応用統計学 2018 第1回

2018 年 10 月 3 日 (水) ver. 1
清 智也 sei@mist.i.u-tokyo.ac.jp
http://www.stat.t.u-tokyo.ac.jp/~sei/lec-j.html

注意：次週以降は資料を配布しません。講義ページから各自で事前にダウンロードしてください。

統計学は，大きく分類すると

- データを代表値や図表によって分かりやすい形に整理すること
- モデルを立て推論すること

の 2 つに分けられる。前者は記述統計，後者は推測統計と呼ばれる。この講義では，はじめに基本的な記述統計の手法を概観した後，確率論に基づく推測統計の手法を学習する。

1 記述統計の基礎

1.1 データ

データとは，広義には「推論の根拠となる資料」を指し，狭義には「データ行列」を指す。データ行列とは各行が個体，各列が変量[1]を表す行列であり，例えば表 1 のようなものである。この講義ではデータと言ったら通常はデータ行列を意味するものとする。

表 1: データの例

ゲーム名	発売年	発売元	スクロール方向	ステージ数
ゼビウス	1983	ナムコ	縦	16
スターフォース	1984	テーカン	縦	25
1942	1984	カプコン	縦	32
ツインビー	1985	コナミ	縦	5
グラディウス	1985	コナミ	横	7
沙羅曼蛇	1986	コナミ	縦横両方	6
ダライアス	1986	タイトー	横	7
R-TYPE	1987	アイレム	横	8

資料：Wikipedia 記事「シューティングゲーム」，一部のみ

[1]変量は変数とほぼ同義語である。ランダムネスを意識した場合には変量と呼ぶことが多い。

データ行列は，線形代数学で扱う意味での行列と比較すると，量的変数 (quantitative variable) と質的変数 (qualitative variable) という 2 つのオブジェクトが同時に含まれるという点で異なっている。表 1 の例では発売年とステージ数が量的変数であり，ゲーム名，発売元，スクロール方向が質的変数である。

データ行列を扱うには何らかのソフトウェアを用いるのが便利である。この講義ではソフトウェアとして主に R を利用するが，各人の環境や興味に合わせてどんなものを用いてもよい。

上述のデータは講義ページに CSV ファイルとしてアップロードされている。ただし漢字コードの問題を避けるため，変数名や質的変数を ASCII コードで表している。このファイルを R で読み込むと，data.frame というクラスのオブジェクトが生成される：

```
> X = read.csv("lec01-game.csv")
> X
       title year company  direction stage
1    XEVIOUS 1983   NAMCO   vertical    16
2 STAR FORCE 1984  TEHKAN   vertical    25
3       1942 1984  CAPCOM   vertical    32
4    TwinBee 1985  KONAMI   vertical     5
5    GRADIUS 1985  KONAMI horizontal     7
6 SALAMANDER 1986  KONAMI       both     6
7     DARIUS 1986   TAITO horizontal     7
8     R-TYPE 1987    IREM horizontal     8

> class(X)
[1] "data.frame"
```

data.frame オブジェクトの要約を表示したい場合は，summary という関数が便利である。量的変数の場合は平均値や四分位点（後述）が表示され，質的変数の場合は度数分布（後述）が表示される。

```
> summary(X)
       title        year        company       direction     stage
 1942      :1   Min.   :1983   CAPCOM:1   both      :1   Min.   : 5.00
 DARIUS    :1   1st Qu.:1984   IREM  :1   horizontal:3   1st Qu.: 6.75
 GRADIUS   :1   Median :1985   KONAMI:3   vertical  :4   Median : 7.50
 R-TYPE    :1   Mean   :1985   NAMCO :1                  Mean   :13.25
 SALAMANDER:1   3rd Qu.:1986   TAITO :1                  3rd Qu.:18.25
 STAR FORCE:1   Max.   :1987   TEHKAN:1                  Max.   :32.00
 (Other)   :2
```

