Lab 1: Evaluating Assumptions

Legislators

2023-03-04

```
library(tidyverse)
legislators <- read_csv('../datasets/legislators-current.csv', show_col_types = FALSE)
```

There are three assumptions for a Wilcoxon rank-sum test:

1. **Metric scale.** Data have to be measured on the same metric scale.
2. **IID data.** Each pair Xi, Yi is drawn from the same distribution, independently of all other pairs.
3. **Paried data** Each unit of observation is represented by a pair of random variables (X,Y)

Not all of the assumptions for the Wilcoxon rank-sum test are met, as reviewed below:

Assumption 1: Metric scale

As we are measuring the age of senators, this requirement is meet.

Assumption 2: IID data.

The data is independent as each senator are independently chosen by the voters. But they may not be identically distributed as the senator's party in the distribution does have a trend based on their state. Some states have a tendency to choose a particulat party.

Assumption 3: Paried data

The data is not paired as we can see from the count of data points from two parties:

```
# Keep senators only
senators = legislators[legislators$type == "sen",]
# Republican senators age
republican_birthday = subset(senators[senators$party == "Republican",],
                             select = c(birthday))
republican_age = trunc(as.numeric(
  difftime(Sys.Date(), republican_birthday$birthday, units = "weeks" ))/52.25)
# Democratic senators age
democrat_birthday = subset(senators[senators$party == "Democrat",],
                           select = c(birthday))
democrat_age = trunc(as.numeric(
  difftime(Sys.Date(), democrat_birthday$birthday, units = "weeks" ))/52.25)
age_count = c(as.numeric(length(republican_age)),
              as.numeric(length(democrat_age)))
party = c("Republican", "Democrat")
colors = c("blue", "red")
barplot(age_count, names.arg=party, xlab="Party", ylab="Count", col=colors,
        main = "Number of Senators from Democrat and Republican")
```