R 言語の基本

操作の基本

- 使い方はテキストベースのコマンド
- 行の最後は「Enter」キー
- 入力中の修正は ← → キーで修正位置に移動可能
- ↑ ↓ キーで過去に入力したコマンドの呼び出し可能
- 無意味な文字列は「エラー」となる
- 1 行で完結しない場合はプロンプトが「+」に変化→複数行に分けた入力も可能。

「変数（オブジェクト)」の定義

- **変数 = データ表現**
- データ表現は「関数」，「式」，「数値」など
- = の代わりに <- （2 文字） を使ってもよい。
- **データ表現 ->**（2 文字） **変数** とすることも可能（向きと順序に注意）

変数やデータフレームの参照

- 変数名やデータフレーム名のみ ： 値・内容の表示・参照
- データフレーム名$カラム名 ： 特定のカラム（列）のみを表示・参照

主要な関数（使用したもの）

汎用

- floor ： **小数部の切り捨て**

分布・記述統計量（要約統計量）・統計的検定

- rnorm ： **正規分布に従った乱数の発生**
- mean ： **（算術）平均の計算**
- sd ： **標準偏差の計算**
- var.test ： **分散比の検定（分散の等値性を検証する F 検定）**
- t.test ： **平均値の差の検定（t 検定，オプションにより等分散を仮定した検定にも対応可能）**
- summary ： データの要約統計や，詳細な情報の表示

データ入出力

read.csv ： テキスト形式（CSV 形式）のファイルからのデータの読み込み（読込結果は「データフレーム」となる）

実行結果の処理

実行結果はテキストとして「コピー」できるので，そのまま Word などへ貼り付けることで利用できる。Word の「段落」パネルで改行ピッチ（行間）を「固定値」で「12pt」などにすることで，見やすく整形できる。フォントも等幅系（「MS ゴシック」や「MS 明朝」あるいは「Courier new」など）に変更しておくと，数字の位置も揃うので，さらに見やすくなる。

R による分析例（１）

初回の授業で宿題としてあった qa-01ex.pdf の第 6 問を R で処理してみる。

R を起動後，x1 と x2 という 2 種類のデータを，以下のように定義しておく。各行の先頭の「> 」は，プロンプトなので入力する必要はない。

```
> x1 = floor( rnorm( 30, mean=60, sd=10 ) + 0.5 )
> x2 = floor( rnorm( 30, mean=65, sd=15 ) + 0.5 )
```

それぞれ，1 学期の疑似データ（乱数を使った人工的なデータ）と 2 学期の疑似データとして使用する。内容を確認するには，それぞれの名前のみを入力する。青字は，R からの出力を示している（rnorm は乱数を発生するので，実行ごとに異なる結果が出てくるので厳密に一致している必要はない）。

```
> x1
[1] 76 73 70 54 55 54 63 71 46 55 68 53 74 59 58 63 47 50 56 53
[21] 63 53 51 72 50 58 60 65 57 59
> x2
[1] 81 66 77 68 37 73 95 52 52 22 64 61 71 87 89 48 63 61 62 54
[21] 85 77 38 66 63 87 63 46 50 85
```

データの平均・標準偏差を確認してみる。

```
> mean(x1)
[1] 59.53333
> sd(x1)
[1] 8.398413
> mean(x2)
[1] 64.76667
> sd(x2)
[1] 17.26804
```

平均値・標準偏差ともやや条件とは異なるが，この状態で検定を行なってみる（値が異なり過ぎると感じた場合には，rnorm によるデータの生成を何度か繰り返して，ほどよい疑似データになるように調整すればよい）。

まず，分散（標準偏差の平方値）の等値性を F 検定で確認する。

```
> var.test(x1,x2)

        F test to compare two variances

data:  x1 and x2
F = 0.23654, num df = 29, denom df = 29, p-value = 0.0002155
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval: 信頼区間の確率（1-α：有意水準の値）
 0.1125857 0.4969741 信頼区間
sample estimates:
ratio of variances
         0.2365421
```

検定対象の F 値（分散比）

p 値:（有意確率）
帰無仮説が有意な確率
→小さいと帰無仮説を棄却
→大きいと帰無仮説を採択

対立仮説の内容
（分散比が 1 に等しくない）

結果は，「帰無仮説：分散比が 1 である」を採択できるほど，p 値（有意確率）が大きく

はない（有意水準 5%を下回っている）ので，帰無仮説を棄却して（「対立仮説：分散比が 1 に等しくない」を採択して），「分散は等しくない」が妥当であることを示している。

