

中級ミクロデータサイエンス期末課題

Problem Set 3

横浜国立大学経済学部 3 年
学籍番号 2125178
廣江友哉

2024 年 2 月 5 日

1 データセットのシミュレーション

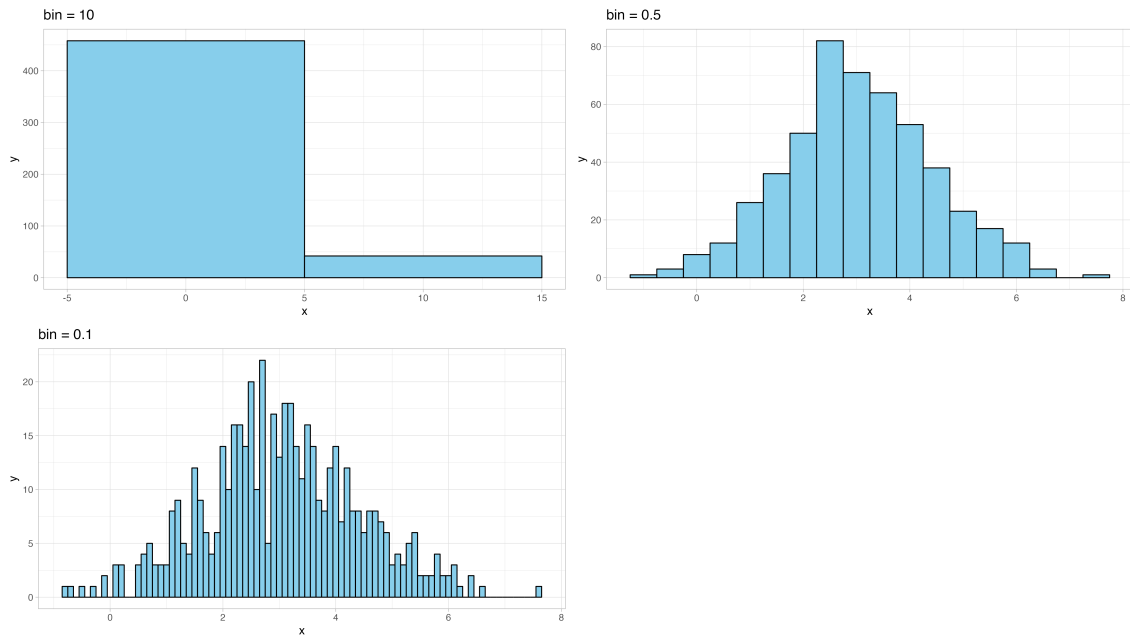
ソースコードは、<https://github.com/tomoyahiroe/replication-project> にある。リポジット
リページ下部の README.md ファイルを参照いただきたい。各 Problem Set についての説明を記述
している。

データセットは以下のように生成した。

```
1 main <- function() {  
2  
3   # 運動神経 X のデータを生成  
4   set.seed(123)  
5   X = rnorm(500, mean = 3, sd = sqrt(2))  
6  
7   # Y (試合で得点できるか) X + ε が4以上なら 1を返す  
8   set.seed(456)  
9   Y = ifelse(X + rnorm(500, mean = 0, sd = 1) >= 4, 1, 0)  
10  
11  # Speed, Shoot, Height, Age, Teamwork のデータを生成  
12  set.seed(789)  
13  Speed = X + rnorm(500, mean = 0, sd = 1)  
14  
15  set.seed(101112)  
16  Shoot = X + rnorm(500, mean = 0, sd = 1)  
17  
18  set.seed(131415)  
19  Height = X + rnorm(500, mean = 0, sd = 1)  
20  
21  set.seed(161718)  
22  Age = X + rnorm(500, mean = 0, sd = 1)  
23  
24  set.seed(192021)  
25  Teamwork = X + rnorm(500, mean = 0, sd = 1)  
26  
27  # 収入Z = Teamwork + 0.5 * Shoot + 0.5 * Speed + η  
28  set.seed(222324)  
29  Z = Teamwork + 0.5 * Shoot + 0.5 * Speed + rnorm(500, mean = 0, sd = 1)  
30  
31  # データフレームを作成  
32  simulate_dataset = data.frame(X, Y, Speed, Shoot, Height, Age, Teamwork, Z)  
33  
34  save(simulate_dataset, file = basics$get_absolute_path("src/analyze/output/data/simulate_dataset.rda"))  
35  
36 }  
37  
38 box::use(functions/basics)  
39 box::use(functions/plot_modules)  
40 box::use(functions/df_modules)  
41  
42 main()  
43
```

