STA302 - Lecture 1

Cedric Beaulac

May 7, 2019

Introduction

Syllabus

- Hey Cedric! You should look at the syllabus.

Concept Pre-requisite

- I will assume you have a good understanding on some fundementals of Distribution theory.

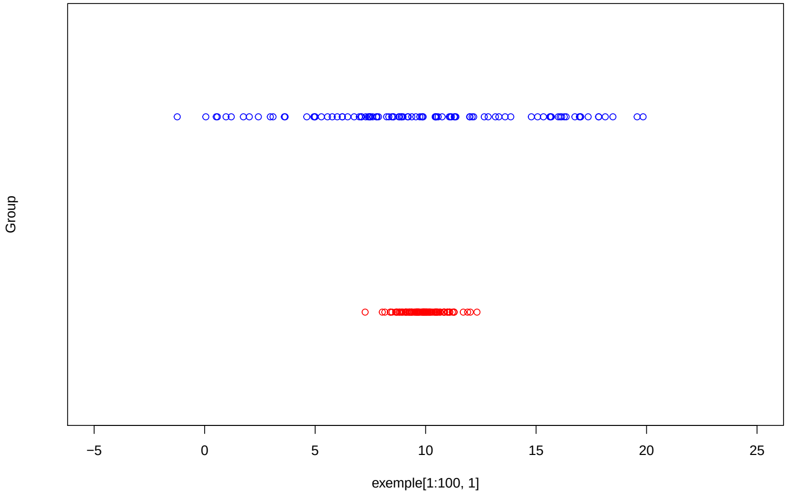
Statistical analysis

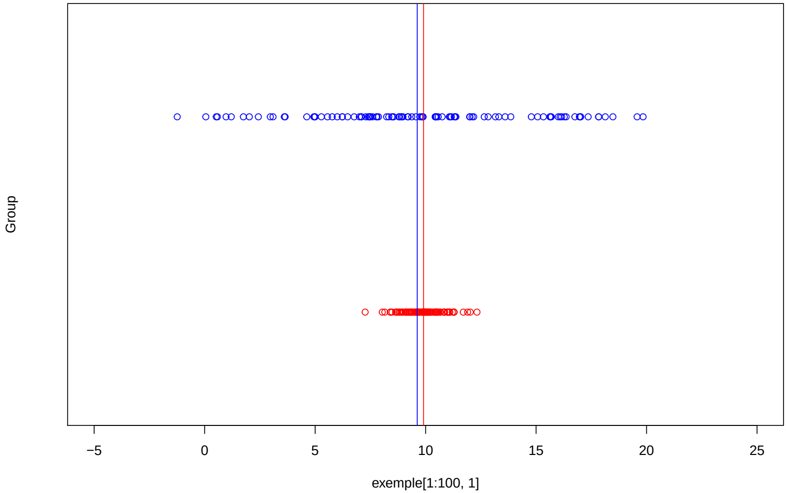
- Data analysis that rellies on Probability theory to account for the variability of the data.

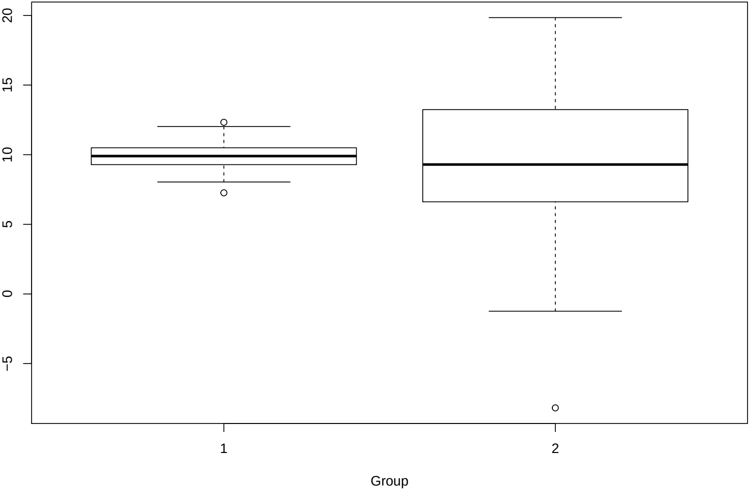
Variability

Variability

- Typically we only have a sample of the true population.
- High variability leads to uncertainty.

