# 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

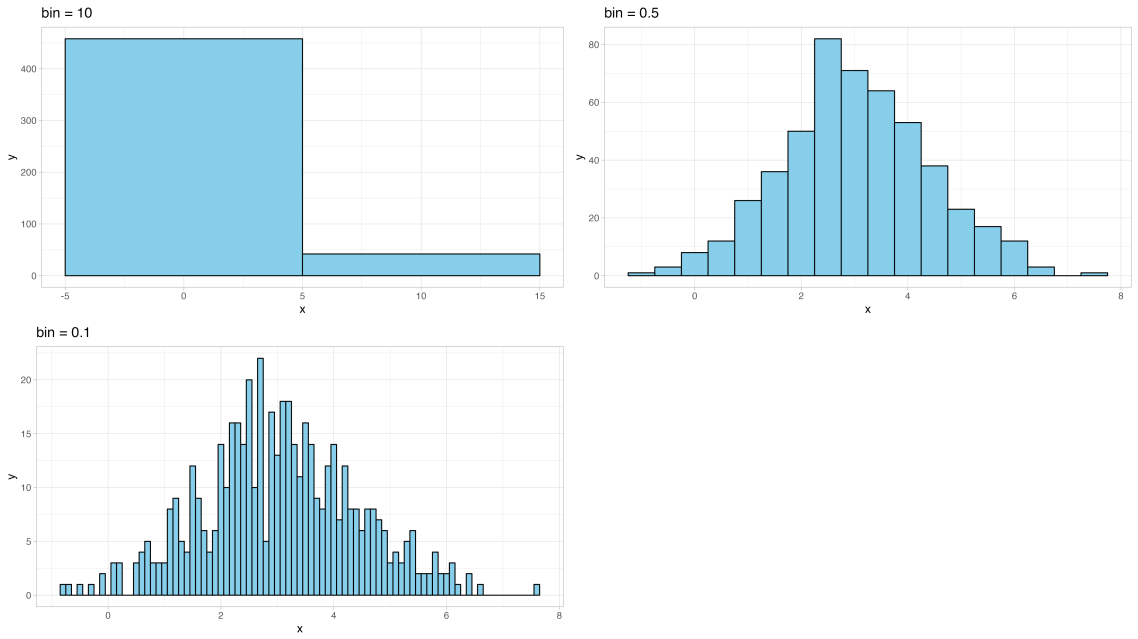
ソースコードは、`https://github.com/tomoyahiroe/replication-project` にある。リポジトリページ下部の README.md ファイルを参照いただきたい。

データセットは以下のように生成した。

```r
main <- function() {

  # 運動神経 X のデータを生成
  set.seed(123)
  X = rnorm(500, mean = 3, sd = sqrt(2))

  # Y(試合で得点できるか) X + ε が4以上なら 1を返す
  set.seed(456)
  Y = ifelse(X + rnorm(500, mean = 0, sd = 1) >= 4, 1, 0)

  # Speed, Shoot, Height, Age, Teamwork のデータを生成
  set.seed(789)
  Speed = X + rnorm(500, mean = 0, sd = 1)

  set.seed(101112)
  Shoot = X + rnorm(500, mean = 0, sd = 1)

  set.seed(131415)
  Height = X + rnorm(500, mean = 0, sd = 1)

  set.seed(161718)
  Age = X + rnorm(500, mean = 0, sd = 1)

  set.seed(192021)
  Teamwork = X + rnorm(500, mean = 0, sd = 1)

  # 収入Z = Teamkwork + 0.5 * Shoot + 0.5 * Speed + η
  set.seed(222324)
  Z = Teamwork + 0.5 * Shoot + 0.5 * Speed + rnorm(500, mean = 0, sd = 1)

  # データフレームを作成
  simulate_dataset = data.frame(X, Y, Speed, Shoot, Height, Age, Teamwork, Z)

  save(simulate_dataset, file = basics$get_absolute_path("src/analyze/output/data/simulate_dataset.rda"))

}

box::use(functions/basics)
box::use(functions/plot_modules)
box::use(functions/df_modules)

main()

```

# 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、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、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