Central Limit Theorem (CLT)

Computer Lab 2

Insert Your Name Here

2025-03-16

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

- illustrate the Central Limit Theorem (CLT) using simulation
- apply CLT in a practical situation
- fit models and make conclusions

Table of contents

Preparation Tasks 1

1. Central Limit Theorem **2**

2. CLT in Practice **3**

3. Provtagningstider - problemställningar från labbet **4**
Background 4
Questions 5
Suggested workflow 6
Instruction for reporting 6

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:

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_files/Hodgkindata.RData`. Load the data with `load()`. You now have two new variables Hodgkin and NonHodgkin.

```
load("data/lab2_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:

3. Provtagningstider - problemställningar från labbet

Målsättningen med denna uppgift är att du ska

- practice extracting a problem from reality and constructing a reasonable statistical model using collected data,
- critically review the model and its ability to describe reality,
- apply your knowledge and analyze a biostatistical dataset using R, and
- practice presenting assumptions, models, and conclusions from a statistical analysis in writing.

Background

At our chemistry laboratory, we analyse various samples from the nearby sampling center. There have been strong requests for certain patients who are undergoing tests to be able to meet with a doctor during the same visit and not have to schedule different days for sampling and doctor appointments. In order for this request to be fulfilled, we need to understand the time it takes from when the sample is taken from the patient until the analysis results are ready and investigate whether it is reasonable to have patients wait for test results on

the same day. There are several factors to consider: the sample may wait at the sampling center until it is picked up by our laboratory, there is a certain manual handling time for the sample, and finally, we have the processing time in the machine itself. Additionally, we have slightly different handling times depending on the day of the week and whether it is morning or afternoon.

In the data files `data/lab2_filer/proverfm24.RData` and `data/lab2_filer/proverem24.RData`, there are times (in minutes) it took "from the patient's arm to result of the analysis". The sample is for "general chemistry", and as you can see, we have divided the data into two files, one for samples taken in the morning and one for samples taken in the afternoon. When we create a histogram of this data, we see that it does not appear to follow a normal distribution. This seems reasonable since time has a lower bound. This means that we should use an appropriate distribution to model waiting times in this task.

Questions

> **Tip**
>
> Läs igenom avsnitten Suggested workflow and Instruction for reporting before you answer questions 3.1 to 3.4

Question 3.1.

How likely is it that a morning patient waits more than two hours for the analysis result?

Question 3.2.

We want to be able to say that 95% of afternoon patients will have their test result faster than y minutes. What is the statistical term for y? What value does y have according to the model you have chosen for the afternoon times?

Question 3.3.

There are usually slightly more patients in the afternoon compared to the morning. Let's assume that 60% of patients come in the morning and the rest in the afternoon. How likely is it that a patient sample taken at any time during the day takes more than two hours to analyze?

Question 3.4.

If we have 25 patient samples in the morning, what is the probability that the **average time** for these 25 samples exceeds one hour?

Suggested workflow

- Examine the data (histogram, empirical distribution function). Calculate simple measures (mean, standard deviation). Compare morning and afternoon times.
- Set up an appropriate model (probability distribution) for the morning times and estimate the parameters in the model.
- Do the same for the afternoon times.
- Answer questions (3.1)–(3.3) by utilising the fitted models (probability distributions) for morning and afternoon times.
- For question 3.4, consider which probability distribution is suitable as a model for the average time (mean) of 25 morning times. State which theoretical result (refer to a famous theorem) you use to justify your choice of model! Use the "simple summaries" you calculated earlier (mean and standard deviation of the sample) to estimate parameters in your model for the average time.

Instruction for reporting

Please specify (if appropriate):

- what assumptions you make about data,
- which models you are running,
- which kits you use.
- Present the results of the analysis and what interpretations and conclusions you make.
- Summarize your results for sampling times.
- To pass task 3, avoid referring to the R-code to describe what you have done.