the IQ of 20 volunteered college juniors and then randomly assigns them to one of the two groups: 10 of them take the smarts pill and 10 take a placebo (sugar pill). He then measures the IQ of all the volunteers. He would like to compare the mean IQ between the pill group and placebo group.

   Smart pill IQ, 81, 98, 117, 76, 91, 110, 103, 105, 72, 121 Placebo IQ, 79, 80, 89, 94, 99, 106, 111, 109, 73, 119

a) Identify the independent variable and its levels IV = Pill Type
b) Is the independent variable categorical or numeric? IV = Nominal Categorical Variable
c) Identify the dependent variable DV = IQ (Intelligence Quotient)
d) Is the dependent variable categorical or numeric? DV = Numeric Variable Note: In this case, it is a Discrete Numeric Variable. However, IQ is sometimes expressed using decimals, which, in those cases, would make it a Continuous Numeric Variable.

Question 2 Parts e) and f)

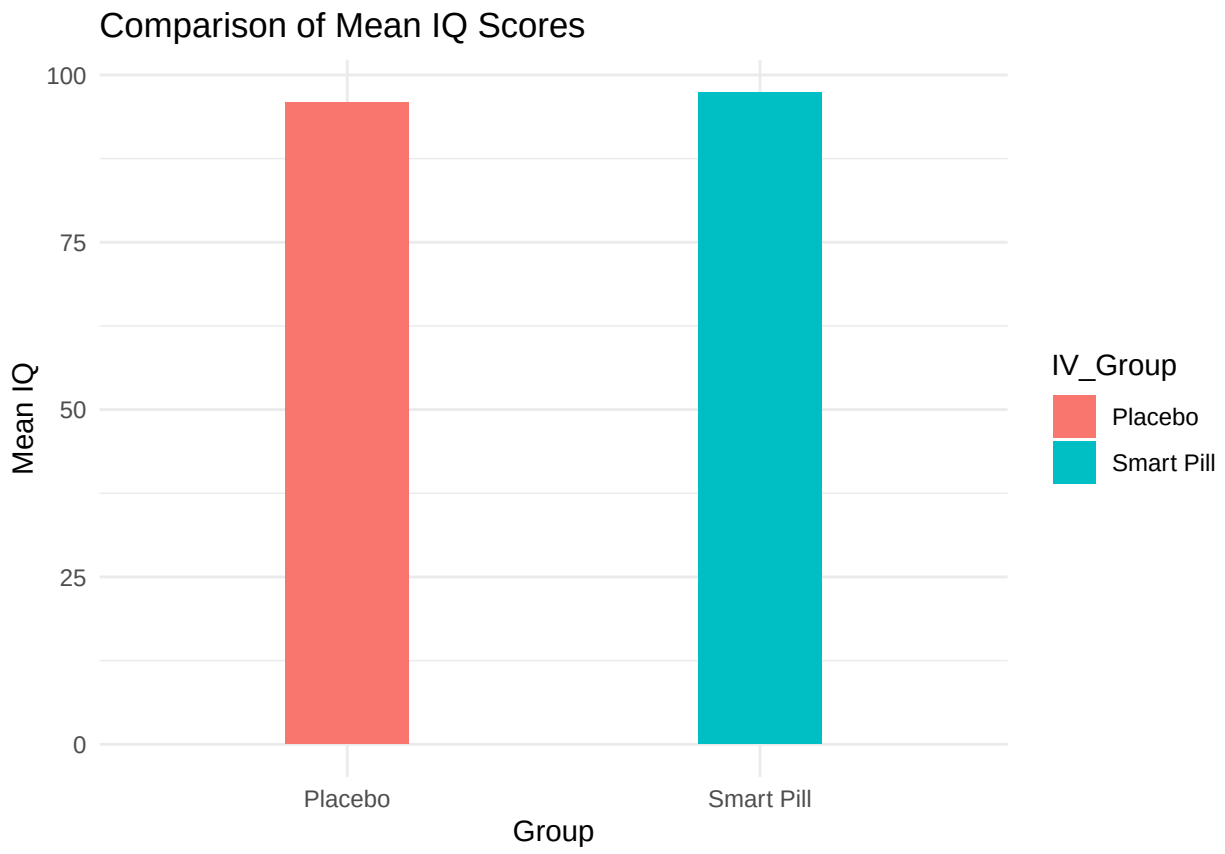
```
# e) Enter the data in R as a data frame. Submit your R Markdown file to d2l
#    drop box

# each = 10 ensures pill type's assignments repeat ten times.
# This aligns the IQ scores with each pill type.
# rep() function replicates values in a vector.
IV_Group <- rep(c("Smart Pill", "Placebo"), each = 10)
DV_IQ <- c(81, 98, 117, 76, 91, 110, 103, 105, 72, 121,
           79, 80, 89, 94, 99, 106, 111, 109, 73, 119)

# create data frame using the vectors
pill_data <- data.frame(IV_Group, DV_IQ)
pill_data
```

```
##      IV_Group DV_IQ
## 1  Smart Pill    81
## 2  Smart Pill    98
## 3  Smart Pill   117
## 4  Smart Pill    76
## 5  Smart Pill    91
## 6  Smart Pill   110
## 7  Smart Pill   103
## 8  Smart Pill   105
## 9  Smart Pill    72
## 10 Smart Pill   121
## 11    Placebo    79
## 12    Placebo    80
## 13    Placebo    89
## 14    Placebo    94
## 15    Placebo    99
## 16    Placebo   106
## 17    Placebo   111
## 18    Placebo   109
## 19    Placebo    73
## 20    Placebo   119
```

```
# f) Generate a graph to visualize the means of IQ from the smart pill group and
#    the placebo group. Submit your R Markdown file.
ggplot(pill_data, aes(x = IV_Group, y = DV_IQ, fill = IV_Group)) +
  geom_bar(stat = "summary", fun = "mean", width = 0.3) +
  labs(title = "Comparison of Mean IQ Scores", x = "Group", y = "Mean IQ") +
  theme_minimal()
```