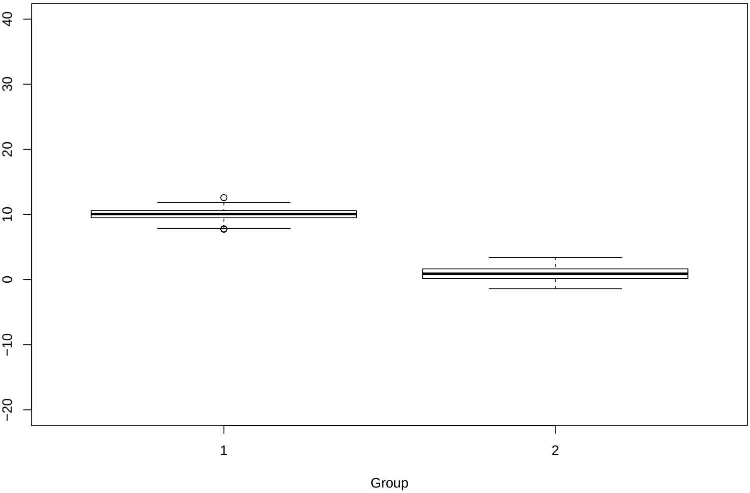

Group
−5
0
5
10
15
20
25
exemple[1:100, 1]

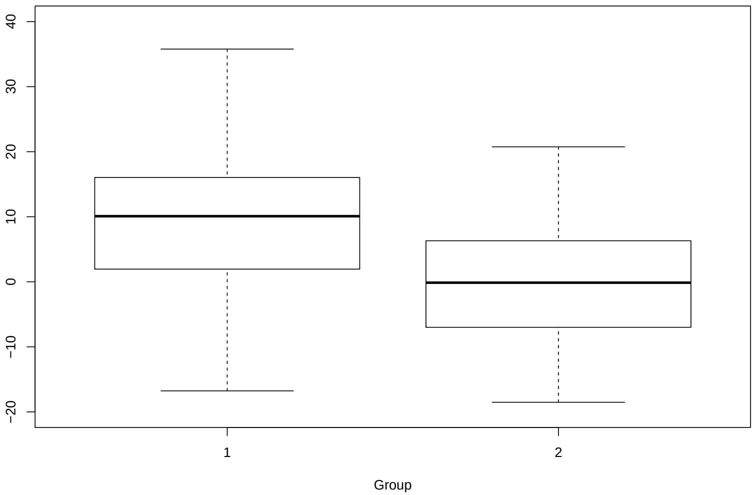

Group
−5
0
5
10
15
20
25
exemple[1:100, 1]


20
15
10
5
0
−5
1
2
Group

Supervised Learning

- Usually we are given a specific statistical task.
- We are interested in the effect of a predictor x (explanatory, independent variable, input) on an response variable y (dependent variable, ouput).
- Both of these can take multiple forms.
- We use linear models when we want to undertsand the relationship between a continuous predictor and a continuous response.
- Let us introduce basic concept using a model where the response is continuous but the predictor is a binary variable (two different groups).
- Our question : Are the two groups different ?


40
30
20
10
0
−10
−20
1
2
Group


40
30
20
10
0
−10
−20
1
2
Group

The need for rigorous tests

- Take two samples of continuous random variables with the same expectation.
- Sample means will be different with probability 1. (IMPORTANT)
- Point estimation is not enough, we need something that account for the variability in the data.

Roll the dices

```r
x1 = sample(seq(1:6),3,replace=TRUE)
x2 = sample(seq(1:6),3,replace=TRUE)
x1
```

```
## [1] 2 6 5
```

```r
x2
```

```
## [1] 1 3 1
```

```r
mean(x1)
```

```
## [1] 4.333333
```

```r
mean(x2)
```

```
## [1] 1.666667
```

```r
x1 = sample(seq(1:6),3,replace=TRUE)
x2 = sample(seq(1:6),3,replace=TRUE)
x1
```

```
## [1] 2 5 4
```

```r
x2
```

```
## [1] 3 5 4
```

```r
mean(x1)
```

```
## [1] 3.666667
```

```r
mean(x2)
```

```
## [1] 4
```

```r
x1 = sample(seq(1:6),3,replace=TRUE)
x2 = sample(seq(1:6),3,replace=TRUE)
x1
```

```
## [1] 6 1 6
```

```r
x2
```

```
## [1] 4 2 5
```

```r
mean(x1)
```

```
## [1] 4.333333
```

```r
mean(x2)
```

```
## [1] 3.666667
```

The need for rigorous tests

- We expect any two samples to have different means.
- How can we confidently claim the different is meaningfull.
- How much of a difference is enough ?
- Let's define statiscal significance with a simple permutation test.

