Chapter 2

統計

2.1 測定量

最初の回の講義でも触れたように、実験的に求まる量はすべて測定量である。測定によって求められた測定量の値を測定値と呼ぶ。個数や金額など、離散的な値を持つ量は別として、長さ、体積、質量など連続的な量には必ず測定誤差がある。(cf. 有効数字) また、サイコロを振って出た目のように、測定するたびに値が変わってくるものもある。

今回の講義では、測定値がたくさんあったとき、どのようにしてその特徴を把握するか (=情報を得るか) を考える。つまり、統計の初歩である。数値がたくさんあって、それらから情報を取り出すという作業は、研究を進める上で必要になる場面がきわめて多いので、本章は将来必ずや復習が必要になる章であろう。

例えば、ある量を測定して次のような値が得られたとする。

$$76, 86, 77, 88, 78, 83, 86, 77, 74, 79, 82, 79, 80, 81, 78, 78, 73, 78, 81, 86, 71, 80, 81, 88,$$

$$82, 80, 80, 70, 77, 81$$

長さを測ってこれだけのバラツキがでたら問題であるが、例えば 10 分あたりの車の通行量などであればこれくらいのバラツキはむしろ少ないと思われる。

もし、**このような数値を分析せよ、この集合から何らかの情報を取り出せ**、と言われたらどうすればいいだろうか? 統計とは、このようにデータを様々な方法で眺めて情報を取り出す方法を研究する学問である。

さて、統計学では、測定して得られたデータのことを標本 (サンプル) と呼ぶ。サンプル数は標本の数のことを指す。ここで注意しなくてはならないのは、統計学においてはサンプルとはあくまでも「あらゆる可能性から選ばれた 1 つの事象」に過ぎないということである。つまり、我々の目の前にあって測定される (ある

いは見たり触ったりする) 対象はあくまで未知の値 (や未知の統計量) をもち、真の姿は決して測定されることがないという見方である。

目の前にある物体が未知、というと不思議な気もするが、サイコロを振って出た目に偏りがあるから「このサイコロはおかしい」と思うとか、今感じる気温から「今年は寒いね」というのが本当かどうか考える、というような場合をイメージするとわかりやすい。実際のところ、統計のルーツの一つは賭け事の必勝法の研究にあった。この場合は、目の前にある状況下で最も大きなポイントを得られる「可能性のある」手を決定するということになる。

ここで、重要な言葉を定義する。2 つの標本分布が「おなじもの」であるとは、それぞれの標本を取り出したそれぞれの母集団に関して、「同じ母集団である」ことをいう。つまり、標本自体をそれぞれ比べて「同じ」ということではなく (上にも書いたようにまずありえない) 同じ母集団か否かを考えるのである。同じ母集団とは、たとえば一つのサイコロから出た目であるとか、同じ性質を持った人の集まり (薬が効かなかった集団、あるいは効いた集団) などである。

2.1.1 乱数

ここで、やや順番は前後するが乱数というものを紹介する。乱数とは文字通り乱れた数のことで、予測ができないような数の並びのことである。サイコロを振って得られた数の並びも乱数である。

例えば、10 個の乱数を求めるには Octave 上では関数 rand() を呼び出す。

```
octave:27> A=rand(10,1)
A =

  0.225704
  0.018580
  0.818762
  0.634118
  0.026280
  0.980303
  0.014780
  0.477392
  0.636399
  0.382046
```

なお、乱数は同じ数列を二度と出力しないので、もし上の例にある値と違っても心配には及ばない。乱数を出力する関数 rand() の値域は (0,1) である。

整数の値の乱数が欲しいときは

```
octave:29> B=ceil(rand(10,1)*10)
B =

  8
  2
  9
  7
  3
  8
  3
  5
  9
  2
```

とする。ceil() は小数点以下を切り捨てる関数である。

与えられた数列が乱数であるかどうかは、分布をとるなどして調べる。

```
octave:30> hist(B)
```

で、ヒストグラムが別のウィンドウ上で出力される。分布の詳細については次回触れる。ここでは、分布のプロットがフラット (一様な分布) になっていることを確認できれば良いとする。**ただし、少ない数の乱数では必ずしもフラットにはならない。**サイコロでも連続して同じ目が出続けることがあるように、乱数が必ずきれいに分布した数を出すとは限らない。あくまで、多数のサンプルをとってはじめて判定できることである。

なお、ここでは乱数であることを知っているから、一様な分布がでることを仮定してよいが、実際のデータは、ランダムであるか否かは未知であり、サンプル数も自由に選ぶことができないことに注意すること。

練習問題

n 個の乱数の分布を求めるには、

```
hist(rand(n,1))
```

とする。様々な n の値に対して (hist(rand(100,1) など) 実行し、一つの n に関しても実行するたびに分布の形が変わることを確認せよ。n の値が大きくなるにつれて、フラットな分布になることを確認せよ。

2.1.2 平均

統計量のなかで最も身近なものは平均である。標本 $A = (a_1, a_2, ...a_n)$ の平均 μ は

$$\mu = (1/n)\sum_{i=1}^{n} a_i \tag{2.1}$$

と定義される。要するに、サンプルの和をとって、サンプル数で割ったものである。Octave では、mean() 関数を用いる。

```
Octave:> mean(A)
```

として実行する。

練習問題

- 皆で適当な数を 10 個挙げ、その平均を求めよ。
- 乱数 n 個の平均を求めよ。n を 10,100,1000,,,, と大きくすると、平均値はどうなるか？

さて、平均にどんな意味があるだろうか？「平均的サラリーマン」という言葉があるように、たぶんサンプルの「最もありそうな値」というのが「日常的な意味での平均」であろう。では、数学的な平均が、上の「日常的な意味での平均」と異なるような極端な場合を考えてみよう。

- 1 歳の子供と 80 歳の祖母:平均年齢 40.5。
- 月給 15 万円の社員 5 人と 200 万円の社長 1 人の会社:平均月給 45.8 万円。

つまり、平均値付近にサンプルの分布がない場合、「日常的な意味での平均」は意味を持たない。とくに、役員が多い会社の平均給与ほど当てにならないものはない。逆に、そのような場合は、**平均を出すことで人を欺く道具としても使える**ことに注意せよ。