1.2　1 変数データの記述

1 つの量的データ $\boldsymbol{x} = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ を考える。ここで記号 $'$ はベクトルの転置を表す。$\boldsymbol{x}$ の代表値として，例えば以下のものが挙げられる。

- 平均 (mean)：
$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$
- 分散 (variance)：
$$\sigma^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$
- 標準偏差 (standard deviation)：
$$\sigma = \Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2\Big)^{1/2}$$
- 歪度（わいど, skewness）：
$$\frac{1}{\sigma^3}\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^3\Big)$$
- 尖度（せんど, kurtosis）：
$$\frac{1}{\sigma^4}\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^4\Big) - 3$$

これらの代表値はいずれも $t(x_1, \ldots, x_n)$ という形の関数として表されることに注意しよう。このようにデータのみに依存する関数のことを統計量 (statistic) と呼ぶ。

表 2 は，表 1 のデータのうち量的変数に対する代表値をまとめたものである。

表 2: 量的変数の代表値

代表値	発売年	ステージ数
平均	1985	13.25
分散	1.50	90.35
標準偏差	1.22	9.50
歪度	0.00	0.97
尖度	−1.00	−0.65
最小値	1983	5.00
第 1 四分位点	1984	6.75
中央値	1985	7.50
第 3 四分位点	1986	18.25
最大値	1987	32.00

この表にある最小値，最大値の意味は明らかであろう。これに対し中央値 (median) や四分位点 (quartile) は以下に述べる分位点 (quantile) によって定義される。分位点の正確な定義は一通りではないが，実質的な定義は以下のように一義的に述べられる。まず，$x_1, \ldots, x_n \in \mathbb{R}$ が与えられたとき，

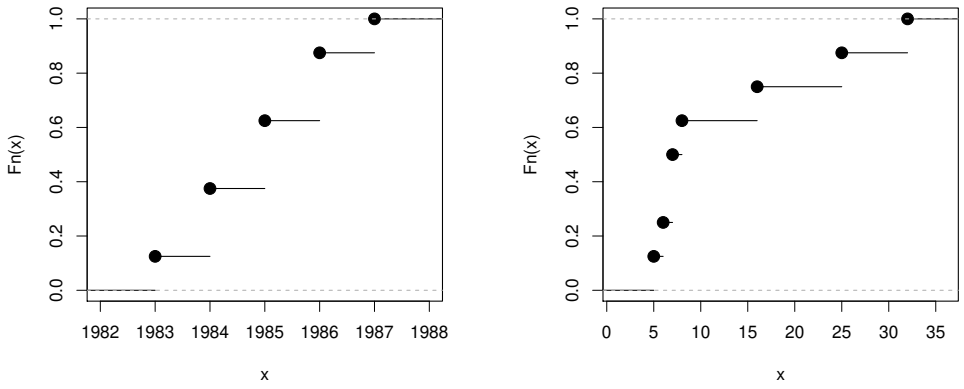
$$\hat{F}_n(x) = \frac{1}{n} \sum_{t=1}^{n} I_{\{x_t \leq x\}}, \quad x \in \mathbb{R},$$

と定義される $\mathbb{R}$ 上の関数のことを経験分布関数 (empirical distribution function) と呼ぶ。このとき，確率 $u \in [0, 1]$ に対する分位点は

$$\hat{F}_n(x) \approx u \tag{1}$$

を満たす x として定義される。第 1 四分位点は $u = 0.25$ に対する分位点であり，中央値は $u = 0.5$ に対する分位点，第 3 四分位点は $u = 0.75$ に対する分位点である。式 (1) で用いた記号 $\approx$ の意味は「両辺がほぼ等しい」ということだが，これをぴったり等号にすることは（多くの場合）不可能である。その補正の方法が複数存在するのである。

