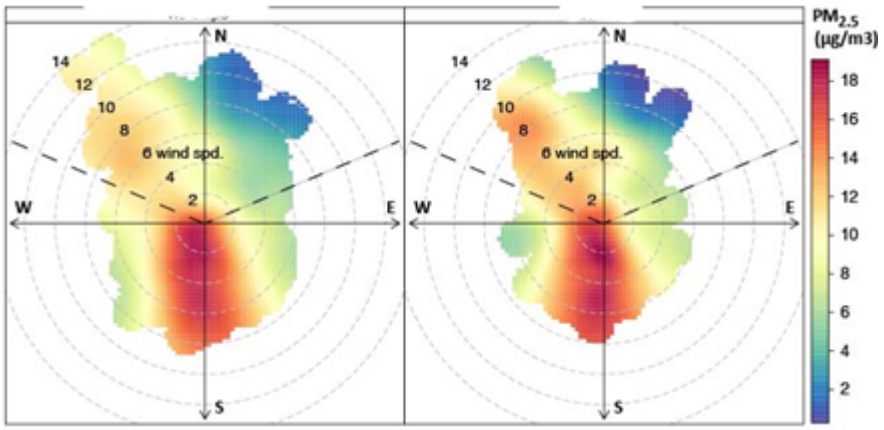
Working with air quality and meteorological data Exercises (Part-1)



Atmospheric air pollution is one of the most important environmental concerns in many countries around the world, and it is strongly affected by meteorological conditions. Accordingly, in this set of exercises we use openair package to work and analyze air quality and meteorological data. This packages provides tools to directly import data from air quality measurement network across UK, as well as tools to analyse and producing reports. In this exercise set we will import and analyze data from MY1 station which is located in Marylebone Road in London, UK.

Answers to the exercises are available here.

Please install and load the package openair before starting the exercises.

Exercise 1
Import the MY1 data for the year 2016 and save it into a dataframe called my1data.

Exercise 2
Get basic statistical summaries of myd1 dataframe.

Exercise 3
Calculate monthly means of:
a. pm10

b. pm2.5
b. nox
c. no
d. o3

You can use Air Quality Data and weather patterns in combination with spatial data visualization, **Learn more** about spatial data in the online course [[Intermediate] Spatial Data Analysis with R, QGIS & More](). this course you will learn how to:

- Work with Spatial data and maps
- Learn about different tools to develop spatial data next to R
- And much more

Exercise 4

Calculate daily means of:

a. pm10
b. pm2.5
b. nox
c. no
d. o3

Exercise 5

calculate daily maximum of:

b. nox
c. no

Sending Emails from R

Exercises

When monitoring a data source, model, or other automated process, it's convienent to have method for easily delivering performance metrics and notifying you whenever something is amiss. One option is to use a dashboard; however, this requires active time and effort to grab numbers and catch errors. An alternative approach is to send an email alert on the performance of the process. In this exercise set, we will explore the email approach using the mailR package.

Exercises in this section will be solved using the mailR package as well as basic HTML and CSS. It is recommended to take a look at the mailR documentation before continuing.

Answers to the exercises are available here.

Exercise 1
Let's begin by sending "Hello World. This is my email!" as the body parameter from yourself to yourself.

Exercise 2
By passing in a vector for the to parameter, you can send the email to multiple recipients. Send the above email to yourself and a friend.

Exercise 3
So far, your emails have had no subject. Send the email from Exercise 1 to yourself with "Email Testing" for the subject parameter.

Exercise 4

With this package, we can take full advantage of CSS when constructing the body of an email. Send the email from the previous exercise from yourself to yourself where “Hello World.” is now red and “This is my email!” is now blue.

Note: make sure that html = TRUE.

Learn more about html functionality and web connection in the online course A complete journey to web analytics using R tool. In this course you will learn how to:

- Perform a web based analytic question start to end
- Learn how to import data from different online platforms such as twitter
- And much more

Exercise 5
If you write a complex email containing images, dynamic elements, etc. as an HTML file, then you can reference this file with the body parameter. Create an HTML file containing “Hello World. This is my email!” called *my_email.html*. Send this email to yourself.