2.1.3 分散

平均だけでは、特徴をうまくつかめないことがある。ある 2 つの集団の平均値が同じでも、上に挙げた極端な分布をしているような場合は、その 2 つの性質は全く異なるものかもしれない。

このような場合に問題となる、「ばらつき」を定量化する統計量が分散である。まず下の例題を考えよう。

例題

次の 2 組の数字の組の平均をそれぞれ計算せよ。これら 2 組は、同じもの、あるいは類似したものといえるか考えよ。

$$A = [10, 11, 10, 11, 10, 10, 11, 9, 11, 9, 11, 9, 9, 9, 11, 11, 12, 11, 11, 10]$$

$$B = [4, 6, 5, 5, 6, 6, 6, 4, 5, 5, 18, 17, 17, 16, 16, 16, 15, 15, 13, 11]$$

Octave では、A の平均は mean(A) という関数で求めることができる。さて、上の例では A と B の平均は同じである。では、おなじものとみなすことは出来るだろうか？

ここで、分散という量を定義する。標本 $A = (a_1, a_2, ...a_n)$ の分散 σ^2 は平均 μ を用いて、

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)^2 \tag{2.2}$$

である。これは、式を変形することにより

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(a_i^2 - 2a_i\mu + \mu^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(a_i^2) - \frac{1}{n}2\mu\sum_{i=1}^{n}a_i + \frac{1}{n}\cdot n\cdot\mu^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(a_i^2) - \mu^2
\end{aligned} \tag{2.3}$$

となる。「2 乗の平均マイナス平均の 2 乗」と覚えると良い。なお、Octave では var() という関数で分散を求められる。

練習問題

上で示した A,B の分散を計算せよ。A と B はおなじものと見なすことが出来るだろうか。

2.1.4 コイントス

有名な例であるコイントスを考えてみよう。コイントスとはその名の通りコインを投げて裏表のいずれかの標本を得るものである。ここでは、表を+1、裏を-1 として、その合計値をスコアとする。そのスコアの分布を考えてみたい。

なお、コイントスは独立事象 (以前でた値とは無関係) であるため、n 個のコインを同時に投げて裏表の数を数えるのと一つのコインを n 回投げて集計することは同じである。ただし、これは母集団が等しいということを示すのみであるから、もちろん、全く同じ値は再現しないということに注意すること。また、癖のあるコイン (裏表の出る確率に偏りがあるように細工されたコイン) の場合を除く。

さて、コイントスの分散を求めてみよう。Octave でコイントスを 1 セット行うプログラムは以下の通りである。

```
sum(round(rand(100,1))*2-1)
```

上のプログラムを 10 回実行し、平均と分散を求めよ。100 回ではどうか？

ちょっと賢いプログラム例:

```
#dist1.m
n=10000;
A=zeros(n,1);
for(i=1:n)
   A(i,1)=sum(round(rand(100,1))*2-1);
 endfor
mean(A)
var(A)
```

このプログラムを “dist1.m” として保存し、コマンドライン上から

```
Octave:> source("dist1.m");
```

と入力すれば実行可能になる。n をいろいろな値に変化させてみること。また、hist(A) によりヒストグラムを表示してみること。なお、ファイルのディレクトリに関してはクイックリファレンスを参照のこと。

なお、round() は小数点以下四捨五入を行う関数である。round(rand()) の出力は 0.5 を境にして 0 か 1 となる。[0,n-1] の整数の乱数を求めるテクニックとして、round(rand()*n) というものがある。同様なもので切り上げを行うものを ceil(), 切り捨てを行うものを floor() というものがある。これを応用して、サイコロを作るときは ceil(rand(n,1)*6) とするとよい。(n は求めたい試行回数)

コイントスに話題を戻そう。n を大きくするにつれ釣り鐘状の一山の分布が現れただろうか? これは正規分布 (ガウス分布とも呼ばれる) と呼ばれるもので、非常に多くの偏りのないランダムな現象にみられている (正確には、$n \to \infty$ の極限で正規分布になる)。このことは有限の平均と分散が存在するいかなる分布をもつ

母集団から得られた標本でも、その「標本平均の分布」が正規分布になるという「中心極限定理」により証明されている。詳しくは 2.3.7 章を参照されたい。

なお、コントスでは「平均」ではなく「和」を用いたが、回数が固定されているため、意味は変わらない。また、正規分布に近づくのは標本平均の平均と分散であり、分散に関しては、平均を取らないで求めた分布に関しては、母集団の性質を反映していることに注意。

2.1.5 関連する話題:場合の数とくじ引きの確率

コイントスと関連して、くじ引きに関して述べておこう。コイントスは以前出た値と無関係であるため、常に公平のように思えるが、一人一人順番に引くくじ引きは、もしかしたら不公平だと思う人がいるかもしれない。

例えば、4 人で引くくじで、一人だけ当たりがあるくじを順番に引くとする。引く順番と当たりの確率に関係があるだろうか？全員で一斉に引いたり、全員が引き終わるまでくじを見ないことにすると、何となく公平さが増すように思えないだろうか？しかし、ここではくじは順番に公開され、誰かが当たった時点でくじ引きを終了するものとする。

一人目が当たりを引く確率は、４つあるうちから一つだけ選ぶのであるから

$$\frac{1}{4}$$

である。次に、二人目の場合、既に一人目で終わっている可能性が 1/4。つまり、二人目まで回ってくる確率が 3/4。残り３つの中に１つだけ当たりがあるから、

$$(1-\frac{1}{4})\cdot(\frac{1}{3}) = \frac{3}{4}\cdot\frac{1}{3} = \frac{1}{4}$$

三人目の場合、一人目または二人目で当たりが出ている確率は $1/4+1/4$。のこり 2 つの中に 1 つだけ当たりがある。

$$(1-(\frac{1}{4}+\frac{1}{4}))\cdot\frac{1}{2} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

