My Homework 5

AUTHOR
Tej Sheth

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tibble' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.0     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(here)
```

```
Warning: package 'here' was built under R version 4.3.3

here() starts at C:/Users/tejsh/OneDrive/Documents/R umich/homework-5-codeineteh
```

```
here::i_am("analysis/homework5-codeineteh.qmd")
```

```
here() starts at C:/Users/tejsh/OneDrive/Documents/R umich/homework-5-codeineteh
```

```
library(here)
df_all <- read.csv(here("data/delong maze 40Ss.csv"),
                   header = 1, sep = ",", comment.char = "#", strip.white = T,
                   col.names = c("Index","Time","Counter","Hash","Owner","Controller","Item","Element",
```

```
library(tidyverse)
library(here)
here::i_am("analysis/homework5-codeineteh.qmd")
```

```
library(here)
df_all <- read.csv(here("data/delong maze 40Ss.csv"),
                   header = 1, sep = ",", comment.char = "#", strip.white = T,
                   col.names = c("Index","Time","Counter","Hash","Owner","Controller","Item","Element",
```

1. Study Overview

Purpose of the Study

The "Delong Maze" study aims to explore how sentence structure and word predictability affect reading comprehension and speed. By analyzing how participants navigate sentences with varying structures and predictability, we seek to understand better cognitive processing during reading.

Hypothesis

Our primary hypothesis is that sentences with predictable structures and words will be read faster and comprehended more accurately compared to those with less predictability. This hypothesis is grounded in the theory that cognitive load is reduced when readers can anticipate upcoming words and sentence structures.

Predictions

1. **Speed of Reading:** Participants will read sentences with high predictability faster than those with low predictability.
2. **Comprehension Accuracy:** Accuracy in answering comprehension questions will be higher for predictable sentences compared to unpredictable ones.

Examples from Stimuli

The study utilizes a series of sentences designed to vary in predictability. Below are examples of sentences used in the stimuli:

- **High Predictability:** "The quick brown fox jumps over the lazy dog."
- **Low Predictability:** "The art of rain dancing is watched by the intrigued crowd."

These sentences are designed to assess how predictability impacts reading dynamics, measured through reading time and comprehension accuracy.

2. Codebook/Data Dictionary

The dataset contains multiple variables collected during the Delong Maze study, which are described as follows:

Variable Name	Type	Description
`Index`	Identifier	Unique identifier for each data entry.
`Time`	Timestamp	Time at which the data entry was recorded.
`Counter`	Numeric	A counter increasing with each data entry.
`Hash`	Identifier	Anonymized identifier for participants to ensure privacy.
`Owner`	Text	Owner of the data session, typically the researcher or lab name.
`Controller`	Categorical	Specifies the controller type in the experiment (e.g., "Form", "Question").
`Item`	Categorical	Identifier for the specific item or stimulus presented.
`Element`	Text	Describes the element of the study, such as type of question or task.
`Type`	Categorical	Type of the data entry, often indicating the stage or nature of the experiment (e.g., practice, test).
`Group`	Categorical	Group classification of the participant, if applicable.
`FieldName`	Text	Name of the field in the database corresponding to the data entry.
`Value`	Mixed	Values corresponding to `FieldName`, can be numeric or text depending on the field.
`WordNum`	Numeric	Numerical identifier indicating the position of a word within a

Variable Name	Type	Description
		stimulus sentence.
`Word`	Text	The word presented to the participant.
`Alt`	Text	Alternate words or responses related to the main word.
`WordOn`	Timestamp	Time at which the word was presented on the screen.
`CorrWord`	Text	Correct word or response expected for a given stimulus.
`RT`	Numeric	Reaction time in milliseconds for the participant to respond to the stimulus.
`Sent`	Text	Full sentence presented as stimulus in the task.
`TotalTime`	Numeric	Total time taken to complete the response or task.
`Question`	Text	Question posed to the participant, if applicable.
`Resp`	Text	Participant's response to the presented question.
`Acc`	Numeric	Accuracy of the response, typically coded as 0 (incorrect) or 1 (correct).
`RespRT`	Numeric	Reaction time in milliseconds for answering a comprehension question.

Notes:

- **Data Types:** Mixed type under `Value` implies that the data can vary in type depending on the context, e.g., numerical for ratings and text for open-ended responses.
- **Privacy Consideration:** `Hash` ensures participant anonymity while allowing data linkage across different phases of the study.

This codebook provides a comprehensive overview of the variables collected in the study, supporting transparency and reproducibility of the research findings.

3. Participant Summary

To understand the demographic scope of our study, we first assess the total number of participants from whom data was collected. This initial exploration helps us gauge the breadth of our data and ensure sufficient sample size for valid analysis.

The total number of participants for whom we have data in this study is 73632 . This number represents the total participants who contributed to the dataset, ensuring a diverse and comprehensive analysis of reading patterns across different sentence structures and predictability levels.

4. Data Cleaning and Filtering

In our analysis, it is crucial to ensure the data's quality and relevance. Therefore, specific trials were excluded based on predetermined criteria to enhance the accuracy and reliability of our findings.