Exercise 6
Using knitr, you can compile HTML files. Compile the default knitr document that uses the mtcars dataset to an HTML file and email this to yourself.

Exercise 7
Create a new R script called *mailr_six.R* containing your code from the above exercises and attach that to your email by referencing the file path to *mailr_six.R* in the attach.files parameter. Send this email from yourself to yourself.

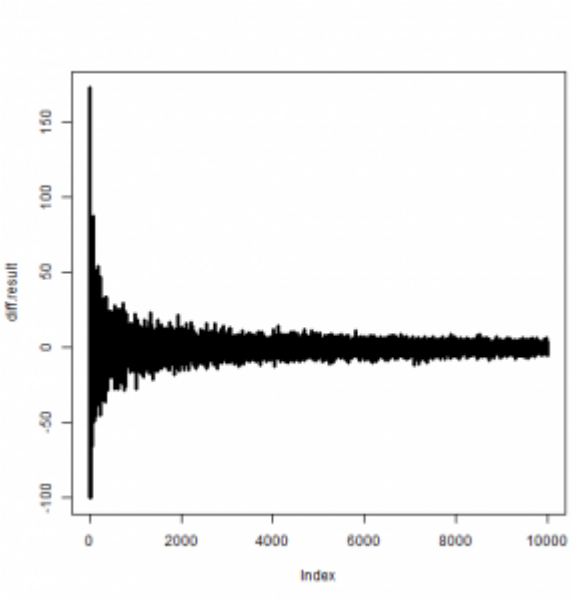
Exercise 8
The attached R script above does not have a description or a name. Add these in the file.descriptions and file.names

parameters, respectively. Send the resulting email to yourself.

Exercise 9
Just as with the recicipents, you can attach multiple files, descriptions, and names by passing in vectors to the respective parameters. Create a new R script called *mailr_eight.R* containing your code from the above exercises and attach both *mailr_six.R* and *mailr_eight.R* to your email. Send the resulting email to yourself.

Exercise 10
Create a new R script where a random integer called *important_number* is generated. If *important_number* is even, then send an email to yourself notifying you that *important_number* is even.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-6)



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 previous set, we've seen how to compute probability based on certain density distributions, how to simulate situations to compute their probability and use that knowledge make decisions in obvious situation. But what is a probability? Is there a more scientific way to make those decisions? What is the P-value xkcd keep talking about? In this exercise set, will learn the answer to most of those question and more!

One simple definition of the probability that an event will occur is that it's the frequency of the observations of this event in a data set divided by the total number of observations in this set. For example, if you have a survey where 2 respondents out of 816 says that they are interested in a potential partner only if they are dressed in an animal costume, you can say that the probability that someone in the population is a furry is about 2/816 or 1/408 or 0.00245… or 0.245%.

Answers to the exercises are available here.

Exercise 1
The average height of males in the USA is about 5 foot 9

inches with a standard deviation of 2.94 inches. If this measure follow a normal distribution, write a function that takes a sample size as input and compute the probability to have a subject taller than 5 foot 8 and smaller than 5 foot 9 on this sample size. Then, set the seed to 42 and compute the probability for a sample size of 200.

Exercise 2

We can deduce a lot from that definition. First, the probability is always a fraction, but since we are usually not used to high number and have a hard time doing division in our head 3968/17849 is not a really useful probability. In consequence, we will usually use a percentage or a real number between o and 1 to represent a probability. Why 0 and one? If an event is not present in the data set, his frequency is 0 so whatever is the total number of observations his probability is 0 and if all the observations are the same, the fraction is going to be equal to 1. Also, if you think about the example of the furries in the survey, maybe you think that there's a chance that there are only two furries in the entire population and they both take the survey, so the probability that an individual is a furry is in reality a lot lower than 0.0245%. Or maybe there's a lot more furries in the population and only two where surveyed, which makes the real probability much higher. You are right token reader! In a survey, we estimate the real probability and we can never tell the real probability from a small sample (that's why if you are against the national survey in your country, all the statisticians hate you in silence). However, the more the sample size of a survey is high the less those rare occurrences happen.