四人目は引く必要がない。誰も当たってなければ自動的に当たりである。誰かが当たっている確率は、$1/4+1/4+1/4$ であるから、

$$(1-(\frac{1}{4}+\frac{1}{4}+\frac{1}{4})) = \frac{1}{4}$$

である。

なお、確率の計算では、「かつ/and」の関係がある場合は掛け算、「または/or」の関係がある場合は足し算になる。「それ以外」のばあいは 1 からその確率を引けば良い。

2.2 レポート問題3

以下の問題から 2 問以上を選択して解答すること。

1. サイコロを用意し、1000 回振ってその目を記録せよ (時間がなければ回数を減らしてもよい。最低 100 回は振ること)。分布はどうなるか？また、サンプル数ごと (100、200、300 等) の比較をせよ。可能であれば次の問題も解答し、結果を比較せよ。

2. 自分でサイコロのふりをして、1 から 6 までの数を 100 個以上可能な限り記録し、分布を求めよ。友人等にも同様に数をつくってもらい、複数人の比較ができるとなお良い。

3. (理系向き) 計算機で乱数が生成できるのは何故か?あるいは、計算機で生成された乱数は信用できるか?(ヒント:「疑似乱数」を調べよ)

2.3 確率分布

前回までは、標本をそのまま扱ってきた。しかし、人間が把握できる数とその量には限界がある。そのため、把握できないくらい多くの数をひと塊として扱うために、確率分布およぶ分布関数という概念を導入する。確率分布とは、ある確率変数の値に対応する起こりやすさを関数の形で表現したものである。これを確率分布関数と呼ぶ。この講義では、正規分布を最低限理解して欲しい。いきなり関数をみても理解しがたいので、まず標本の分布から考えることにする。

2.3.1 頻度分布

まず、標本の分布を調べることからはじめる。すでに何度かOctaveでhist()という関数を使用した。この関数が出力するグラフが頻度分布(ヒストグラム histogram)と呼ばれるものである。これは、たくさんある標本がそれぞれ短冊状に並んでいる長方形の幅のなかにどれだけの個数入っているかを高さとして表示したものである。

しかし、何でもヒストグラムをプロットすればいいというものではない。Octaveでは、hist()にスカラー量の引数(「()」の中にいれる文字)を与えることで、ヒストグラムの間隔(=短冊の幅)を調整できる。正規分布の乱数を作る関数randn()を用いて、以下のような分布を作る。つぎのOctaveプログラムを入力せよ。
行末に「;」セミコロンがある行はセミコロンを入れることを忘れないこと。忘れると、コンソールが大量の数字で溢れてしまう。(Octaveは一行ずつ実行の度に結果を出力するが、行末に「；」セミコロンがあると出力をしない。)

```
A= randn(50,1);
hist(A)
hist(A,100)
```

短冊の数がサンプル数と同程度になるまで間隔を狭めると、何も読み取れない状態になってしまう。

```
B1= randn(10000,1)+2.4;
B2= randn(10000,1);
B=[B1;B2];
hist(B)
hist(B,100)
```

最初のhist(B)は10個の区切りであるから、一山に見えるであろう。次のhist(B,1000)は100個に区切ってあるために2つの山が見える。これは間隔が大きすぎる場合である。つまり、適切な区切りを入れてないと、結論を間違う可能性がある。

2.3.2 一様な分布

サンプルがあらゆる値を等しい確率でとりうるときを一様分布という。一様乱数 rand() は (0,1) の区間内の数を等確率で出力するので、多く集めて分布をとれば、フラットな分布が表れる。

練習問題

Octave で rand() を 100 個、1000 個、10000 個と求め、分布がどうなるか確認せよ。

2.3.3 確率

この講義では、これまで確率を紹介せずにきた。確率とは、ある事象がおきる可能性の度合いのことである。「明日何が起きるか」というような漠然としたことがらではなく、「サイコロを振って 4 の目が出る確率は 1/6 である」とか、**おきうる事象がそれぞれ明確に定義され、確実に事象が確認される**場合で定義される。

ある分布 $f(x)$ が与えられ、それが (本来は測定できないが) 母集団の真の分布であると考えたとき、その全面積に対する $f(x)$ の値の比が、事象 x が起きる確率である。ただし全面積は

$$\int_{-\infty}^{\infty} f(x)dx \tag{2.4}$$

によって求められる。なお、一般に「確率分布」と呼ばれる関数は上の積分が 1 になるようになっている。これを規格化と呼ぶ。

規格化がなされた確率分布では、求める事象の起きる確率を求めるには、同様に確率を積分によって求める。事象を $x \in [a, b]$ とすると

$$\int_{a}^{b} f(x)dx \tag{2.5}$$

により求められる。

なお、微分積分を履修しなかった、ないし忘却してしまった人は、ある関数のグラフを思い浮かべて、それと x 軸で囲まれた領域の面積が積分であると理解していれば良い。

フラットな分布の場合は、すべての値が等確率で表れる可能性を持つ。つまり、標本数が無限に大きくなれば、水平な直線が表れる。このように、確率分布は、標本数が無限に大きくなった場合に表れる分布の形を関数として表したものである。なお、標本数無限としたため、もはや関数に与える変数 x は標本ではない。この変数を確率変数と呼ぶ。

練習問題

[0,1] の区間の数を与える一様乱数が $0.3 \leq x < 0.4$ となる確率を求めよ。ただし $f(x) = 0.1$ である。

2.3.4 平均と分散

確率分布をもとにすると、確率変数の平均 $E(X)$ と分散 $V(X)$ が計算できる。平均は μ, 分散は σ^2 として表記されることも多い。

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{2.6}$$

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \tag{2.7}$$

なお、$\sigma = \sqrt{V(X)}$ は標準偏差と呼ばれる。式 2.13 と同様な変形により

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - E(X)^2 \end{aligned} \tag{2.8}$$

