Rで統計解析入門

(3) 1つのデータの要約

本日のメニュー

1. データの読み込み
   - データ「DEP」の概要と読み込み
   - 薬剤 A の QOL のデータの取り出し
2. 1 つのデータの要約
   - 要約統計量の一覧
   - グラフの作成
3. 検定と信頼区間について

架空のデータ「DEP」

- うつ病を患っている患者さんに薬剤治療を行った後，QOLの点数を測定
- QOL（Quality of Life；生活の質）の点数：以下の架空のアンケート票を使って患者さんに回答してもらい，各質問項目で回答した番号を合計したものを当該患者さんの点数とする

No	質問	当てはまらない（1点）	あまり当てはまらない（2点）	やや当てはまる（3点）	当てはまる（4点）
1	起床時に気分が良い		○		
2	朝食は美味しい				○
3	学校/会社に行きたい	○			
:	:	:	:	:	:

架空のデータ「DEP」の変数

- GROUP：薬剤の種類（A，B，C）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- EVENT：改善の有無（1：改善あり，2：改善なし）
  ⇒ QOLの点数が 5 点以上である場合を「改善あり」とする
- DAY：観察期間（数値，単位は日）
- PREDRUG：前治療薬の有無（YES：他の治療薬を投与したことあり，
  NO：投与したことなし）
- DURATION：罹病期間（数値，単位は年）

架空のデータ「DEP」

GROUP	QOL	EVENT	DAY	PREDRUG	DURATION
A	15	1	50	NO	1
A	13	1	200	NO	3
A	11	1	250	NO	2
A	11	1	300	NO	4
A	10	1	350	NO	2
A	9	1	400	NO	2
A	8	1	450	NO	4
A	8	1	550	NO	2
A	6	1	600	NO	5
A	6	1	100	NO	7
A	4	2	250	NO	4
A	3	2	500	NO	6
A	3	2	750	NO	3
A	3	2	650	NO	7
A	1	2	1000	NO	8
A	6	1	150	YES	6
A	5	1	700	YES	5
A	4	2	800	YES	7
A	2	2	900	YES	12
A	2	2	950	YES	10
B	13	1	380	NO	9
B	12	1	880	NO	5
B	11	1	940	NO	2
B	4	2	20	NO	7
B	4	2	560	NO	2
B	5	1	320	YES	11
B	5	1	940	YES	3
B	4	2	80	YES	6
B	3	2	140	YES	7
B	3	2	160	YES	13

架空のデータ「DEP」

GROUP	QOL	EVENT	DAY	PREDRUG	DURATION
B	3	2	240	YES	15
B	2	2	280	YES	9
B	2	2	440	YES	8
B	2	2	520	YES	7
B	2	2	620	YES	9
B	2	2	740	YES	8
B	2	2	860	YES	2
B	1	2	880	YES	10
B	0	2	920	YES	8
B	0	2	960	YES	4
C	9	1	170	NO	1
C	7	1	290	NO	4
C	5	1	430	NO	2
C	3	2	610	NO	4
C	2	2	110	NO	5
C	2	2	410	NO	2
C	1	2	530	NO	7
C	1	2	580	NO	2
C	0	2	810	NO	3
C	0	2	990	NO	10
C	6	1	30	YES	1
C	5	1	830	YES	6
C	3	2	70	YES	16
C	2	2	310	YES	9
C	2	2	370	YES	18
C	1	2	490	YES	7
C	1	2	690	YES	10
C	0	2	730	YES	3
C	0	2	770	YES	12
C	0	2	910	YES	8

データ「DEP」の読み込み

1. データ「DEP」を以下からダウンロードする
   http://www.cwk.zaq.ne.jp/fkhud708/files/dep.csv
2. ダウンロードした場所を把握する ⇒ ここでは「c:/temp」とする
3. R を起動し，2. の場所に移動し，データを読み込む

```
> setwd("c:/temp")              # dep.csv がある場所に移動
> getwd()                       # 移動できたかどうか確認
> DEP <- read.csv("dep.csv")    # dep.csv を読み込む
> head(DEP)                     # データ DEP の中身を確認
  GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
3     A  11     1 250      NO        2
4     A  11     1 300      NO        4
:     :   :     :   :       :        :
```

薬剤 A のQOL スコアの要約

- データ「DEP」から薬剤 A のデータのみ抽出

```
> A <- subset(DEP, GROUP=="A")
> head(A)
  GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
3     A  11     1 250      NO        2
4     A  11     1 300      NO        4
5     A  10     1 350      NO        2
6     A   9     1 400      NO        2
```

薬剤 A のQOL スコアの要約

- データ「DEP」から薬剤 A のデータのみ抽出した後，
  変数 QOL の変数のみデータフレームに格納

```
> A <- subset(DEP, GROUP=="A", select=QOL)
> head(A)
  QOL
1  15
2  13
3  11
4  11
5  10
6   9
```

薬剤 A のQOL スコアの要約

- データ「DEP」から薬剤 A のデータのみ抽出した後，
  変数 QOL の変数のみベクトルに格納
  ⇒ 以降はベクトル A を使用

```
> A <- subset(DEP, GROUP=="A")$QOL
> A
[1] 15 13 11 11 10  9  8  8  6  6  4  3  3  3  1  6  5  4  2  2
```

本日のメニュー

1. データの読み込み
   - データ「DEP」の概要と読み込み
   - 薬剤 *A* の QOL のデータの取り出し
2. 1 つのデータの要約
   - 要約統計量の一覧
   - グラフの作成
3. 検定と信頼区間について