3. Listed below are the lead concentrations (in ug/g) measured in different Ayurveda medicines. Ayurveda is a traditional medical system commonly used in India. The lead concentrations listed here are from medicines manufactured in the United States. The data below are based on the article "Lead, Mercury, and Arsenic in US and Indian Manufactured Ayurvedic Medicines Sold via the Internet," by Saper, et al., Journal of the American Medical Association.

   3.0 6.5 6.0 5.5 20.5 7.5 12.0 20.5 11.5 17.5

   What is the mean, standard deviation, range and median of this sample data?

Question 3

```
numbers <- c(3.0, 6.5, 6.0, 5.5, 20.5, 7.5, 12.0, 20.5, 11.5, 17.5)
mean(numbers)
```

```
## [1] 11.05
```

```
sd(numbers)
```

```
## [1] 6.461209
```

```
range(numbers)
```

```
## [1]  3.0 20.5
```

```r
median(numbers)
```

```
## [1] 9.5
```

4. The data Orange in R recorded the age and circumferences of 5 different types of orange trees. The variable "Tree" recorded the five types as 1, 2, 3, 4, and 5. The variable "age" recorded the ages of the trees and the variable "circumference" recorded the circumferences of the trees. Use R pipeline to

a) Keep only tree 1, 2 and 3.
b) Find the mean and standard deviation of circumferences of each type of the trees for tree type 1, 2 and 3. List the means and standard deviations:
c) Generate boxplots for tree type 1, 2 and 3. Submit your R Markdown file.