となる。ただし、定義 $\mu = E(X)$ および

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

を用いた。

以上の定義は連続分布のものである。では、離散分布の場合をサイコロを例として考えてみよう。確率変数 X をサイコロの出た目と考える。このとき、X のとりうる値は $X = 1, 2, 3, 4, 5, 6$ である。理想的なサイコロであれば、どの目も同じ確率で出現するから、分布確率は $f(X) = 1/6$ である。

X の平均 $E(X)$ を求めてみよう。**確率変数の平均値は値の和ではなく、それぞれの値と確率の積で与えられる**

$$\begin{aligned} E(X) &= \sum_{k=1}^{6} k \cdot f(k) \\ &= (\frac{1}{6} \times 1) + (\frac{1}{6} \times 2) + (\frac{1}{6} \times 3) + (\frac{1}{6} \times 4) + (\frac{1}{6} \times 5) + (\frac{1}{6} \times 6) \\ &= \frac{1}{6}(1+2+3+4+5+6) \\ &= \end{aligned} \tag{2.9}$$

である。一般に、一様分布の場合、1,2,...,n の値をとりうる確率変数の平均は

$$E(X) = \frac{n+1}{2} \tag{2.10}$$

である。

次に分散 $V(X)$ を考えてみよう。式 2.13 でみたように、X^2 の平均から $(E(X)^2)$ を引けば良い。公式

$$\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1) \tag{2.11}$$

を用いて一般解から先に求める。

$$\begin{aligned} E(X^2) &= \frac{1}{n}\sum_{k=1}^{n} k^2 \\ &= \frac{1}{n}\cdot\frac{1}{6}n(n+1)(2n+1) \\ &= \frac{1}{6}(n+1)(2n+1) \end{aligned} \tag{2.12}$$

であるから、

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 && (2.13) \\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2 \\ &= \frac{1}{12}(n^2-1) && (2.14) \end{aligned}$$

となる。$n = 6$ を代入すると「　　」であることがわかる。

練習問題

1 から 10 までの整数値を出す乱数の平均と分散を求めよ。

2.3.5 2 項分布

これまでみてきたように、コイントスの和 (平均) の分布は正規分布 (詳細は 2.3.7 章を参照) であった。しかしコイントスの持つ性質はこれだけではない。コイントスの確率分布としての側面を、もう少し詳しく見てみよう。即ちコインを投げて、何回表がでたかという確率を考える。

表の出る確率 p、裏の出る確率 $q = (1-p)$ のコインを n 回投げ、表の出る回数を k とすると、その分布は 2 項分布と呼ばれ、b(k;n,p) と表される。

$$b(k;n,p) = \binom{n}{k} p^k q^{n-k} \tag{2.15}$$

ここで、

$$\binom{n}{k} = {}_nC_k = \frac{n!}{k!(n-k)!}$$

は 2 項係数と呼ばれ、「n 個から k 個選ぶ場合の数」となる。

Octave には 2 項分布を求める関数が内蔵されている。

```
n=10;
p=0.5;
x=1:n;
pd=binomial_pdf(x,n,p);
plot(pd);
```

というプログラムで、2 項分布の確率分布を表示することができる。3 行目の $1:n$ は、1 から n までの、整数値の行ベクトルを作る関数である。(各行の動作をみるには、セミコロン「；」をとって実行してみるといい)

また、binomial_cdf(x,n,p) により確率分布関数の積分値を求めることができる。上のプログラムで、binomial_cdf() を置き換えて実行してみるといい。

詳細は省くが、2 項分布の平均値は np、分散は npq となる。

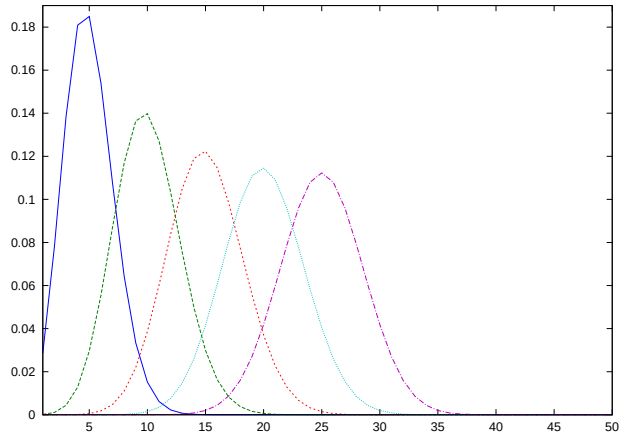

Figure 2.1: 2 項分布の確率密度関数。n=50, 左から p=0.1,0.2, 0.3, 0.4, 0.5

練習問題

不良品率が 5% のとき、100 個の製品のうち不良品が 10 個以内に収まる確率を求めよ。また、95%の確率で、何個以下の不良品を覚悟しなくてはならないか。

2.3.6 ポアソン分布

コイントスのように、n が少ない対象に関する分布は 2 項分布により表現できる。しかし、n が非常に大きい場合はポアソン分布を用いる。また、電話の発呼や、放射性同位体の崩壊 (放射能を持つ物質からの放射線の発生) など、「離散事象であり、発生する時刻がランダム」であると考えられている現象は「ある時間の間に、その事象が発生した回数 x」がポアソン (Poisson) 分布に従うことが知られている。

ポアソン分布はパラメータ λ により $p(x, \lambda)$ と表され、

$$p(x, \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{2.16}$$

である。平均 λ、分散 λ である。なお、n を大きくしたときに ($\lambda = np$ という条件で)2 項分布と一致する。「ある単位時間内に、平均 λ の事象が x 回おきる」確率が、ポアソン分布で与えられるのである。ただし、現在では計算機により計算できるので、こういう近似ができる、と覚えておけばよい。

Octave では、同様に poisson_pdf(x,lambda),poisson_cdf(x,lambda) で確率分布と確率分布の積分値を求めることができる。

