# Lab 04 – Central limit theorem

## ENVX1002 Handbook

Semester 1, 2025

## Facing challenges

Project 1 is due in week 5 and for many of you this may be your first university assignment; some may be nervous, while others may be more relaxed. Your demonstrators will use the start of this practical to discuss how you may be feeling for this first assessment and share ways you might want to approach and prepare for the assessment.

This is the last reflective activity for now, thank you all for contributing so far and we hope you have found some benefit in the activities. Now for some probability!

Learning outcomes

At the end of this computer practical, students should be able to:

- calculate tail, interval and inverse probabilities associated with the Normal distribution
- calculate probabilities associated sampling distribution of the sample mean by using simulation in R and using R commands.

Link to data is below:

- ENVX1002_Data4.xlsx
- Alternatively download from Canvas

## R commands for normal distribution

- **Probability density function:** $f(x) = dnorm(x, \mu, \sigma)$

Here are the main R functions we'll use to work with normal distributions:

- **Probability density function (height of the curve at x):**
  $f(x) =$`dnorm(x, mean, sd)`
- **Cumulative probability (area under curve up to x):**
  $F(x) = P(X \leq x) =$`pnorm(x, mean, sd)`
- **Interval probability (area between two points):**
  $P(a \leq X \leq b) =$`pnorm(upper, mean, sd) - pnorm(lower, mean, sd)`

- **Finding values for given probabilities:**
  $P(X \leq x) =$`qnorm(probability, mean, sd)`
- **Generate random values from a normal distribution:**
  `rnorm(n, mean, sd)`

# Normal distribution

## Exercise 1 - Class activity - How tall is ENVX1002??

1. Make sure you have an existing or new project for Lab 4 and create a quarto document called `Lab_4.qmd` (suggestion). Save it in your project directory.
2. This is an anonymous exercise. Your demonstrator will create a google sheet (or similar) for you to enter your height and sex (M or F only so that the data is easy to analyse).
3. Once all data is collected, manually enter the data into R or export the data as a .csv to be read into R, for example:

```
female <- data.frame(heights = c(176, 180, 187, 168, 160, 170))
male <- data.frame(heights = c(175, 183, 163, 190, 179))
```

4. Calculate the mean and standard deviation using R and graph the distribution for both genders, for example:

```
# calculate the mean and standard deviation of males
m_mean <- mean(male$heights)
m_sd <- sd(male$heights)

library(tidyverse) # needed for ggplot2

# create a histogram of the data
ggplot(male, aes(x = heights)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black")
```

```
# now do the same for females
```

5. Discuss with your neighbour and class why the normal distribution is a good model for height data. Is it a good model for all data?
6. How does your class compare to the Australian statistics. For this we will look at the mean and standard deviation of measured heights for men and women aged 18 - 24 from the ABS for 1995 see page 13 of the pdf

**Note** that both reported and measured heights are provided and (not surprisingly) reported heights are bigger that the measured heights. What do you think the reason for this is?

## Simulating height distribution

Using the ABS stats on female and male heights from the previous exercise:

i) Use R to simulate samples of female **or** male heights for 10, 100, 1000, 10000 simulations and report the mean and standard deviation and draw a histogram for each.

ii) Discuss with your neighbour what happens to the shape of the histogram as your number of simulations increases. An example is given below for 10 simulations of female height. You can make a table in excel to record your observations by hand using the headings:

| Sample size | Mean | SD | Shape of histogram |
|---|---|---|---|
| 10 | | | |
| 100 | | | |
| 1000 | | | |
| 10000 | | | |

```
set.seed(1) # means that we all generate the same set of random numbers so we
can compare
sample10 <- rnorm(10, 163.9, 6.6) # generates a random sample of size 10 from
N(163.9,6.6^2)
sample10 # prints the 10 simulated values (you may not want this with the 10K
simulations!!)
mean(sample10)
sd(sample10)
ggplot(data.frame(heights = sample10), aes(x = heights)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black")
```

iii) Find the probability $P(X \geq 180)$ for each of the simulations and also the exact value. What do you notice in the difference between each of the simulations and the exact probability? An example is given below for the 10 simulations above and the exact probability:

**Actual probability**

```
1 - pnorm(180, 163.9, 6.6)
## OR
pnorm(180, 163.9, 6.6, lower.tail = FALSE)
```

**Simulated probability**

```
length(which(sample10 >= 180)) / 10 ## simulated p for sample size 10
```

# Sampling distributions

## Exercise 2 - Milkfat example

### Part 1

The milkfat content in milk (in %) for 120 cows are presented in the worksheet called ENVX1002_Data4.xlsx. Copy the file into your project directory and:

i) Import the data into R.

```
library(readxl)
milkfat <- read_excel("data/ENVX1002_Data4.xlsx", sheet = "Milkfat")
```

ii) Calculate the summary statistics of Milkfat (mean, median and sd)

**Note** that we use $ColumnName to select a column from the data

```
mean(milkfat$Milkfat)
```

- What type of cows could they be? Compare your data to the table in the following link:

https://lactalis.com.au/info-center/different-breeds-of-cows/

- What state could they be from? Check some of the recent Milk Production reports from Dairy Australia. The data can be found in the Average Milkfat & Protein (%) section of the PDF report: The reports can be found at the following link:

https://www.dairyaustralia.com.au/resource-repository/2020/09/25/milk-production-report

- Could the data be normally distributed?

iii) Create a histogram and boxplot of the milk fat data. Is the data "normally distributed"?

```
require(ggplot2)
ggplot(milkfat, aes(x = Milkfat)) +
  geom_histogram(binwidth = 0.1, fill = "lightblue", color = "black") +
  xlab("Milkfat (%)")
```

iv) In the UK, `breakfast milk' (or`Channel Island milk') has 5.5% fat content. What percentage of the cows in this data set is yielding breakfast milk with $\geq$ 5.5%?

```
s <- sort(milkfat$Milkfat) # Sorts the data
s # Look at the sorted data
length(s[s >= 5.5]) # Counts how many are >= 5.5
```

v) In Australia, full cream milk has greater than 3.2% milk fat content. What percentage of these cows is yielding full cream milk?

```
## Your turn
```

**Part 2**

Let $X$ represent the milk fat content for the population of this breed of cows.

i) Assuming the population is normal, use the sample mean and standard deviation from the previous question as estimates of the population parameters. So $X \sim (\mu = ..., \sigma^2 = ...)$.

ii) Draw a picture of the curve representing $X$. The below example uses ggplot2 to draw the curve for $N(4.16, 0.30^2)$.

```
library(ggplot2)
ggplot(data.frame(x = c(4.16 - 4 * 0.3, 4.16 + 4 * 0.3)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = 4.16, sd = 0.30)) +
  xlab("x") +
  ylab(expression(N(4.16, 0.30^2) ~ pdf))
```

iii) What is the probability that 1 cow has a fat content less than 4%? We will adapt the ggplot command above a picture of this probability and then use R to find the probability.

Hint: You may need to use the `stat_function` command to draw the curve and then use the `pnorm` command to find the probability.

```
ggplot(data.frame(x = c(4.16 - 4 * 0.3, 4.16 + 4 * 0.3)), aes(x = x)) +
  stat_function(
    fun = dnorm, args = list(mean = 4.16, sd = 0.30),
    geom = "area", fill = "white"
  ) +
  stat_function(
    fun = dnorm, args = list(mean = 4.16, sd = 0.30),
    xlim = c(4.16 - 4 * 0.3, 4), geom = "area", fill = "red"
  ) +
  xlab("x") +
  ylab(expression(N(4.16, 0.30^2) ~ pdf))
```

```
pnorm(4, 4.16, 0.30)
```

iv) What is the probability that 1 cow (randomly sampled) has a fat content greater than 4.5%? Try and adapt the ggplots above to draw a picture of this probability and then use R to find the probability.

v) For a sample of 10 cows (randomly sampled), what is the probability that the sample mean milk fat content is greater than 4.2%?