先述のデータの場合，経験分布関数は図 1 のような関数となる。この関数と，先の表 2 に示した四分位点の値を比較されたい。

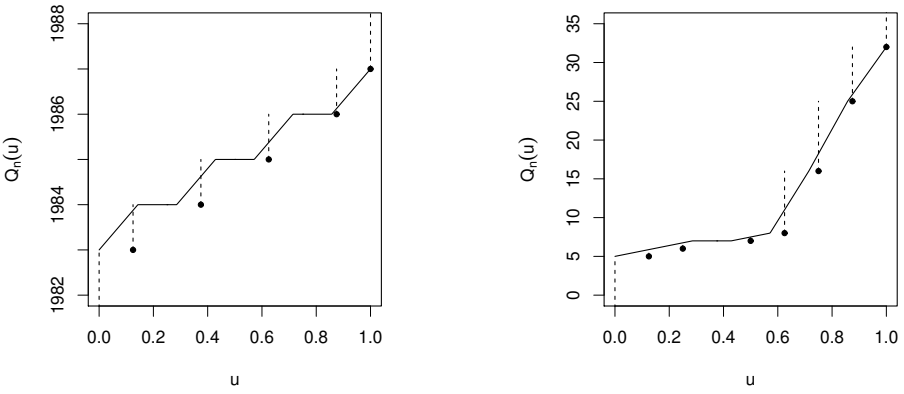


(a) 発売年　　(b) ステージ数

図 1: 経験分布関数

参考までに R の `quantile` 関数においてデフォルトで採用されている分位点の定義を与えておく。$(x_1, \ldots, x_n)$ を小さい順に並べ替えたものを $(x_{(1)}, \ldots, x_{(n)})$ と記す。このとき，確率 u に対する分位点を

$$\hat{Q}_n(u) = (1 - \gamma) x_{(j)} + \gamma x_{(j+1)}, \quad j = \lfloor nu + 1 - u \rfloor, \quad \gamma = nu + 1 - u - j, \tag{2}$$



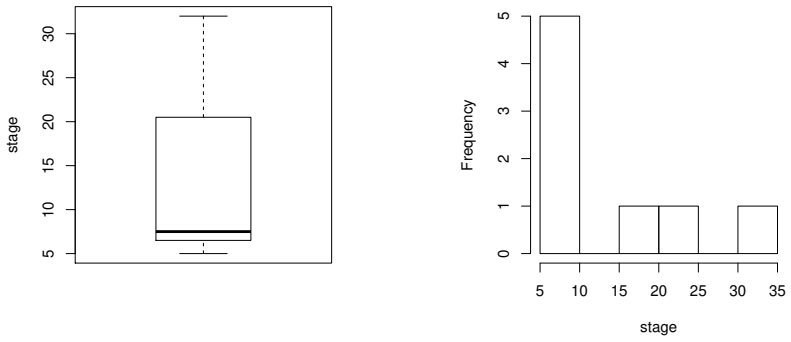
(a) 発売年　　(b) ステージ数

図 2: 式 (2) で定義される分位点関数。点線は経験分布関数を反転させたもの。

と定義する。 図 2 は，図 1 の経験分布関数の横軸と縦軸を反転して得られるグラフに，式 (2) で定義される分位点関数を重ね描きしたものである。図から，分位点関数とは概ね経験分布関数の逆関数であることが見て取れる。

四分位点を元に描かれるのが箱ひげ図 (box plot) である（図 3 (a)）。「箱」の下端は第 1 分位点，上端は第 3 四分位点であり，箱の内側に引かれた横線は中央値を表す。「ひげ」は最大値と最小値まで引かれるのが基本であるが，R では箱の両端から「箱の高さの 1.5 倍」を超える位置にある観測値を外れ値として明示する。図 3 (a) では外れ値は検出されておらず，結果として最大値・最小値までひげが引かれている。なお，R では箱ひげ図を描くときの四分位点の定義が式 (2) と異なるため，表 2 の値とは若干異なっている。

ヒストグラム (histogram) は同図 (b) のように，いくつかの等間隔な区間（階級）ごとに観測値の度数を数えて棒グラフにしたものである。視覚的に分かりやすいが，区間の選び方によって見え方が大きく変わってしまうという問題点がある。



