データ分析基礎
確率・統計の基礎

京都大学 国際高等教育院 附属データ科学イノベーション教育研究センター

せきど ひろと
關戸 啓人

sekido.hiroto.7a@kyoto-u.ac.jp

統計と確率の基礎知識

統計とは

★ 統計学とは，大雑把に言って，データの扱い方を考える学問
- ★ データをどうやって得るか
- ★ データをどうやって解析するか
- ★ データを使って何か主張できるか

★ なんとなく，ではなく，できるだけ数学的（客観的）に議論する

★ 確率論を道具として用いることが多い
- ★ データには誤差が付きもの（データを取る対象の選び方・測定誤差など）
- ★ 正しいモデルは不明．適当な近似を行いモデル化（モデル化誤差）
- ★ 誤差の要因，振る舞いは複雑怪奇 $\rightarrow$ 確率的なものとして扱う
- ★ 知りたいことにフォーカスを当て，わからないことは確率で暈す

確率とは

★ ○○を行ったとき，△△が起きる確率

★ ○○を**試行**，△△を**事象**と呼ぶ

★ **確率は，試行を行ったとき，その事象の「起こりやすさ」を実数で表したもの**

★ P(事象) や P(事象 a が起こる) のように表す

★ 確率は0以上1以下の実数で値が大きいほど起こりやすい

★ 確率が p とは，十分大きな N に対して，試行を独立に N 回行ったとき，だいたい Np 回ぐらいその事象が起こるだろうと期待されるということ

確率とは

★ 確率として捉えるには

★ 6月28日に東京で雨になる確率は53%である

★ 気象データが残っている年を無作為に1つ選ぶとその年の6月28日の東京の天気が雨である確率が0.53である

★ 日本人成人男性のうち身長が190cm以上の人の割合は0.06%である

★ 日本人成人男性を無作為に1人選ぶと，その人の身長が190cm以上である確率が0.06%である

★ 日本人男性が最近の環境で育つと，その人の身長が190cm以上となる確率が0.06%である

★ 明日京都で雨が降る確率は10%である

★ 観測しうる範囲で，今日と同じ気候条件であれば，その次の日京都で雨が降る確率が0.1である

条件付き確率

★ 試行を行い，事象 B が起こったとき，事象 A が起こる確率 $P(A|B)$ または $P_B(A)$

- ★ A も B も起こる確率 ÷ B が起こる確率，$P(A \cap B)/P(B)$ が定義

★ 試行にある種の制限をかけていると思うこともできる

- ★ 今日と同じような気象条件になったとき，次の日が晴れる確率
- ★ ランダムに1日選び，選んだ日が今日と同じような気象条件であったときに，その次の日が晴れである確率

★「確率」を考えるときは，(あるならば) 得られている全ての情報を用い，条件付き確率を考える

余談1：モンティ・ホール問題

- ★ クイズ番組の優勝者が賞品を選ぶゲーム
- ★ ドアが3つあり，開けるとそのうちの2つにはヤギ（外れ），1つには車
- ★ 以下の手順を行う
  - ★ 優勝者は1つのドアを無作為に選ぶ
  - ★ 司会者はどれが外れか知っており，優勝者が選ばなかった外れのドアのうち1つを無作為に開く
  - ★ 優勝者は選ぶドアを変えることができる
  - ★ 最終的に優勝者が選んだドアに車があればもらえる
  - ★ 優勝者はどうすれば良いか？　車が貰える確率は？

余談1：モンティ・ホール問題

★ クイズ番組の優勝者が賞品を選ぶゲーム

★ ドアが3つあり，開けるとそのうちの2つにはヤギ（外れ），1つには車

★ 以下の手順を行う

- ★ 優勝者は1つのドアを無作為に選ぶ
- ★ 司会者はどれが外れか知っており，優勝者が選ばなかった外れのドアのうち1つを無作為に開く
- ★ 優勝者は選ぶドアを変えることができる
- ★ 最終的に優勝者が選んだドアに車があればもらえる
- ★ 優勝者はどうすれば良いか？　車が貰える確率は？