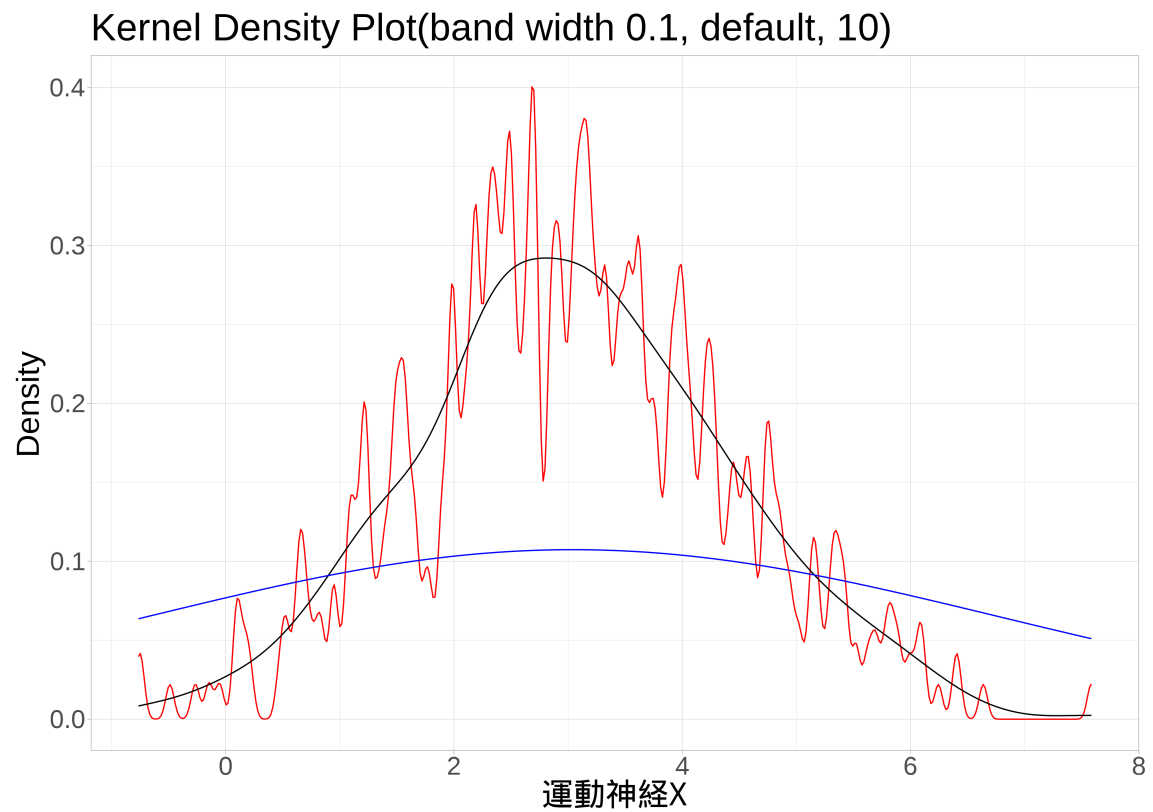
2 分布の推定

2.1 ヒストグラムを用いることの難しさを議論せよ



ヒストグラムはビンの幅によってデータの分布が著しく変わってしまう。運動神経 X のデータでヒストグラムを作成すると、 $\text{bin} = 10$ のときは幅が大きすぎてデータの分布が見えにくく、 $\text{bin} = 0.1$ のときは幅が小さすぎるためにヒストグラムが凸凹としていてデータの傾向がわかりにくい。このように、ヒストグラムで適切にデータの傾向を把握することは難しく、手動でビンの幅を何度も調整する必要がある。

2.2 カーネル密度



上図では、カーネルの幅を赤 0.1, 黒デフォルト, 青 10 として表時している。デフォルトが最も当てはまりが良いように見える。カーネル密度では、カーネルの幅が広がるほど曲線の滑らかさが増し広がりすぎるとデータの傾向が掴めなくなる。この傾向はヒストグラムがビンの幅を大きくするとデータの分布が見えにくくなることと似ている。

2.3 分位回帰