RStudio_082923

2023-08-29

In this worksheet, we will go over Single-Subject Research Designs. In particular, we will examine various design considerations, create plots, and perform visual/statistical analysis.

Problem Statement

At a school that you run, there is a study period at the end of the day. Ideally, this time is used for studying and completing homework, but the teachers notify you that students are mostly off-task and are falling behind on homework. You decide to implement a program that rewards studying. You propose that if students study for more than 75% of the time, they will receive a small treat (e.g. candy or a sticker). Such a program would be somewhat costly, but if it works, it will help students tremendously. You decide to implement the program into a single classroom but are unsure of how to move forward. What kind of single-subject study would you implement and how would you consider doing any analysis?

Preliminaries

We will start by importing any libraries that may be useful.

```r
# Import libraries
library(ggplot2)
```

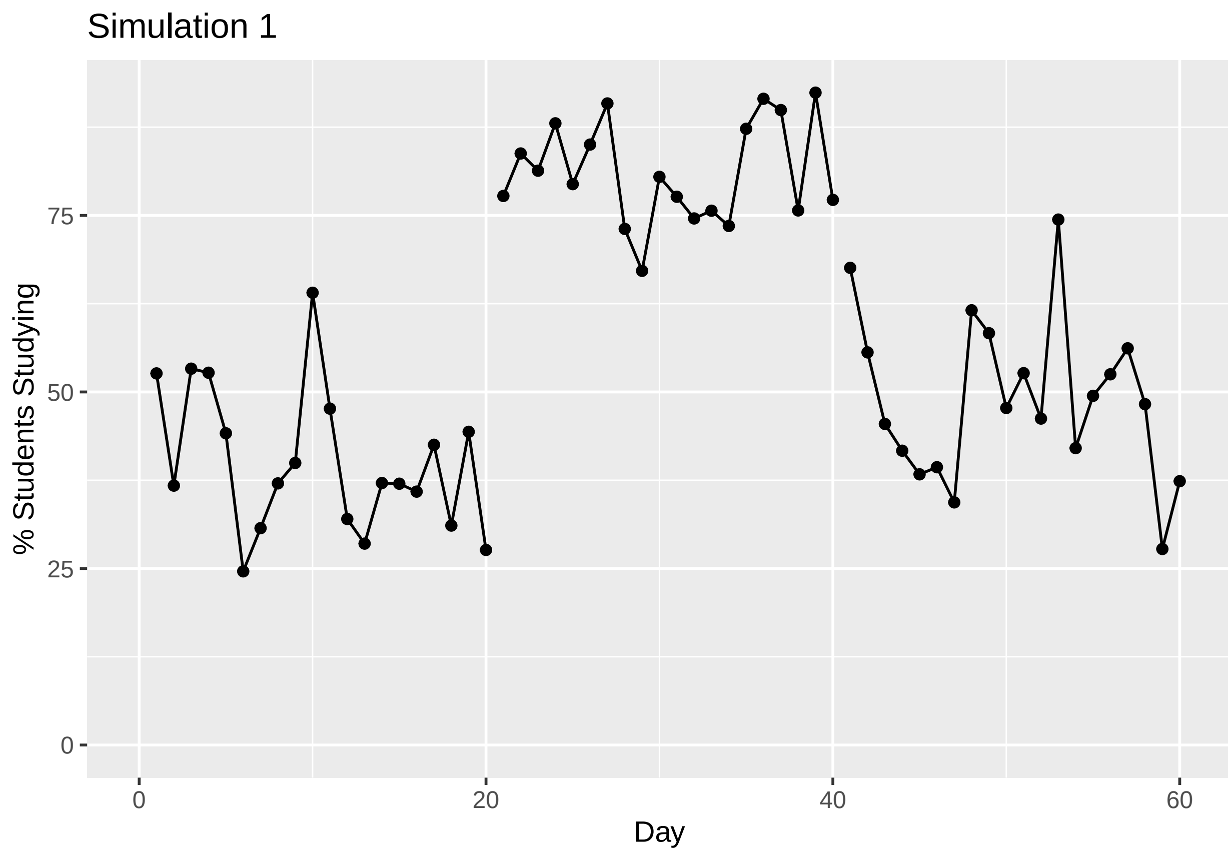
The Reversal Design

For ease of notation, we designate A as the default condition as B as the treatment condition. In the reversal design, across three phases, the conditions implemented are A, B, and A, respectively. This is known as the **ABA** or **Reversal** design since the treatment is "reversed" in the third phase.

