# NDH802 - Assignment 2

Group no.

---

- The assignment includes 2 questions, 6 points each.
- Question 1 is empirical (based on the data provided). Question 2 is purely theoretical (based on the assumption in the question).
- Submit your assignment via Canvas before 10:00 CET, April 27, 2021.
- Your submission should be an RMarkdown file with your solutions in words and/or R code. If you handwrite parts of your assignment, insert it as an image near the corresponding question(s). Name the file NDH802_Assignment2_GroupNumber.
- You should work in groups andcontribute equally.
- You can copy my code, but make sure you understand it.
- You should not have the exact solutions and/or results with other groups.
- Results without code/justifications will not be graded.

---

**Set things up**

Set your working directory

```
#setwd("")
```

Run this code chunk to load data into your R Environment. The command will randomly select 1,000,000 rows of data from the original data set. Fill in your group number within `set.seed()`. For example, if you are group 3, make it `set.seed(3)`. Hereby each and every group should have a unique `df`. Accordingly, your results should be different from other groups' and you should not be comparing them.
*Note.* If you make the wrong seed, your assignment will **not** be graded.

```
inference_dataset <- read.csv("https://cda.hhs.se/inference_dataset.csv")
set.seed(1); df <- inference_dataset[sample(1:nrow(inference_dataset),
                           size = 1000000,
                           replace = FALSE), -1]

rm(inference_dataset)
```

Please refer to Canvas, Hand-in 2 for more details about the data set.

**Question 1. Discrete distribution**

(a) What are the probabilities of loyal and not loyal customers in your data set?
Formally, compute $P(\text{loyal} = 1)$ and $P(\text{loyal} = 0)$. (1p)

(b) Imagine you go to a store and meet five independent customers. What is the probability that one of them is loyal customer (i.e., the other four are not loyal customers). *Hint*: You can try `dbinom()`, the argument `prob` is your result from (a). (1p)

(c) What is the probability of being loyal customers, given the customers who only shop offline?
Formally, compute $P(\text{loyal} = 1 \mid \text{channeltype} = \text{offlineonly})$. (1p)

(d) What is the probability of shopping only offline, given the loyal customers?
Formally, compute $P(\text{channeltype} = \text{offlineonly} \mid \text{loyal} = 1)$. (1p)

(e) What is the probability that a randomly chosen customer will be a loyal customer who only shop offline?
Formally, compute $P(\text{channeltype} = \text{offlineonly} \cap \text{loyal} = 1)$. (1p)

(f) Compare the results of (c), (d) and (e). Are they similar/different? Should they? Why/why not? You can use mathematical formula, words, venn diagrams or the combination of them, whichever expresses your rationales the best. If you find handwriting is more convenient, feel free to do so and attach a (readable) photo in the submission.(1p)

**Question 2. Continuous distribution**

Let $X$ (measured in cm) denote the height of all the Swedes in 2021. Assume $X \sim N(\mu = 175, \sigma^2 = 20)$. Be careful, the argument `sd` in `pnorm()` is standard deviation. Standard deviation is *different* from variance.

(a) What is the probability that a random Swede is shorter than 172cm? (0.5p)

What is the probability that a random Swede is taller than 178cm? (0.5p)

(b) Compare the results you got from (a). Are they similar/different? Why? (1p)

(c) What is the probability that a random Swede is from 172cm to 178cm tall? (1p)

(d) What is the probability that a random Swede is exactly 175 cm tall? Justify your answer. *Hint*: This is a trick question. (1p)

(e) What is the cut point of the top 5% tallest Swedes (i.e., find the height of the shortest Swede among the top 5% tallest)? (1p)

(f) Find the shortest range such that the probability is 90% that the height of Swedes will fall in this range. (1p)

*Have fun and good luck!*
*Huong and Emelie*