

STA312 Tutorial 8

Matthew Scicluna

Tuesday, March 17, 2015

What we are going to cover

A bit more information about this

Today we are going to talk about

- ▶ knitr
- ▶ Assignment 4

knitr

Lets talk more about this

- ▶ A hip new way of using \LaTeX
- ▶ Lets you make reproducible research
- ▶ What does that mean???

This is some reproducible research

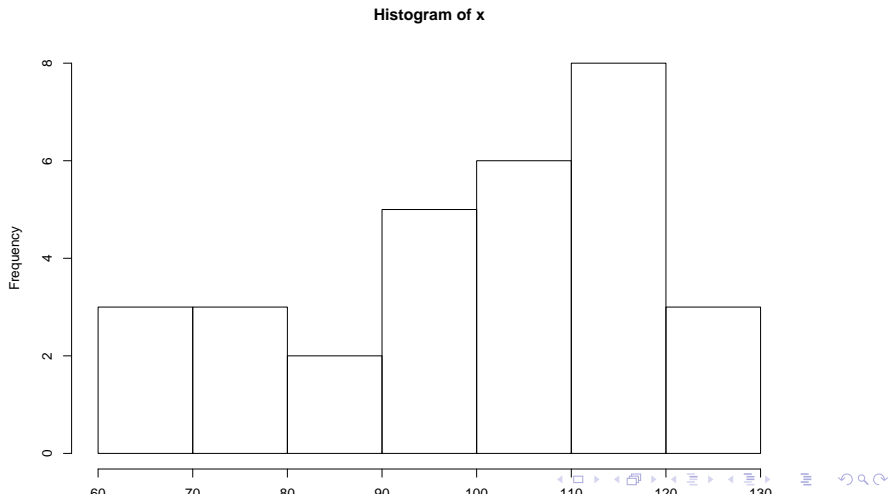
Lets take a random sample of 30 random variables $X \sim N(100, 15)$

```
##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,]  83.20106 115.6169  62.54273  66.40733 117.0235 103.83
##           [,8]      [,9]     [,10]     [,11]     [,12]     [,13]
## [1,]  98.13381  99.90343 115.9583 113.4668 118.3535 106.60
##           [,15]     [,16]     [,17]     [,18]     [,19]     [,20]
## [1,] 103.2001 116.8811 121.7256 103.6967  89.75755  76.471
##           [,22]     [,23]     [,24]     [,25]     [,26]     [,27]
## [1,]  70.98312 123.0206  68.54824 124.9559 104.9909  93.350
##           [,29]     [,30]
## [1,] 117.4804 101.6965
```

This is some reproducible research

```
## [1] "Lets see the sample mean, variance and a plot!"
```

```
## [1] 99.83714 336.34224
```



On reproducible research

- ▶ Each time I run this script the code generates a new random sample.
- ▶ I get different results each time, which can be a good thing.
- ▶ Assignment 4 will be looking at some data which is rapidly changing, so this approach will be essential to completing it.

Knitr References

- ▶ A great tutorial on Knitr can be found at http://kbroman.org/knitr_knutshell/
- ▶ A tutorial on Markdown can be found here http://kbroman.org/knitr_knutshell/pages/markdown.html
- ▶ Note that in your .Rmd file you can use ordinary syntax from \LaTeX
- ▶ You can use the relatively simpler markdown syntax as well!

On Assignment 4

- ▶ For this assignment you are going to analyze twitter data
- ▶ This may or may not involve cyberstalking celebrities
- ▶ You will use knitr to present your findings in an interactive program

On Assignment 4

- ▶ install twitteR package from CRAN
- ▶ Reference manual here: <http://cran.r-project.org/web/packages/twitteR/twitteR.pdf>
- ▶ Dont forget to read the source material before installing a package!
- ▶ You dont need to understand how to use twitteR well as most of the code will be provided for you

Assignment 4 - Mining tweets using twitterR

- ▶ When the starter code is executing it will harness tweets with some particular keyword in them. It will produce a different set of tweets each time!
- ▶ Usually you would have to make a twitter account, register as a developer and get the access code and token, but. . .
- ▶ I made a twitter account and set this up for you. The authentication procedure has been built into the code.

Assignment 4 - What you will deliver

- ▶ You will create a Markdown file that gathers tweets from two or more sources.
- ▶ These can be tweets from separate users or tweets with separate keywords (I do both in the starter code).
- ▶ *e.g. of 1 tweets from Obama v.s. tweets from Rand Paul*
- ▶ *e.g. of 2 tweets containing 'Democrat' v.s. tweets containing 'Republican'*

Assignment 4 - What you will deliver

- ▶ you will compare differences in the frequencies of the words using a statistical test.
- ▶ you don't have to do anything more advanced than a t-test.
- ▶ you are welcome to try to do LSA or anything else you have learned in this course (or on your own)
- ▶ Creativity is always rewarded!

Assignment 4 Grading Scheme

- 3 Marks You will collect tweets from 2 or more sources and display the most popular words from both sources.
- 2 marks Explain why you think these 2 sources will be distinguishable on the level of word frequencies of tweets.
- 3 marks Plot these tweets in feature space using tsne as I did in the demo and come up with some statistical tests to see if there is a significant difference between the word frequencies of your sources. Display as much as you can visually.
- 2 marks 2 marks will be awarded based on quality and thoughtfulness of discussion included in report.

Assignment 4 - Final Remarks

- ▶ You will email me your assignment as a markdown file and I will run it myself.
- ▶ My email is `matthew.scicluna@mail.utoronto.ca`
- ▶ Notice that because of this you cannot directly comment on your data, but are welcome to mention previous results you found and any expectations you may have.
- ▶ I will provide feedback via email. There is **NO** hard copy submission.
- ▶ **You will lose all 10 marks if I cannot run your code**