R と RStudio の使い方

芳賀敏郎（2011）医薬品開発のための統計解析　第 1 部　基礎
2 1 組のデータの解析
2.2 データのグラフ表示と外れ値

テキストと利用上の注意

●テキスト

芳賀敏郎（2011）医薬品開発のための統計解析
第１部　基礎　改訂版、サイエンティスト社、p.275
（サイトへアップすることに対して、サイエンティスト社の了解を得ています）

●R による解析事例を紹介

R スクリプトの出力結果を紹介します（tidyverse 系には次期バージョンで対応します）
R スクリプト（文字コードUTF-8に設定）を、このサイトからダウンロードできます
R スクリプトを [Compile Report] することにより、Word または HTML で見ることができます
R と RStudio の設定と基本的な使い方は「R と RStudio の使い方」を参照してください
R の出力結果の見方は、テキストとそれを解説した PDF ファイルを参照してください
グラフ表示は、解析手段として、必要最小限の表現に止めています

●自己責任で利用

上記のことを理解した上で、自己責任により利用してください

第１部　基礎

1. 統計の基礎・・・・・1.1 宝くじの期待値と分散、1.2 サイコロの目の数の期待値と分散
1.3 分散の加法性・中心極限定理・正規分布、1.4 統計的推測、1.5 モデル

2. １組のデータの解析　2.1 データの特徴の記述、**2.2 データのグラフ表示と外れ値**
2.3 対数変換と対数正規分布、2.4 平均に関する推測（母標準偏差 σ 既知）
2.5 分散に関する推測、2.6 平均に関する推測（母標準偏差 σ 未知）

3. ２組のデータの解析　3.1 データのグラフ化、3.2 平均値の差の t 検定、3.3 分散の違いの検定
3.4 分散が異なる場合の平均値の差の比較
3.5 対応のある場合の平均値の差の t 検定、3.6 検出力と n の決め方
3.7 ノンパラメトリック検定

4. 相関・回帰・・・・・4.1 散布図、4.2 相関係数、4.3 回帰モデルとモデルの推定
4.4 誤差を考慮した推定、4.5 回帰分析適用上の諸問題

テキストの内容

ヒストグラム

p.72

- 表示2.2.2 Excel による度数表とヒストグラム

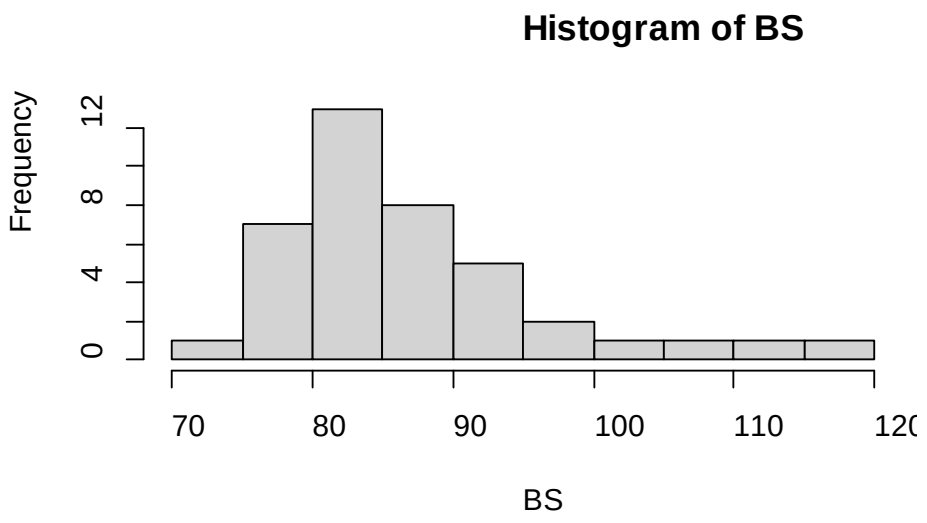
  スクリプトファイル：Green1-2-2a.R

  利用した関数：hist

  方法：BS（血糖値）のベクトルから作成

```
df <- read_excel("Green1-2.xlsx",
                 sheet = "2-ensyu")
df <- data.frame(df)
vt <- df$BS  # 40人の血液検査、血糖(BS)

hist(x = vt,
     breaks = "Sturges",
     freq = TRUE,
     right = FALSE,
     col = "lightgray",
     border = "black",
     main = "Histogram of BS",
     xlab = "BS",
     plot = TRUE)
```



ヒストグラム

- 表示2.2.2 Excel による度数表とヒストグラム

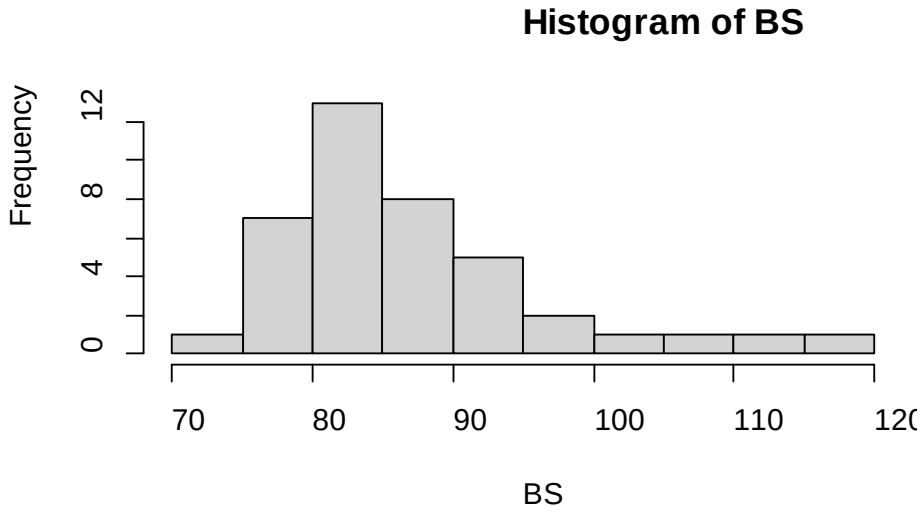
  スクリプトファイル：Green1-2-2a.R

  利用した関数：hist

  方法：BS（血糖値）のベクトルから作成

```
df <- read_excel("Green1-2.xlsx",
                 sheet = "2-ensyu")
df <- data.frame(df)
vt <- df$BS  # 40人の血液検査、血糖(BS)

hist(x = vt,
     breaks = "Sturges",
     freq = TRUE,
     right = FALSE,
     col = "lightgray",
     border = "black",
     main = "Histogram of BS",
     xlab = "BS",
     plot = TRUE)
```



breaks
= "Sturges"（デフォルト）
= "Scot"
= "FD"（Freedman-Diaconis）
= 20 （20 等分割）
= c(70, 80, 100, 120)

