Statistic with Java

程式設計一期末報告
統計教學暨研究

董宸賓

2024-01-04

本次報告相關的java class

都可以在以下網站透過動態互動的方式來了解

當然也包含所有完整程式碼的githhub

目錄

1.class tour

2.架構分析

3.應用篇1-教學用途

4.應用篇2-輔助教育研究與分析

5.Design pattern

Class tour

1.DescriptiveStatistics

計算每個數據的基礎統計量

CONSTRUCTOR Method Sample

code

```
DescriptiveStatistics
(double[] data, String name)
```

description

data: 輸入的數據

name: 輸入的數據名稱

2.Anova(繼承DescriptiveStatistics)

分析多組數據集之間的均值是否存在顯著差異

CONSTRUCTOR Method Sample

code

```
public Anova(double[][] groups, String name) {
        super(flatten(groups), name);
        this.groups = groups;
    }
```

description

1.為了處理多項數據之比較，接受一個二維數組作為參數

2.通過flatten函數將二維數組轉換為一維數組，讓其也可以在DescriptiveStatistics中使用

3.LinearRegression(一樣繼承)

計算線性回歸模型的斜率和截距，並使用模型進行預測。

CONSTRUCTOR Method Sample

code

```
public LinearRegression(double[] xData, double[] yData, String name) {
        super(yData, name); // 使用因變量初始化 DescriptiveStatistics
        this.xData = xData;
        this.yData = yData;
    }
```

description

1.以自變量和因變量的數據，以及自變量數據集的名稱來構造物件

2.通常以一個DescriptiveStatistics物件作為因變量，以及一個double[]作為自變量

架構分析

基礎架構

1.宣告Scanner與存放DescriptiveStatistics物件的ArrayList

```
Scanner sc = new Scanner(System.in);
List<DescriptiveStatistics> statsList = new ArrayList<>();
```

2.選擇輸入方式，並將每份資料以物件形式加入List裡

```
System.out.println("請選擇數據來源 ? 1:CSV 2:手動輸入");
...
switch(choice1){
  case 1:
    ...//讀取csv檔案
     break;
   case 2:
     ...//手動輸入
     break;
}
```

基礎架構

3.對每份數據作基礎之分析，檢視數據

```
for (DescriptiveStatistics stats : statsList) {
  System.out.println(stats.information());
}
```

4.Sample output1-from csv

```
請選擇數據來源 ? 1:CSV 2:手動輸入
1
Enter the CSV file path:
data.csv
數據名稱: "StudentID"
平均值: 54732.87
中位數: 56340.0
標準偏差: 27148.35057628915
樣本大小: 100
母體方差: 7.370329390130996E8
母體標準偏差: 27148.35057628915
```

```
數據名稱: "LibraryHours"
平均值: 4.927089057526639
中位數: 4.619102264084435
標準偏差: 1.6412256355060801
樣本大小: 100
母體方差: 2.6936215866423368
母體標準偏差: 1.6412256355060801

數據名稱: "Grade"
平均值: 2.44
中位數: 2.0
標準偏差: 1.1429785649783635
樣本大小: 100
母體方差: 1.3063999999999991
母體標準偏差: 1.1429785649783635
```

基礎架構

3.對每份數據作基礎之分析，檢視數據

```
for (DescriptiveStatistics stats : statsList) {
  System.out.println(stats.information());
}
```

4.Sample output1- 手動

```
請選擇數據來源 ? 1:CSV 2:手動輸入
2
數據欄位名:
程式設計一段考成績
請問有幾個數據:
3
輸入數據:
60 78 35
Do you want to enter another set of da
Y
數據欄位名:
程式設計一段考讀書時間
請問有幾個數據:
3
輸入數據: 10 3 62
```