★ 適切な仮定のもとで…

- ★ 選ぶドアを変えなければ車の確率1/3
- ★ 選ぶドアを変えれば車の確率2/3

余談2：2つの封筒問題

★2つの封筒があり，中にはお金が入っている

★片方の封筒の中には片方の封筒の倍額入っているがどちらかわからない

★以下の手順を行う

- ★1つの封筒を無作為に選び，中に何円入っているか確認する
- ★選ぶ封筒を変更することができる
- ★最終的に選んだ封筒の中身がもらえる

★無作為に選んだから最初選ばなかった方に倍額入っている確率は1/2

- ★選ぶ封筒を変えなければ確認した額x円もらえる
- ★選ぶ封筒を変えれば貰える額の期待値は$\frac{1}{2}x + \frac{1}{2}(2x) = 1.25x > x$
  - ★何がおかしいか？

余談3：為替？

★ 日本円をx円持っていて，1ドルx円でドルを買った

★ 1ドルが$2x$円になるか，$x/2$円になったら1ドルを売る

★ 為替はランダムに動くとすると

- ★ 円に対してドルの価値が2倍になることも，ドルに対して円の価値が2倍になることも，どちらが先に起こるかは同じ確率
- ★ 1ドルが$2x$円になるか，1ドルが$x/2$円になるか，どちらが先に起こるかは同じ確率

★ 最終的に手元に残る日本円の期待値は$\frac{1}{2}\frac{x}{2} + \frac{1}{2}(2x) = 1.25x > x$

- ★ 何かおかしいのか？

母集団と標本

- ★ **母集団**
  - ★ 知りたい調査対象全体
  - ★ （確率）分布で表される
- ★ **標本**
  - ★ 母集団から抽出された集団
  - ★ n 個の実数（の組）で表される
    - ★ n は標本サイズ

確率変数とは

★ **確率変数**とは，試行を行うと値が定まるもの

- ★ サイコロを振ったときに出る目
- ★ 全日本人の中から無作為に１人選んだとき，その人の身長

★ どんな値を取りうるか，またそれぞれの値はどれぐらい起こりやすいか，を両方表している

- ★ **確率分布**

確率変数とは

★ 取る値が連続的なものと離散的なものがある

- ★ サイコロの出目は$1,2,3,4,5,6$のどれかなので離散的（飛び飛びの値）
- ★ 身長は「この値またはこの値または…（可算無限）」と言えないので連続的と見なすことが多い
  - ★ 少なくても連続的であると仮定して議論したほうが便利
  - ★ ちょうど値がxになる確率，ということを考えないものは連続的

★ 複数の確率変数を考えても1つの試行

- ★ 全日本人の中から無作為に1人選んだとき，その人の身長Hと体重W
- ★ この場合，身長と体重は無関係ではない
  - ★ 身長が$170\ \mathrm{cm}$のときに体重が$60\ \mathrm{kg}$以上である確率$P(W \geq 60 | H = 170)$，などと条件付き確率を考えることもできる

確率密度関数，確率質量関数

★ 連続的な確率変数 X は確率密度関数 $f(x)$ で記述

- ★ $\displaystyle\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1, \quad f(x) \geq 0$
- ★ $\displaystyle P(a \leq X \leq b) = \int_{a}^{b} f(x)\,\mathrm{d}x \quad (a \leq b)$

★ 離散的な確率変数 X は確率質量関数 $f(x)$ で記述

- ★ $\displaystyle\sum_{x} f(x) = 1, \quad f(x) \geq 0$
- ★ $P(X = a) = f(a)$：$f(a)$ はその確率変数が a となる確率

★ 2つ以上の確率変数を同時に考えれば，多変数関数になる

- ★ $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$

離散的な確率変数と確率質量関数の例

★（例）理想的なサイコロの出目の場合

★離散分布で，$f(1) = f(2) = f(3) = f(4) = f(5) = f(6) = 1/6$

★サイコロを振って出る目 X が4である確率は $P(X = 4) = f(4) = 1/6$．これは，何回もサイコロを振ると，その振った回数の $1/6$ 倍ぐらいの回数だけ4の目が出るという意味である．

★注意：通常サイコロは厳密に各目が出る確率が $1/6$ となるわけではないと思われるが，$1/6$ に非常に近いことが経験的に知られており，厳密にやるのは面倒，または，不可能なのでこういう風にしましょう，と考えている．

連続的な確率変数と確率分布関数の例

★（例）無作為に20歳の男性を1人選んだとき，その人の体重(kg)