薬剤 A の QOL スコアの要約

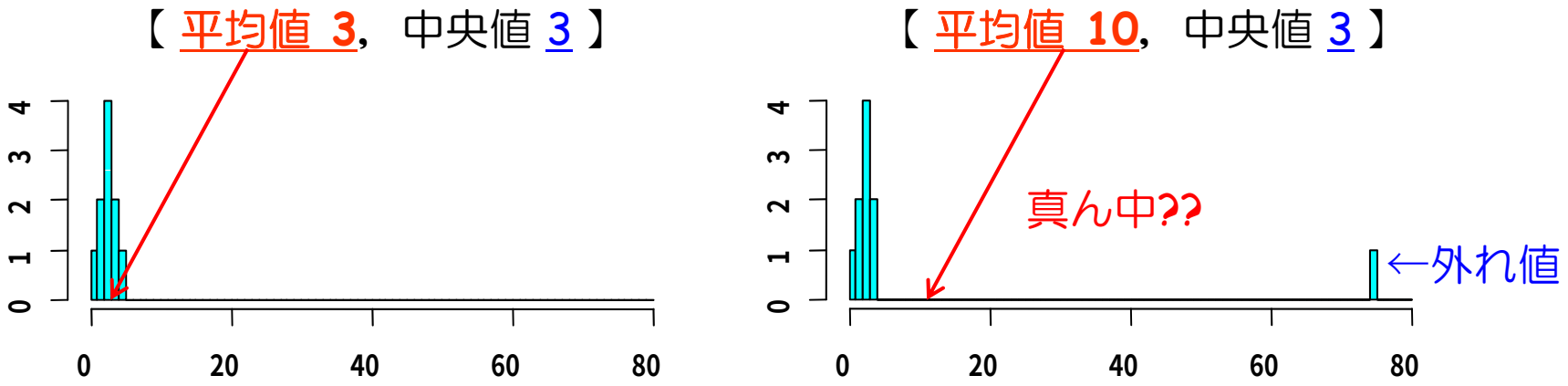
```
> summary(A)    # 最小値，25%点，中央値，平均値，75%点，最大値
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    6.00    6.50    9.25   15.00

> var(A)        # 分散
[1] 15.84211

> sd(A)         # 標準偏差
[1] 3.980214

> range(A)      # 範囲
[1]  1 15

> IQR(A)        # 四分位範囲
[1] 6.25
```

要約統計量の一覧

- 最小値（Min.）：データの中で一番小さい値
- 25%点（1st Qu.）：最小値から数えて全体の 1/4 であるデータ
- 中央値（50%点，Median）：最小値から数えて全体の半分であるデータ
- 平均値（Mean）：データの合計をデータの数で割った値
- 75%点（3rd Qu.）：最小値から数えて全体の 3/4 であるデータ
- 最大値（Max.）：データの中で一番大きい値
- 分散：「データとデータの平均値との差」を 2 乗したものを足し算し，
  「データの個数－ 1 」で割った値
- 標準偏差：分散の平方根（ルート）
- 範囲：最小値～最大値
- 四分位範囲：75%点から25%点を引いた値

「真ん中」を表す指標

- 平均値（Mean）
  - 「各データとの差の 2 乗和」を最小としている
  - 外れ値（極端な値）があると意味のない値になる可能性がある
- 中央値（Median）
  - 「各データとの差の絶対値の和」を最小としている
  - 外れ値（極端な値）の影響を受けにくい



【参考】前の頁のグラフを作成するプログラム

```
> x <- c(1,2,2,3,3,3,3,4,4,5)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.25    3.00    3.00    3.75    5.00
> hist(x,breaks=seq(0,80,1),col="cyan")   # 左の図

> x <- c(1,2,2,3,3,3,3,4,4,75)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.25    3.00   10.00    3.75   75.00
> hist(x,breaks=seq(0,80,1),col="cyan")   # 右の図
```

- 「QOL の平均値が●である」という情報だけでは，「真ん中はどこ」という情報だけなので心もとない（例えば「ばらつき」の情報が不足）
- 「ばらつき」をふまえる ⇒ 区間推定・信頼区間の登場（後述）

「ばらつき」を表す指標

- 分散，標準偏差
  - 外れ値（極端な値）の影響を受けやすい
  - 標準偏差は元のデータと次元が同じなので，解釈がしやすい
    （分散はデータを 2 乗しているので元のデータと次元が異なる）
- 四分位範囲
  - 外れ値（極端な値）の影響を受けにくい