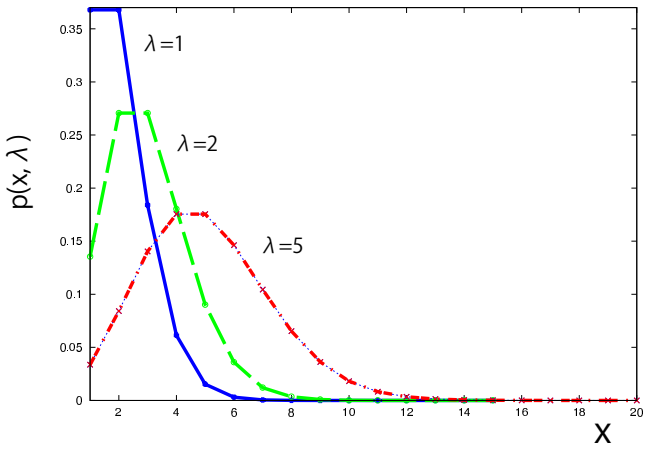

Figure 2.2: ポアソン分布。λ=1,2,5 の場合をプロットした。

練習問題

ある橋には、1 分間に平均 6 台の車が通行するとする。

1. 30 秒間一台も通らない確率を求めよ。
2. 1 分間に 6 台以上通る確率を求めよ。

3. 1時間当たりの平均の収入を求めよ。1台当たりの通行料は100円とする。

2.3.7 正規分布

上でも一度紹介した正規分布は、$N(\mu, \sigma^2)$ と表記される。μ、σ^2 はそれぞれ平均と分散である。正規分布は、変数 x に対して以下の関数で示される

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.17}$$

グラフが指数関数になっていることに注意してほしい。これは、平均から離れるに従って急激に減少することを意味する。この分布は、独立で同じ確率分布をもつ変数の標本値の和を、標本数を無限に大きくしたときにもたらされる分布である。n が大きいとき、2 項分布は正規分布に近くなる。さらに、正規分布では x は連続関数であるために使いやすい。

正規分布のグラフは以下のようになっている。

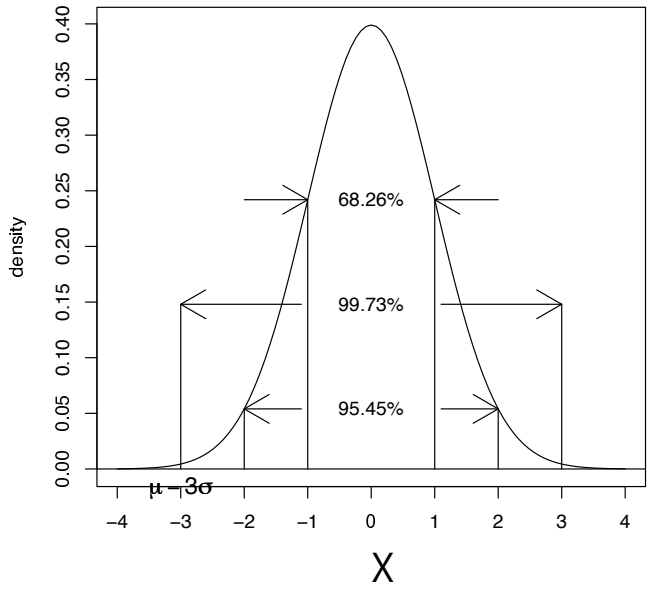

Figure 2.3: 正規分布

では、これが正しいとすると、数学の試験の点数の頻度分布などではありがちな山が 2 つある分布とかは絶対に観測されないのだろうか？実際には観測されることがある。しかし、この理論はそんなことには全く構わない。それには以下の理由が考えられる。

- 標本数が充分でない

- 異なる確率分布を持つ変数を混ぜて観測している (確率変数に独立性がない)

後者に関しては統計的検定により判定することが出来る。後者では判定ができなかった場合は、前者の理由であるか、また分布が本当に特殊な場合である。

正規分布の確率密度関数は normal_pdf(X,m,v), normal_cdf(X,m,v) により与えられる。X は値、m,v はそれぞれ平均と分散である。

中心極限定理

さて、なぜランダムな現象の標本値の和の分布は正規分布に近づくのだろうか? このことは、中心極限定理から導かれる。この定理は、「お互いに独立で同じ分布をもつ n 個の標本の平均値の分布は、$n \to \infty$ の極限で正規分布に近づく」というものである。ここでは、標本の真の分布が正規分布である必要はなく、平均と分散が有限の値であることが要求されているだけである。つまり、1 セットのコイントスは、一様分布の確率変数の n=100 の場合の和を求めたことになる (個数が固定なので、平均値をとっても値が 1/100 になるだけである)。その和 (平均の 100 倍) の分布が正規分布になることは、まさしく中心極限定理を表現したものである。

練習問題

- Octave で正規分布の乱数を出力する関数 randn() を用いて、この平均と分散を求めよ。
- normal_cdf(x,m,v) を用いて、平均 0, 分散 1 の標準正規分布で x=-1, 1.5, 2.0 においてそれぞれ上位何%に位置するか求めよ。
- normal_cdf(x,m,v) を用いて、平均 0, 分散 1 の標準正規分布で x= [0.3, 0.5] の区間には何%の分布があるか求めよ。
- normal_cdf(x,m,v) を用いて、平均 0, 分散 1 の標準正規分布で x= [1.0, 1.2] の区間には何%の分布があるか求めよ。

なお、ここで紹介した様々な分布の間の関係は、以下の URL にまとめられている。

`http://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/toukei.htm`

2.3.8 偏差値

偏差値とは学生諸子には耳の痛い言葉であったであろう。まず、偏差値 Z の定義は以下のように与えられる。

$$Z = \frac{X - \mu}{\sigma} \times 10 + 50 \tag{2.18}$$

μ は平均、σ は分散の平方根で、これを「標準偏差」と呼ぶ。10 倍したり、50 を足しているのは 30-70 という、人間が把握しやすいスケールの範囲に変換しているのに過ぎない。**変数 X の値が σ だけ平均より上であれば偏差値 Z は 10 上がる。**

