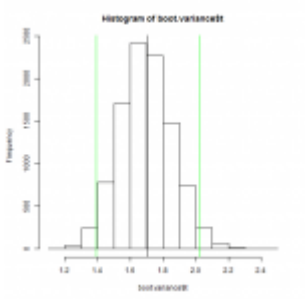
Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-3)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In the first two part of this series, we've seen how to identify the distribution of a random variable by plotting the distribution of a sample and by estimating statistic. We also seen that it can be tricky to identify a distribution from a small sample of data. Today, we'll see how to estimate the confidence interval of a statistic in this situation by using a powerful method called bootstrapping.

Answers to the exercises are available here.

Exercise 1
Load this dataset and draw the histogram, the ECDF of this sample and the ECDF of a density who's a good fit for the data.

Exercise 2
Write a function that takes a dataset and a number of iterations as parameter. For each iteration this function must

create a sample with replacement of the same size than the dataset, calculate the mean of the sample and store it in a matrix, which the function must return.

Exercise 3
Use the t.test() to compute the 95% confidence interval estimate for the mean of your dataset.

Learn more about bootstrapping functions in the online course Structural equation modeling (SEM) with lavaan. In this course you will learn how to:

- Learn how to develop bootstrapped confidence intervals
- Go indepth into the lavaan package for modelling equations
- And much more

Exercise 4
Use the function you just wrote to estimate the mean of your sample 10,000 times. Then draw the histogram of the results and the sampling mean of the data.

The probability distribution of the estimation of a mean is a normal distribution centered around the real value of the mean. In other words, if we take a lot of samples from a population and compute the mean of each sample, the histogram of those mean will look like one of a normal distribution center around the real value of the mean we try to estimate. We have recreated artificially this process by creating a bunch of new sample from the dataset, by resampling it with replacement and now we can do a point estimation of the mean by computing the average of the sample of means or compute the confidence interval by finding the correct percentile of this distribution. This process is basically what is called bootstrapping.

Exercise 5

Calculate the value of the 2.5 and 97.5 percentile of your sample of 10,000 estimates of the mean and the mean of this sample. Compare this last value to the value of the sample mean of your data.

Exercise 6
Bootstrapping can be used to compute the confidence interval of all the statistics of interest, but you don’t have to write a function for each of them! You can use the boot() function from the library of the same name and pass the statistic as argument to compute the bootstrapped sample. Use this function with 10,000 replicates to compute the median of the dataset.

Exercise 7
Look at the structure of your result and plot his histogram. On the same plot, draw the value of the sample median of your dataset and plot the 95% confidence interval of this statistic by adding two vertical green lines at the lower and higher bounds of the interval.

Exercise 8
Write functions to compute by bootstrapping the following statistics:

- Variance
- kurtosis
- Max
- Min

Exercise 9
Use the functions from last exercise and the boot function with 10,000 replicates to compute the following statistics:

- Variance
- kurtosis
- Max
- Min

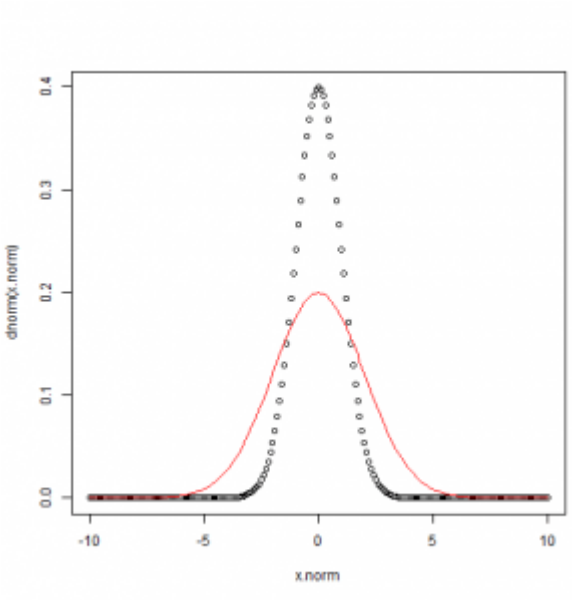
Then draw the histogram of the bootstrapped sample and plot

the 95% confidence interval of the statistics.

Exercise 10
Generate 1000 points from a normal distribution of mean and standard deviation equal to the one of the dataset. Use the bootstrap method to estimate the 95% confidence interval of the mean, the variance, the kurtosis, the min and the max of this density. Then plot the histograms of the bootstrap samples for each of the variable and draw the 95% confidence interval as two red vertical line.

Two bootstrap estimate of the same statistic of two sample who are distributed by the same density should be pretty similar. When we compare those last plots with the confidence interval we drawn before we see that they are. More importantly, the confidence interval computed in exercise 10 overlap the confidence interval of the statistics of the first dataset. As a consequence, we can't conclude that the two sample come from different density distribution and in practice we could use a normal distribution with a mean of 0.4725156 and a standard deviation of 1.306665 to simulate this random variable.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-2)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In the last exercise set we've seen that random variable can be described by mathematical functions called probability density and that when we know which one describe a particular random process we can use it to compute the probability of realization of a given event. We have also seen how to use an histogram and an ECDF plot to identify which function express the random variable. Today, we will see which mathematical properties of those function we can compute to help us find the probability density who fit a sample. Those properties are called statistics and our job today is to estimate the real value of those properties by using a small sample of data.