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    6.00    6.50    9.25   15.00
```

四分位範囲

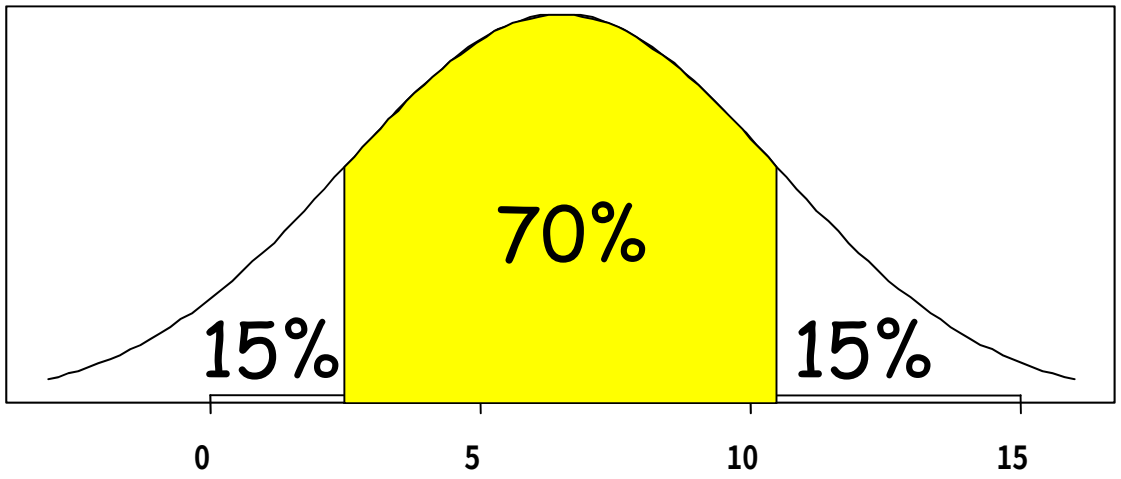
（この範囲の中に，全体の50%のデータ（下から1/4～上から1/4）が含まれる）

「ばらつき」を表す指標

（データが正規分布に従っていると仮定すると・・・）

- 全体の約 70 %のデータが平均値±標準偏差（2.5～10.5）に含まれる

⇒ 端から 3 個（ 20×0.15 ＝ 3 ）が外れることになる（大体合っている）



1	2	2	3	3	3	4	4	5	6	6	6	8	8	9	10	11	11	13	15

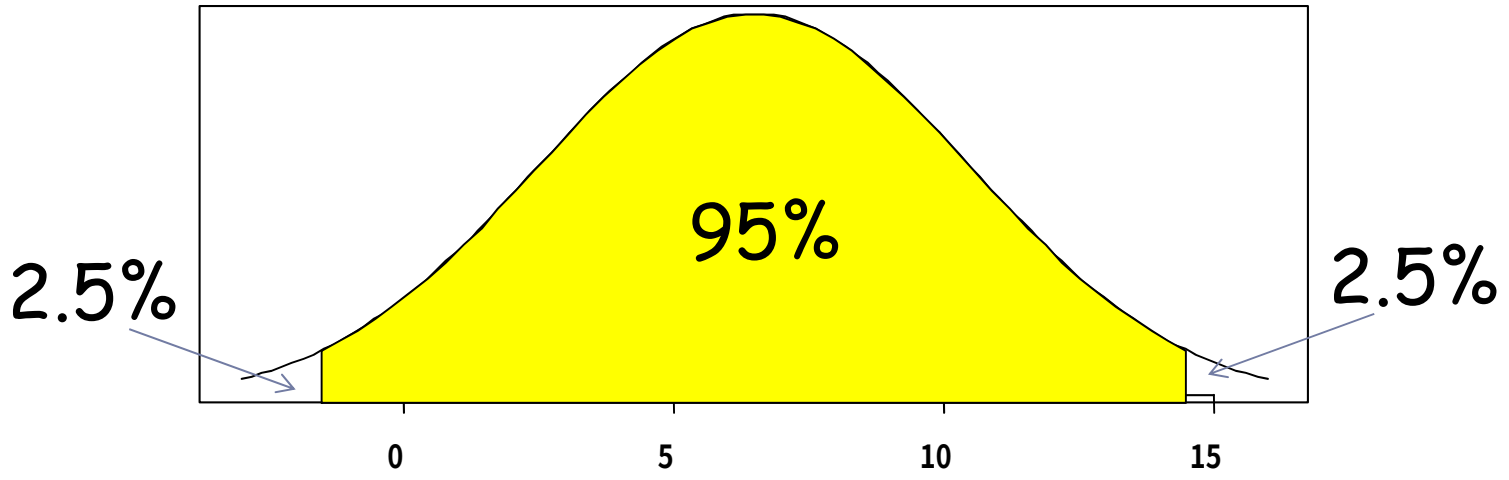
【 薬剤 A の QOL スコア（小さい順に並べ替えたもの） 】

「ばらつき」を表す指標

（データが正規分布に従っていると仮定すると・・・）

- 全体の約 95 %のデータが平均値±2×標準偏差（-1.5～14.5）に含まれる

⇒ 端から 0 個か 1 個（ 20×0.025 ＝ 0.5 ）が外れることになる（大体合っている）



1	2	2	3	3	3	4	4	5	6	6	6	8	8	9	10	11	11	13	15

【 薬剤 A の QOL スコア（小さい順に並べ替えたもの） 】

【参考】前の頁のグラフを作成するプログラム

```
> curve(dnorm(x, 6.5, 4), -3, 16)                    # 平均 6.5，標準偏差 4 の正規分布
> xvals <- seq(2.5, 10.5, length=50)                 # 領域をx軸方向に30個の多角形(台形)に等分割
> dvals <- dnorm(xvals, 6.5, 4)                      # 対応するグラフの高さ
> polygon(c(xvals,rev(xvals)),
+         c(rep(0,50), rev(dvals)), col="yellow")    # 塗りつぶす