★ 連続分布で，$f(60)$ とか $f(70)$ ぐらいが大きくて，そこから外れると小さくなるであろう

★ 注意：20歳の男性の人数は有限であるので，これは本当は離散分布であると思われる．しかし，厳密に体重を量れるわけでもなく，そもそも対象がかなり曖昧（今この瞬間の体重分布か？ それとも人間が通常の生活を20年行った後の体重の分布か？）また，実際に必要となる値は，体重が60kg台の人口比率 $P(60 \leq X < 70)$ などが多く，連続分布と考えたほうが便利な場合が多い．

★ 実際の現象をできるだけうまく説明できる，数学的に扱うのが便利，などを考えて，うまく設定する必要がある．

期待値，平均，分散，標準偏差

★ 確率[密度|質量]関数 $f(x)$ に対応する確率変数 X の関数 $s(X)$ の**期待値**

- ★ 連続分布の場合：$\mathrm{E}[s(X)] = \int_{-\infty}^{\infty} s(x)f(x)\,\mathrm{d}x$
- ★ 離散分布の場合：$\mathrm{E}[s(X)] = \sum_{x} s(x)f(x)$

★ **平均（mean, average），分散（variance），標準偏差（standard deviation）**

- ★ 平均：$\mathrm{E}[X]$
- ★ 分散：$\mathrm{V}[X] = \mathrm{E}[(X - E[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$
- ★ 標準偏差：$\sqrt{\mathrm{V}[X]} = \sqrt{\mathrm{E}[(X - \mathrm{E}[X])^2]} = \sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}$

★ 分散の項の等式は $\mathrm{E}[\alpha s(X) + \beta t(X)] = \alpha\mathrm{E}[s(X)] + \beta\mathrm{E}[t(X)]$ や $\mathrm{E}[1] = 1$ などから導かれる

期待値，平均，分散，標準偏差

★ 平均，分散などは，その確率分布を知るための手がかりとなることがある

★ もしくは，確率分布関数を直接的に扱うのが難しいが，平均や分散は単なる実数なので扱いやすい

★ **平均**：確率変数がだいたいどのぐらいの値になるかという目安の1つ．
$X_1, X_2, \ldots, X_N$ が独立で，それぞれその分布に従うなら，$(X_1 + X_2 + \ldots + X_N)/N$ は N が大きいとき，平均に近づく

★ **分散**：確率変数が平均からどれぐらい離れた値になるか，という目安の1つ．（平均からの距離の2乗）の平均，であるので，オーダーは「距離」の2乗

★ **標準偏差**：分散の正の平方根．オーダーは「距離」

母集団 (population) と標本 (random sample)

★ $X_1, X_2, \ldots, X_N$ が独立で，すべて同じ確率分布に従うとき，これをサイズ N の標本という

★ 独立であるとは，荒っぽく言うと，X_1 がとある値のとき，X_2 はとある値になりやすい，などといった関係がないという事．標本が従う確率分布を母集団分布という

★ (例) サイコロを N 回振るとき，i 回目に出た目の数を X_i で表すと，$X_1, X_2, \ldots, X_N$ はサイズ N の標本となる

★ 1回目のサイコロの出目は，2回目のサイコロの出目に影響しないであろうから，これらは独立である

★ **注意**：統計処理を行う際には，標本 X_i には，与えられたデータなどの実数が入っている．ただし，数学的に，標本や，その標本から得られる量はどのような性質があるかなどを議論したいときは，標本を確率変数と考えている

標本平均，標本分散，不偏分散

★ サイズ N の標本 $X_1, X_2, \ldots, X_N$ に対して，以下が定義される

★ **標本平均**：$\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_N}{N}$

★ 標本分散：$\dfrac{1}{N}\sum_{i=1}^{N}(\overline{X} - X_i)^2 = \overline{X^2} - (\overline{X})^2$

★ **不偏分散**：$\dfrac{1}{N-1}\sum_{i=1}^{N}(\overline{X} - X_i)^2$

★ **注意**：標本平均の期待値は母集団分布の平均に一致する．また，不偏分散の期待値は，母集団分布の分散に一致する．標本に対して分散を調べるときは標本分散を，標本を通じて，母集団分布の分散を調べるときは不偏分散を用いることが多い．

