XuChris_ZhangMengru_Assignment3

Chris Xu, Mengru Zhang

2025-11-01

```
# Setting the working directory
# Commented out because we each have different working directories
# setwd("~/Personal/Brown/BHDS2010/Assignment 3/Assignment3")

# Import libraries used in this analysis
library(tidyverse) # To use ggplot, filter, mutate, etc. tidyverse functions

# Read in the data from the csv
text_msg <- read.csv("TextMessages.csv")

# Visually examine the structure of the data
head(text_msg)
```

```
##   Group Baseline Six_months Participant
## 1     1       52         32           1
## 2     1       68         48           2
## 3     1       85         62           3
## 4     1       47         16           4
## 5     1       73         63           5
## 6     1       57         53           6
```

```
# The Group column shows the two groups of participants. The Baseline and Six_months
↪  columns contain the number of text messages measured at two separate time points.
↪  Lastly, the participant column contains the subject number.

# For visualization, we will create a long version of the text_msg dataset,
text_msg_long <- text_msg %>%
  pivot_longer(
    cols = c("Baseline", "Six_months"),
    names_to = "Timepoint",
    values_to = "n_Msg"
  )

# Print out the fist few rows of text_msg_long to verify the pivot is corrected processed
head(text_msg_long)
```

```
## # A tibble: 6 x 4
##   Group Participant Timepoint  n_Msg
##   <int>       <int> <chr>      <int>
## 1     1           1 Baseline      52
## 2     1           1 Six_months    32
## 3     1           2 Baseline      68
## 4     1           2 Six_months    48
## 5     1           3 Baseline      85
## 6     1           3 Six_months    62
```