この意味は、図 2.3 を見るとわかりやすい。偏差値 40〜60 の範囲には全体の 68.26%が属する一方、70 以上には「　　」の割合しかないということを意味する。つまり、測定してえられた分布を一度 ($\mu = 0, \sigma^2 = 1$) に押し込め (下図)、それからの偏差を分散を基準として求めているのである。

ただし、様々な理由により、測定値はきれいな正規分布をしているとは限らない。偏差値はあくまでも目安である。

練習問題

- normal_cdf(x,m,v) を用いて、平均 0, 分散 1 の標準正規分布で 偏差値 x=40, 55, 70 においてそれぞれ上位何%に位置するか求めよ。

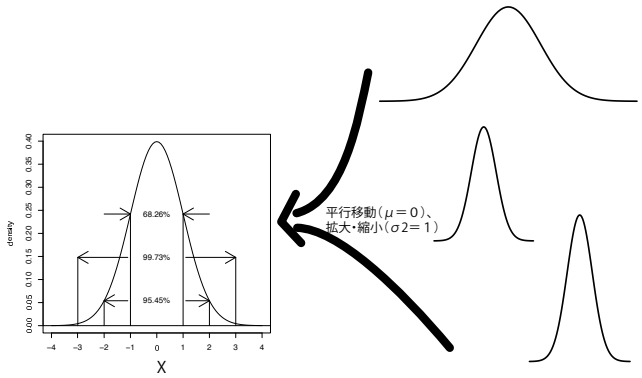

Figure 2.4: 正規分布の N(0,1) への変換のイメージ。

2.3.9 べき分布

世界のあらゆるものが正規分布により支配されているかというと、必ずしもそうではない。例えばジップ (Zipf) 則として知られている、双曲線的な分布がその 1 例である。$1/x(x^{-1})$ を代表とした、$x^{-\alpha}$ 的な分布が、自然現象や人間の介在する現象にも多く見いだされている。

2.4 レポート問題 4

次の問題から 2 問以上選んで解答せよ。

1. 正 12 面体のサイコロの出た目を確率変数 X とする。X の平均と分散を求めよ。
2. ある橋の上で記念写真の撮影をしたい。そのためには 1 分間車がない時間が必要とする。任意の 1 分間を選び、85%以上の確率で成功させるためには、橋の交通量は 1 時間当たり何台以下でなくてはならないか？ 95%以上の確率の場合はどうか？
3. 撮影のやり方を変えて半分ずつ記念写真を撮ることにした。2 回に分けて、各 30 秒で撮影が可能になった。 3 分間のうち車のこない 30 秒間が 2 回必要とする。成功確率は同じ (85%) として、交通量が多くてもチャンスがあるだろうか？その理由も述べよ。(簡単のため、3 分間を 30 秒 x6 個に区切って考えるとよい)
4. 10000 人が参加したテストで、あなたの得点のスコアは偏差値 68 だった。上位何番以内にいるだろうか？
5. ベキ分布のあらわれる現象を調べよ。提出された中で、なるべく珍しいものを高く評価する。
6. (理系向け) ベキ分布の現れる原因を考察せよ。一般ではなく、特殊な場合についてでよい。

2.5 統計的検定

測定により求まった 2 つの分布が、異なる原因によるものか、それとも偶然の一致かと言うことを判定するのが統計的検定である。生物学や物理学の実験的研究や、社会調査などでは必須の技術である。

2.5.1 1 変数の場合

まず、前回の講義で触れた正規分布を思い出そう。確率変数 X の観測値 x(大文字小文字に注意) が、平均プラスマイナス 2σ に収まる確率は約 0.95 である。(この 0.95 という数字の根拠は便宜的なものである。95%の雨の確率で傘を忘れる人はほとんどいないであろう。というくらいの感覚である。) **実のところ、偏差値と考え方は同じである。**偏差値と同様に、Z スコアというものを導入する。

$$Z = \frac{(X - \mu)}{\sigma} \tag{2.19}$$

X は確率変数、μ, σ はそれぞれ平均と標準偏差である。**ただし，母集団の平均と分散は実世界では不明である。**これは、サンプルの平均と分散で「代用」する。この Z の値が [-2,2] の区間にあるか否かで、X の測定値 x が μ, σ で規定されている分布に属している「確からしさ」を求めることができる。これが、**予め決めた信頼区間 95%の内外にあるか否か**で検定することである。

検定とは、例えばこういうことである。

> ある子供について、ある科目についてテストで平均で 60 点 (分散 100) 程度しかとっていなかったのに、今回のテストで 95 点を叩き出した。これはその子供にとって「偶然」の出来事であるか否か？

その子ががんばったからなのか、それとも不正をしたのか、あるいは単にテストが易しすぎたのかは、統計学は答えてくれない。すくなくとも偶然の出来事ではないらしいことを示してくれるのみである。

このように、統計的検定とは、ある事柄が理由のある現象か、全くの偶然の出来事かを判断するものである。それぞれを仮説と呼ぶ。つまり、「この高得点は偶然の出来事ではない」というのが検証される仮説で**「対立仮説」**と呼ぶ。、偶然性を主張するのは**「帰無仮説」**と呼ばれる。また、**検定とは、帰無仮説が成立するか否か (真か偽か) を判定して棄却するか否かを決定するものである。**

2.5.2 検定のエラー

統計的検定はあくまでも「確からしさ」を与えるのみであり、何かを断定するのはあくまでも人間の主観である。また、与える仮説 (後述) によってその確からしさの検出能力が異なる。上の例では、「偶然か否か」を判定したのみであり、「が