1. Compute the probability that an American male is taller than 5 foot 8 and smaller than 5 foot 9 with the pnorm function.
2. Write a function that draws a sample of subject from this distribution, compute the probability of observing a male of this height and compute the percentage of

difference between that estimate and the real value. Make sure that you can repeat this process for all sample size between two values.
3. Use this function to draw sample of size from 1 to 10000 and store the result in a matrix.
4. Plot the difference between the estimation of the probability and the real value.

This plot show that the more the sample size is big, the less the error of estimation is, but the difference of error between an sample of size 1000 and 10000 is quite small.

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Exercise 3

We have already seen that density probability can be used to compute probability, but how?

For a standard normal distribution:

1. Compute the probability that x is smaller or equal to zero, then plot the distribution and draw a vertical line at 0.
2. Compute the probability that x is greater than zero.
3. Compute the probability that x is less than -0.25, then plot the distribution and draw a vertical line at -0.25.
4. Compute the probability that x is smaller than zero and greater than -0.25.

Yeah, the area under the curve of a density function between two points is equal to the probability that an event is equal to a value on this interval. That's why density are really useful: they help us to easily compute the probability of an event by doing calculus. Often we will use the cumulative distribution function (cdf), which is the antiderivative of the density function, to compute directly the probability of an event on an interval. The function pnorm() for example, compute the value of the cdf between minus infinity and a value x. Note that a cdf return the probability that a random variable take a value smaller.

Exercise 4

For a standard normal distribution, find the values x such as:

1. 99% of the observation are smaller than x.
2. 97.5% of the observation are smaller than x.
3. 95% of the observation are smaller than x.
4. 99% of the observation are greater than x.
5. 97.5% of the observation are greater than x.
6. 95% of the observation are greater than x.

Exercise 5

Since probability are often estimated, it is useful to measure how good is the estimation and report that measure with the estimation. That's why you often hear survey reported in the form of "x% of the population with a y% margin 19 times out of 20". In practice, the size of the survey and the variance of the results are the two most important factors that can influence the estimation of a probability. Simulation and bootstrap methods are great way to find the margin of error of an estimation.

Load this dataset and use bootstrapping to compute the interval that has 95% (19/20) chance to contain the real probability of getting a value between 5 and 10. What is the margin of error of this estimation?

This process can be used to any statistics that is estimated,

like a mean, a proportion, etc.

When doing estimation, we can use a statistic test to draw conclusion about our estimation and eventually make decisions based on it. For example, if in a survey, we estimate that the average number of miles traveled by car each week by American is 361.47, we could be interested to know if the real average is bigger than 360. To do so, we could start by formulation a null and an alternative hypothesis to test. In our scenario, a null hypothesis would be that the mean is equal or less than 360. We will follow the step of the test and if at the end we cannot support this hypothesis, then we will conclude that the alternative hypothesis is probably true. In our scenario that hypothesis should be that the mean is bigger than 360.

Then we choose a percentage of times we could afford to be wrong. This value will determine the range of possible values for which we will accept the null hypothesis and is called the significance level (α).

Then we can use a math formula or a bootstrap method to estimate the probability that a sample from this population would create an estimate of 361.47. If this probability is less than the significance level, we reject the null hypothesis and go with the alternative hypothesis. If not, we cannot reject the null hypothesis.

So basically, what we do is we look at how often our estimation should happen if the null hypothesis is true and if it's rare enough at our taste, significance level, we conclude that it's not a random occurance but a sign that the null hypothesis is false.

Exercise 6

This dataset represents the survey of the situation above.

1. Estimate of the mean of this dataset.
2. Use the bootstrap method to find 10000 estimations of the mean from this dataset.

3. Find the value from this bootstrap sample that is bigger than 5% of all the others values.This value is called the critical value of the test and correspond to α.
4. From the data we have, should be conclude that the mean of the population is bigger than 360? What is the significance level of this test?

Exercise 7

We can represent the test visually. Since we reject the null hypothesis if the percentage of bootstrapped mean smaller than 360 is bigger than 5%, we can simply look where the fifth percentile lie on the histogram of the bootstrapped mean. If it's at the left of the 360 value, we know that more than 5% of bootstrapped means are smaller than 360 and we don't reject the null hypothesis.