> curve(dnorm(x, 6.5, 4), -3, 16)                    # 平均 6.5，標準偏差 4 の正規分布
> xvals <- seq(-1.5, 14.5, length=50)                # 領域をx軸方向に30個の多角形(台形)に等分割
> dvals <- dnorm(xvals, 6.5, 4)                      # 対応するグラフの高さ
> polygon(c(xvals, rev(xvals)),
+         c(rep(0,50), rev(dvals)), col="yellow")    # 塗りつぶす
```

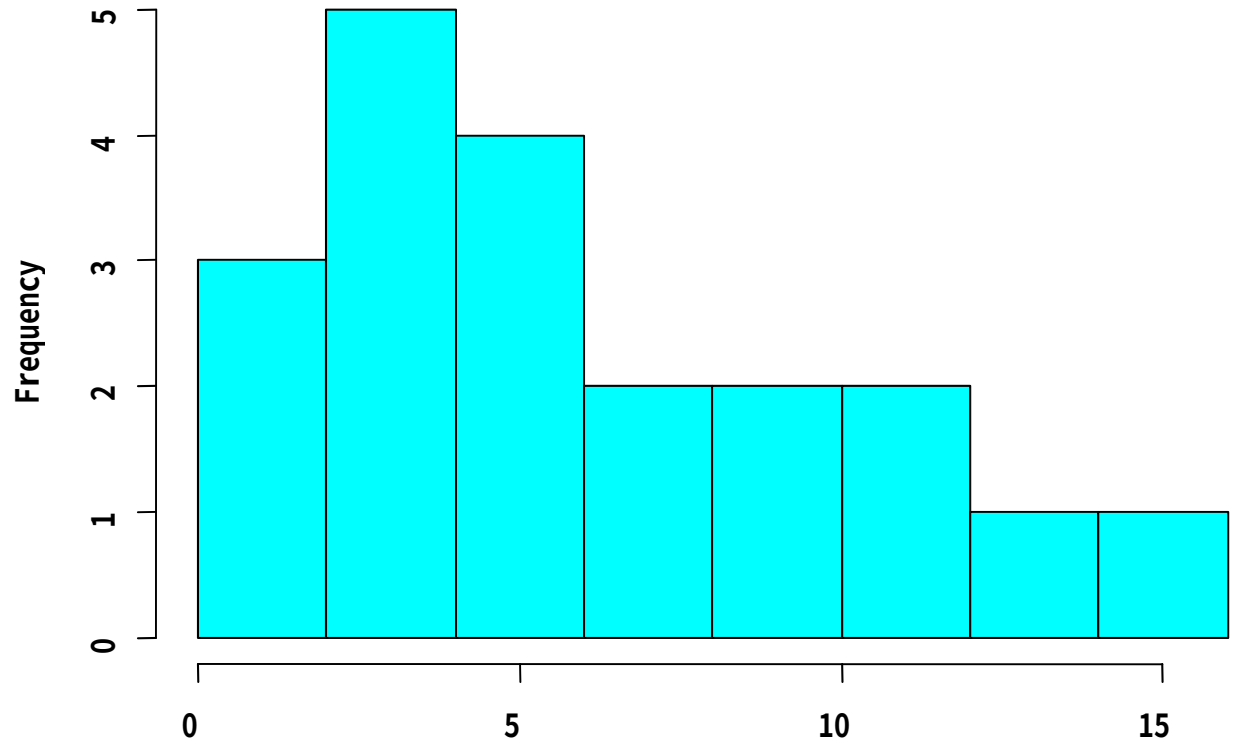
本日のメニュー

1. データの読み込み
   - データ「DEP」の概要と読み込み
   - 薬剤 *A* の QOL のデータの取り出し
2. 1 つのデータの要約
   - 要約統計量の一覧
   - グラフの作成
3. 検定と信頼区間について

薬剤 A の QOL スコアのヒストグラム

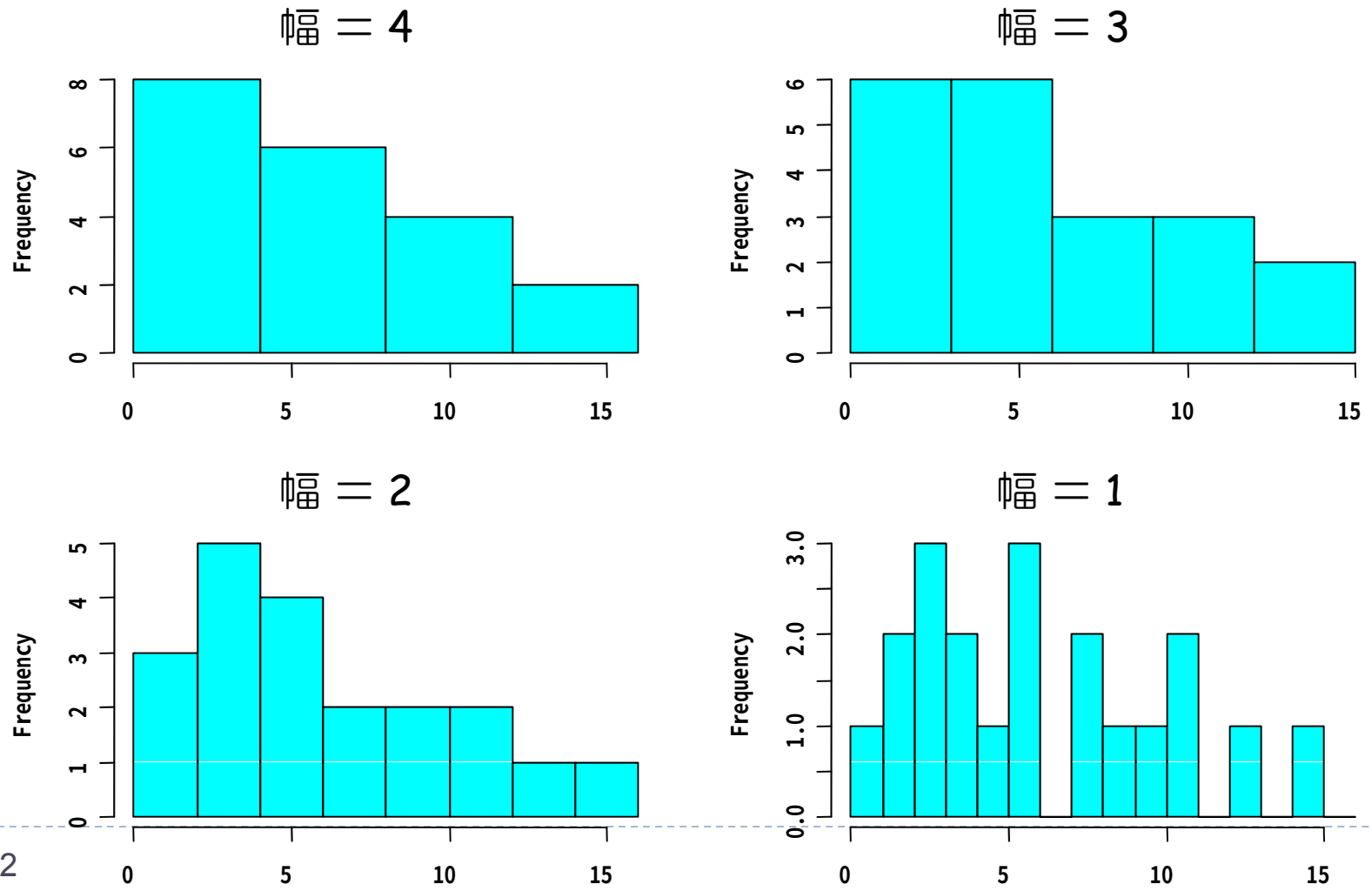
- 分布をパッと確認する場合はヒストグラムが手っ取り早い