ヒストグラム

- 表示2.2.2 Excel による度数表とヒストグラム

  スクリプトファイル：Green1-2-2a.R

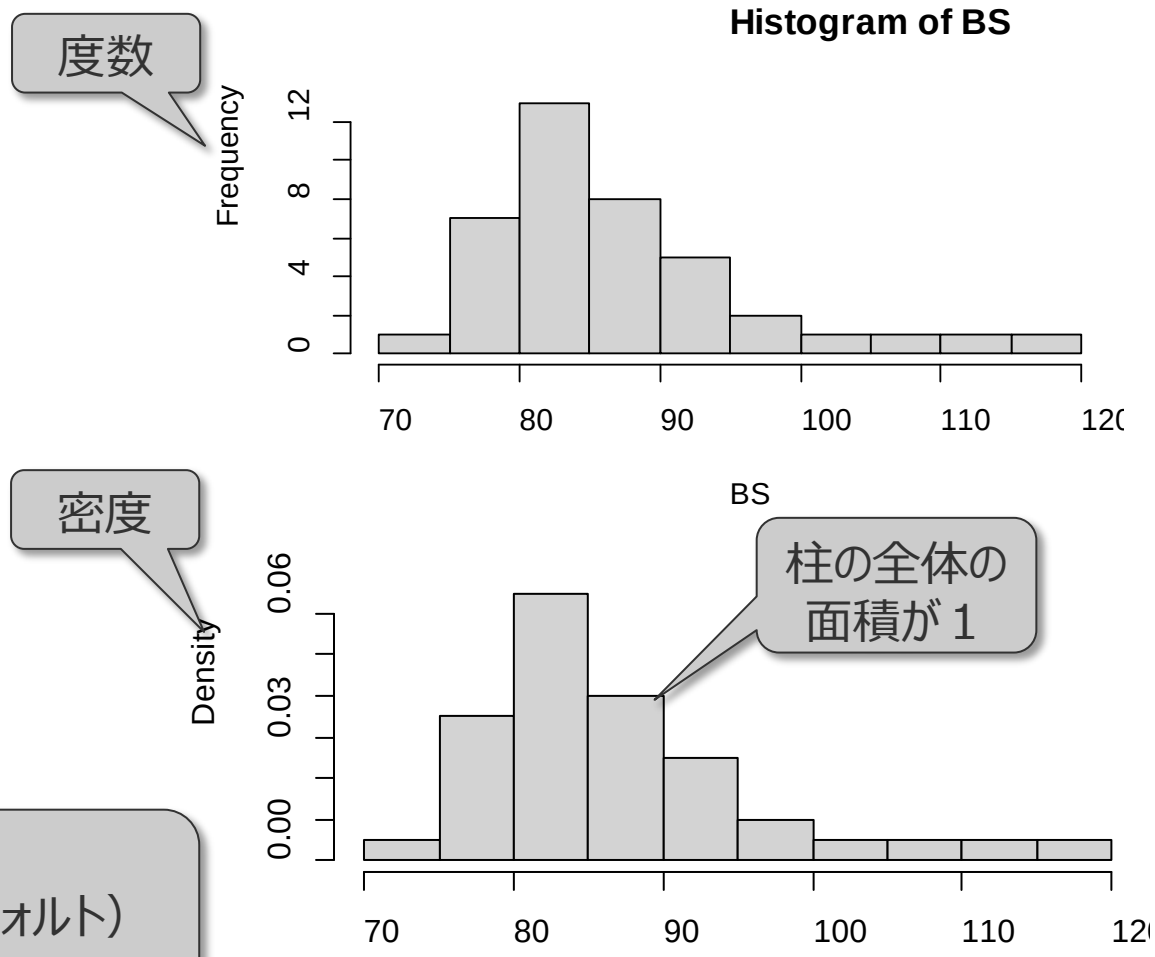
  利用した関数：hist

  方法：BS（血糖値）のベクトルから作成

```
df <- read_excel("Green1-2.xlsx",
                 sheet = "2-ensyu")
df <- data.frame(df)
vt <- df$BS  # 40人の血液検査、血糖(BS)

hist(x = vt,
     breaks = "Sturges",
     freq = TRUE,
     right = FALSE,
     col = "lightgray",
     border = "black",
     main = "Histogram
     xlab = "BS",
     plot = TRUE)
```

freq
= TRUE（度数、デフォルト）
= FALSE（密度）

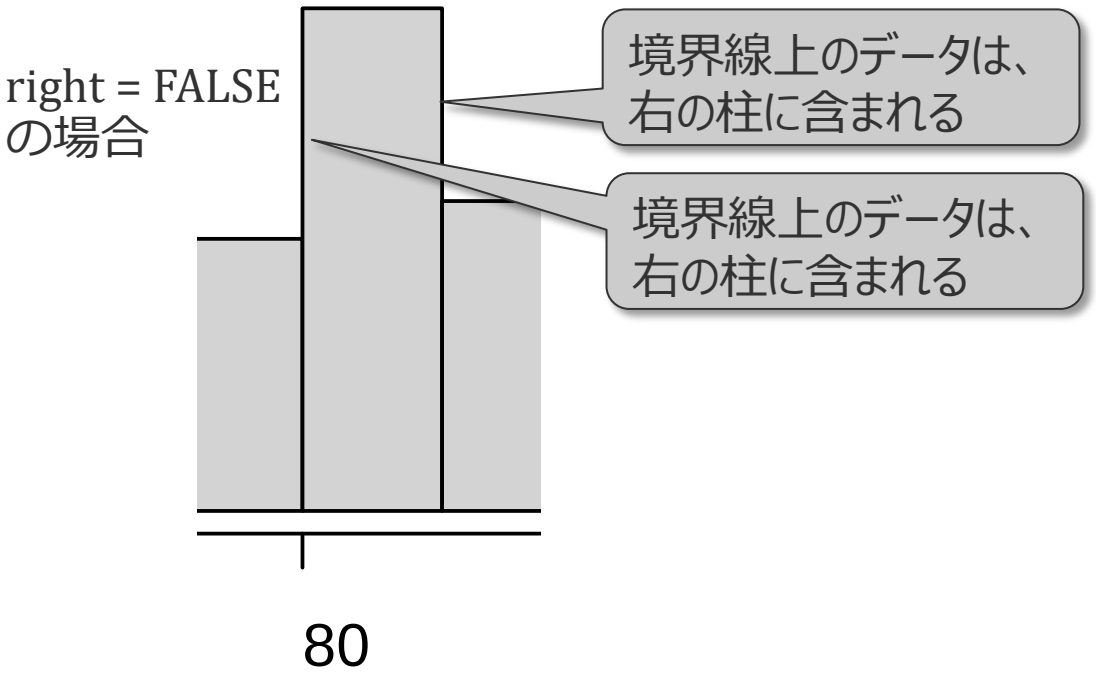


ヒストグラム

- 表示2.2.2 Excel による度数表とヒストグラム
  スクリプトファイル：Green1-2-2a.R
  利用した関数：hist
  方法：BS（血糖値）のベクトルから作成

```
df <- read_excel("Green1-2.xlsx",
                 sheet = "2-ensyu")
df <- data.frame(df)
vt <- df$BS  # 40人の血液検査、血糖(BS)

hist(x = vt,
     breaks = "Sturges",
     freq = TRUE,
     right = FALSE,
     col = "lightgray",
     border = "black",
     main = "Histogram of BS",
     xlab = "BS",
     plot = TRUE)
```



right
= TRUE（デフォルト）
左空き右閉じの区間（超えて、以下）
= FALSE（密度）
左閉じ右空きの区間（以上、未満）・・・テキスト

ヒストグラム

p.72

- 表示2.2.2 Excel による度数表とヒストグラム

  スクリプトファイル：Green1-2-2a.R

  利用した関数：hist

  方法：BS（血糖値）のベクトルから作成

```
df <- read_excel("Green1-2.xlsx",
                 sheet = "2-ensyu")
df <- data.frame(df)
vt <- df$BS  # 40人の血液検査、血糖(BS)

hist(x = vt,
     breaks = "Sturges",
     freq = TRUE,
     right = FALSE,
     col = "lightgray",
     border = "black",
     main = "Histogram of BS",
     xlab = "BS",
     plot = TRUE)
```

FALSE で
度数表を出力

度数表

```
## $breaks
##  [1]  70  75  80  85  90  95 100 105
110 115 120
##
## $counts
##  [1]  1  7 13  8  5  2  1  1  1  1
##
## $density
##  [1] 0.005 0.035 0.065 0.040 0.025
0.010 0.005 0.005 0.005 0.005
##
## $mids
##  [1]  72.5  77.5  82.5  87.5  92.5
  97.5 102.5 107.5 112.5 117.5
```

箱ひげ図

p.76

- 表示2.2.6 箱ひげ図

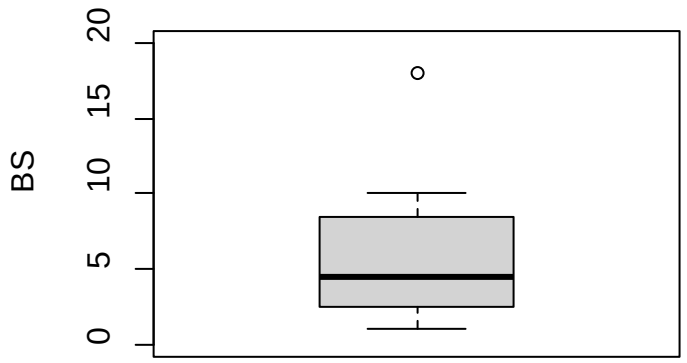
  スクリプトファイル：Green1-2-2a.R

  利用した関数：boxplot

  方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



箱ひげ図

p.76

● 表示2.2.6 箱ひげ図

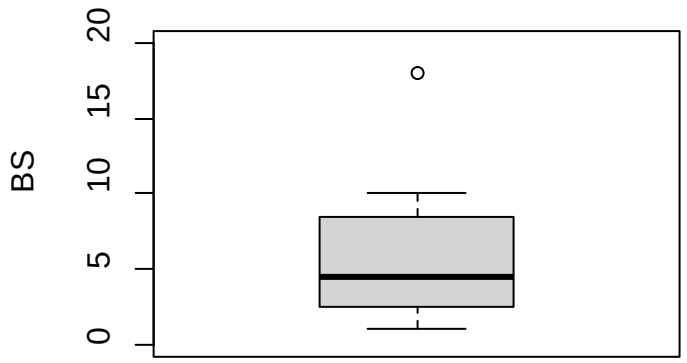
スクリプトファイル：Green1-2-2a.R

利用した関数：boxplot

方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



数値を設定
箱の幅 W の range 倍離れた位置の
外側のデータを外れ値とし、
内側のデータで最大値と最初値まで
「ひげ」を描く
デフォルトは 1.5 ・・・テキスト
表示2.2.6

