Master of Public Health, specialization: AI4PH
DESU: AI4PH

TU PHS-PRIM: Principles and methods of public health sciences

General Guidelines

- The quality of the comments and the writing in general will be considered in the scoring.
- The "Practice" section (Part 3) must be done on R.
    - In your answers, the number of each question must be indicated.
    - Make sure that the script you send me is executable, i.e., that there are no errors and that I can run it directly.
      If you want to write a notebook, it is possible to do so.
    - If some of your functions don't run and you can't find the errors, specify it in a comment.
      Don't hesitate to comment!
- I am waiting for you: A Zip folder with the name "Firstname_LastName ", in which there must be two files:
    - The first one will have to contain the answers to parts 1 and 2 "Firstname_LastName_Parties_1_2.pdf"
    - The second file the part 3, which will be a script. "Firstname_LastName_Parties_3.R" or if you decided to use a notebook: "Firstname_LastName_Parties_3.Rmd"

Attention on AMETICE, you will be able to deposit only one file, thus the Zip folder.

Good luck!

I) MQC: Several choices are possible

Question 1: A survey is conducted in a population of 7,500 subjects of whom 653 have been diagnosed with breast cancer. The proportion 653/7,500 represents:

a. The incidence
b. Prevalence
c. Lethality
d. None of the above

Question 2: The standard deviation of a series of values:

a. Is a central measurement parameter
b. Is expressed in the same units as the values in the series
c. Has a different value if measured on a sample or in a population
d. Is calculated from the variance
e. Is small when values are scattered

Question 3: The variance of a series of values:

a. Is a parameter used to measure the dispersion of values
b. Is expressed in the same units as the values in the series
c. Is independent of the standard deviation
d. Is obtained by calculating the average of the squares of the deviations from the mean
e. Is high when the values of the series are widely dispersed

Question 4: In a case-control survey, the confidence interval (CI) of the odds ratio (OR) is [0.7-0.9].

Which of the following statements are correct? This result means that the factor studied:

a. Does not play a role in the occurrence of the disease
b. Is a risk factor with a weak effect
c. Is a protective factor
d. May be a protective factor but is not significant
e. Invalidates the study as it lacks power

Question 5:

In a study comparing the effectiveness of two types of dressings for skin wounds, the authors concluded that the performance (healing speed) of dressings A was superior to that of dressings B with a risk of error of less than 2%.

Question: Which of the following statements are correct?

This figure of 2% corresponds:

f. An alpha risk
g. A beta risk
h. A significance level p

II) Exercises of application

Exercise 1:

During 2010, 2,346 cases of angina were identified in children under 10 years of age. The population of children under 10 years of age was 16,745 on January 1, 2010 and 21,345 on January 1, 2011.

What is the incidence of measles in 2010 in children under 10 years of age?

Exercise 2: Confidence interval of percentage

To know the frequency of scabies in a region of 250 000 inhabitants, a survey was carried out on a representative sample of 4 327 persons. Among them, 913 people were found to have scabies. Calculate the estimated frequency of scabies in this region and its 95% confidence interval.

Exercise 3:

The following table shows the distribution of a population in 2021 by age and gender.

Category age (years)	Women	Men	Total
	n	n	N
0-14	5347	5123	10470
15-29	3236	4276	7512
30-44	6239	5349	11588
45-59	3459	5302	8761
60-74	4302	3999	8301
>74	3290	3333	6623
Total	25873	27382	53255

1) What is the frequency of women?

2) What is the frequency of subjects over 74 years old?

3) What is the frequency of men among 30-44 year olds?

4) What is the frequency of 15-29 year olds among women?

5) What is the ratio of females/males among subjects over 45 years old?

III) Practice

VICAN (VIe après le CANcer), is a five-year national survey of cancer survivors in France. Data were collected from telephone interviews with patients 5 years after diagnosis. The objective of this survey is to provide information on the living conditions of cancer survivors.

You will find here a sample of this survey. The main objective will be to identify factors associated with the physical and mental quality of life of cancer survivors.

- The patient's ID : Variable name = ID
- Health insurance: 1 = CNAMTS; 2 = MSA; 3= RSI
  Variable name = FC_Caisse
- Location of the pathology: 1 = Breast; 2 = Lung; 3 = Colon and Rectum; 4 = Prostate; 5 = VADS; 6 = Bladder; 7 = Kidney; 9 = Thyroid; 10 = Lymphoma; 11 = Melanoma; 81 = Cervix; 82 = Body of Uterus
  Variable name = MS_CodCancer
- Age : Variable name = FC_AgeDiag_R0
- Marital status : 1 = Married/Partnered/Concubine; 2 = Single/Divorced/Separated/Widowed
  Variable name: Q5_sd4_r1
- Do you have children? : 1 = Yes; 2 = No;
  Variable name: Q5_sd5
- Level of study : 1 = No degree; 2 = Less than Bachelor's degree; 3 = High school diploma or more
  Variable name: Q5_sd10_r2
- Physical quality of life score: Variable name = Q5_pcs12_r1
- Mental quality of life score: Variable name: Q5_mcs12_r1
- Pain : 0 = No; 1 = Yes;
  Variable name: Q5_Pain
- Anxiety : 0 = No anxiety; 1 = Questionable anxiety state; 2 = Certain anxiety state
  Variable name: Q5_anxiete
- Depression : 0 = No depression; 1 = Questionable depression state; 2 = Certain depression state
  Variable name: Q5_depression
- Total net remuneration at time of survey : 1 = <1500€; 2 = >=1500€; 3 = Not employed
  Variable name : Q5_jobv5.36_R1
- Socio-professional category at the time of the survey (3 classes) : 1 = Executives (farmers, craftsmen, workers, employees); 2 = Managerial occupations (executives and senior managers, company directors, intermediate occupations); 3 = Not employed at the time of the survey
  Variable Name : MS_CSP_ENQ_3C_R1
- Sequels (More generally, do you have any after-effects following the management of your disease?) : 1 = YES and they are very important; 2 = YES and they are important; 3 = YES but moderate; 4 = YES but very moderate; 5 = NO, I have no after-effects;
  Variable Name : Q5_med23.1

1. Import the file "BDD_VICAN.csv".

2. Display the first lines of the dataset. Display the lines 1; 4; 18; 103 of the dataset.

3. How many variables and observations are there?

4. Does this file contain any missing values?

5. What is the nature of the variables studied?

6. Some of the variables are in the wrong format, for example, a qualitative variable in "numeric" format. Based on the description of each variable (found at the beginning of this exercise), recode the variable(s) into the correct format.

7. Definition of clinically significant fatigue score: score ≥ 40 on the fatigue scale included in the survey, the threshold at which a fatigue condition was shown to be clinically significant.
Create a categorical variable based on this definition.

8. Group the modalities of the variable sequelae into 3 modalities.
This new variable, named "Q5_med23.1_rec" will be considered in the following analyses instead of "Q5_med23.1".

9. Display the frequency table for this new variable.

10. Concerning age: What is the average age of our study population, then that of breast cancer. Determine the 95% confidence intervals (CI) for each of the calculated means.

Calculate the variance, standard deviation of the sample, then that of breast cancer.

11. Draw a graph that will represent the distribution of age by location of the pathology. Choose the most appropriate graph. Export the graph in a pdf format.

12. Determine the factors associated with physical and then mental quality of life, including variables with a p-value < 0.2. Which model will you use? How will you proceed?
Interpret the final result.

13. Export the new database in ".csv" format. It should be attached to the file that will be sent to me.