```
> hist(A, col="cyan")
```



ヒストグラムの問題点

- 棒の横幅を変えると印象が変わる・・・



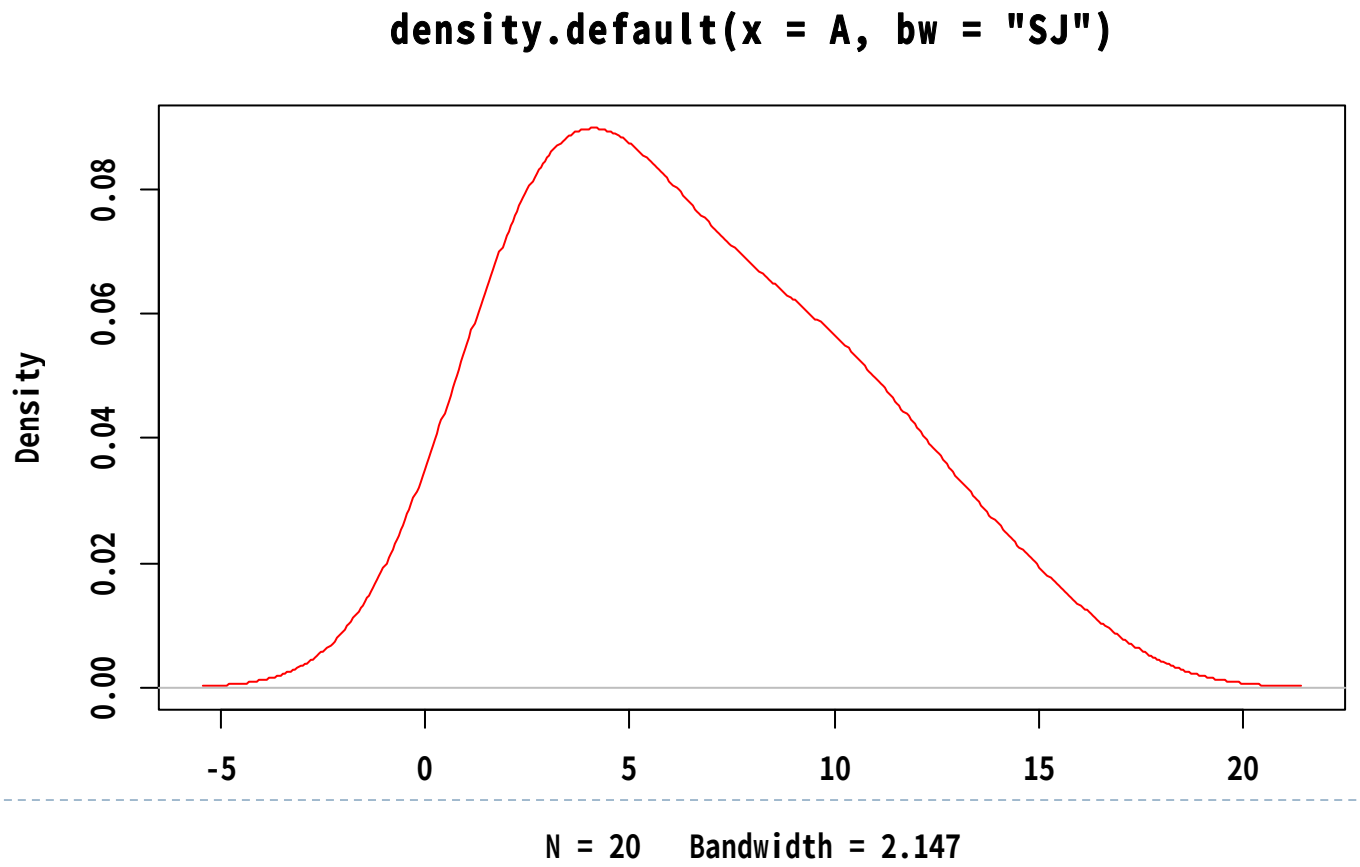
【参考】前の頁のグラフを作成するプログラム