(a) 箱ひげ図　　(b) ヒストグラム

図 3: ステージ数の箱ひげ図とヒストグラム

さて，これまでは 1 変数の量的データの記述を扱ってきた。一方で，質的データの記述は度数分布表 (frequency table)，すなわち各値が観測された回数をまとめるくらいしかバリエーションがない。これは，裏を返せば，量的データの定義域である実数体 $\mathbb{R}$ が四則演算と大小関係（と完備性）という豊かな構造を持っているためである。次節で述べるように，質的変数は多変数データの層別において役立つことが多い。

1.3 多変数データの記述

2 つの量的データ $\boldsymbol{x} = (x_1, \ldots, x_n)'$, $\boldsymbol{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ を考える。このような 2 変数データについては，まず散布図 (scatter plot) を描くことが大事である。散布図とは，(x_t, y_t) のペアを座標平面にプロットして得られる図のことである（図 4)。

代表値としては共分散と相関係数が基本的である。

- 共分散 (covariance)：

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y}).$$

  ここで $\bar{x}$ は $\boldsymbol{x}$ の平均である。

- 相関係数 (correlation coefficient)：

$$\mathrm{Corr}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathrm{V}(\boldsymbol{x})}\sqrt{\mathrm{V}(\boldsymbol{y})}}.$$

  ここで $\mathrm{V}(\boldsymbol{x}) = \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{x})$ は $\boldsymbol{x}$ の分散である。相関係数は $\mathrm{V}(\boldsymbol{x}), \mathrm{V}(\boldsymbol{y}) > 0$ のときに限って定義され，-1 から 1 の範囲に値を取る（コーシー・シュワルツの定理より)。

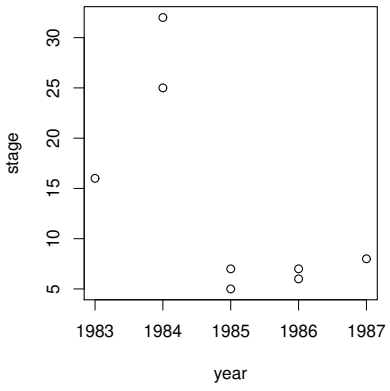


図 4: 発売年とステージ数の散布図（相関係数 -0.64）

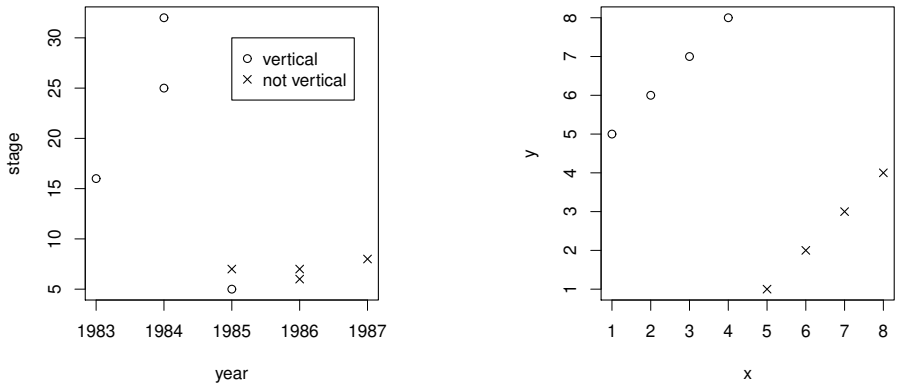
次に 2 つの質的データ $\boldsymbol{x}, \boldsymbol{y}$ の記述を考える。これについては分割表（contingency table, クロス集計表）が基本的である。分割表とは，フォーマルに定義するならば以下のようになる。まず第 1 変数の定義域を $\{1, \ldots, I\}$，第 2 変数の定義域を $\{1, \ldots, J\}$ に対応させる。次に $(x_t, y_t) = (i, j)$ となる t の個数を n_{ij} とおく。このようにして作られた $I \times J$ 行列 (n_{ij}) が分割表である。実際には表 3 のように元の定義域の値や行和・列和を付記することが多い。R では `table` という関数を使って求めることができる。変数が 3 つある場合は 3 元分割表と言って，添字が 3 つある配列に対応する（4 つ以上も同様）。