箱ひげ図

p.76

● 表示2.2.6 箱ひげ図

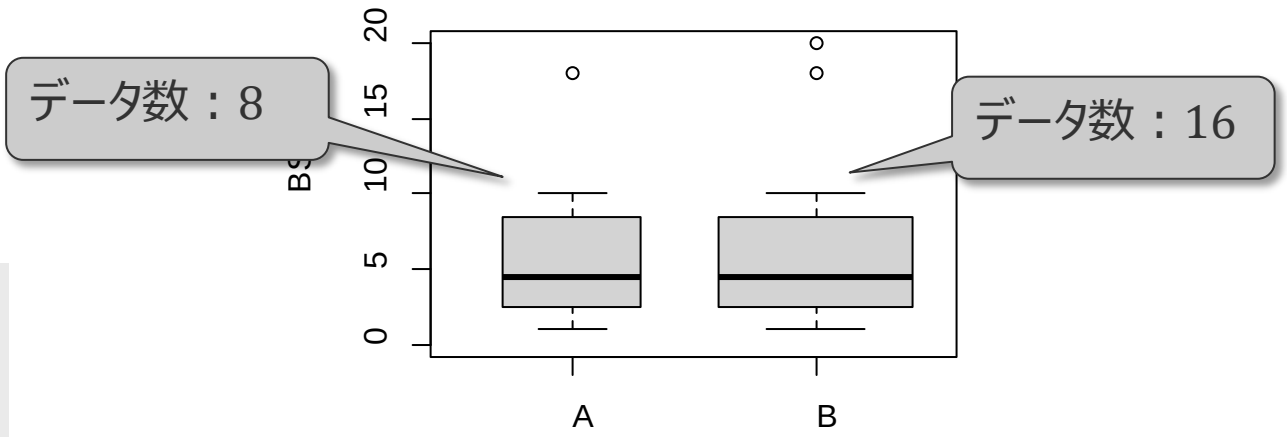
スクリプトファイル：Green1-2-2a.R

利用した関数：boxplot

方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



複数の群があり、
複数の箱ひげ図が並ぶ場合に有効
TRUE にすると観測値の数を反映して
箱の幅が変化する

箱ひげ図

● 表示2.2.6 箱ひげ図

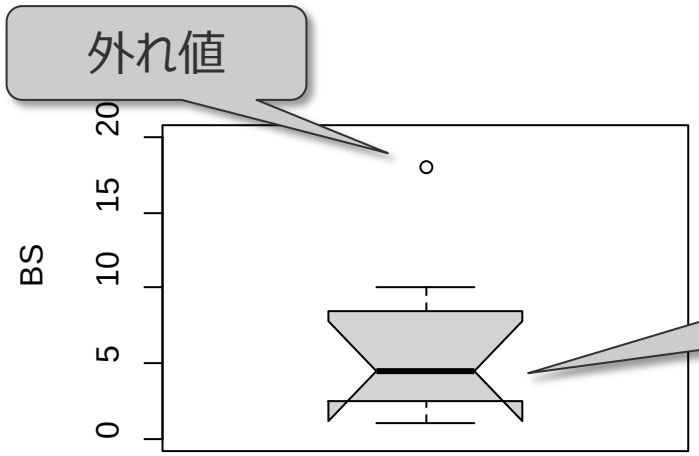
スクリプトファイル：Green1-2-2a.R

利用した関数：boxplot

方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



notch = TRUE の場合
箱に切り欠け（notch）を入れる
デフォルトは FALSE

outline = FALSE の場合
外れ値を描かない
デフォルトは TRUE

箱ひげ図

p.76

● 表示2.2.6 箱ひげ図

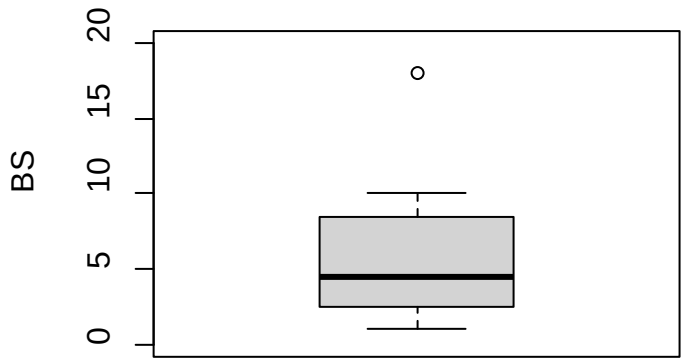
スクリプトファイル：Green1-2-2a.R

利用した関数：boxplot

方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



plot = FALSE の場合
箱ひげ図に必要な情報の要約を出力

border は箱の境界線の色指定

col は箱の中の色指定

箱ひげ図

p.76

● 表示2.2.6 箱ひげ図

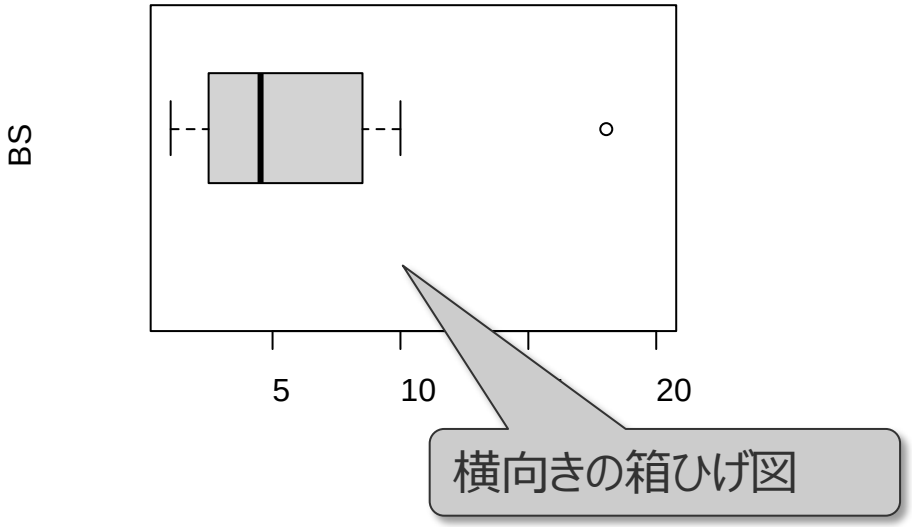
スクリプトファイル：Green1-2-2a.R

利用した関数：boxplot

方法：データを付値したベクトルから作成

```
vt2 <- c(1, 2, 3, 4, 5, 7, 10, 18)

boxplot(x = vt,
        range = 1.5,
        varwidth = FALSE,
        notch = FALSE,
        outline = TRUE,
        plot = TRUE,
        border = "black",
        col = "lightgray",
        log = "",
        ylim = c(70, 120),
        ylab = "BS",
        horizontal = FALSE)
```