```
> par(mfrow=c(2,2))                                    # 2×2に画面分割
> hist(A, breaks=seq(0,16,4), col="cyan")
> hist(A, breaks=seq(0,16,3), col="cyan")
> hist(A, breaks=seq(0,16,2), col="cyan")
> hist(A, breaks=seq(0,16,1), col="cyan")
```

薬剤 A の QOL スコアの密度推定

- ヒストグラムの代わりに密度推定曲線を描く（欠点は概ね解消）

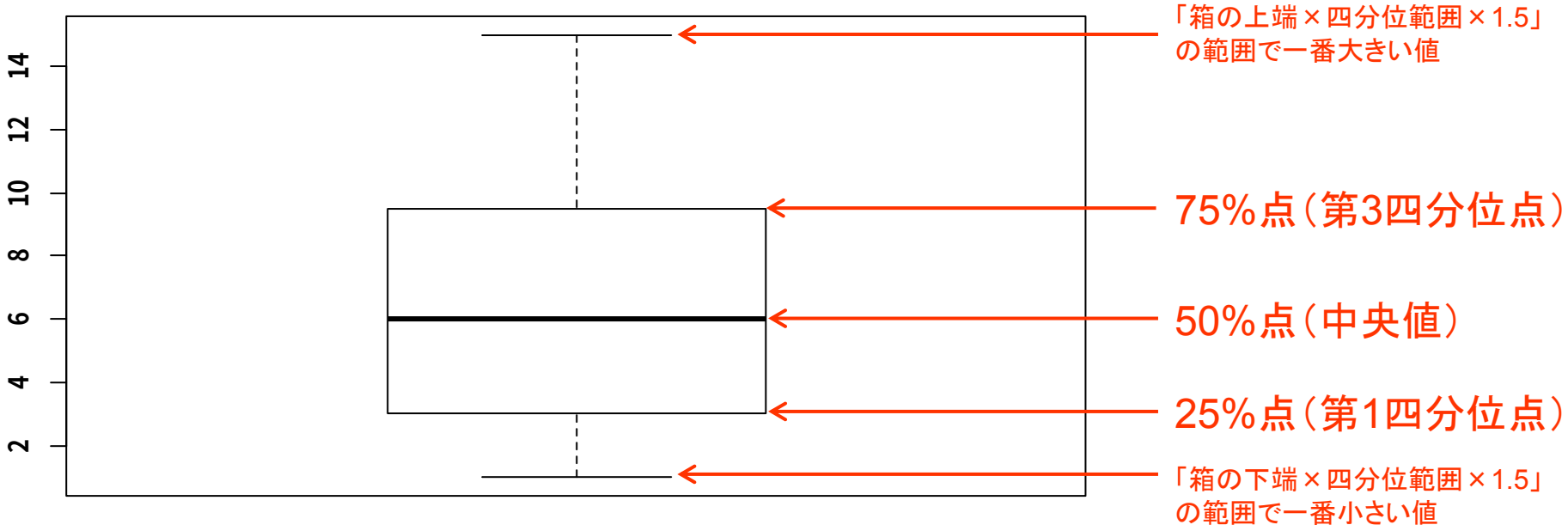
```
> plot(density(A, bw="SJ"), col="red")
```



薬剤 A の QOL スコアの箱ひげ図

- 要約統計量をグラフ化する場合は箱ひげ図

```
> hist(A, col="cyan")
```



本日のメニュー

1. データの読み込み
   - データ「DEP」の概要と読み込み
   - 薬剤 *A* の QOL のデータの取り出し
2. 1 つのデータの要約
   - 要約統計量の一覧
   - グラフの作成
3. 検定と信頼区間について

薬剤 A の QOL スコアに関する 1 標本 t 検定

- 薬剤 A の QOL スコアの平均が 4 であるかどうかを検定する
  - p = 1.12% なので結果は有意
  - 有意なので QOL スコアの平均は 4 ではない

```
> t.test(A, mu=4)

        One Sample t-test
data:  A
t = 2.809, df = 19, p-value = 0.0112 ← 検定結果（ p 値 = 約 1 %）
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.637202 8.362798
sample estimates:
mean of x
      6.5
```

疑問

- 「検定」って何？
- 「p　1%（0.0112）」の「p」って何？
- 「有意」って何？
- どうして「有意」になったら
  「薬剤 A の QOL スコアは 4 ではない」となるの？

検定の手順

1. 比較の枠組みを決める
2. 比較するものの間に差がないという仮説（帰無仮説 H_0）を立てる
3. 帰無仮説とは裏返し（差がある）の仮説（対立仮説 H_1）を立てる
4. 帰無仮説が成り立つという条件の下で，手元にあるデータ（よりも極端なこと）が起こる確率（ p 値）を計算する
5. 計算した確率が非常に小さい場合は「珍しいデータが得られた」と考えるのではなく「そんな珍しいことは通常起こらない・・・」
   ⇒「帰無仮説 H_0（差がないという仮説）自体が間違っている」と考え，対立仮説 H_1 が正しいと結論付ける
6. 計算した確率が小さくない場合は「帰無仮説 H_0 が間違っている」といえないので「帰無仮説 H_0 が間違っているとはいえない」と考える

検定の手順（薬剤 *A* の QOL スコアの場合）

1. 比較の枠組み ⇒「薬剤 *A* の QOL スコア」と「4」を比較する
2. 比較するものの間に差がないという仮説（帰無仮説 H_0）を立てる
   ⇒ 帰無仮説 H_0 ：薬剤 *A* の QOL スコア = 4 である
3. 帰無仮説とは裏返しの仮説（対立仮説 H_1）を立てる
   ⇒ 対立仮説 H_1 ：薬剤 *A* の QOL スコア ≠ 4 である
4. 帰無仮説が成り立つという条件の下で，手元にあるデータ（よりも極端なこと）が起こる確率（= p 値）を計算 ⇒ p = 0.0112（約 1 %）
6. 「確率が 1 %の珍しいデータが得られた」と考えずに
   「帰無仮説 H_0が間違っている」と考え，対立仮説 H_1 が正しいと結論
   ⇒「薬剤 *A* の QOL スコア ≠ 4 である」と結論付ける

疑問に対する回答

- 「検定」って何？

  ⇒ 前頁までの手順

- 「p　1%（0.0112）」の「p」って何？

  ⇒ 帰無仮説が成り立つという条件の下で手元にあるデータが起こる確率

- 「有意」って何？

  ⇒ p 値（帰無仮説が成り立つという条件の下で手元にあるデータ（よりも極端なこと）が起こる確率）が非常に小さい状態

- どうして「有意」になったら「薬剤 A の QOL スコアは 4 ではない」となるの？

  ⇒ p 値が非常に小さい場合は「珍しいデータが得られた」と考えずに「帰無仮説 H_0（差がないという仮説）が間違っている」と考える

検定のまとめ

- 「差がある」ことを証明する目的で「差がない」という帰無仮説 H_0 を設定する

  ⇒ 背理法の考え※

- p 値は「帰無仮説が成り立つという条件の下で，手元にあるデータ（よりも極端なこと）が起こる確率」

  ⇒ p 値が小さい場合（通常は 5% 未満）は帰無仮説 H_0 が誤りとする

- 逆に，p 値が小さくない場合（通常は 5% より大きい場合）は帰無仮説 H_0 が誤りではないとする

  ⇒ ややこしいが「帰無仮説 H_0 が正しい」とするのは間違い！あくまで p 値が小さい場合は背理法の考えが適用できるが，p 値が小さくない場合は背理法が成り立っていないので，何も結論は出ないことになる

※例：k と n は自然数とする ⇒ k^2 が奇数であるとき k も奇数となる」ことが示したい ⇒ 背理法で示す
⇒ k が偶数（=2n）と仮定すると $k^2 = (2n)^2 = 2 \times (2n^2)$ は偶数（k^2 は奇数のはず）⇒ 矛盾 ⇒ k は奇数

続・QOL スコアに関する 1 標本 t 検定

- 薬剤 A の QOL スコアの平均が 6 であるかどうかを検定する

  ⇒ 帰無仮説 H_0：薬剤 A の QOL スコアの平均が 6 である

  ⇒ p = 58% なので p 値は大きい（有意でない）

  ⇒「QOL スコアの平均は 6 ではないとはいえない」と結論