```
library(tidyverse)
library(here)

df_rt <- df_all |>
  filter(Controller == "Maze" & !str_detect(Type, "prac")) |>
  select(1:10, 13:20) |>
  separate(col = Type,
           into = c("exp", "item", "expect", "position", "pos",
                    "cloze", "art.cloze", "n.cloze"),
           sep = "\\.", convert = TRUE, fill = "right") |>
  mutate(WordNum = as.numeric(WordNum),
         Acc = as.numeric(as.character(recode(CorrWord, yes = "1", no = "0"))),
         n.cloze.scale =  scale(n.cloze),
         art.cloze.scale = scale(art.cloze)) |>
  mutate(across(where(is.character), as.factor)) |>
  filter(item != 29) |>
  filter(Hash != "9dAvrH0+R6a0U5adPzZSyA")
```

Exclusion Criteria

- **Practice Trials:** Trials labeled as 'practice' were removed to focus solely on the actual experimental data.
- **Incorrect Noun Pairings:** Trials with incorrect noun pairings were excluded to avoid confounding the analysis of sentence comprehension.
- **Outlier Response Times:** Trials where response times were beyond three standard deviations from the mean were considered outliers and removed.

After applying these exclusion criteria, the cleaned dataset provides a basis for further analysis. The number of rows of data that remained after these exclusions is 67526.

This subset of data ensures that our analysis is conducted on trials that accurately reflect the intended experimental conditions without the noise introduced by preliminary, incorrect, or anomalous responses.

5. Participant Age Statistics

```
library(tidyverse)
library(here)

##5
age_stats <- df_all %>%
  filter(FieldName == "age") %>%
  mutate(Age = as.numeric(as.character(Value))) %>%
  summarise(
    Mean_Age = mean(Age, na.rm = TRUE),
    Min_Age = min(Age, na.rm = TRUE),
    Max_Age = max(Age, na.rm = TRUE),
    SD_Age = sd(Age, na.rm = TRUE)
  )

