確率・統計の基礎知識

2018 年 前期講義資料 （作成者：大塚 淳）

1 記述統計

統計処理の第一歩は、観察・計測を繰り返して得られるデータを、わかりやすい形にまとめることである。これを記述統計 (descriptive statistics) という。例えば教室に n 人の学生がいるとして、その身長を変数 X で表すとしよう。それぞれの学生の身長を測定して得られたデータを、$x_1, x_2, \ldots, x_n$ と表すことにする。ただしここで x_i は i 番目の学生の身長であり、たとえば彼女が 155cm だったら $x_i = 155$ である。他方、別の変数 Y で年齢を表すとしたら、y_i は i さんの年齢（例えば $y_i = 23$）となる。変数（大文字）は観察される特徴のカテゴリーを表し、その値（小文字）は観察された特定の状態を表す。[*1]

1.1 単変数統計量

データは単に集めるだけではダメで、そこから有益な情報を引き出すためにはそれを適切な仕方で要約する必要がある。データを要約する種々の指標を、**統計量** (statistics) という。代表的な統計量として、平均、分散、標準偏差などがある。

変数 X の**標本平均** (sample mean) は、観測された X の値の総和を標本数 n で割ったものである

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_i^n x_i$$

標本平均はデータの「中心」を与えるという意味で、標本を要約するものである。

もう一つの指標は、**標本分散** (sample variance) であり、以下のように定義される

$$\mathrm{Var}(X) = \frac{1}{n}\sum_i^n (x_i - \bar{X})^2$$

分散は中心周りのデータのバラツキを示す。つまり、各データが中心近に固まって分布しているほど分散は小さくなり、逆に広範囲に散らばっていると大きくなる。

[*1] 哲学的に言えば、変数は determinable、変数の値は determinate に対応する。

バラツキの指標には、分散の平方根である**標準偏差** (standard deviation) を用いることもある

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{n}\sum_{i}^{n}(x_i - \bar{X})^2}$$

1.2 多変数統計量

二つ以上の変数があるとき、それらの間の関係性を知りたいことがある。例えば身長 X がどれくらい年齢 Y に伴って変化しているかは、その**共分散** (covariance) によって調べることができる：

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

共分散をそれぞれの変数の標準偏差で割ったものを、**相関係数** (correlation coefficient) という

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

相関係数はつねに $-1 \leq \mathrm{corr}(X,Y) \leq 1$ の範囲にある。相関係数が 0 以下のとき負の相関、0 以上のときは正の相関という。

ある変数 Y によって別の変数 X の変化を予測したいことがある。例えば、年齢が一つ上がるにつれ、平均身長はどれだけ上がる（あるいは下がる）のだろうか？これに応えるのが**回帰係数** (regression coefficient) であり、次によって表される

$$b_{x,y} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$$

これを X の Y への回帰係数とよび、Y の単位あたりの X の変化を表す。例えば上の例では、年齢 Y が 1 歳増えるごとに、身長 X は平均して $b_{x,y}$ だけ上がる。

1.3 離散変数

我々が観測する「特徴」は、必ずしも身長のように連続的な数で表される必要はない。例えば、コイン投げの結果を X としたとき表を 1、裏を 0 で表すことができる。この時、n 回のコイン投げの結果は 0,1 からなる数列 $(x_1, x_2, \ldots, x_n)$ で表される。ちなみにその平均 $\bar{X}$ は、n 回の試行のうち表が出た割合である。

2 確率

これまで、「確率」という言葉が出てこなかったことに注意してほしい。これはなぜかというと、単に観測されたデータ（標本）を記述・要約するためには、実は確率という概念は必要ないからだ。確率とはむしろ、データの背後にあって、我々がそこからデータを取ってくる源として想定されるような世界 — これを**母集団** (population) という — に属する概念なのだ。以下では、母集団と確率について、簡単に説明する。

2.1 標本空間

母集団あるいは**標本空間** (sample space; Ω) とは、関心のある試行や集団について、起こりうる・観測しうるすべての結果や個体の集合である。例えば、サイコロを 1 回投げる試行の標本空間は $\Omega = \{1, 2, 3, 4, 5, 6\}$ である。一方、選挙結果予想などで想定される標本空間は、投票した全有権者の集合である。

我々が事象 (event) と呼ぶのは、この標本空間の部分集合である。例えばサイコロを投げたとき偶数がでるという事象は $\{2, 4, 6\}$ で表すことができ、これは上述の Ω の部分集合になっている。

次にこの標本空間における任意の事象に対し、**確率** (probability) を割り当てる。[*2]確率とは、ありていに言えば標本空間全体に占める事象=部分集合の「大きさ」を測るもの (measure) であり、次の 3 つの公理を満たす関数として定義される。