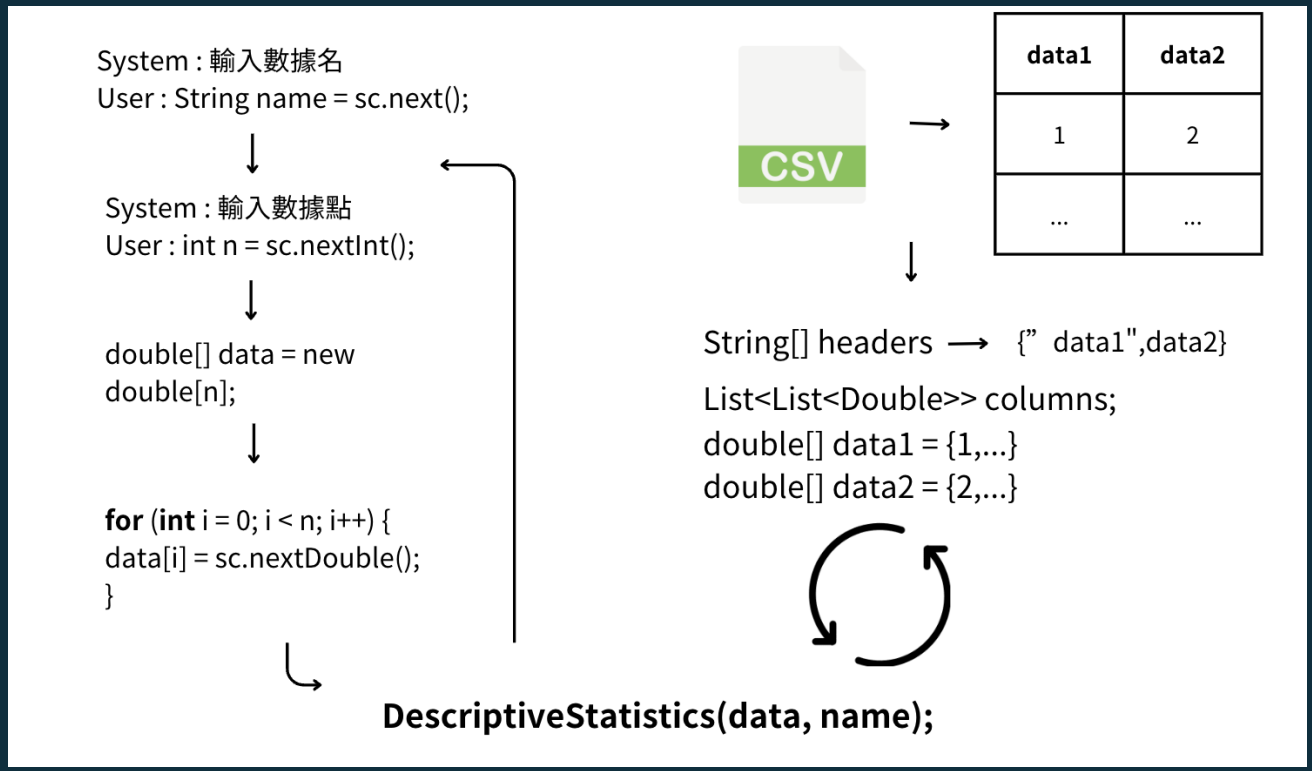
```
Do you want to enter another set of da
N
數據名稱: 程式設計一段考成績
平均值: 57.666666666666664
中位數: 60.0
標準偏差: 17.632041540584257
樣本大小: 3
母體方差: 310.8888888888889
母體標準偏差: 17.632041540584257

數據名稱: 程式設計一段考讀書時間
平均值: 25.0
中位數: 10.0
標準偏差: 26.318561257535844
樣本大小: 3
母體方差: 692.6666666666666
母體標準偏差: 26.318561257535844
```

資料處理

流程圖

from csv

手動輸入



進階分析

應用篇1-教學用途

1.數據與流程

Sample output

應用篇2-輔助教育研究與分析

課程長度與成績之分析

本數據是某大學在期末時，拿出兩門相同課程、相同老師、相同教材但課程長度不一樣之學生成績數據。

Show 5 entries Search:

	score	course_length
1	55.95044554414471	16
2	66.88829846859782	16
3	88.24004321299603	16
4	81.15636849302675	16
5	85.91668955535361	16

Showing 1 to 5 of 100 entries

Previous 1 2 3 4 5 … 20 Next

Anova

```
已選擇 score
已選擇 course_length
ANOVA 分析 - ANOVA Test
總體平方和 (SST): 204458.51488688402
組間平方和 (SSB): 190421.12282211994
組內平方和 (SSW): 14037.392064764077
F 值: 2685.9250026520804
```

LinearRegression

```
已選擇 score作為因變量
已選擇 course_length作為自變量
線性回歸模型 - score vs. course_length
斜率 (beta): 0.03159793293103698
截距 (alpha): 14.997051348946353
輸入預測值：98
預測結果：17.227
```