Answers to the exercises are available here.

Exercise 1

The most commonly used statistics is the mean, which is the center of mass of the distribution, i.e. the point on the x axis where the weighted relative position of each observation sum to zero. For example, draw the density of a standard normal distribution and add to the plot a vertical line to indicate the mean of this distribution. Then, draw another plot, but this time of an exponential distribution with a rate

of 1 and his mean.

From the density plot of the standard normal distribution we can see how the mean represent the center of mass of the distribution: the normal distribution is symmetric, so the mean is in the center of the plot of the function. The exponential function is not symmetric, in this case the mean is the point where all the points with a small y value, at the right of the mean on the plot, counterbalance the few points with a high y value at the left of the mean. Since the value of the mean is at the center of the distribution, we often use the mean to represent a typical value of a probability distribution. The mean also give us the ability to put a number on the location of a probability distribution on the axis.

Exercise 2
In practice, we don’t have access to the probability density function of a random variable and can’t compute directly the mean of the distribution. We must estimate it using a sample of observations of that random variable. Since it’s random, all samples will be different and our estimations of the mean, will all be different.

Generate 500 points from an exponential distribution with a rate of 0.5. Draw the histogram of the sample and compute the sample mean of this distribution. Then write a function that repeat this process for n iterations, store the sample mean in a vector and return this vector. Use this function to compute 10,000 sample means, plot the histogram of the sample means and compute the mean of those estimations.

Exercise 3
From the histogram of the sample mean, we can see that the estimations follow a normal distribution centered around the real value of the mean. We can use this fact to compute the interval which have a certain probability of containing the real value of the mean. This interval is called the confidence

interval of the estimate and the probability that this interval contain the real mean of the distribution is called the confidence level. In the next exercise set, we will see methods to compute this interval directly from a sample without knowing the probability density function of the random variable.

Use the quantile() function to compute the 2.5 and 97.5 percentile from the sample of estimations of the mean, then use the t.test() function to compute the confidence interval with a level of 95% of the original distribution and compare those values.

Learn more about density functions in the online course Learning Data Mining with R. In this course you will learn how to:

- Work with clustering methods, KNN classification algorithms and density functions
- Go indepth into different data mining tools available in R
- And much more

Exercise 4

Load this dataset and use the t.test() function to compute the confidence interval of the mean for both variables with a level of 95%. Does those random variables seems to follow distributions who have the same means?

We see that the confidence intervals doesn't overlap. This is an indication that the real value of the mean of the first variable is not in the same interval as the distribution mean of the second variable. As a consequence, we can safely suppose than both mean are different and that they don't have the same probability distribution.

Exercise 5

Another useful statistics is the variance. This statistic is an indication of how the data are spread around the mean. So if two distributions have the same mean, the one with the smallest variance has the most homogeneous value, while the one with the highest variance has more small and high value far from the mean. A related statistics is the standard deviation, which is defined as the square root of the variance.

Draw the density of a standard normal distribution and of a normal distribution of mean equal to zero and with a standard deviation of 5 to see the effect of a change of variance on a density.

Exercise 6

In the case of the variance, we cannot directly compute the confidence interval without making assumption on the type of distribution the sample come from or use some fancy method we will introduce in the next exercise set. Luckly for us we can use thevar.test() function to verify if the variances are equal. Use this function on the dataset of exercise 4 three time, once with the alternative parameter set to “two.sided”, then to “less” and finally to “greater”. What is the signification of the three test?

Exercise 7

If the mean is a good representation of the typical value of a random variable defined by a density, this statistics can be skew by outliers. When a sample has outlier a better statistics to use is the median, which is the value that separate the range of observations that can be generated by a random variable in two equal parts.

Generate 200 points from a log-normal distribution with a parameter meanlog = 0 and sdlog = 0.5. Then plot the histogram of those points and represent the mean and the median of this sample by two vertical lines.

Exercise 8

The median is a special case of a more general statistics called quantile, which are cutpoints dividing the domain of a probability density function into sub-interval containing the same amount of observations. So the 2-quantile is the median, since this statistics separate the domain of a probability distribution in two sub-interval containing 50% of the observations. Other quantile statistics often used are the 4-quantile, called quartile, which are the values on the domain of a probability distribution that separates it in four sub-interval containing 25% of the observations and the 100-quantile, called percentille, which are the values that separate this domain in 100 part containing 1% of the observations.

Compute the median, the quantile and the 5 and 95% percentile on the variables of the dataset of exercise 4. Then compute the interquartile range which is the difference between the 25% and the 75% quartile. Does those statistics suggest that the two samples have the same distribution?