んばったか」「不正をしたか」「試験問題が易しすぎたか」検定のエラーに関しては下の 2 通り考えられる。

タイプ I エラー 帰無仮説が真なのにも関わらず棄却され、対立仮説を採択する。false positive, 偽陽性ともいう。

タイプ II エラー 帰無仮説が偽なのにも関わらず採択され、対立仮説を棄却する。false negative, 偽陰性ともいう。

ただし、真偽はあくまでも統計の外の話である。つまり、薬の臨床検査で、効かない薬を誤って効くとしてしまったのを、タイプ I エラーと呼ぶ。

仮説の作り方が悪いと、このようなエラーを良く招く。また、検定の結果と結論にも論理の飛躍があることがある。上の試験の点数の例では、あくまでも「点数が偶然でなく高い」ことがわかっただけなのに、一足飛びにその「原因」のみ考えてしまいがちである。子供の不正を疑うよりは、まず努力したか否か、あるいは「先生の教え方が良くなった」という仮説を立ててみるべきである。もちろん保護者は子供の普段の行動も考慮の基準に入れている。ただし、そういうものは数量化が難しい。**数量化可能なデータから無理なく推論できる仮説を設定すること**こそが研究遂行に必要な技術である。

2.5.3 仮説検定のポイント

ここで仮説検定の考え方のポイントをまとめておこう。

- 仮説検定は、得られた 2 つの標本集団がおなじ母集団から得られたもの (帰無仮説) か否か (対立仮説) を検定する
- 危険率とは、「同じ母集団から得られた可能性」を示す
- 帰無仮説が棄却出来ないことは、たとえ 2 者の平均や分散が大きく見えても、それはサイコロの目と同様に偶然得られた結果にすぎない
- 対立仮説が成立するか否かを積極的に示すことはない
- 対立仮説の解釈は、実験計画に依存する。つまり検定以前に YES/NO で簡単に判定出来る程度にまで論理を構築しておかねばならない

実のところ、研究において重要なのは、最後に述べたように、実験計画および仮説のデザインである。間違った実験計画からは無意味な検定結果しか得られない。

2.5.4 t 検定

一般には、母集団の平均も分散も未知であるし、正規分布をしているという保証もない。仕方がないので、正規分布であるという仮定の下に検定を行うことがよくある。これをパラメトリック検定法と呼ぶ。パラメトリック検定法のうち、最もよく使われる t 検定を紹介する。なお、「仕方がない」とは書いたが、正規分布であるか否かを判定する検定方法もある。

t 検定とは、2 つの標本の集団が、同じ平均値を持つ母集団に属するか否かを検定するものである。つまり、「2 つの集団は同じ母集団からのサンプリングである」という帰無仮説を検証するものである。対立仮説は「平均が等しくない」(「大きい」か「小さい」の場合も可能) となる。つまり、2 つのデータの集団の平均値が違ったとして、これはサンプリングが少ないからたまたま起こった現象なのか否かを判定するものである。

例えば、ある薬の効果を調べる実験を考えよう。薬を与えた集団と、与えない集団 (見かけは同じで薬効のない砂糖のかたまりなどを与えたりする) の、ある検査でのデータを t 検定にかけて、帰無仮説が棄却された場合は効果があると判断することができる。

さらに、t 検定は標本の数が少ない場合にも対応が可能である。なお、ここでは標本の数のことを「自由度」と呼ぶ。これは、t 検定を定義づけたゴセット (Gosset) がビールの醸造技術者だったことによる。ビールの品質の分析には時間がかかり、標本の数が限られているからである。つまり、彼は自分の仕事のためにこの t 検定というものを作ったのである。

t 検定では、t スコアを以下のように定義する。$\bar{X}$ は標本平均、σ を標本の標準偏差とすると、

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{2.20}$$

である。Z スコアと比較すると、母集団の平均、分散が標本のものに置き換えられていることがわかる。t 分布は、正規分布の曲線ではなく、**自由度によって異なる** t 分布と呼ばれる分布関数である。これは複雑な関数であり、統計の本には付録として t スコアとその分布が表になって載っている。それをもとに、信頼区間の中に収まっているか否かを判定できる。現在では Octave をはじめ、計算機の数値計算アプリケーションソフトウェアに表が内蔵されている。

2.5.5 一般の場合

解こうとしている問題に応じて、また、仮説によっても使う分布は異なる。しかし、一般的に言えるのは、統計量 (スコア) が累積確率で信頼区間 (95%など) の中に収まっているか否か、を判定すればよいということである。他に代表的なものは

- カイ 2 乗検定 (χ^2 検定)
  想定した分布との差を求め、想定した分布と同じ分布か否かを調べる。つまり標本が正規分布か否かを調べることもできる。

- F 検定
  正規分布に従う 2 群の分散が等しいか否かを調べる検定。(上で紹介したのは等しい場合)。

などが挙げられる。それぞれの分布関数は計算が複雑なので、統計の本にたいてい付録として与えられている表か計算機を使うとよい。これらの検定は Octave に含まれている。

2.5.6 χ^2 検定

t 検定は値の分布に関する検定であったが、サイコロやカテゴリデータのような、多項の変数の標本の分布の検定をするのには χ^2(カイ 2 乗) 検定を用いる。

χ^2 検定は、標本の期待値からのずれが有意か否かを検定するものである。たとえば、サイコロの出た目に偏りがあるか、2 つの集団のアンケート結果に差があるか、といったことの検定に使える。

χ^2 スコアは以下のようにして求める。2 つの分布である標本 A(度数:$a_0, a_1, ...a_k$, 大きさ (度数の総和)n) と期待値 P(確率:$p_0, p_1, ..., p_k$) が与えられたとき、スコア χ^2 は自由度 (k-1) の χ^2 分布に従うことがわかっている。

スコア χ^2 は以下のようにして与えられる。

$$\chi^2 = \sum_{i=1}^{n} \frac{(a_i - np_i)^2}{np_i} \tag{2.21}$$

本来 χ^2 検定は B に当たる標本は「(推定された) 母集団の期待値」として定義されている。これを両方とも標本 (B とする) として扱う場合は、期待値を

$$\hat{p}_i = \frac{a_i + b_i}{n_a + n_b} \tag{2.22}$$

とする。ただし、n_a、n_b は A,B それぞれの標本数の大きさである。

では、サイコロを振った場合の数を求めて、それが正しいサイコロであったか否かを検定してみよう。サイコロの場合では、P は [1/6,1/6,1/6,1/6,1/6,1/6] である。