Let's simulate what the data for this design may look like. We note that each condition is implemented for 20 days each.

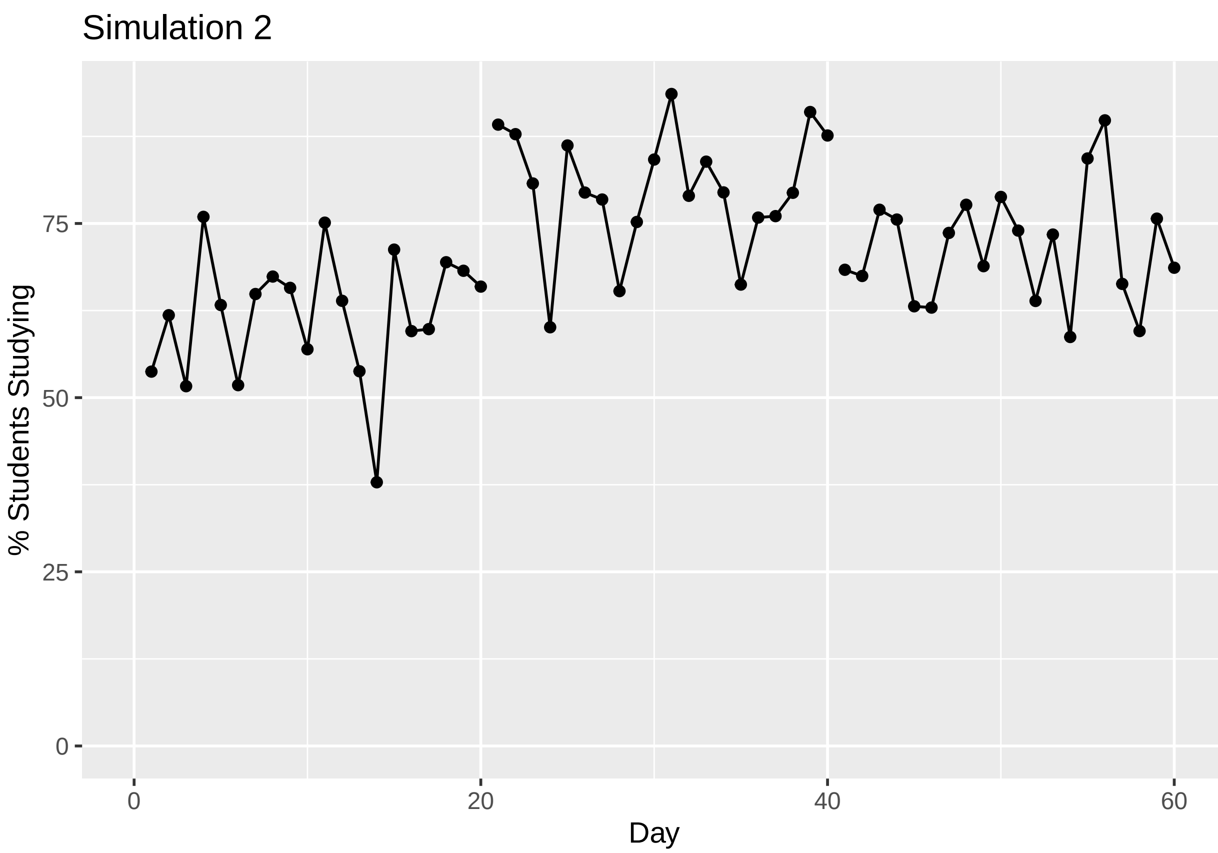
```r
# Generate random data
set.seed(0) # Set seed to make data reproducible
n = 20 # Number of days
A1 = rnorm(n, mean = 0.4, sd = 0.1) * 100
B2 = rnorm(n, mean = 0.8, sd = 0.1) * 100
A3 = rnorm(n, mean = 0.5, sd = 0.1) * 100
data = c(A1, B2, A3)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n))
df = data.frame(data, time, condition)
```

```
# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
  ylab("% Students Studying") +
  ggtitle("Simulation 1")
```




```
# Generate random data
set.seed(1) # Set seed to make data reproducible
n = 20 # Number of days
A1 = rnorm(n, mean = 0.6, sd = 0.1) * 100
B2 = rnorm(n, mean = 0.8, sd = 0.1) * 100
A3 = rnorm(n, mean = 0.7, sd = 0.1) * 100
data = c(A1, B2, A3)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n))
df = data.frame(data, time, condition)

# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
```

```
  ylab("% Students Studying") +
  ggtitle("Simulation 2")
```



Question 1

Between the 2 simulations, which suggests that the treatment condition B was effective?