```
> t.test(A, mu=6)

        One Sample t-test
data:  A
t = 0.5618, df = 19, p-value = 0.5808 ← 検定結果（ p 値 = 58 %）
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval:
 4.637202 8.362798
sample estimates:
mean of x
      6.5
```

「平均は6である」
といってはダメ

続・QOL スコアに関する 1 標本 t 検定

- 平均が 7 であるかどうかの 1 標本 t 検定 ⇒ p = 58%（有意でない）
- 平均が 6 であるかどうかの 1 標本 t 検定 ⇒ p = 58%（有意でない）
- 平均が 5 であるかどうかの 1 標本 t 検定 ⇒ p = 11%（有意でない）
- 平均が 4 であるかどうかの 1 標本 t 検定 ⇒ p= 1%（有意）
- 平均が 3 であるかどうかの 1 標本 t 検定 ⇒ p = 0.001%（有意）
- 平均が 3 や 4 ではないようだが，5～7 ではないとはいえない（？）
  ⇒「いったい平均がどの位なのか」という情報は得られない

- 「QOL スコアの平均は●と▼の間にありそう」という情報が欲しい
  ⇒ 95%信頼区間の登場

薬剤 A の QOL スコアに関する 95% 信頼区間

- 薬剤 A の QOL スコアの平均が 4 であるかどうかを検定したときの結果を再度見てみる ⇒ 95%信頼区間が表示されている！

```
> t.test(A, mu=4)

        One Sample t-test

data:  A
t = 2.809, df = 19, p-value = 0.0112
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.637202 8.362798
sample estimates:
mean of x
      6.5
```

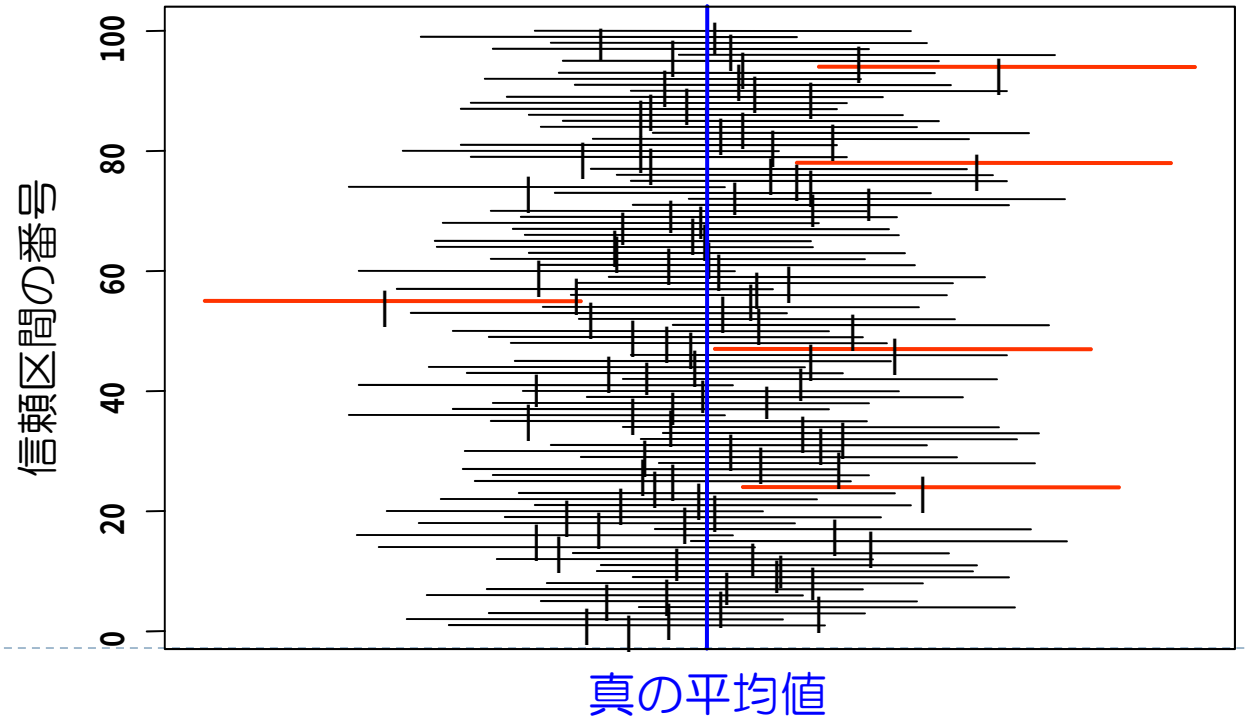
← 95%信頼区間：[4.63, 8.36]

薬剤 A の QOL スコアに関する 95% 信頼区間

- ちゃんとした「平均の 95% 信頼区間」の意味：

  「同じような状況で薬剤 A の QOL スコアの信頼区間を求める」ことを繰り返した場合，100 個の信頼区間のうち 95 個は真の平均値を含む

  ⇒「 [4.63, 8.36] は 95% の確率で真の平均値を含む」は間違い！



【参考】前の頁のグラフを作成するプログラム

