Continuous Data (4 of 5)

Aug 3, 2023. V1.2 -=- This chapter is very much Work in Progress

Continuous Data Across Categories using ggplot2

1. This chapter takes us a step further in our exploration of continuous data. Here, we delve into the use of ggplot2.
2. **Data**: Let us work with the same `mtcars` data from the previous chapter. Suppose we run the following code to prepare the data for subsequent analysis. The data is now in a tibble called `tb`:

```
# Load the required libraries, suppressing annoying startup messages
library(tibble)
suppressPackageStartupMessages(library(dplyr))
# Read the mtcars dataset into a tibble called tb
data(mtcars)
tb <- as_tibble(mtcars)
# Convert several numeric columns into factor variables
tb$cyl <- as.factor(tb$cyl)
tb$vs <- as.factor(tb$vs)
tb$am <- as.factor(tb$am)
tb$gear <- as.factor(tb$gear)
# Directly access the data columns of tb, without tb$mpg
attach(tb)
```

Summarizing Continuous Data across one Factor, using ggplot2

1. We investigate the bivariate Relationship between Miles Per Gallon (`mpg`) and Cylinders (`cyl`) using ggplot2.

```r
library(dplyr)

agg <- tb %>%
  group_by(cyl) %>%
  summarise(Avg_MPG = mean(mpg, na.rm = TRUE),
            SD_MPG = sd(mpg, na.rm = TRUE))
agg
```

```
# A tibble: 3 x 3
  cyl   Avg_MPG SD_MPG
  <fct>   <dbl>  <dbl>
1 4        26.7   4.51
2 6        19.7   1.45
3 8        15.1   2.56
```

2. Discussion:

- In this code, we are using the pipe operator %>% to perform a series of operations. We first group the data by the cyl column using the group_by() function. We then use summarise() to apply both the mean() and sd() functions to the mpg column.
- The results are stored in new columns, Avg_MPG and SD_MPG.
- Note that na.rm = TRUE is used in both mean() and sd() to remove missing values before calculation. This ensures that our calculations won't be disrupted by any missing data. If you are sure your data has no missing values, you may omit this.

Visualizing Continuous Data across one Factor, using ggplot2

Summarizing Continuous Data across two Factors using ggplot2

Visualizing Continuous Data across two Factors using ggplot2