log = "y" で、y 軸が対数目盛になる

horizontal = TRUE で
横向きの箱ひげ図が出力される

ヒストグラムと箱ひげ図

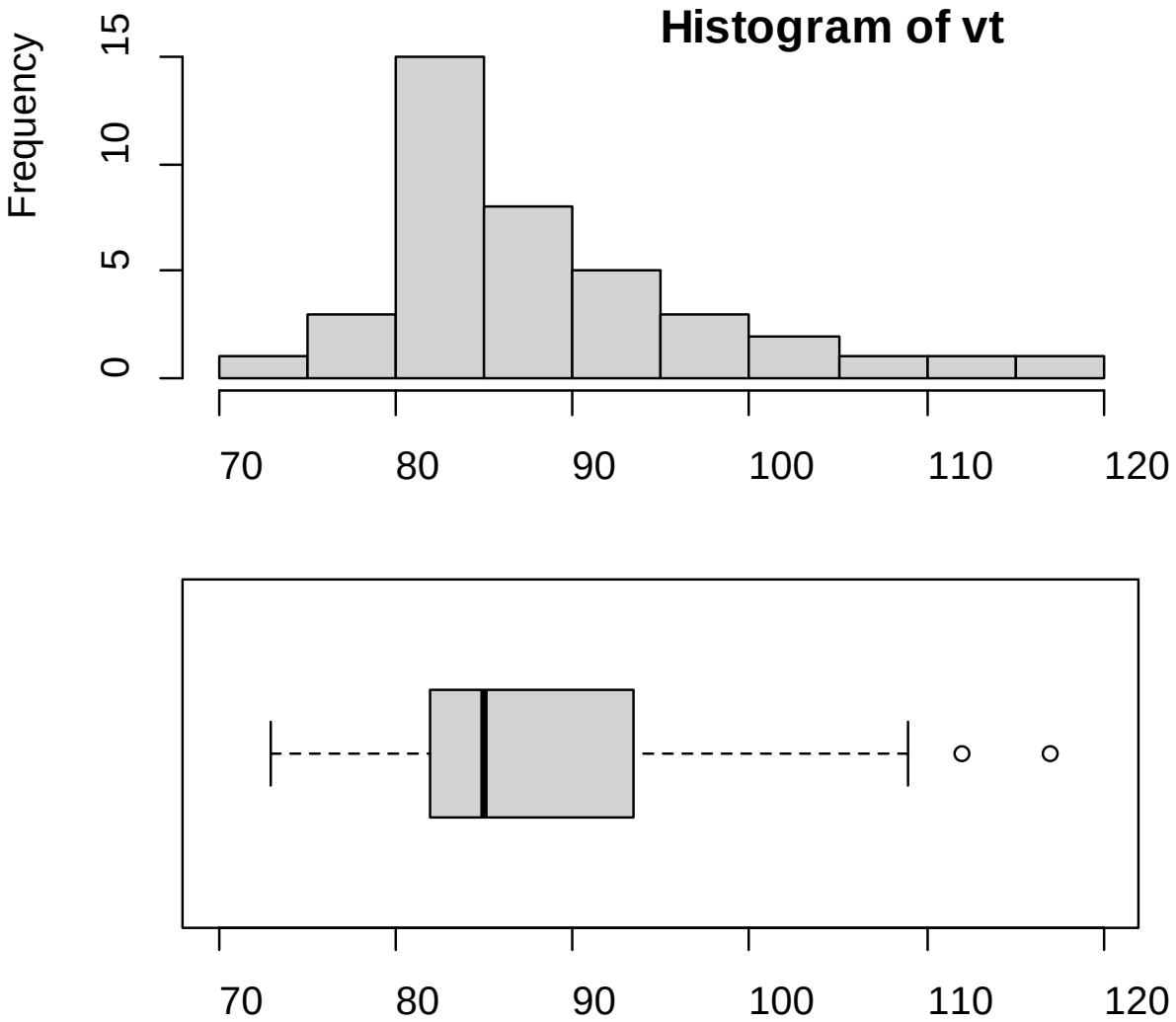
p.73

- 表示2.2.3 表示2.2.4 JMP によるヒストグラム
  - スクリプトファイル：Green1-2-2a.R
  - 利用した関数：hist、boxplot
  - 方法
    - 画面を２分割、余白を調整（§4.1 参照）
    - BS（血糖）のベクトル vt から
    - グラフィックス 関数を使って
    - ヒストグラムと箱ひげ図を描画

```
par(mfcol = c(2,1))
par(mar = c(3, 4, 2, 1))

hist(vt, breaks = 17, right = FALSE,
     xlim = c(70, 120))

boxplot(vt, horizontal = TRUE,
     ylim = c(70, 120))
```

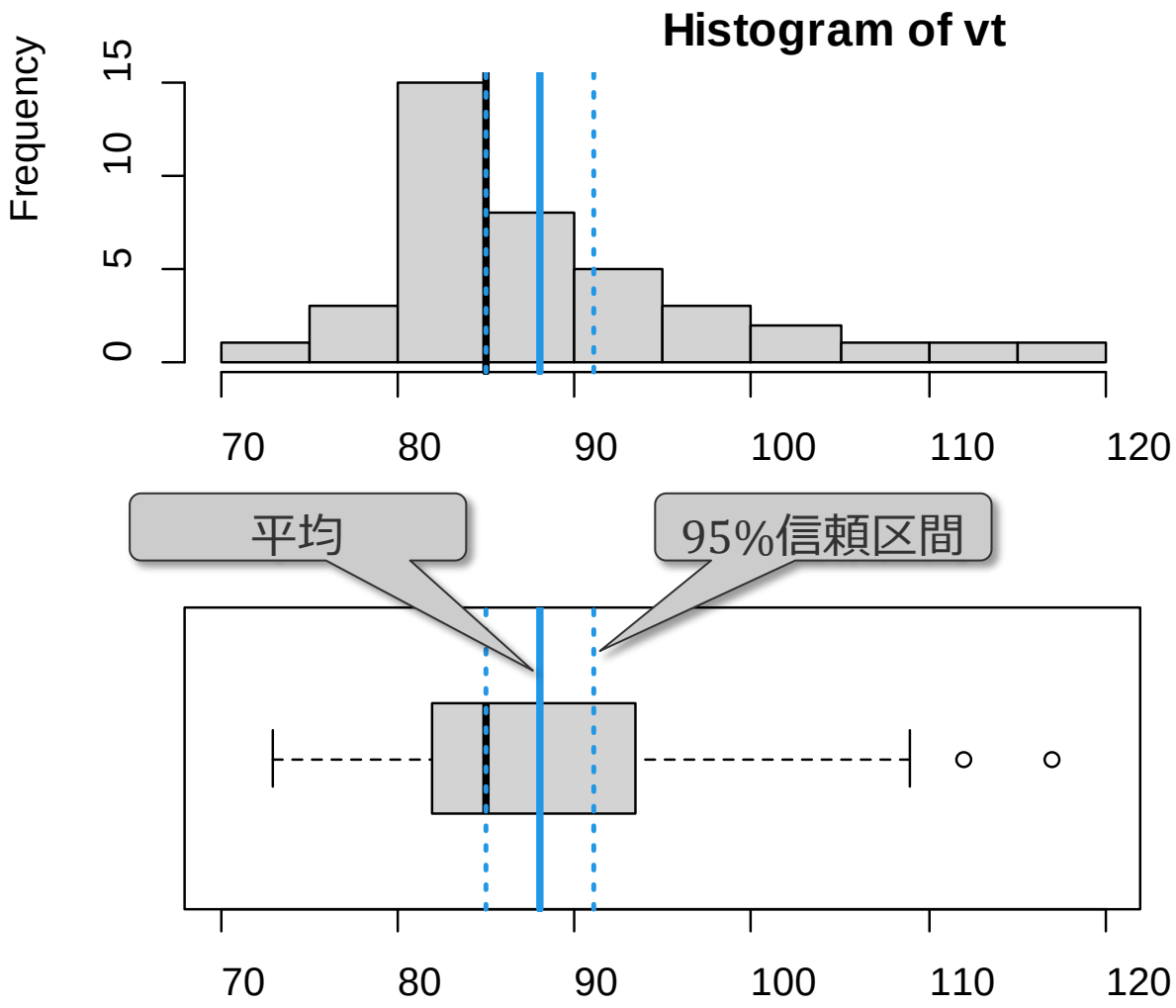


ヒストグラムと箱ひげ図

p.73

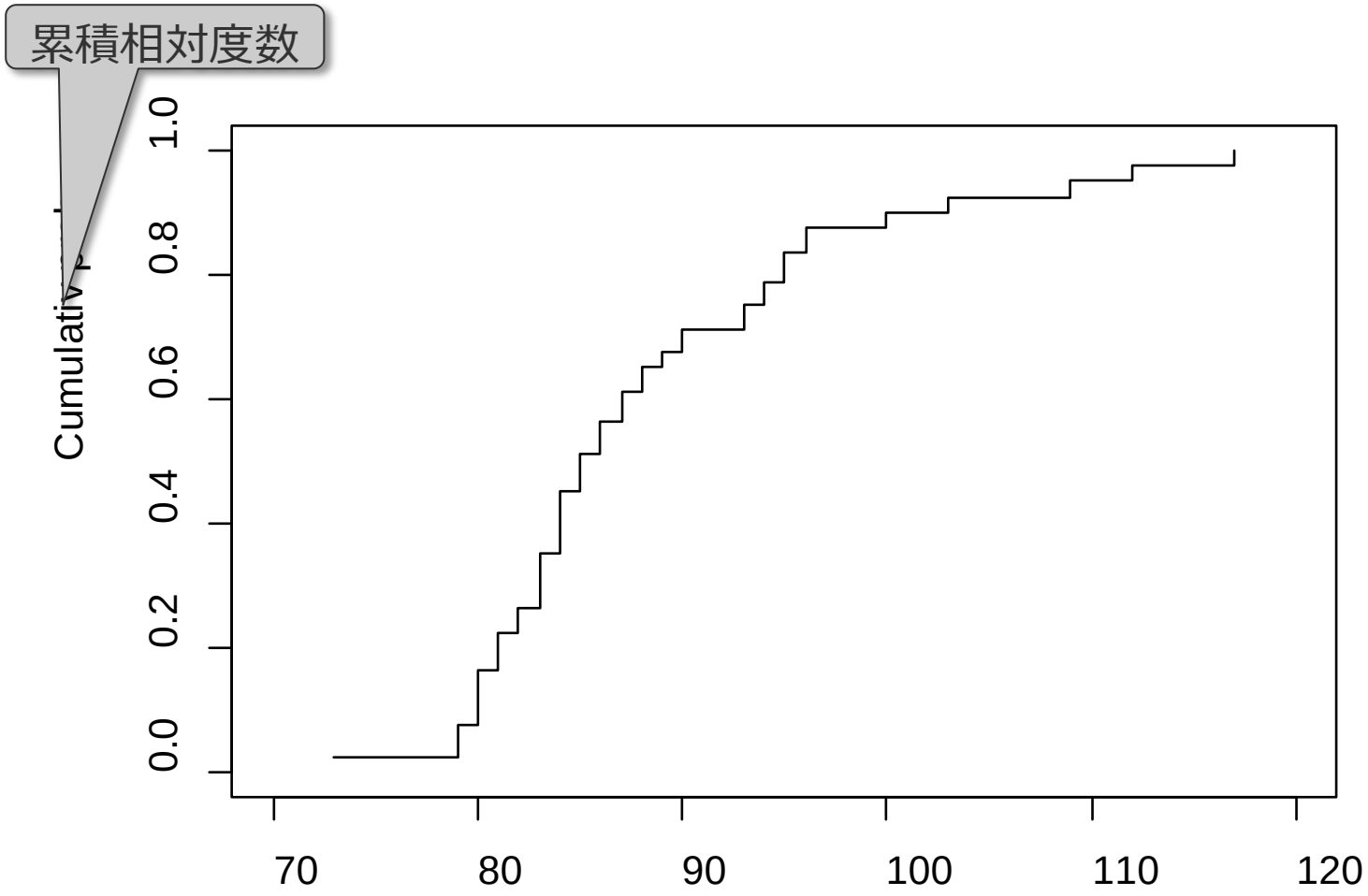
- 表示2.2.3 表示2.2.4 JMP によるヒストグラム
  - スクリプトファイル：Green1-2-2a.R
  - 利用した関数：hist、boxplot、t.test、
  - 方法
    - 画面を２分割、余白を調整（§4.1 参照）
    - BS（血糖）のベクトル vt から
    - グラフィックス 関数を使って
    - ヒストグラムと箱ひげ図を描画

