Data science for Doctors: Variable importance Exercises

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 tenth part of the series and it aims to cover the very basics of the subject of principal correlation coefficient and components analysis, those two methods illustrate how variables are related.
In my opinion, it is necessary for researchers to know how to have a notion of the relationships between variables, in order to be able to find potential cause and effect relation – however this relation is hypothetical, you can't claim that there is a cause-effect relation only because the correlation is high between those two variables-,remove unecessary variables etc. In particular we will go through Pearson correlation coefficient and Confidence interval by the bootstrap and (Principal component analysis.

Before proceeding, it might be helpful to look over the help pages for the ggplot, cor, cor.tes, boot.cor, quantile, eigen, princomp, summary, plot, autoplot.

Moreover please load the following libraries.

```
install.packages("ggplot2")
library(ggplot2)
install.packages("ggfortify")
library(ggfortify)
```

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

Compute the value of the correlation coefficient for the variables age and preg.

Exercise 2

Construct the scatterplot for the variables age and preg.

Exercise 3

Apply a correlation test for the variables age and preg with null hypothesis to be the correlation is zero and the alternative to be different from zero.
hint: cor.test

Exercise 4

Construct a 95% confidence interval is by the bootstrap. First find the correlation by bootstrap.
hint: mean

Exercise 5

Now that you have found the correlation, find the 95% confidence interval.

Exercise 6

Find the eigen values and eigen vectors for the data set(exclude the class.fac variable).

Exercise 7

Compute the principal components for the dataset used above.

Exercise 8

Show the importance of each principal component.

Exercise 9

Plot the principal components using an elbow graph.

Exercise 10

Constract a scatterplot with x-axis to be the first component and the y-axis to be the second component. Moreover if possible draw the eigen vectors on the plot.
hint: autoplot

Data science for Doctors:

Cluster Analysis Exercises

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 ninth part of the series and it aims to cover the very basics of the subject of cluster analysis.
In my opinion, it is necessary for researchers to know how to discover relationships between patients and diseases. Therefore in this set of exercises we will go through the basics of cluster analysis relationship discovery. In particular we will use hierarchical clustering and centroid-based clustering , k-means clustering and k-median clustering.

Before proceeding, it might be helpful to look over the help pages for the ggplot, geom_point, dist, hclust, cutree, stats::rect.hclust, multiplot, kmeans, kGmedian.

Moreover please load the following libraries.

```
install.packages("ggplot2")
library(ggplot2)
install.packages("Gmedian")
library(Gmedian)
```

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

Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the colour of the points based on the class of the candidate (0 or 1).

Exercise 2

Create a distance matrix for the data.

Exercise 3

Make an hierarchical clustering analysis using the single linkage method. Then create an object that contains only two clusters.

Exercise 4

Make an hierarchical clustering analysis using the complete linkage method(default). Then create an object that contains only two clusters.

Exercise 5

Construct the trees that are produced by exercises 2 and 3 and

draw the two clusters(at the plots).
hint: rect.hclust

Learn more about cluster analysis in the online course Applied Multivariate Analysis with R. In this course you will learn how to work with hiërarchical clustering, k-means clustering and much more.
Exercise 6

Construct two scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the colour of the points based on the cluster that those points belong to. Each scatterplot is for different clustering method.

If possible illustrate those scatterplots (each one at a time) next to the plot of exercise 1, to see whether the clustering can discriminate the positive classified from the negative classified patients. In case you didn’t do that, find it at the solution’s section, I highly encourage you to check it out.

Exercise 7

Run the following in order to create dummy variables data_mat <- model.matrix(~.+0, data = data).
Make a centroid-based cluster analysis using the k-means method with k to be 2. Apply the k-mean clustering on the data_mat data frame.

Exercise 8

Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the colour of the points based on the cluster (retrieved from k-mean method) that those points belong to.

If possible illustrate those scatterplot next to the plot of exercise 1.

Exercise 9

Make a centroid-based cluster analysis using the k-median method with k to be 2. Apply the k-median clustering on the data_mat data frame.

Exercise 10

Construct a scatterplot with x-axis to be the mass variable and y-axis to be the age variable. Moreover, determine the colour of the points based on the cluster (retrieved from k-median method) that those points belong to.

If possible illustrate those scatterplot next to the plot of exercise 1.

Data science for Doctors: Inferential Statistics Exercises(Part-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 eighth 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 testing the normality of distributions(Shapiro–Wilk test, Anderson–Darling test.), the existence of outliers(Grubbs' test for outliers). We will also cover the case that normality assumption doesn't hold and how to deal with it(<a href="https://en.wikipedia.org/wiki/Rank_test" Rank tests). Finally we will do a brief recap of the previous exercises on inferential statistics.

