Data Science for Doctors – Part 4 : Inferential Statistics (1/5)

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 here.

This is the fourth 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.

Before proceeding, it might be helpful to look over the help pages for the sample, mean, sd , sort, pnorm. Moreover it is crucial to be familiar with the Central Limit Theorem.

You also may need to load the ggplot2 library.

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

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

Generate (10000 iterations) a sampling distribution of sample size 50, for the variable mass.

You are encouraged to experiment with different sample sizes and iterations in order to see the impact that they have to the distribution. (standard deviation, skewness, and kurtosis) Moreover you can plot the distributions to have a better perception of what you are working on.

Exercise 2

Find the mean and standard error (standard deviation) of the sampling distribution.

You are encouraged to use the values from the original distribution (data$mass) in order to comprehend how you derive the mean and standard deviation as well as the importance that the sample size has to the distribution.

Exercise 3

Find the of the skewness and kurtosis of the distribution you generated before.

Exercise 4

Suppose that we made an experiment and we took a sample of size 50 from the population and they followed an organic food diet. Their average mass was 30.5. What is the Z score for a mean of 30.5?

Exercise 5

What is the probability of drawing a sample of 50 with mean less than 30.5? Use the the z-table if you feel you need to.

Exercise 6

Suppose that you did the experiment again but to a larger sample size of 150 and you found the average mass to be 31. Compute the z score for this mean.

Exercise 7

What is the probability of drawing a sample of 150 with mean less than 31?

Exercise 8

If everybody would adopt the diet of the experiment. Find the margin of error for the 95% of sample means.

Exercise 9

What would be our interval estimate that 95% likely contains what this population mean would be if everyone in our population would start adopting the organic diet.

Exercise 10

Find the interval estimate for 98% and 99% likelihood.