医学統計学演習：資料 3

芳賀昭弘 *

1　復習

R を起動し、まずは作業ディレクトリをデスクトップの（これまで作業してきたディレクトリ）フォルダに設定してください。
manaba から Ohtani.csv のファイルをダウンロードしてそのフォルダに入れてください。次のようにデータを読み込んでみよう。

```
> data = read.csv(“Ohtani.csv”)
> NLB = subset(data, NLBorMLB == “N”)
> MLB = subset(data, NLBorMLB == “M”)
> data
> NLB
> MLB
```

まず Ohtani.csv を R で読み込んで data という名前で格納する。そのうち NLBorMLB の項目で N となっている方を NLB、M となっている方を MLB として分けて格納する。これは野球の大谷翔平選手の日本のプロ野球（NLB）とアメリカ・メジャー（MLB）での年度別の主要成績をまとめたデータです。これを使って簡単な統計量の演算を本日は行います。
では、その中の列項目 Year を横軸に、Average を縦軸にデータをプロットしてみましょう。

```
> plot(NLB$Year,NLB$Average,“l”)
> plot(MLB$Year,MLB$Average,“l”)
```

Average（打率）の年次推移がプロットされたでしょうか。Average は Hits（ヒット数）/AtBats（打数）で計算できます。実際にそのようになっているか R で計算させて確かめてみよ。（ちなみに x を有効数字 3 桁で表示させるには format(x,digit=3) だったのを思い出そう。）

例題：NLB と MLB での年次毎のホームラン数の推移を図示せよ。

* Electronic address: haga@tokushima-u.ac.jp

2　データの代表値の計算

2.1　平均 mean

大谷の日本とアメリカでの年間のホームラン数の平均値を算出してみよう。

```
> mean(MLB$HomeRuns)
> mean(NLB$HomeRuns)
```

例題：１ホームラン（HomeRuns）あたりに要した打数（AtBats）の年度別のデータと MLB, NLB ごとの平均値を求めよ。

上で述べた平均は、算術平均と言われるもの。平均と言えば、普通は算術平均をいう。しかし、平均にはこの他にも幾何平均、重み付け平均などがある。

算術平均；

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

幾何平均 (geometric mean)；

$$m_g = (\prod_{i=1}^{n} x_i)^{1/n} \tag{2}$$

重み付け平均；

$$E_w[x] = \frac{1}{n}\sum_{i=1}^{n} w_i x_i \tag{3}$$

w_i は x_i の重み。

2.2　中央値 median 及び最頻値

中央値はデータを小さい順（もしくは大きい順）に並べた時に中央に位置する値（データ数が偶数の場合は、中央２つの平均値）のこと。

```
> median(MLB$HomeRuns)
> median(NLB$HomeRuns)
```

ヒストグラムのデータで最大度数（最大頻度）を持つ階級値のことを、最頻値という。

2.3　最大値 max と最小値 min

```
> max(MLB$HomeRuns)
> max(NLB$HomeRuns)
> min(MLB$HomeRuns)
> min(NLB$HomeRuns)
```

2.4 分散・不変分散と標準偏差

ホームラン数の不偏分散と標準偏差はどうなっているでしょう？

```
> var(MLB$HomeRuns)
> sd(MLB$HomeRuns)
```

R の関数 var と sd は、不偏分散と不偏標準偏差を与える。標本分散と標本標準偏差を与えるには、前に作成した varp 関数を利用する。

```
> source("varp.R")
> varp(MLB$HomeRuns)
> sqrt(varp(MLB$HomeRuns))
```

R では、summary 関数によるデータの代表値の計算が用意されている。summary 関数では、最大・最小値、中央値・平均値の他、不偏標準偏差 σ と 3σ の値が計算される。

```
> summary(NLB)
> summary(MLB)
```

2.5 共分散 cov

共分散 (covariance) は 2 つのデータ間の相関（関係性）を見る時に有効な指標となる。不偏共分散は、

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{4}$$

で与えられ、R では不偏共分散を計算するための cov 関数が用意されている。

```
> cov(MLB$HomeRuns, MLB$Average)
```

標本共分散 (covariance)；

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{5}$$

は、R では特定の関数を用意していないが、次のように cov 関数の値から簡単に変換できる。

```
> cov(MLB$HomeRuns, MLB$Average) * (length(MLB$HomeRuns)−1)/(length(MLB$HomeRuns))
```

例題：標本共分散を与える新しい関数を作成せよ。

共分散が正のとき正の相関（一方が大きくなると一方も大きくなる）、負の時は負の相関（一方が大きくなると一方は小さくなる）があると言える。しかし、その値の大きさは、データで扱っている値の大きさに依存してしまうため、相関を見る場合には次の相関係数を用いる。

2.6 相関係数 cor

相関係数（correlation coefficient）は２つの変量の関係を表す統計量で、

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{6}$$

と定義される。R では cor 関数が用意されている。

```
> cor(MLB$Hits, MLB$Average)
> cor(MLB$Hits, MLB$HomeRuns)
```

安打数と打率は相関がありそう。ホームラン数とも相関がある？

3 演習

1. 前回、前々回に使用した EOM2.csv を R で読み込み、投与前 before と投与後 just after、投与後１時間 one hour after それぞれの算術平均、中央値、最大値、最小値、不変分散と標本分散、不変標準偏差と標本標準偏差を求めよ。
2. EOM2.csv の投与前（Before）と投与後（Just.after）の相関係数を求めよ。
3. 上記 1 と 2 を出力するスクリプトを作成せよ。
   ヒント: for ループを使うと便利。例えば data の２列目から n 列目までそれぞれの平均を表示させるには、

```
for ( j in 2:n) {
    Data_mean <- mean(data[,j])
    cat("Mean:",Data_mean,\n)
}
```

   とする。この出力 (print(Data_mean)) 前に、何の平均なのかが出力させよう（colnames(data)[j] で項目名を取り出すことができる)。このループの中で平均だけでなく計算したい統計量を出力させるように書くと良い。
   最後の問題で作成したスクリプトは、manaba のレポートシステムに貼り付け、解説してください。