Introduction to Data Science Assignment 1

Due time: 11:59 pm, September 28, 2023
6 questions and 100 points in total

This assignment covers the sections of “introduction”, “knowing your data”, “sampling” and “data quality, cleaning, and integration”.

Every student is expected to complete the assignment independently. Discussion between students is encouraged, but no plagiarism will be tolerated.

Please submit your answers in a PDF file. If you want to also submit your code separately, you can pack all files, including a readme file, into a compressed file.

Question 1 (20 points)

Let us consider the Pima Indians Diabetes Database, which can be downloaded from Canvas as a CSV file. We will practice data quality assessment and data cleaning. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage. The dataset consists of several medical predictor variables (also known as attributes) and one **target variable**, *Outcome*. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

On the attributes *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, and *BMI*, value 0 is used as an indicator where the value is missing. We call those attributes the **attributes with missing values**.

When we start to work on a dataset that contains missing values, an important first step is to understand how the missing values are distributed. This can be done by collecting some statistics about the missing values. You will conduct the exercise in this question by answering the following questions.

1. (5 points) We may investigate how missing values are distributed in records. Calculate in each record, how many values are missing in those attributes with missing values. Summarize your findings by plotting a figure that reports your results, where the x-axis is the number of missing values in a record, and the y-axis is the number of records having that many missing values.
2. (5 points) As the dataset contains data in two classes indicated by the target variable *Outcome*, it is interesting to observe how the missing values appear in those two classes. To report the results, please draw a figure using the same configuration in Question 1.1 but report the numbers of records having the number of missing values in class 0 and class 1 separately.

3. (5 points) Compute the conditional probability $P(X = 0|Y = 0)$ where X and Y are the attributes with missing values. You can report your results using a table where each row and each column represents an attribute, respectively, and the entry [X, Y] in the table reports the conditional probability $P(X = 0|Y = 0)$.
4. (5 points) For each pair (X, Y) of attributes with missing values, test whether they are independent from each other. The null hypothesis here is that the two attributes have missing values independently. Please compute the Chi-square statistics and set the p-value to 0.05. Report whether you can reject the null hypothesis for each pair of attributes with missing values. Hint: here we are only concerned about whether a value is a missing or not. Thus, you should compute a 2 x 2 contingency table and then compute the Chi-square statistics.

Question 2 (10 points)

We want to draw a uniform sample of size n from a population of size N. Consider the following procedure. We consider every unit of the population and take the unit into the sample by probability $\frac{n}{N}$ and discard the unit by probability $\left(1 - \frac{n}{N}\right)$.

1. (5 points) Does this procedure guarantee that we obtain a uniform sample of size n from the population? Justify your answer analytically.
2. (5 points) Write a simple program to verify the procedure. Set $n = 100$ and $N = 10000$. Run your program 100 times to observe how well the procedure produces a sample satisfying the size requirement. Report the distribution of the numbers of units in the samples, that is, draw a figure where the x-axis is the number of units in a sample and the y-axis is the number of samples that each has x units.

Question 3 (10 points)

1. (5 points) Mathematically prove that, in simple random sampling, sample mean is an unbiased estimator of the population mean.
2. (5 points) A policy maker is interested in the proportion of homeless people living in a large metropolitan area who are mentally ill. One suggested sampling design is to take a sample of homeless people who receive medical care at several randomly selected clinics in the area that treat homeless individuals. Let us assume that patients visiting these clinics are truthful about their mental health condition, and that the clinics' detection of mental illness is accurate. However, this design may be subject to some potential biases. Identify two possible biases in this design.

Question 4 (15 points)

Consider three groups of people. Group A has 84 people, group B has 9 people and group C has 7 people. We want to draw a uniform sample of size 10 to represent the 100 people and, at the same time, each group can be also fairly represented. An important requirement is that each group should have at least one representative in the sample.

One simple and intuitive idea is to allocate to each group a sample size that is 10% of its population. We need to either truncate or round up the sample size allocated to a group if it is not a whole number.

1. (5 points) What problem would we have if we truncate the sample size allocated? Use an example to illustrate.
2. (5 points) What problem would we have if we round up the sample size allocated? Use an example to illustrate.
3. (5 points) The Huntington-Hill method allocates sample quota to groups as follows. It starts with assigning each group one sample. Then, for each group, it calculates the quotient $A = \frac{n}{\sqrt{m(m+1)}}$, where n is the population of the group and m is the number of samples currently allocated to the group. The group with the highest quotient will get one more sample. The iteration continues until all samples are allocated to groups. Write a program to compute the number of samples assigned to each group using the Huntington-Hill method. Report the procedure how the sample quotas are assigned to the groups. By the way, the Huntington-Hill method is used by the United States House of Representatives to allocate seats to states.

Question 5 (25 points)

In this question, we will practice sampling using the Pima Indians Diabetes Database. Please treat the value 0s as valid here, that is, in this question we do not treat them as missing data. The statistic we want to estimate is the mean of the attribute Outcome. You should write programs to conduct the following tasks.

1. (5 points) Use simple random sampling on the data set using sample rates 5% and 10%, respectively, to estimate the mean of Outcome. Report your estimates and the corresponding 95% confidence intervals for the population mean.
2. (10 points) Conduct stratified simple sampling on the data set. Use the attributes, Pregnancies, Glucose (discretized into groups according to the highest digit, i.e., [0, 9], [10, 19], …, [190, 199]), BloodPressure (discretized into groups according to the highest digit), and age (discretized into groups according to the highest digit), respectively, to partition the data into strata. Set the sample rate to 10%. Which attribute gives you the best estimate? Discuss why. Hint: you need to design how to handle the small strata that have less than 10 units. You can either truncate or round the sample size in each stratum.
3. (10 points) Conduct cluster simple sampling on the data set. Use the attributes, Pregnancies, Glucose (discretized into groups according to the highest digit, i.e., [0, 9], [10, 19], …, [190, 199]), BloodPressure (discretized into groups according to the highest digit), and age (discretized into groups according to the highest digit), respectively, to divide the data into primary units. Randomly choose 3 primary units in your sampling. Which attribute gives you the best estimate? Discuss why.

Question 6 (20 points)

In this question, we practice drawing samples in different distributions and examine the relations among the distributions.

First, let us draw a sample S1 of size 10000 from a random normal distribution with a mean of 100 and standard deviation of 20. This can be implemented as follows in R.

```
hist(
  rnorm( 10000, 100, 20 )
)
```

In Python you can easily do this using numpy as follows.

```
import numpy as np


# Set the mean and standard deviation for the normal
distribution

mean = 100

std_dev = 20


# Draw 10000 samples from the normal distribution

samples = np.random.normal(mean, std_dev, 10000)
```

Second, draw another sample S2 of size 10000 from a lognormal distribution as specified in the following R statement.

```
hist(
  exp(1)^rnorm(10000,3,.9) + 100,
  breaks = 39
)
```

Third, draw the third sample S3 of size 10000 from an exponential distribution as specified in the following R statement.

```
hist(
  rexp(10000, 1/10) +100
)
```

Last, draw the fourth sample S4 of size 10000 from a power law distribution as specified in the following R statement.

```
hist(
  rnorm(10000, 3, 1)^exp(1) + 100
)
```

Write programs to complete the following tasks.

1. (5 points) Plot the above four distributions, where the x-axis is the value and the y-axis is the number of sampled units in a sample.
2. (5 points) Calculate and plot the median and the Pythagorean means of those four distributions.
3. (10 points) Using the lognormal distribution, demonstrate that the geometric mean is the antilog of the arithmetic mean of the log transformed values of the dataset. Recall that antilog(x) = 10^x.