他の代表値（確率変数の大体の大きさを表す）

★ 中央値（median）

- ★ 標本（データ）に対して，小さい順に並び替えたときに真ん中の値（真ん中が2つあるなら，足して2で割る）
- ★ 確率変数 X に対して，$P(X \leq c) = 0.5$ なる c
  - ★ 連続分布に対して，

$$\int_{-\infty}^{c} f(x)\,\mathrm{d}x = 0.5$$

なる c

  - ★ 確率変数 X に対して，$P(X \leq c) = \alpha$ なる c を上側 α 点，もしくは，$100\alpha\%$ 点という

★ 最頻値（mode）

- ★ 標本に対して，$X_1, X_2, \ldots, X_n$ のうち，最も多く登場する値
  - ★ 実際には，度数分布において，最も該当する数が多い階級とすることが多い
- ★ 連続分布，離散分布に対して，$f(x)$ の最大値を取る x

Rを用いた演習

使用するデータ

★ 独立行政法人 統計センターが作成した**一般用ミクロデータ**を利用する
- ★ 全国消費実態調査（平成21年）詳細品目 全世帯
- ★ 集計表などから作成された各世帯に関する疑似データ

★ 以下より入手しました
- ★ https://www.nstac.go.jp/services/ippan-microdata.html

★ ただし，以下の前処理を行ったものを利用しています
- ★ Rの read.csv で簡単に読み込めるように最初の数行を削除
- ★ 後半の詳細品目に関するデータを削除（動作を軽くするため）

★ スライドでは細かい説明は省いている部分がありますので注意してください

データの読み込み

★ データを作業フォルダにおいてある場合は以下で読み込み完了

★ 作業フォルダなどについては **R 言語の基礎知識** のスライドを確認してください

```
> x <- read.csv("ippan_2009zensho_s_dataset.csv")
```

★ 読み込んだ内容を確認する次ページのようにすると良い（最初の6行のみ表示）

データの読み込み

```
> head(x)
  X3City T_SeJinin T_SyuJinin T_JuSyoyu T_Syuhi T_Age_5s T_Age_65
1      1         2          1         1       1        1        1
2      1         2          1         1       1        1        1
3      1         2          1         1       1        1        1
4      1         2          1         1       1        1        1
5      1         2          1         1       1        1        1
6      1         2          1         1       1        1        1
    Weight Y_Income L_Expenditure  Food Housing   LFW Furniture
1 895.2667     3917         201649 47756   16028  9652      6702
2 895.2667     6675         166381 34054    7416 26313     17062
3 895.2667     6706         259736 84501    1927 10082      6741
4 895.2667     2790         114511 41664     730 22358      5413
5 895.2667     2577         193505 56981    3779 28747      4812
6 895.2667     3452         152109 34924    3418  8131      4164
  Clothes Health Transport Education Recreation OL_Expenditure
1    8088    726     21546         0      14433          76719
2    6989   7637     20773         0      19048          27089
(以下略)
```

世帯年収の推定

★ データ数（標本サイズ，列の数）を調べるには以下のようにできます

★ 1列目だけ取り出してきて要素数を調べています

```
> length(x[,1])
[1] 45811
```

★ 日本全国にある世帯の中からランダムに45811世帯を抜き出してきて調査したと思うことにする

★ この標本から日本全国の世帯年収を推定してみよう

★ まず，世帯年収に関するデータを抜き出してみよう

★ 9列目にあることに着目して抜き出す方法

```
> z <- x[,9]
```

★ 列の名前 Y_Income に着目して抜き出す方法（上と同じ意味）

```
> z <- x$Y_Income
```

世帯年収の平均・分散・標準偏差を推定

★ 標本平均・不偏分散・標準偏差を求めるには次のようにします

```
> mean(z)
[1] 6401.921
> var(z)
[1] 14522100
> sd(z)
[1] 3810.787
```