表 3: 分割表

	カプコン	アイレム	コナミ	ナムコ	タイトー	テーカン	合計
横	0	1	1	0	1	0	3
縦	1	0	1	1	0	1	4
両方	0	0	1	0	0	0	1
合計	1	1	3	1	1	1	8

最後に層別とシンプソンのパラドックスについて説明する。

層別 (stratification) とは，特定の質的変数の値によってデータを分けることである。図 5(a) は，図 4 の散布図を他の変数（ここではスクロール方向）によって層別して描き直した図である。もともとの相関係数は -0.64 であったが，層別によって縦スクロールのゲームの相関係数は -0.38，他のゲームの相関係数は 0.50 となった。



(a) 散布図の層別　　(b) シンプソンのパラドックス

図 5: データの層別とシンプソンのパラドックス

さらに同図 (b) のような極端な例を考えれば，層別によって相関係数の符号が変わってしまうことが分かる。このような現象はシンプソンのパラドックス (Simpson's paradox) と呼ばれ，因果関係を探る際に注意を払うべき点の一つとして知られている。もしこの図

の横軸が「ある薬の投与量」，縦軸が「投与による血圧の増加量」，層別する変数が「ある病気の有無」だった場合の解釈を考えてみるとよい。

層別は量的変数をもとに行うこともある。再び図 4 の散布図を考えると，明らかに 1984 年以前と 1985 年以降でステージ数の変化が見られる[2]。そこでこれらの期間によってデータを 2 つに層別し，各期間におけるステージ数の変動係数（標準偏差を平均値で割った値）を求めるとそれぞれ 0.33 と 0.17 となる。これは層別しない場合の変動係数 0.77 に比べて小さくなっている。いまの例は単にゲームデータの特徴を調べたに過ぎないが，製品の性能のばらつきを抑えたい品質管理などの文脈では，層別による変動係数の変化を観察することは重要な視点の一つである。

用語のまとめ

- 1 変数：平均，分散，標準偏差，度数分布表，経験分布関数，分位点関数，中央値，四分位点，箱ひげ図，ヒストグラム。
- 多変数：共分散，相関係数，散布図，分割表，層別，シンプソンのパラドックス。

演習問題

問題 1.1. 次のデータは，あるクラスの 23 人の学生に 1 円玉の所持枚数を聞いた結果である。枚数の平均，中央値，度数分布表を求めよ。また箱ひげ図を描け。

3, 3, 2, 5, 0, 6, 3, 0, 4, 7, 2, 0, 1, 33, 0, 2, 3, 13, 2, 14, 4, 13, 7

問題 1.2. 平均，分散，標準偏差，歪度，尖度は，経験分布関数が与えられれば一意的に定まることを示せ。

問題 1.3. R の関数 `boxplot` で採用されている四分位点の定義を調べよ。

問題 1.4. 次の 3 元分割表に対し，ベクトル $(a_1, b_1), (c_1, d_1), (a_2, b_2), (c_2, d_2)$ を 2 次元平面に描き，シンプソンのパラドックスが生ずる条件を幾何学的に説明せよ[3]。

n_{ij1}	1	2
1	a_1	b_1
2	c_1	d_1

n_{ij2}	1	2
1	a_2	b_2
2	c_2	d_2

ただし質的データにおけるシンプソンのパラドックスとは次のような場合を指す：

$$a_1 d_1 > b_1 c_1, \quad a_2 d_2 > b_2 c_2, \quad (a_1 + a_2)(d_1 + d_2) < (b_1 + b_2)(c_1 + c_2).$$

[2] と言っても個体の数が 8 個と少ないのであまり信頼性のない考察である。
[3] 宮川「統計技法」7 章。