Draw the histogram of the bootstrapped mean and draw two vertical lines: one at 360 and one at the fifth percentile.

Exercise 8

There are two ways that a mean can be not equal to a value: when the mean is bigger than the value and when it's smaller than this value. So if we want to test the equality of the mean to a specific value we must verify if most of our estimations lie around this value or if a lot of them are far from it. To do so, we create an interval who has for endpoints our mean and another point that is at the same distance from this value that the mean. Then we can compute the probability to get an estimation outside this interval. This way, we test if the value is not bigger or smaller than the value 1-α of the time.

Here's the steps to test the hypothesis that the mean of the dataset of exercise 6 is equal to 363:

1. To simulate that our distribution has a mean of 363, shift the dataset so that this value become the mean.
2. Generate 10000 bootstrapped means from this

distribution.
3. Compute the endpoints of the test interval.
4. Compute the probability that the mean is outside this interval.
5. What conclusion can we make with a α of 5%?

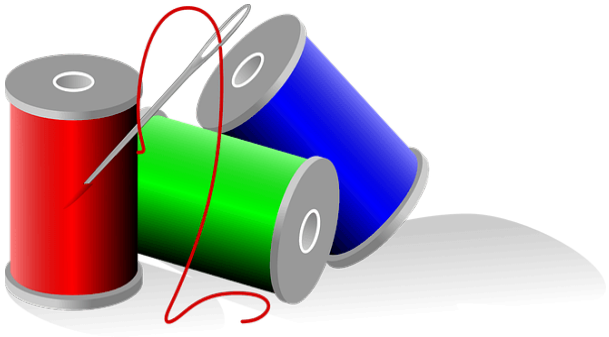
Exercise 9
Repeat the step of exercise 8, but this time test if the mean is smaller than 363.

This show that a one direction test is more powerful than a two direction test in this situation since there's less wiggle room between the value of reference and the critical region of the test. So if you have prior knowledge that could make you believe that an estimation is bigger or smaller than a value, testing for than would give you more assurance of the validity of your results.

Exercise 10
The p-value of a test is the probability that we would observe a random estimation as the one we made if the null hypothesis is true. This value is often used in scientific reports since it's a concise way to express statistics finding. If we know the p-value of a test and the significance level α we can deduce the result of the test since the null hypothesis is rejected when $p<\alpha$. In another word: you have been using the p-value all this time to make conclusion!

Load the dataset of exercise 5 and compute the p-value associated to the test that the mean is equal to 13 if α is equal to 5%.

More string Hacking with Regex and Rebus



For a begineer in R or any language,regular expression might seem like a daunting task . Rebus package in R gives a lowers the barrier for common regular expression tasks and is useful for a begineer or even for advanced users for most of the common regex skills in a more intuitive yet verbose way .Check out the package and try this exercises to test your knowledge .
Load stringr/stringi as well for this set of exercise . I encourage you to do this and
this before working on this set .
Answers 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 two strings
Suppose you have a vectore
x <- c("stringer","stringi","rebus","redbus")

use rebus and find the strings starting with st .Hint use START from rebus

Exercise 2

Use the same string vectore and find the strings which ends

with bus.

Exercise 3

you have a vector like

```
m <- c("aba","aca","abba","accdda")
```

find the strings which starts and ends with a and have a single character in between

Hint – use ANY_CHAR

Exercise 4

```
y <- c("brain","brawn","rain","train")
```

find all the strings that starts with br and ends with n .

Hint – use any_char with hi=Inf to build the regex

Learn more about Text analysis in the online course [Text Analytics/Text Mining Using R](). In this course you will learn how create, analyse and finally visualize your text based data source. Having all the steps easily outlined will be a great reference source for future work.

Exercise 5

Use the same vector as previous exercise and find strings starting with br or tr .

Hint – or

Exercise 6

Now we turn our attention to character class,if you are familiar with character classes in regex , you will find it pretty easy with rebus and if you are starting with regex .you might find it easy to remember with rebus

Suppose you have a vector

```
l <- c("Canada","america","france")
```

Find string with C or m in it .so your answer should be Canada and America

Exercise 7

From the string 123abc ,find the digits ,using rebus .