1. 任意の事象 A について、$0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. 互いに排反な事象 $A_1, A_2, \ldots$ に対し、

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

この公理から、以下が導かれる（できれば証明してみよう）

4. $P(A^c) = 1 - P(A)$ （ただし A^c は A の補集合）
5. どんな事象 A, B であっても（それらが排反でなくても）

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

つまり「A もしくは B」の確率は、A の確率、B の確率を合わせたものから、ダブるところ（A かつ B である確率）を引いたものに等しい。なお以下では簡便のため $P(A \cap B)$ を $P(A, B)$ と記す。

2.2 条件付確率と独立

「B という事象が生じた条件のもとでさらに A という事象が生じる確率」を B のもとでの A の**条件付確率** (conditional probability) と呼び、以下のように定義する

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

一般に、条件付ける前と後では事象の確率は異なる。しかしこれが変化しない、つまり $P(A|B) = P(A)$ のとき、A と B は**独立** (independent) であるという。独立のとき、B について情報が得られたとしても A についてのことは何もわからない、つまり両者には関係がない。A と B が独立でないとき、両者は**従属** (dependent) であるという。

独立関係について、次が成り立つ（理由は上の定義から考えてみよ）。

[*2] 本当は標本空間上に代数を定義しなければならないのだが、ややこしくなるので割愛する。

- 対称性。つまり $P(A|B) = P(A)$ ならば $P(B|A) = P(B)$ であり、逆もまたしかり。
- A と B が独立ならば、$P(A, B) = P(A)P(B)$、つまり両者がともに成立する確率はそれぞれが成立する確率をかけたものである。

■全確率の定理　条件付確率の定義と公理 3 より、互いに排反かつ exhaustive な $B_1, B_2, \ldots, B_n$, such that $\bigcup_i^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$ に対し次が導かれる

$$P(A) = \sum_i^n P(A|B_i)P(B_i)$$

3　確率変数と分布

前述したように、事象は標本空間の部分集合である。よってあるさいころを振って偶数が出る確率は、$P(\{2, 4, 6\})$ と表記される。しかし例えば、日本国民全員からなる標本空間の中から 18 歳以上の有権者のみを抜き出したいときなど、いちいち個々の要素を列挙していたのでは面倒である。そこで、標本空間中の興味ある事象を抜き出す方法として、**確率変数** (random variables) を導入する。確率変数とは、上述の変数と同じように、対象の性質を表す。Y を年齢を表す確率変数とすると、「18 歳以上の国民」という部分集合は、$Y \geq 18$ と表される。よってその確率は $P(Y \geq 18)$ である。また X で身長を表すとすると、$P(X = 165)$ は身長が 165cm であることの確率である。一般に、ある確率変数 X について、その値が x を取る確率は $P(X = x)$ で表される。このとき $P(X)$ を X の**確率分布** (probability distribution) と呼ぶ。単一変数 X の確率分布は、X を横軸にとり、$P(X = x)$ の値を縦軸にプロットしたグラフによって表すことができる。もちろん、我々はこのグラフを直接「観察」することはできない。統計学の主要な目的は、興味ある確率変数について、手持ちのデータからこのグラフ=確率分布について推測することである。

3.1　母数

我々は 1 節で、平均や標本分散などの統計量を学んだ。これらは得られたデータ（標本）を要約する指標であった。さて、いま我々が扱っているのは手元にある有限データではなく、それを取ってくる源としての標本空間とその上で定義された確率分布である。通常、この確率分布の全体像は我々には隠されている（だから我々は推論するのである）。しかし依然として、こうした「真なる」分布を要約する値を考えることはできるはずだ。このように、「真なる」（しかし我々には隠された）確率分布を特徴づける値を、**母数** (parameter) という。

代表的な母数が、母平均 (population mean) である[*3]

$$\mu = \sum_x xP(X = x)$$

[*3] ここで $\sum_x$ は X の可能な値すべてについての和を意図している。X が連続値を取るときは、和のかわりに積分をとり $\mu = \int_{-\infty}^{\infty} x \cdot P(X = x)dx$ となる。

つまり平均は、X のそれぞれの値 x にその確率 $P(X=x)$ を掛けたものを、全部足していくことで得られる。こうした母平均ないし平均は、確率分布の「中心」を与える。他方、分布のバラツキは母分散 (population variance) によって与えられる

$$\sigma^2 = \sum_x (x-\mu)^2 P(X=x)$$

これは言葉で説明すると、X のそれぞれの値 x について、平均からのズレ $(x-\mu)^2$ を計算して、それにその確率 $P(X=x)$ を掛けたものをすべて足したものである。