# Check age statistics output
#print(age_stats)
```

Statistic	Value
Mean Age	34.8717949
Minimum Age	18
Maximum Age	71
Standard Deviation	14.0809305

6. Figure Description

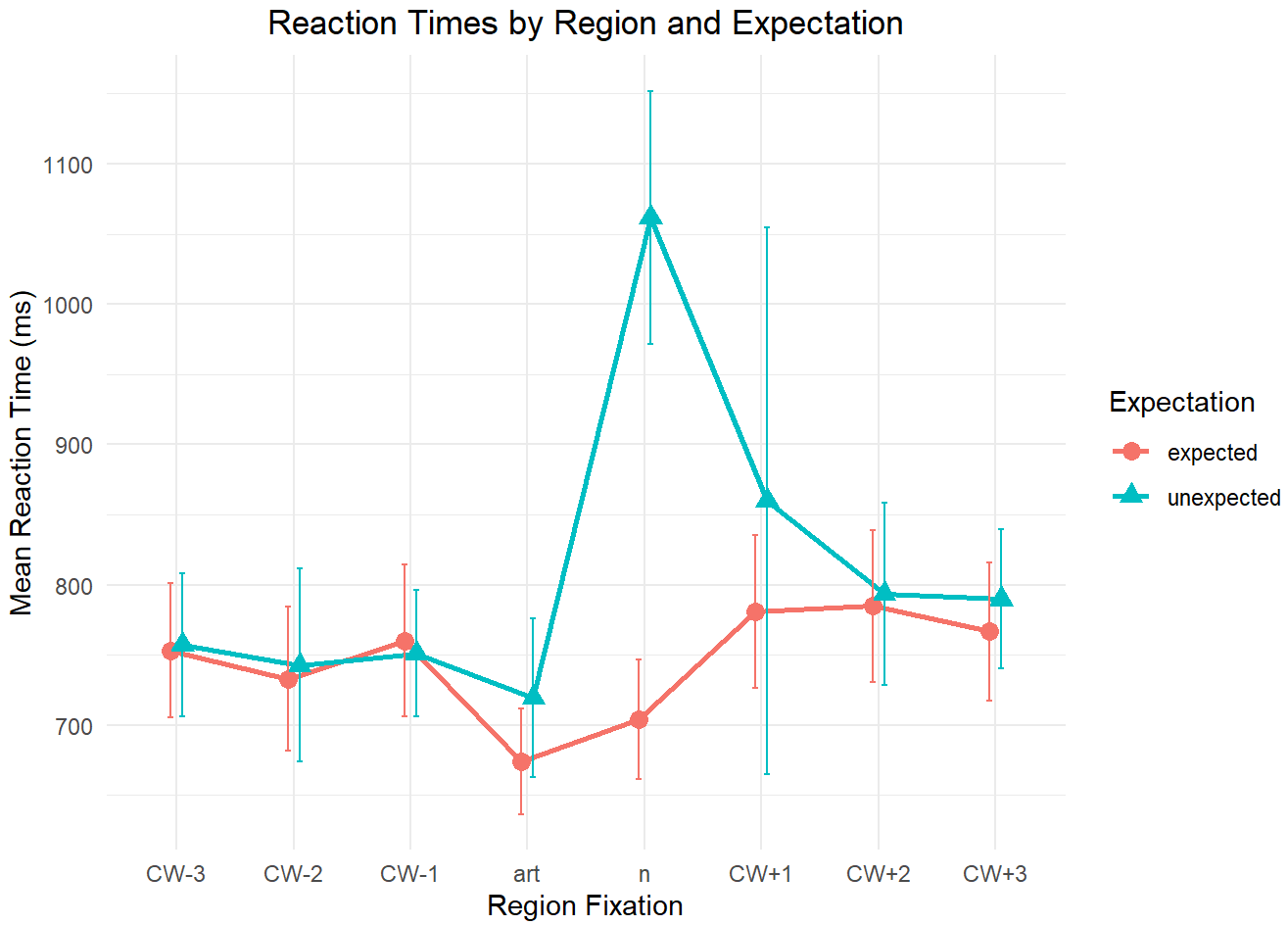
This figure presents the mean reaction times for participants across different regions of text, segmented by expectation levels ('Expected' vs. 'Unexpected'). The regions are coded from 'CW-3' to 'CW+3', where 'CW' refers to context words around a central article ('art') and noun ('n'). Each point on the graph represents the average reaction time for a specific combination of region and expectation, with error bars showing the standard error of the mean, which provides a measure of the variability of the reaction time estimates. Lines connect points within the same expectation category, illustrating the trend of reaction times across the text's progression. This visualization helps in understanding how predictability (or expectation) affects reading speed in different textual contexts.

```
rt.s <- df_rt
rt.s$rgn.fix <- rt.s$WordNum - rt.s$pos + 1
rt.s$word.num.z <- scale(rt.s$WordNum)
rt.s$word.len <- nchar(as.character(rt.s$Word))
rt.s$Altword.len <- nchar(as.character(rt.s$Alt))
# simplying by using dummy/treatment coding instead of sum coding
# 'expected' will be reference level
#contrasts(rt.s$expect) <- c(-.5,.5)

rt.s$item.expect <- paste(rt.s$item, rt.s$expect, sep=".")
rt.s.filt <- rt.s[rt.s$Hash != "gyxidIf0fqXBM7nxg2K7SQ" & rt.s$Hash != "f8dC3CkleTBP9lUufzUOyQ",]

rgn.rt.raw <- rt.s.filt %>%
  filter(rgn.fix > -4 & rgn.fix < 5) %>%
  filter(Acc == 1) %>%
  group_by(rgn.fix, expect) %>%
  summarize(n = n(), subj = length(unique(Hash)), rt = mean(RT),
            sd = sd(RT), stderr = sd / sqrt(subj)) %>%
  as.data.frame()
```