Exercise 8
Create a character class for vowels and find all the Vowels in the vector
vow <- c("blue","sue","CLUE","TRUE")

Exercise 9
Find the characters other than vowels from above vector .

Exercise 10
Now create a new vector
vow1 <- c("blue","sue","CLUE","TRUE","aue")

find the string which is made of only vowels

Soccer data sparring: Scraping, merging and analyzing exercises

While understanding and spending time improving specific techniques, and strengthening indvidual muscles is important, occasionally it is necessary to do some rounds of actual sparring to see your flow and spot weaknesses. This exercise sets forces you to use all that you have practiced: to scrape links, download data, regular expressions, merge data and then analyze it.

We will download data from the website football-data.co.uk that has data on some football/soccer leagues results and odds quoted by bookmakers where you can bet on the results.

Answers are available here.

Exercise 1

Use R to scan the German section on football-data.co.uk for any links and save them in a character vector called all_links. There are many ways to accomplish this.

Exercise 2

Among the links you found should be a number pointing to comma-separated values files with data on Bundesliga 1 and 2 separated by season. Now update all_links vector so that only links to csv files remain. Use regular expressions.

Learn more about Data Pre-Processing in the online course R Data Pre-Processing & Data Management – Shape your Data!. In this course you will learn how to:

- import data into R in several ways while also beeing able to identify a suitable import tool
- use SQL code within R
- And much more

Exercise 3

Again, update all_links so that only links to csv tables 'from Bundesliga 1 from Season 1993/1994 to 2013/2014 inclusive' remain.

Exercise 4

Import to a list in your workspace all the 21 remaining csv files in all_links, each one as a data.frame. Use read.csv, with the url and na.strings = c("", "NA"). Not that you might

need to add a prefix for them, so the links are complete.

Exercise 5

Take the list and generate a one big data.frame with all the data.frames previously imported. One way to do this is using rbind.fill function from a well-known package. Name the new data.frame as bundesl.

Exercise 6

Take a good look at the new dataset. Our read.csv did not work perfectly on this data: it turns out that there are some empty rows and empty columns, identify and count them. Update the bundesl so it no longer has empty rows m nor columns.

Exercise 7

Format the Date column so R understands using as.Date().

Exercise 8

Remove all columns which are not 100% complete, and the variable Div as well.

Exercise 9

Which are the top 3 teams in terms of numbers of wins in Bundesliga 1 for our period? You are free to use base-R functions or any package. Be warned that his task is not as simple as it seems due the nature in the data and small inconsitency in the data.

Exercise 10

Which team has held the longest winning streak in our data?

Data wrangling : Transforming (3/3)

Data wrangling is a task of great importance in data analysis. Data wrangling, is the process of importing, cleaning and transforming raw data into actionable information for analysis. It is a time-consuming process which is estimated to take about 60-80% of analyst's time. In this series we will go through this process. It will be a brief series with goal to craft the reader's skills on the data wrangling task. This is the third part of the series and it aims to cover the transforming of data used.This can include filtering, summarizing, and ordering your data by different means. This also includes combining various data sets, creating new variables, and many other manipulation tasks. At this post, we will go through a few more advanced transformation tasks on mtcars data set, in particular table manipulation.

Before proceeding, it might be helpful to look over the help pages for the iner_join, full__join, left_join, right_join, semi_join, anti_join, intersect, union, setdiff, bind_rows.

Moreover please load the following libraries and run the following link.

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

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 new object named *car_inner* containing the observations that have matching values in both tables *mtcars* and *cars_table* using as key the variable ID.

Exercise 2

Create a new object named *car_left* containing all the observations from the left table (*mtcars*), and the matched records from the right table (*cars_table*) using as key the variable ID.

Learn more about Data Pre-Processing in the online course R Data Pre-Processing & Data Management – Shape your Data!. In this course you will learn how to:

- Work with popular libraries such as dplyr
- Learn about methods such as pipelines
- And much more

Exercise 3

Create a new object named *car_right* containing all the observations from the right table (*cars_table*), and the matched records from the right table (*mtcars*) using as key the variable ID.