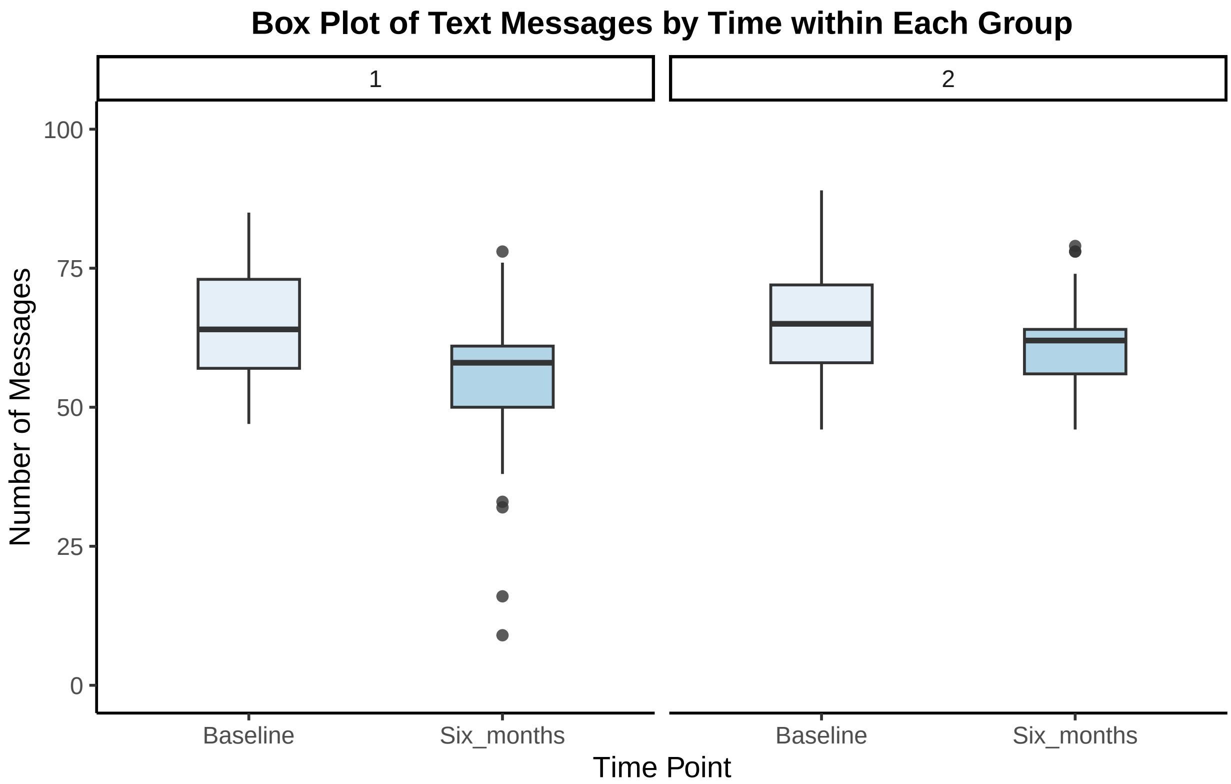
```
# Summary statistics: We compute the summary statistics of the data, including mean,
↪  median, count, standard deviation, standard error, minimum, and maximum.
summary_stats <- text_msg_long %>%
  group_by(Group, Timepoint) %>%
  summarise(
    n = n(),
    mean = mean(n_Msg, na.rm = TRUE),
    median = median(n_Msg, na.rm = TRUE),
    stdev = sd(n_Msg, na.rm = TRUE),
    se = stdev / sqrt(n),
    min = min(n_Msg, na.rm = TRUE),
    max = max(n_Msg, na.rm = TRUE)
  )

# Next we print out the results
print(summary_stats)
```

```
## # A tibble: 4 x 9
## # Groups:   Group [2]
##   Group Timepoint      n  mean median stdev    se   min   max
##   <int> <chr>      <int> <dbl>  <int> <dbl> <dbl> <int> <int>
## 1     1 Baseline      25  64.8     64 10.7   2.14    47    85
## 2     1 Six_months    25  53.0     58 16.3   3.27     9    78
## 3     2 Baseline      25  65.6     65 10.8   2.17    46    89
## 4     2 Six_months    25  61.8     62  9.41  1.88    46    79
```

```
# There are 25 data points in each group at each timepoint.
# Group 1 and Group 2 showed similar means and medians around 65, with standard deviation
↪  of 10.7-10.8. The minimum values of both groups are 46-47, and maximum are 85 and 89,
↪  respectively. Overall, the two groups display similar summary statistics at the
↪  Baseline timepoint. At the Six Months timepoint, the statistics are quite different.
↪  Group 1 has a mean of 53 and median of 58, while Group 2 has a mean of 61.8 and
↪  median of 62.
# For Group 1,  the standard deviation increased to 16.3, while for Group 2, the standard
↪  deviation slightly decreased to 9.41. Group 1 sees more extreme values on the
↪  downside, with a minimum of 9 text messages. Both groups saw maximum number of text
↪  messages near 80.

# Visualization 1: We first create box plots of text messages stratified by Group and
↪  Time
# Caption for chart to explain the data
vis_1_caption = "n = 25 for each group. The number of text messages a person typed were
↪  captured at two time points: baseline, and six months."
# Start a blank canvas, clarify the data on two axes
# Box plot, set width and color opacity
# Define faceted chart
# Choose color palette
# Add axis labels, title, and caption
# Choose theme
# Set title and caption location
text_msg_long %>% ggplot (aes(x = Timepoint, y = n_Msg, fill = Timepoint)) +
  geom_boxplot(width = 0.4, alpha = 0.8) +
  facet_grid(. ~ Group, switch = "y") +
  coord_cartesian(ylim = c(0, 100)) +
  scale_fill_brewer("Paired") +
  labs(x = "Time Point", y = "Number of Messages",
       title = "Box Plot of Text Messages by Time within Each Group",
       caption = str_wrap(vis_1_caption, width = 100)) +
  theme_classic() +
  theme(legend.position = "none",
        plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
        plot.caption.position = "plot",
        plot.caption = element_text(hjust = 0))
```

