Exercise: Grouping and Summarizing Data Frames

Table of Contents

[1 Overview 2](#_Toc159501700)

[2 Practice with the Symmetry Span Task 2](#_Toc159501701)

[2.1 Reading Data 2](#_Toc159501702)

[2.2 Taking Inventory of Variables 2](#_Toc159501703)

[2.3 Grouping Data Frames 2](#_Toc159501704)

[2.4 Mutating Variables without Grouping 2](#_Toc159501705)

[2.5 Grouping and Mutating Data Frames 2](#_Toc159501706)

[2.6 Grouping and Summarizing Data Frames 3](#_Toc159501707)

[2.7 Grouping, Mutating, and Summarizing 3](#_Toc159501708)

[2.8 Grouping, Summarizing, Grouping, and Summarizing 3](#_Toc159501709)

[3 Bonus 3](#_Toc159501710)

[4 Practice with the Go/No-Go Task 5](#_Toc159501711)

[4.1 Reading Data 5](#_Toc159501712)

[4.2 Taking Inventory of Variables 5](#_Toc159501713)

[4.3 Grouping Data Frames 5](#_Toc159501714)

[4.4 Mutating Variables without Grouping 5](#_Toc159501715)

[4.5 Grouping and Mutating Data Frames 5](#_Toc159501716)

[4.6 Grouping and Summarizing Data Frames 5](#_Toc159501717)

[4.7 Grouping, Mutating, and Summarizing 6](#_Toc159501718)

[4.8 Grouping, Summarizing, Grouping, and Summarizing 6](#_Toc159501719)

[5 Bonus 6](#_Toc159501720)

# 1 Overview

In a future exercise, you will work on pivoting the wide data produced by your sub-setting script from a previous exercise into long format. Once you have that long file, you will need to verify it’s content, clean it up by cleaning rows or by computing or selecting relevant variables. You will also then need to filter data in some way and produce descriptive statistic summaries at very least.

For this exercise, you will read data, select variables to subset the data, filter cases/observations, group, and and summarize the data of a long file.

Data files:

* “gng\_long\_clean.Rds”
* “sspan\_long\_clean.Rds”

# 2 Practice with the Symmetry Span Task

## 2.1 Reading Data

Read “sspan\_long\_clean.Rds” and assign to an object.

## 2.2 Taking Inventory of Variables

Examine the structure of your data frame using **{dplyr}** so that you know variables that you might summarize and the variables you might group by. Load your necessary libraries.

Make a note of the “grouping structure”. Also examine for any oddities that you might need to address.

## 2.3 Grouping Data Frames

Take the data frame and pipe it to a grouping variable (e.g., id\_wave) and then pipe to examine the structure. Make note of how the tibble’s grouping structure has changed.

## 2.4 Mutating Variables without Grouping

Add a new variable to the existing data frame that represents the mean of sspan. Code your function to exclude NA values when computing the mean. Do not overwrite the data frame.

*Describe what the mean represents:*

## 2.5 Grouping and Mutating Data Frames

Using that same grouping structure used earlier (e.g., id\_wave), group the data and then pipe the data frame and add a new variable to the data frame that represents the mean of sspan. Code your function to exclude NA values. Do not overwrite the data frame.

*Describe what the mean represents:*

*Compare the values of the mean variable added to the data frame with and without the grouping. Is the calculated variable in one of the data frames more similar to what you expected when you mutated the variable? If so, which one?*

## 2.6 Grouping and Summarizing Data Frames

Using that same grouping structure that you just used, rather than add a new variable to the data frame that represents the mean of sspan, summarize the data frame. Code your function to exclude NA values. Do not overwrite the data frame.

## 2.7 Grouping, Mutating, and Summarizing

Using that same (a) grouping structure that you just used, (b) add a new variable to the data frame that represents the mean of sspan, (c) then summarize the data frame by that *same* variable. Code your function to exclude NA values. Do not overwrite the data frame.

## 2.8 Grouping, Summarizing, Grouping, and Summarizing

In many cases, you may need to group data frames in one way in order to obtain data summaries which will you will further summarize at a more general level. For example, you may need to aggregate your data in order to obtain average performance at a participant level so that you can further aggregate individuals within a group in service of obtaining group-level summaries.

In order to understand the difference in aggregation techniques, we will group the data two ways.