Exercise 4

Create a new object named *car_full* containing all the observations when there is a match in either left (*cars_table*) or right (*mtcars*) table observation using as key the variable ID.

Exercise 5

Create a new object named *car_semi* containing all the observations from *mtcars* where there are matching values in

cars_table using as key the variable ID.

Exercise 6
Create a new object named *car_anti* containing all the observations from *mtcars* where there are not matching values in *cars_table* using as key the variable ID.

Exercise 7

Create a new object named *cars_inter* which contains rows that appear in both tables *mtcars* and *cars*.

Exercise 8

Create a new object named *cars_union* which contains rows appear in either tables *mtcars* and *cars*.

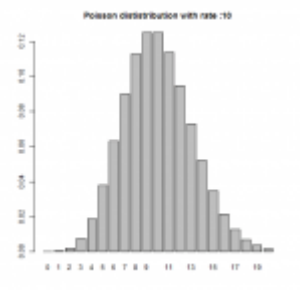
Exercise 9

Create a new object named *cars_diff* which contains rows appear in table *mtcars* and not *cars*.

Exercise 10

Append *mtcars* to *cars* and assign it at the object *car_rows*.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-5)



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 today's set you will have to use the stuff you've seen in the first fourth installment of this series of exercise set, but in a more practical setting. Take this as a fun test before we start learning cool stuff like A/B testing, conditional probability and the Bayes theorem. I hope you will enjoy doing it!

Answers to the exercises are available here.

Exercise 1

A company makes windows who should be able to withstand wind of 120 km/h. The quality assurance department of that company has for mandate to make sure that the failure rate of those windows is less than 1% for each batch of windows produced by their factory. To do so, they choose randomly 10 windows per batch of 150 and place them in a wind tunnel where they are tested.

1. Which probability function should be used to compute the number of failing engine in a QA test if the failure rate is 1%?
2. What is the probability that a windows work correctly during the QA test?

3. What is the probability that no windows breaks during the test?
4. What is the probability that up to 3 windows breaks during the test?
5. Simulate this process to estimate the average amount of engine failure during the test

Exercise 2

A team of biologist is interested in a type of bacteria who seems to be resistant to extreme change to their environment. In a particular study they put a culture of bacteria in an acidic solution, observed how many days 250 individual bacteria would survive and created this dataset. Find the 90% confidence interval for the mean of this dataset.

Exercise 3

The MNIST database is a large dataset of handwritten digits used by data scientist and computer science experts as a reference to test and compare the effectiveness of different machine learning and computer vision algorithms. If a state of the art algorithm can identify the handwritten digits in this dataset 99,79% of the time and we use this algorithm on a set of 1000 digits:

1. What is the probability that this algorithm doesn't recognize 4 digits?
2. What is the probability that this algorithm doesn't recognize 6 or 7 digits?
3. What is the probability that this algorithm doesn't recognize 3 digits or less?
4. If we use this algorithm on a set of 3000 digits, what is the probability that it fails more than 10 times?

Exercise 4

A custom officer in an airport as to check the luggage of every passenger that goes through custom. If 5% of all passenger travels with forbidden substances or objects:

1. What is the chance that the fourth traveler who is checked has a forbidden item in his luggage?
2. What is the probability that the first traveler caught with forbidden item is caught before the fourth traveler?

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Exercise 5

A start-up want to know if their marketing push in a specific market has been successful. To do so, they interview 1000 people in a survey and ask them if they know their product. Of that number, 710 where able to identify or name their product. Since the start-up has limited resource, they decided that they would reallocate half the marketing budget to their data science department if more than 70% of the market knew about their product.

1. Simulate the result of the survey by creating a matrix containing 710 ones representing the positive response and 290 zeros representing the negative response to the survey.
2. Use bootstrapping to compute the proportion of positive answer that is smaller than 95% of the other possible proportion.
3. What is the percentage of bootstrapped proportion smaller than 70%?
4. As a consequence of your last answer, what the start-up

should do?