Question 2

What would be the advantages and disadvantages of expanding the reversal design (i.e. run an ABAB or ABABA design)?

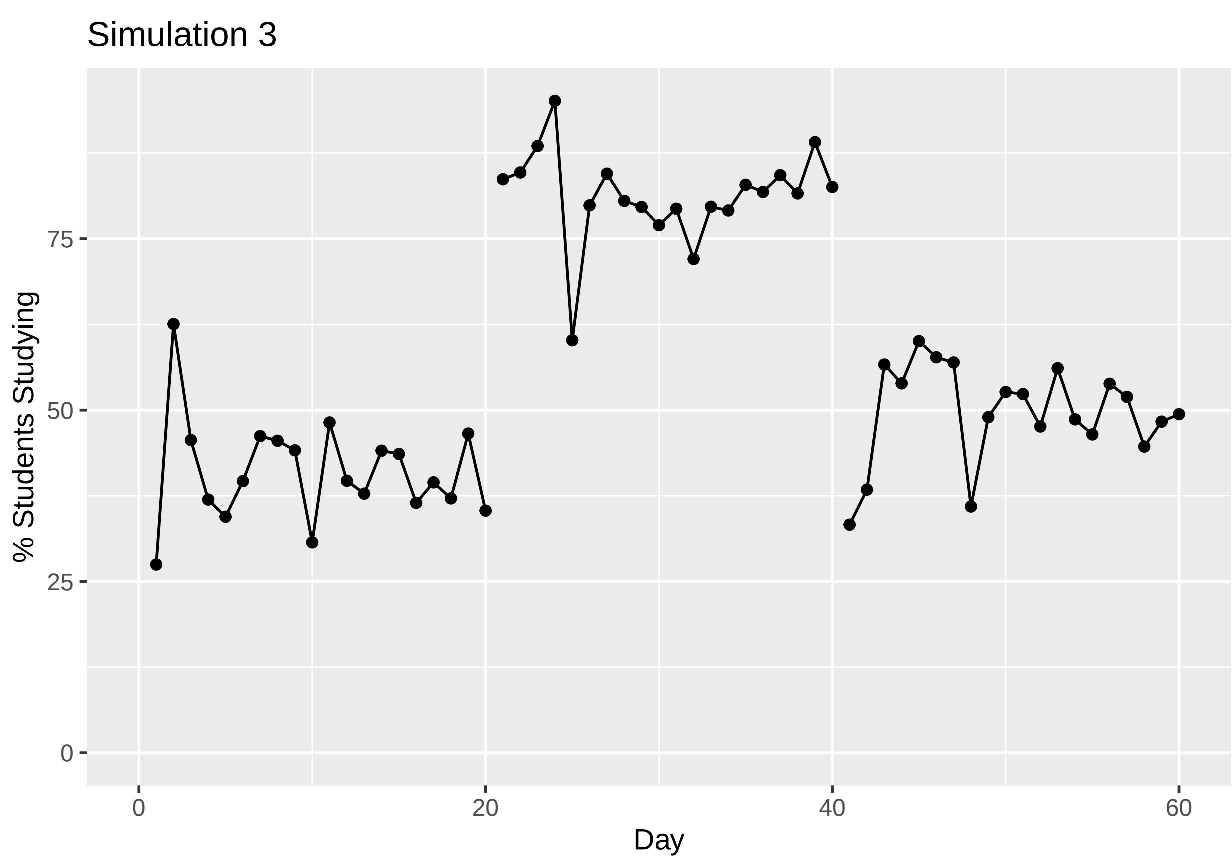
Question 3

Consider the simulated data below. What makes it different from the simulated data above?