    t.test 関数により、平均値、95%信頼区間を得て、これを abline 関数で図の中に追加



累積確率プロット

- 表示2.2.5 JMP によるヒストグラム
  - スクリプトファイル
    - Green1-2-2a.R
  - 利用した関数
    - plot、sort、rank、length
  - 方法
    - BS（血糖値）のベクトル vt を、
    - 昇順にソート
    - 平均順位に変換
    - サンプルサイズで割って
    - 累積確率を計算
    - plot 関数（type="s"）で描画



累積確率プロット

● 表示2.2.5 JMP によるヒストグラム

データフレームの
１列をベクトルに付値

ベクトルを昇順に
並び替え

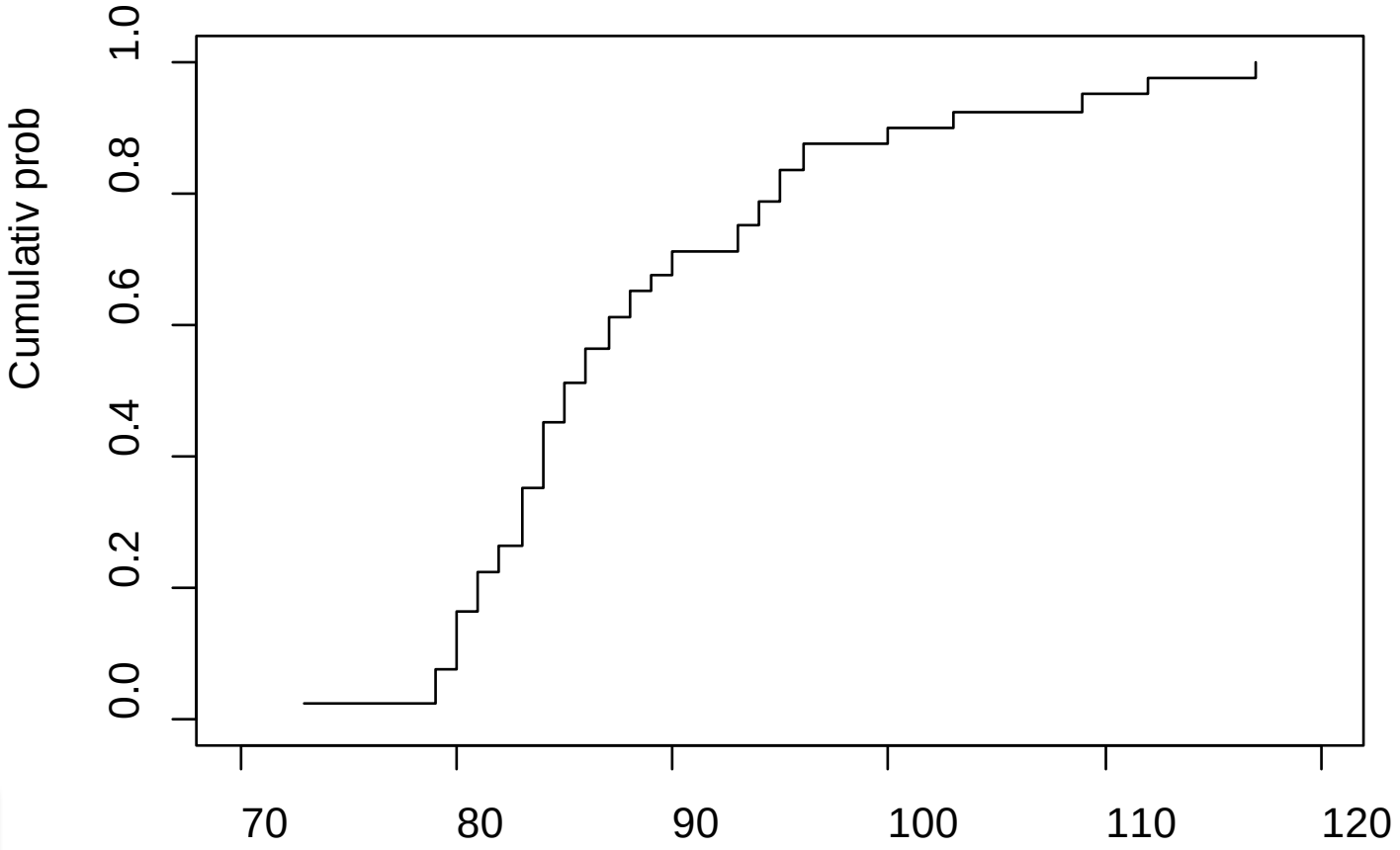
平均順位
／ サンプルサイズ

階段状に折れ線で
表示

```
vt <- df$BS
vt <- sort(vt)

cumu_p <- rank(vt) / length(vt)


plot(x = vt,
     y = cumu_p,
     ylab = "Cumulativ prob",
     xlim = c(70, 120),
     ylim = c(0, 1),
     type = "s")
```



テキストの補足

ローレンツ曲線とジニ係数

ローレンツ曲線とジニ係数

●ローレンツ曲線、ジニ係数

ローレンツ曲線：格差（不平等の程度）を視覚的に表したグラフ
ジニ係数　　：格差（不平等の程度）の指標
（累積相対度数の応用）

●事例

ある会社で、賞与 1,000 万円を 40 人の社員に分配する
分配の方法で生じる格差の程度を、ローレンツ曲線とジニ係数で表す

分配の方法(1)：1人当たり 24 万円～ 26 万円を支払う・・・格差は小
分配の方法(2)：1人当たり　5 万円～ 50 万円を支払う・・・格差は中
分配の方法(3)：1人当たり　1 万円～125 万円を支払う・・・格差は大

ローレンツ曲線とジニ係数

●ローレンツ曲線

（1）40人を賞与額で昇順に並べて、4人ずつの区間にまとめ、区間ごとの賞与計を算出
（境界値を決めた区間ではない）

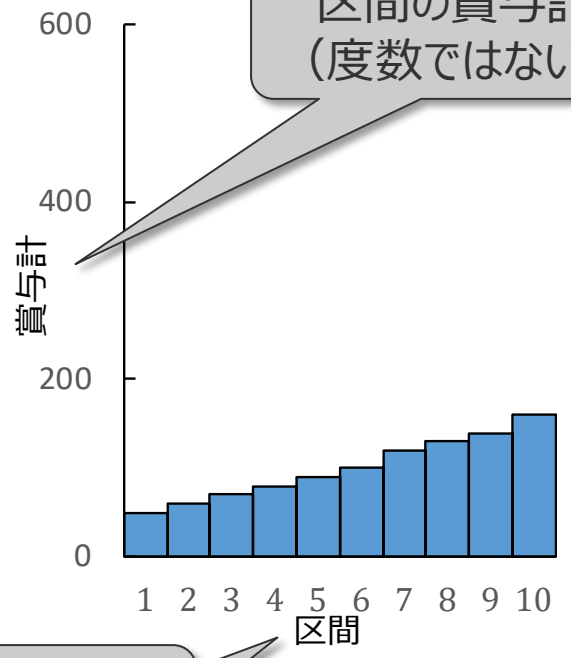
賞与の額 (万円) で
40人を
昇順に並べる

4 人ずつの区間

区間	賞与	賞与計
1	12	50
	12	
	13	
	13	
2	14	60
	15	
	15	
	16	
3	17	70
	17	
	17	
	19	
・	・	・
・	・	・

区間	人数	賞与計
1	4	50
2	4	60
3	4	70
4	4	80
5	4	90
6	4	100
7	4	120
8	4	130
9	4	140
10	4	160
計	40	1,000