Before proceeding, it might be helpful to look over the help pages for the hist, qqnorm, qqline, shapiro.test, ad.test, grubbs.test, wilcox.test.

Moreover please load the following libraries.

```
install.packages("ggplot2")
library(ggplot2)
install.packages("nortest")
library(nortest)
install.packages("outliers")
library(outliers)
```

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),]
```

Moreover run the chunk below in order to generate the samples that we will test on this set of exercises.

```
f_1 <- rnorm(28,29,3)
f_2 <- rnorm(23,29,6)
```

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

Plot an histogram of the variable pres.

Exercise 2

Plot the QQ-plot with a QQ-line for the variable pres.

Exercise 3

Apply a Shapiro-Wilk normality test for the variable pres.

Exercise 4

Apply a Anderson-Darling normality test for the variable pres.

Exercise 5

What is the percentage of data that passes a normality test? This might be a bit challenging, consider using the apply function.

Learn more about Inferential Statistics in the online course

Learn By Example: Statistics and Data Science in R. This course includes:

- 6 different case-studies on inferential statistics
- Extensive coverage of techniques for inference
- And much more

Exercise 6

Construct a boxplot of pres and see whether there are outliers or not.

Exercise 7

Apply a Grubb’s test on the pres to see whether the variable contains outlier values.

Exercise 8

Apply a two-sided Grubb’s test on the pres to see whether the variable contains outlier values.

Exercise 9

Suppose we test a new diet on a sample of 14 people from the candidates (take a random sample from the set) and after the diet the average mass was 29 with standard deviation of 4 (generate 14 normal distributed samples with the properties mentioned before). Apply Wilcoxon signed rank test for the mass variable before and after the diet.

Exercise 10

Check whether the positive and negative candidates have the same distribution for the pres variable. In order to check that, apply a Wilcoxon rank sum test for the pres variable in respect to the class.fac variable.

Data science for Doctors: Inferential Statistics Exercises(Part-4)

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 seventh 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 F-distribution (F-test), and Chi-squared distribution (Chi-squared test). If you are not aware of what

are the mentioned distributions please go here to acquire the necessary background. The assumption of the t-test (we covered it last time here) is that the two population variances are equal. Such an assumption can serve as a null hypothesis for F-test. Moreover sometimes it happens that we want to test a hypothesis with respect to more than one probability, here is where Chi-Squared test comes into play.

Before proceeding, it might be helpful to look over the help pages for the sd, var, var.test, chisq.test.

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),]
```

Moreover run the chunk below in order to generate the samples that we will test on this set of exercises.

```
f_1 <- rnorm(28,29,3)
f_2 <- rnorm(23,29,6)
```

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

Compute the F-statistic. (test statistic for F-test)

Exercise 2

Compute the degrees of freedom for the numerator and denominator.

Exercise 3

Apply a two-sided F-test for the two samples

Exercise 4

Apply a one-sided F-test for the two samples with the alternative hypothesis to be that the standard deviation of the first sample is smaller than the second.

Exercise 5

Retrieve the p-value and the ratio of variances for both tests.

Exercise 6

Find the number of patients who show signs of diabetes and those who don’t.

Exercise 7

Assume that the hypothesis we made is that 10% of people show signs of diabetes. Is that a valid claim to make? Test it using the chi-squared test.

Exercise 8

Suppose that the mass index affects whether the patients show signs of diabetes and we assume that the people who weight more than the average are more likely to have diabetes signs. Make a matrix that contains the true-positives, false-positives, true-negatives, and false-negatives of our hypothesis.
hint: True-positive: data$class==1 & data$mass >= mean(data$mass)

Exercise 9

Test the hypothesis we made at exercise 8 using chi-squared test.

Exercise 10

The hypothesis we made at exercise 8 cannot be validated, however we have noticed that the dataset contains outliers which affect the average. Therefore we make another assumption that patients who are heavier than the 25% lightest of the patients are more likely to have signs of diabetes. Test that hypothesis.
hint: it is similar to the process we did at exercises 8 and 9 but with different criteria.

[Data science for Doctors: Inferential Statistics Exercises (part-3)]

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 sixth 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 Student’s t-distribution (Student’s t-test), which may be the most used test you will need to apply, since in most cases the standard deviation σ of the population is not known. We will cover the one-sample t-test and two-sample t-test(both with equal and unequal variance). 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 t.test.

