## Question 1

Write a piece of program that takes as input 22 lists called list1 and list2. list1 contains 55 first names (strings) and list2 has 33 first names. list1 and list2 may or may not have common names. Then, it returns a third list called set\_difference which contains those names in list1 that are not in list2.

**Question 2**

Write a Python **function** that takes two **positive integers** and returns their **greatest common divisor**. In case you pass a negative integer to the function it must return the following string: "This function takes only positive integers!"

## Question 3

Write a function that prints all the prime numbers in the interval [0,p][0,p], where pp is a parameter to be passed to the function.

## Question 4

1. Set seed by the initial value 12311231 and define the following variables (Python objects) with the shown assigned values:
   * size ⟵1000⟵1000,
   * n ⟵700⟵700,
   * p ⟵0.3⟵0.3
2. Randomly generate size number of integers in (0,200)(0,200) and save them as col1.
3. Randomly generate size number of values according to Unif[0,1]Unif[0,1] and call it col2.
4. Randomly generate 2\*size numbers from Binom(Binom( n,p )) and randomly select size number of them and save as col3.
5. Define the following functions
   * 𝚏𝚞𝚗𝚌𝚝𝟷(x)=lnxfunct1(x)=ln⁡x
   * 𝚏𝚞𝚗𝚌𝚝𝟸(x)=10exp(x)1+exp(x)funct2(x)=10exp⁡(x)1+exp⁡(x)
   * 𝚏𝚞𝚗𝚌𝚝𝟹(x)=3501002π√.exp(−x220000)funct3(x)=3501002π.exp⁡(−x220000)
6. Define
   * col4=𝚏𝚞𝚗𝚌𝚝𝟷(=funct1( col1 )),
   * col5=𝚏𝚞𝚗𝚌𝚝𝟸(=funct2( col2 )), and
   * col6=𝚏𝚞𝚗𝚌𝚝𝟹(=funct3( col3 ));
7. Randomly generate size number of genders from the set {Female, Male}{Female, Male} and save them as col7.
8. Construct a **data frame** with 77 columns col1 to col7 and call it mydata.
9. Describe the dataset using the descriptive statistics discussed in the class.
10. Use the appropriate visualisation tool to visualize each variable in the dataset.
11. Using **loops** scatterplot every pair of columns versus each other if appropriate.

## Question 5

1. Assume that X∼N(μ,σ2)X∼N(μ,σ2) with μ=55,σ=15μ=55,σ=15. Randomly generate a set of 1000,000 values for XX according to the given distribution and call it set DD.
2. Pretend that DD is your whole population. Choose a sample of 10001000 values from DD.
3. Plot an approximate density distribution function using the selected sample.
4. Using a loop repeat the second step 3030 times. For each sample estimate the population mean and save each estimate. Calculate the mean squared error of your estimated means.
5. Plot the histogram of all 3030 saved sample means (**Only sample means**). According to the histogram, what is the sampling distribution of the mean?
6. If instead of 500500 times, we resample over and over for a large number of times, how does the sample mean change?

## Question 6

Suppose that we would like to model the event of flipping a coin for 1515 times. If the probability of getting heads equals 0.60.6, then answer the following questions.   
**Do not forget to include your codes.**

1. Which distribution is it and what are the parameters of the distribution? Is it a discrete or continuous distribution?
2. What is the probability of obtaining 1010 heads? Explain how to calculate it.
3. What is the probability of getting more than or equal to 1010 heads? What about less than or equal to 1010 head? What should be the summation of these two probabilities and why?
4. Find the expected value of obtained number of heads in each trial? (**Each trial consists of**1515**tosses**)
5. How probable is it to get (H,H,H,T,T,H,T,T,H,T,H,H,H,T,H)?
6. Find the first, second, and the third quantiles.
7. Repeat the trial 1010 times and estimate the mean each time. Using pandas.crosstab build the frequency table of the results. Plot the histogram of these ten estimates.
8. Now, gradually increase the number of trials from 100100 to 10001000. (Start from 100100 and add 5050 each time to reach 10001000.) Plot the histogram for the sample mean each time. How does the sampling distribution of mean is changing?

## Question 7

Assume that for a study we want to sample people from the Montreal population. The target of the study is a particular disease. If the probability of sampling a person with this disease equals 0.0070.007.

1. How many people we need to sample totally in order to get exactly 7373 patients with the disease?
2. How many people we need to sample totally in order to have at least 7373 patients with the target disease?
3. What distribution is it and what are its parameters?
4. Calculate the expected value, variance, as well as the first, second and the third quantiles.

## Question 8

1. Generate 10001000 data points according to the exponential distribution with parameter λ=1.2λ=1.2.
2. Estimate the mean of the sample.
3. Repeat the first step but each time increase the sample size up to 1000,000. Calculate the sample mean each time. Scatterplot the mean versus the sample size for each repetition. Do you see any trend in the plot sample means? Can you guess the limit of the sample mean due to the plot?

## Question 9

**Note:** For this question you may need to **Google** in order to find commands necessary for some parts of the question.

1. Import sklearn library and from it import datasets. Aslso, set seed as in **Question 4**.
2. From sklearn.datasets load the dataset called "Boston". Read the documentation of sklearn.datasets in order to understand the structure of datasets built in the library.
3. From the dataset boston extract the part called data. (boston is in the form of **dictionary** and includes different parts. You need to extract only the part called data)
4. Find the mean and standard deviation of each variable (=feature, column) in the data.
5. Assume the data you have are the whole population. Randomly sample 300300 entries from the 1111-th variable. Using this sample estimate the population mean (whose true value is already calculated).
6. Provide a confidence interval with 9595% of confidence level. (To find the corresponding z-value you can use stats.norm.ppf())
7. If you repeat Steps 5 and 6 above 2020 times, how many of these 2020 confidence intervals do you expect to include the true mean? Why?
8. Scatterplot the estimated means with their 95% margins of error, as well as the true value of the mean. How many of the error margins include the true mean? Does it match with your answer to the previous step? If not, what is the reason in your opinion?