★ 説明を読むと，単位は 千円 であることがわかるので，世帯年収の

- ★ 平均は約 640 万円
- ★ 分散は約 14522100 (千円)^2
- ★ 標準偏差は約 381 万円

★ と推定された

平均値と中央値の比較

★ 他の代表値の例として中央値も求めてみよう

★ 中央値は以下のように求めることができる

```
> median(z)
[1] 5504
```

★ 推定結果を見ると

★ 平均値は約 640 万円

★ 中央値は約 550 万円

★ と両方ともだいたいどのぐらいの値化という目安に使われるものだが約90万円のずれがある

★ このように，同じような意味でも，細かいニュアンスは違うので，**自分の目的にあった代表値を用いるのが大切**

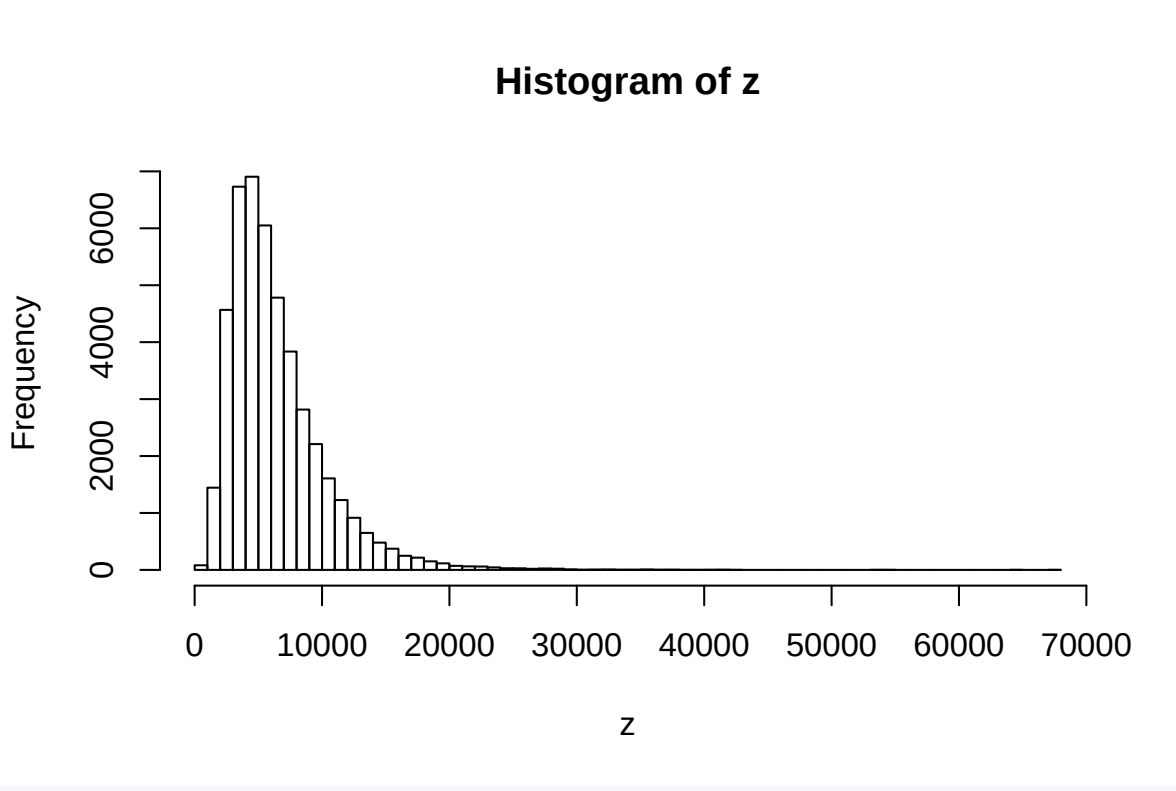
ヒストグラムの確認

★ 一番情報量が多いのは確率密度関数なので，それに対応するヒストグラムを確認してみよう

★ 代表値だけではわからないこともたくさんあるので，**分布を確認するのも重要**

★ Rでは例えば以下のようにすると良い

```
> hist(z, nclass = 50)
```



考察

★ ヒストグラムを確認すると左右対称とは程遠い

- ★ 年収が高い方に向けて緩やかに数は減るが，かなり高年収の世帯も相当数ある
  - ★ 平均値を押し上げるが，中央値にはさほど影響ない

★ 世帯年収や貯蓄などはこういう傾向があり，ときどき平均値は当てにならないので，中央値の方が直感に合うといわれる

- ★ 直感とは何なのか？（数学的・客観的な意味は？）
- ★ バイアスはないか？（その人の周辺だけの偏った事例から言ってないか？）

★ 兎にも角にも，多角的に捉えて，目的と状況に応じて意思決定などを行うべき