中級ミクロデータサイエンス
Problem set3

学籍番号：2125095

氏名：佐々木 椋

0. コードのURL

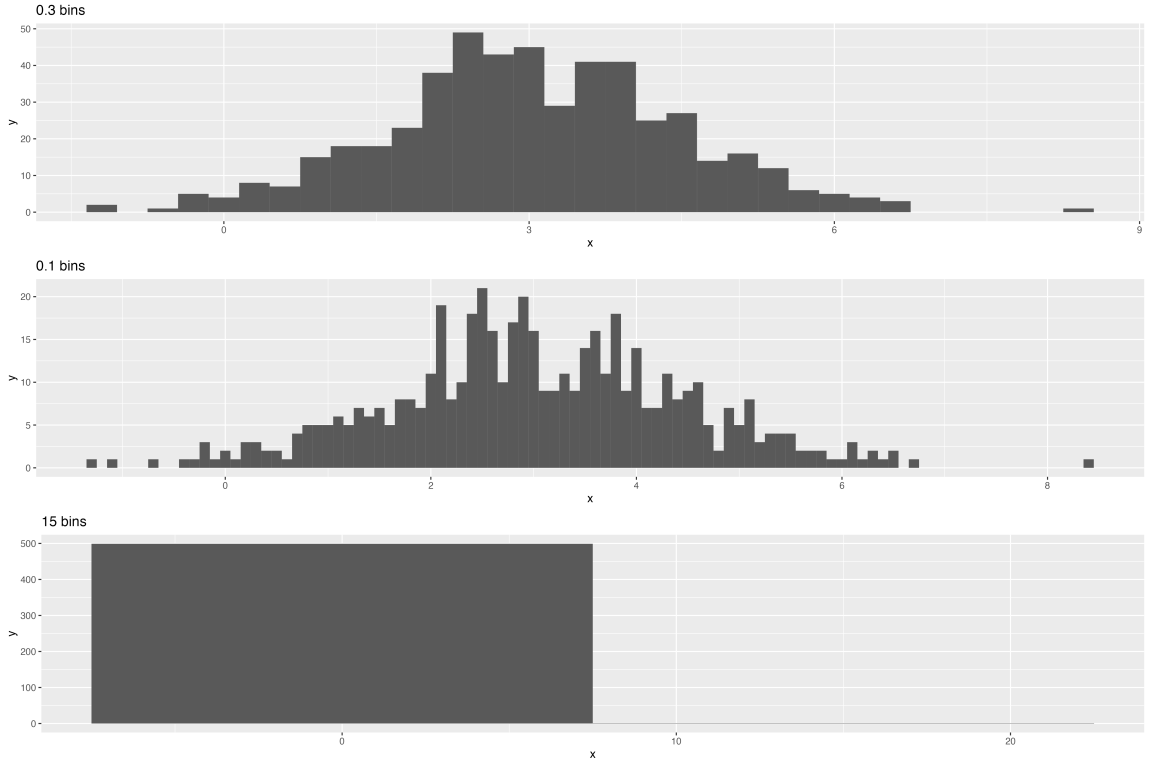
problemset2と同様https://github.com/Ryo-Sasaki-xxx/problemsetにプッシュしてあります。