(Hint: Consider the variation in the data over time)

```
# Generate random data
set.seed(1) # Set seed to make data reproducible
n = 20 # Number of days
A1 = c()
B2 = c()
A3 = c()
```

```
for (i in 1:n){
  A1 = c(A1, rnorm(1, mean = 0.4, sd = 0.2 / (i^(1/2))) * 100)
  B2 = c(B2, rnorm(1, mean = 0.8, sd = 0.2 / (i^(1/2))) * 100)
  A3 = c(A3, rnorm(1, mean = 0.5, sd = 0.2 / (i^(1/2))) * 100)
}
data = c(A1, B2, A3)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n))
df = data.frame(data, time, condition)

# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
  ylab("% Students Studying") +
  ggtitle("Simulation 3")
```



Question 4

Based on your response from Question 3, what is one consideration you could make when selecting data to keep for any statistical analysis?

Variations of the Reversal Design

There may be cases where you want to try out different treatments. This is known as the **multiple-treatment reversal design**. Let's consider the conditions A, B, and C, where C is another treatment. One potential design is the *ABCACB* design. In this design, we start with a baseline, then introduce conditions B and C, in that order. We then return to the baseline, then introduce conditions C and B, in that order.

Question 5

Why might it be important to reverse the order of the two treatments?

There may also be the case where you will have to implement a treatment no matter what. In this case, you may have to choose between two treatments. So, a baseline may not be needed. This is known as the **alternating treatments design**. One potential design is the *BCBC* design. However, since implementing a treatment is costly, the treatments should be switched frequently.

Question 6

In the alternating treatment design, one criteria is that the treatments should be fast acting. Why?

Question 7

Based on the problem statement in the beginning, come up with a new treatment condition. Then, propose a way to implement a reversal design, a multiple-treatment reversal design, and an alternating treatments design. Note how many days each condition should be implemented.

Analysis in Single-Subject Research Designs

It turns out that single-subject research relies on visual inspection more than the traditional inferential statistics. This is because in order to generalize results for a population, we would need to examine multiple subjects and look at group data. Nonetheless, some inferential statistics can still be used in tandem with visual inspection.

Visual Inspection

One possible feature to look out for is **level**. This is perhaps the most obvious feature to look out for in which we look for how much higher or lower the dependent variable is between the conditions.

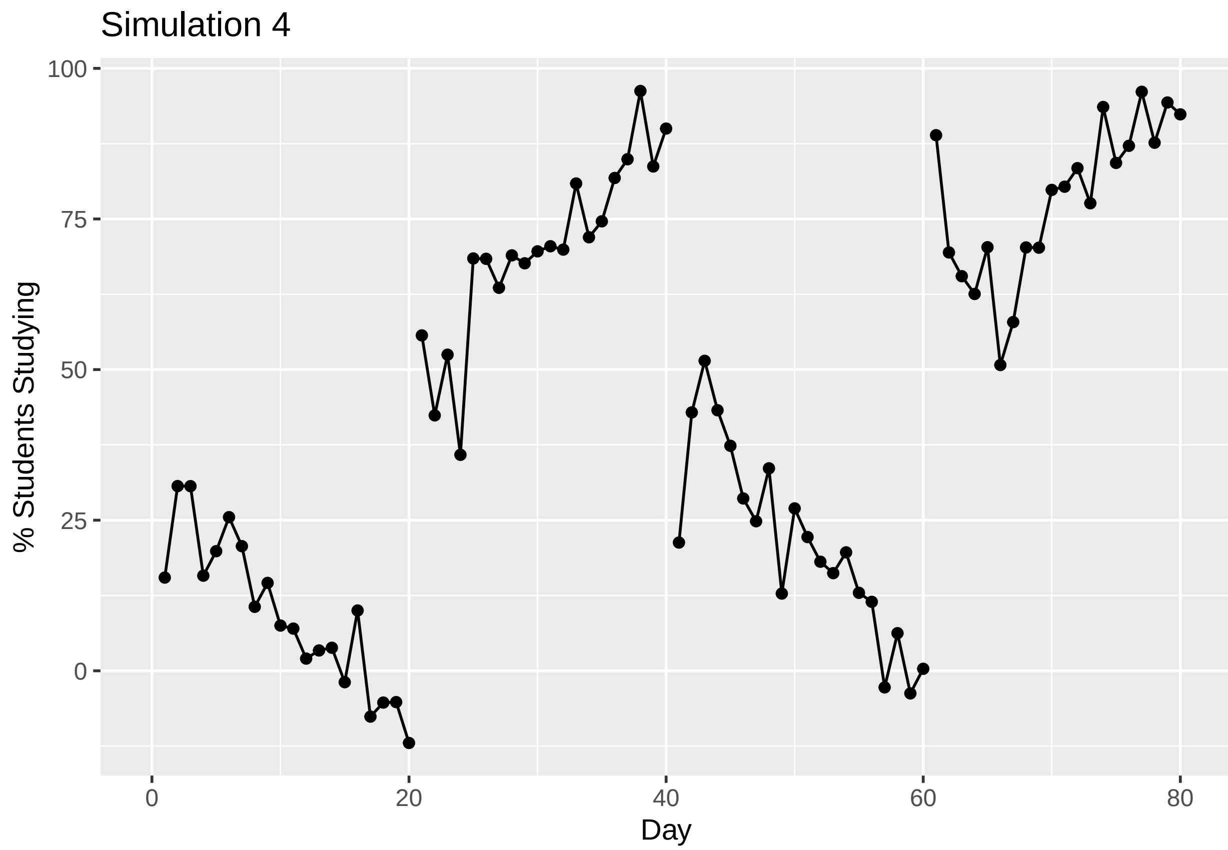
Another possible feature is to look for a **trend**. We would look for cases where the dependent variable increases or decreases within a condition. This is especially notable when the trends go in opposite directions for varying conditions.