Octave では、サイコロを検定する場合の χ^2 スコアは以下のようにして求めることができる。ただし、ここでは行ベクトルを想定している。 χ^2 スコアを検定するのは、分布表を用いるが、Octave では p 値を求めるのに (1-chisquare_cdf(X,k)) という関数を用いる。ここで X はスコア、k は自由度である。ここでは、以下の関数を定義しよう。 “chiSQ.m” という名前で保存する。

```
#chiSQ.m
function x = chisq(A)
  length=size(A)(1,2);
  t0=round(sum(A)/length);
  x=0;
  for(i=1:length)
    x=x+ (A(1,i)-t0)^2/t0;
  endfor
endfunction
```

```
function p = doChiSQTest(A)
  k=size(A)(1,2)-1;
  p = 1-chisquare_cdf(chisq(A),k);
endfunction
```

ある学生が求めた、サイコロを 100 回、200 回、300 回振って求めた度数分布をそれぞれ D1,D2,D3 とする。

```
D1=[11,20,9,16,19,25];
D2=[26,38,22,34,37,43];
D3=[44,52,39,56,53,56];
```

まず、χ^2 スコアを求めてみよう。

```
Octave:134> source("chiSQ.m");
octave:135> chisq(D1)
ans = 10.471
octave:141> chisq(D2)
ans = 9.4545
octave:142> chisq(D3)
ans = 4.8400
```

ちなみに、k=5 のときの 95%信頼区間は

```
octave:138> chisquare_inv(0.95,5)
ans = 11.070
```

である。χ^2 スコアは単調増加であることを考えると、スコアが 11.07 より小さければ、帰無仮説 (P:等確率) が「棄却されない」。そのため、すべての標本は「等確率のサイコロから得られた結果と等しい」ことになる。

では、検定をしてみよう。

```
octave:143> doChiSQTest(D1)
ans = 0.062948
octave:144> doChiSQTest(D2)
ans = 0.092251
octave:145> doChiSQTest(D3)
ans = 0.43572
```

すべて p 値が 0.05 より「大きい」ため、正しいサイコロであったと考えられる。なお、差が「ある」ということを主張した場合は p 値が 0.05 より小さければよい。よく、論文で有意差をしめすのに (p<0.05) などという表現があるのはこのためである。

2.5.7 統計的検定:練習問題

Octave では、t 検定を t_test(x,m,alt), t_test_2(x,y,alt) という関数を用いて実行し、p-value と呼ばれる「帰無仮説が成立する確からしさ」を求めることができる。ただし、x,y はベクトル (標本)、m は平均値、alt は対立仮説で、平均が等しくないとする場合は ” <> ”(省略可)、$\mu_x < \mu_y$ のとき ” < ”、$\mu_x > \mu_y$ のとき ” > ” を文字列として代入する。

正規乱数を用いて、平均値に関する検定をしてみよう。例えば、

```
A1=randn(100,1)+10;
A2=randn(200,1);
```

として 2 つのベクトルを作る。

以下は実行例である。

```
octave:173> t_test(A1,10)
  pval: 0.779652
ans = 0.77965
octave:174> t_test(A1,9)
  pval: 0
ans = 0
octave:175> t_test(A1,9)
  pval: 0
ans = 0
octave:176> t_test(A1,11)
  pval: 1.13369e-19
ans =  1.1337e-19
```

2 つのベクトルの比較は以下の通りである。

```
octave:177> t_test_2(A1,A2)
  pval: 0
ans = 0
octave:178> t_test_2(A1,A2,">")
  pval: 0
ans = 0
octave:179> t_test_2(A1,A2,"<")
  pval: 1
ans = 1
octave:180> mean(A1)
ans = 9.9747
```

```
octave:181> mean(A2)
ans = 0.085481
```

パラメーターを様々に変えて、平均値がどれだけ検定されうるか調べてみよう。

```
octave:183> A3=randn(10,1)+4;
octave:184> t_test(A3,3)
  pval: 0.0177336
ans = 0.017734
octave:185> t_test(A3,3.5)
  pval: 0.203474
ans = 0.20347
octave:186> t_test(A3,3.8)
  pval: 0.658434
ans = 0.65843
octave:187> t_test(A3,4.1)
  pval: 0.65837
ans = 0.65837
octave:188> t_test(A3,4.3)
  pval: 0.313878
ans = 0.31388
octave:189> t_test(A3,4.5)
  pval: 0.128008
ans = 0.12801
octave:190> t_test(A3,4.6)
  pval: 0.0789323
ans = 0.078932
octave:191> t_test(A3,5)
  pval: 0.0108304
ans = 0.010830
```

2.5.8 レポート問題 5

気象庁の日本の平均気温の平年差からの差が以下の URL から取得できる。

http://www.data.kishou.go.jp/climate/cpdinfo/temp/list/an_jpn.html

1. 1995 年以降とそれ以前では、平均気温からの差の平均値が等しいと言えるか？ t 検定により判定せよ。
2. 10 年ごとにグループ分けし、それぞれについて平均と分散を求め、前後の 10 年と平均気温からの差が等しいか調べよ。t 検定により判定せよ。

3. その他、どのような特徴がこの数値から見いだせるか、考えることを述べよ。

2.5.9 特別レポート問題

注意:このレポート問題は最後の授業終了時まで、随時受け付ける。複数回提出可。

新聞・雑誌・ウェブサイト上の記事や広告など、一般の人の注目度が高いメディアにおいて、統計が不適切に使われていると思えるものを指摘せよ。理由を簡潔に示すこと。また、オリジナルの記事等のコピーを添付のこと。なるべく最近の記事が望ましい。

参考文献

ダレル・ハフ「統計でウソをつく法」講談社ブルーバックス
谷岡一郎「『社会調査』のウソ―リサーチ・リテラシーのすすめ」 文春新書
S.D. レヴィット、S.J. ダブナー「ヤバい経済学–悪ガキ教授が世の裏側を探検する」 東洋経済新報社