1. データセットのシミュレーション

problemset3_based/01_analysis/code/generate_data.Rでシミュレーションデータを作成しています。

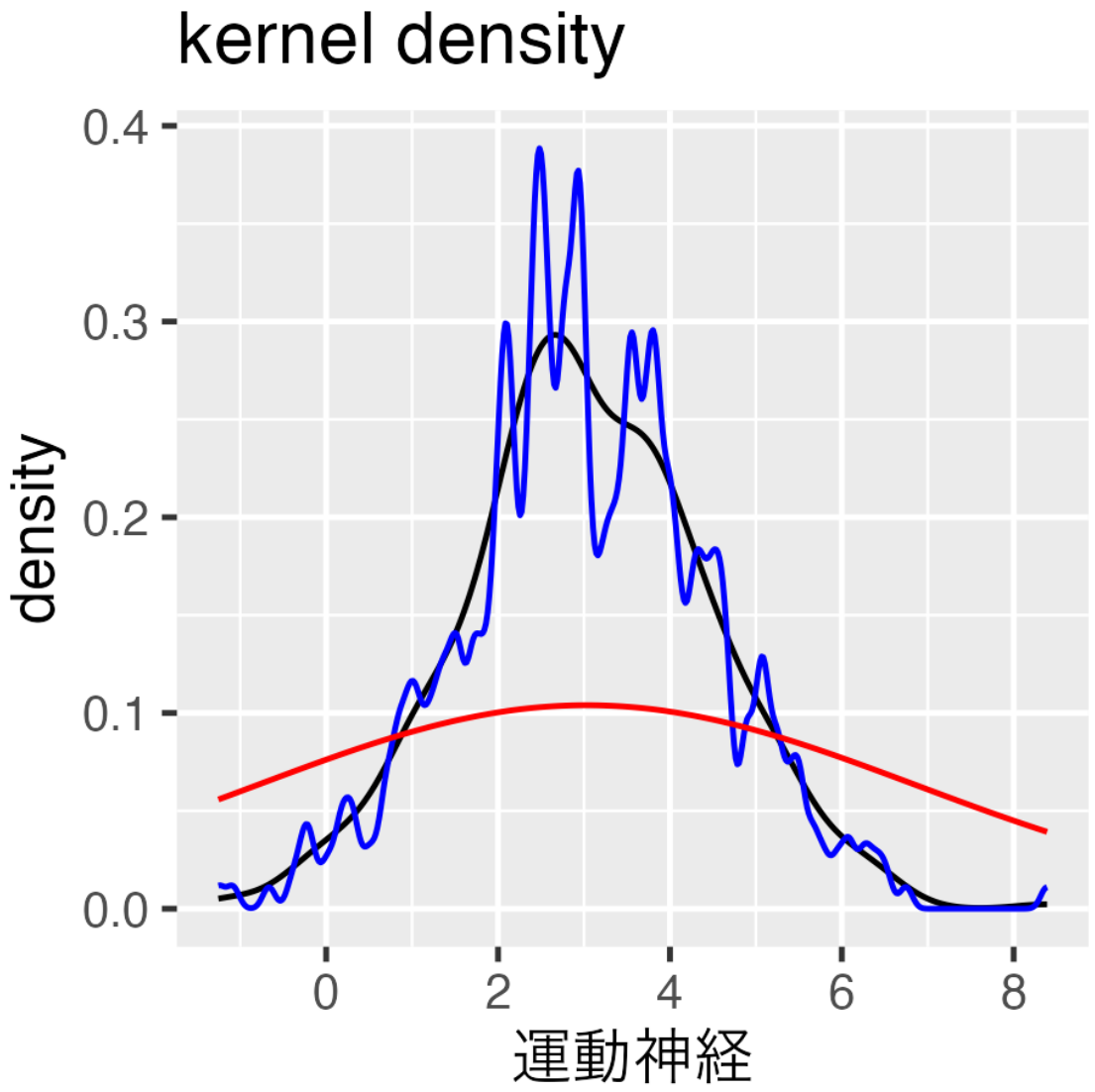
2. 分布の推定

a. ヒストグラム



上図は上から順にデータ区間が0.3、0.1、15のヒストグラムとなっている。明らかにデータ区間が15の場合は、グラフに問題があることがわかる。一方でデータ区間が0.3や0.1のヒストグラムは意見が分かれるだろう。例えば、ある人は0.1の方をより詳細な分布が見れるという理由で評価するかもしれないし、別の人は0.3の方を大まかな傾向だけが表されているという理由で評価するかもしれない。このようにデータの区間をどのように決めるのかについて明確な答えが見当たらないことがある。またデータの区間を意図を持って決めたとしても、そこには客観性が保たれているのかという問題も残る。このようにヒストグラムはデータの区間によって印象が変わるので取り扱いが非常に難しい。

b. カーネル密度



上図はカーネル密度の図であり、赤が幅10、青が幅、黒がデフォルト幅となっている。ヒストグラムと比較すると、幅が広がるにつれて、傾向がより滑らかとなっていることがわかる。

c. 分位回帰

head	beta1	beta0
mean	2.022681	-0.07941144
median	2.027443	-0.14477083
25%	1.982857	-1.01553026
75%	2.057617	0.87930774

最小二乗法によるβ1の推定値は、medianと比較するとあまり差はない、そして第1四分位点のデータと比較するとβ1は少し小さい、また第3四分位点のデータと比較するとβ1は少し大きいという結果になった。