Exercise 6

A data entry position need to be filed at a tech company. After doing the interview process, human resource selected the two ideal candidate to do a final test where they had to complete a sample day of work (they take data entry really seriously in this company). The first candidate did his work with an average time of 5 minutes for each form and a variance of 35 minutes while the second did it with a mean of 6.5 minutes and a variance of 25. Assuming that the time needed by an employer to fill in a form follow a normal distribution:

1. Simulate the work of both candidates by generating 200 points of data from both distributions.
2. Use bootstrapping to compute the 95% confidence interval for both means.
3. Can we conclude that a candidate is faster than the other?

Exercise 7

A business wants to launch a product in a new market. Their study show that to be viable a market must be composed of at least 60% of potential consumer making more than 35 000$. If the last census show that the salary of this population follow an exponential distribution with a mean of 60000 and that the rate of an exponential distribution is equal to 1/mean, should this business launch their product in this market?

Exercise 8

A batch of 1000 ohms resistance are scheduled to be solder to two other 200 ohms resistance to create a serial circuit of 1400 ohms. But no manufacturing process is perfect and no resistance has perfectly the value it supposed to have. Suppose that the first resistance is made following a normal process that makes batch of resistance with a mean of 998 ohms and a standard deviation of 5.2 ohms, while the two other come from another process who produce batch of resistance with a

mean of 202 and a variance of 2.25. What is the percentage of circuits will have for resistance a value in the interval [1385,1415]? (Note: you can use bootstrap to solve this problem or you can use the fact that the sum of two normal distributions is equal to another normal distribution whose mean is equal to the sum of their two means. The variance the new distribution is calculated the same way. You can learn more here)

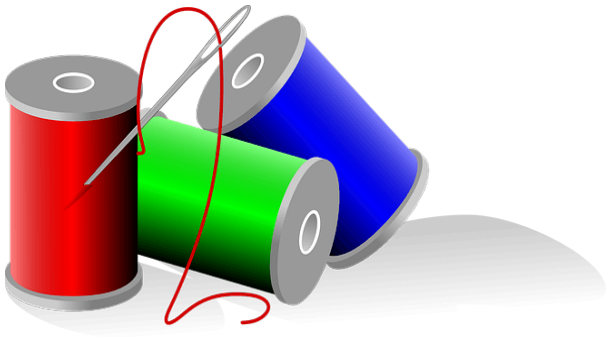
Exercise 9
A probiotic supplement company claim that three kinds of bacteria are present in equal part in each of their pill. An independent laboratory is hired to test if this company respects this claim. After taking a small sample of five pills, they get the following dataset where the numbers are in millions.

In this dataset, the rows represent pills used in the sample and each column represents a different kind of bacteria. For each kind of bacteria:

1. Compute the mean.
2. Compute the variance.
3. Compute the quartile.
4. Compute the range which is define by the maximum value minus the minimum value.

Exercise 10
A shipping company estimate that the delivery delays of his shipment, in hours, follow a student distribution with a parameter of 6. What is the proportion of delivery that are between 1 hours late and 3 hours late?

Hacking Strings with stringi



In the last set of exercises, we worked on the basic concepts of string manipulation with stringr. In this one we will go further into hacking strings universe and learn how to use stringi package.Note that stringi acts as a backend of stringr but have many more useful string manipulation functions compared to stringr and one should really know stringi for text manipulation .

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 two strings
c1 <- "a quick brown fox jumps over a lazy dog"
c2 <- "a quick brown fox jump over a lazy dog"
Now stringi comes with many functions and wrappers around functions to check if two string are equivalent. Check if they are equivalent with
stri_compare, %s<=% and try to reason about the answers.

Learn more about Text analysis in the online course Text Analytics/Text Mining Using R. In this course you will learn how create, analyse and finally visualize your text based data source. Having all the steps easily outlined will be a great reference source for future work.

Exercise 2

How would you find no of words in c1 and c2 . Its pretty easy with stringi.Find it out .

Exercise 3

Similarly How would you find all words in c1 and c2 . Again its pretty straight forward with stringi.Find it out .

Exercise 4

Lets say you have a vector which contains famous mathematicians

genius <- c(Godel,Hilbert,Cantor,Gauss, Godel, Fermet,Gauss)

Find the duplications .

