Central Limit Theorem (CLT)

Computer Lab 2

Insert Your Name Here

2024-04-12

The purpose of today's lab is for you to:

- illustrate the Central Limit Theorem (CLT) using simulation
- apply CLT in a practical situation
- work on Mini Project I for the course

Table of contents

Preparation Tasks 1

1 Central Limit Theorem **2**

2 CLT in Practice **3**

Preparation Tasks

Review the concepts of *normal distribution, expected value, and variance for a sum (or average) of independent random variables.*

You should have completed the following tasks **before** attending the lab.

Homework 1

Complete **Dig:3.3.2_5**, **3.96**, and **Dig:3.4.5_1** in Exercise 5.

Homework 2

Attempt to describe in your own words what the Central Limit Theorem states.

Homework 3

Review what you did in Lab 1 and the methods you use for simulating random numbers and fitting a distribution to data.

1 Central Limit Theorem

If one adds (or calculates the mean) of several independent normally distributed random variables, the sum is also normally distributed. But what happens if one adds several variables that are all uniformly distributed? What distribution is obtained if one adds exponentially distributed variables?

The Central Limit Theorem (CLT) states that if a large number of independent variables from any distribution are added, the sum (or mean) becomes approximately normally distributed. In formulas: if n is sufficiently large, $Z_n = X_1 + X_2 + ... + X_n$ is approximately normally distributed regardless of the distribution of $X_1, ..., X_n$. You will investigate whether this seems to be true through some simulations, where we start with some different distributions for the X variables.

Task 1.1

Simulate 1000 random numbers from a uniform distribution, $R(0, 1)$, and store them in the variable `unif1`. Use `hist()` and `qqnorm()` to confirm that the random numbers are uniformly distributed and definitely do not fit a normal distribution.

```
unif1 <- runif(1000,0,1)
# write your R code here
```

Task 1.2

Simulate 1000 new random numbers from a uniform distribution, $R(0, 1)$, and store them in the variable `unif2`. Then, sum the old and new random numbers using `sum12 <- unif1+unif2`. The result is 1000 random numbers from $Z_2 = X_1 + X_2$. Use `hist()` and `qqnorm()` to investigate the distribution of this sum.

```
# write your R code here
```

Task 1.3

Also create variables `unif3`, `unif4`, and `unif5` in a similar way and study the distribution for $X_1 + X_2 + ... + X_5$. Does it seem reasonable that the larger n is, the better the distribution for the sum can be fitted to a normal distribution?

```
# write your R code here
```

Answer:

:::{.callout-note icon=false}

Task 1.4

Try what happens when you sum exponentially distributed random variables with mean 1 (`rexp(1000,1)`). How many variables need to be summed before the sum can be approximated with a normal distribution?

```
# write your R code here
```

Answer:

2 CLT in Practice

In 35 patients with Hodgkin's disease, the number of T4 cells in the blood (number/mm3) was measured. At the same time, the corresponding number was measured in 35 patients with other diseases (Non-Hodgkin). Data are in the file `data/lab2_mini1_files/Hodgkindata.RData`. Load the data with `load()`. You now have two new variables Hodgkin and NonHodgkin.

```
load("data/lab2_mini1_filer/Hodgkindata.RData")
```

Task 2.1

Investigate whether the number of cells in the blood is normally distributed for both groups.

```
# Write your R code here
```

Answer:

Task 2.2

It may be interesting to compare the groups by forming the difference between the two group means. Use the Central Limit Theorem to comment on the distribution of the difference in sample means.

Model:

Let $\mu_{Hodgkin}$ and $\mu_{NonHodgkin}$ be the means in the two groups.

Let $\sigma^2_{Hodgkin}$ and $\sigma^2_{NonHodgkin}$ be the variances in the two groups.

We have two samples $(x_1, ..., x_{35})$ and $(y_1, ..., y_{35})$ from the groups with Hodgkin's and Non-Hodgkin's diseases, respectively.

The difference between sample means is $\delta = \bar{x} - \bar{y}$ and has the following distribution:

$\delta \sim$ ... fill in the description!

> Task 2.3
>
> Estimate the parameters (mean and variance) in the distribution of the difference δ.

```
# Write your R code here
```

Answer:

> Task 2.4
>
> Is it a significant issue that the number of blood cells is not normally distributed in both groups initially? Imagine a situation where it's a problem and suggest a solution.

Answer:

Run (Render) this file to an HTML (open in a browser and print to PDF) or a PDF and upload your PDF as a lab report on Canvas. Then proceed with Mini Project I during lab 2!