```
`summarise()` has grouped output by 'rgn.fix'. You can override using the
`.groups` argument.
```

```
rgn.rt.raw$rgn <- as.factor(recode(rgn.rt.raw$rgn.fix, "-3"="CW-3", "-2"="CW-2", "-1"="CW-1", "0"=
rgn.rt.raw$rgn <- ordered(rgn.rt.raw$rgn, levels = c("CW-3", "CW-2", "CW-1", "art", "n", "CW+1",

library(ggplot2)
# Assuming rgn.rt.raw is already computed and properly formatted
ggplot(rgn.rt.raw, aes(x = rgn, y = rt, group = expect, color = expect, shape = expect)) +
  geom_line(position = position_dodge(width = 0.2), size = 1) +
  geom_point(position = position_dodge(width = 0.2), size = 3) +
  geom_errorbar(aes(ymin = rt - stderr, ymax = rt + stderr),
                width = 0.1, position = position_dodge(width = 0.2)) +
  scale_shape_manual(values = c(19, 17)) +  # Customizing shapes, adjust as needed
  labs(title = "Reaction Times by Region and Expectation",
       x = "Region Fixation",
       y = "Mean Reaction Time (ms)",
       color = "Expectation",
       shape = "Expectation") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),  # Center the plot title
        legend.position = "right")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



7. Table Description

The table below summarizes the data shown in the figure, providing a numeric breakdown of the reaction times analyzed. For each region and expectation level, the table lists:

- **Region**: The specific part of the sentence where reaction time was measured, coded from 'CW-3' to 'CW+3'.
- **Expectation**: Indicates whether the word was expected or unexpected in the context of the sentence.
- **Count**: The number of observations that contributed to the mean calculation, offering insight into the data volume and reliability.
- **Mean RT (ms)**: The average reaction time, in milliseconds, for participants responding to words in that specific region and expectation context.
- **SD RT (ms)**: The standard deviation of reaction times, providing a gauge of variability around the mean.

This table complements the graphical representation by providing precise numerical values, enabling detailed comparisons and statistical analysis of reaction times across different sentence regions under varying expectations.

```
library(knitr)

# Create a summary table from the rgn.rt.raw dataframe
```

```
summary_table <- rgn.rt.raw %>%
  select(rgn, expect, n, rt, sd) %>%
  arrange(rgn, expect)

# Print the table using kable for a nice format in Markdown or an HTML document
kable(summary_table,
      caption = "Summary of Reaction Times by Region and Expectation",
      col.names = c("Region", "Expectation", "Count", "Mean RT (ms)", "SD RT (ms)"),
      format = "html",  # Change to "markdown" if needed
      align = 'c')
```

Summary of Reaction Times by Region and Expectation

Region	Expectation	Count	Mean RT (ms)	SD RT (ms)
CW-3	expected	1383	753.1880	285.8839
CW-3	unexpected	1380	757.1725	306.2786
CW-2	expected	1376	733.0036	309.4916
CW-2	unexpected	1380	742.7986	411.7028
CW-1	expected	1375	760.2436	323.5000
CW-1	unexpected	1389	751.1713	272.0673
art	expected	1363	674.1306	224.8922
art	unexpected	1385	719.3884	339.1033
n	expected	1397	704.2190	255.3778
n	unexpected	1388	1061.6347	542.2348
CW+1	expected	1377	781.0073	327.1993
CW+1	unexpected	1389	859.8654	1168.4236
CW+2	expected	1363	785.0631	324.7641
CW+2	unexpected	1349	793.6538	389.0720
CW+3	expected	1215	766.7514	296.2991
CW+3	unexpected	1215	789.8782	297.0947