Exercise 5

Find the number of characters in genius vector by stri function.

Exercise 6

Its important to keep the character's of a set of strings in same encoding .Suppose you have a vector

Genius1 <- c("Godel","Hilbert","Cantor","Gauss", "Gödel", "Fermet","Gauss")

Now basically Godel and Gödel are same person but the encoding of the characters are different . but if you try to compare them in a naive way they will act as different .So for the sake of consistency,we should really translate it to similar encoding .Find it how .

Hint – use "Latin-ASCII" transliterator in stri_trans* like function.

Exercise 7

How do we collapse the LETTER vector in R such that it looks like this

"A-B_C-D_E-F_G-H_I-J_K-L_M-N_O-P_Q-R_S-T_U-V_W-X_Y-Z_"

Exercise 8
Suppose you have a string of words like c1 that we have created earlier . You might want to know the starting and end index of the first word, last word.which is obvious for start index of first word and last word but not so obvious for the end index of first word and start index of last word. How would you find this .

Exercise 9
Suppose I have a string
pun <- "A statistician can have his head in an oven and his feet in ice, and he will say that on the average he feels fine"
Suppose I want to replace statistician and average with mathematician and median in the string pun .How can I achieve that .
Hint -use a stri_replace* method.

Exercise 10
My string x is like
x <- "I AM SAM. I AM SAM. SAM I AM"
replace last SAM with ADAM.

Data wrangling : Transforming (2/3)

Data wrangling is a task of great importance in data analysis. Data wrangling, is the process of importing, cleaning and transforming raw data into

actionable information for analysis. It is a time-consuming process which is estimated to take about 60-80% of analyst's time. In this series we will go through this process. It will be a brief series with goal to craft the reader's skills on the data wrangling task. This is the third part of the series and it aims to cover the transforming of data used.This can include filtering, summarizing, and ordering your data by different means. This also includes combining various data sets, creating new variables, and many other manipulation tasks. At this post, we will go through a few more advanced transformation tasks on mtcars data set.

Before proceeding, it might be helpful to look over the help pages for the group_by, ungrpoup, summary, summarise, arrange, mutate, cumsum.

Moreover please load the following libraries.
install.packages("dplyr")
library(dplyr)

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 new object named *cars_cyl* and assign to it the *mtcars* data frame grouped by the variable *cyl*
Hint: be careful about the data type of the variable, in order to be used for grouping it has to be a factor.

Exercise 2

Remove the grouping from the object *cars_cyl*

Exercise 3

Print out the summary statistics of the *mtcars* data frame

using the summary function and pipeline symbols %>%.

Learn more about Data Pre-Processing in the online course R Data Pre-Processing & Data Management – Shape your Data!. In this course you will learn how to:

- Work with popular libraries such as dplyr
- Learn about methods such as pipelines
- And much more

Exercise 4

Make a more descriptive summary statistics output containing the 4 quantiles, the mean, the standard deviation and the count.

Exercise 5

Print out the average *hp* for every *cyl* category

Exercise 6

Print out the *mtcars* data frame sorted by *hp* (ascending oder)

Exercise 7

Print out the *mtcars* data frame sorted by *hp* (descending oder)

Exercise 8

Create a new object named *cars_per* containing the *mtcars* data frame along with a new variable called *performance* and calculated as performance = hp/mpg

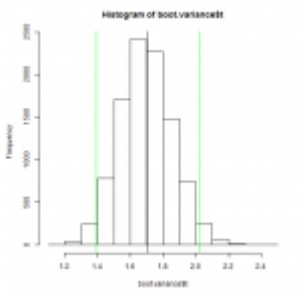
Exercise 9

Print out the cars_per data frame, sorted by *performance* in descending order and create a new variable called *rank* indicating the rank of the cars in terms of performance.

Exercise 10

To wrap everything up, we will use the *iris* data set. Print out the mean of every variable for every *Species* and create two new variables called *Sepal.Density* and *Petal.Density* being calculated as Sepal.Density = Sepal.Length Sepal.Width and Petal.Density = Sepal.Length Petal.Width respectively.

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.

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.