1. Take your data frame and (a) group by id\_school, and then (b) summarize the data frame so that your new data frame contains the mean of sspan at the school level. Do not overwrite the data frame.
2. Next, take your data frame and (a) group by id\_school and id\_subject, (b) summarize the data frame so that your new data frame contains the mean of sspan for each participant in each school, (c) group again but only by the school, (d) summarize the data frame so that your new data frame contains the mean of sspan at the school level. Do not overwrite the data frame.

*Describe what the differences in the summaries and why they exist.*

# 3 Bonus

1. Practice summarizing the data using different metrics (e.g., standard deviation, standard error of the mean, median, etc.).
2. Practice summarizing the data using different variables errors, speed errors, etc.
3. Practice summarizing the data by grouping the data different ways (e.g., cue, target, soa, etc.).
4. All of the above approaches will be appropriate for working with your data frame. Your liaison may want the data summarized in different ways, so you should certainly want to verify from them how they want their data grouped and summarized. Additionally, only some aggregation approaches are relevant for certain statistical models which will later run. You should certainly start to consider obtaining clarification about ways they want the data summarized so that you are able to summarize the data appropriately when the time arises.
5. Consider whether your data make sense. Do you need to create any variables or filter your data frame in order to take care of any problems?

# 4 Practice with the Go/No-Go Task

## 4.1 Reading Data

Read “gng\_long\_clean.Rds” and assign to an object.

## 4.2 Taking Inventory of Variables

Examine the structure of your data frame using **{dplyr}** so that you know variables that you might summarize and the variables you might group by. Load your necessary libraries.

Make a note of the “grouping structure”. Also examine for any oddities that you might need to address.

## 4.3 Grouping Data Frames

Take the data frame and pipe it to a grouping variable (e.g., id\_wave) and then pipe to examine the structure. Make note of how the tibble’s grouping structure has changed.

## 4.4 Mutating Variables without Grouping

Add a new variable to the existing data frame that represents the mean of accuracy. Code your function to exclude NA values when computing the mean. Do not overwrite the data frame.

*Describe what the mean represents:*

## 4.5 Grouping and Mutating Data Frames

Using that same grouping structure used earlier (e.g., id\_wave), group the data and then pipe the data frame and add a new variable to the data frame that represents the mean of sspan or accuracy. Code your function to exclude NA values. Do not overwrite the data frame.

*Describe what the mean represents:*

*Compare the values of the mean variable added to the data frame with and without the grouping. Is the calculated variable in one of the data frames more similar to what you expected when you mutated the variable? If so, which one?*

## 4.6 Grouping and Summarizing Data Frames

Using that same grouping structure that you just used, rather than add a new variable to the data frame that represents the mean of accuracy, summarize the data frame. Code your function to exclude NA values. Do not overwrite the data frame. Do not overwrite the data frame.

## 4.7 Grouping, Mutating, and Summarizing

Using that same (a) grouping structure that you just used, (b) add a new variable to the data frame that represents the mean of accuracy, (c) then summarize the data frame by that *same* variable. Code your function to exclude NA values. Do not overwrite the data frame.

## 4.8 Grouping, Summarizing, Grouping, and Summarizing

In many cases, you may need to group data frames in one way in order to obtain data summaries which will you will further summarize at a more general level. For example, you may need to aggregate your data in order to obtain average performance at a participant level so that you can further aggregate individuals within a group in service of obtaining group-level summaries.

In order to understand the difference in aggregation techniques, we will group the data two ways.

1. Take your data frame and (a) group by id\_school, and then (b) summarize the data frame so that your new data frame contains the mean of accuracy at the school level. Do not overwrite the data frame.
2. Next, take your data frame and (a) group by id\_school and id\_subject, (b) summarize the data frame so that your new data frame contains the mean of accuracy for each participant in each school, (c) group again but only by the school, (d) summarize the data frame so that your new data frame contains the mean of sspan at the school level. Do not overwrite the data frame.

*Describe what the differences in the summaries and why they exist.*

# 5 Bonus

1. Practice summarizing the data using different metrics (e.g., standard deviation, sample size, standard error of the mean, median, etc.).
2. Practice summarizing the data using different variables (e.g., rt).
3. Practice grouping the data different ways (e.g., cue, target, soa, etc.).
4. All of the above approaches will be appropriate for working with your data frame. Your liaison may want the data summarized in different ways, so you should certainly want to verify from them how they want their data grouped and summarized. Additionally, only some aggregation approaches are relevant for certain statistical models which will later run. You should certainly start to consider obtaining clarification about ways they want the data summarized so that you are able to summarize the data appropriately when the time arises.
5. Consider whether your data make sense. Do you need to create any variables or filter your data frame in order to take care of any problems?