# Class Activity 13

Arthur Viegas Eguia

April 22 2024

## Group Activity 1

### a. Variance and Skewness

The variance of a random variable $x$ is defined as:

$$\text{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $x_i = (\sum_i^n x_i)/n$ is the sample mean. Also, the skewness of the random variable $x$ is defined as:

$$\text{Skew}(x) = \frac{\frac{1}{n-2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{\text{Var}(x)^{3/2}}$$

Please write functions to calculate the variance and skewness of $\{12, 45, 54, 34, 56, 30, 67, \text{NA}\}$.

```
x <- c(12, 45, 54, 34, 56, 30, 67, NA)
```

```
# function to calculate the variance of a vector
var <- function(x){
  x <- na.omit(x)
  res <- (sum((x - mean(x))^2))/(length(x) - 1)
  return(res)
}
```

```
var(x)
```

```
[1] 346.619
```

```
# function to calculate the skewness of a vector
skewness <- function(x){
  x <- na.omit(x)
  res <- sum((x - mean(x))^3)/((length(x) - 2) * var(x)^(3/2))
  return(res)
}
```

```
skewness(x)
```

```
[1] -0.3930586
```

## Group Activity 2

```
musical_instr_reviews <-  read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/musicrev
  rename(ratingOverall = overall)
glimpse(musical_instr_reviews)
Rows: 10,261
Columns: 3
$ reviewerName  <chr> "cassandra tu \"Yeah, well, that's just like, u...", "Ja~
$ reviewText    <chr> "not much to write about here but it does exactly what i~
$ ratingOverall <dbl> 5, 5, 5, 5, 5, 5, 5, 3, 5, 5, 5, 4, 3, 5, 5, 2, 4, 5, 5,~
```

**a. Write a function to filter the dataset based on the provided rating:**

```
filter_reviews_by_rating <- function(dataset, review){
  dataset %>% filter(ratingOverall == review)
}
filtered_data <- filter_reviews_by_rating(musical_instr_reviews, 4)
```

**b. Write a function to process the text and remove stop words:**

```
process_text <- function(data){
  data %>%
    select(reviewText) %>%
    unnest_tokens(output = word,
                  input = reviewText) %>%
    anti_join(stop_words)
}
process_text(filtered_data) -> process_data_4
```

**c. Write a function to join the processed text with sentiment data and create a word count table.**

```
create_word_count_table <- function(data){
  data %>%
    inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  reshape2::acast(word ~ sentiment, value.var = "n", fill = 0)
}
```

**d. Create the final function that takes the rating and number of words as input arguments and returns a word cloud plot.**

```
library(wordcloud)
library(reshape2)

word_cloud <- function(rating, num.words) {
 rating = as.numeric(rating)
 num.words = as.numeric(num.words)
 if(rating >= 1 & rating <= 5){
   filtered_data <- filter_reviews_by_rating(musical_instr_reviews, rating)
```

```
   processed_data <- process_text(filtered_data)
   word_table <- create_word_count_table(processed_data)
   wordcloud::comparison.cloud(word_table, colors = c("blue", "green"),
                               scale = c(2, 0.5),
                               max.words = num.words,
                               title.size = 2)
 }
 else warning("Please enter a number from 1 to 5")
}
word_cloud(rating = "4", num.words = 50)
```

negative


positive