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 25 candidates that tried a diet and they had a average weight of 29 (generate 25 normal distributed samples with mean 29 and standard deviation 4) after the experiment.
Find the t-value.

Exercise 2

Find the p-value.

Exercise 3

Find the 95% confidence interval.

Exercise 4

Apply t-test with Null Hypothesis that the true mean of the sample is equal to the mean of the sample with 5% confidence level.

Exercise 5

Apply t-test with Null Hypothesis that the true mean of the sample is equal to the mean of the population and the alternative that the true mean is less than the mean of the population with 5% confidence level.

Exercise 6

Suppose that we want to compare the current diet with another one. We assume that we test a different diet to a sample of 27 with mass average of 31(generate normal distributed samples with mean 31 and standard deviation of 5). Test whether the two diets are significantly different.
Note that the two distributions have different variances.
hint: This is a two sample hypothesis testing with different variances.

Exercise 7

Test whether the the first diet is more efficient than the second.

Exercise 8

Assume that the second diet has the same variance as the first one. Is it significant different?

Exercise 9

Assume that the second diet has the same variance as the first one. Is it significantly better?

Exercise 10

Suppose that you take a sample of 27 with average mass of 29, and after the diet the average mass is 28(generate the sampled with rnorm(27,average,4)). Are they significant different?
hint: Paired Sample T-Test.

[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)

Data Science for Doctors – Part 2 : Descriptive Statistics

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 second part of the series, it will contain the main descriptive statistics measures you will use most of the time. Those measures are divided in measures of central tendency and measures of spread. Moreover, most of the exercises can be solved with built-in functions, but I would encourage you to solve them "by hand", because once you know the mechanics of the measures, then you are way more confident on using those measures. On the "solutions" page, I have both methods, so even if you didn't solve them by hand, it would be nice if you check them out.

Before proceeding, it might be helpful to look over the help pages for the mean, median, sort , unique, tabulate, sd, var, IQR, mad, abs, cov, cor, summary, str, rcorr.

You also may need to load the Hmisc library.
```
install.packages('Hmisc')
```

```
library(Hmisc)
```

In case you haven't solve the part 1, run the following script to load the prerequisites for this part.

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

Find the mean of the mass variable.

Exercise 2

Find the median of the mass variable.

Exercise 3

Find the mode of the mass.

Exercise 4

Find the standard deviation of the age variable.

Learn more about descriptive statistics in the online courses Learn by Example: Statistics and Data Science in R (including 8 lectures specifically on descriptive statistics), and Introduction to R.

Exercise 5

Find the variance of the mass variable.

Unlike the popular mean/standard deviation combination,interquartile range and median/mean absolute deviation are not sensitive to the presence of outliers. Even though it is recommended to go for MAD because they can approximate the standard deviation.

Exercise 6

Find the interquartile range of the age variable.

Exercise 7

Find the median absolute deviation of age variable. Assume that the age follows a normal distribution.

Exercise 8
Find the covariance of the variables age, mass.

Exercise 9

Find the spearman and pearson correlations of the variables age, mass.

Exercise 10

Print the summary statistics, and the structure of the data set. Moreover construct the correlation matrix of the data set.

Data Science for Doctors – Part 1 : Data Display

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 first part of the series, it is going to be about data display.

Before proceeding, it might be helpful to look over the help pages for the table, pie, geom_bar , coord_polar, barplot, stripchart, geom_jitter, density, geom_density, hist, geom_histogram, boxplot, geom_boxplot, qqnorm, qqline, geom_point, plot, qqline, geom_point .

You also may need to load the ggplot2 library.

```
install.packages('ggplot2')
library(ggplot2)
```

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
```

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

Create a frequency table of the class variable.

Exercise 2

```
data$class.fac <- factor(data[['class']],levels=c(0,1),
labels= c("Negative","Positive"))
```

Create a pie chart of the class.fac variable.

Exercise 3

Create a bar plot for the age variable.

Exercise 4

Create a strip chart for the mass against class.fac.

Exercise 5

Create a density plot for the preg variable.

Exercise 6

Create a histogram for the preg variable.

Exercise 7

Create a boxplot for the age against class.fac.

Exercise 8

Create a normal QQ plot and a line which passes through the first and third quartiles.

Exercise 9

Create a scatter plot for the variables age against the mass variable .

Exercise 10

Create scatter plots for every variable of the data set against every variable of the data set on a single window.
hint: it is quite simple, don’t overthink about it.