Question 4 Parts a - c

```r
# view the data
Orange
```

```
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
```

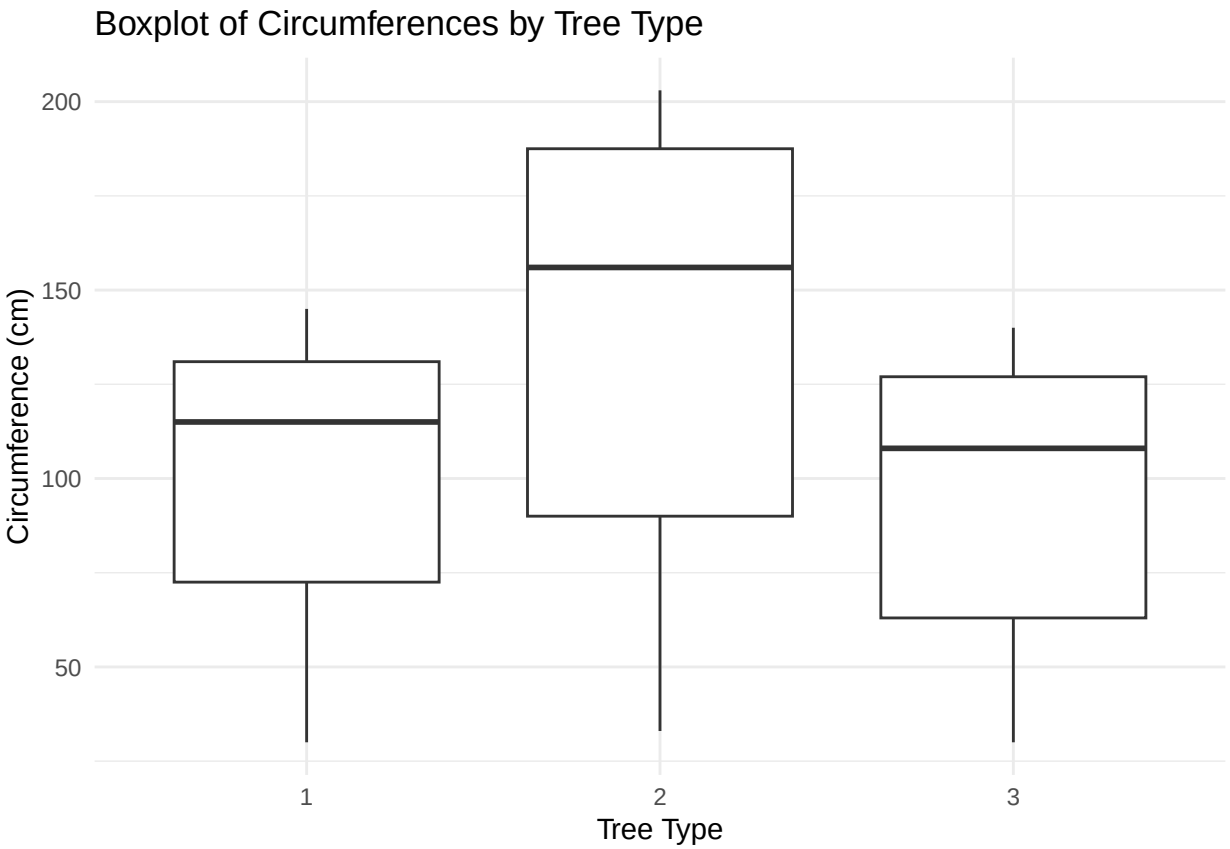
```
## 34    5 1372           174
## 35    5 1582           177
```

```r
# Part a: keep 1, 2, and 3
first_three <- Orange %>%
  filter(Tree == 1 | Tree == 2 | Tree == 3) # OR (|) operator

# Part b: find mean and standard deviation
stats <- first_three %>%
  group_by(Tree) %>%
  summarize(mean = mean(circumference), sd = sd(circumference))
stats
```

```
## # A tibble: 3 x 3
##   Tree   mean    sd
##   <ord> <dbl> <dbl>
## 1 3      94    43.0
## 2 1      99.6  43.3
## 3 2     135.   66.3
```

```r
# Part c: generate boxplots with ordered tree types
ggplot(first_three, aes(x = factor(Tree, levels = c(1, 2, 3)), # ascending order
                        y = circumference)) +
  geom_boxplot() +
  labs(title = "Boxplot of Circumferences by Tree Type", x = "Tree Type",
       y = "Circumference (cm)") +
  theme_minimal()
```


Boxplot of Circumferences by Tree Type
Circumference (cm)
200
150
100
50
1
2
3
Tree Type