Data science for Doctors: Inferential Statistics Exercises (part-2)

Data science enhances people's decision making. Doctors and researchers are making critical decisions every day. Therefore, it is absolutely necessary for those people to have some basic knowledge of data science. This series aims to help people that are around medical field to enhance their data science skills.

We will work with a health related database the famous "Pima Indians Diabetes Database". It was generously donated by Vincent Sigillito from Johns Hopkins University.
Please find further information regarding the dataset there.

This is the fifth part of the series and it aims to cover partially the subject of Inferential statistics.
Researchers rarely have the capability of testing many patients,or experimenting a new treatment to many patients, therefore making inferences out of a sample is a necessary skill to have. This is where inferential statistics comes into play.
In more detail, in this part we will go through the hypothesis testing for binomial distribution (Binomial test)
and normal distribution (Z-test). If you are not aware
of what are the mentioned distributions please go here to acquire
the necessary background.

Before proceeding, it might be helpful to look over the help pages for the binom.test, mean,sd ,sqrt, z.test.
Moreover it is crucial to be familiar with the Central Limit Theorem.

```
install.packages("TeachingDemos")
library(TeachingDemos)
```

Please run the code below in order to load the data set and transform it into a proper data frame format:

```
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data"
data <- read.table(url, fileEncoding="UTF-8", sep=",")
names <- c('preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class')
colnames(data) <- names
data <- data[-which(data$mass ==0),]
```

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Suppose that we take a sample of 30 candidates that tried a medicine and 5 of them are positive.
The null hypothesis is H_{0}: p = average of classes, is to be tested against H1: p != average of classes.
This practically means whether the drug had an effect on the patients

Exercise 2

Apply the same test as above but instead of writing the number of samples try to apply the test in respect to the number of successes and failures (5,25).

Exercise 3

Having the same null hypothesis as the exercises 1,2 apply a one-sided test where H1: p < average of classes.

Exercise 4

At the previous exercises we didn’t specified the confidence interval, so it applied it with the default 0.95. Run the test from exercise 3 but instead of having confidence interval of 0.95 run it with confidence interval 0.99.

Exercise 5

We have created another drug and we tested it on other 30 candidates. After having taken the medicine for a few weeks only 2 out of 30 were positive. We got really excited and decided to set the confidence interval to 0.999. Does that drug have an actual impact?

Exercise 6

Suppose that we establish a new diet and the average of the sample,of size 30, of candidates who tried this diet had average mass 29 after the testing period. Find the confidence interval for significance level of 0.05. Keep in mind that we run the test and compare it in respect to the data$mass variable

Exercise 7

Find the Z-score of the sample.

Exercise 8

Find the p-value for the experiment.

Exercise 9

Run the z-test using the z.test function with confidence level of 0.95 and let the alternative hypothesis be that the diet had an effect. (two-sided test)

Exercise 10

Let’s get a bit more intuitive now, let the alternative hypothesis be that the diet would lead to lower average body mass with confidence level of 0.99. (one-sided test)