```
> curve(dnorm(x, 6.5, 4), -3, 16)                    # 平均 6.5，標準偏差 4 の正規分布
> xvals <- seq(2.5, 10.5, length=50)                 # 領域をx軸方向に30個の多角形(台形)に等分割
> dvals <- dnorm(xvals, 6.5, 4)                      # 対応するグラフの高さ
> polygon(c(xvals,rev(xvals)),
+         c(rep(0,50), rev(dvals)), col="yellow")    # 塗りつぶす

> curve(dnorm(x, 6.5, 4), -3, 16)                    # 平均 6.5，標準偏差 4 の正規分布
> xvals <- seq(-1.5, 14.5, length=50)                # 領域をx軸方向に30個の多角形(台形)に等分割
> dvals <- dnorm(xvals, 6.5, 4)                      # 対応するグラフの高さ
> polygon(c(xvals, rev(xvals)),
+         c(rep(0,50), rev(dvals)), col="yellow")    # 塗りつぶす
```

薬剤 A の QOL スコアに関する 95% 信頼区間

- ちゃんとした「平均の 95% 信頼区間」の意味：
  「同じような状況で薬剤 A の QOL スコアの信頼区間を求める」ことを繰り返した場合，100 個の信頼区間のうち 95 個は真の平均値を含む
  ⇒ [4.63, 8.36] は 95% の確率で真の平均値を含む，という解釈は間違い！

- ただ，ちゃんとした定義で考えるとまどろっこしい場合が多いので，実用上は以下のようにざっくりと解釈する

- 薬剤 A の QOL スコアの平均の 95% 信頼区間は [4.63, 8.36]
  ⇒ ざっくりとした意味は「平均はだいたい 4.63～8.36 の間にある」
  ⇒「平均値が 6.5 である」という情報には「ばらつき」の情報が無いので「ばらつき」をふまえて区間で平均値の推定をする（区間推定）

薬剤 A の QOL スコアに関する 95% 信頼区間

「平均値が 6.5 である」という情報には「ばらつき」の情報が無い

- 「平均値が 6.5 である」ことが分かれば十分，という考えもあるが…

以下の 2 つの例を見てみる

1. 平均値が 6.5，95%信頼区間が [-30, 43]（信頼区間が広い場合）
   ⇒ 平均は -30～43 の間にあるといわれてもあまり有用な情報でない
   ⇒「平均値が 6.5」という値は精度が悪い ⇒ データ数が少ない？

2. 平均値が 6.5，95%信頼区間が [6.3, 6.7]（信頼区間が狭い場合）
   ⇒ 平均は 6.3～6.7 の間にあるという情報はかなり有用
   ⇒「平均値が 6.5」という値は精度が良い ⇒ 確証が持てる

雑談

QOL スコアに関する 1 標本 t 検定の場合については，以下が成り立つ

- 「 QOL スコアの平均の 95% 信頼区間」が「QOL スコアの平均と比較する値（ 4 とか 5 ）」を含んでいる場合は，1 標本 t 検定※の結果は有意にならない
- 「 QOL スコアの平均の 95% 信頼区間」が「QOL スコアの平均と比較する値（ 4 とか 5 ）」を含んでいない場合は，1 標本 t 検定※の結果は有意
- QOL スコアの平均の 95% 信頼区間は [4.63, 8.36] だが・・・
  - 「H_0：QOL スコアの平均 = 4.6」とした 1 標本 t 検定 ⇒ p = 0.04602（有意）
  - 「H_0：QOL スコアの平均 = 4.7」とした 1 標本 t 検定 ⇒ p = 0.05743 （有意でない）
  - 「H_0：QOL スコアの平均 = 8.35」とした 1 標本 t 検定 ⇒ p = 0.05144 （有意でない）
  - 「H_0：QOL スコアの平均 = 8.37」とした 1 標本 t 検定 ⇒ p = 0.04921 （有意）

※有意水準が 5% の場合

【参考】QOL スコアに関する 1 標本 Wilcoxon 検定

- 薬剤 A の QOL スコアの中央値が 4 であるかどうかを検定する

  ⇒ 帰無仮説 H_0：薬剤 A の QOL スコアの中央値が 4 である

  ⇒ p = 1.9% なので p 値は小さい（有意でない）

  ⇒「QOL スコアの中央値は 4 ではない」と結論

```
> wilcox.test(A, mu=4)

        Wilcoxon signed rank test with continuity correction

data:  A
V = 139.5, p-value = 0.01934← 検定結果（ p 値 = 58 %）
alternative hypothesis: true location is not equal to 4
```

本日のメニュー

1. データの読み込み
   - データ「DEP」の概要と読み込み
   - 薬剤 *A* の QOL のデータの取り出し
2. 1 つのデータの要約
   - 要約統計量の一覧
   - グラフの作成
3. 検定と信頼区間について

参考文献

- 統計学（白旗 慎吾 著，ミネルヴァ書房）
- The R Tips 第 2 版（オーム社）
- R 流！イメージで理解する統計処理入門（カットシステム）

Rで統計解析入門

終