Permutation test

Permutation test

- Insert random premise
- Here is the data :
  - Group A : 48, 56, 58
  - Group B : 44, 46, 51
  - $\bar{y}_A = 54$ and $\bar{y}_B = 47$
- Given our fascinating premise, we want to know if Group A has larger values than Group B.

Permutation test

- Given the previous slides, we know we have to be carefull.
- If groups have no effect, the means would still be different.
- If groups have no effect, What are the reasonnable differences we could observe ?
- If groups have no effect, we could have observed any permutations of the data.

Permutation test

- We observed :
  - Group A : 48, 56, 58
  - Group B : 44, 46, 51
- but if groups have no effect we might as well have observed :
  - Group A : 48, 46, 58
  - Group B : 44, 56, 51
- Or :
  - Group A : 44, 56, 58
  - Group B : 48, 46, 51

Permutation test

- There exist $\binom{6}{3} = 20$ ways to divided the observations into 2 groups.
- If the groups have no effect, all of them are equally likely.

Permutation test

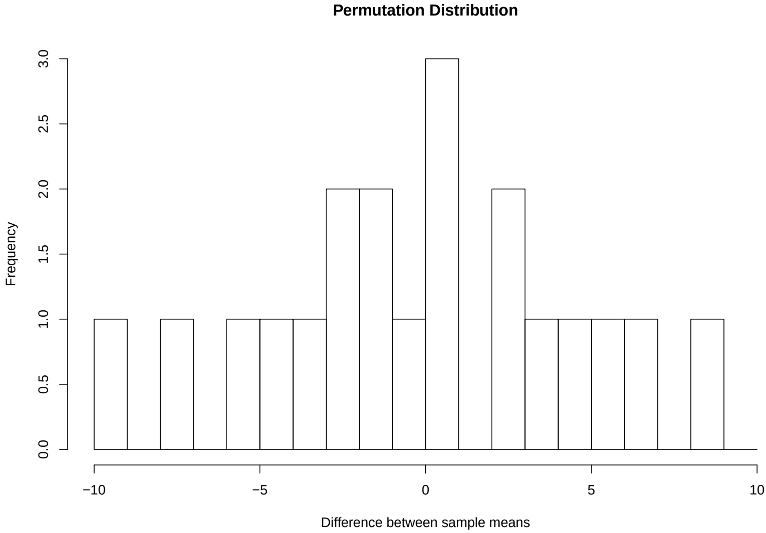
Group A			Group B		
48	56	58	44	46	51
48	56	44	58	46	51
48	56	46	44	58	51
48	56	51	44	46	58
48	44	58	56	46	51

You get the idea.

Permutation test

- We observe :
    - Group A : 48, 56, 58
    - Group B : 44, 46, 51
    - $\bar{y}_A = 54$ and $\bar{y}_B = 47$
- The observed difference is : $\bar{y}_A - \bar{y}_B = 7$

```
##  A  B
## 54 47
```


Permutation Distribution
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−10
−5
0
5
10
Difference between sample means

Permutation test

- If groups have no effect, we could have observed a totel of 20 differences :

```
 [1] -9.00 -7.00 -5.67 -4.33 -3.67 -2.33 -2.33 -1.00 -1.00 -0.33  0.33
[12]  1.00  1.00  2.33  2.33  3.67  4.33  5.67  7.00  9.00
```

Permutation test

- We have a difference of 7
- Out of the 20 equally likely difference if there is no group effect, only 7 and 9 are ≥ 7.
- Out of the 20 equally likely difference if there is no group effect, only 10% of them are equally large as ours.

P-value

P-value

- We have a p-value $= 0.1$
- Under the assumption that groups have no effect, the observed difference is larger or equal than 10% of all the possible differences.
- Under the assumption that groups have no effect, the probability of sampling a data set with a difference between groups as extreme as what we observed is 10%.

Statistical significance

- Perhaps we think 10% is plausible. Perhaps the groups are the same ?
- Perhaps we think something that happens 10% of the time is exceptionnal. Then it implies the groups must not be the same.
- If we think this is exceptional, we fix our significance level at 10%.
- Then we claim that the difference is statistically significant.

Statistical significance

- We say a difference is statistically significance if it's less probable than our pre-determined significance level.
- We say the groups have a signficant effect if it causes the variable of interest to be significantly different.

Permutation test

- Involves simple probability theory.
- Distribution-free.
- Listing all the permutation for large data set is almost impossible.

Practice problems

- Nothing today
- You can read about p-values

External ressources (references I used)

- Wikipedia
- Craig Burkett's STA 2101/442 introduction slides
- Nathan Taback's STA305 slides