母平均、母分散はそれぞれ上でみた標本平均、標本分散の概念を標本空間全体に広げたものだと理解できる。母数は我々には不可知だが、対応する統計量をもとに推測することはできる。実際、我々は日本人全体の身長平均を知りたかったら、日本人からの有限サンプルをとってきて、その平均身長をもとに推測するだろう。つまり標本平均は母平均に対しての自然な推定量となっている。[*4]

3.2 さまざまな分布族

母数は分布の中心やバラツキを表す。しかしこれだけでなく、古典的な統計的推論では通常、分布が取る「カタチ」の種類に関しても何らかの想定をする。この分布の「カタチ」の種類を分布族と呼ぶ。関心ある確率変数の性質に応じて、どの分布族が想定できるかが変わってくる。いったん分布族を決めてしまえば、後は対応する母数を推定するだけで、分布が一意的に定まる。

■一様分布

ある確率変数 X が取りうる値 $x_1, x_2, \ldots$ にすべて同じ確率を割り当てる分布を一様分布 (uniform distribution) と呼ぶ。例えば公平なサイコロのそれぞれの目がでる確率は $P(X=x)=1/6$ の一様分布である。また X が α から β までの連続値を取る場合、その一様分布は

$$P(X=x) = \frac{1}{\beta-\alpha}$$

となり、母数 α, β から一律に決まる。

■ベルヌーイ分布

コインを 1 回投げる結果を X とし、裏がでることを $X=0$、表が出ることを $X=1$ で表す。表が出る確率を $P(X=1)=p$ とすると、X の分布は次の式で表せる

$$P(X=x) = p^x(1-p)^{1-x}$$

x は 0 か 1 であり、上式は x が 0（裏）のとき $1-p$、1（表）のとき p となることを確認せよ。このとき X はベルヌーイ分布 (Bernoulli distribution) に従うという。ベルヌーイ分布の平均は p、分散は $p(1-p)$ であり、つまり母数 p だけで決定される。

[*4] 母数の良い推定がどの統計量によって与えられるかは、必ずしも自明ではない。例えば、母分散の推定には上で見た標本分散ではなく、不偏分散 $\frac{1}{n-1}\sum_i^n (x_i-\bar{X})^2$ を用いる（理由は割愛するが、こうでないと推定にバイアスが入る）。ここから、こちらの不偏分散を「標本分散」として定義することもある。

■二項分布

次に同じコインを 1 回でなく、複数回、例えば 10 回連続して投げ、表が出た回数を記録する実験を考える。つまり X を 10 回の試行のうち表が出た回数とする。さて、X の分布はどのようなものだろうか。つまり、10 回投げて 1 回も表が出ない確率、1 回だけ表が出る確率、2 回出る確率・・・はそれぞれどれくらいだろうか。まず、$X = 0$ つまり表が一回も出ない確率は、

$$(\text{表が出る確率})^0(\text{裏が出る確率})^{10} = p^0(1-p)^{10} = (1-p)^{10}$$

である。次に $X = 1$ を考える。例えば 1 回目だけが表だった場合の確率は

$$(\text{表が出る確率})^1(\text{裏が出る確率})^9 = p^1(1-p)^9 = p(1-p)^9$$

である。これは 2 回目、3 回目、...10 回目だけが表である確率と同じだから、「10 回中 1 回だけ表が出る」確率は上の確率を全 10 ケース分かけて得られる。一般に $X = x$ の場合は、10 から x 個のものを選ぶ場合の数は ${}_{10}\mathrm{C}_x = \frac{10!}{x!(10-x)!}$ となることを考慮して、

$$P(X = x) = {}_{10}\mathrm{C}_x p^x (1-p)^{10-x}$$

さらに上式の 10 を n で置き換えれば、n 回のコイン投げで表が x 回でる確率が与えられる。

これを二項分布 (binomial distribution) という。X が二項分布に従うとき、その平均は np、分散は $np(1-p)$ で与えられる。つまり二項分布は個々の試行の確率 p と試行の回数 n によって決まる。

■正規分布

次に、コイン投げの回数 n をどんどん大きくして、無限回に近づけていってみよう。この膨大な試行においても、表が出る回数 X は上述の二項分布に従う。しかし n をどんどん大きくしていくと、その分布は次の正規分布に近づいていく