信頼区間（有意水準）を変更して検証するには，conf.level= オプションを加えて実行する。標準では信頼区間の確率が 95%となっているので，90%に変えた例を示す。

```
> var.test(x1,x2,conf.level=0.9) # 信頼区間の値は小数で与える

        F test to compare two variances

data:  x1 and x2
F = 0.23654, num df = 29, denom df = 29, p-value = 0.0002155
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval: 信頼区間の確率(90%になっていることに注意)
 0.1271177 0.4401603 信頼区間(95%の時より狭まっている)
sample estimates:
ratio of variances
         0.2365421
```

続いて，平均値の差を検証するために，t 検定を行なう。分散が等しくない場合には，t.test をそのまま使用する。

```
> t.test(x1,x2)

        Welch Two Sample t-test

data:  x1 and x2
t = -1.4928, df = 41.992, p-value = 0.143
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.308356   1.841689
sample estimates:
mean of x mean of y
 59.53333  64.76667
```

検定対象の t 値

p 値:（有意確率）
帰無仮説が有意な確率
→小さいと帰無仮説を棄却
→大きいと帰無仮説を採択

対立仮説の内容
（平均の差が 0 ではない）

仮に，分散が等しいと判定された場合には，var.equal= オプションを加えて実行する。

```
> t.test(x1,x2,var.equal=TRUE)

        Two Sample t-test

data:  x1 and x2
t = -1.4928, df = 58, p-value = 0.1409
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.250949   1.784282
sample estimates:
mean of x mean of y
 59.53333  64.76667
```

いずれの結果も，p 値（有意確率）が有意水準 5%（信頼区間の確率 95%）より大きいの

で,「帰無仮説：平均の差が 0 である」を採択して（「対立仮説：平均の差が 0 ではない」を棄却して），平均は差がない（1 学期の平均点と 2 学期の平均点には「有意に差があるとは認められない」）と判定することになる。

なお，解答例として示した qa-ex03.xlsx の「平均値の差の検定」シートにあるデータを使った実行例を以下に示す（データは授業の HP に置いてあるものを URL で指定して，直接読み込ませている）。

```
> Q6 = read.csv("http://sakkun.cc.yokohama-cu.ac.jp/text/soc/Q6.csv")
> head(Q6)
  No x1 x2
1  1 70 51
2  2 70 80
3  3 70 80
4  4 69 79
5  5 51 51
6  6 51 50
> mean(Q6$x1)
[1] 60
> sd(Q6$x1)
[1] 10.04129
> mean(Q6$x2)
[1] 65
> sd(Q6$x2)
[1] 14.99195
> var.test(Q6$x1,Q6$x2) # 分散比の検定

        F test to compare two variances

data:  Q6$x1 and Q6$x2
F = 0.4486, num df = 29, denom df = 29, p-value = 0.03465 # 有意に 1 ではない
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2135196 0.9425148
sample estimates:
ratio of variances
         0.4486039

> t.test(Q6$x1,Q6$x2) # 分散が等しくないとされたので Welch の t 検定

        Welch Two Sample t-test

data:  Q6$x1 and Q6$x2
t = -1.5177, df = 50.66, p-value = 0.1353 # 「帰無仮説：平均差が 0 である」を採択
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.614801   1.614801
```

```
sample estimates:
mean of x mean of y
       60        65
```

この場合も，結果として，平均点に「有意な差があるとは認められない」とみるのが妥当な結果となっている。なお，qa-ex03.xlsx に示した t 検定では「5 点差がある」ことを検定しており，結果として「5 点差がある」ことを採択できないので，「5 点差があると有意には言えない」結果となっている。同様の検定をするには，mu=オプションを加えて次のように実行する。

```
> t.test(Q6$x1,Q6$x2,mu=5) # 5 点差であることが妥当かどうかを検定

        Welch Two Sample t-test

data:  Q6$x1 and Q6$x2
t = -3.0355, df = 50.66, p-value = 0.003786 #「対立仮説：5 点差である」を採択
alternative hypothesis: true difference in means is not equal to 5
95 percent confidence interval:
 -11.614801   1.614801
sample estimates:
mean of x mean of y
       60        65
```