区間の賞与計
（度数ではない）



4 人ずつの区間
（境界値はない）

ローレンツ曲線とジニ係数

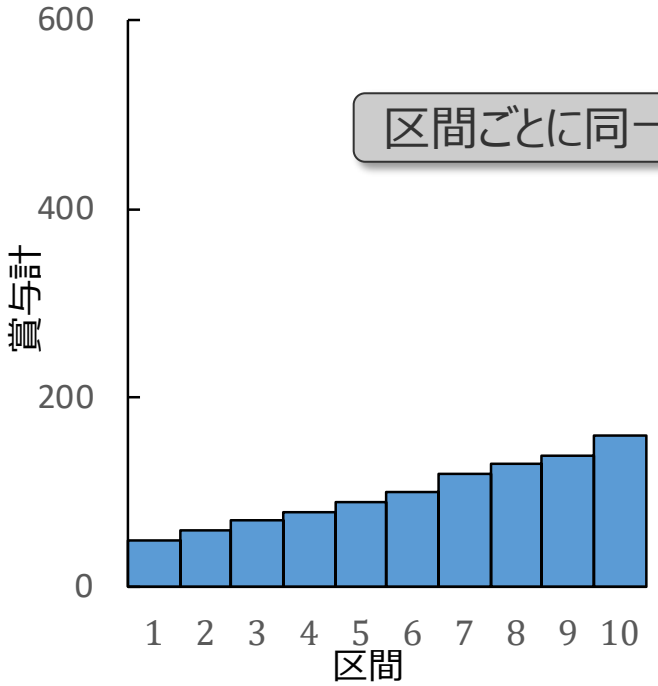
●ローレンツ曲線

（1）40人を賞与額で昇順に並べて、4人ずつの区間にまとめ、区間ごとの賞与計を算出

（2）人数と賞与計の比率、累積比率を算出

（3）横軸：人数の累積比率、縦軸：賞与計の累積比率をとってグラフ化

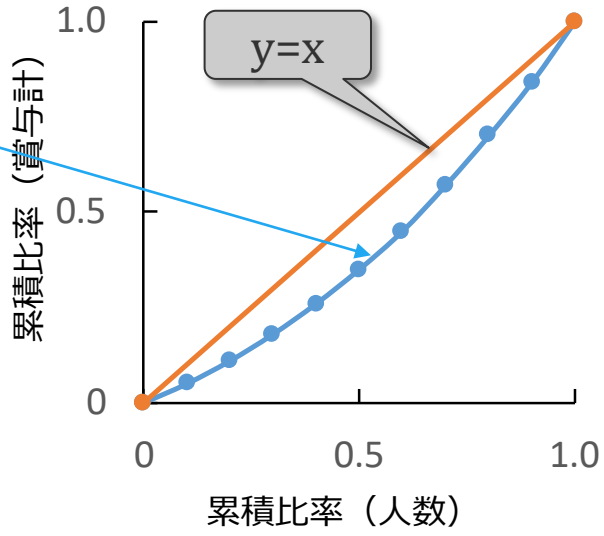
区間	人数	賞与計
1	4	50
2	4	60
3	4	70
4	4	80
5	4	90
6	4	100
7	4	120
8	4	130
9	4	140
10	4	160
計	40	1,000



区間ごとに同一

区間	比率 人数	比率 賞与計
1	0.1	0.05
2	0.1	0.06
3	0.1	0.07
4	0.1	0.08
5	0.1	0.09
6	0.1	0.10
7	0.1	0.12
8	0.1	0.13
9	0.1	0.14
10	0.1	0.16
計	1	1

区間	累積比率 人数	累積比率 賞与計
	0	0.00
1	0.1	0.05
2	0.2	0.11
3	0.3	0.18
4	0.4	0.26
5	0.5	0.35
6	0.6	0.45
7	0.7	0.57
8	0.8	0.70
9	0.9	0.84
10	1	1



ローレンツ曲線とジニ係数

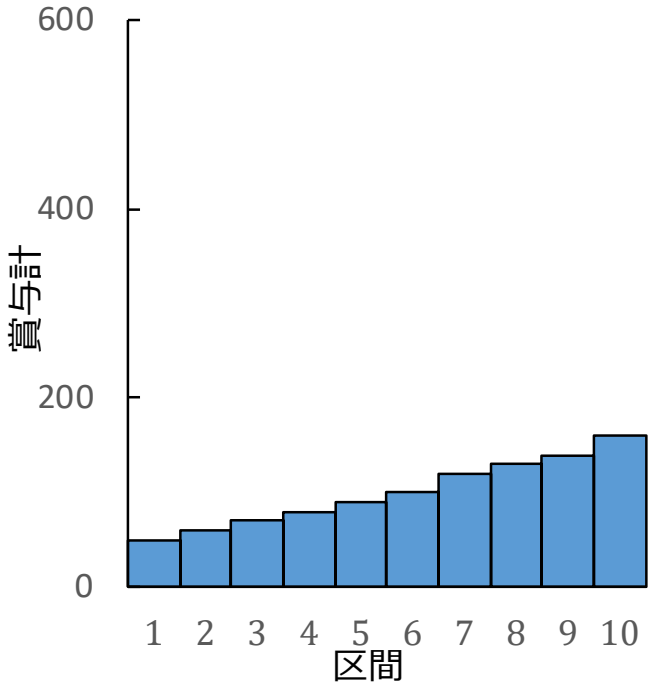
●ローレンツ曲線

（1）40人を賞与額で昇順に並べて、4人ずつの区間にまとめ、区間ごとの賞与計を算出

（2）人数と賞与計の比率、累積比率を算出

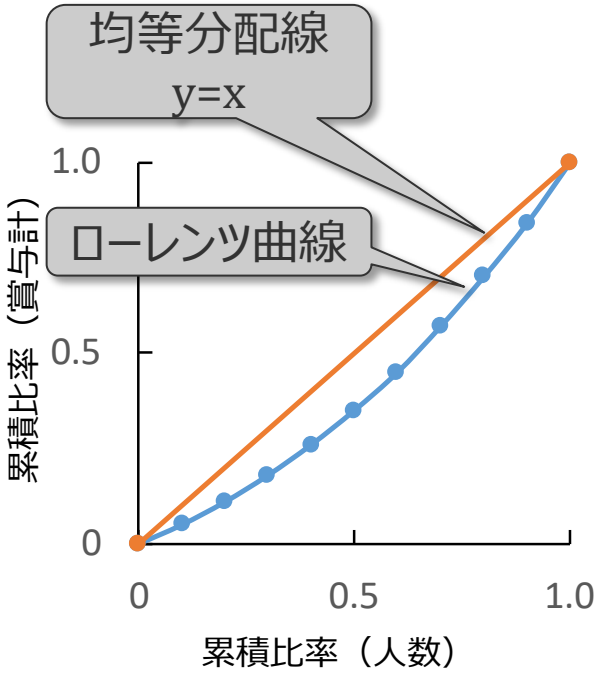
（3）横軸：人数の累積比率、縦軸：賞与計の累積比率をとってグラフ化

区間	人数	賞与計
1	4	50
2	4	60
3	4	70
4	4	80
5	4	90
6	4	100
7	4	120
8	4	130
9	4	140
10	4	160
計	40	1,000



区間	比率	
	人数	賞与計
1	0.1	0.05
2	0.1	0.06
3	0.1	0.07
4	0.1	0.08
5	0.1	0.09
6	0.1	0.10
7	0.1	0.12
8	0.1	0.13
9	0.1	0.14
10	0.1	0.16
計	1	1

区間	累積比率	
	人数	賞与計
	0	0.00
1	0.1	0.05
2	0.2	0.11
3	0.3	0.18
4	0.4	0.26
5	0.5	0.35
6	0.6	0.45
7	0.7	0.57
8	0.8	0.70
9	0.9	0.84
10	1	1