Exercise 9

Another statistics that can be used to differentiate two probability distribution is the skewness. As his name imply, the skewness is a measure of how much there is an imbalance between the observations at the right of the mean and at the left of the mean. A negative skewness indicate that the distribution is skew to the left, a positive value indicate that the distribution is skew to the right and a skewness of zero tell us that the distribution is perfectly symmetric.

Load the moment package and use the skewness() function to compute the skewness of three samples you must create:

- 150 points sample from a standard normal distribution
- 1000 points sample from a standard normal distribution
- 200 points sample from a exponential distribution with a rate of 5

Exercise 10

The last statistic we will use today is the kurtosis, which describe the general shape of the probability distribution. When the kurtosis is greater than zero, the probability distribution has heavy tail and a pointy shape. Both of those characteristics are proportional to the magnitude of the kurtosis. If the kurtosis is less than zero, the distribution has a more regular shape with light tails. When this statistic has a value of zero, the distribution's shape look a lot like the normal distribution.

Use the kurtosis() function to compute the kurtosis of those samples:

- 500 points sample from a standard normal distribution
- 500 points sample from a exponential distribution with a rate of 5
- 500 points sample from uniform distribution

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.

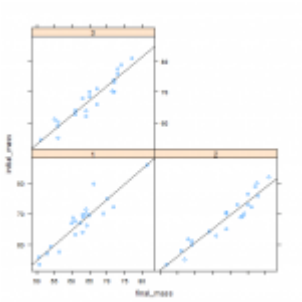
Exercise 9

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

Experimental Design Exercises



In this set of exercises we shall follow the practice of conducting an experimental study. Researcher wants to see if there is any influence of working-out on body mass. Three groups of subjects with similar food and sport habits were included in the experiment. Each group was subjected to a different set of exercises. Body mass was measured before and after workout. The focus of the research is the difference in body mass between groups, measured after working-out. In order to examine these effects, we shall use paired t test, t test for independent samples, one-way and two-ways analysis of variance and analysis of covariance.

You can download the dataset here. The data is fictious.

Answers to the exercises are available here.

If you have different solution, feel free to post it.

Exercise 1

Load the data. Calculate descriptive statistics and test for the normality of both initial and final measurements for whole sample and for each group.

Exercise 2

Is there effect of exercises and what is the size of that effect for each group? (Tip: *You should use paired t test.*)

Exercise 3

Is the variance of the body mass on final measurement the same for each of the three groups? (Tip: *Use Levene's test for homogeneity of variances*)

Exercise 4

Is there a difference between groups on final measurement and what is the effect size? (Tip: *Use one-way ANOVA*)

Exercise 5

Between which groups does the difference of body mass appear after the working-out? (Tip: *Conduct post-hoc test.*)

Exercise 6

What is the impact of age and working-out program on body mass on final measurement? (Tip: *Use two-way between groups ANOVA.*)

Exercise 7

What is the origin of effect of working-out program between subjects of different age? (Tip: *You should conduct post-hoc test.*)

Exercise 8

Is there a linear relationship between initial and final measurement of body mass for each group?

Exercise 9

Is there a significant difference in body mass on final measurement between groups, while controlling for initial measurement?

Exercise 10

How much of the variance is explained by independent variable? How much of the variance is explained by covariate?

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.

- 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)

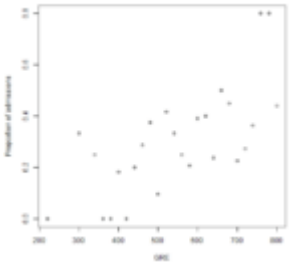
Exercise 9

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.

Logistic Regression Exercises



In the exercises below we cover some material on logistic regression in R. Logistic regression is used when our response variable is binary (or dichotomous), i.e., takes on two categories (usually ‘yes’ and ‘no’), and we usually have one continuous predictor variable.

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.

We will be using the dataset at http://www.ats.ucla.edu/stat/data/binary.csv, which includes

the variables admit (whether a student was admitted to the school), gre (the student’s GRE score), gpa (the student’s GPA), and rank (school’s prestige; 1=highest, 4=lowest).

First, let’s obtain our data:

```
admissions <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
```

Exercise 1
Suppose that we would like to predict admission based on a student's GRE score, using a logistic regression model. We can check whether a logistic regression model is a good idea by plotting proportions of admissions (by value of GRE) vs. our predictor variable. Compute these proportions.

Exercise 2
Create the plot mentioned in Exercise 1.

Exercise 3
Fit the model specified in Exercise 1.

Exercise 4
Extract the model coefficients.

Exercise 5
What is the estimated increase in the odds of admission for each additional 100 points on GRE score?

Exercise 6
Give a 95% confidence interval for the estimate in Exercise 5. Interpret.

Exercise 7
Use the value of residual deviance to test the fit of the model.

Exercise 8
Plot the fitted logistic regression curve (i.e., probability of admission vs. GRE).

Exercise 9
Extract the predicted values.

Exercise 10
Predict the probability of admission for a student whose GRE score is 580.

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)