ANOVA 分析顯示 course_length 對 score 有顯著的影響

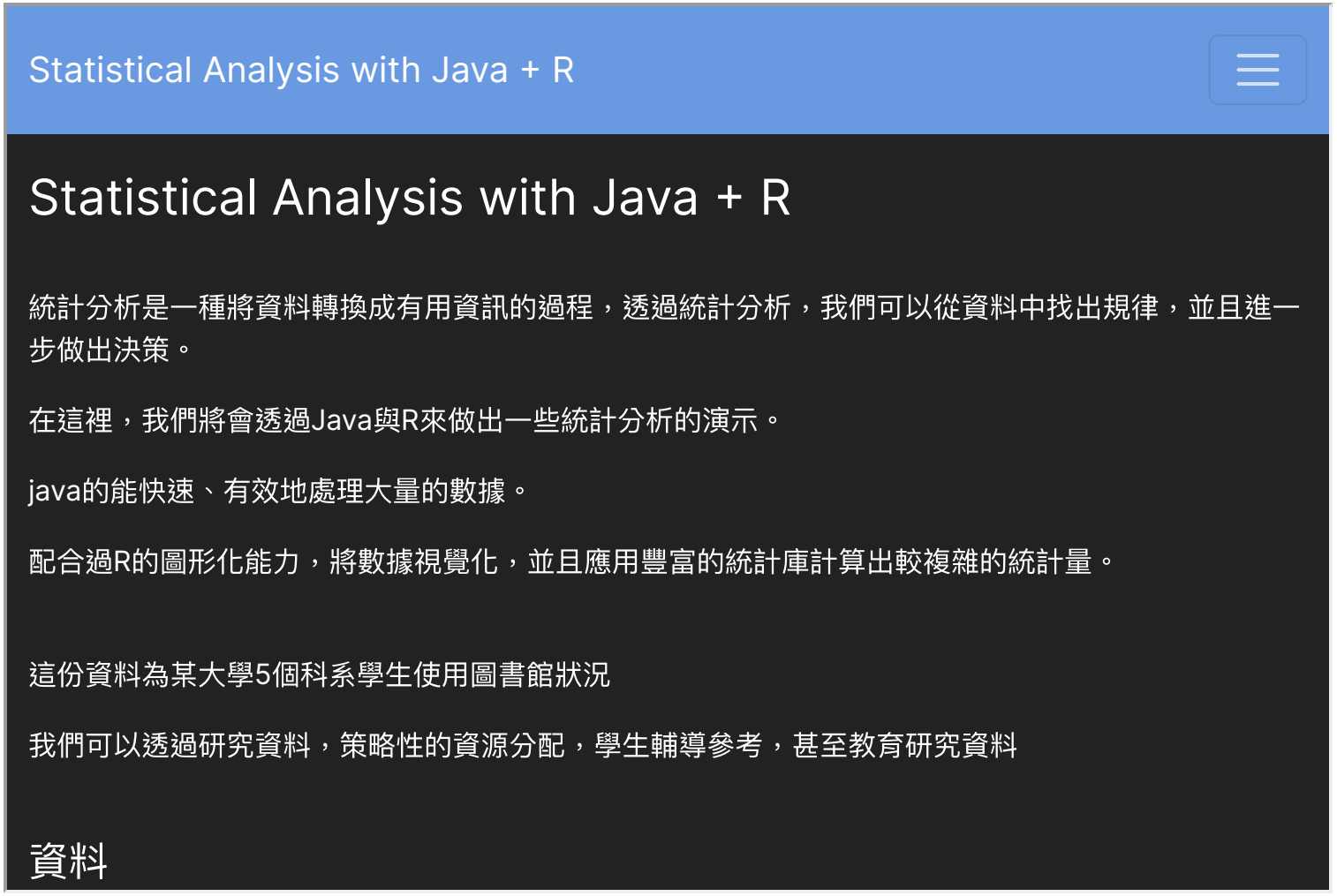
意味著不同的課程長度對學生成績有顯著的影響。

線性回歸模型提供了一種量化這種影響的方式

表明隨著課程長度的增加，成績也有所提高，儘管幅度不大。

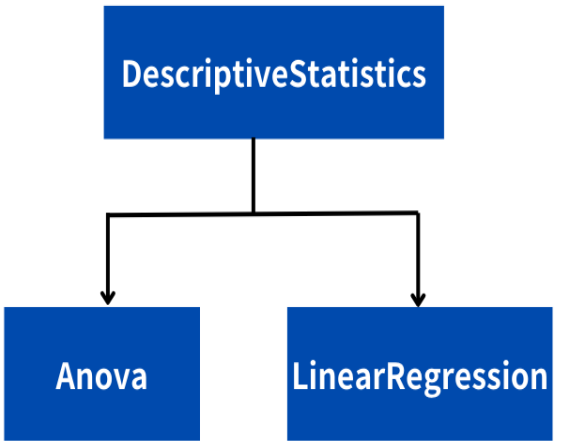
藉由以上的資料分析，能為教育研究提供一些參考，甚至可以...

extend


Statistical Analysis with Java + R
Statistical Analysis with Java + R
統計分析是一種將資料轉換成有用資訊的過程，透過統計分析，我們可以從資料中找出規律，並且進一步做出決策。
在這裡，我們將會透過Java與R來做出一些統計分析的演示。
java的能快速、有效地處理大量的數據。
配合過R的圖形化能力，將數據視覺化，並且應用豐富的統計庫計算出較複雜的統計量。
這份資料為某大學5個科系學生使用圖書館狀況
我們可以透過研究資料，策略性的資源分配，學生輔導參考，甚至教育研究資料
資料

Design pattern

1繼承 (Inheritance)



共用
mean() method

ds.mean()
anova.mean()
lr.mean()

2.多型 (Polymorphism)

Override Summary()

DescriptiveStatistics

數據名稱: score
平均值: 79.21241
中位數: 79.37917
標準偏差: 11.795
樣本大小: 100
母體方差: 139.123
母體標準偏差: 11.79

Anova

線性回歸模型
- score vs. course_length
斜率 (beta): 0.031597
截距 (alpha): 14.9970513
輸入預測值：98
預測結果：17.227

More info

本投影片是使用R package **xaringan**.

查克拉來自**remark.js**, **knitr**, and **R Markdown**.

點擊進入**網站**，**Github**，**本簡報**

謝謝聆聽

THE END