ローレンツ曲線とジニ係数

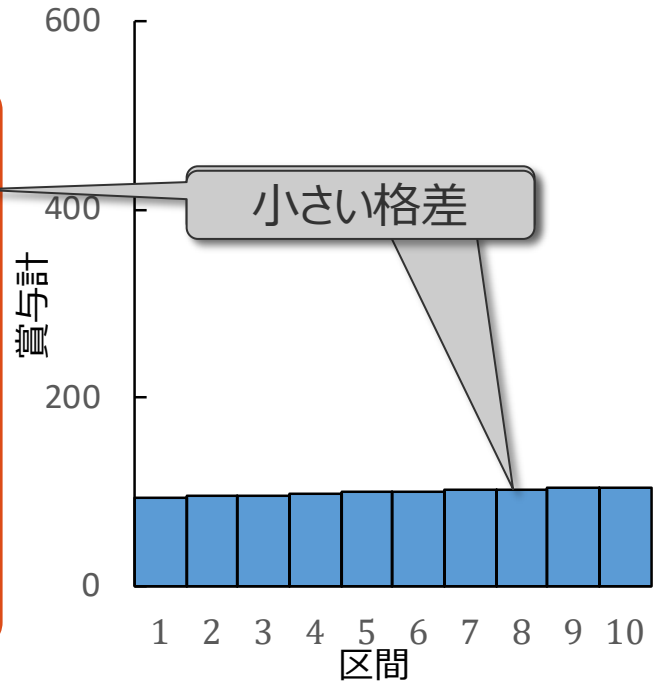
●ローレンツ曲線

（1）40人を賞与額で昇順に並べて、4人ずつの区間にまとめ、区間ごとの賞与計を算出

（2）人数と賞与計の比率、累積比率を算出

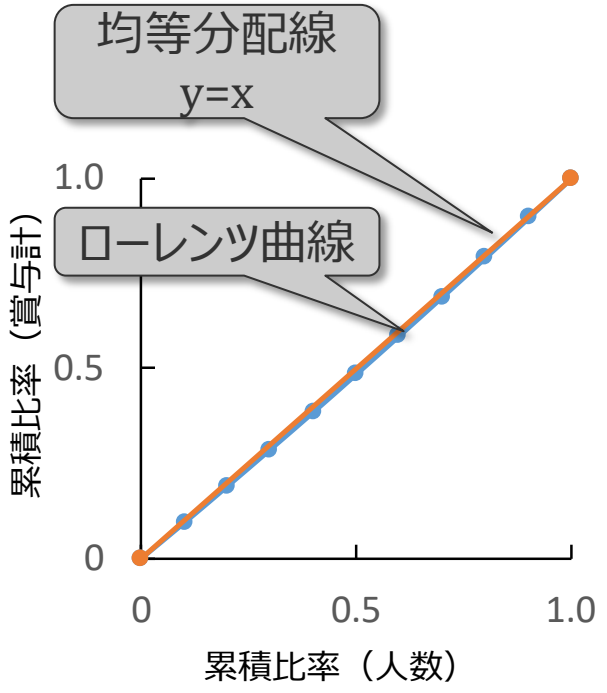
（3）横軸：人数の累積比率、縦軸：賞与計の累積比率をとってグラフ化

区間	人数	賞与計
1	4	94
2	4	96
3	4	97
4	4	98
5	4	100
6	4	101
7	4	102
8	4	103
9	4	104
10	4	105
計	40	1,000



区間	比率	
	人数	賞与計
1	0.1	0.09
2	0.1	0.10
3	0.1	0.10
4	0.1	0.10
5	0.1	0.10
6	0.1	0.10
7	0.1	0.10
8	0.1	0.10
9	0.1	0.10
10	0.1	0.11
計	1	1

区間	累積比率	
	人数	賞与計
	0	0.00
1	0.1	0.09
2	0.2	0.19
3	0.3	0.29
4	0.4	0.39
5	0.5	0.49
6	0.6	0.59
7	0.7	0.69
8	0.8	0.79
9	0.9	0.90
10	1	1



ローレンツ曲線とジニ係数

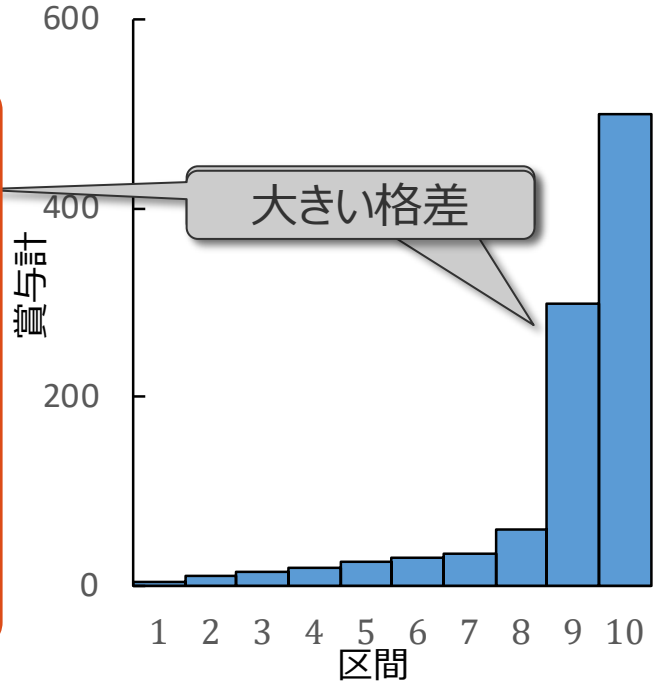
●ローレンツ曲線

（1）40人を賞与額で昇順に並べて、4人ずつの区間にまとめ、区間ごとの賞与計を算出

（2）人数と賞与計の比率、累積比率を算出

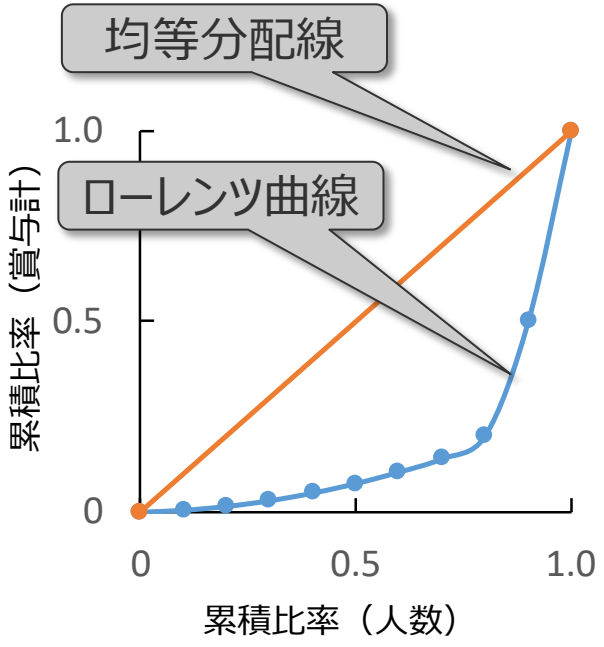
（3）横軸：人数の累積比率、縦軸：賞与計の累積比率をとってグラフ化

区間	人数	賞与計
1	4	5
2	4	10
3	4	15
4	4	20
5	4	25
6	4	30
7	4	35
8	4	60
9	4	300
10	4	500
計	40	1,000



区間	比率	
	人数	賞与計
1	0.1	0.01
2	0.1	0.01
3	0.1	0.02
4	0.1	0.02
5	0.1	0.03
6	0.1	0.03
7	0.1	0.04
8	0.1	0.06
9	0.1	0.30
10	0.1	0.50
計	1	1

区間	累積比率	
	人数	賞与計
	0	0.00
1	0.1	0.01
2	0.2	0.02
3	0.3	0.03
4	0.4	0.05
5	0.5	0.08
6	0.6	0.11
7	0.7	0.14
8	0.8	0.20
9	0.9	0.50
10	1	1

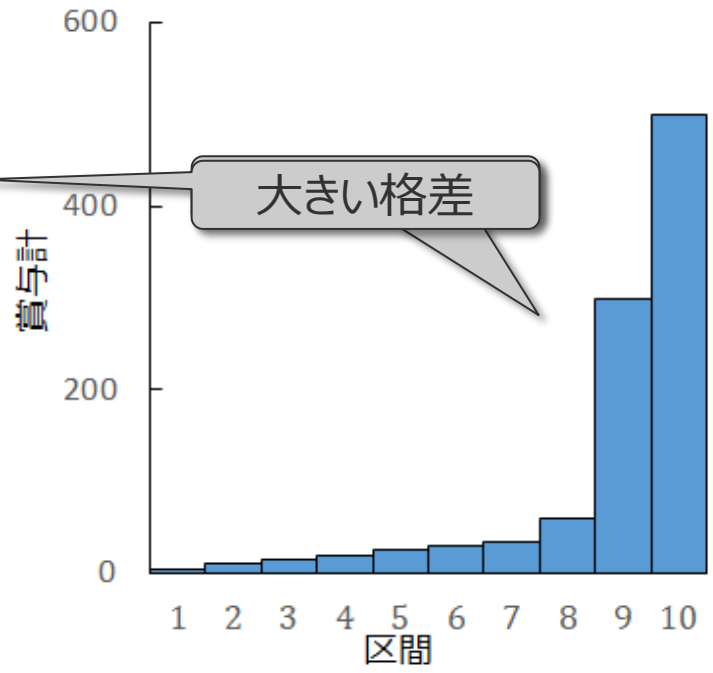


ローレンツ曲線とジニ係数

●ローレンツ曲線とジニ係数

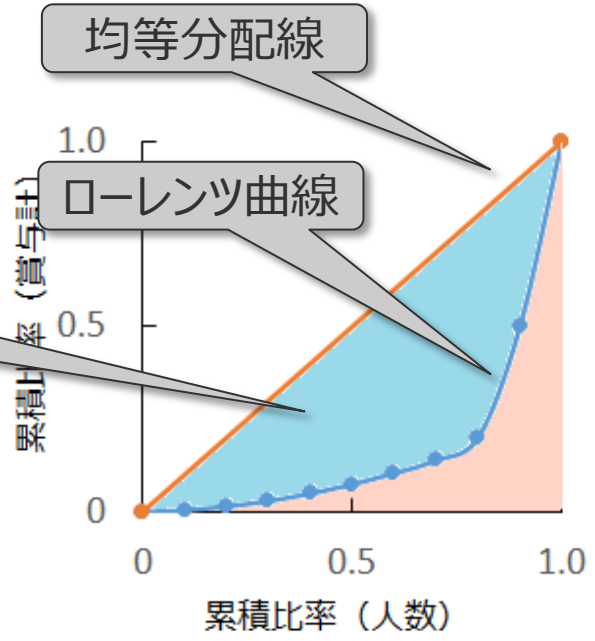
ジニ係数：均等配分線とローレンツ曲線で囲まれる面積（ブルー）が、
均等分配線よりも下の三角形の面積に占める割合、0～1 の値をとる