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

ただし平均 $\mu = np$、分散 $\sigma^2 = np(1-p)$ である。これを、X は正規分布に漸近的に従う、という。

正直、「は？なんで？」と思うかもしれないが、なるんだからしょうがない。そして実はこの正規分布、統計学において最も重要な分布である。というのも、自然界の多くの事象が、どうやら正規分布に従っているらしい、ということがわかっているからだ。例えば身長や体重などといった生物学的形質の多くは正規分布に従う。

そしてなにより正規分布がすごいのは、どのような分布に従う確率変数であろうと、それを複数回繰り返した和をとると、回を増せば増すほど和の分布は正規分布に近づいていく、という事実である。[*5]これを**中心極限定理** (central limit theorem) という。この定理は、統計学、とりわけ検定理論において極めて重要となってくる。

[*5] 実際のところ上で示したのは、二項分布はベルヌーイ試行の繰り返しの和の分布にすぎないのだから、ベルヌーイ分布に従う確率変数の和をとることによって、その和が漸近的に正規分布に近づいていく、ということである。

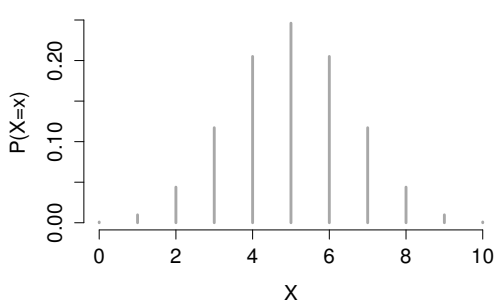

図 1　$p = 0.5, n = 10$ としたときの二項分布。横軸は表の回数、縦軸はその確率を表す。

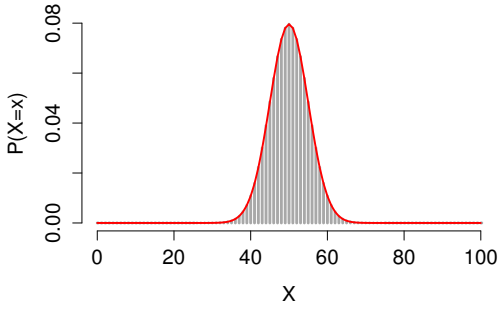

図 2　$p = 0.5, n = 100$ としたときの二項分布に、$\mu = 50, \sigma^2 = 25$ の正規分布（赤線）を重ねあわせたもの。

4　まとめ

まとめると、統計学には 2 つの層がある。1 つは得られたデータとそれを要約する記述統計の層であり、もう 1 つはそのデータの裏にある、直接的にはアクセス不可能な「真なる世界」における確率分布である。統計学の主眼は、得られたデータから、この真なる世界=確率分布について推測を行うことにある。しかし全くなんの前提なしでは歯がたたないので、通常はこの確率分布が何らかの分布族に従うと仮定して、対応する母数についての推論を行うことで「真なる世界」への接近を試みる。この推論の仕方、接近の仕方には様々なやり方があって、授業でも見るように、それがベイズ主義、統計的検定、モデル選択などといった種々の立場の違いとなって現れてくる。

練習問題

1. ある草野球チームの年齢と身長を調べたところ、次を得た。

背番号	1	2	3	4	5	6	7	8	9	10
身長 (X)	155	165	160	170	150	160	165	170	165	155
年齢 (Y)	13	15	14	15	12	13	14	15	15	12

   （a）このチームの身長の平均と分散、標準偏差を求めよ。
   （b）身長と年齢の相関係数を求め、それが正ないし負に相関しているか調べよ。
   （c）年齢が 1 歳上がると、平均して身長はどれだけ変化するか、求めよ。

2. 前述したように、相関係数は $[-1, 1]$ の範囲をとる。ある確率変数 X が与えられたとき、これとの相関係数が (a) 1 になる変数、(b) 0 になる変数、(c) -1 になる変数をそれぞれ求めよ。

3. ある自動車工場には二つのラインがあって、ライン A は工場全体の 6 割、ライン B は 4 割を生産する。ただしライン A は 1 割の確率で不良品が生じ、ライン B では 2 割の確率で不良品が生じる。この工場全体の不良品率はどれくらいか。（ヒント：全確率の定理を用いる）

4. 図 1, 2 を参考に、次の問に答えよ
   （a）公平なコインを 10 回投げたときに 7 回以上表が出る確率と、同じコインを 100 回投げたときに 70 回以上表がでる確率は、同じだろうか。違うとしたら、どちらが大きいか。
   （b）公平なコインを 10 回投げたときの表の数を X とする。「表がでる回数が 2 回より少ないか、あるいは 8 回より多い」という事象の確率を $P(\cdots)$ の形で表現せよ。また、この確率は 0.1 より大きいか、図 1 をもとに判断せよ。