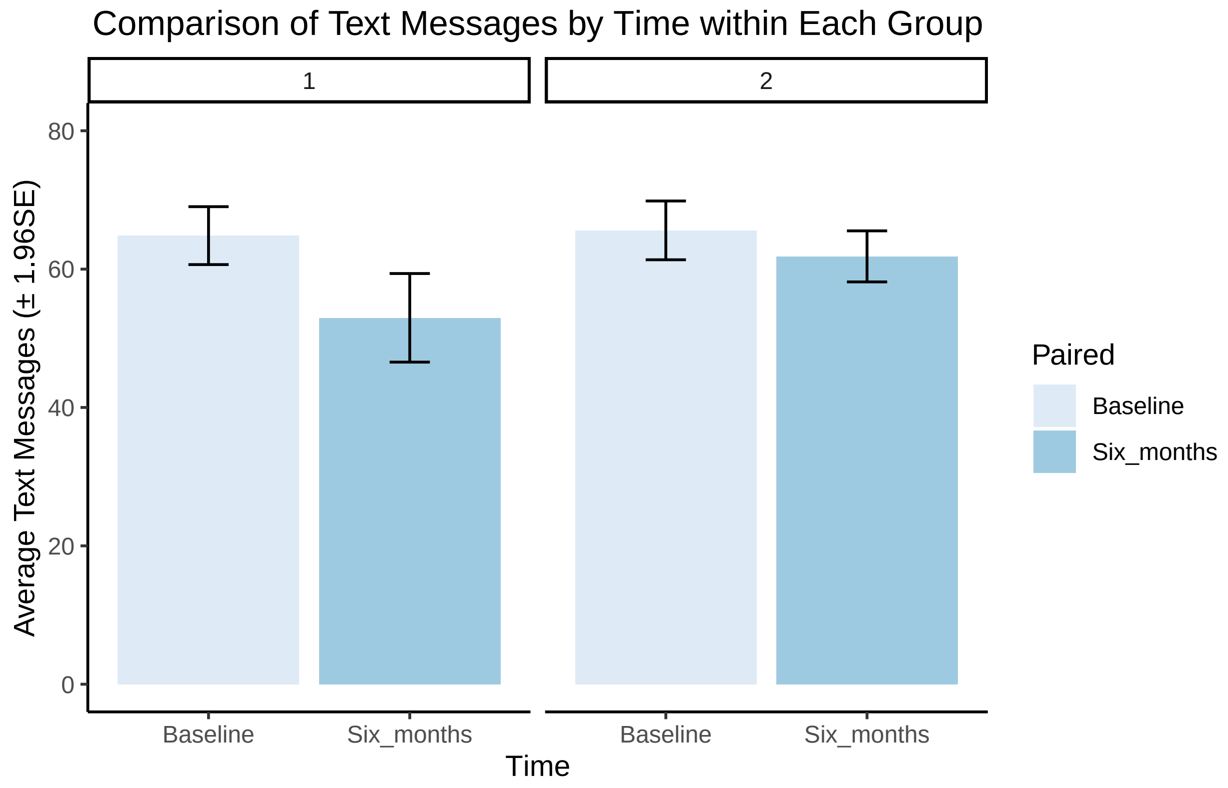


n = 25 for each group. The number of text messages a person typed were captured at two time points: baseline, and six months.

```
# The box plot shows that the number of text messages sent decreased from the Baseline
↪   observation to that six months later. The decrease appears more significant in Group
↪   1 than Group 2, which we will show later in the summary statistics section.
# The number of text messages six months later for Group 1 contains a fair amount of
↪   outliers on the downside, with the minimum being 9 messages.

# Visualization 2: We then create bar charts of text messages stratified by Group and
↪   Time
# Caption for chart to explain the data
vis_2_caption = "n = 25 for each group. The number of text messages a person typed were
↪   captured at two time points: baseline, and six months. Error bars indicating 95%
↪   confidence interval."
# Start a blank canvas, clarify the data on two axes
# Define a bar chart, where the height of the bars represents the means
```

```r
# Define error bars to be +-1.96 standard error, i.e. 95% CI
# Define faceted chart
# Add axis labels, title, and caption
# Set title and caption location
ggplot(summary_stats, aes(x = Timepoint, y = mean, fill = Timepoint)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(aes(ymin = mean - 1.96*se,
                    ymax = mean + 1.96*se),
                width = 0.2, position = position_dodge(width = 0.9)) +
  facet_wrap(~ Group) +
  coord_cartesian(ylim = c(0, 80)) +
  scale_fill_brewer("Paired") +
  labs(
    title = "Comparison of Text Messages by Time within Each Group",
    x = "Time",
    y = "Average Text Messages (± 1.96SE)",
    fill = "Time Period",
    caption = str_wrap(vis_2_caption, width = 100)
  ) +
  theme_classic() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "right",
    plot.caption.position = "plot",
    plot.caption = element_text(hjust = 0)
  )
```



n = 25 for each group. The number of text messages a person typed were captured at two time points: baseline, and six months. Error bars indicating 95% confidence interval.

```
# Group 1 showed a significant decrease in the average number of text messages sent,
↪  dropping from approximately 65 at Baseline to about 53 at the six-month mark. The
↪  error bars (which represent 95% CI) do not overlap, suggesting this change is
↪  statistically significant.
# Group 2 showed only a slight decrease in average text messages, from approximately 66
↪  at Baseline to 62 at six months. The error bars for these two time points overlap,
↪  suggesting this small drop is likely not statistically significant.


# In summary, while both groups started with a similar average, Group 1 experienced a
↪  much larger and more statistically significant reduction in text messages after six
↪  months compared to Group 2.
```