The Mean: Takeaways

Syntax

- Generating randomly a distribution of integers:

```
set.seed(1)

distribution <- sample.int(value_max, size_of_distribution)
```

- Computing the mean of any numerical vector using R base functions sum() and length():

```
mean  <- sum(vector) / length(vector)
```

- Computing the mean of any numerical vector using R base function mean():

```
mean  <- mean(vector)
```

- Creating your own compute_mean() function from algebraical definition of mean:

```
compute_mean <- function(distribution) {

    N <- length(distribution)

    sum_of_the_distribution = 0

    for ( i in 1:N) {

        sum_of_the_distribution <- sum_of_the_distribution + distribution[i]

    }

    sum_of_the_distribution / N

}
```

- Checking if the mean is equidistant from below and above values:

```
round(sum(distribution - mean)) == 0
```

- Reading a TSV (tab-separated value) dataset:

```
library(readr)

data <- read_tsv(file_name)
```

- Generating a scatter plot to represent visually how the sampling error changes as the sample size increases:

```
library(ggplot2)

ggplot(data = df, aes(x = sample_sizes, y = sampling_errors)) +

    geom_point() +

    geom_hline(yintercept = 0) +

    geom_vline(xintercept = length(sampling_sizes)) +

    labs(x = "Sample size",

         y = "Sampling error")
```

- Generating a histogram to represent visually the distribution of sample means:

```
library(ggplot2)

ggplot(data = tibble(mean_points), aes(x = mean_points)) +

    geom_histogram(bins = 10,

                   position = "identity",

                   alpha = 0.5) +

    geom_vline(aes(xintercept = population_mean)) +

    xlab("Sample mean") +

    ylab("Frequency")
```

Concepts

- We can summarize the distribution of a numerical variable by computing its **mean**.
- The mean is a single value and is the result of taking into account **equally** each value in the distribution.
- The mean is **the balance point** of a distribution — the total distance of the values below the mean is equal to the total distance of the values above the mean.
- The mean of a population can be defined algebraically in several equivalent ways:

- The mean of a sample can be defined algebraically in several equivalent ways:

- The sample mean $\bar{x}$ is an unbiased estimator for the population mean μ.

Resources

- The Wikipedia entry on the mean.
- Useful documentation:
    - sample.int()
    - mean()
    - replicate()