Hint: First find the distribution of the sample mean $\overline{X}$. Then find $P\left(\overline{X} > 4.2\right)$

v. What is the probability that the sample mean milk fat content is greater than 4.2%?

```r
# For a sample of 10 cows, we need to find P(X_bar > 4.2)
# Mean of X_bar = population mean = 4.16
# Standard error of X_bar = population sd / sqrt(n) = 0.30 / sqrt(10)

# Calculate the standard error
se <- 0.30 / sqrt(10)

# Calculate the probability
pnorm(4.2, 4.16, se, lower.tail = FALSE)

# Visualization
ggplot(data.frame(x = c(4.16 - 4 * se, 4.16 + 4 * se)), aes(x = x)) +
  stat_function(
    fun = dnorm, args = list(mean = 4.16, sd = se),
    geom = "area", fill = "white"
  ) +
  stat_function(
    fun = dnorm, args = list(mean = 4.16, sd = se),
    xlim = c(4.2, 4.16 + 4 * se), geom = "area", fill = "blue"
  ) +
  xlab("Sample mean milk fat content") +
  ylab("Probability density") +
  ggtitle("Distribution of sample mean (n=10)")
```

# Standard Error of the mean

## Exercise 3 - Skin cancer

A dermatologist investigating a certain type of skin cancer induced the cancer in nine rats and then treated them with a new experimental drug. For each rat she recorded the number of hours until remission of the cancer. The rats had a mean remission time of 400 hours and a standard deviation of 30 hours. From this data, calculate the standard error of the mean.

## Exercise 4 - Soil carbon

An initial soil carbon survey of a farm based on 12 observations found that the sample mean $\overline{X}$ was 1.2% and the standard deviation s was 0.4%. How many observations would be needed to estimate the mean carbon value with a standard error of 0.1%?

## Exercise 5 - What's in the media - looming state election

An article was published in the Sydney Morning Herald on Saturday 20.3.2010 about statistics related to opinion polls. Read it and find the sentences related to (i) populations versus samples (ii) standard error formula (iii) the effect of sample size on standard errors.

http://www.smh.com.au/national/demystifying-the-dark-art-of-polling-20100319-qmai.html

## Exercise 6 - Extra practice

The average Australian woman has height (in cms) of 161.8 with a standard deviation of 6.

i) The Australian Institute of Sport ran a netball training camp for the best Australian young players. How tall were the goal position players? http://www.abc.net.au/news/2015-06-14/tall-athletes-get-support-at-ais-to-stand-as-proud-netballers/6544642

ii) What is the probability of finding an Australian woman of this height or taller?

**Hints**:

**Step 1**: Using ggplot, draw a sketch of the Normal curve with the probability identified. You may need to draw a section of the right tail as the probability is small! We have provided the solution for the plotting to assist you.

**Step 2**: Calculate the probability in R.

```
1 - pnorm(189, 161.8, 6)

ggplot() +
  stat_function(
    fun = dnorm, args = list(mean = 161.8, sd = 6),
    geom = "area", fill = "white", xlim = c(180, 161.8 + 4 * 6)
  ) +
  stat_function(
    fun = dnorm, args = list(mean = 161.8, sd = 6),
    geom = "area", fill = "red", xlim = c(161.8 + 4 * 6, 189)
  ) +
  xlab("x") +
  ylab(expression(N(161.8, 6^2) ~ pdf)) +
  scale_x_continuous(breaks = 189)
```

iii) Dharshani Sivalingam is the tallest netball player in the world. How tall is Dharshani? https://en.wikipedia.org/wiki/Tharjini_Sivalingam What is the probability of finding an Australian woman of Dharshani's height?

iv) Madison Brown is one of the the shortest Australian International players. How tall is Madison? https://en.wikipedia.org/wiki/Madison_Browne What percentage of Australian women are between Madison and Dharshani's heights?

v) If 80% of Australian women are above a certain height, what is that height?