区間	人数	賞与計
1	4	5
2	4	10
3	4	15
4	4	20
5	4	25
6	4	30
7	4	35
8	4	60
9	4	300
10	4	500
計	40	1,000

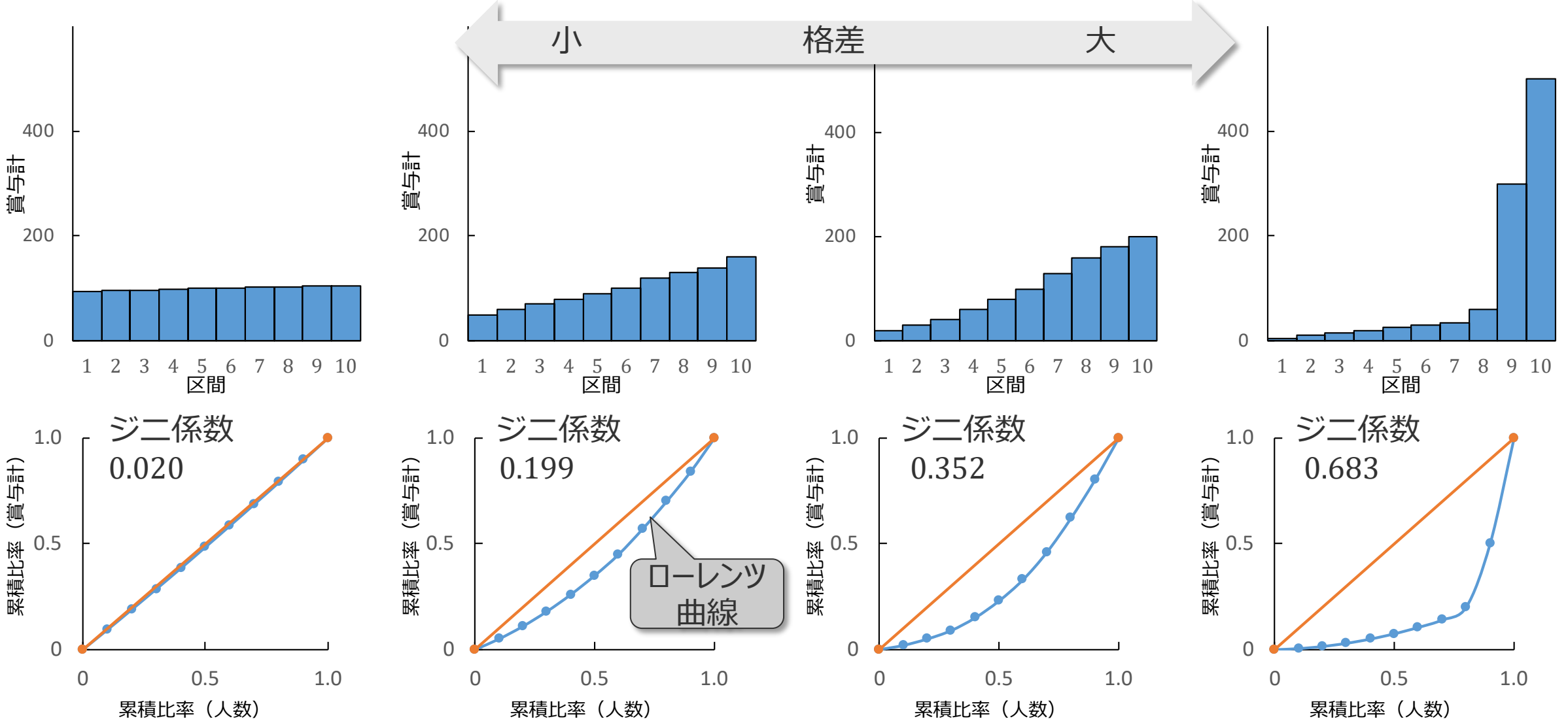


区間	比率	
	人数	賞与計
1	0.1	0.01
2	0.1	0.01
3	0.1	0.02
4	0.1	0.02
5	0.1	0.03
6	0.1	0.03
7	0.1	0.04
8	0.1	0.06
9	0.1	0.30
10	0.1	0.50
計	1	1

区間	累積比率	
	人数	賞与計
	0	0.00
1	0.1	0.01
2	0.2	0.02
5	0.5	0.08
6	0.6	0.11
7	0.7	0.14
8	0.8	0.20
9	0.9	0.50
10	1	1



ローレンツ曲線とジニ係数



ローレンツ曲線とジニ係数

- ローレンス曲線とジニ係数

  スクリプトファイル：Green1-2-2b.R

  利用した関数：ineq::Gini、ineq::Lc

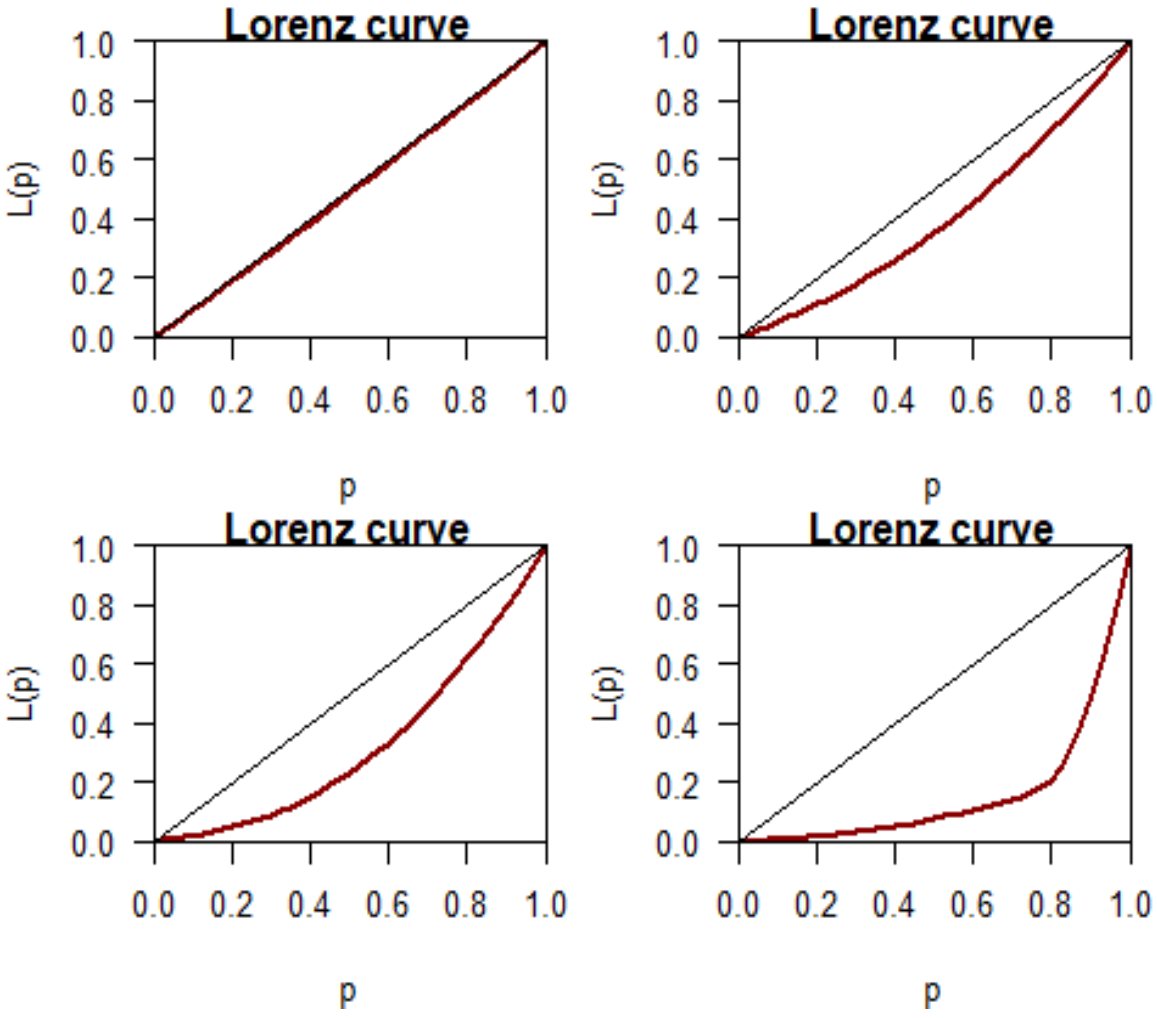
  方法：Excel ファイルの「incme」シートに40人へ賞与を分配する事例がある

  bonus1、bonus2、bonus3、bosum4

```
df <- read_excel("Green1-2.xlsx",
      sheet = "income")
df <- data.frame(df)

Gini(df$bonus4)
## [1] 0.6833
plot(Lc(df$bonus4),
     col = "darkred", lwd = 2)
```

ジニ係数



- 作成　　片瀬雅彦
- 作成時期　　2021年8月26日
- 改定　　2022年5月11日