One last feature to look for is **latency**. This is probably the most tricky feature to look out for. If we notice a sudden change with the introduction of a condition, then this suggests that the change in condition was responsible for the change. But if the change occurs more slowly, then it becomes more likely that a combination of other external factors are responsible for the change.

Question 8

Comment on the level, trend, and latency in the simulated data for the ABAB design.

```r
# Generate random data
set.seed(1) # Set seed to make data reproducible
n = 20 # Number of days
A1 = c()
B2 = c()
A3 = c()
B4 = c()
for (i in 1:n){
  A1 = c(A1, rnorm(1, mean = 0.3 - i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  B2 = c(B2, rnorm(1, mean = 0.5 + i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  A3 = c(A3, rnorm(1, mean = 0.4 - i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  B4 = c(B4, rnorm(1, mean = 0.55 + i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
}
data = c(A1, B2, A3, B4)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n), rep("B4", n))
df = data.frame(data, time, condition)

# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
  ylab("% Students Studying") +
  ggtitle("Simulation 4")
```



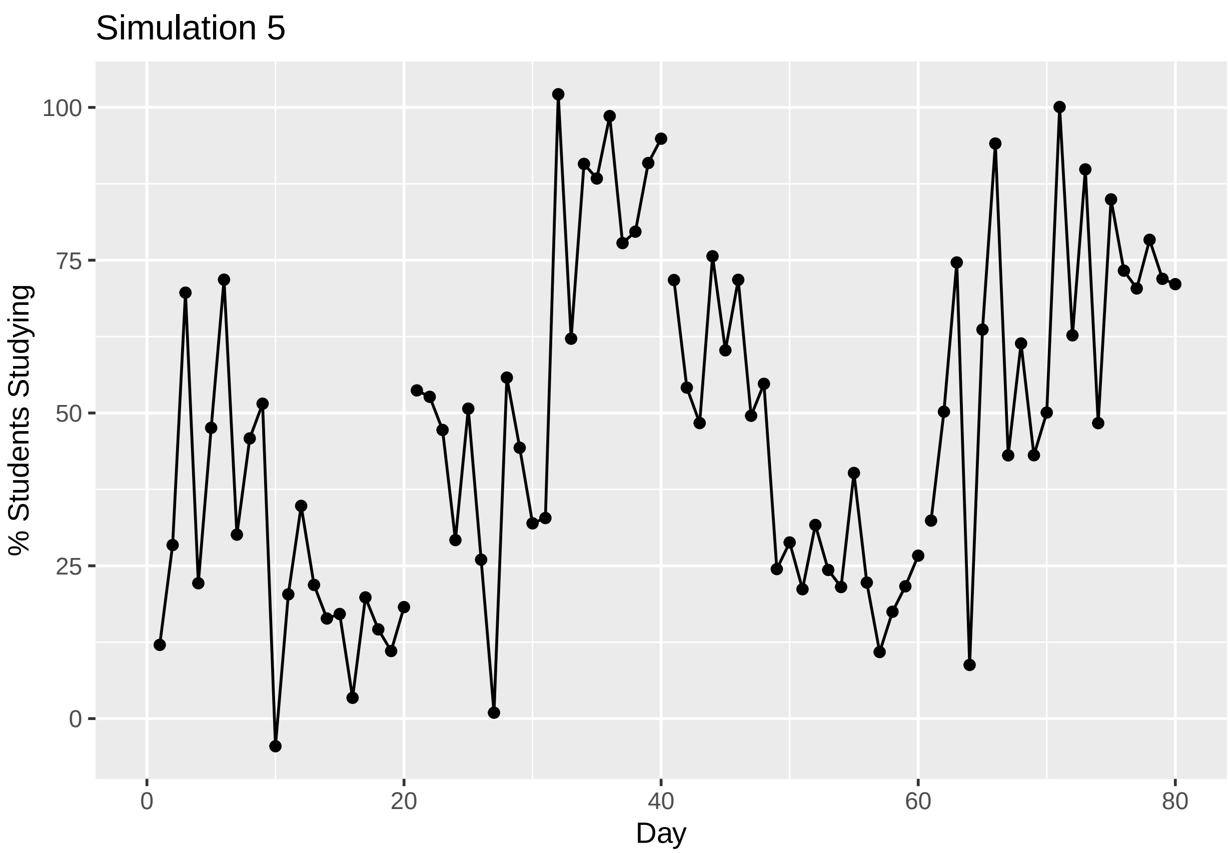
Question 9

Comment on the level, trend, and latency in the simulated data for the ABAB design.

(Hint: For latency, think about how long it takes for a trend to become easy to spot)

```
# Generate random data
set.seed(2) # Set seed to make data reproducible
n = 20 # Number of days
A1 = c()
B2 = c()
A3 = c()
B4 = c()

for (i in 1:(n/2)){
  A1 = c(A1, rnorm(1, mean = 0.3, sd = 0.2) * 100)
  B2 = c(B2, rnorm(1, mean = 0.5, sd = 0.2) * 100)
  A3 = c(A3, rnorm(1, mean = 0.4, sd = 0.2) * 100)
  B4 = c(B4, rnorm(1, mean = 0.55, sd = 0.2) * 100)
}

for (i in 1:(n/2)){
  A1 = c(A1, rnorm(1, mean = 0.3 - i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  B2 = c(B2, rnorm(1, mean = 0.7 + i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  A3 = c(A3, rnorm(1, mean = 0.4 - i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
  B4 = c(B4, rnorm(1, mean = 0.6 + i * 0.02, sd = 0.2 / (i^(1/2))) * 100)
```

```
}

data = c(A1, B2, A3, B4)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n), rep("B4", n))
df = data.frame(data, time, condition)

# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
  ylab("% Students Studying") +
  ggtitle("Simulation 5")
```



Percentage of Nonoverlapping Data (PND)

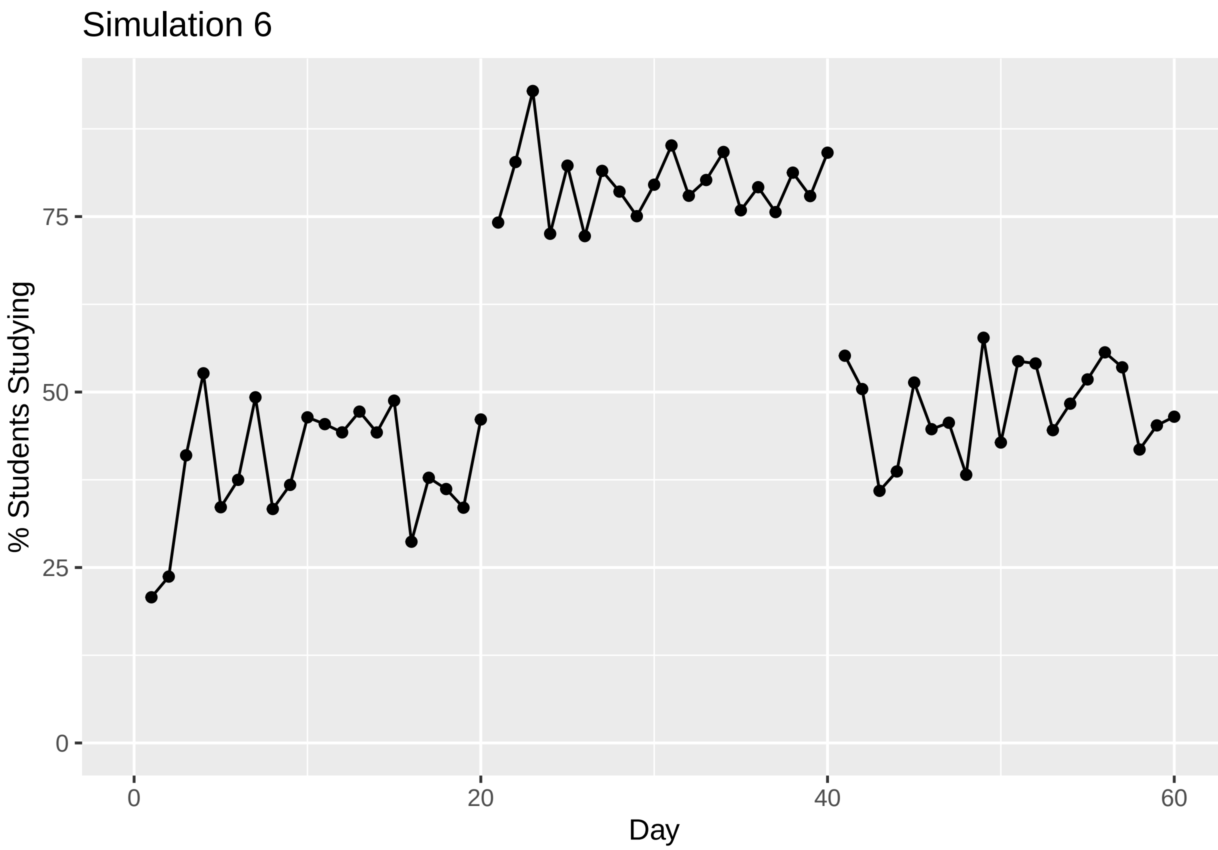
One way to quantify the difference between conditions is the **Percentage of Nonoverlapping Data (PND)**. To calculate PND, we determine the percentage of data in the treatment condition that is more extreme than the most extreme baseline condition or another treatment condition. Let's do an exammple calculation below:

```
# Generate random data
set.seed(3) # Set seed to make data reproducible
```

```r
n = 20 # Number of days
A1 = c()
B2 = c()
A3 = c()
for (i in 1:n){
  A1 = c(A1, rnorm(1, mean = 0.4, sd = 0.2 / (i^(1/2))) * 100)
  B2 = c(B2, rnorm(1, mean = 0.8, sd = 0.2 / (i^(1/2))) * 100)
  A3 = c(A3, rnorm(1, mean = 0.5, sd = 0.2 / (i^(1/2))) * 100)
}
data = c(A1, B2, A3)
time = seq(from = 1, to = length(data))
condition = c(rep("A1", n), rep("B2", n), rep("A3", n))
df = data.frame(data, time, condition)

# Calculate percentage of non-overlapping data
percentage = sum((B2 > max(A1)) & (B2 > max(A3))) / length(B2) * 100
print(paste0("PND = ", percentage, "%"))
```

```
## [1] "PND = 100%"
```

```r
# Make plot
ggplot(data = df, aes(x = time, y = data, group = condition)) +
  geom_line() +
  geom_point() +
  expand_limits(x = 0, y = 0) +
  xlab("Day") +
  ylab("% Students Studying") +
  ggtitle("Simulation 6")
```



Question 10

The data above indicates a PND of 100%. How would you interpret cases where the PND is lower, say, 90%? In your opinion, what would be a sufficiently low PND such that you cannot say there is a significant difference between two conditions?

Answering the Problem Statement

Based on your responses to the previous 10 questions, what kind of single-subject study would you implement and how would you consider doing any analysis?

Credits

The material for this worksheet was based on this source.

(Click on the word "source" to access the link)